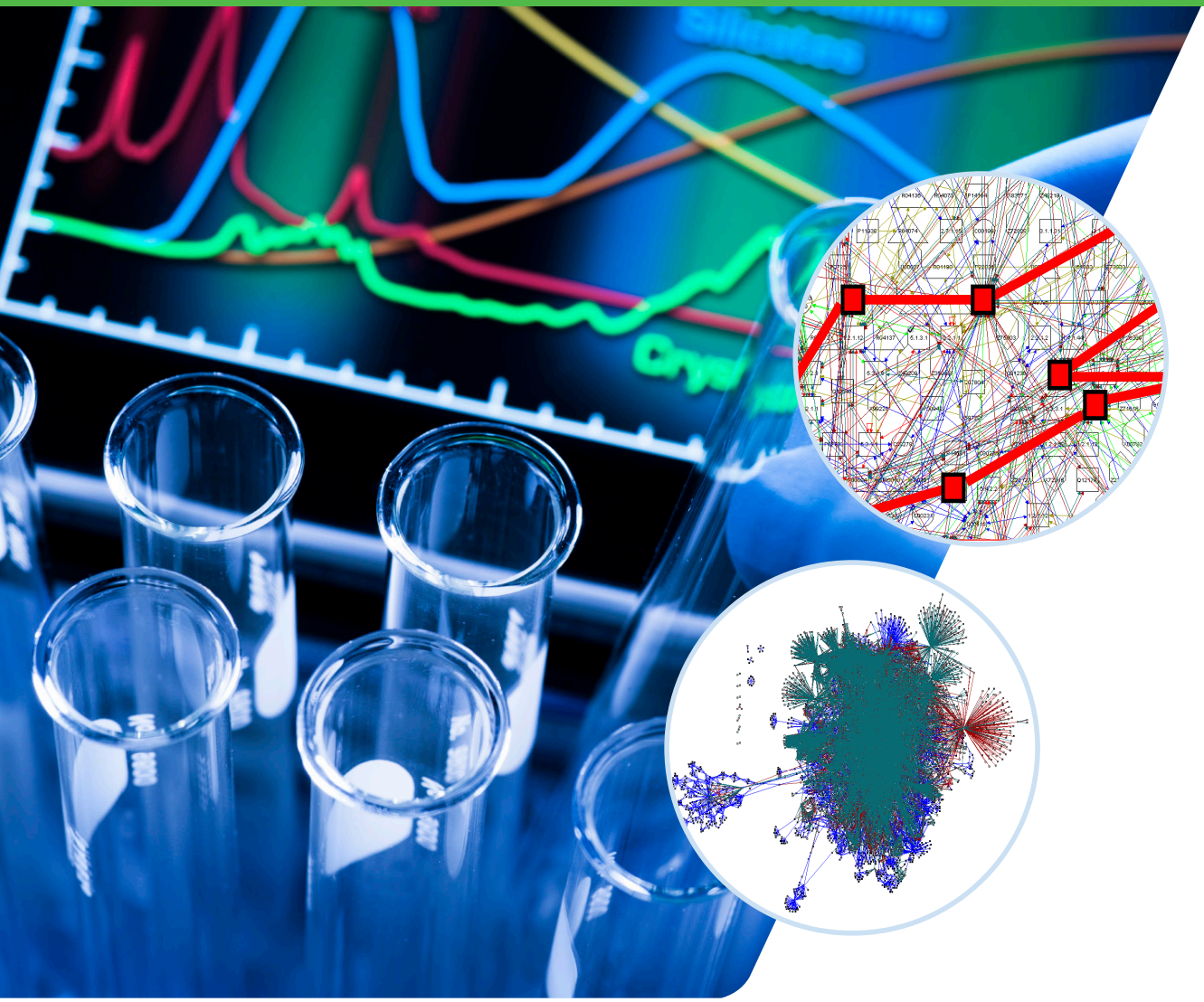
Dissertation

VTT PUBLICATIONS 774



Erno Lindfors

Network Biology

Applications in medicine and biotechnology

VTT PUBLICATIONS 774

Network Biology

Applications in medicine and biotechnology

Erno Lindfors

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 Aalto Doctoral Programme in Science, The Aalto University School of Science and Technology, for public examination and debate in Auditorium Y124 at Aalto University (E-hall, Otakaari 1, Espoo, Finland) on the 4th of November, 2011 at 12 noon.

ISBN 978-951-38-7758-3 (soft back ed.)
ISSN 1235-0621 (soft back ed.)

ISBN 978-951-38-7759-0 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)

JULKAISIJA – UTGIVARE – PUBLISHER

VTT, Vuorimiehentie 5, PL 1000, 02044 VTT
puh. vaihde 020 722 111, faksi 020 722 4374

VTT, Bergsmansvägen 5, PB 1000, 02044 VTT
tel. växel 020 722 111, fax 020 722 4374

VTT Technical Research Centre of Finland, Vuorimiehentie 5, P.O. Box 1000, FI-02044 VTT, Finland
phone internat. +358 20 722 111, fax + 358 20 722 4374

Technical editing Marika Leppilahti

Kopijyvä Oy, Kuopio 2011

Erno Lindfors. Network Biology. Applications in medicine and biotechnology [Verkkobiologia. Lääketieteellisiä ja bioteknisiä sovelluksia]. Espoo 2011. VTT Publications 774. 81 p. + app. 100 p.

Keywords network biology, systems biology, biological data visualization, type 1 diabetes, oxidative stress, graph theory, network topology, ubiquitous complex network properties

Abstract

The concept of systems biology emerged over the last decade in order to address advances in experimental techniques. It aims to characterize biological systems comprehensively as a complex network of interactions between the system's components. Network biology has become a core research domain of systems biology. It uses a graph theoretic approach. Many advances in complex network theory have contributed to this approach, and it has led to practical applications spanning from disease elucidation to biotechnology during the last few years.

Herein we applied a network approach in order to model heterogeneous biological interactions. We developed a system called megNet for visualizing heterogeneous biological data, and showed its utility by biological network visualization examples, particularly in a biomedical context. In addition, we developed a novel biological network analysis method called Enriched Molecular Path detection method (EMPath) that detects phenotypic specific molecular paths in an integrated molecular interaction network. We showed its utility in the context of insulitis and autoimmune diabetes in the non-obese diabetic (NOD) mouse model. Specifically, ether phospholipid biosynthesis was down-regulated in early insulitis. This result was consistent with a previous study (Orešič et al., 2008) in which serum metabolite samples were taken from children who later progressed to type 1 diabetes and from children who permanently remained healthy. As a result, ether lipids were diminished in the type 1 diabetes progressors. Also, in this thesis we performed topological calculations to investigate whether ubiquitous complex network properties are present in biological networks. Results were consistent with recent critiques of the ubiquitous complex network properties describing the biological networks, which gave motivation to tailor another method called Topological Enrichment Analysis for Functional Subnetworks (TEAFS). This method ranks topological activities of modules of an integrated biological network under a dynamic response to external stress. We showed its utility by exposing an integrated yeast network to oxidative stress. Results showed that oxidative stress leads to accumulation of toxic lipids.

Erno Lindfors. Network Biology. Applications in medicine and biotechnology [Verkkobiologia. Lääketieteellisiä ja bioteknisiä sovelluksia]. Espoo 2011. VTT Publications 774. 81 s. + liitt. 100 s.

Avainsanat network biology, systems biology, biological data visualization, type 1 diabetes, oxidative stress, graph theory, network topology, ubiquitous complex network properties

Tiivistelmä

Järjestelmäbiologian käsite syntyi yli kymmenen vuotta sitten vastauksena kokeellisten menetelmien kehitystyöhön. Tämä lähestymistapa pyrkii kuvaamaan biologisia järjestelmiä kattavasti kompleksisena vuorovaikutusverkkona, joka koostuu järjestelmän komponenttien välisistä vuorovaikutuksista. Verkkobiologiasta on tullut tärkeä järjestelmäbiologian tutkimuskohde, ja se käyttää graafiteoreettista lähestymistapaa. Kompleksisten verkkojen teorian kehitystyö on edistänyt tätä lähestymistapaa, ja se on johtanut moniin käytännön sovelluksiin aina sairauksien selventämisestä bioteknologiaan viimeisten parin vuoden aikana.

Tässä väitöskirjassa sovellettiin verkkobiologista lähestymistapaa heterogeenisten biologisten vuorovaikutusten mallintamiseen. Siinä kehitettiin heterogeenisen biologisen tiedon visualisointityökalu megNet, jonka hyödyllisyys osoitettiin biologisten verkkojen visualisointiesimerkein, erityisesti biolääketieteellisessä kontekstissa. Tämän lisäksi väitöstutkimuksessa kehitettiin uusi biologisten verkkojen analysointimenetelmä, rikastettujen molekyylipolkujen havaitsemismenetelmä, joka havaitsee fenotyyppikohtaisia molekyylipolkuja integroidusta molekyylivuorovaikutusverkosta. Tämän menetelmän hyödyllisyys osoitettiin insuliitiksen ja autoimmuunidiabeteksen kontekstissa käyttäen laihojen diabeteshiirien mallia. Erityisesti eetterifosfolipidibiosynteesi oli alisäädelty insuliitiksen varhaisessa vaiheessa. Tämä tulos oli yhteensopiva aikaisemman tutkimuksen (Orešič et al., 2008) kanssa, jossa mitattiin myöhemmin tyypin 1 diabetekseen sairastuneiden lasten ja pysyvästi terveiden lasten seerumin aineenvaihduntatuotteidenpitoisuuksia. Tässä tutkimuksessa havaittiin, että eetterilipidipitoisuudet olivat sairastuneilla lapsilla alhaisemmat kuin terveillä lapsilla. Tässä väitöskirjassa laskettiin myös topologialaskuja, joiden avulla voitiin selvittää, noudattavatko biologiset verkot kaikkialla läsnä olevia kompleksisten verkkojen ominaisuuksia. Tulokset olivat yhteensopivia kaikkialla läsnä olevien kompleksisten verkkojen ominaisuuksiin viime aikoina kohdistuneen kritiikin kanssa. Tämä loi motivaatiota räätälöidä topologista rikastamisanalyysia funktionaalisille aliverkoille, joka etsii topologisesti aktiivisimmat moduulit integroidusta biologisesta verkosta dynaamisen stressin alaisuudessa. Tä-

män menetelmän hyödyllisyys osoitettiin altistamalla integroitu hiivaverkko oksidatiiviselle stressille. Tulokset osoittivat, että oksidatiivinen stressi aiheuttaa toksisten lipidien kasaantumisen.

Preface

This thesis was carried out in the Quantitative Biology and Bioinformatics (QBIX) group at VTT Technical Research Centre of Finland from 2006 to 2010. The main funding sources were National Graduate School in Informational and Structural Biology (ISB) that provided me three-year graduate student grant from 2007 to 2010, TRANSCENDO project of the Tekes MASI Program that funded my six-month exchange visit to International Computer Science Institute (ICSI) Berkeley (CA, USA) in 2006 and 2007, and DIAPREPP EU FP7 project that provided additional funding for my research. I am grateful to all of these funding organizations.

I am indebted to many people that have contributed to this thesis both scientifically and non-scientifically. The biggest gratitude goes to my instructor Research Professor Matej Orešič for making me a scientist. Without his persistent encouragement and enthusiasm I would never have dared to embark on my PhD thesis. During the whole thesis work he has professionally supervised my work on daily basis and maintained scientifically stimulating atmosphere in the whole QBIX group and provided solid funding for us. Also, I am grateful to my supervisor Professor Kimmo Kaski, Head of the Centre of Excellence in Computational Complex Systems Research, Vice Dean of Aalto School of Science, for accepting me as a PhD student at Aalto University, and for his invaluable help in finalizing the thesis and wrapping up everything into covers, and also for helping me with many practical issues. Also, I would like to thank the pre-examiners of this thesis Docent Juho Rousu and Docent Tero Aittokallio for carefully reading the manuscript and for their invaluable comments that helped improve the quality of the thesis. I am also grateful to Professor Samuel Kaski and Dr. Jari Saramäki for being on my advisory board in the ISB graduate school. Both of them have provided invaluable comments in annual meetings. From VTT management level I would like to thank Technology Manager Dr. Richard Fager-

ström, Vice President (R&D) Dr. Anu Kaukovirta-Norja, former Vice President (R&D) (currently Vice President, Business Development) Dr. Juha Ahvenainen, Professor Hans Söderlund, and Professor Johanna Buchert for providing excellent research environment.

The QBIX group was founded by Matej, and in the beginning of 2009 it was split in two groups: Metabolomics group and Bio systems Modeling group. I work in the latter group. I would like to thank all people from these groups for excellent scientific company. Especially, I would like to thank my group leader Dr. Marko Sysi-Aho and my former group leaders Dr. Mika Hilvo, Mr. Pekka Savolahti and Dr. Kim Ekroos for their continuous support and for pushing me to finish my PhD thesis. Also, I am deeply indebted to my close colleague Dr. Venkata Gopalacharyulu Peddinti for his excellent work during the years, especially his contribution to megNet's databases has been crucial. Also, many discussions with him have been very invaluable opening up always new scientific aspects, and he has been always very helpful and showed capability to explain challenging issues in simple way. I would also like to thank my other close colleague Laxmana Rao Yetukuri for fruitful collaboration on lipid pathway reconstruction, and continuously pushing me to finish my PhD thesis. Also, I would like to thank Dr. Tuulia Hyötyläinen and Dr. Tuulikki Seppänen-Laakso for their collaboration on lipidomics studies, and Ms. Sandra Castillo, Mr. Artturi Koivuniemi, Mr. Matti Kankainen, Dr. Tijana Marinković, Dr. Jing Tang, and Mr. Brudy Han Zhao for excellent company in daily life at VTT, and Ms. Anna-Kaarina Hakala and Ms. Sirpa Nygrén for their secretarial help with practical issues.

I have continuously been exposed to working with people from different background at VTT, which has been very rewarding. First of all, I would like to thank Dr. Jyrki Lötjönen and Mr. Jussi Mattila from VTT Signal and Image Processing group, as well the other members of the group for fruitful collaboration on studying biological networks in the context of medical images. Especially, I would like to thank Jussi for developing a desktop user interface for megNet and teaching me many useful aspects in software engineering. Also, I would like to thank Research Professor Merja Penttilä, Dr. Laura Ruohonen, Dr. Mikko Arvas, Dr. Juha-Pekka Pitkänen, Dr. Merja Oja, Dr. Paula Jouhten and Dr. Eija Rintala from VTT Cell Factory for collaboration on studying biological networks in the context of metabolic engineering, and Dr. Harri Siitari, Dr. Arho Virkki, Dr. Vidal Fey, Dr. Sampo Sammalisto and Dr. Timo Pulli for collaboration efforts to commercialize VTT's bioinformatics tools.

This thesis is composed of six jointly published scientific publications. I would like to thank all coauthors of these publications. I have mentioned most of them earlier in this preface. Those not mentioned I would like to thank Dr. Eran Halperin, Dr. Catherine Bounsaythip, Dr. Teemu Kivioja, Dr. Jaakko Hollmén, Mr. Jarkko Miettinen, Dr. Antti Pesonen, and Dr. Vidya R. Velagapudi for their contribution, especially Eran for supervising my work while visiting his group at ICSI Berkeley, and Jaakko for supervising my Master's thesis which initiated the research topic of this thesis.

In addition, I would like to thank all other people of this world. We are composed of a complex network of interactions, so all of you have directly or indirectly interacted with me, and thus made this thesis a reality. Thank you all very much!

September 23, 2011, Espoo, Finland

Erno Lindfors

List of publications

I. **Erno Lindfors**, Peddinti V. Gopalacharyulu, Eran Halperin, and Matej Orešič (2009). Detection of molecular paths associated with insulitis and type 1 diabetes in non-obese diabetic mouse. PLoS ONE, 4(10), e7323. 9 p.

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

III. Peddinti V. Gopalacharyulu (*), **Erno Lindfors** (*), Jarkko Miettinen, Catherine Bounsaythip, and Matej Orešič (2008). An integrative approach for biological data mining and visualization. International Journal of Data Mining and Bioinformatics, 2(1)1:54–77.

IV. Catherine Bounsaythip, **Erno Lindfors**, Peddinti V. Gopalacharyulu, Jaakko Hollmén, and Matej Orešič (2005). 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, Espoo, Finland, Jun 2005. Helsinki University of Technology, Laboratory of Computer and Information Science. 6 p.

V. **Erno Lindfors**, Jussi Mattila, Peddinti V. Gopalacharyulu, Antti Pesonen, Jyrki Lötjönen, and Matej Orešič. Heterogeneous Biological Network Visualization System: Case Study in Context of Medical Image Data. Advances in Experimental Medicine and Biology. (In press.)

VI. Peddinti V. Gopalacharyulu (*), Vidya R. Velagapudi (*), **Erno Lindfors**, Eran Halperin, and Matej Orešič (2009). Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast. Molecular BioSystems, 5(3):276–287.

(*) Equal contribution

Author's contribution

I. Publication **I** introduces the Enriched Molecular Path detection method (EMPath), and shows its utility in the context of type 1 diabetes mouse models leading to interesting findings in terms of medical biology. The author of this thesis designed the method together with Eran Halperin (EH). The author implemented the method, and used it in a type 1 diabetes case study. The author and Matej Orešič (MO) wrote the main parts of the manuscript. Also, Peddinti V. Gopalacharyulu (PVG) and EH contributed to the writing. PVG designed and performed functional and gene set enrichment analyses for the type 1 diabetes case study. MO interpreted the results of the type 1 diabetes case study. EH and MO supervised and conceived the study.

II. Publication **II** introduces a heterogeneous data integration and visualization system called megNet. The utility of this system is demonstrated by two examples: an example in which there is cross-talk[1] between two different stages of metabolism and an example in which a conceptual graph is mapped into two dimensions. The author designed and implemented the algorithm logic in the middle tier, integrated biological entities and modeled them as a biological network representation, and implemented the Sammon's mapping method. Also, he implemented a user interface for the system, and wrote these parts in the manuscript. PVG designed the system, performed data modeling, developed the schemas for the databases, and acquired and incorporated most of the data into the databases. Also, he wrote the first draft of the manuscript which was then improved by the other authors. Catherine Bounsaythip (CB) designed the conceptual spaces for the system. Laxman Yetukuri (LY) acquired the compound data and incorporated it into the databases. Teemu Kivioja (TK) participated in database design and discussed efficiencies of database queries. Jaakko Hollmén (JH) participated in discussion of mapping methods. MO conceived and supervised the study, and interpreted the results.

---

[1] The concept of cross-talk will be used widely in this thesis. In broad sense, this concept means connections between different biological processes (e.g. stages of metabolism). In usual case, more than one 'omics' technologies are involved in this, for example protein-protein interactions can make signaling between different stages of metabolism or between transcriptional regulation and metabolism.

III. Publication **III** extends Publication **II** by introducing new mapping methods and methods for topological calculations and co-expression network construction. The utility of these methods is shown by three practical examples: a generic topological study in a yeast metabolic network, a mapping example in the context of a specific biological process and a co-expression network example in which transcriptomics data is integrated with interaction data. The author designed and implemented the topological study, implemented and designed most of the middle tier, and wrote some parts of the manuscript. PVG developed the ideas concerning integration of transcriptomics data to networks and implemented the analyses of these networks, and wrote the first draft of the manuscript. The author and PVG contributed equally to this work. Jarkko Miettinen (JaM) implemented the Curvilinear Component Analysis (CCA) and Curvilinear Distance Analysis (CDA) mapping methods and improved the Sammon's mapping method. Also, he improved the user interface and middle tier software design and implementation, and wrote the mapping method part of the manuscript. CB designed the conceptual spaces and contributed to the writing. MO conceived and supervised the study, interpreted the results and contributed to the writing.

IV. Publication **IV** describes the details of network representation and the distances used in the megNet's network. It contains three practical examples: an example demonstrating how megNet retrieves and visualizes a metabolic network, an example that demonstrates how a mapping can be used to study the structure of an integrated metabolic and protein-protein interaction network, and a context based mapping example demonstrating how distances between biological entities change based on the biological context. The author designed the network representation and distance matrix, implemented the Sammon's mapping method, and created the practical examples. The author and CB wrote the main parts of the manuscript. All authors contributed to the writing. PVG provided biological details of the data. JH participated in discussion of mapping methods. MO conceived and supervised the study.

V. Publication **V** describes the latest status of the megNet system. It extends Publications **II** and **III** by introducing a desktop user interface for visualizing biological networks in three dimensions, and a web user interface for taking input parameters from the user, and an in-house text mining system

that utilizes existing knowledge. The practical utility of the latest megNet is demonstrated by a case study in which lipidomics data from our laboratory is integrated with interaction data from many sources leading to interactions that could possibly explain our previous associations between biological data and medical images. The author created the practical examples, interpreted the results, designed and implemented most of the algorithm logic in the middle tier, designed and implemented the web user interface, and wrote the main parts of the manuscript. The author and Jussi Mattila (JuM) designed interfaces between the middle tier and user interfaces. JuM designed and implemented the desktop application, and contributed to the writing. PVG maintained the databases, designed and implemented correlation calculations and gene expression data normalization in the middle tier, incorporated UMLS annotation into gene expression data sets, and contributed to the writing. Antti Pesonen (AP) designed and implemented the in-house text mining system. Jyrki Lötjönen (JL) and MO conceived and supervised the study, and contributed to the writing. MO finalized the manuscript.

VI. Publication **IV** introduces the Topological Enrichment Analysis of Functional Subnetworks method (TEAFS), and shows its utility by a case study in which a yeast biological network is exposed to oxidative stress in dynamic manner. The author constructed the networks for the case study, performed topological calculations on reconstructed networks under the dynamic stress, implemented topological calculations in megNet's middle tier that were used in parts of the TEAFS method, implemented the statistical test of the TEAFS method and contributed to the writing. PVG developed the main ideas and implemented parts of the TEAFS method, performed the data analyses and wrote the manuscript. Vidya R. Velagapudi (VRV) performed metabolic experiments and data analysis, and wrote the experimental methods and biological details in the manuscript. PVG and VRV contributed equally to this publication. EH provided ideas for the statistical test, and contributed to the writing. MO conceived and supervised the study and contributed to the writing.

Contents

Appendices V–VI of this publication are 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

API	Application Programming Interface
BIND	Biomolecular Interaction Network Database
BioGRID	Biological General Repository for Interaction Datasets
CCA	Curvilinear Component Analysis
CDA	Curvilinear Distance Analysis
DIP	Database of Interacting Proteins
DNA	DeoxyriboNucleic Acid
EC	Enzyme Commission
EMBL	European Molecular Biology Laboratory
EMPath	Enriched Molecular Path detection
FDR	False Discover Rate
GO	Gene Ontology
JDBC	Java Database Connectivity
JVM	Java Virtual Machine
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes

megNet	Matej Erno Gopalacharyulu Network[2]
MINT	Molecular Interaction Database
MR	Magnetic Resonance
NOD	Non-Obese Diabetic
OAT	Ontology Aided Text mining system
SANDY	Statistical Analysis of Network Dynamics
SCID	Severe Combined Immunodeficiency
SOAP	Simple Object Access Protocol
SRB2	Suppressor of RNA polymerase B II
TEAFS	Topological Enrichment Analysis for Functional Subnetworks
TransFac	Database of Transcription Factors
TransPath	Database of Signal Transduction Pathways
UMLS	Unified Medical Language System
UniProt	Universal Protein Resource
XML	eXtensible Markup Language

[2] This name is based on the inventors of the megNet system: Matej Orešič, Erno Lindfors, Peddinti V. Gopalacharyulu.

1. Introduction

The systems biology approach considers the biological system such as cell as a holistic system that comprises several types of molecules and interactions (Ideker et al., 2001; Kitano 2002a, b). This approach has been developed over the past decade, with network biology emerging as one of its core domains (Chuang et al., 2010). The network approach has already led to practical applications for example in disease elucidation (Chuang et al., 2007; Ideker & Sharan, 2008; Schadt, 2009) and in biotechnology (Luscombe et al., 2004). The basic idea is to model biological phenomena as networks in which nodes are biological entities (e.g. proteins, genes, metabolites) and edges interactions (e.g. protein-protein interactions, metabolic reactions). These methods are based on advances in complex network methods across many fields (Barabási & Albert, 1999; Shen-Orr et al., 2002; Milo et al., 2002, 2004). Ubiquitous complex network properties stemmed from this work have lately obtained some critiques but they have remained as a powerful framework for network biology (Lima-Mendez & Helden, 2009).

One challenge of systems biology is the heterogeneity of biological data: there have been many advances in biological measurement techniques over the past decade, which has generated a huge amount of heterogeneous biological data (Demir et al., 2010). In order to translate this into practical utility, it is necessary to integrate data from various sources into an integrated platform and enable an easy visualization of this data (Gehlenborg et al., 2010; O'Donoghue et al., 2010).

1.1 Aims of the thesis

The aim of this thesis is to address the above-mentioned challenges of systems biology. More specifically the main aims are listed below, and they are summarized in Figure 1.1.

- We set up a system called megNet for visualizing heterogeneous biological data in order to model various types of biological interactions as holistic networks (Publications **II**–**V**) and assign an appropriate distance metric for the biological entities (Publication **IV**). More specifically, the author of this thesis has designed and implemented most of the algorithm logic of this system. Also, he implemented the first desktop user interface of this system, and a web interface for taking input parameters from the user. The practical utility of this system is demonstrated first by a cross-talk example via different stages of yeast metabolism (Publication **II)** and by a context based mapping example in a yeast metabolic network (Publication **III**). Then we used similar approaches to study biological networks in the context of medical images, and we found interactions that could possibly explain our previous associations between lipidomics profiles and medical image parameters (Publication **V**).
- As a main methodological contribution we develop a graph theoretic method called Enriched Molecular Path detection method (EMPath). We show the utility of this method by using it in the context of type 1 diabetes mouse models leading to interesting results in terms of medical biology (Publication **I**).
- This thesis contributes to topological analyses of biological networks. We first performed topological calculations on a generic yeast metabolic network (Publication **III**), and then on reconstructed yeast networks under dynamic stress (Publication **VI**) to investigate whether ubiquitous complex network properties are present in these networks. These results showed that these laws are not present, which is consistent with the recent critiques to them. It thus indicated that we cannot gain our biological understanding much from generic topological studies and thus gave motivation to tailor the Topological Enrichment Analysis for Functional Subnetworks method (TEAFS) so that it analyzes modules of networks. This method was developed in Publication **VI**. In this publication we showed the utility of this method by exposing a yeast biological network to oxidative stress. As a result we found that toxic lipids were accumulated under dynamic response to oxidative stress, which was validated by in-house metabolomic analysis. In the development of this method the author of this thesis provided help in network construction, and in statistical and topological calculations.

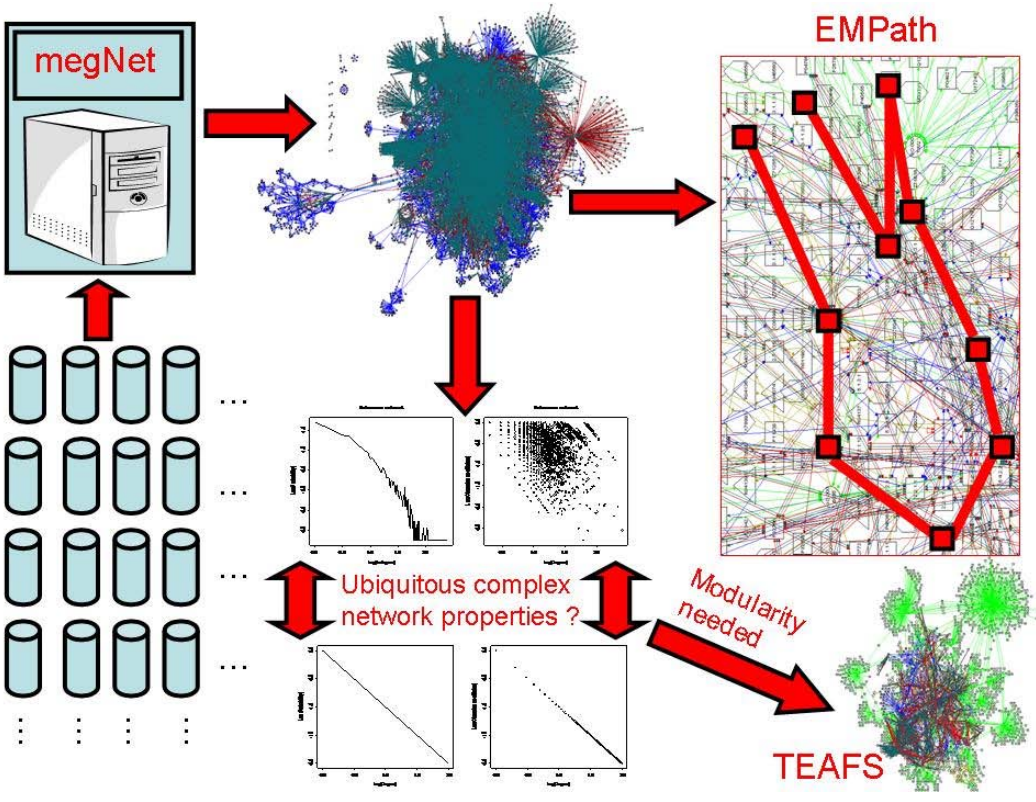


Figure 1.1. Schematic diagram summarizing the main aims of this thesis.

2. Literature review

In order to better understand the background of network biology, in this chapter we describe how it has evolved during the last few decades. We can roughly divide this process in three main parts as illustrated in Figure 2.1. In the first part solid theory for complex networks was created. In the beginning not much computational resources were available. Some preliminary models were created, but they were mainly based on intuition while lacking practical evidence. Then gradually more computational power became available. This enabled testing models on real data, which introduced ubiquitous complex network properties across many fields. In the second part a huge amount of experimental data became available. This enabled considering several components simultaneously as a holistic system leading to 'systems biology' (Ideker et al., 2001; Kitano 2002a, b). During the last few years these models have been used in real biological contexts. This has led to some critiques towards the ubiquitous complex network properties. However, specific tools and concepts of complex network theory have remained as a powerful framework in network biology leading to many practical applications.

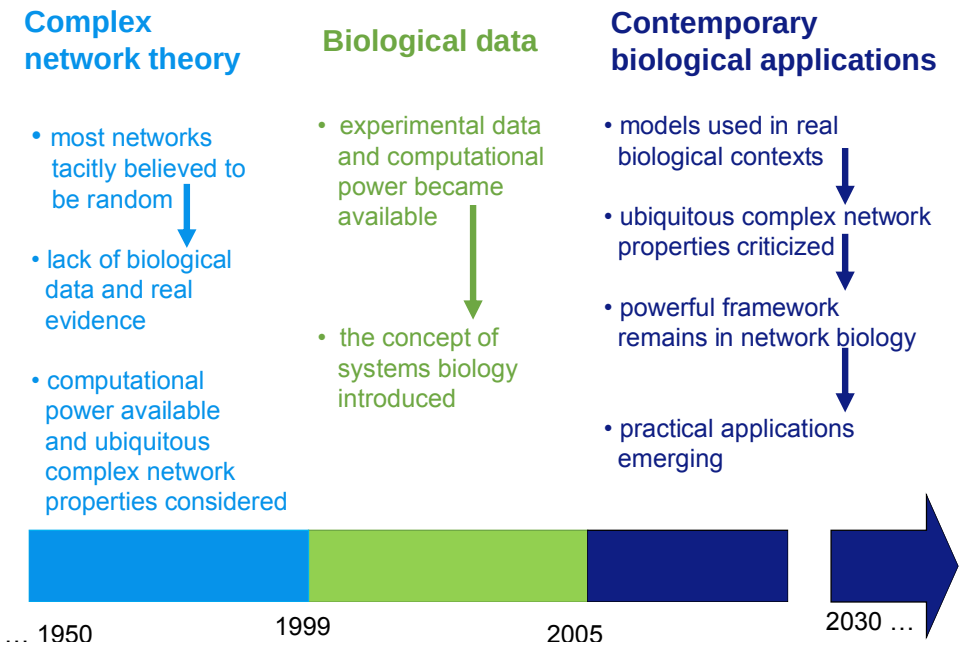


Figure 2.1. Main parts of network biology.

2.1 Complex network theory

During the last decade there have been many advances in complex network theory (Albert & Barabási, 2002). In these efforts phenomena from many fields are modeled by networks. In biology these networks comprise nodes that are biological entities (e.g. proteins, metabolites) and edges that are interactions (e.g. protein-protein interactions, metabolic reactions).

Until 1999 most networks were tacitly believed to follow an Erdős-Rényi random network model (Erdős & Rényi, 1959, 1960). Mathematical details of this model are described in Section 3.3. Briefly the idea is that nodes are connected randomly to each other. However, the assumption that most networks follow this model was mainly based on intuition: there were not practical applications to validate this assumption.

In the beginning of this millennium more computational power became available, which enabled testing models on real data. It led to a power-law degree distribution model which was first demonstrated by practical examples from outside biology (Barabási & Albert, 1999) and then also in biological networks such as in metabolic networks (Jeong et al., 2000) and in protein-protein interac-

tion networks (Jeong et al., 2001; Wagner, 2001; Giot et al., 2003; Li et al., 2004). Then another model called hierarchical network model was introduced (Ravasz et al., 2002; Ravasz & Barabási, 2003), and it was shown that biological networks such as metabolic networks (Ravasz et al., 2002) and protein-protein interaction networks (Yook et al., 2004) follow this model, as well many networks from outside biology (Ravasz & Barabási, 2003). Therefore, some scientists considered the power-law degree distribution and hierarchical models as ubiquitous complex network properties, since they were applied across many fields. The mathematical details of these models are also described in Section 3.3.

The ubiquitous complex network properties introduced important concepts for network biology. For example robustness: a power-law network is robust to a random attack to a node and lethal to a targeted attack to a highly connected hub node (Jeong et al., 2000, 2001). The network can thus keep its structure if a random node is collapsed, but it gets fragmented if a highly connected hub node is collapsed. Another important concept is modularity: biological networks tend be organized in modules, and inside each module biological entities interact with each other in order to carry out a distinct biological function (Hartwell et al., 1999; Qi & Ge, 2006). However, this is not usually ideally the case, for example there are connections between modules via hierarchy levels (Ravasz et al., 2002; Ravasz & Barabási, 2003). Also, as an important concept to study the biological meaning of modules a network motif[3] was introduced as a significantly recurring pattern in a network about ten years ago, first by showing that a transcriptional interaction network in *Escherichia coli* is composed of biologically meaningful motifs (Shen-Orr et al., 2002). Then this concept was generalized by showing that complex networks from many other fields (e.g. neurology, ecology, and engineering) are also composed of meaningful motifs (Milo et al., 2002). A few years later the universality of this concept was shown: similar motifs across many fields were found, for example in transcription networks in microorganisms, World Wide Web and social networks, and word adjacency networks from different languages (Milo et al., 2004). However, the concept of network motif has been criticized by stating that some motifs tend to be results from spatial clustering rather than ubiquitous evolutionary properties (Artzy-Randrup et al., 2004).

[3] Analogously the concept of motif had been used in sequence analysis as recurring nucleotide or amino-acid patterns.

A growth and preferential attachment process is another interesting concept related to the ubiquitous complex network properties (Yule, 1925; Simon, 1955; Price, 1976; Barabási & Albert, 1999; Newman, 2005). It is a stochastic process that is assumed to generate the power-law degree distribution model. In brief, it is based on the following two assumptions.

1. The network grows over time: new nodes continuously join the network.
2. A new node prefers to link to a highly connected node: the higher number of links a node has the higher probability is that it gets a new link.

In a network biology review Barabási & Oltvai (2004) they explain how the growth and preferential attachment process is associated with gene duplication in protein-protein interaction networks. Briefly, the idea is that in gene duplication one or several genes are copied twice. This is manifested as a new interacting partner in protein-protein interaction network. The more links a protein has the higher probability is that it interacts with a protein of duplicated genes, and thus gets a new interacting partner.

In Albert & Barabási (2002) they mention that the growth and preferential attachment process could generate networks also in other fields. For example, when we create a new page in the World Wide Web, we tend to create a link to a popular page (e.g. Google Web Search page). Therefore a highly connected page tends to get linked to a new page when the World Wide Web grows. In a citation network a highly cited publication tends to get a new citation, since it is well known and thus has scientific credibility.

2.2 Biological data

Gradually early this millennium many high-throughput technologies emerged for many types of interactions. As a result, we have a huge amount of heterogeneous biological interaction data available, which has revolutionized the biological research. Traditionally we were interested in single molecules (e.g. genes), whereas now it is possible to consider several components simultaneously in integrated manner via several types of interactions. This approach has led to a new concept called ‘systems biology’ (Ideker et al., 2001; Kitano 2002a, b).

As high-throughput technology examples, two techniques for detecting protein-protein interactions were developed: a yeast two-hybrid method (Uetz et al., 2000; Ito et al., 2000; Fields, 2005) and affinity purification coupled with mass spectrometry (Ho et al., 2002; Gavin et al., 2002, 2006; Krogan et al., 2006).

Both of these technologies enable detecting thousands of protein-protein interactions simultaneously. The former detects binary interactions. The later detects interaction complexes. These methods have generated a huge amount of protein-protein interaction data. Many databases have been established to collect this data, for example DIP (Xenarios et al., 2002), MINT (Ceol et al., 2010), and BIND (Bader et al., 2003). Though these databases provide promising initial framework for studying networks in protein level, they still have many challenges ahead, for example it has been estimated that protein interaction maps are 50% complete for a model organism *Saccharomyces cerevisiae* yeast and 10% complete for human, and they contain a high number of false-positive interactions (Hart et al., 2006).

During the last 10–20 years many genomes have been completed, most notably the human genome project (Lander et al., 2001; Venter et al., 2001). Many organism specific metabolic models have been constructed from these genomes. For example, KEGG is a database comprising metabolic pathway maps for more than one hundred species (Kanehisa et al., 2004). Also, many genome-wide metabolic models have been constructed for model organisms such as yeast *Saccharomyces cerevisiae* (Förster et al., 2003; Duarte et al., 2004; Herrgård et al., 2008), *Escherichia coli* (Feist & Palsson, 2008), mouse (Sheikh et al., 2005; Quek & Nielsen, 2008), and also for human (Duarte et al., 2007; Ma et al., 2007).

Also, many microarray technologies emerged by the early millennium (Schulze & Downward, 2001). This has enabled simultaneous study of several genes in a phenotypic context by taking gene expression measurements for example from disease and healthy samples. Some systematic efforts have been made to collect this data. For example, GEO is a public database where biologists can submit their gene expression experiments (Barrett et al., 2009). As a result, there are several thousands of samples from different conditions that researchers can freely use. In addition, several other biological databases have been established during the last decade. More extensive list of these databases is presented for example in Demir et al. (2010).

2.3 Contemporary biological applications

Since the concept of systems biology has existed for a while, biologically meaningful applications have emerged, which in turn has shed also some critiques towards the ubiquitous complex network properties that were made in the early times of complex network theory. Especially, the presence of the power-law

degree distribution[4] in biological networks has been criticized. For example, in Khanin & Wit (2006) they took a rigorous approach to this question. This was based on an observation that it is usually tempting to come up with a conclusion that a distribution follows the power-law always when it is decreasing. They used a maximum likelihood method to investigate rigorously whether distributions of 10 biological networks (e.g. protein-protein interactions, gene interactions, synthetic lethal interactions, metabolic interactions) follow the power-law. As a result, none of these distributions followed ideally the power-law degree distribution model. In addition, they investigated how consistent the same 10 biological networks are with a truncated power-law degree distribution model which defined rigorously in Equation 3.4 in Section 3.3. The results were more promising: all networks followed the truncated power-law degree distribution model with quite small cut-off coefficients. This gave a hint that it seems that biological networks follow the power-law degree distribution model only in very small degrees. Actually already in Jeong et al. (2001) there was supporting evidence stating that biological networks follow better the truncated power-law degree distribution model than the 'normal' power-law degree distribution model. In addition, some other alternative models to the power-law degree distribution model have emerged. For example, in Pržulj et al. (2004) they introduced a geometric random model. In Pržulj (2007) they showed that many protein-protein interaction networks are more consistent with this model than with the power-law degree distribution model. Based on all of these findings we can conclude that it seems that the power-law degree distribution model is not present in the ideal form suggested by the theory in biological networks, and also there has been evidence stating that these models contain sampling artifacts, i.e. if a subnetwork follows the power-law degree distribution model, it does not imply that the whole network follows it (Aittokallio & Schwikowski, 2006).

A recent network biology review (Lima-Mendez & Helden, 2009) points out the above-mentioned weaknesses of ubiquitous complex network properties but it also points out that complex network theory has created important tools and concepts such as hub, robustness and modularity that have turned out to be a powerful framework in practical applications in network biology. Especially, it points out the importance of local modules and motifs. The same issue is elevat-

[4] This distribution is defined formally in Section 3.3 in a bullet entitled "Power-law degree distribution model".

ed also in another network biology review (Qi & Ge, 2006) in which they point out that the modularity is an important concept when studying biological networks in dynamic manner.

During the last few years useful biological applications have emerged. For example Luscombe et al. (2004) developed a method called Statistical Analysis of Network Dynamics (SANDY). This method has biological novelty, since it handles a biological network in dynamic manner: previously biological networks were studied in static manner. This method uses time-varying transcriptomics data from multiple conditions. For each condition it calculates topological measures (e.g. node degrees), identifies most important hubs and motifs. They showed the utility of the method by a case study in which a cell was exposed to inter-cellular processes in two conditions and to environmental changes in three conditions. They found that transcription factor combinations are complex and highly inter-connected under inter-cellular processes, whereas they are simple and loosely connected under environmental changes.

As a local modularity approach Chuang et al. (2007) developed a method that searches sub-networks in the context of gene expression data. They used this method to search sub-networks in a protein-protein interaction network to discriminate patients with breast cancer metastasis. As a result, they detected sub-networks that provided novel hypotheses for pathways involved in tumor progression. These networks contained genes that were not differentially expressed whereas they importantly interconnected differentially expressed genes. This indicated the importance of the network approach: the gene expression data alone would not have been able to detect the interconnecting genes.

In addition, visualization has been an important topic during the last few years. There is a huge amount of heterogeneous biological data available and there are several good single tools for visualizing and analyzing heterogeneous biological data, for example Cytoscape (Cline et al., 2007), PATIKA (Demir et al., 2002), ONDEX (Köhler et al., 2006), Medusa (Hooper & Bork, 2005), Osprey (Breitkreutz et al., 2003), BioLayout Express(3D) (Freeman et al., 2007), ProViz (Iragne et al., 2005), PIVOT (Orlev et al., 2004), COPASI (Hoops et al., 2006), GEPASI (Mendes, 1993, 1997), E-CELL (Tomita et al., 1999), COBRA Toolbox (Becker et al., 2007). However, the basic problem that the biologist faces is the usability: databases and tools tend to be separated from each other (Gehlenborg et al., 2010; O'Donoghue et al., 2010), and they are usually quite difficult to use in a real biological context (Saraiya et al., 2005; Pavlopoulos et al., 2008). Therefore, there is need for integrated platforms that allow easy visu-

alization and analysis of heterogeneous data (e.g. signaling, regulatory, metabolic) across multiple levels (e.g. from molecular to anatomical level) in different contexts (e.g. cellular localizations, disease versus healthy state). Traditionally this has been quite a formidable challenge, but efforts towards this direction are underway.

3. Methods

In this chapter we describe the methods used in this thesis. In Section 3.1 we describe a heterogeneous biological data visualization system called megNet that constitutes the set up for the research of this thesis. In Section 3.2 we describe the Enriched Molecular Path detection method (EMPath) that is the main method developed in this thesis. In Section 3.3 we go through the most commonly used topological methods of biological networks and briefly describe how we use them in this thesis. In Section 3.4 we describe the Topological Enrichment Analysis for Functional Subnetworks method (TEAFS) to which this thesis contributes.

3.1 megNet – Heterogeneous biological data visualization system

In Publications **II–V** we have developed a heterogeneous biological visualization system called megNet in order to address the needs of systems biology: model various biological interaction types as holistic systems (Ideker et al., 2001; Kitano 2002a, b). The main aim is to provide easy visualization of heterogeneous biological data (Gehlenborg et al., 2010; O'Donoghue et al., 2010). This system is described in detail in these publications. In this chapter we describe it briefly. More specifically, in Section 3.1.1 we present its overall idea. In Section 3.1.2 we briefly describe its technical architecture and main algorithms.

3.1.1 Overall idea

An overall conceptual framework of megNet is presented in Figure 1 of Publication **V**. Several single biological databases exist. The basic idea is to integrate these databases into an integrated platform, and thus translate the work made on these databases into practical utility. Once the data is integrated, the user then

models it as a network: biological entities as nodes (e.g. proteins, metabolites) and interactions as edges (e.g. protein-protein interactions, metabolic reactions).

Once the user has created the network model, he or she then uses megNet to construct networks that are usually quite large for reasonable interpretation. He or she therefore needs to study them in a specific context that can be for example a medical image or a physiological condition from a yeast culture. Then he or she uses computational methods to extract context specific information from the network. He or she can use for example a context based mapping that we will briefly describe in Section 3.1.2. Alternatively he or she can export the network to other tools for example to the Enriched Molecular Path detection method (EMPath) (Section 3.2), or to the Topological Enrichment Analysis of Functional Subnetworks method (TEAFS) (Section 3.4). In addition, he or she can browse the network manually, and use the human eye to detect for example cross-talk between different stages of biological processes. The utility of this approach is demonstrated by practical examples in Sections 4.1.1 and 4.1.3. Also, we have made an online demo in http://sysbio.vtt.fi/megNet_demo/index.html[5] that briefly shows a few use-case examples.

3.1.2 Technical architecture and main algorithms

The technical architecture of megNet is described in detail in Publications **II–V**. It can be divided in three main components: client, middle tier and database tier that are presented in Figure 3.1. Next we will describe how the middle tier implements the main algorithms of megNet. Also, we will briefly describe the basic functionalities of the client and the overall content of the database tier.

[5] If this link expires, please contact the author of this thesis (Erno.Lindfors@vtt.fi) to request an updated link.

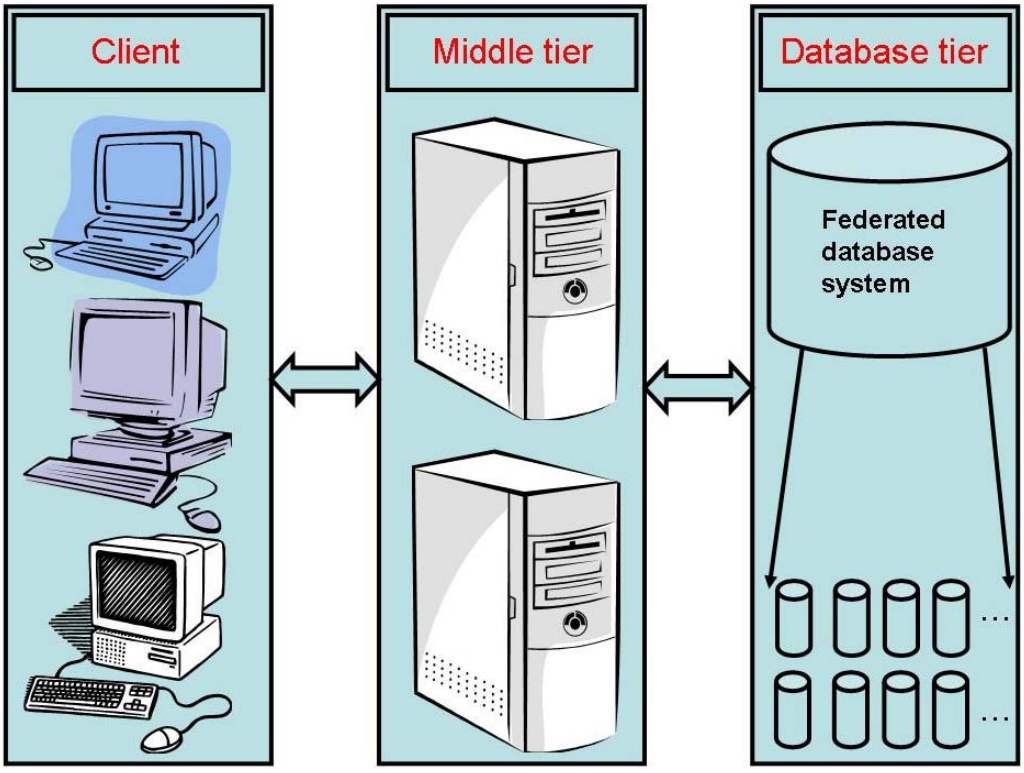


Figure 3.1. Main components of megNet.

Middle tier

The purpose of middle tier is to process the algorithm logic of megNet. More specifically, it constructs networks, performs text mining, context based mapping and topology calculations. In this section we will describe how megNet implements these algorithms.

The middle tier is implemented in Java programming language by using JVM v.1.6.16 (Oracle, Inc.), and it is running on a JBoss Application Server (JBoss, Inc.). It uses a Tamino Java API and Oracle JDBC Thin drivers to communicate with the databases, and Simple Object Access Protocol (SOAP) messages to communicate with the user interfaces by using internal XML schemas that are represented as diagrams in Figures 3.2–3.12.

Network construction

Network construction is the most central algorithm that the middle tier implements, since most of the other algorithms use the network. It takes a graph construction request (Figure 3.2) as in put. This message comprises many elements which enables constructions of networks of many types. Most of these elements are optional which means that the middle tier can construct the network from only a few input parameters. Next we will briefly describe each of these elements.

- *QueriedDatabases*. This element comprises the names of the databases from which the middle tier retrieves interactions and reactions for the network.
- *Species*. This element comprises the species in which the middle tier constructs the network.
- *UniProtAccessionNumbers*. This element comprises the UniProt accession numbers (UniProt Consortium, 2010) of proteins for which the middle tier retrieves interactions and reactions.
- *UniProtEntryNames*. This element comprises the UniProt entry names (UniProt Consortium, 2010) of proteins for which the middle tier retrieves interactions and reactions.
- *EcNumbers*. This element comprises the EC numbers (Webb, 1992) of proteins for which the middle tier retrieves interactions and reactions.
- *EmblIds*. This element comprises the EMBL identifiers (Cochrane & Galperin, 2010) of genes for which the middle tier retrieves interactions and reactions.
- *KeggMetabolicPathways*. This element comprises the names of metabolic pathways that the middle tier retrieves from KEGG (Kanehisa et al., 2004) and integrates them with other selected databases.
- *YeastNetMetabolicPathways*. This element comprises the names of metabolic pathways that the middle tier retrieves from Yeast 1.0 (Herrgård et al., 2008) and integrates them with other selected databases.
- *GeneNames*. This element comprises the names of genes for which the middle tier retrieves interactions and reactions.

- *GoAccessions*. This element comprises the GO (Gene Ontology Consortium, 2008) accessions of biological processes for which the middle tier retrieves interactions and reactions.
- *CompoundNames*. This element comprises the names of compounds for which the middle tier retrieves interactions and reactions.
- *KeggCompoundIds*. This element comprises the KEGG identifiers (Kanehisa et al., 2004) of compounds for which the middle tier retrieves interactions and reactions.
- *Depth*. This element comprises the depth of the network construction, which means how many nearest neighbors the middle tier retrieves for given proteins, genes and/or metabolic pathways.
- *CorrCoeffs*. This element comprises correlation coefficients for gene pairs for which the middle tier constructs a co-expression network and integrates it with interactions and reactions retrieved from other selected databases.
- *BarDataSets*. This element comprises gene expression datasets that the middle tier associate with genes so the client visualizes them as bars inside gene nodes.
- *UseComp*. This element defines whether the middle tier constructs a compartmentalized or non-compartmentalized network.

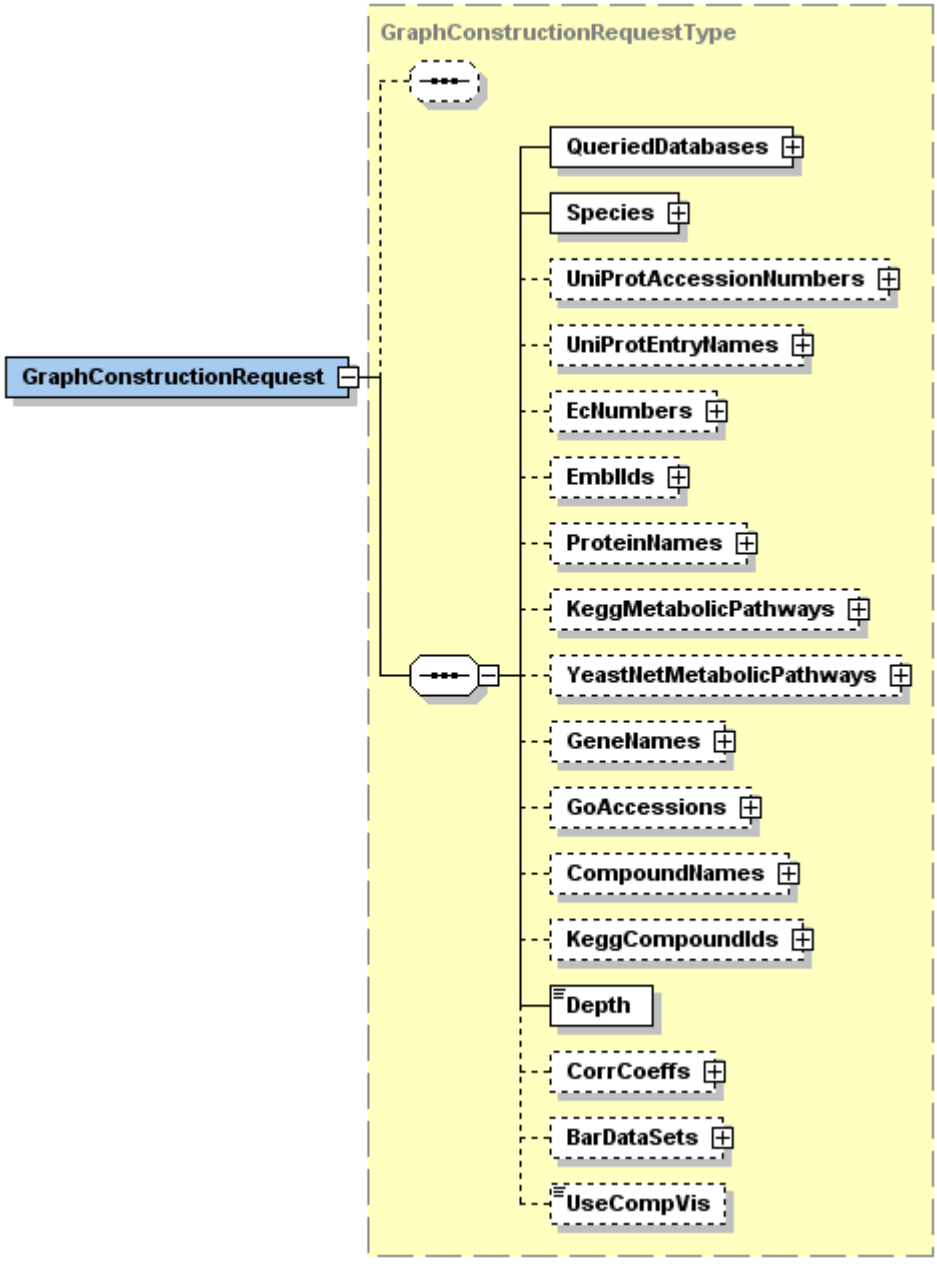


Figure 3.2. XML schema for graph construction request.

Once the middle tier has constructed the network, it returns it as a graph construction response (Figures 3.3–3.5). This message comprises three main elements that we will briefly describe below.

- *ConnectionTypes*. This element comprises connection types that the network comprises. It has three attributes: the first one defines whether the connection is uni-, bi-, or non-directional, the second one defines a shortened name for the connection type (e.g. PROT_INT) and the third one defines a longer name for the connection type (e.g. "protein interaction").
- *Nodes*. This element comprises nodes that the network comprises (Figure 3.4). Each sub-element represents one node type (e.g. protein, gene). Each of these elements comprises more specific data about the node. For example, the protein comprises many identifiers that describe it in detail (e.g. UniProt Identifiers, EC number) as described in Figure 3.4.
- *Edges*. This element comprises edges that the network comprises (Figure 3.5). Each sub-element represents one edge type (e.g. protein-protein interaction, KEGG). Each of these elements comprises more specific data about the edge. For example, the protein-protein interaction comprises source databases from which the interaction was retrieved as described in Figure 3.5.

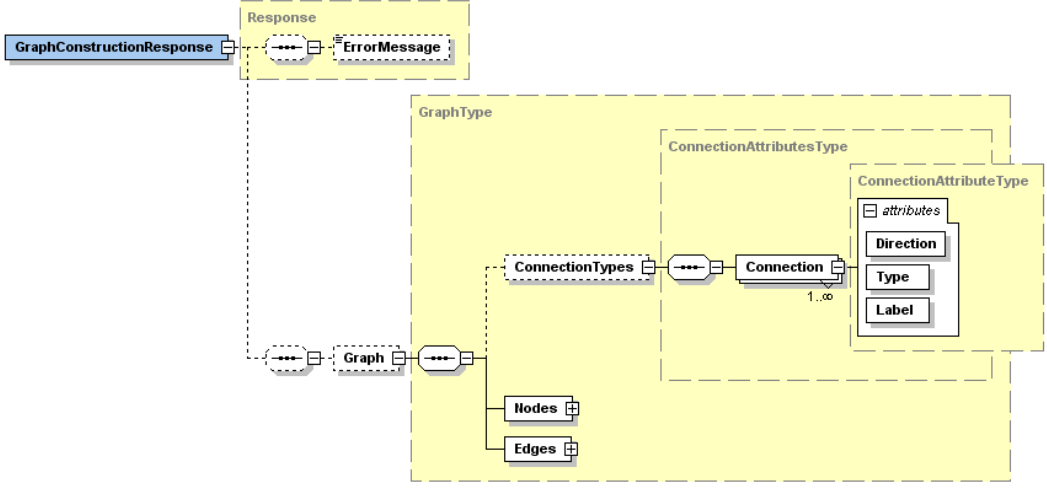


Figure 3.3. XML schema for the main elements of graph construction response. The nodes and edges elements are opened in Figures 3.4 and 3.5 respectively.

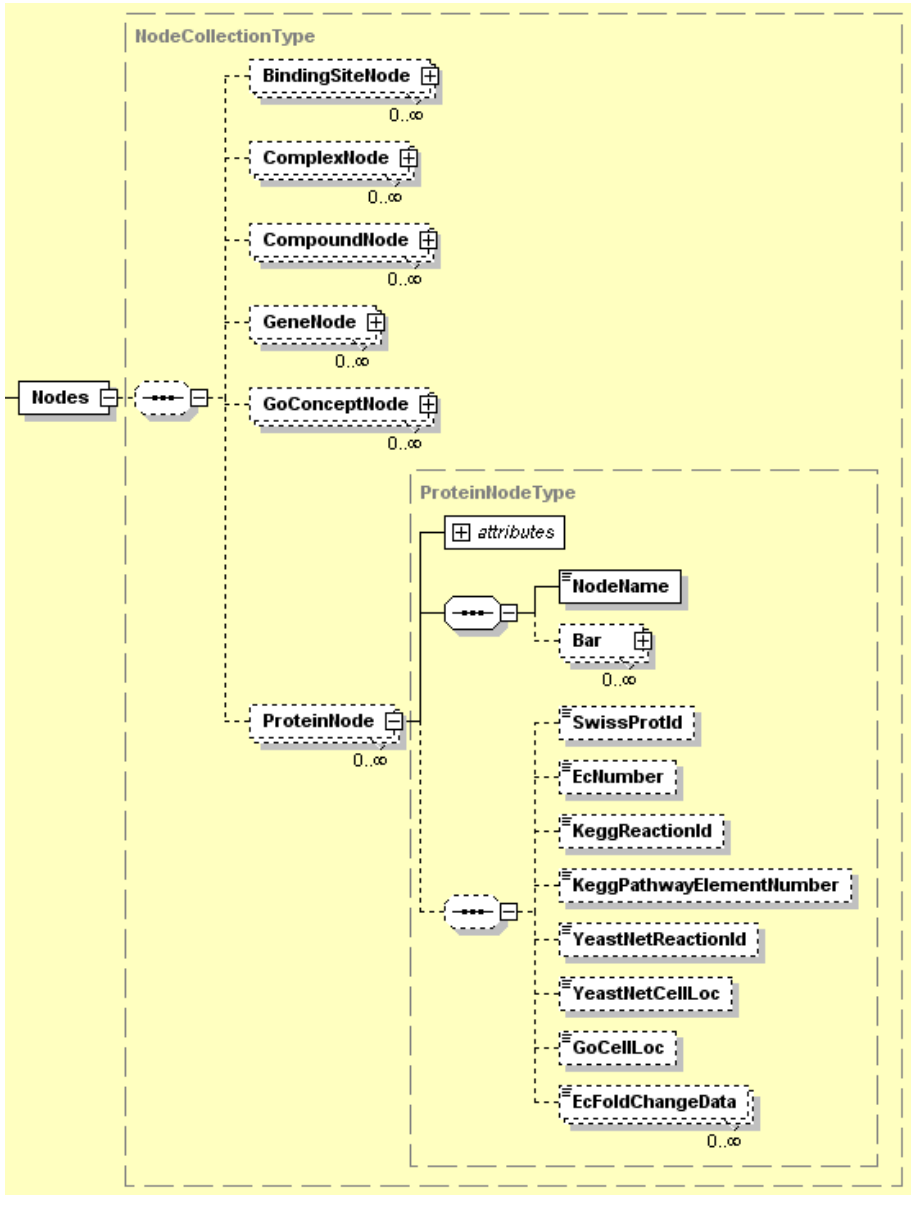


Figure 3.4. XML schema for the nodes element of graph construction response.

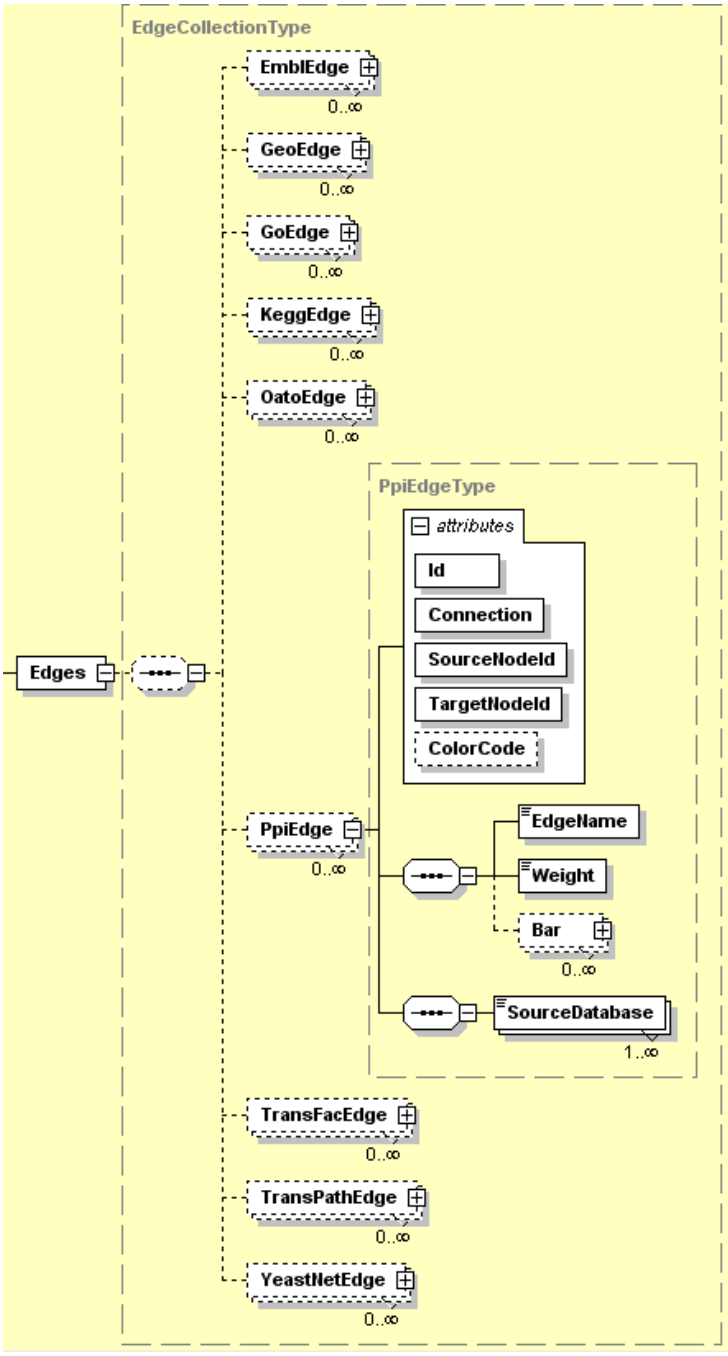


Figure 3.5. XML schema for the edges element of graph construction response.

Text mining

The text mining algorithm takes a text mining request (Figure 3.6) as input. This message comprises elements for databases and species. The purposes of these elements are similar as in the graph construction request: they define from which database and in which species the middle tier retrieves data. Also, there is an element that defines keyword(s) (e.g. diabetes, oxygen) for the retrieval.

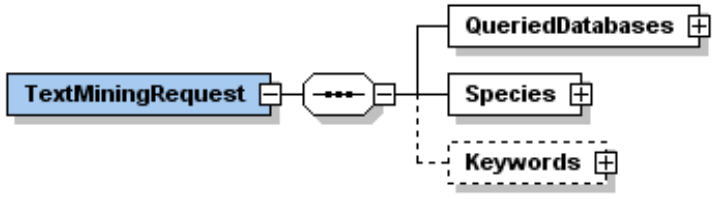


Figure 3.6. XML schema for text mining request.

The middle tier retrieves gene expression microarray data sets and proteins that are annotated with the keyword from GEO (Barrett et al., 2009) and UniProt (UniProt Consortium, 2010) respectively, and includes them in the text mining response (Figure 3.7). The retrieved proteins are included the *ProteinNodes* element, which is identical to this element in the graph construction response (Figure 3.4). The retrieved datasets are included in the *DataSets* element. This element comprises a data type called *ExperimentDataType*. This data type comprises an experiment specific data (e.g. textual description, title, keywords, medical annotations). In addition, the *DataSets* element comprises a *Samples* element that contains also the *ExperimentDataType* which in turn defines a sample specific data. In the *DataSets* element there is a *Channel* attribute that defines whether the data set is of single (Lockhart et al., 1996) or of dual (Schena et al., 1995) channel microarray.

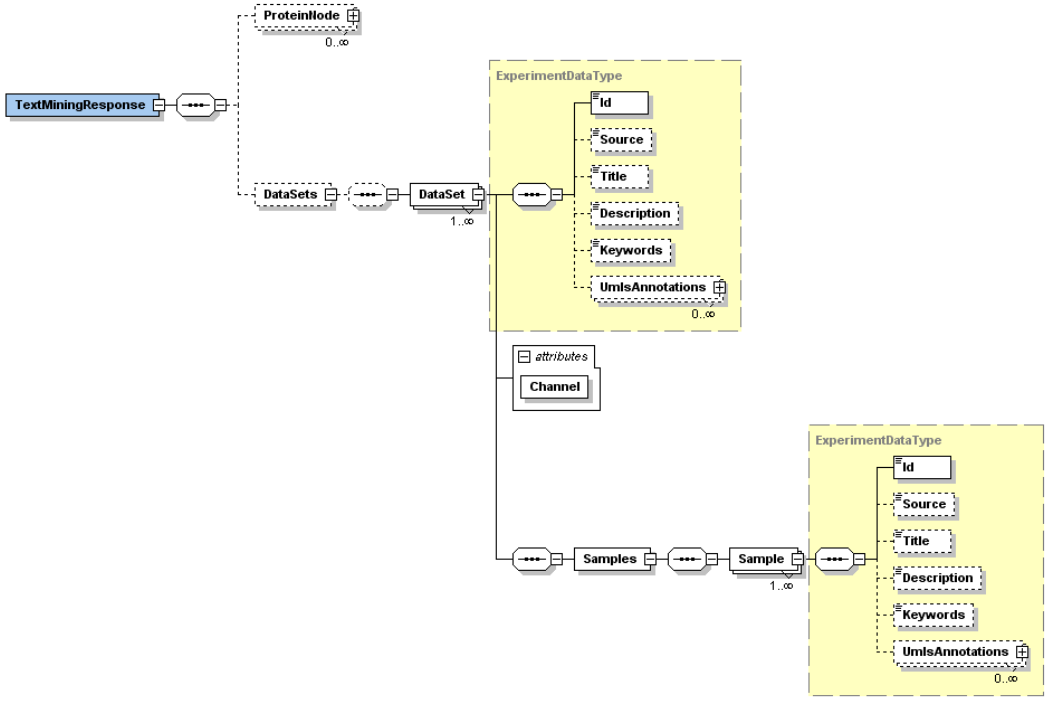


Figure 3.7. XML schema for text mining response.

Context based mapping

The purpose of the context based mapping algorithm is to map internal distances of nodes of a biological network into low a dimensional output space (usually two or three). Figure 1 of Publication **IV** illustrates how the internal distances are calculated. The internal distances and the output space have some discrepancy that we call mapping error. The purpose of the mapping algorithm is to iterate the output space so that the mapping error is minimized. The middle tier comprises three mapping methods: Sammon's Non-Linear Mapping (Sammon, 1969), CCA (Demartines & Hérault, 1997) and CDA (Lee et al., 2004). The mapping algorithm comprises three messages: initialize mapping request (Figure 3.8), mapping response (Figure 3.9) and iterate mapping (Figure 3.10). Next we will briefly describe the content of each of these messages and how the middle tier interacts with them.

The purpose of the initialize mapping request is to initialize a mapping for a network. It comprises a *Graph* element, which is identical to this element in the graph construction response (Figure 3.3), and it comprises a network for which

the mapping will be initialized. This network comprises weights of the edges as illustrated in the graph construction response (Figure 3.5). They are taken into account when calculating the internal distances of the nodes. Also, the initialize mapping request comprises input parameters elements for each mapping types: *CdaParameters*, *CcaParameters* and *SammonsParameters* element. All of these elements comprise a *ResponseDimension* attribute that defines the dimension of the output space and a *StartingIterations* attribute that defines how many times the mapping is iterated in the initialization. The *CdaParameters* and *CcaParameters* elements comprise additional mapping parameters that are described in detail in Publication **III**.

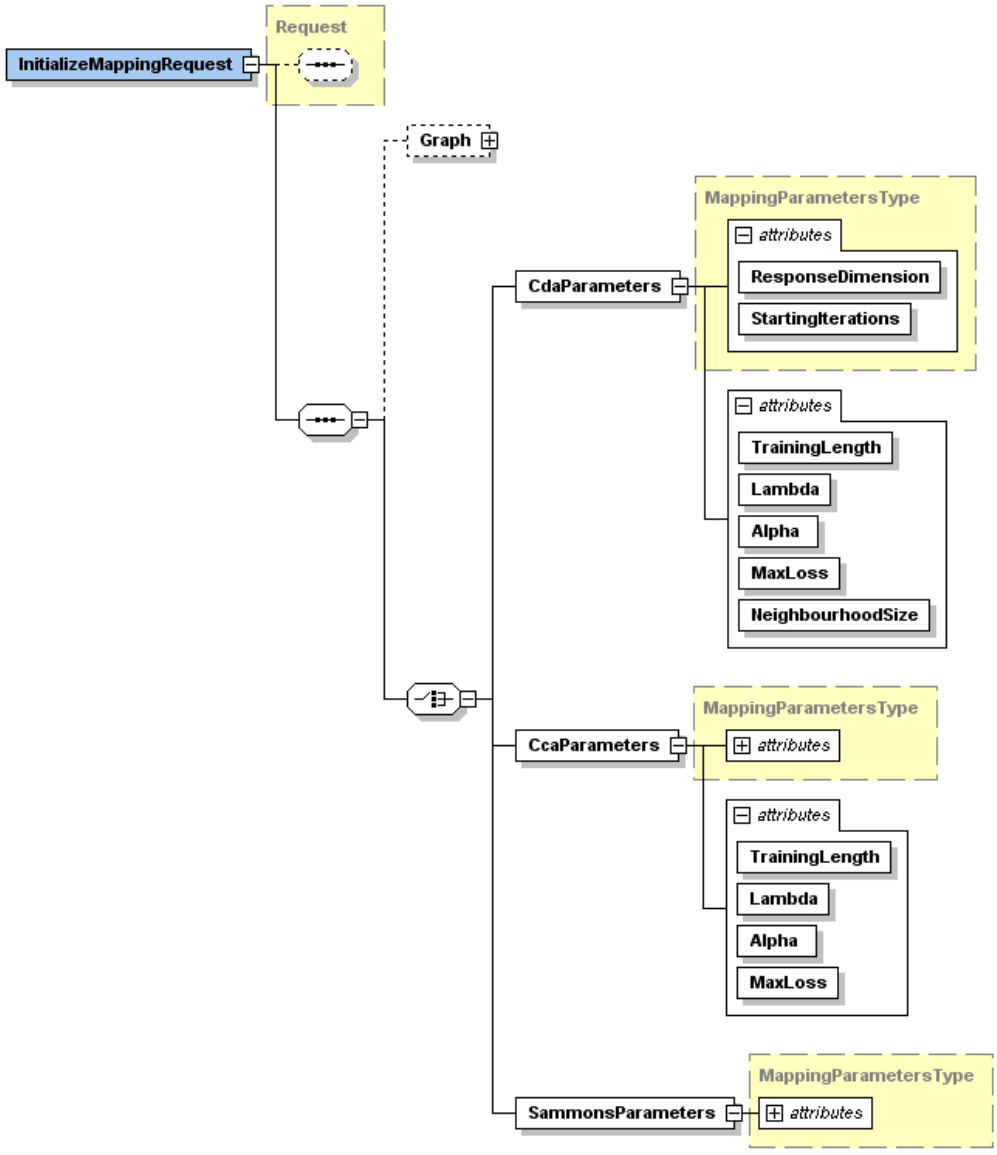


Figure 3.8. XML schema for initialize mapping request.

When receiving an initialize mapping request, the middle tier first calculates the internal distances, and then initializes the output space based on the mapping parameters. It includes the mapping error between the initialized output space and internal distances in a *MappingError* element and the coordinates of the

initialized output space in a *Coordinates* element (Figure 3.9). This element has a *Coordinate* child element that defines coordinates for one node of the biological network of which internal nodes are being mapped. *PosX*, *PosY* and *PosZ* attributes defines the position of the node in the output space. The *NodeId* attribute links the node to the *Graph* element of the initialize mapping request (Figure 3.8).

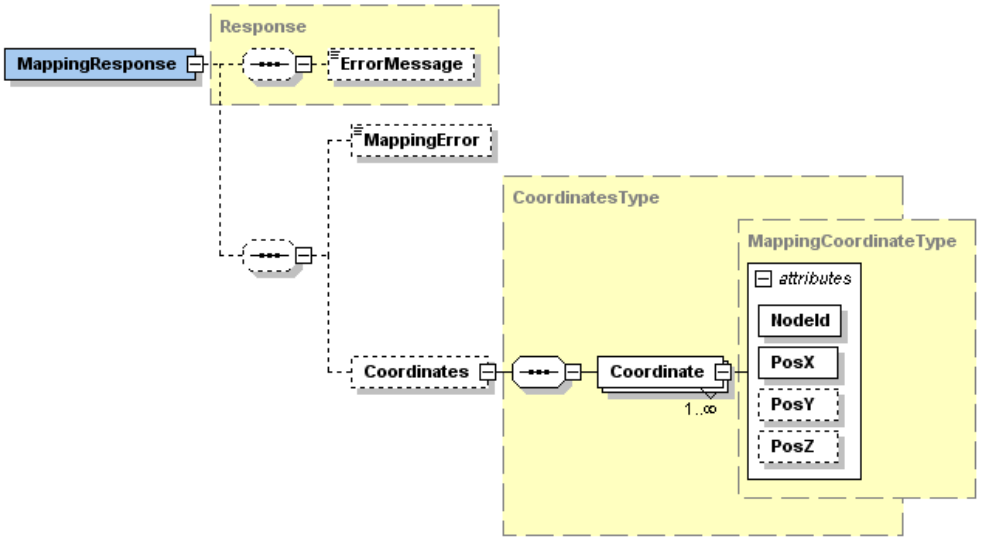


Figure 3.9. XML schema for mapping response.

The purpose of the iterate mapping request (Figure 3.10) is to request the middle tier to iterate the output space. It comprises elements for coordinates and mapping parameters that are identical to the corresponding element in the mapping response (Figure 3.9). These elements comprise the coordinates of the output space before these iterations and mapping parameters that will be used in these iterations. In addition, the iterate mapping request comprises an *Iterations* element and a *MappingType* element. The former defines the number of iterations that will be performed and the latter defines the type of the mapping method that will be used in these iterations. When the middle tier has performed the iterations, it includes the iterated output space in a mapping response (Figure 3.9).

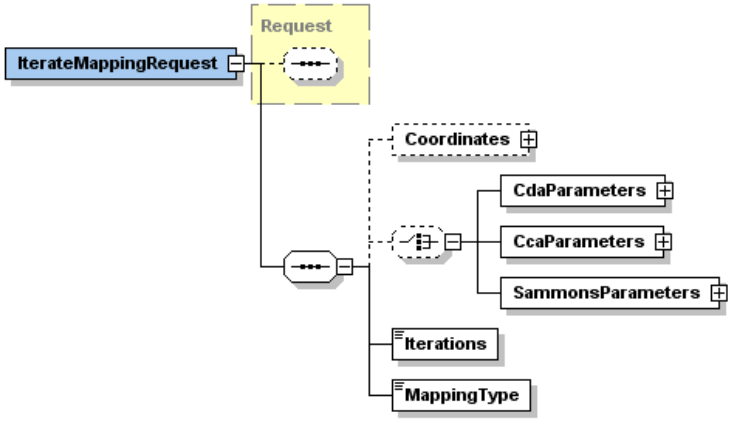


Figure 3.10. XML schema for iterate mapping request.

Topology calculations

The purpose of the topology calculation algorithm is to calculate the clustering coefficient, in- and out-degree distributions for a generic biological network. The mathematical details of these distributions are described in Equations 3.2 and 3.3 in Section 3.3. This algorithm was used in a topology example in a yeast metabolic network (Section 4.3.1) and in a topological enrichment example under oxidative stress (Section 4.3.2). The topology calculation algorithm comprises a topology calculation request and response. Next we will briefly describe these messages and how the middle tier interacts with them.

The topology calculation request (Figure 3.11) comprises a *Graph* element, which is identical to this element in the graph construction response (Figure 3.3), and it comprises a network for which the topology calculation will be performed. Also, it comprises a *TopologyCalculationParameters* element that comprises a Boolean attribute describing whether the distribution will be calculated for in- and out-degrees and another Boolean attribute describing whether the distribution will be calculated for clustering coefficients.

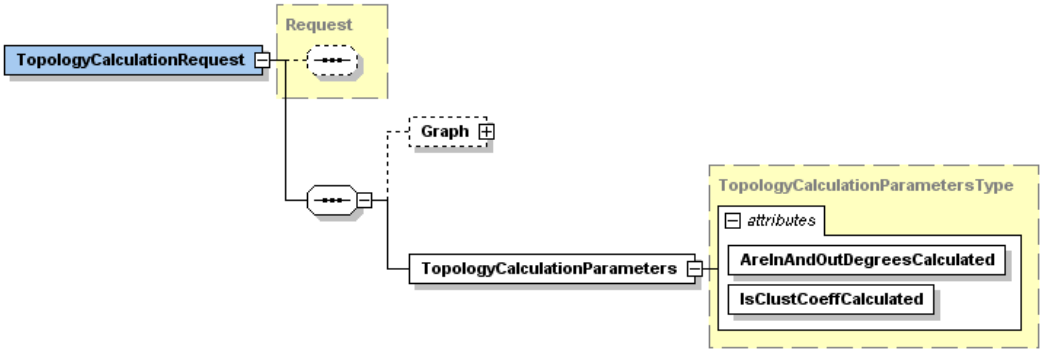


Figure 3.11. XML schema for topology calculation request.

When receiving a topology calculation request, the middle tier calculates selected distribution type(s), and includes the calculated distribution(s) in the topology calculation response (Figure 3.12). More specifically it includes degree and clustering coefficient pairs in a *DegreeAndClustCoeffPair* element and in- and out-degree occurrences in *InDegree* and *OutDegree* elements. All of these elements comprise attributes for node ids that link them to the nodes in the *Graph* element of the topology calculation request (Figure 3.11).

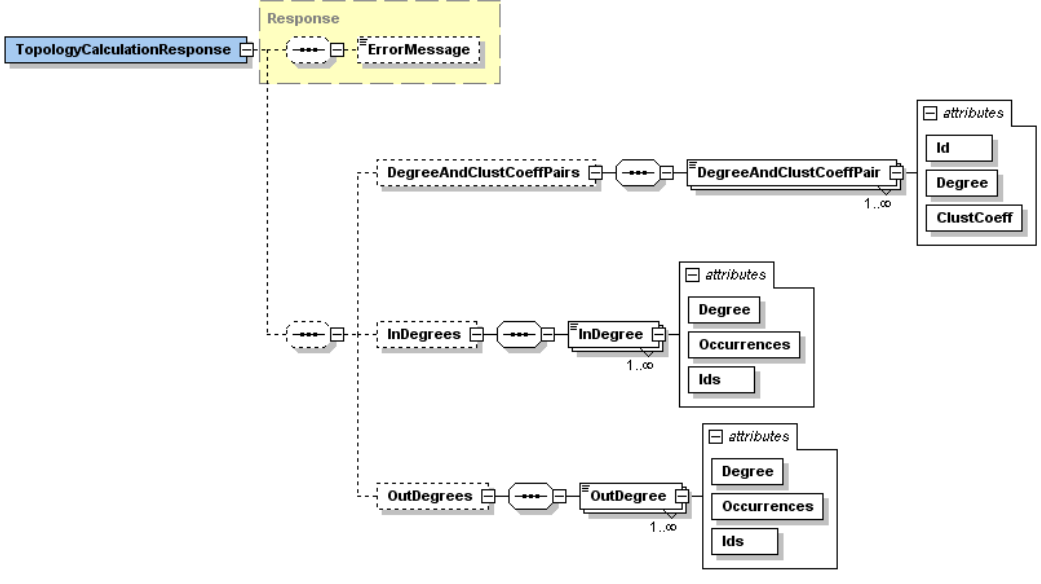


Figure 3.12. XML schema for topology calculation response.

Client

The purpose of the client component is to provide user interfaces for visualizing networks and for performing a context based mapping. We have had three separate user interfaces. In Publications **II–IV** we developed a desktop user interface implemented in Java environment, and the network visualization was implemented by Tom Sawyer Visualization Toolkit 6.0 (Tom Sawyer, Inc.). In Publication **V** we developed an improved user interface. This is also a desktop user interface but it visualizes networks in three dimensions. It is a Microsoft Windows application developed in C# 2.0. It uses Microsoft .NET Framework Version 2.0. The three dimensional visualization is implemented in Microsoft's DirectX 9.0c platform. Also, in Publication **V** we developed a web user interface by using Google Web Toolkit (http://code.google.com/intl/fi/webtoolkit). This user interface takes input parameters from the user, and then uses the middle tier for network construction. The constructed network can be exported to the desktop user interface for visualization or alternatively to Cytoscape (Cline et al., 2007) which a popular generic biological network visualization tool.

Database tier

The database tier comprises all databases that are incorporated in megNet. Most of them are presented in an XML format and they are stored in a Tamino XML server (Software AG) in a Redhat Linux Advanced Server v2.1 environment. In addition, some of the data is presented in a relational database format, and they are stored in an Oracle 10g database server (Oracle, Inc.). In Publications **II–V** we have described in detail for example how the databases have been incorporated, and how the middle tier retrieves data from them. In Table 3.1 we briefly list all databases we currently have in megNet. More extensive description of this data is presented in Peddinti V. Gopalacharyulu's PhD dissertation (Gopalacharyulu, 2010).

Table 3.1. megNet's databases.

Database type	Database names
Protein-protein interaction databases	• BioGRID (Reguly et al., 2006) • DIP (Xenarios et al., 2002) • MINT (Ceol et al., 2010) • BIND (Bader et al., 2003)
Metabolic pathway databases	• KEGG (Kanehisa et al., 2004) • genome-scale yeast metabolic models (Herrgård et al., 2008; Dobson et al., 2010)
Transcriptional regulatory databases	• TransFac (Matys et al., 2003)
Signal transduction databases	• TransPath (Krull et al., 2006)
Compound databases	• PubChem (Wang et al., 2009) • KEGG compounds (Kanehisa et al., 2004)
Ontological databases	• GO (Gene Ontology Consortium, 2008) • OAT (Timonen & Pesonen, 2008)
Gene expression databases	• GEO (Barrett et al., 2009)
Protein and gene sequence databases	• UniProt (UniProt Consortium, 2010) • EMBL (Cochrane & Galperin, 2010)

3.2 EMPath – Enriched Molecular Path detection method

In Publication **I** we have developed the Enriched Molecular Path detection method (EMPath) and showed its utility in the context of type 1 diabetes mouse models. Figure 3.13 shows a schematic pipeline of this method.

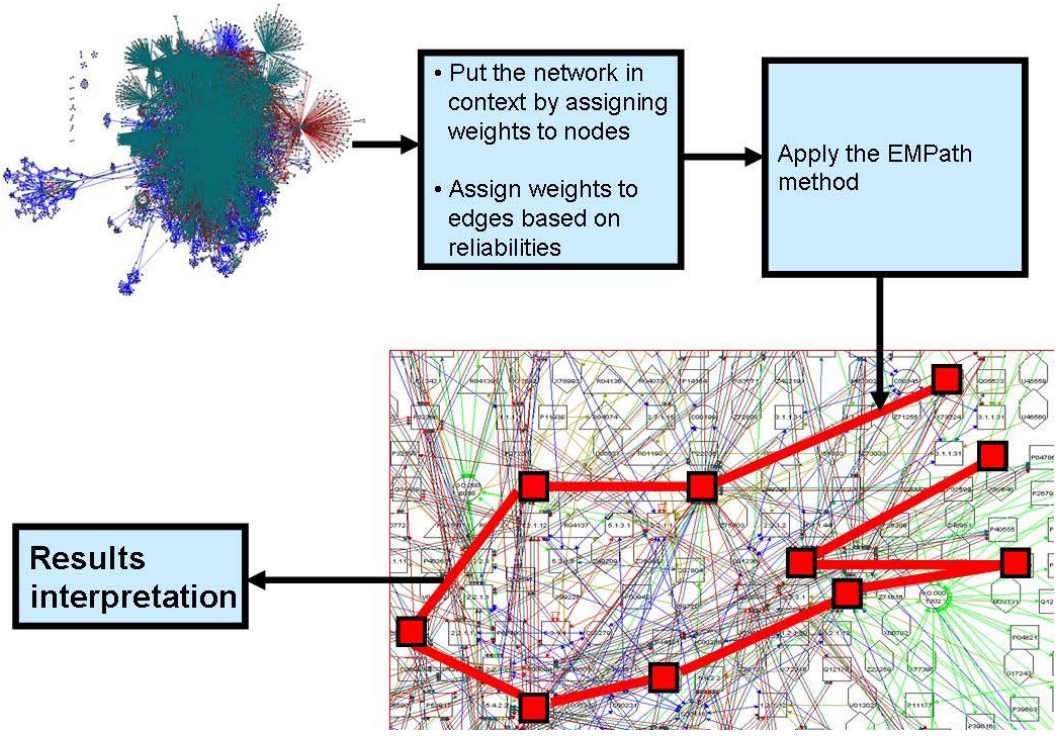


Figure 3.13. The schematic pipeline of the EMPath method.

This method is based on a molecular interaction network that is described in detail in Publications **II**–**V**. Briefly the idea is that the nodes are biological entities (e.g. proteins, metabolites) and the edges are interactions (e.g. protein-protein interactions, metabolic reactions).

We put the network in a phenotypic context by assigning weights to the nodes. Usually this is based on transcriptomics data since it is most easily available, but it can be based on any phenotypic specific data. Also, we assign weights to the edges based on their reliabilities (e.g. reliabilities of protein-protein interactions).

The actual path detection is based on a color coding algorithm (Alon et al., 1995) that was developed to detect optimal paths in a complex network. This method is generic and it is applicable to be used in a complex network of many types. To my knowledge in biology it was first used to detect signaling cascades in a protein-protein interaction network in yeast *Saccharomyces cerevisiae* (Scott et al., 2006). In Publication **I** we tailored this method so that it is suitable for detecting paths in a phenotypic context. Next we will briefly describe our approach to use this method.

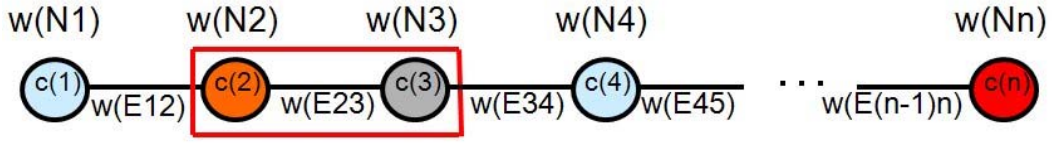


Figure 3.14. The scoring and coloring of the EMPath method.

In the beginning we define the length of the path that will be detected. It can be any integer. Let us denote it by k. In order to score the path, we assign the phenotypic weights to the nodes and the reliability weights to the edges as illustrated in Figure 3.14. Exact scoring formulas are presented in Equations (1–3) of Publication **I** as follows. First we multiply the edge we ights, so a long cascade of unreliable edges gets enough penalty. Then we sum up the node weights. In the end we calculate the total weight by multiplying the edge product and the node sum.

The basic idea of the path search strategy is that we assign colors to the nodes (Figure 3.14) and we allow the detected path to contain a same color only once, which guarantees that the detected path is simple and makes the search algorithm non-greedy since it does not go through all possible branches which would be time-consuming especially in a dense network. The path search strategy is described rigorously in the equation on the next page.

```
// initialize the network by assigning colors
for each node(i) in the network {
 assign a random integer number c(i) from [1, k] to node(i)
}

initialize empty sets :
- P for the detected path
- D for the denied colors

add a node with maximum weight to P, so it will be the first node in the detected path

// add more nodes to P as described in the following loop
for (int i = 2; i ≤ k; i++) {
 initialize a maximum node to be null
 for each neigbor node(n) of the most recently added node {
  assign node(n) to be the maximum node if it satisfies the following conditions :
  - node(n) would lead to a greater total scor of P than the current maximum node
  - the color of node(n) is not in D
 }
 add the maximum node to P and its color to D
}
```

(3.1)

If we do not manage to detect a path by using the procedure described in the previous paragraph, we use a sliding window (Figure 3.14). The idea is that when we are detecting a path, we have a window in which we have most recently taken nodes. The single color requirement applies only to the nodes that are inside the window. For example in Figure 3.14 we have a window of size two that contains grey and pink colors. We have blue in the detected path but the corresponding node is outside the window, so we can add another blue to the detected path. The sliding window makes the path detection faster since there are less denied colors. However, in the end we have to check that the detected path does not contain any cycle, and discard it if it contains. We first try the path detection by using $k-1$ as window size. If we do not manage to find a path, we

decrease the window size by one. We continue this until the window size is one. If we do not manage to find a path with this window size, we conclude that we did not manage to detect a path.

In order to assess the statistical significance of the detected path, we calculate a p-value for it. We shuffle the edge and node weights of the original network 10 000 times. After each shuffle we use the same path detection procedure to detect an optimal in the shuffled network. However, it does not make sense to make all 10 000 shuffles for paths for which the p-value does not look promising. Therefore after each shuffle we check how promising the p-value looks by calculating the percent of shuffles in which the optimal path score is higher in the shuffled network than in the original network. If the percent is greater than 0.025, we discard the path and jump to the next path.

In the end we calculate the p-value for a path for which we managed to perform all 10 000 permutations in the same way as described in the preceding paragraph. If the obtained p-value is less than 0.025, we conclude that the path is statistically significant. Otherwise, we discard the path.

We consider that the network is *harvested* if its all statistically significant paths are detected. However, there is not any rigorous way to investigate this. Therefore, we take a heuristic approach by assuming that the network is *harvested* if we come up with a predefined number (e.g. 50) of consecutive iterations in which the detected path is already detected. Also, we restrict the algorithm to take only a limited number of significant paths (e.g. 2), since it is time-consuming to calculate a p-value for a significant path. We therefore quit detecting paths if we come up with a conclusion that the network is *harvested* or if we have detected enough statistically significant paths.

We can perform the above-described path detection procedure by using different path lengths (e.g. from 3 to 12). After that we can interpret results by studying the detected paths individually and by performing a functional enrichment analysis to associate the detected paths with previously known pathways.

3.3 Topological methods of biological networks

The purpose of this section is to introduce most commonly used complex network concepts in the context of biological networks. In mathematical terms we model a biological network as a graph $G=\{N,E\}$ in which N is a set of nodes and E is a set of edges that connect two elements of N: $E \subseteq [N]^2$. The bio-

logical network can be directed or undirected: in a directed network the order of edge's nodes matters, whereas undirected network it is irrelevant.

Next, I will briefly describe most commonly used topological measures of biological networks that have been summarized for example in a network biology review (Barabási & Oltvai, 2004).

- *Degree*. This measure defines how many edges a node has. Let us denote it by k. In a directed network we usually use two separate measures: *in-degree* and *out-degree*. Let us denote them by k_{in} and k_{out} respectively. The former stands for the number of edges that are targeted to the node, and the latter stands for the number of edges starting from the node.

- *Clustering coefficient*. This measure describes the density of node's neighborhood connections. Let us denote it by C. More specifically, for a node i it is obtained by dividing the number of edges that connect the neighbor nodes of the node i (henceforth n_i) by the number of all possible edges between the neighbor nodes of the node i. In mathematical terms it is defined by $C_i = 2n_i / [k * (k-1)]$. In extreme case this measure obtains one if there are edges between all neighbor nodes, and in the opposite extreme it obtains zero if there is not any edge between the neighbor nodes.

Based on the above-mentioned topological measures we can derive the following distributions that have been commonly used in topological analyses of biological networks. These concepts are also summarized in Barabási & Oltvai (2004).

- *Degree distribution*. This distribution defines the probability that a randomly selected node from a network has a certain degree. It is usually defined separately for in-degrees and out-degrees. These distributions $\mathrm{P}_{\mathrm{in}}(k)$ and $\mathrm{P}_{\mathrm{out}}(k)$ are defined more formally in the equation below.

$$
\begin{aligned}
&N_{tot} = \text{The total number of nodes in the graph} \\
&N_{in}(k) = \text{The number of nodes that have k in - degrees} \\
&N_{out}(k) = \text{The number of nodes that k out - degrees} \\
&\mathrm{P}_{\mathrm{in}}(k) = \frac{N_{in}(k)}{N_{tot}} \\
&\mathrm{P}_{\mathrm{out}}(k) = \frac{N_{out}(k)}{N_{tot}}
\end{aligned}
\tag{3.2}
$$

- *Clustering coefficient distribution*. This distribution stands for the probability that a randomly selected node from the network has a certain clustering coefficient. It is defined only for an undirected network. This distribution $\mathrm{C}(k)$ is more formally presented in the equation below.

$$
\begin{aligned}
&C_n(k) = \text{The number of nodes of which clustering coefficient is k} \\
&\mathrm{C}(k) = \frac{C_n(k)}{N_{tot}}
\end{aligned}
\tag{3.3}
$$

Next, I will briefly describe a few widely used biological network models that use the above-mentioned distributions. These models are also described in detail in Barabási & Oltvai (2004) except that the truncated power-law is described in Khanin & Wit (2006).

- *Erdős-Rényi random network model.* In the Erdős-Rényi random network model (Erdős & Rényi, 1959; 1960) N_{tot} nodes are connected randomly to each other with probability p. The degree distributions of this model $\mathrm{P}_{\mathrm{in}}(k)$ and $\mathrm{P}_{\mathrm{out}}(k)$ are rapidly increasing and decreasing bell shaped curves having a small average value (e.g. 2–3). This means that almost all nodes have only a few links, and there are no highly connected nodes. The clustering coefficient distribution $\mathrm{C}(k)$ is a straight horizontal line in this model, which means that the clustering coefficient is independent of a node's degree.

- *Power-law degree distribution model*[6]. In the power-law degree distribution model (Barabási & Albert, 1999) the degree distributions $P_{in}(k)$ and $P_{out}(k)$ differ from the degree distributions of the Erdős-Rényi random network model, and they are of form $k^{-\lambda} * e^{-k}$, in which λ is a degree exponent. These degree distributions are linearly decreasing in log-log scale. Like in the Erdős-Rényi random network model the clustering coefficient distribution $C(k)$ is a straight horizontal line meaning that also in this model the clustering coefficient is independent of a node's degree.

- *Truncated power-law degree distribution model*. This distribution is a truncated version of the power-law degree distribution model: it follows the power-law only in small numbers, which means that the network follows the power-law within the range $1 \leq k < k_c$. This distribution is defined more rigorously in the equation below.

$$
\begin{aligned}
&k_c = \text{The cut-off value } (\succ 1) \\
&P_{in}(k) = k^{-\lambda} * e^{-(k/k_c)} \\
&P_{out}(k) = k^{-\lambda} * e^{-(k/k_c)}
\end{aligned}
\tag{3.4}
$$

- *Hierarchical network model.* The hierarchical network model (Ravasz et al., 2002; Ravasz & Barabási, 2003) combines the power-law degree distribution, modularity and local clustering into one model. The basic idea is that the network has a pyramid structure in which modules are organized in a hierarchical manner: in the low level there are highly connected modules, and in the upper level there are loosely connected modules. The clustering coefficient distribution $C(k)$ is thus linearly decreasing in log-log scale. The degree distributions $P_{in}(k)$ and $P_{out}(k)$ are also linearly decreasing in log-log scale since in the high level there are only few highly connected nodes, whereas in the lower level there are quite many loosely connected nodes.

---

[6] In some contexts this model is called scale-free network model. However, it is pointed out that the concept of scale-free tends to be ambiguous (Lima-Mendez & Helden, 2009), so I do not use it in this thesis.

3.4 TEAFS – Topological Enrichment Analysis for Functional Subnetworks

In Publication **VI** we have developed the Topological Enrichment Analysis of Functional Subnetworks method (TEAFS) and showed its utility in the context of oxidative stress in yeast *Saccharomyces cerevisiae*. Figure 3.15 shows a schematic pipeline of this method.

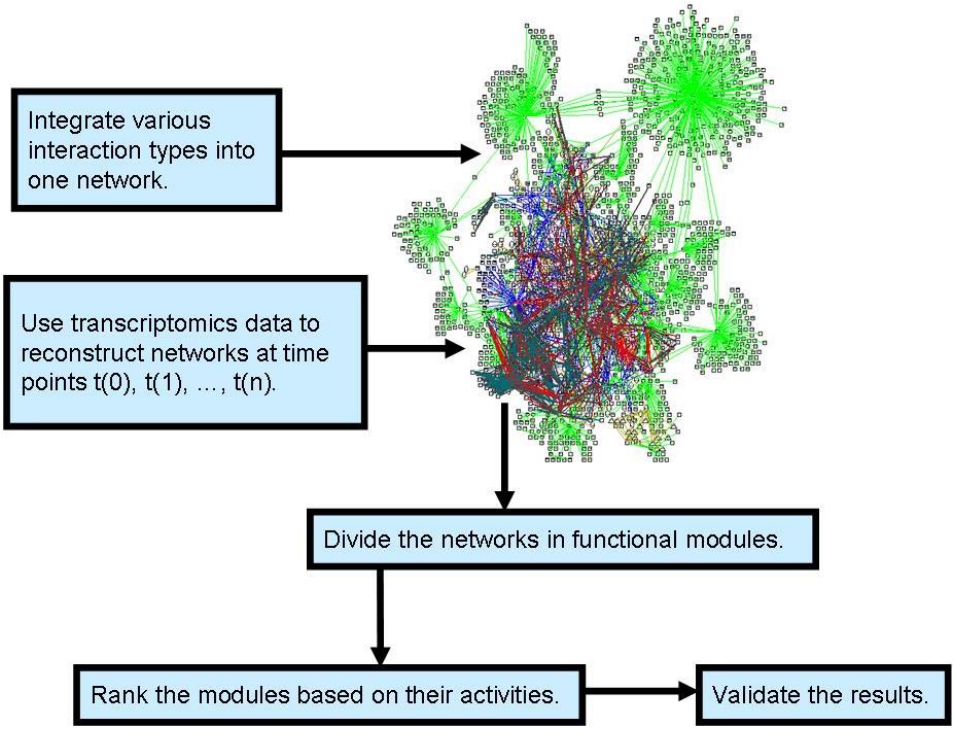


Figure 3.15. The schematic pipeline of the TEAFS method.

The TEAFS pipeline starts from a construction of a megNet network: integration various interaction types into one network. This network can comprise any type of molecular interactions, for example protein-protein interactions, metabolic reactions, transcriptional regulations.

We reconstruct networks at time points by using a time series of a transcriptomics data set. This is based on a method that was introduced in a dynamic network topology study (Luscombe et al., 2004). We first reconstruct a reference network at time point t(0) by taking all protein nodes of which encoding genes

are in the transcriptomics data set. Then at each time points t(1), t(2), …, t(n) we reconstruct a network by removing protein nodes and their incident edges based on the expressions of their encoding genes. This requires that the transcriptomics data set is of dual channel (Schena et al., 1995). In order to decide whether we remove a protein node and its incident edges, we first divide the log-transformed values of the control channel intensities in high, medium and low by using a k-means clustering algorithm (Lloyd, 1982). Then we use a change between the case and control intensities, and deduce that it is either up, constant or down. Then based on the control condition intensity level and change between case and control intensities we use Table 4 of Publication **VI** to decide whether we remove the protein node and its incident edges.

We divide the networks in functional modules based on a biological criterion. It can be for example based on protein's and gene's involvement in GO biological processes (Gene Ontology Consortium, 2008) or in metabolic pathways (Kanehisa et al., 2004).

We rank the functional modules based on their activities in terms of three topological measures: in-degree, out-degree and clustering coefficient that are described in more detail in Section 3.3 More specifically, we first calculate a de-activation ratio for each module at each time interval [t(i), t(i+1)] by dividing the sum of a topological measure of proteins that are present at time t(i) but absent at time t(i+1) by the sum of proteins that are present at time point t(i). Then for each module we perform 10 000 permutations in terms of each topological measure in order to calculate p-values rejecting the null hypothesis stating that proteins are deactivated uniformly in the whole network. In each permutation we create a 'random module' by removing each protein at each time interval with probability of the corresponding de-activation ratio. The p-value is obtained by dividing the number of permutations in which the activity of the topological measure in the random module is at least as much as it is in the original module by the number of all permutations (10 000). Then we correct the p-values from multiple comparisons by using Bonferroni correction, and calculate False Discover Rate (FDR) q-values. We consider modules of which q-value is less than 0.05 as statistically significant.

In the end we validate the results: figure out if the detected activities of functional modules under the given condition make sense. We can do this for example by in-house metabolomic experiments or by literature survey.

4. Results and discussion

In this chapter we present the main results of this thesis. In Section 4.1 we show a few integrative biological data visualization examples in megNet. In Section 4.2 we show the utility of the Enriched Molecular Path detection method (EMPath) in the context of type 1 diabetes. In Section 4.3 we show network topology studies carried out in this thesis.

4.1 Integrative biological data visualization in megNet

In this section we show the basic idea of megNet: the ability to visualize biological data across multiple interaction levels and the ability to enable context based inference. In Section 4.1.1 we show that megNet has potential for interesting novel hypotheses by an example in which a protein-protein interaction connects two enzymes that are from each other in metabolic level in yeast *Saccharomyces cerevisiae*. In Section 4.1.2 we show that megNet can be used for context based mapping by an example in which a Gene Ontology biological process (Gene Ontology Consortium, 2008) categorizes biological entities involved in yeast metabolism into two groups. In Section 4.1.3 we apply these approaches to a medical context: we show cross-talk and context based mapping examples in the context of medical image data leading to interesting associations between biological networks and medical image data.

4.1.1 Cross-talk in yeast metabolism

There has been evidence that between different biological interaction levels there is cross-talk leading to interesting phenotypes (Papin & Palsson, 2004; Lee et al., 2008; Li et al., 2010). In Publication **II** we showed how megNet can be used to find this kind of cross-talk by constructing an integrated metabolic (KEGG; Kanehisa et al., 2004) and protein-protein interactions (MINT; Ceol et

al., 2010, BIND; Bader et al., 2003) network in yeast *Saccharomyces cerevisiae*. We included *Glycolysis/Gluconeogenesis*, *Pentose phosphate pathway* and *Citrate cycle* metabolic pathways along with their protein-protein interactions in this network. As a result we obtained a network that is visualized in Figure 5 of Publication **II**. We can see that there are quite much protein-protein interactions making cross-talk between different stages of metabolism. For example, there are two enzymes: *phosphoglycerate kinase* and *acetate-CoA ligase* that are quite far from each other in metabolic level: the former is involved in the starting point of *citrate cycle*, whereas the latter is involved in the second phase of *glycolysis*. However, both of these enzymes interact with an *SRB2 protein* detected by the yeast two-hybrid method (Uetz et al., 2000; Ito et al., 2000; Fields, 2005). There is evidence that the *SRB2 protein* is involved in transcriptional initiation (Thompson et al., 1993), which could be a sign that the se two enzymes are co-regulated at different stages of metabolism. However, it is good to keep in mind that the yeast two-hybrid method notoriously produces quite much false-positive protein-protein interactions (Mrowka et al., 2001). However, we believe that this cross-talk can shed light on novel hypotheses.

4.1.2 Context based visualization in yeast metabolism

In Publication **III** we integrated Gene Ontology biological process terms (Gene Ontology Consortium, 2008) with a metabolic pathway network (KEGG; Kanehisa et al., 2004) in yeast *Saccharomyces cerevisiae* by using megNet. In Figure 6 of Publication **III** there is a zoomed region from the neighborhood of a *citrate cycle* biological process term. We performed a context based mapping by assigning low weights to the incident edges of the *citrate cycle* biological process term and then mapping the internal distances into two dimensions by using the CDA mapping method. The results are presented in Figure 7 of Publication **III**. We can see that there are two clusters. This may be a sign that the *citrate cycle* biological process divides metabolic reactions in two main groups: one group of reactions that are strongly involved in *citrate cycle* and another group of reactions that are weakly involved in *citrate cycle*.

4.1.3 Network visualization in context of medical image data

It is becoming clear that there is need to integrate biological networks with medical images (Walter et al., 2010), and as a practical example it recently came out

a publication in which biological networks were studied in the context of human brain images (Bassett et al., 2011). In Publication **V** we continued these directions by visualizing biological networks in megNet in the context of Lamin A/C image data. As a background study, we had previously derived Magnetic Resonance (MR) image parameters from Lamin A/C mutation patients (Koikkalainen et al., 2008). In a follow-up study we had performed lipidomics analysis in the same patients, and developed a statistical model to find associations between the lipidomics profiles and medical image parameters (Sysi-Aho et al., 2011). In order to understand these associations better, in Publication **V** we used megNet to construct a biological network in the context of the same lipidomics profiles. More specifically, we first constructed *glycerophospho-*, *glycero-* and *sphingolipid* metabolic pathways from KEGG (Kanehisa et al., 2004) in *homo sapiens*, and mapped molecular lipid species to their generic lipid names on these pathways by using the biochemical knowledge of the side chain length and saturation, as described in Yetukuri et al. (2007). Then we integrated these pathways with protein-protein interactions from BioGrid (Reguly et al., 2006), DIP (Xenarios et al., 2002) and MINT (Ceol et al., 2010), ontological relationships from OAT (Timonen & Pesonen, 2008) and GO (Gene Ontology Consortium, 2008), and gene-protein relationships from EMBL (Cochrane & Galperin, 2010). The constructed network is visualized in Figure 6 of Publication **V**. In the same vein as in the example in Section 4.1.1 we can see that also between metabolic reactions in this figure there is quite dense cross-talk via many interaction levels.

A cross-talk example is visualized in Figure 7 of Publication **V**. There seems to be signaling between two isoforms of *phospholipase A2* (Coffey et al., 2004). One of these isoforms catalyzes a metabolic reaction in which a product comprises molecular lipid species that correlated quite strongly with image parameters in our previous case study (Sysi-Aho et al., 2011), whereas the other isoform catalyzes a metabolic reaction in which a substrate comprises molecular lipid species for which the correlation was not so obvious. Maybe the signaling between the isoforms of *phospholipase A2* has some role in these correlations. For example, it may regulate the activities of the phospholipases.

Another cross-talk example is visualized in Figure 8 of Publication **V**. In this figure there are two isoforms of *endothelial lipase*: one of them breaks down *1,2-Diacyl-sn-glycerol* and the other one breaks down *triacylglycerol*. Both of these lipases are involved in the *cholesterol transport and homeostasis* biological processes. In our previous case study (Sysi-Aho et al., 2011) triglyceride molecular lipid species were associated with increased end-diastolic wall thick-

ness. This may be a sign that cholesterol metabolism has some role in this association: it may be associated with the increased end-diastolic wall thickness. Also, from this figure we can see that between the endothelial lipases there are associations that have been detected by our in-house text mining system OAT (Timonen & Pesonen, 2008). This system detected one article suggesting that these lipases are associated with diabetes prevention (Mizuno et al., 2004), and another article suggesting that they are associated with maintenance of cell homeostasis (Mi et al., 2004). From the former observation we could make tentative conclusion that the end-diastolic wall thickness prevents type 1 diabetes, and from the latter observation we could conclude that the end-diastolic wall thickness may have important role in the maintenance of cell homeostasis in diabetes development.

In order to gain our understanding of the role cholesterol metabolism in the association between *triacylglycerol* and end-diastolic wall thickness, we performed a mapping in the context of cholesterol metabolism, in the same vein as we performed a mapping in the context of *citrate cycle* in Section 4.1.2. More specifically, we assigned low weights to the incident edges of the nodes corresponding to the cholesterol biological processes that were associated with the endothelial lipases in the previous paragraph. The results of this mapping are presented in Figure 9 of Publication **V** in which there is a zoom from the neighborhood of *triacylglycerol*. This figure comprises for example a kinase and a *receptor signaling* biological process, which could give a hint that maybe a receptor signaling cascade stimulates the *triacylglycerol* to participate in cholesterol metabolism and in turn associates it with the increased end-diastolic wall thickness. Also, this figure comprises a *'regulation of macrophage activation'* biological process. As supporting evidence there has been discussion that macrophages may play critical role in the pathogenesis of type 1 diabetes (Yang, 2008). Also, this could be related to the observation that we made in the previous paragraph suggesting that the end-diastolic wall thickness might prevent type 1 diabetes.

4.2 Enriched molecular path detection case study in type 1 diabetes

In Publication **I** we used the Enriched Molecular Path detection method (EMPath) in an integrated protein-protein interaction (BIND; Bader et al., 2003, MINT; Ceol et al., 2010, DIP; Xenarios et al., 2002), signal transduction

(TransPath; Krull et al., 2006) and metabolic network (KEGG; Kanehisa et al., 2004) in the context of transcriptomics data from Non-Obese Diabetic (NOD) mouse models (Vukkadapu et al., 2005). This data set comprises measurements from pancreas of four NOD mouse strains from 3 week old animals: BDC2.5/NOD, NOD, BDC2.5/NOD.scid, and NOD.scid. These strains have differences in terms of insulitis[7] and type 1 diabetes development. We detected molecular paths in two case-control settings. In one case-control setting we compared BDC2.5/NOD versus NOD, since the BDC2.5/NOD has more accelerated insulitis development. In the other case-control setting we compared BDC2.5/NOD.scid versus NOD.scid, since BDC2.5/NOD.scid has more accelerated type 1 diabetes development. So, in these case-control settings our purpose was to detect pancreas specific paths that are associated with early insulitis and type 1 diabetes development. In both case-control settings we detected separately up- and down-regulated paths. In Vukkadapu et al. (2005) these strains were studies in the context of type 1 diabetes related genes. Our purpose was to gain understanding of these genes by detecting their interactions.

The mathematical details of this method are described in Section 3.2. In this case study we obtained the node weights for protein nodes by calculating gene expression intensities between case and control strains of their encoding genes. We obtained the edge weights based on the evidence that a protein interaction from BIND (Bader et al., 2003) is quite unreliable (Futschik et al., 2007), and interactions and reactions from the other databases are reliable. Therefore, we assigned 0.33 to a protein-protein interaction edge if the interaction was curated only into the BIND database (Bader et al., 2003). We assigned 1.0 to edges from the all other databases (MINT; Ceol et al., 2010, DIP; Xenarios et al., 2002, KEGG; Kanehisa et al., 2004, TransPath; Krull et al., 2006). In the network harvesting we used 50 as the maximum number of consecutively detected paths and 2 as the maximum number of statistically significant paths.

As a result we obtained several statistically significant up- and down-regulated paths in both case-control settings. As a most surprising finding many lipid paths were down-regulated in early insulitis. Especially, an ether phospholipid synthesis path was down-regulated. This is an interesting finding, since serum ether lipids were diminished children who later progressed to type 1 diabetes in comparison with healthy children in a previous study (Orešič et al.,

[7] Pre-state of type 1 diabetes when pancreatic beta cells get inflammated.

2008). The ether phospholipids synthesis path contained plasmalogens that have previously found to protect cellular functions from oxidative damage (Zoeller et al., 1999; Zoeller et al., 2002). Also, there is evidence that pancreatic beta cells are particularly susceptible to oxidative damage (Lenzen et al., 1996; Cnop et al., 2005). Maybe this is a sign that oxidative stress destroys pancreatic beta cells during the progression to type 1 diabetes.

In order to elucidate the biological meaning of the detected paths, we associated their enrichment with previously known pathways in a Molecular Signature Database (Subramanian et al., 2005). As a result we obtained a summary for the whole case study. In early insulitis phosphorilation pathways were up-regulated that is probably associated with altered cell signaling, and lipid metabolism was down-regulated. In type diabetes development paths related to cell communication were up-regulated, and nucleotide and nucleoside metabolism were down-regulated that was probably related to cell cycle and DNA repair.

4.3 Network topology studies

In this section we go through network topology studies carried out in this thesis. In Section 4.3.1 we show an example in which we performed topological calculations on a static yeast metabolic network to investigate whether ubiquitous complex network properties are present. In Section 4.3.2 we describe how we develop the Topological Enrichment Analysis for Functional Subnetworks method (TEAFS). We first show how we investigated whether ubiquitous complex network properties are present in reconstructed yeast networks under a time series of an oxidative stress gene expression data set. Also in this section we describe how these results gave motivation to tailor the TEAFS method in order to gain our biological understanding by analyzing modules of networks.

4.3.1 Topology example in yeast metabolism

In Publication **III** we constructed a complete metabolic network for yeast *Saccharomyces cerevisiae* from KEGG (Kanehisa et al., 2004). The constructed network is visualized in Figure 3 of Publication **III**. As briefly mentioned in Section 3.3 linearly decreasing degree distribution in log-log scale and constant clustering coefficient are considered to imply that a biological network follows the power-law degree distribution model, and linearly decreasing degree and clustering coefficient distributions as the hierarchical network model. Therefore

in Publication **III** we calculated these distributions for the yeast metabolic network, which are presented in Figures 4 and 5 of this publication. We can see that the degree distribution is not linearly decreasing, and that the clustering coefficient distribution is not linearly decreasing and not constant. It thus seems that this network does not follow the power-law degree distribution and hierarchical network models that were initially observed to be present in many biological networks: metabolic networks (Jeong et al., 2000) and protein-protein interaction networks (Jeong et al., 2001; Wagner, 2001; Giot et al., 2003; Li et al., 2004; Yook et al., 2004). Our observation supports the critiques presented in Khanin & Wit (2006) stating that most biological networks actually do not ideally follow the ubiquitous complex network properties.

4.3.2 Topological enrichment in yeast under oxidative stress

In the previous section we demonstrated that ubiquitous complex network properties cannot really be applied to biological networks. In this section we use the Topological Enrichment Analysis for Functional Subnetworks method (TEAFS) to study topological properties of a yeast network. This method is biologically more meaningful than the example in the previous section. Firstly, the example in the previous section was done in static manner. However, in reality in biology everything is dynamic, so the current trend is to study network properties in dynamic manner (Luscombe et al., 2004; Klipp, 2007). The TEAFS method addresses this issue by enabling using a time series of a transcriptomics data set when studying topological properties. More specifically, we used a transcriptomics data set from oxidative stress (Gasch et al., 2000). In addition, another limitation of the example in the previous section was the fact that it was done solely on metabolic level. However, there has been evidence that in biology phenotypes usually result from interplay of many interaction levels (Papin & Palsson, 2004; Lee et al., 2008; Li et al., 2010). We also addressed this issue by taking protein-protein interactions and transcriptional regulations along with metabolic level. More specifically, we took all metabolic reactions from KEGG (Kanehisa et al., 2004), transcriptional regulations from TransFac (Matys et al., 2003) and protein-protein interactions from DIP (Xenarios et al., 2002) in yeast *Saccharomyces cerevisiae*. In this network nodes are proteins, metabolites, genes and DNA binding sites, and edges are interactions and reactions.

We first reconstructed a reference network and networks at time points in the way as described in Section 3.4. We investigated whether these networks follow

the power-law degree distribution and hierarchical network models by studying their degree and clustering coefficient distributions. We came up with the same observation as in the example in the previous section: none of these networks followed the above-mentioned models. The results are visualized in Figure 4.1–4.3[8] comprising in- and out-degree and clustering coefficient distributions for the reference and networks at time points. From all of these networks we can see the same result as we saw in the static yeast metabolic network in the previous section: the degree distribution is not linearly decreasing, and the clustering coefficient distribution is not linearly decreasing and not constant. We therefore concluded that we cannot apply the previous findings related to the ubiquitous complex network properties (Barabási & Oltvai, 2004) to this case study, and we realized that it is good to tailor the method. Therefore, we decided to divide the network in functional modules based on their Gene Ontology biological process (Gene Ontology Consortium, 2008) memberships in the way as described in Section 3.4. The modularity has been shown to be an important concept when studying biological networks in dynamic manner (Qi & Ge, 2006).

[8] These results are not included in Publication **III** because of lack of space. They have been placed here in order to elevate their importance.

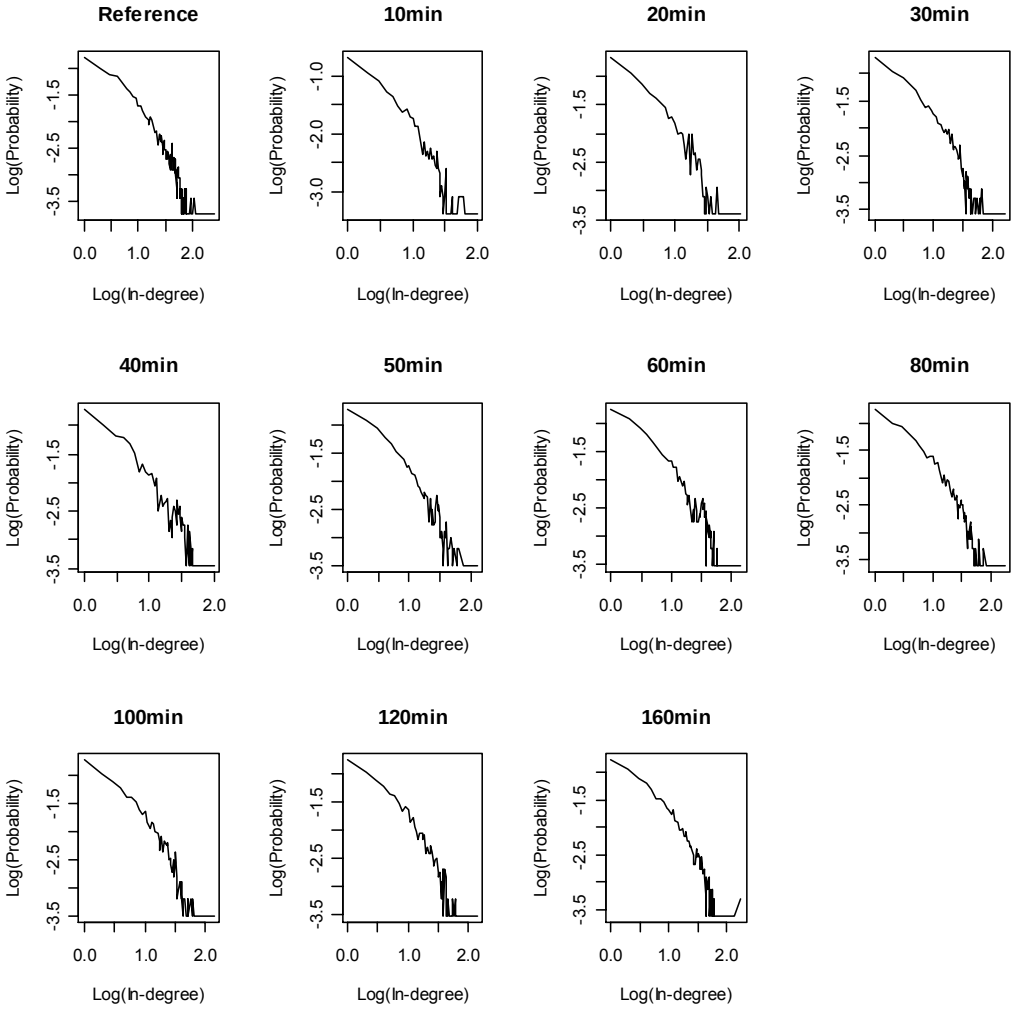


Figure 4.1. In-degree distributions for reference and networks at time points.

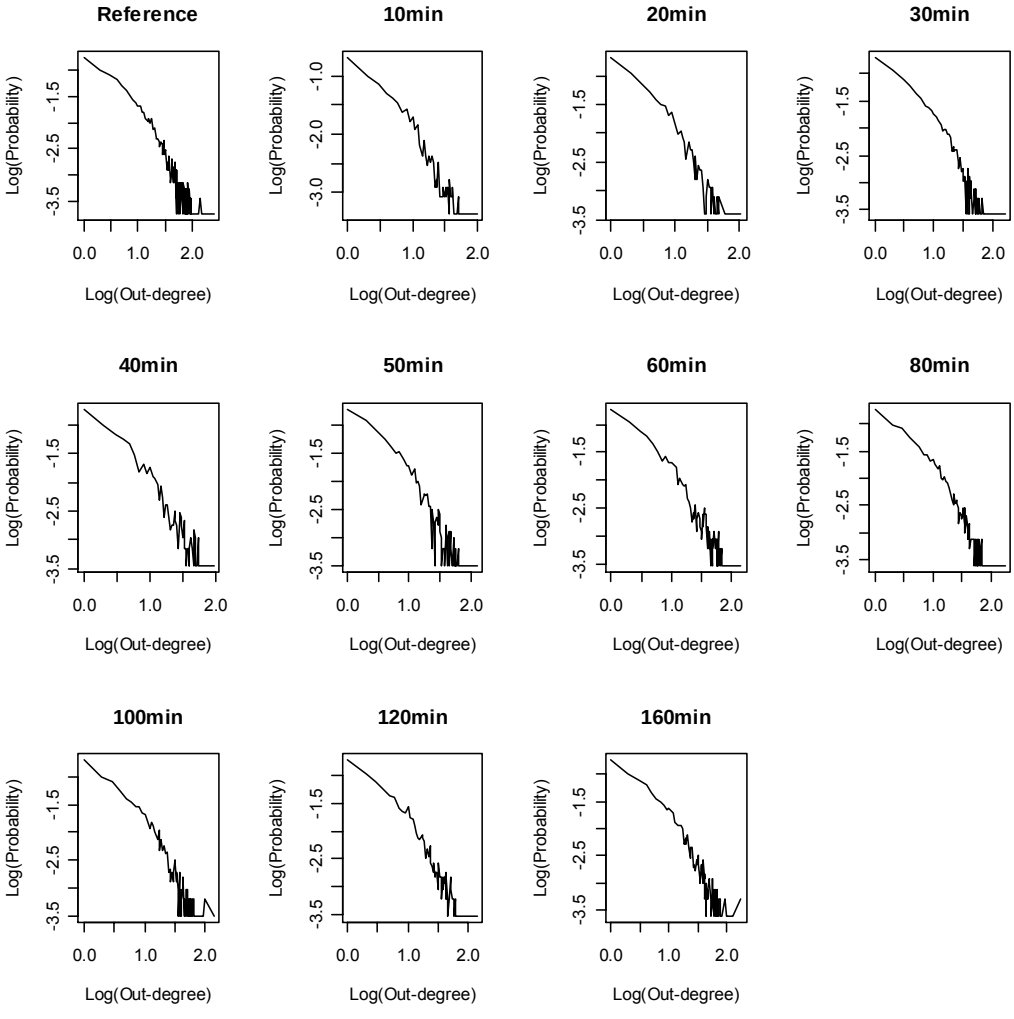


Figure 4.2. Out-degree distributions for reference and networks at time points.

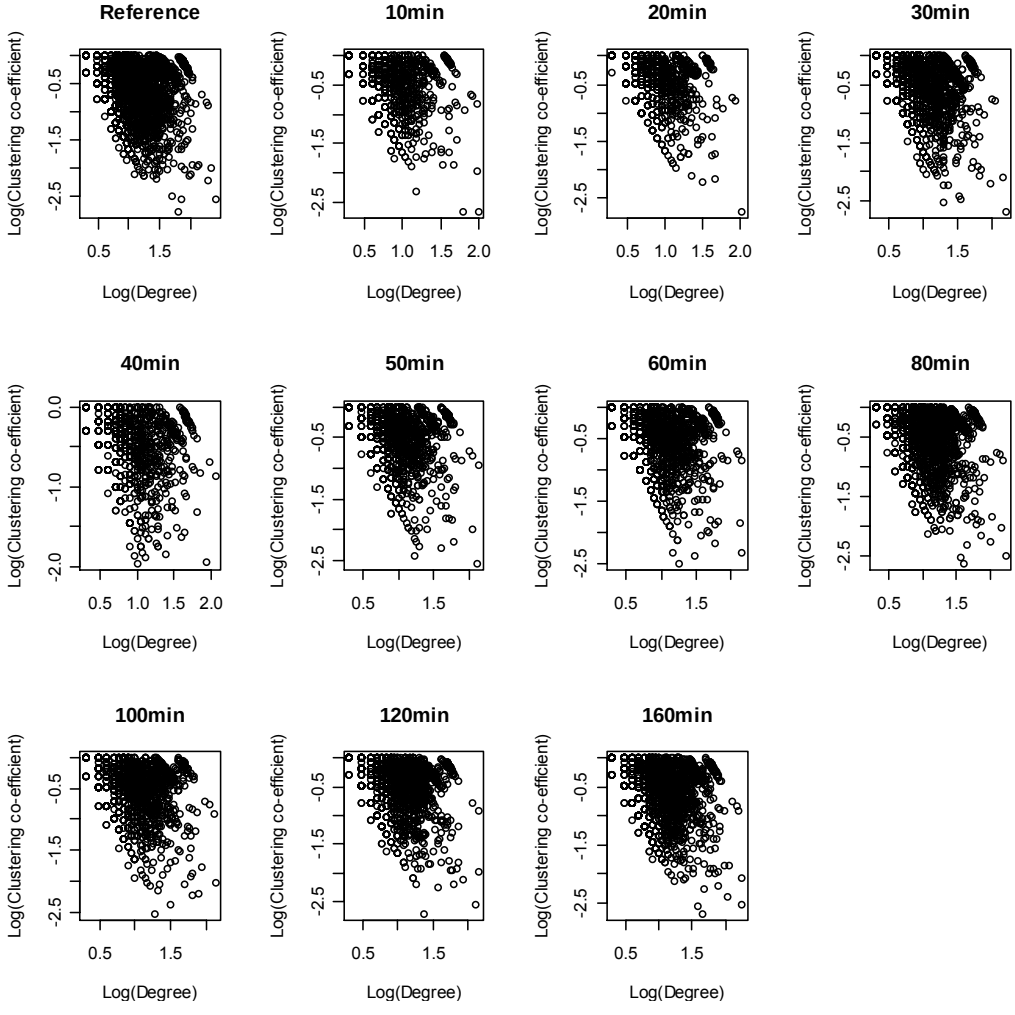


Figure 4.3. Clustering coefficient distributions for reference and networks at time points.

Before starting the actual TEAFS method we calculated average clustering coefficient over the time series for each module. We selected modules of which average clustering coefficient were significantly more than zero for further analysis. After that we performed the TEAFS method for the remaining modules in the way as described in Section 3.4.

As a result of the module activity analysis, we found for example that lipid metabolism and phospholipid biosynthesis modules were highly active. We validated our results by performing in-house metabolomic analysis under dynamic

response to oxidative stress in our laboratory. As a result, we found that the concentrations of precursors of ceramide biosynthesis increased over time. We may thus conclude that it seems that dynamic modules lead to the accumulation of toxic lipids such as ceramides under oxidative stress.

5. Summary and conclusions

In the research related to this thesis we used a network biological approach to address various present day challenges of systems biology. We set up a visualization system for heterogeneous biological data to address biologists' need for integrative visualization (Gehlenborg et al., 2010; O'Donoghue et al., 2010). We showed the utility of this system by a few examples. First we showed how protein-protein interactions make cross-talk between different stages of yeast metabolism leading to novel hypotheses. In the second example we used a context based mapping to show how a Gene Ontology biological process term (Gene Ontology Consortium, 2008) categorizes yeast metabolism into two parts. Then we applied these approaches to a medical context: we showed a case study in which we integrated our in-house lipidomics data into a biological network. We showed two examples demonstrating how interactions between metabolic reactions could possibly explain our previous associations between biological data and medical images, and one example demonstrating how biological entities are related to each other in a medical context.

In addition, we developed the Enriched Molecular Path detection method (EMPath). We showed a case study in which this method was used in the context of type 1 diabetes mouse models. As a most interesting result, we found that ether phospholipid biosynthesis was down-regulated in early insulitis, consistently with a previous study in which serum ether lipids were diminish in children who later progressed to type 1 diabetes in comparison with healthy children, which indicates that this method is capable for novel findings in molecular level. In addition, we performed topological calculations on biological networks to investigate whether they follow ubiquitous complex network properties, and in contrast to initial tentative findings in complex network theory we observed that the ubiquitous complex network properties are not present in these networks, which is consistent with recent critiques to the ubiquitous complex network

properties (Lima-Mendez & Helden, 2009). We therefore tailored a method called Topological Enrichment Analysis of Functional Subnetworks (TEAFS) so that it analyzes modules of networks. We showed that this method is capable of predicting the accumulation of toxic lipids in yeast *Saccharomyces cerevisiae*, which we validated by in-house metabolomic analysis.

Naturally there are many remaining challenges. For example, megNet has potential to be extended to other usages. One possible direction is to progress in integration with lipid pathway reconstruction methods that are presented in Laxmana R. Yetukuri's PhD dissertation (Yetukuri, 2010). We have already done some preliminary work in this direction, for example in the medical data image data case study (Section 4.1.3) we used megNet to integrate lipidomics data into a molecular interaction network.

Also, I believe the EMPath method can be used in the context of any phenotype. In this thesis we showed its utility in the context of type 1 diabetes mouse models but the same should work in many other case studies. We have already been using it in the context of microbial and other type 1 diabetes mouse strains. Preliminary results have shown that this method seems to be capable of making interesting findings also in these studies. For example, we have used it to detect metabolic paths associated with the correlation of gene expression and protein production rate in a fungal species (Arvas et al., submitted).

In addition, I think megNet would benefit from being publicly available as pointed out in Publication **V**. It is probably not reasonable to make the whole megNet publicly available because of e. g. restrictions in database licenses. However, it would make sense to make parts of megNet publicly available, for example network construction could be implemented as an open source Cytoscape plug-in, which could lead to good complementary efforts between Cytoscpape (Cline et al., 2007) and megNet: Cytoscape is a popular generic network visualization tool and megNet would provide a data integration framework for Cytoscape. Also, the EMPath method would probably benefit from being publicly available. This would enable anybody in the systems biology community to use the method in the context of his or her data, which would probably lead to many novel findings. For example, Gene Set Enrichment Analysis method (GSEA) (Subramanian et al., 2005) is publicly available, and it is widely used in the systems biology community.

In addition, megNet would probably benefit from better usability. In order to address this challenge, we have been implementing user interfaces as web applications. As first step towards this effort, we separated a part of the user interface into a web application in Publication **V**.

References

Aittokallio, T., and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. Briefings in Bioinformatics, 7(3):243–255.

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

Alon, N., Yuster, R., and Zwick, U. (1995). Color-coding. J. ACM, 42:844–856.

Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004). Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". Science, 305(5687):1107.

Arvas, M., Pakula, T., Smit, B., Rautio, J., Koivistoinen, H., Jouhten, P., Lindfors, E., Wiebe, M., Penttilä, M., and Saloheimo, M. Correlation of gene expression and protein production rate – a system wide study. Submitted.

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

Barabási, A.-L., and Albert, R. (1999). Emergence of Scaling in Random Networks. Science, 286(5439):509–512.

Barabási, A.-L., and Oltvai, Z. N. (2004). Network Biology: Understanding the Cells's Functional Organization. Nature Reviews/Genetics, 5:101–113.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., and Edgar, R. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Research, 37(Database issue):D885–D890.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. Proceedings of the National Academy of Sciences of the United States of America.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgård, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature Protocols, 2(3):727–738.

Breitkreutz, B. J., Stark, C., and Tyers, M. (2003). Osprey: a network visualization system. Genome Biology, 4(3):R22.

Ceol, A., Chatr, A. A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. Nucleic Acids Research, 38:532–539.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. Molecular Systems Biology, 3:140.

Chuang, H. Y., Hofree, M., and Ideker, T. (2010). A decade of systems biology. Annual Review of Cell and Developmental Biology, 26:721–744.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. (2007). Integration of Biological Networks and Gene Expression Data using Cytoscape. Nature Protocols, 2(10):2366–2382.

Cnop, M., Welsh, N., Jonas, J. C., Jörns, A., Lenzen, S., and Eizirik, D. L. (2005). Mechanisms of pancreatic beta-cell death in type 1 and type 2 diabetes: many differences, few similarities. Diabetes, 54:S97–S107.

Cochrane, G. R., and Galperin, M. Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Research, 38(Database issue):D1–D4.

Coffey, M. J., Coles, B., Locke, M., Bermudez-Fajardo, A., Williams, P. C., Jarvis, G. E., and O'donnell, V. B. (2004). Interactions of 12-lipoxygenase with phospholipase A2 isoforms following platelet activation through the glycoprotein VI collagen receptor. FEBS Letters, 576(1–2):165–168.

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

Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., and Ozturk, M. (2002). PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics, 18(7):996–1003.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R.,

Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Reubenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K. H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Le Novère, N., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. Nature Biotechnology, 28(9):935–942.

Dobson, P. D., Smallbone, K., Jameson, D., Simeonidis, E., Lanthaler, K., Pir, P., Lu, C., Swainston, N., Dunn, W. B., Fisher, P., Hull, D., Brown, M., Oshota, O., Stanford, N. J., Kell, D. B., King, R. D., Oliver, S. G., Stevens, R. D., and Mendes, P. (2010). Further developments towards a genome-scale metabolic model of yeast. BMC Systems Biology, 4(1):145.

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 Research, 14:1298–1309.

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. Proceedings of the National Academy of Sciences of the United States of America, 104(6):1777–1782.

Erdős, P., and Rényi, A. (1959). On random graphs. Publicationes Mathematicae, 6:290–297.

Erdős, P., and Rényi, A. (1960). On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy of Science, 5:17–61.

Feist, A. M., and Palsson, B. Ø. (2008). The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nature Biotechnology, 26:659–667.

Fields, S. (2005). High-throughput two-hybrid analysis. The promise and the peril. FEBS J., 272(21):5391–5399.

Freeman, T. C., Goldovsky, L., Brosch, M., van Dongen, S., Mazière, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. J. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Computational Biology, 3(10):2032–42.

Futschik, M. E., Chaurasia, G., and Herzel, H. (2007). Comparison of human protein–protein interaction maps. Bioinformatics, 23(5):605–611.

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. (2003). Genome-Scale Reconstruction of the *Saccharomyces cerevisiae* Metabolic Network. Genome Research, 13(2):244–253.

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. Molecular Biology of the Cell, 11(12):4241–4257.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415:141–147.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084):631–636.

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., and Gavin, A.-C. (2010). Visualization of omics data for systems biology. Nature Methods, 7(3):56–68.

Gene Ontology Consortium (2008). The Gene Ontology project in 2008. Nucleic Acids Research, 36(Database issue):D440–D444.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L. Jr., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shim-

kets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. Science, 302(5651):1727–1736.

Gopalacharyulu, P. V. (2010). Data integration, pathway analysis and mining for systems biology. Espoo 2011. VTT Publications, 732. http://www.vtt.fi/inf/pdf/publications/2010/P732.pdf (30.9.2011).

Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? Genome Biology, 7(11):120.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. Nature, 402(6761):47–52.

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., Le Novère, N., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasić, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kırdar, 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 obtained from a community approach to systems biology. Nature Biotechnology, 26:1155–1160.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature, 415:180–183.

Hooper, S. D., and Bork, P. (2005). Medusa: a simple tool for interaction graph analysis. Bioinformatics, 21(24):4432–4433.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). COPASI – a COmplex PAthway SImulator. Bioinformatics 22(24):3067–3074.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: Systems biology. Annual Review of Genomics and Human Genetics, 2:343–272.

Ideker, T., and Sharan, R. (2008). Protein networks in disease. Genome Research, 18(4):644–652.

Iragne, F., Nikolski, M., Mathieu, B., Auber, D., and Sherman, D. (2005). ProViz: protein interaction visualization and exploration. Bioinformatics, 21(2):272–274.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2000). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of the United States of America, 98(8):4569–4574.

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

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

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

Khanin, R., and Wit, E. (2006). How scale-free are biological networks. Journal of Computational Biology, 13(3):810–818.

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

Kitano, H. (2002b). Computational systems biology. Nature, 420:206–210.

Klipp, E. (2007). Modelling dynamic processes in yeast. 24(11):943–959.

Koikkalainen, J. R., Antila, M., Lötjönen, J. M., Heliö, T., Lauerma, K., Kivistö, S. M., Sipola, P., Kaartinen, M. A., Kärkkäinen, S. T., Reissell, E., Kuusisto, J., Laakso, M., Orešič, M., Nieminen, M. S., and Peuhkurinen, K. J. (2008). Early familial dilated cardiomyopathy: identification with determination of disease state parameter from cine MR image data. Radiology, 249(1):88–96.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature, 440(7084):637–643.

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(R): an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucleic Acids Research, 34(Database issue):D546–D551.

Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Rüegg, A., Rawlings, C., Verrier, P., and Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics, 22(11):1383–1390.

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., Blöcker, 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., and Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822):860–921.

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

Lee, J. M., Gianchandani, E. P., Eddy, J. A., and Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS Computational Biology, 4(5):e1000086.

Lenzen, S., Drinkgern, J., and Tiedge, M. (1996). Low antioxidant enzyme gene expression in pancreatic islets compared with various other mouse tissues. Free Radical Biology and Medicine, 20(3):463–466.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans. Science, 303(5657): 540–543.

Li, X., Gianoulis, T. A., Yip, K. Y., Gerstein, M., and Snyder, M. (2010). Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. Cell, 143(4):639–650.

Lima-Mendez, G., and Helden, J. (2009). The powerful law of the power law and other myths in network biology. Molecular BioSystems, 5(12):1482–1493.

Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2):129–137.

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. Nature Biotechnology, 14(13):1675–1680.

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, 431:308–312.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007). The Edinburgh human metabolic network reconstruction and its functional analysis. Molecular Systems Biology, 3:135.

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., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucleic Acids Research, 31(1):374–378.

Mendes, P. (1993) GEPASI: A software package for modelling the dynamics, steady states and control of biochemical and other systems. Computer Applications in the Biosciences, 9:563–571.

Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. Trends in Biochemical Sciences, 22:361–363.

Mi, Q. S., Ly, D., Zucker, P., McGarry, M., and Delovitch, T. L. (2004). Interleukin-4 but not interleukin-10 protects against spontaneous and recurrent type 1 diabetes by activated CD1d-restricted invariant natural killer T-cells. Diabetes, 53(5):1303–1310.

Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. Science, 298:824–827.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S. S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of designed and evolved networks. Science, 303:1538–1542.

Mizuno, M., Masumura, M., Tomi, C., Chiba, A., Oki, S., Yamamura, T., and Miyake, S. (2004). Synthetic glycolipid OCH prevents insulitis and diabetes in NOD mice. Journal of Autoimmunity, 23(4):293–300.

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? Genome Research, 11(12):1971–1973.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46:323–351.

O'Donoghue, S. I., Gavin, A.-C., Gehlenborg, N., Goodsell, D. S., Hériché, J.-K., Nielsen, C. B., North, C., Olson, A. J., Procter, J. B., Shattuck, D. W., Walter, T., and Wong, B. (2010). Visualizing biological data – now and in the future. Nature Methods, 7(3):2–4.

Orešič, M., Simell, S., Sysi-Aho, M., Näntö-Salonen, K., Seppänen-Laakso, T., Parikka, V., Katajamaa, M., Hekkala, A., Mattila, I., Keskinen, P., Yetukuri, L., Reinikainen, A., Lähde, J., Suortti, T., Hakalax, J., Simell, T., Hyöty, H., Veijola, R., Ilonen, J., Lahesmaa, R., Knip, M., and Simell, O. (2008). Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. Journal of Experimental Medicine, 205(13):2975–2984.

Orlev, N., Shamir, R., and Shiloh, Y. (2004). PIVOT: Protein Interacions VisualizatiOn Tool. Bioinformatics, 20(3):424–425.

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

Pavlopoulos, G. A., Wegener, A. L., and Schneider, R. (2008). A survey of visualization tools for biological network analysis. BioData Mining, 1:12.

Price, D. J. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27:292–306.

Pržulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: Scale-free or geometric? Bioinformatics, 20(18):3508–3515.

Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. Bioinformatics, 23(2):e177–e183.

Qi, Y., and Ge, H. (2006). Modularity and dynamics of cellular networks. PLoS Computational Biology, 2(12):e174.

Quek, L. E., and Nielsen, L. K. (2008). On the reconstruction of the Mus musculus genome-scale metabolic network model. Genome Informatics, 21:89–100.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. Science, 297:1551–1555.

Ravasz, E., and Barabási, A.-L. (2003). Hierarchical organization in complex networks. Physical Review, E 67.

Reguly, T., Breitkreutz, A., Boucher, A., Bobby-Joe 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*. Journal of Biology, 5(4):11.

Sammon, J. W. J. (1969). A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers, C-18(5):401–409.

Saraiya, P., North, C., and Duca, K. (2005). Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. Information Visualization, 4(3):191–205.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature, 461(7261):218–223.

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.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays – a technology review, Nature Cell Biology, 3:E190–E195.

Scott, J., Ideker, T., Karp, R. M., and Sharan, R. (2006). Efficient algorithms for detecting signaling pathways in protein interaction networks. Journal of Computational Biology, 13(2):133–144.

Sheikh, K., Förster, J., and Nielsen, L. K. (2005). Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. Biotechnology Progress, 21(1):112–121.

Shen-Orr, S. S., Milo, R., Mangan, M., and Alon, U. (2002), Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics, 31:64–68.

Simon, H. A. (1955). On a class of skew distribution functions. Biometrika, 42: 425–440.

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. Proceedings of the National Academy of Sciences of the United States of America, 102(43):15545–15550.

Sysi-Aho, M., Koikkalainen, J., Seppänen-Laakso, T., Kaartinen, M., Kuusisto, J., Peuhkurinen, K., Kärkkäinen, S., Antila, M., Lauerma, K., Reissell, E., Jurkko, R., Löttjönen, J., Heliö, T., and Orešič, M. (2011). Serum lipidomics meets cardiac magnetic resonance imaging: profiling of subjects at risk of dilated cardiomyopathy. PLoS ONE, 6(1):e15744.

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(7):1361–1375.

Timonen, M., and Pesonen, A. (2008). Combining context and existing knowledge when recognizing biological entities – Early results. Advances in Knowledge Discovery and Data Mining, 5012:1028–1034.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C., and Hutchison, C. A. (1999). E-CELL: software environment for whole-cell simulation. Bioinformatics, 15(1):72–84.

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, 403:623–627.

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research, 38(Database issue):D142–D148.

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., Gabor Miklos, G. L., 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., Di Francesco, V., 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., Guigó, 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.

Vukkadapu, S. S., Belli, J. M., Ishii, K., Jegga, A. G., Hutton, J. J., Aronow, B. J., and Katz, J. D. (2005). Dynamic interaction between T cell-mediated beta-cell damage and beta-cell repair in the run up to autoimmune diabetes of the NOD mouse. Physiological Genomics, 21(2):201–211.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Molecular Biology and Evolution, 18(7):1283–1292.

Walter, T., Shattuck, D. W., Baldock, R., Bastin, M. E., Carpenter, A. E., Duce, S., Ellenberg, J., Fraser, A., Hamilton, N., Pieper, S., Ragan, M. A., Schneider, J. E., Tomancak, P., and Hériché, J. K. (2010). Visualization of image data from cells to organisms. Nature Methods, 7(3):26–41.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., and Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Research, 37(Web Server issue):W623–W633.

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. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

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 Research, 30(1):303–305.

Yang, L. J. (2008). Big Mac Attack: Does It Play a Direct Role for Monocytes/Macrophages in Type 1 Diabetes? Diabetes, 57(11):2922–2923.

Yetukuri, L., Katajamaa, M., Medina-Gomez, G., Seppänen-Laakso, T., Vidal-Puig, A., and Orešič, M. (2007). Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. BMC Systems Biology 1:12.

Yetukuri, L. R. (2010). Bioinformatics approaches for the analysis of lipidomics data. Espoo 2011. VTT Publications, 741. http://www.vtt.fi/inf/pdf/publications/2010/P741.pdf (30.9.2011).

Yook, S. H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. Proteomics, 4(4):928:942.

Yule, G. U. (1925). A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F. R. S., Philosophical Transactions of the Royal Society of London, Ser. B, 213:21–87.

Zoeller, R. A., Lake, A. C., Nagan, N., Gaposchkin, D. P., Legner, M. A., and Lieberthal, W. (1999). Plasmalogens as endogenous antioxidants: somatic cell mutants reveal the importance of the vinyl ether. Biochemical Journal, 338(Pt3):769–776.

Zoeller, R. A., Grazia, T. J., LaCamera, P., Park, J., Gaposchkin, D. P., and Farber, H. W. (2002). Increasing plasmalogen levels protects human endothelial cells during hypoxia. American Journal of Physiology Heart and Circulatory Physiology, 283(2):H671–H679.

Appendices V–VI of this publication are not included in the PDF version. Please order the printed version to get the complete publication (http://www.vtt.fi/publications/index.jsp).

PUBLICATION I

Detection of molecular paths associated with insulitis and type 1 diabetes in non-obese diabetic mouse

In: PLoS ONE 2009, 4(10), e7323. 9 p.

OPEN ACCESS Freely available online

PLoS one

Detection of Molecular Paths Associated with Insulitis and Type 1 Diabetes in Non-Obese Diabetic Mouse

Erno Lindfors[1], Peddinti V. Gopalacharyulu[1], Eran Halperin[2], Matej Orešič[1]*

1 VTT Technical Research Centre of Finland, Espoo, Finland, **2** International Computer Science Institute, Berkeley, California, United States of America

Abstract

Recent clinical evidence suggests important role of lipid and amino acid metabolism in early pre-autoimmune stages of type 1 diabetes pathogenesis. We study the molecular paths associated with the incidence of insulitis and type 1 diabetes in the Non-Obese Diabetic (NOD) mouse model using available gene expression data from the pancreatic tissue from young pre-diabetic mice. We apply a graph-theoretic approach by using a modified color coding algorithm to detect optimal molecular paths associated with specific phenotypes in an integrated biological network encompassing heterogeneous interaction data types. In agreement with our recent clinical findings, we identified a path downregulated in early insulitis involving dihydroxyacetone phosphate acyltransferase (DHAPAT), a key regulator of ether phospholipid synthesis. The pathway involving serine/threonine-protein phosphatase (PP2A), an upstream regulator of lipid metabolism and insulin secretion, was found upregulated in early insulitis. Our findings provide further evidence for an important role of lipid metabolism in early stages of type 1 diabetes pathogenesis, as well as suggest that such dysregulation of lipids and related increased oxidative stress can be tracked to beta cells.

Citation: Lindfors E, Gopalacharyulu PV, Halperin E, Orešič M (2009) Detection of Molecular Paths Associated with Insulitis and Type 1 Diabetes in Non-Obese Diabetic Mouse. PLoS ONE 4(10): e7323. doi:10.1371/journal.pone.0007323

Editor: Wasif N. Khan, University of Miami, United States of America

Received January 15, 2009; **Accepted** September 13, 2009; **Published** October 2, 2009

Funding: The study was supported by the National Graduate School in Informational and Structural Biology (ISB), TRANSCENDO project of the Tekes MASI Program, DIAPREPP project of the Seventh Framework Program of the European Community (HEALTH-F2-2008-202013), and the research program "White Biotechnology - Green Chemistry" (Academy of Finland; Finnish Centre of Excellence programme, 2008–2013, Decision number 118573). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: matej.oresic@vtt.fi

Introduction

Type 1 diabetes (T1D) is an autoimmune disease that results in destruction of insulin-producing beta cells of the pancreas [1]. The early stages of T1D pathogenesis are characterized by insulitis, an inflammation of the islets of Langerhans of the pancreas caused by the lymphocyte infiltration. Although the seroconversion to islet autoantibody positivity has been the first detectable signal for the onset of autoimmunity and progression towards diabetes [2], the initiators of autoimmune response, mechanisms regulating progress toward beta cell failure and factors determining time of presentation of clinical diabetes are poorly understood.

We recently investigated changes in the serum metabolome prospectively in a unique cohort of children at genetic risk for T1D. Intriguingly, we detected multiple changes related to dysregulation of lipid and amino acid metabolism preceding the autoimmunity and overt T1D [3]. In order to better understand the early diabetes pathogenesis, it would have been therefore of great importance to study the molecular mechanisms behind the early metabolic dysregulation as related to the autoimmune response, an area so far neglected in T1D research.

Motivated by our clinical findings, here we study molecular paths associated with the incidence of type 1 diabetes (T1D) and insulitis in the Non-Obese Diabetic (NOD) mouse model using the available gene expression data from young pre-diabetic mice [4]. The NOD mouse is a strain whose immune system shares many similarities with human's immune system as well as the autoimmune response [5]. It is therefore widely used in studies aiming to elucidate T1D, although it is also clear that this experimental model may only in part reflect the the immune system and T1D pathogenesis in human [6]. We introduce a method EMPath (*E*nriched *M*olecular *Path* detection) for detection of molecular paths of physical interactions in an integrated network of protein-protein interactions, signal transduction maps and metabolic pathways by applying a modified version of the color coding algorithm [7]. The color coding algorithm was applied previously to detect signaling pathways derived from protein interaction networks [8]. In our approach the phenotype context is achieved by the introduction of path weights based on the network structure combined with the mRNA expression data. Our aim is to detect paths in an integrated network such that up- or down-regulated protein nodes, as estimated by the gene expression data, are significantly over-represented on the path in comparison with the rest of the network (Figure 1).

Results and Discussion

Detection of molecular paths associated with insulitis and type 1 diabetes incidence

We applied the EMPath method to an integrated network of protein-protein interactions, signal transduction maps and metabolic pathways where the nodes are proteins or metabolites and the edges are interactions or reactions. In order to study the network in the biological context, we used gene expression information to weight the corresponding protein nodes.

Since our primary aim as related to T1D was to study tissue-specific changes of molecular paths during the early disease

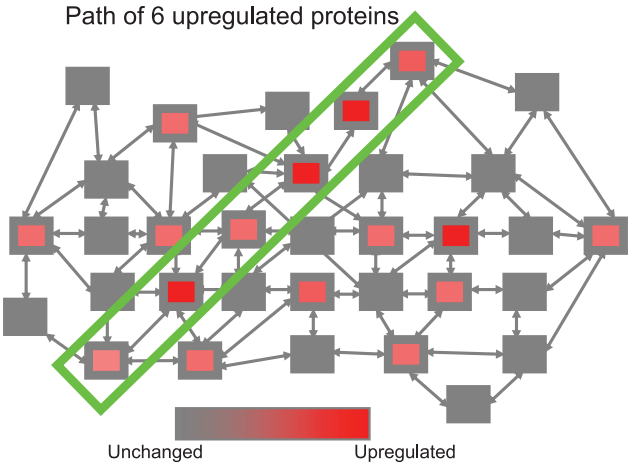


Figure 1. Enriched molecular path detection concept. Illustrative example of path detection in a complex network of interacting entities. An enriched path of 6 entities is highlighted.
doi:10.1371/journal.pone.0007323.g001

pathogenesis, the appropriate study design should include young pre-diabetic mice with selected controls. We searched the T1DBase [9] which hosts T1D related genetic and expression data and identified the study by Vukkadapu *et al.* [4] as the only suitable for our analysis. In addition to that study, there were two other studies available in T1DBase; Chaparro *et al.* [10] and Stanford RoadMap of NOD Type 1 Diabetes (http://fathmanlab.stanford.edu/roadmap_study_design.html). However, we found that Vukkadapu *et al* is more suitable for our analysis than these studies. Chaparro *et al.* contains data from 6-, 9- and 15 week old mice, whereas Vukkadapu *et al.* investigated 3 week old mice. The young mice are more informative for the goals of our study since insulitis is known to occur until 3 or 4 week of age [5]. Standford RoadMap has not yet been published in any journal as of August 2009. However, once available this data will include young mice and will probably provide relevant data in the context of early disease pathogenesis in NOD mice.

In the study by Vukkadapu *et al.* [4], the pancreatic tissue gene expression data is available for four NOD mouse strains from 3 week old animals: BDC2.5/NOD, NOD, BDC2.5/NOD.scid, and NOD.scid. The data analysis in the primary publication was focusing primarily on known T1D-related genes associated with the autoimmune response and inflammation [4]. The four experimental models studied by Vukkadapu *et al.* have differences in the incidence of insulitis and T1D. The BDC2.5/NOD and NOD mice have accelerated and slow insulitis development, respectively. Therefore, comparison of these mouse models may provide information about the pathways associated with early insulitis although as a limitation one should also keep in mind that this not an ideal comparison since genetic factors associated with *e.g.* age and growth are not controlled for. The BDC2.5/NOD.scid model has extremely high diabetes incidence, which develops already at 3–4 weeks of age, whereas the NOD.scid does not develop diabetes. The pathways associated with differences between these two mouse models may thus provide information about mechanisms specific to late insulitis and T1D.

We performed path detection for the two comparisons: (1) BDC2.5/NOD *vs.* NOD (early insulitis) and (2) BDC2.5/NOD.vs. NOD.scid (late insulitis and early T1D). We detected multiple optimal paths at $p<0.025$ threshold in both case-control combinations (Figures S2–S5). Selected high scoring paths are shown in Figure 2. Two serine/threonine-protein phosphatases, 2A (PP2A) and 5 (PP5) were members of the most upregulated paths in early insulitis (Figure S2). PP2A and PP5 are known to interact [11], and PP2A is associated with the autoimmune response in systemic lupus erythematosus [12]. Interestingly, PP2A is also a regulator of insulin secretion in pancreatic beta cells [13] and its activation is required for repression of PPARα, a key regulator of genes involved in beta cell fatty acid oxidation [14].

Several paths including lipid metabolism enzymes were found downregulated in early insulitis (Figure S3, Table S1). Lipid phosphate phosphohydrolase 3 (LPP3) hydrolizes specific phospholipids in the lipid membrane, leading to production of *e.g.* diacylglycerols and ceramides [15]. Two of the enzymes of carnitine metabolism, carnitine O-palmitoyltransferase I (CPT1) and 4-trimethyl aminobutyraldehyde dehydrogenase (TMA-BADH), were also downregulated in the BDC2.5/NOD mice. Interestingly, the dihydroxyacetone phosphate acyltransferase (DHAPAT; Uniprot ID P98192), a key regulator of ether phospholipid synthesis [16], was found in a downregulated path in close proximity of CPT1 (Figure S3).

Two interacting members of the cytochrome P450 family, CYP1B1 and CYP1A1, were found upregulated and present in multiple paths associated with late insulitis and T1D (Figure S4), while basigin (CD147 antigen, also named extracellular matrix metalloproteinase inducer) was found in several downregulated paths (Figure S5). CD147 is a receptor of cyclophilins and is an important messenger of intercellular communication involved also in recruitment of leukocytes from the periphery into tissues during inflammatory responses [17].

As a potential limitation of our approach, in the path detection method presented here we assign weights to nodes based on mRNA expression data and not on protein concentration or direct interaction data. The protein-level data would be ideal for our approach, but such data is generally not available at the global scale such as in transcriptomics studies. We thus use the protein encoding mRNA expression as an approximation, although it is well known that mRNA and corresponding protein level do not always correlate [18]. Although approximate, we believe that use of mRNA expression when protein-level data is unavailable or too sparse is justified and can still provide useful hints about the molecular paths associated with the investigated phenotypes.

Functional characterization of molecular paths

To better understand the paths detected by EMPath in the context of known pathways, we assessed the functional enrichment of detected paths similarly as previously described [8]. We cross-classified the proteins from a molecular path according to whether or not their encoding genes belong to gene sets obtained from the Molecular Signature Database (MSigDB) [19] and tested if the number of those genes associated with the path is larger than expected by chance using the hypergeometric test. We corrected the *p*-values for multiple comparisons using the False Discovery Rate (FDR) *q*-values. By setting the statistical significance level at FDR $q<0.05$, we identified multiple gene sets over-represented among the detected molecular paths (Table S2). As a summary, the top ten enriched pathways in each of the case-control settings are shown in Table 1.

It is evident from Table 1 that early insulitis (*i.e.* BDC2.5/NOD strain, as compared to NOD) is associated with altered cell signaling since multiple (de)phosphorilation pathways are affected. In contrast, the lipid metabolism is diminished. The paths associated with late insulitis and T1D in BDC2.5/NOD.scid strain are related to cell communication and related processes,

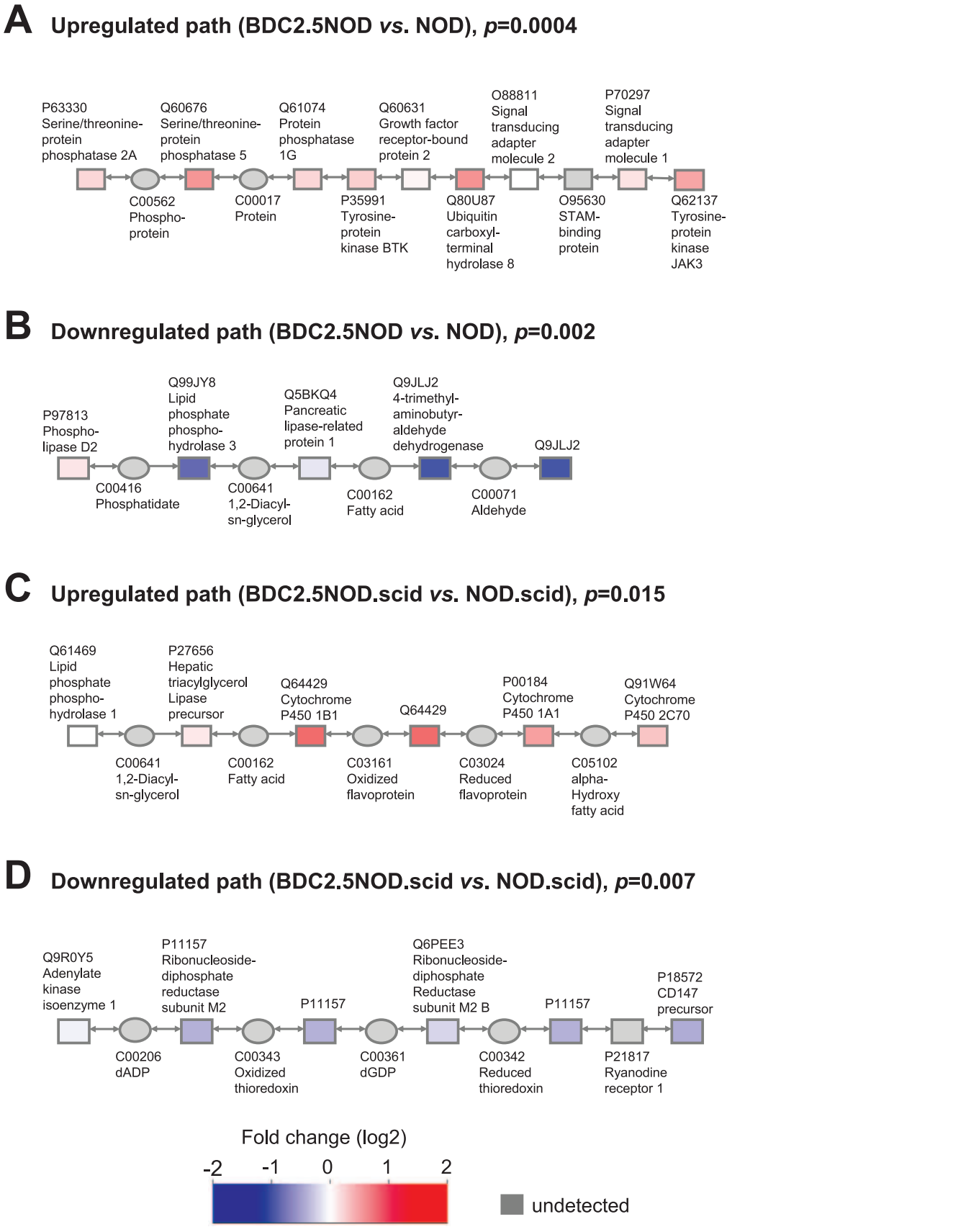


Figure 2. Selected paths significant in different case-control settings. Upregulated (A) and downregulated (B) paths related to insulitis. Upregulated (C) and downregulated (D) paths related to late insulitis and T1D.
doi:10.1371/journal.pone.0007323.g002

Table 1. Top enriched pathways in insulitis and type 1 diabetes as derived from detected paths.

Gene set	Source	n(P & G)	n(G)	Nominal *p*-value	FDR *q*-value
Enriched in upregulated paths (BDC2.5/NOD *vs.* NOD)					
PTDINSPATHWAY	BioCarta	3	19	0.000004	0.000103
HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM	KEGG	3	27	0.000012	0.000155
HSA00530_AMINOSUGARS_METABOLISM	KEGG	2	16	0.000025	0.000280
GALACTOSE_METABOLISM	GenMAPP	2	20	0.000596	0.003099
HSA00052_GALACTOSE_METABOLISM	KEGG	2	24	0.000863	0.003738
GLUCONEOGENESIS	GenMAPP	2	39	0.002286	0.006604
GLYCOLYSIS	GenMAPP	2	39	0.002286	0.006604
HSA04630_JAK_STAT_SIGNALING_PATHWAY	KEGG	4	100	0.000118	0.008023
HSA04664_FC_EPSILON_RI_SIGNALING_PATHWAY	KEGG	3	62	0.000150	0.008426
GHPATHWAY	BioCarta	2	24	0.000863	0.016104
Enriched in downregulated paths (BDC2.5/NOD *vs.* NOD)					
GLYCEROLIPID_METABOLISM	GenMAPP	3	24	$<10^{-6}$	0.000011
STATIN_PATHWAY_PHARMGKB	GenMAPP	2	16	0.000152	0.000557
HSA00565_ETHER_LIPID_METABOLISM	KEGG	2	21	0.000441	0.002093
HSA00071_FATTY_ACID_METABOLISM	KEGG	2	29	0.000847	0.002311
HSA00120_BILE_ACID_BIOSYNTHESIS	KEGG	2	20	0.000399	0.002311
HSA00220_UREA_CYCLE_AND_METABOLISM_OF_AMINO_GROUPS	KEGG	2	21	0.000441	0.002311
HSA00310_LYSINE_DEGRADATION	KEGG	2	29	0.000847	0.002311
HSA00340_HISTIDINE_METABOLISM	KEGG	2	19	0.000359	0.002311
HSA00410_BETA_ALANINE_METABOLISM	KEGG	2	17	0.000286	0.002311
HSA00620_PYRUVATE_METABOLISM	KEGG	2	28	0.000789	0.002311
Enriched in upregulated paths (BDC2.5/NOD.scid *vs.* NOD.scid)					
EGFPATHWAY	BioCarta	4	25	$<10^{-6}$	0.000040
HSA04630_JAK_STAT_SIGNALING_PATHWAY	KEGG	5	100	0.000002	0.000102
HSA05213_ENDOMETRIAL_CANCER	KEGG	4	42	0.000004	0.000128
HSA05223_NON_SMALL_CELL_LUNG_CANCER	KEGG	4	43	0.000004	0.000128
CTLA4PATHWAY	BioCarta	3	15	0.000008	0.000131
ERK5PATHWAY	BioCarta	3	16	0.000010	0.000131
HSA05214_GLIOMA	KEGG	4	50	0.000007	0.000131
PTENPATHWAY	BioCarta	3	16	0.000010	0.000131
NGFPATHWAY	BioCarta	3	17	0.000012	0.000140
IGF1PATHWAY	BioCarta	3	18	0.000014	0.000149
Enriched in downregulated paths (BDC2.5/NOD.scid *vs.* NOD.scid)					
PYRIMIDINE_METABOLISM	GenMAPP	3	43	0.000010	0.000061
HSA00230_PURINE_METABOLISM	KEGG	3	90	0.000096	0.000334
NDKDYNAMINPATHWAY	BioCarta	2	16	0.000898	0.006367
HSA05110_CHOLERA_INFECTION	KEGG	1	31	0.039815	0.046451

Top ten enriched gene sets at FDR $q<0.05$ defined in the Molecular Signature Database [19], using the gene lists derived from the detected paths (Figures S2–S5). The *p*-value is obtained from the hypergeometric test. Column legend: n(P&G), number of common genes in the detected path and the gene set; n(G), number of genes in the gene set.

doi:10.1371/journal.pone.0007323.t001

while the nucleotide and nucleoside metabolism, *i.e.*, likely related to cell cycle and DNA repair, is impaired.

Comparison of path detection with pathway analysis

We performed Gene Set Enrichment Analysis (GSEA) [19] for both case-control comparisons. Table 2 contains the top scored pathways for each strain at FDR $q<0.05$, while a full list of affected pathways at recommended $q<0.25$ is shown in Tables S3–S6. In agreement with earlier analyses [4], both EMPath (Table 1) and GSEA analyses confirmed multiple inflammatory and T cell activation pathways in pancreatic tissue in late insulitis and early T1D. The cell proliferation, division, as well as nucleotide synthesis pathways were found diminished, confirming increasing cell death and DNA damage at this late stage of disease pathogenesis.

In accordance with path detection results, lipid metabolism related pathways (fatty acid metabolism and bile acid synthesis) are downregulated in insulitis, while the cell cycle related pathways are downregulated in T1D (Table 2). The CPT1 and TMABADH

Table 2. Top scored pathways in GSEA.

Gene set	Size	Enrichment Score	Nominal p-value	FDR q-value	Source
Downregulated paths (BDC2.5/NOD vs. NOD)					
HSA03010_RIBOSOME	44	−0.61	0.000466	0.0027	KEGG
WNTPATHWAY	22	−0.63	0.002375	0.0252	BioCarta
HSA00071_FATTY_ACID_METABOLISM	29	−0.58	0.001845	0.0291	KEGG
CALCINEURINPATHWAY	17	−0.64	0.007370	0.0392	BioCarta
PROTEASOMEPATHWAY	21	−0.61	0.004710	0.0418	BioCarta
BILE_ACID_BIOSYNTHESIS	15	−0.65	0.007466	0.0425	GenMAPP
Upregulated paths (BDC2.5/NOD.scid vs. NOD.scid)					
HSA04610_COMPLEMENT_AND_COAGULATION_CASCADES	52	0.62	<10–5	0.0022	KEGG
HSA04612_ANTIGEN_PROCESSING_AND_PRESENTATION	33	0.66	<10–5	0.0038	KEGG
HSA04620_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	74	0.54	<10–5	0.0107	KEGG
HSA04060_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	169	0.47	<10–5	0.0183	KEGG
NKCELLSPATHWAY	15	0.71	0.002838	0.0310	BioCarta
HSA04940_TYPE_I_DIABETES_MELLITUS	20	0.66	0.002753	0.0353	KEGG
Downregulated paths (BDC2.5/NOD.scid vs. NOD.scid)					
CELL_CYCLE_KEGG	58	−0.57	<10–5	0.0031	GenMAPP
CELL_CYCLE	53	−0.56	<10–5	0.0073	GO
UBIQUITIN_MEDIATED_PROTEOLYSIS	20	−0.67	0.000937	0.0096	GenMAPP
G1_TO_S_CELL_CYCLE_REACTOME	54	−0.53	<10–5	0.0112	GenMAPP
HSA00190_OXIDATIVE_PHOSPHORYLATION	86	−0.49	<10–5	0.0113	KEGG
P53PATHWAY	16	−0.70	0.001388	0.0144	BioCarta
PROTEASOMEPATHWAY	21	−0.64	<10–5	0.0174	BioCarta
HSA04120_UBIQUITIN_MEDIATED_PROTEOLYSIS	25	−0.62	0.000473	0.0177	KEGG
HSA04110_CELL_CYCLE	82	−0.47	0.000553	0.0211	KEGG
CARM_ERPATHWAY	19	−0.63	0.004144	0.0279	BioCarta
MRNA_PROCESSING_REACTOME	83	−0.46	<10–5	0.0312	GenMAPP
HSA00510_N_GLYCAN_BIOSYNTHESIS	24	−0.59	0.003738	0.0356	KEGG
G2PATHWAY	18	−0.62	0.004585	0.0475	BioCarta

This table contains top scored gene sets in GSEA for each strain (FDR $q<0.05$). The gene sets studies are the same as in the analysis for Table 1. None of the pathways were significantly upregulated in the BDC2.5/NOD *vs*. NOD comparison using the FDR $q<0.05$ threshold.
doi:10.1371/journal.pone.0007323.t002

found in downregulated paths associated with early insulitis (Table S1) were both among the leading edge genes in the fatty acid metabolism gene set, while TMABADH was also the leading edge in the bile acid synthesis module.

Meta analysis of findings using T1DBase

To investigate how genes detected by EmPath change in gene expression analyses seen in several other studies, we used the Meta Analysis tool of the T1DBase (http://www.t1dbase.org/page/MetaHome) [9]. As a result, we selected the genes found in the significant molecular paths (Figure 2) and visualized their differential expression across multiple studies available in T1Dbase (Figures S6–S9).

We can see some interesting observations regarding the genes that were involved in our detected paths. DHAPAT (often abbreviated as GNPAT), a gene that was found in paths downregulated in early insulitis in paths detected by EmPath, was also down-regulated in mice deficient for transcriptional regulators FoxA2 and Sox4 [20,21]. PP2 (also abbreviated as PPP2CA), a gene that was upregulated in early insulitis and type 1 diabetes in paths detected by EmPath, was also upregulated in FoxA2 deficient mouse [20]. Another interesting observation is that the up-/down-regulation of molecular paths in early insulitis in our study matches particularly well with the data from the FoxA2 deficient mouse (Figures S6–S7 and reference [20]). FoxA2 is a transcription factor involved in the regulation of insulin sensitivity [22].

Ether lipids and oxidative stress in beta cells

As a most surprising finding from our study, multiple lipid pathways were downregulated in early insulitis (BDC2.5/NOD *vs.* NOD comparison), including the ether lipid metabolism (Table 1). Ether phospholipid synthesis, including synthesis of plasmalogens, starts in peroxisomes and involves esterification of dihydroxyacetone phosphate (DHAP) with a long-chain acyl-CoA ester [16,23] (Figure 3). This first reaction is catalyzed by dihydroxyacetone phosphate acyltransferase (DHAPAT, EC 2.3.1.42). This reaction appears to be affected in early insulitis, since the path involving DHAPAT is diminished (Tables 1 and S1, Figure S3). The plasmalogens are the most abundant ether phospholipids and may protect cellular functions from oxidative damage [24,25]. The ether lipids were also found consistently diminished in serum of children who later progressed to type 1 diabetes [3]. Diminished protection against the reactive oxygen species is relevant for T1D since pancreatic beta cells are particularly susceptible to oxidative damage [26,27]. Further supporting the role of lipids in early

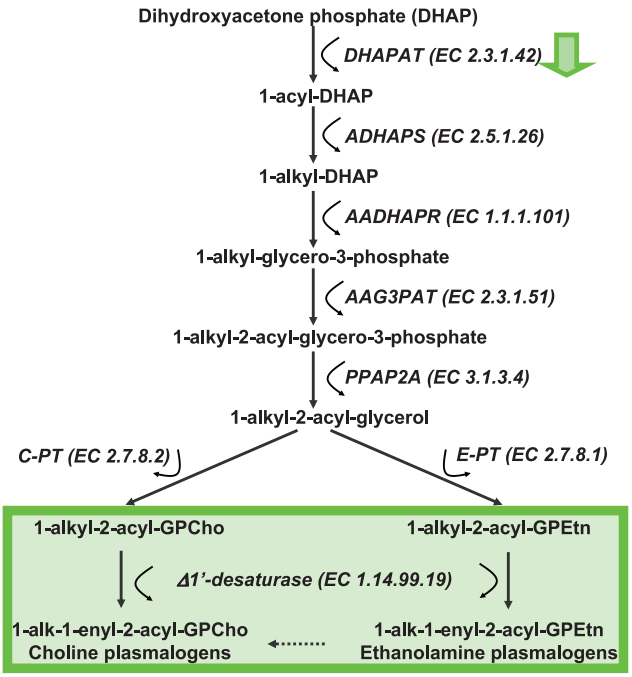


Figure 3. Schematic representation of the steps involved in the biosynthesis of ether phospholipids, including plasmalogens. The lipids found consistently downregulated in serum of children who later developed type 1 diabetes [3] are shown in green box. DHAPAT enzyme is found in the downregulated paths in early insulitis in the present study (green arrow). The first three reactions in the pathway take place in peroxisomes, while the others are catalyzed by microsomal enzyme systems. Other routes for the formation of ether phospholipids may exist [16].
doi:10.1371/journal.pone.0007323.g003

insulitis, the enzymes of carnitine metabolism and fatty acid transport to mitochondria (CPT1 and TMABADH) were found in downregulated paths as well.

Previous genetic studies have shown that defective plasmalogen synthesis associates with impaired membrane trafficking [28] although the implications for type 1 diabetes remain to be established [29]. Plasmalogen synthesis-related genes such as DHAPAT clearly need to be evaluated as potential type 1 diabetes susceptibility genes. The complete depletion of ether lipids via a genetic DHAPAT knock-out model leads to a severe phenotype, including arrest of spermatogenesis, development of cataract and defects in central nervous system myelination [30]. In order to study the physiological consequences of altered ether lipid levels as observed in pre-diabetes, one would therefore need to establish experimental models with partial depletion of ether lipids.

Conclusions

We demonstrated that graph-theoretic approaches such as EMPath are a useful tool for detecting pathways of physical interactions associated with specific disease phenotypes. Our findings from the study of paths associated with early insulitis and T1D are consistent with recent findings from a large scale clinical metabolomics study, suggesting an important role of lipid metabolism in the early stages of T1D pathogenesis. We provide evidence that such dysregulation of lipid metabolism and related oxidative stress may be tracked to beta cells and may thus explain the beta cell loss due to increased oxidative stress. The genes identified as important in early insulitis such as DHAPAT or PP2A clearly need to be investigated further in the context of early T1D pathogenesis as well as for their therapeutic potential.

Materials and Methods

Construction of integrated network

We constructed an integrated interaction network by combining protein-protein interactions, signal transduction maps and metabolic pathways in mouse as described previously [31,32]. The integrated network nodes stand for proteins or metabolites, and edges stand for interactions between nodes. We retrieved protein-protein interactions from BIND [33], MINT [34] and DIP [35], signal transduction interactions from TransPath [36] and biochemical reactions from KEGG [37]. We excluded highly connected cofactors from the network since they do not participate in the actual metabolic conversions as substrates or products. Therefore, their inclusion would connect many metabolically distant enzymes. The excluded cofactors are listed in the Supplementary Table S7.

Gene expression data

We obtained normalized gene expression data from the T1D dataset [4] from NCBI Gene Expression Omnibus (GEO) database [38] series accession number: GSE1623. We used the samples GSM27446 (BDC2.5/NOD1), GSM27451 (BDC2.5/NOD.scid_1), GSM27453 (NOD.scid1) and GSM27456 (NOD1) in all the analyses presented in this paper. In the source mouse model experiments [4], RNA hybridization was done on Affymetrix gene chip platform MGU74AV2.

Edge and node weights

The color coding algorithm used in [8] was not suitable for detecting paths in phenotypic context, since they did not have any phenotypic weights. Their weights were solely based on reliabilities of interactions. We modified the color coding algorithm so that it works in phenotypic manner by assigning weights to nodes. We did the weight assignment for each mouse model comparison separately. In order to find the up-regulated paths, we assigned case-control ratios. And to find down-regulated paths, we assigned control-case ratios as weights to nodes. We can thus use the color coding algorithm to find maximum paths in both cases.

We assigned equal weights of 1.0 to all edges from MINT, DIP, KEGG and TransPath, while the edges from BIND were set to 0.33, reflecting large database size of BIND and its reliability of interactions [39].

Path scoring

The path score is computed as follows. In order to give high penalty for a cascade of unreliable edges, we first multiply all edge weights. In order to reward inclusion of high weight nodes, we sum up all node weights. In the end, we multiply the edge product and the node sum. More precisely, the path scoring scheme is presented in Figure S10 and Formulas (1)–(3) below. We thus move forward on a path by selecting a node and edge so that the total weight is maximized. However, we are not allowed to move forward to a node if its color is inside the sliding window (read more in the next paragraph).

$$w(edge\mathrm{Prod}) = w(E12) * w(E23) * w(E23) * \ldots * w(E(n-1)n) \quad (1)$$

$$w(nodeSum) = w(N1) + w(N2) + w(N3) + \ldots + w(Nn) \quad (2)$$

$$w(tot) = w(edge\mathrm{Prod}) * w(nodeSum) \quad (3)$$

We used a color coding algorithm for detecting optimal paths [7]. The basic idea of this algorithm is to assign colors (*i.e.*, integers)

to nodes randomly and detect paths which do not contain same color twice. The restriction on colors guarantees that the detected path is a simple path. When the network is very large, the applicability of this algorithm is challenged by the large computer memory requirements. To address this limitation, we extended the algorithm by using a sliding window so that the distinct color requirement applies only to nodes that are inside the window (Figure S1). That is, unlike the original algorithm which allows no two nodes in a path to have the same color, our algorithm allows no two nodes within the length of the sliding window to have the same color. We first tried to detect a path by using a window length that is equal to the length of detected path. If we did not find a path, we decreased the window length by 1 until we found a path or the window length became 1. This modification improves the performance because it avoids storing of the whole path in computer memory. The algorithm is thus faster and it is capable of detecting longer paths. It is thus more applicable to integrated networks that are usually very large. However, in principle the original version could be used in integrated networks, but it is more probable that there appear memory problems.

Statistical significance of a path

In order to test for the null hypothesis that the detected path is obtained by chance, we calculated the *p*-values. In order to calculate one *p*-value, we shuffled node and edge weights 10,000 times. For the purpose of computational efficiency, we first tested how promising the *p*-value looks after each shuffle based on the pre-specified cutoff criterion (*p*-value <0.025), then jumped into the next path if the criterion was not met. The full algorithm for the *p*-value calculation is described in the Supplementary Text S1.

Network harvesting

A network is considered *harvested* if all optimal paths in the network are detected. However, there is not any rigorous way to define when the network is *harvested*, so we took a heuristic approach by assuming that the network is harvested if we come up with 50 consecutive iterations in which the detected path is previously detected. However, since the *p*-value calculation for an optimal path is computationally expensive, we also limited ourselves to finding at most two optimal paths of the same length in each network (*i.e.*, in each mouse model comparison). It is easy to increase this number of paths if required. The algorithm is described in the Supplementary Text S2.

Characterization of paths

We used a hypergeometric test to identify gene sets from the MSigDB [19] that are over-represented in the molecular paths detected by the EMPath method. First, as a quality control criterion, we restricted the searches to gene sets compiled from pathway databases KEGG, BioCarta, GenMAPP, and GO. Next, we defined the *Gene Symbol Universe* by taking the union of all genes in the selected gene sets. Next, we translated the Swissprot accession numbers of protein nodes of the molecular paths to the Gene Symbols of their encoding genes. These translations are done using Affymetrix annotations of the mouse gene chip platform MGU74Av2, the platform used for NOD mice gene expression experiments. Finally, by using the function *phyper* of the R *stats* package [40] we tested for enrichment of each gene set in each molecular path. In order to account for multiple comparisons, the Benjamini and Hochberg's method for controlling the false discovery rate was applied [41].

Gene Set Enrichment Analysis

We performed Gene Set Enrichment Analysis (GSEA) of the T1D gene expression data [4] using Java desktop version of the software (February 2006 release). We performed GSEA separately for the two selected phenotype comparisons. Since there was only one sample per phenotype, giving one gene expression value per gene per phenotype, we used the *ratio of classes* statistic of the GSEA for ranking genes. We accessed the gene sets defined in the MSigDB [19] and annotations for the Affymetrix gene chip platform MGU74AV2 via ftp pages of GSEA from within the software interface. The GSEA statistics were computed using 5,000 gene set permutations.

T1DBase Meta Analysis

First, we selected proteins from the paths detected by EmPath (Figures S2–S5). They were annotated by Uniprot identifiers. We then used EMBL database to find EMBL identifiers for corresponding genes. The NCBI Entrez gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) was then searched to find Entrez gene identifiers for those genes. We used these identifiers on the web user interface of the T1DBase Meta analysis tool (http://www.t1dbase.org/page/MetaHome). We performed the expression comparison by using all studies that were available in the Beta Cell Biology Consortium.

Supporting Information

Text S1 The algorithm for calculating significance of optimal paths detected by EMPath method (p-value calculation).
Found at: doi:10.1371/journal.pone.0007323.s001 (0.04 MB DOC)

Text S2 Network harvesting algorithm.
Found at: doi:10.1371/journal.pone.0007323.s002 (0.03 MB DOC)

Table S1 Genes found in downregulated paths in insulitis.
Found at: doi:10.1371/journal.pone.0007323.s003 (0.04 MB DOC)

Table S2 Significantly enriched pathways in insulitis and type 1 diabetes as derived from detected paths.
Found at: doi:10.1371/journal.pone.0007323.s004 (1.06 MB DOC)

Table S3 Enriched upregulated pathways in insulitis.
Found at: doi:10.1371/journal.pone.0007323.s005 (0.03 MB DOC)

Table S4 Enriched downregulated pathways in insulitis.
Found at: doi:10.1371/journal.pone.0007323.s006 (0.09 MB DOC)

Table S5 Enriched upregulated pathways in type 1 diabetes.
Found at: doi:10.1371/journal.pone.0007323.s007 (0.05 MB DOC)

Table S6 Enriched downregulated pathways in type 1 diabetes.
Found at: doi:10.1371/journal.pone.0007323.s008 (0.08 MB DOC)

Table S7 Excluded cofactors.
Found at: doi:10.1371/journal.pone.0007323.s009 (0.08 MB DOC)

Figure S1 Use of a sliding window to optimize the path detection. The distinct color requirement applies only inside the window. We therefore do not need store the whole path in memory, which makes the detection process faster. In this figure we have an example in which our window size is 2. Our path detection is at a stage in which we have traversed from A- to B to

C. And we have {2,3} in denied colors. We can thus continue to either D or E.
Found at: doi:10.1371/journal.pone.0007323.s010 (1.00 MB DOC)

Figure S2 Upregulated paths in BDC2.5/NOD vs. NOD comparison. The nodes are colored using the same color code as in Figure 2. Edge annotations related to the source database: K, KEGG; M, MINT.
Found at: doi:10.1371/journal.pone.0007323.s011 (1.30 MB DOC)

Figure S3 Downregulated paths in BDC2.5/NOD vs. NOD comparison. The nodes are colored using the same color code as in Figure 2. Edge annotations related to the source database: K, KEGG; M, MINT.
Found at: doi:10.1371/journal.pone.0007323.s012 (1.30 MB DOC)

Figure S4 Upregulated paths in BDC2.5/NOD.scid vs. NOD.scid comparison. The nodes are colored using the same color code as in Figure 2. Edge annotations related to the source database: K, KEGG; M, MINT.
Found at: doi:10.1371/journal.pone.0007323.s013 (1.30 MB EPS)

Figure S5 Downregulated paths in BDC2.5/NOD.scid vs. NOD.scid comparison. The nodes are colored using the same color code as in Figure 2. Edge annotations related to the source database: K, KEGG; M, MINT.
Found at: doi:10.1371/journal.pone.0007323.s014 (1.26 MB EPS)

Figure S6 Meta-analysis for upregulated genes in BDC2.5/NOD vs. NOD comparison. Genes are presented as rows and study group comparisons as columns.
Found at: doi:10.1371/journal.pone.0007323.s015 (1.87 MB EPS)

Figure S7 Meta-analysis for downregulated genes in BDC2.5/NOD vs. NOD comparison. Genes are presented as rows and study group comparisons as columns.
Found at: doi:10.1371/journal.pone.0007323.s016 (1.86 MB EPS)

Figure S8 Meta-analysis for upregulated genes in BDC2.5/NOD.scid vs. NOD.scid comparison. Genes are presented as rows and study group comparisons as columns.
Found at: doi:10.1371/journal.pone.0007323.s017 (2.25 MB EPS)

Figure S9 Meta-analysis for downregulated genes in BDC2.5/NOD.scid vs. NOD.scid comparison. Genes are presented as rows and study comparisons as columns.
Found at: doi:10.1371/journal.pone.0007323.s018 (1.67 MB EPS)

Figure S10 Path scoring method. In order to calculate the score for the path, the edge weights are multiplied. All node weights are then summed up. In the end, the edge product and the node sum are multiplied. The total path score is thus (w(E12)* w(E23)*..* w((n-1)N)))*(W(N1)+ W(N2)+..+ W(Nn)).
Found at: doi:10.1371/journal.pone.0007323.s019 (1.00 MB EPS)

Acknowledgments

We thank Jari Saramäki, Samuel Kaski and Fredrik Bäckhed for helpful discussions and moments, and Richard M. Karp for comments on the use of the color coding algorithm in this study.

Author Contributions

Conceived and designed the experiments: MO. Performed the experiments: EL. Analyzed the data: EL MO. Contributed reagents/materials/analysis tools: PG EH. Wrote the paper: EL MO.

References

1. Notkins AL, Lernmark A (2001) Autoimmune type 1 diabetes: resolved and unresolved issues. J Clin Invest 108: 1247–1252.
2. Achenbach P, Bonifacio E, Koczwara K, Ziegler A-G (2005) Natural history of type 1 diabetes. Diabetes 54: S25–31.
3. Oresic M, Simell S, Sysi-Aho M, Näntö-Salonen K, Seppänen-Laakso T, et al. (2008) Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. J Exp Med: doi: 10.1084/jem.20081800.
4. Vukkadapu SS, Belli JM, Ishii K, Jegga AG, Hutton JJ, et al. (2005) Dynamic interaction between T cell-mediated beta-cell damage and beta-cell repair in the run up to autoimmune diabetes of the NOD mouse. Physiol Genomics 21: 201–211.
5. Anderson MS, Bluestone JA (2005) The NOD mouse: a model of immune dysregulation. Annu Rev Immunol 23: 447–485.
6. Atkinson MA, Leiter EH (1999) The NOD mouse model of type 1 diabetes: As good as it gets? Nature 5: 601–604.
7. Alon N, Yuster R, Zwick U (1995) Color coding. J ACM 42: 844–856.
8. Scott J, Ideker T, Karp RM, Sharan R (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. J Comput Biol 13: 133–144.
9. Hulbert EM, Smink LJ, Adlem EC, Allen JE, Burdick DB, et al. (2007) T1DBase: integration and presentation of complex data for type 1 diabetes research. Nucl Acids Res 35: D742–746.
10. Chaparro RJ, Konigshofer Y, Beilhack GF, Shizuru JA, McDevitt HO, et al. (2006) Nonobese diabetic mice express aspects of both type 1 and type 2 diabetes. Proc Natl Acad Sci U S A 103: 12475–12480.
11. Lubert EJ, Hong Y-l, Sarge KD (2001) Interaction between Protein Phosphatase 5 and the A subunit of Protein Phosphatase 2A. Evidence for a heterotrimeric form of protein phosphatase 5. J Biol Chem 276: 38582–38587.
12. Crispín JC, Kyttaris VC, Juang Y-T, Tsokos GC (2008) How signaling and gene transcription aberrations dictate the systemic lupus erythematosus T cell phenotype. Trends Immunol 29: 110–115.
13. Parameswara VK, Sule AJ, Esser V (2005) Have we overlooked the importance of serine/threonine protein phosphatases in pancreatic beta-cells? Role played by protein phosphatase 2A in insulin secretion. JOP 8: 303–315.
14. Ravnskjaer K, Boergesen M, Dalgaard LT, Mandrup S (2006) Glucose-induced repression of PPAR{alpha} gene expression in pancreatic {beta}-cells involves PP2A activation and AMPK inactivation. J Mol Endocrinol 36: 289–299.
15. Brindley DN, Waggoner DW (1998) Mammalian lipid phosphate phosphohydrolases. J Biol Chem 273: 24281–24284.
16. Nagan N, Zoeller RA (2001) Plasmalogens: biosynthesis and functions. Prog Lipid Res 40: 199–229.
17. Bukrinsky MI (2002) Cyclophilins: unexpected messengers in intercellular communications. Trends Immunol 23: 323–325.
18. Jansen R, Greenbaum D, Gerstein M (2002) Relating Whole-Genome Expression Data with Protein-Protein Interactions. Genome Research 12: 37–46.
19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545–15550.
20. Lantz K, Vatamaniuk M, Brestelli J, Friedman J, Matschinsky F, et al. (2004) Foxa2 regulates multiple pathways of insulin secretion. J Clin Invest 114: 512–520.
21. Wilson M, Yang K, Kalousova A, Lau J, Kosaka Y, et al. (2005) The HMG box transcription factor Sox4 contributes to the development of the endocrine pancreas. Diabetes 54: 3402–3409.
22. Puigserver P, Rodgers JT (2006) Foxa2, a novel transcriptional regulator of insulin sensitivity. 12: 38–39.
23. Lee T-C (1998) Biosynthesis and possible biological functions of plasmalogens. Biochim Biophys Acta 1394: 129–145.
24. Zoeller RA, Lake AC, Nagan N, Gaposchkin DP, Legner MA, et al. (1999) Plasmalogens as endogenous antioxidants: somatic cell mutants reveal the importance of the vinyl ether. Biochem J 338: 769–776.
25. Zoeller RA, Grazia TJ, LaCamera P, Park J, Gaposchkin DP, et al. (2002) Increasing plasmalogen levels protects human endothelial cells during hypoxia. Am J Physiol Heart Circ Physiol 283: H671–679.
26. Cnop M, Welsh N, Jonas J-C, Jorns A, Lenzen S, et al. (2005) Mechanisms of pancreatic {beta}-cell death in Type 1 and Type 2 Diabetes: Many differences, few similarities. Diabetes 54: S97–107.
27. Lenzen S, Drinkgern J, Tiedge M (1996) Low antioxidant enzyme gene expression in pancreatic islets compared with various other mouse tissues. Free Radic Biol Med 20: 463–466.
28. Thai T-P, Rodemer C, Jauch A, Hunziker A, Moser A, et al. (2001) Impaired membrane traffic in defective ether lipid biosynthesis. Hum Mol Genet 10: 127–136.

29. Ewens KG, Johnson LN, Wapelhorst B, O'Brien K, Gutin S, et al. (2002) Linkage and association with type 1 diabetes on chromosome 1q42. Diabetes 51: 3318–3325.
30. Gorgas K, Teigler A, Komljenovic D, Just WW (2006) The ether lipid-deficient mouse: Tracking down plasmalogen functions. Biochim Biophys Acta 1763: 1511–1526.
31. Gopalacharyulu PV, Lindfors E, Bounsaythip C, Kivioja T, Yetukuri L, et al. (2005) Data integration and visualization system for enabling conceptual biology. Bioinformatics 21: i177–i185.
32. Gopalacharyulu PV, Lindfors E, Miettinen J, Bounsaythip CK, Oresic M (2008) An integrative approach for biological data mining and visualisation. Int J Data Min Bioinform 2: 54–77.
33. Bader GD, Betel D, Hogue CWV (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31: 248–250.
34. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. Nucl Acids Res 35: D572–574.
35. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucl Acids Res 32: D449–451.
36. Krull M, Pistor S, Voss N, Kel A, Reuter I, et al. (2006) TRANSPATH(R): an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucl Acids Res 34: D546–551.
37. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucl Acids Res 36: D480–484.
38. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2008) NCBI GEO: archive for high-throughput functional genomic data. Nucl Acids Res: doi:10.1093/nar/gkn1764.
39. Futschik ME, Chaurasia G, Herzel H (2007) Comparison of human protein protein interaction maps. Bioinformatics 23: 605–611.
40. R Development Core Team (2008) R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing.
41. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 57: 289–300.

PUBLICATION II

Data integration and visualization system for enabling conceptual biology

In: Bioinformatics 2005, 21(1):i177–i185.
Reprinted with permission from the publisher.

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 artificia (Blagosklonny and Pardee, 2002). In order to utilize the diverse life science knowledge, one firs needs to address several practical and fundamental challenges of data integration. For example, different domain-specifi 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-specifi 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 definition 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.).

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,

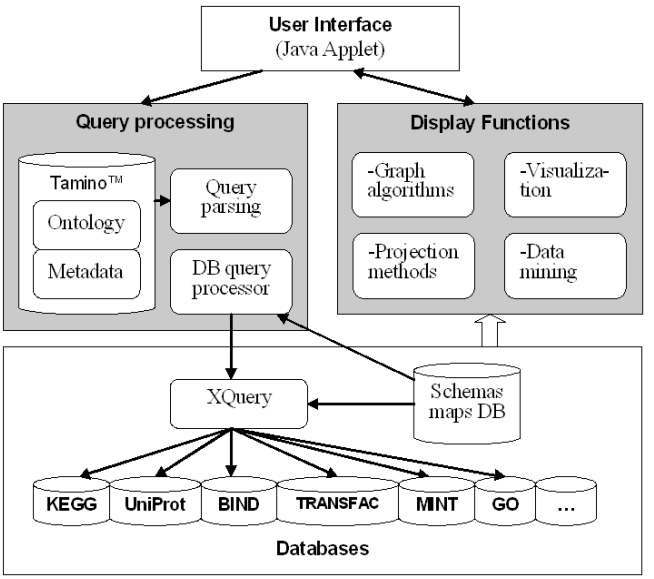


Fig. 1. Architecture of our bioinformatics data integration and visualization system.

- 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 specifi bioinformatics database into our system, it has to be passed firs through a curation stage. A typical data curation fl w 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).

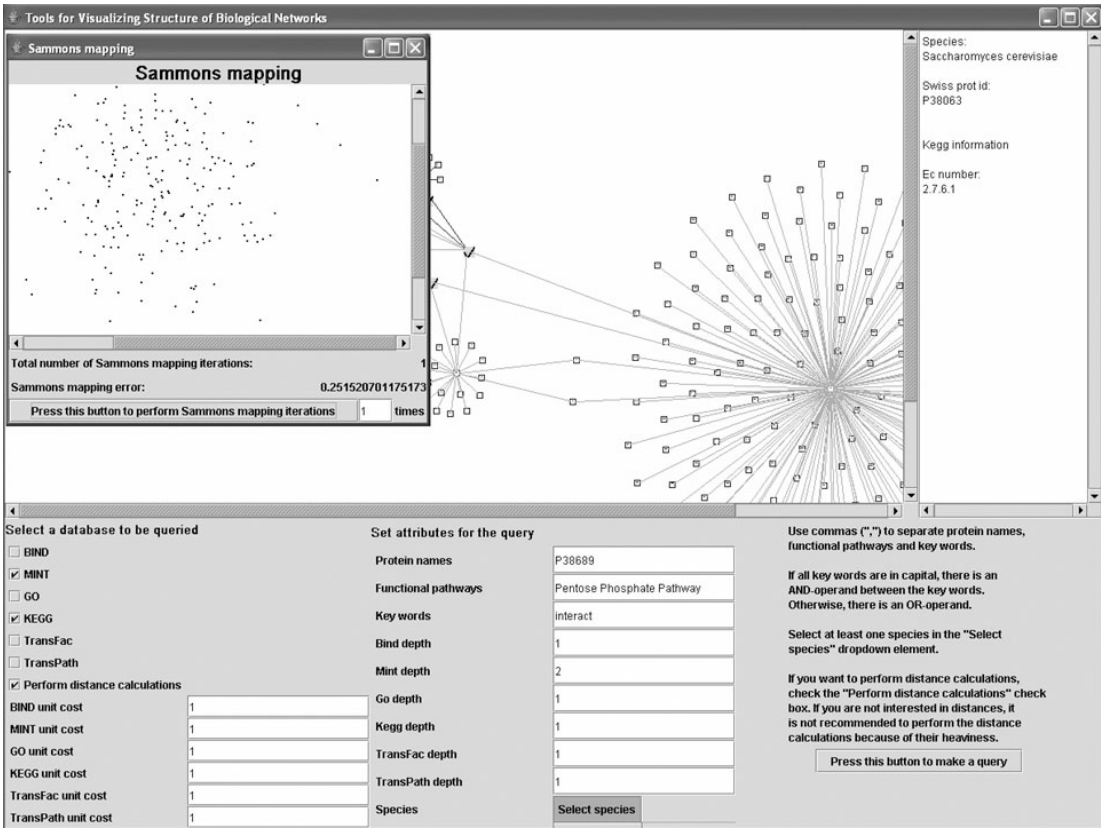


Fig. 2. Screenshot of the megNet network visualization tool. Node shapes represent their types (e.g. protein, gene), and edge colors represent types of relationships. The Sammon's mapping window displays the mapping based on specifie distance metrics.

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

(3) If the document structures have been define 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 fi the requirements of the implementation (e.g. change the target namespace to Tamino namespace and defin a prefi for the original target namespace).

(4) Defin physical properties, such as indices and doctype for the logical schema to construct a Tamino Schema Definitio 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 fiel in the source database is integrated. It is the task of the curator to capture its relevant subparts as well as to defin 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 identifier 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 identifier 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 firs search the maps database to fin 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 difficul 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

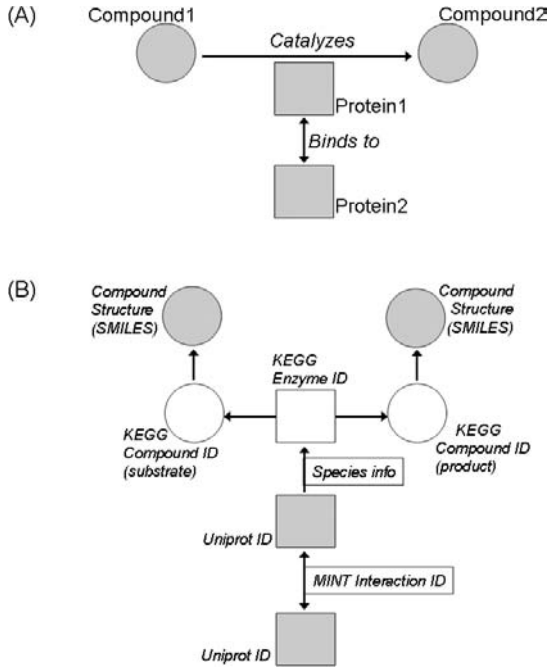


Fig. 3. (**A**) Schematic representation of relationships between two compounds and two proteins. (**B**) Same representation as hypothetically resolved via traversals across multiple databases.

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 specific 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 configuratio 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 configuratio 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.

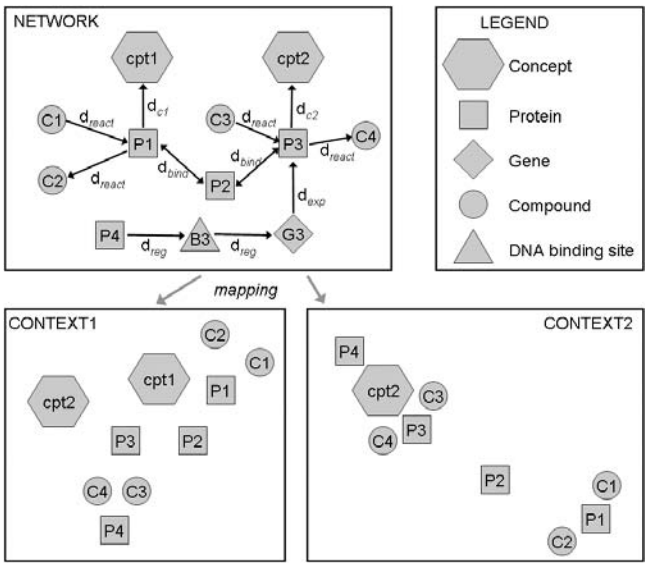


Fig. 4. Illustrative example of using graph projection in exploratory analysis of biological networks. In CONTEXT1 we are weighting all types of relationships similarly, so the nodes are clustered based on shortest path length between the edges. In CONTEXT2, we are interested only in concept cpt2, and 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 firs 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). Specificall , 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 firs 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

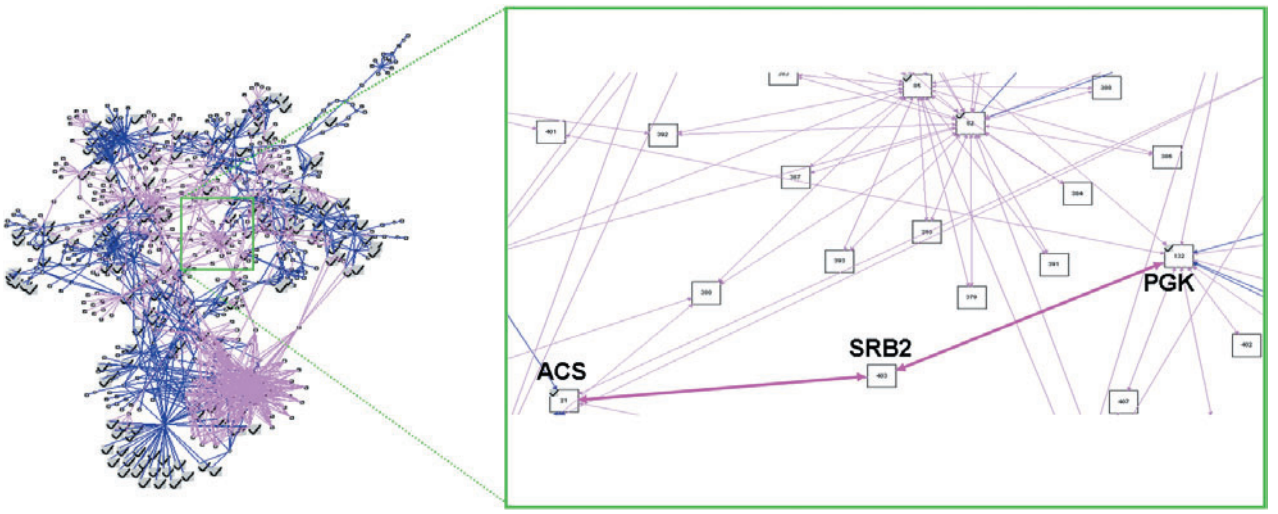


Fig. 5. Integrated pathway retrieval using megNet network visualization tool, with the query for 'Glycolysis/Gluconeogenesis AND Pentose phosphate pathway AND TCA cycle IN *S.cerevisiae*'. Metabolic pathways are shown with blue edges, protein–protein interactions with pink. Proteins are represented with 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 specifi 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.

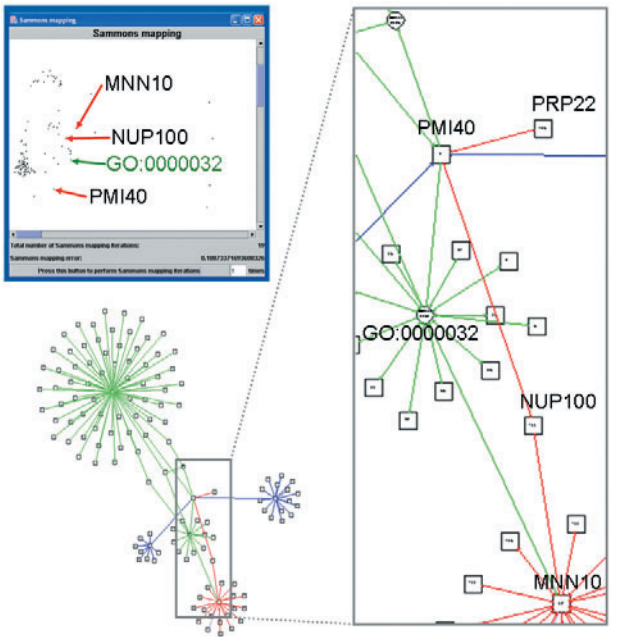


Fig. 6. Network neighborhood of mannose-6-phosphate isomerase (PMI40) in *S.cerevisiae*. Metabolic pathway relationships are shown in blue, protein–protein interactions in red, and GO associations in green. Both PMI40 and MNN10 are involved in cell wall mannoprotein synthesis (GO:0000032). NUP100 protein, 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 firs needs to defin distance measures across heterogeneous types of information. We are taking a pragmatic approach by letting the user defin the distances as a part of the query. This is reasonable since the distance basically define the context of the questions posed by the user and allows biasing the similarity toward particular types of relationships, or toward relationships in a specifi 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 reflec a subjective choice and as such will be subject to debate. It is ultimately the end result of mining that determines the utility of specifi 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. Specificall , 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 profile 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 profil 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 unificatio 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 Unifie 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 scientifi 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 fla fil 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.

PUBLICATION III

An integrative approach for biological data mining and visualization

In: International Journal of Data Mining and Bioinformatics 2008, 2(1)1:54–77.

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.

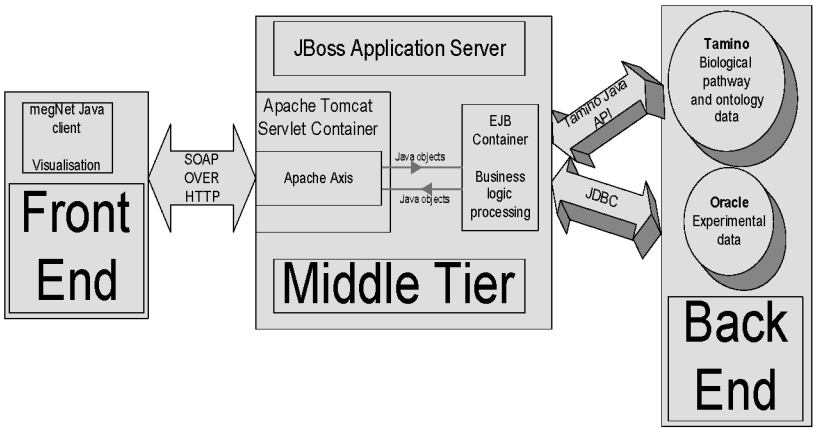
2 Integrated database system

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).

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

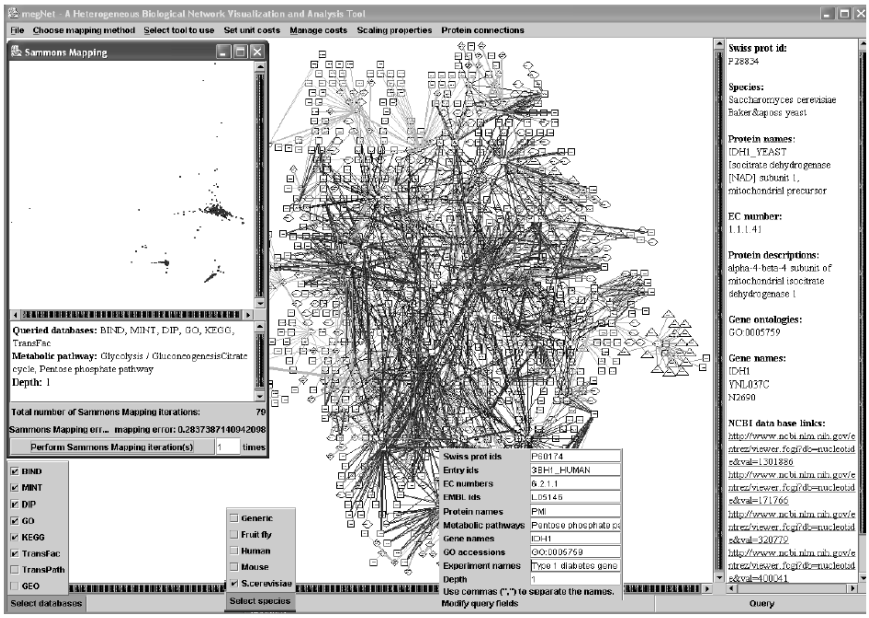
3.1 Network visualisation

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



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:

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

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:

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

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:

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

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*

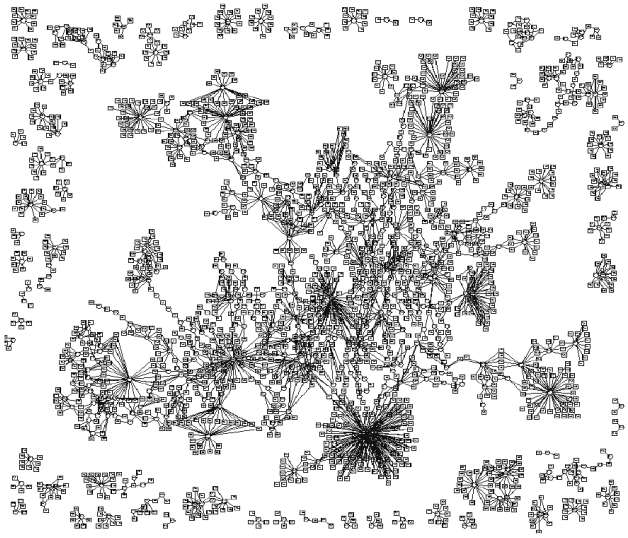


Figure 4 Degree distribution of the yeast metabolic network shown in Figure 3. It appears that the degree distribution does not follow the power law which means that there is no hierarchical modularity in this metabolic network

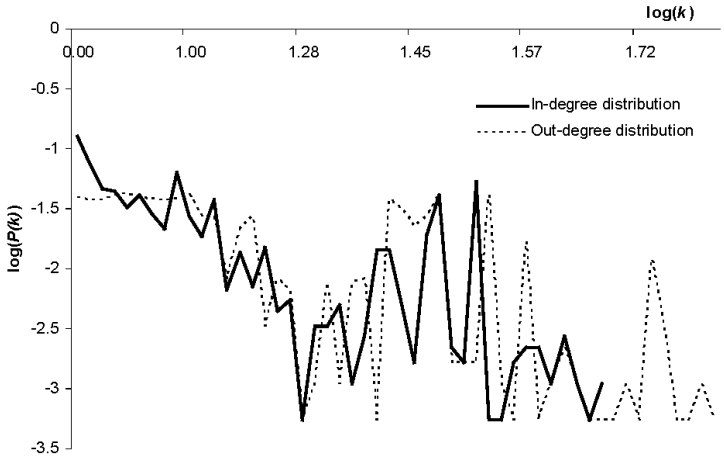
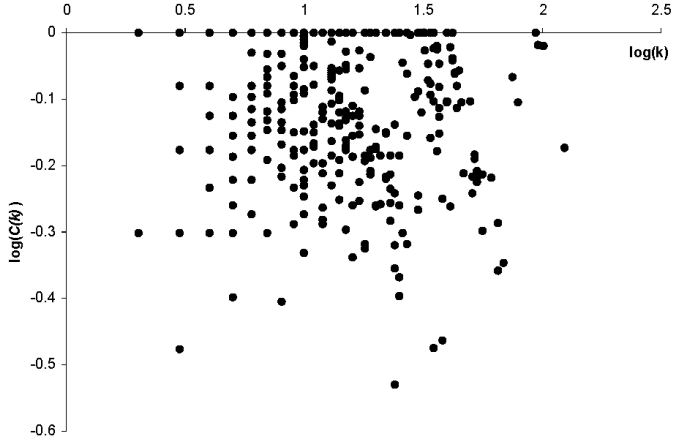


Figure 5 Clustering coefficient as a function of node degree for the yeast metabolic network. Here the clustering coefficient does not seem to follow the power law either, which suggests that there is no hierarchical modularity in our network



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

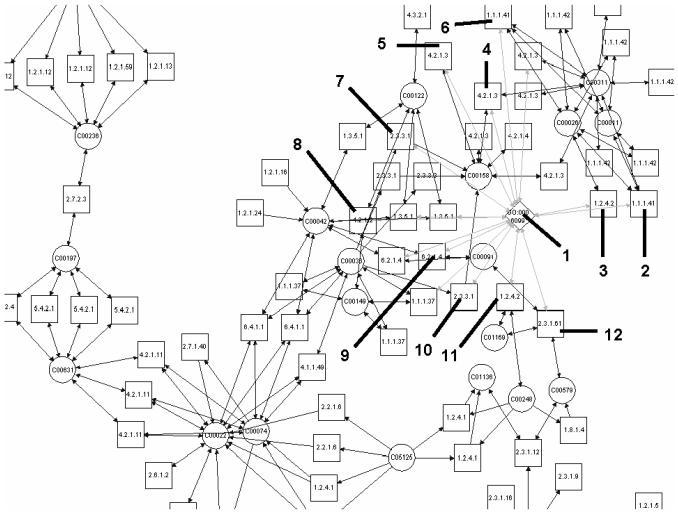
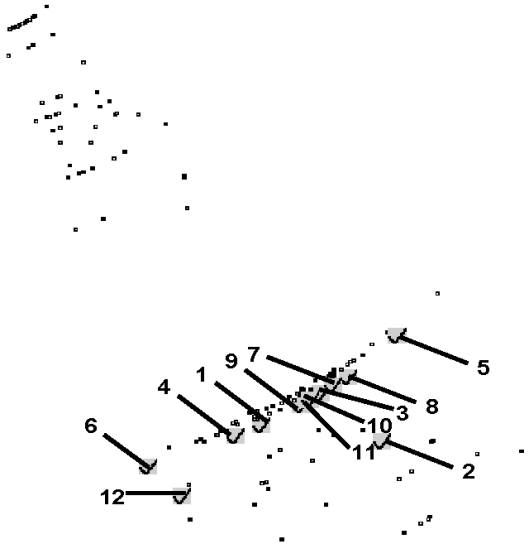


Figure 7 A Curvilinear Distance Analysis projection biasing *tricarboxylci acid cycle.* The projection was obtained by lowering the distance of all connected edges to TCA node (number 1) in the above graph



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

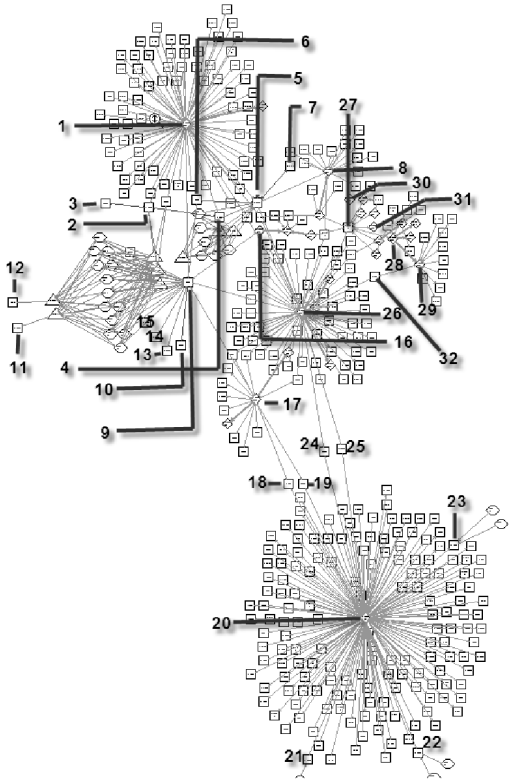


Table 2 Short description of select entities from the network shown in Figure 8

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).

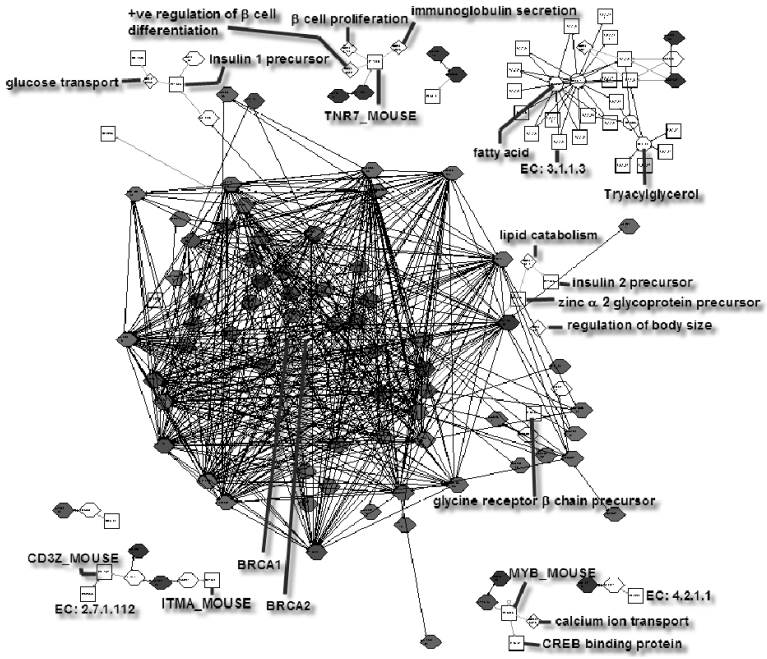
As an illustration of combining gene expression data with the existing pathways and ontologies, we utilised gene expression data from mouse congenic strains in a study related to Type 1 Diabetes (Eaves et al., 2002). We processed this data as explained below in order to construct the query. The resulting 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/.

PUBLICATION IV

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

In: Proceedings of KRBIO'05, International Symposium on Knowledge Representation in Bioinformatics, Espoo, Finland, Jun 2005. 6 p.

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.

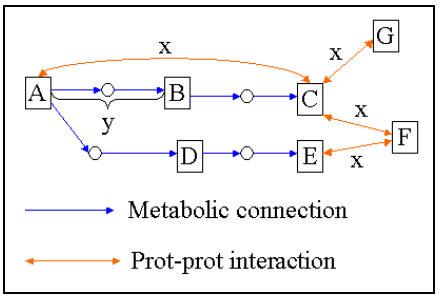


Figure 1: Example of our integrated network representation used. The distance between the entities A and B, is the same as for B to A. If there is not any path between two nodes, we assume that the distance between them is infinity.

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:

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

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.

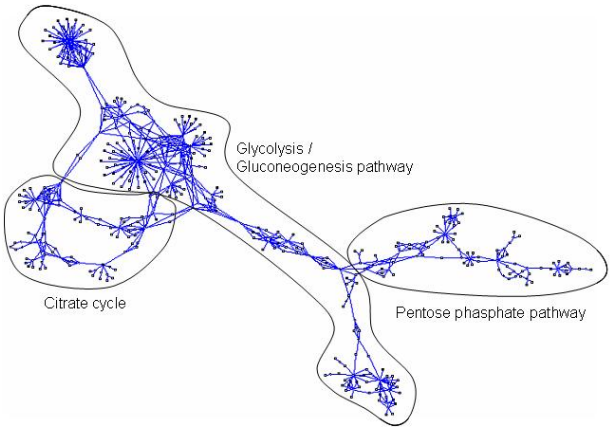


Figure 2: KEGG metabolic pathways for "Glycolysis / Gluconeogenesis , Pentose phosphate and Citrate cycle pathways.

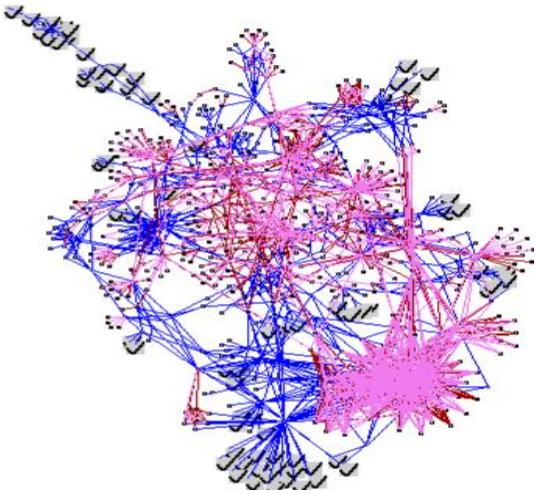


Figure 3: Metabolic pathway (KEGG) enriched with protein-protein interactions from MINT and DIP databases for "Glycolysis / Gluconeogenesis, Pentose phosphate and Citrate cycle pathways,. The proteins loosely connected are highlighted with gray marks.

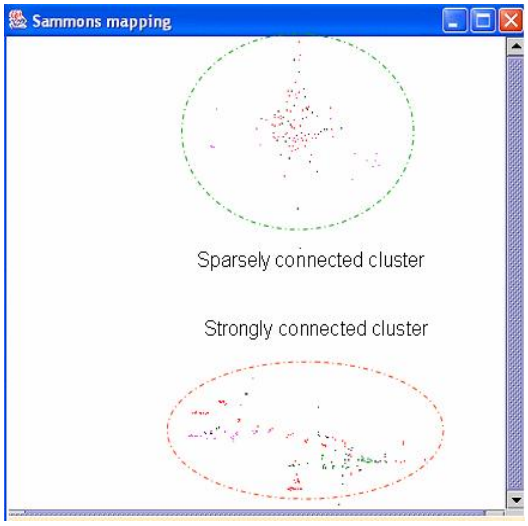


Figure 4: Clusters from Sammon's mapping of the previous graph. Two main clusters emerged, one strongly connected and one loosely connected.

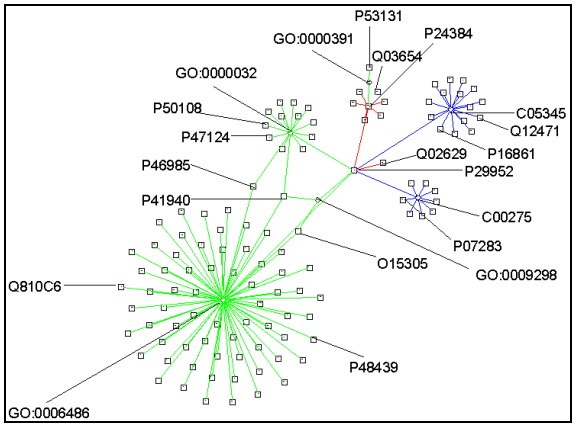


Figure 5: Search result of pathway query for mannose synthesis GO:0009298.

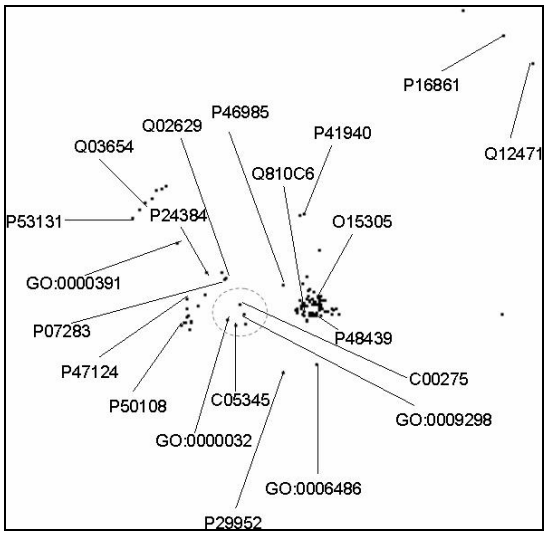


Figure 6: Sammon's mapping of the previous network for "Context 1: Every edge has equal weight".

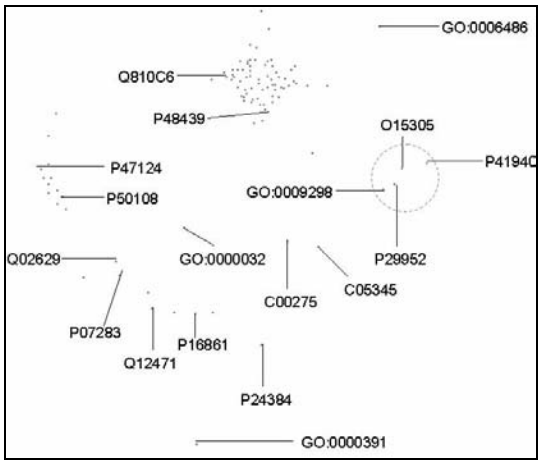


Figure 7: The Sammon's mapping for "Context 2: The neighborhood edges of GO:0009298 have higher weights than the other edges".

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.

Series title, number and
report code of publication

VTT Publications 774
VTT-PUBS-774

Author(s)
Erno Lindfors

Title

Network Biology
Applications in medicine and biotechnology

Abstract

The concept of systems biology emerged over the last decade in order to address advances in experimental techniques. It aims to characterize biological systems comprehensively as a complex network of interactions between the system's components. Network biology has become a core research domain of systems biology. It uses a graph theoretic approach. Many advances in complex network theory have contributed to this approach, and it has led to practical applications spanning from disease elucidation to biotechnology during the last few years.

Herein we applied a network approach in order to model heterogeneous biological interactions. We developed a system called megNet for visualizing heterogeneous biological data, and showed its utility by biological network visualization examples, particularly in a biomedical context. In addition, we developed a novel biological network analysis method called Enriched Molecular Path detection method (EMPath) that detects phenotypic specific molecular paths in an integrated molecular interaction network. We showed its utility in the context of insulitis and autoimmune diabetes in the non-obese diabetic (NOD) mouse model. Specifically, ether phosholipid biosynthesis was down-regulated in early insulitis. This result was consistent with a previous study in which serum metabolite samples were taken from children who later progressed to type 1 diabetes and from children who permanently remained healthy. As a result, ether lipids were diminished in the type 1 diabetes progressors. Also, in this thesis we performed topological calculations to investigate whether ubiquitous complex network properties are present in biological networks. Results were consistent with recent critiques of the ubiquitous complex network properties describing the biological networks, which gave motivation to tailor another method called Topological Enrichment Analysis for Functional Subnetworks (TEAFS). This method ranks topological activities of modules of an integrated biological network under a dynamic response to external stress. We showed its utility by exposing an integrated yeast network to oxidative stress. Results showed that oxidative stress leads to accumulation of toxic lipids.

ISBN
978-951-38-7758-3 (soft back ed.)
978-951-38-7759-0 (URL: http://www.vtt.fi/publications/index.jsp)

Series title and ISSN
VTT Publications
1235-0621 (soft back ed.)
1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)

Project number
74263

Date	Language	Pages
October 2011	English, Finnish abstr.	81 p. + app. 100 p.

Keywords
Network biology, systems biology, biological data visualization, type 1 diabetes, oxidative stress, graph theory, network topology, ubiquitous complex network properties

Publisher
VTT Technical Research Centre of Finland
P.O. Box 1000, FI-02044 VTT, Finland
Phone internat. +358 20 722 4520
Fax +358 20 722 4374

Julkaisun sarja, numero ja raporttikoodi

VTT Publications 774
VTT-PUBS-774

Tekijä(t)
Erno Lindfors

Nimeke

Verkkobiologia
Lääketieteellisiä ja bioteknisiä sovelluksia

Tiivistelmä

Järjestelmäbiologian käsite syntyi yli kymmenen vuotta sitten vastauksena kokeellisten menetelmien kehitystyöhön. Tämä lähestymistapa pyrkii kuvaamaan biologisia järjestelmiä kattavasti kompleksisena vuorovaikutusverkkona, joka koostuu järjestelmän komponenttien välisistä vuorovaikutuksista. Verkkobiologiasta on tullut tärkeä järjestelmäbiologian tutkimuskohde, ja se käyttää graafiteoreettista lähestymistapaa. Kompleksisten verkkojen teorian kehitystyö on edistänyt tätä lähestymistapaa, ja se on johtanut moniin käytännön sovelluksiin aina sairauksien selventämisestä bioteknologiaan viimeisten parin vuoden aikana. Tässä väitöskirjassa sovellettiin verkkobiologista lähestymistapaa heterogeenisten biologisten vuorovaikutusten mallintamiseen. Siinä kehitettiin heterogeenisen biologisen tiedon visualisointityökalu megNet, jonka hyödyllisyys osoitettiin biologisten verkkojen visualisointiesimerkein, erityisesti biolääketieteellisessä kontekstissa. Tämän lisäksi väitöstutkimuksessa kehitettiin uusi biologisten verkkojen analysointimenetelmä, rikastettujen molekyylipolkujen havaitsemismenetelmä, joka havaitsee fenotyyppikohtaisia molekyylipolku-ja integroidusta molekyylivuorovaikutusverkosta. Tämän menetelmän hyödyllisyys osoitettiin insuliitiksen ja autoimmuunidiabeteksen kontekstissa käyttäen laihojen diabeteshiirien mallia. Erityisesti eetterifosfolipidibiosynteesi oli alisäädelty insuliitiksen varhaisessa vaiheessa. Tämä tulos oli yhteensopiva aikaisemman tutkimuksen kanssa, jossa mitattiin myöhemmin tyypin 1 diabetekseen sairastuneiden lasten ja pysyvästi terveiden lasten seerumin aineenvaihduntatuotteidenpitoisuuksia. Tässä tutkimuksessa havaittiin, että eetterilipidipitoisuudet olivat sairastuneilla lapsilla alhaisemmat kuin terveillä lapsilla. Tässä väitöskirjassa lasketaan myös topologialaskuja, joiden avulla voidaan selvittää, noudattavatko biologiset verkot kaikkialla läsnä olevia kompleksisten verkkojen ominaisuuksia. Tulokset olivat yhteensopivia kaikkialla läsnä olevien kompleksisten verkkojen ominaisuuksiin viime aikoina kohdistuneen kritiikin kanssa. Tämä loi motivaatiota räätälöidä topologista rikastamisanalyysia funktionaalisille aliverkoille, joka etsii topologisesti aktiivisimmat moduulit integroidusta biologisesta verkosta dynaamisen stressin alaisuudessa. Tämän menetelmän hyödyllisyys osoitettiin altistamalla integroitu hiivaverkko oksidatiiviselle stressille. Tulokset osoittivat, että oksidatiivinen stressi aiheuttaa toksisten lipidien kasaantumisen.

ISBN
978-951-38-7758-3 (nid.)
978-951-38-7759-0 (URL: http://www.vtt.fi/publications/index.jsp)

Avainnimeke ja ISSN	Projektinumero
VTT Publications 1235-0621 (nid.) 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)	74263

Julkaisuaika	Kieli	Sivuja
Lokakuu 2011	Englanti, suom. tiiv.	81 s. + liitt. 100 s.

Avainsanat	Julkaisija
Network biology, systems biology, biological data visualization, type 1 diabetes, oxidative stress, graph theory, network topology, ubiquitous complex network properties	VTT PL 1000, 02044 VTT Puh. 020 722 4520 Faksi 020 722 4374

Network biology uses a graph theoretic approach to characterize biological systems comprehensively as a complex network of interactions. This approach has led to practical applications spanning from disease elucidation to biotechnology during the last few years.

In this thesis we applied a network approach in order to model heterogeneous biological interactions. We developed a system for visualizing heterogeneous biological data, and showed its utility by biological network visualization examples. In addition, we developed a novel biological network analysis method that detects phenotypic specific molecular paths in an integrated molecular interaction network. We showed the utility of this method in the context of type 1 diabetes mouse models, and found that ether phospholipid biosynthesis was down-regulated in early state of type 1 diabetes, which was consistent with recent clinical findings. Also, we performed topological calculations on biological networks, and obtained consistent results with recent critiques of ubiquitous complex network properties describing the biological networks. This gave motivation to tailor a topological enrichment analysis method. We showed the utility of this method by exposing an integrated yeast network to oxidative stress. Results showed that oxidative stress leads to accumulation of toxic lipids.

ISBN 978-951-38-7758-3 (soft back ed.)
ISSN 1235-0621 (soft back ed.)
ISBN 978-951-38-7759-0 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)