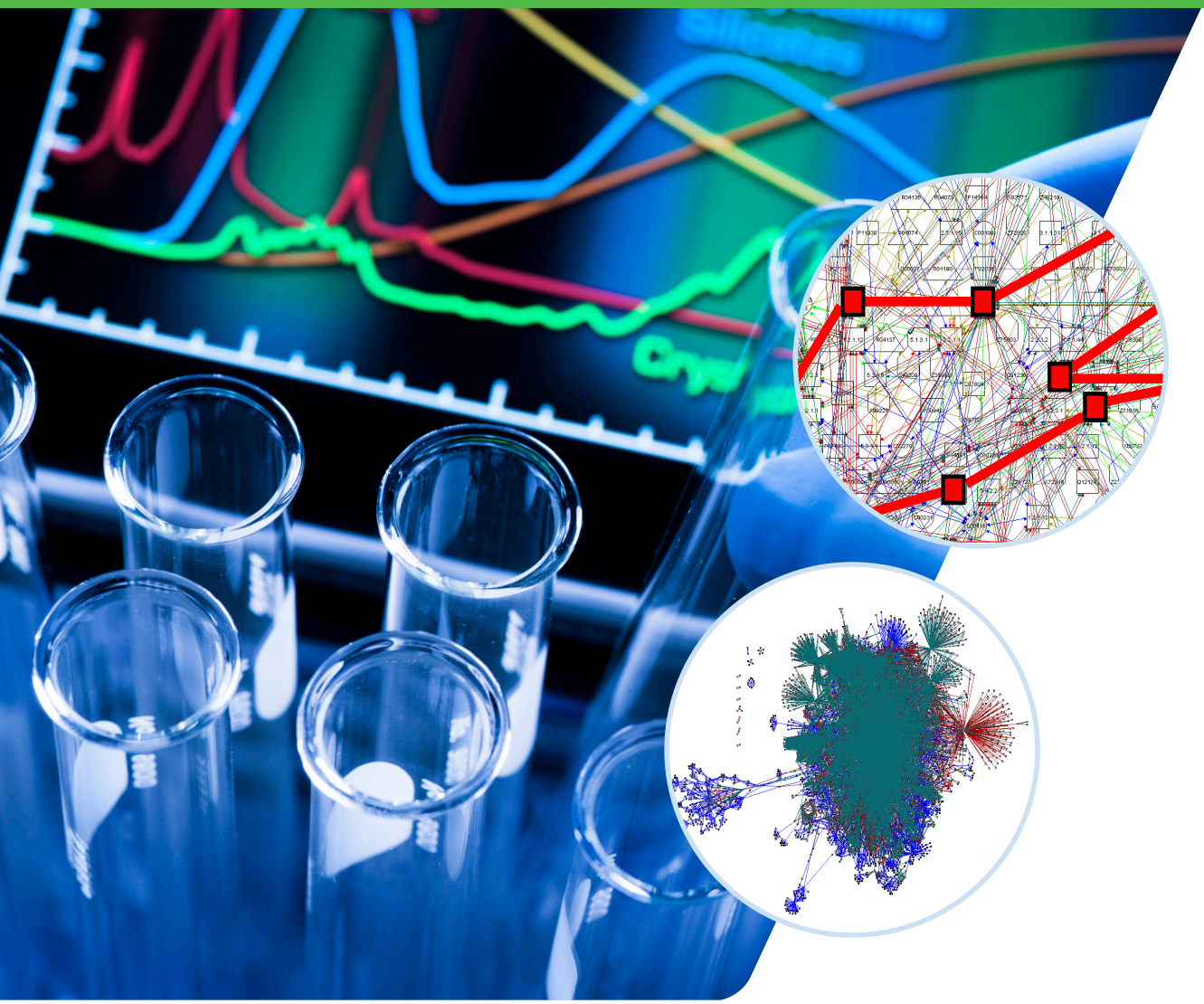
Dissertation VTT PUBLICATIONS 774



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

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

List of abbreviations

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.

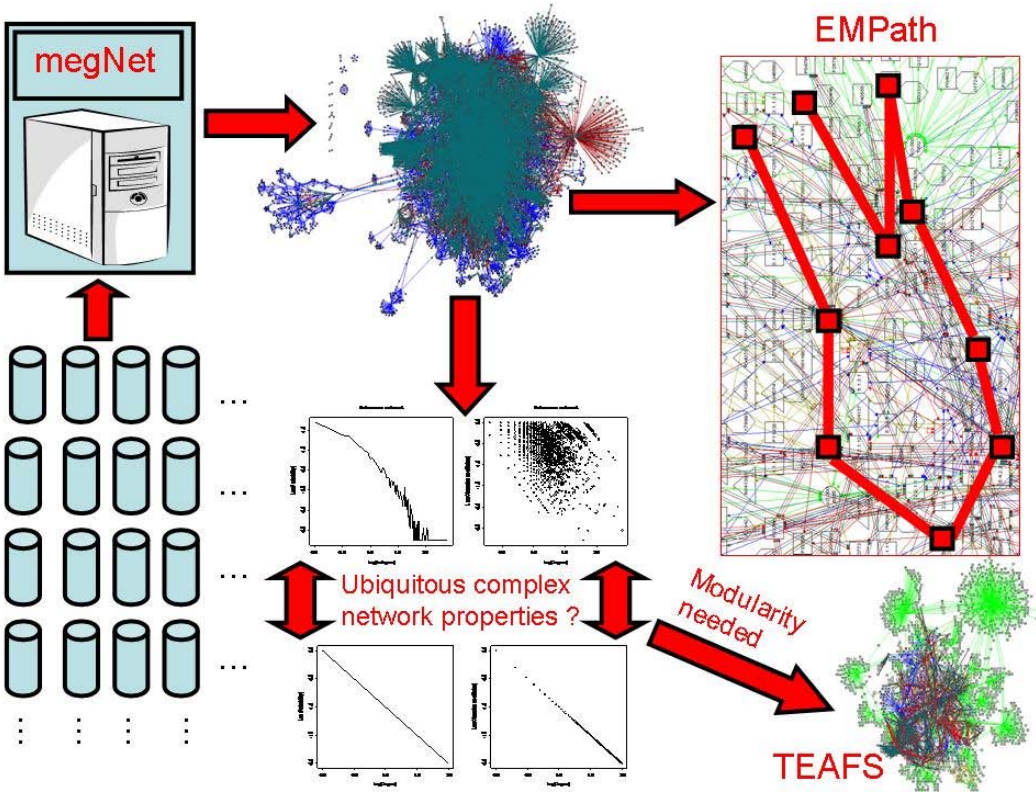


Figure 1.1. Schematic diagram summarizing the main aims of this thesis.

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.

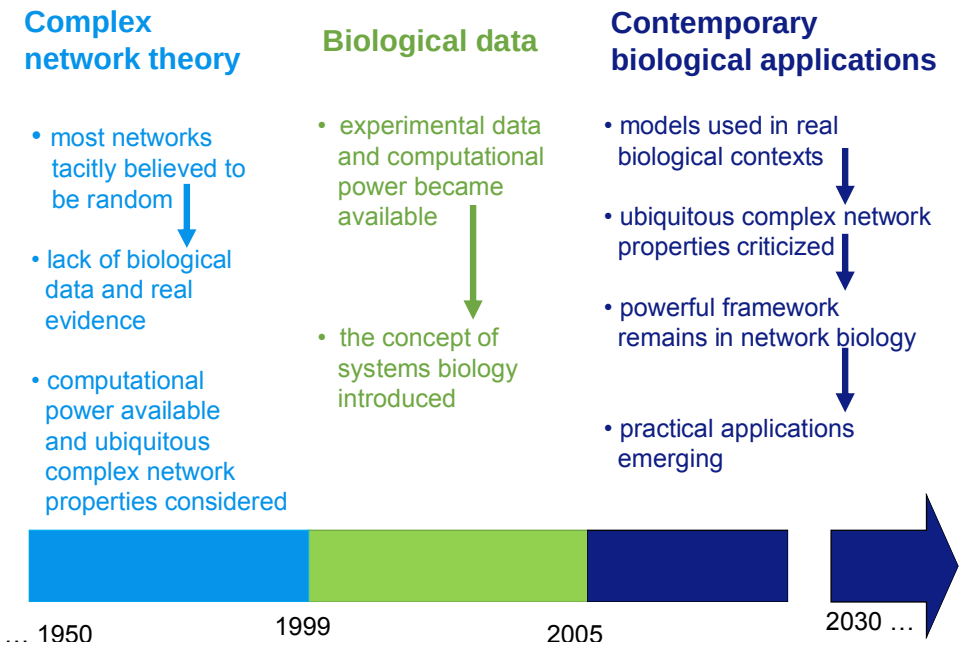


Figure 2.1. Main parts of network biology.

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.

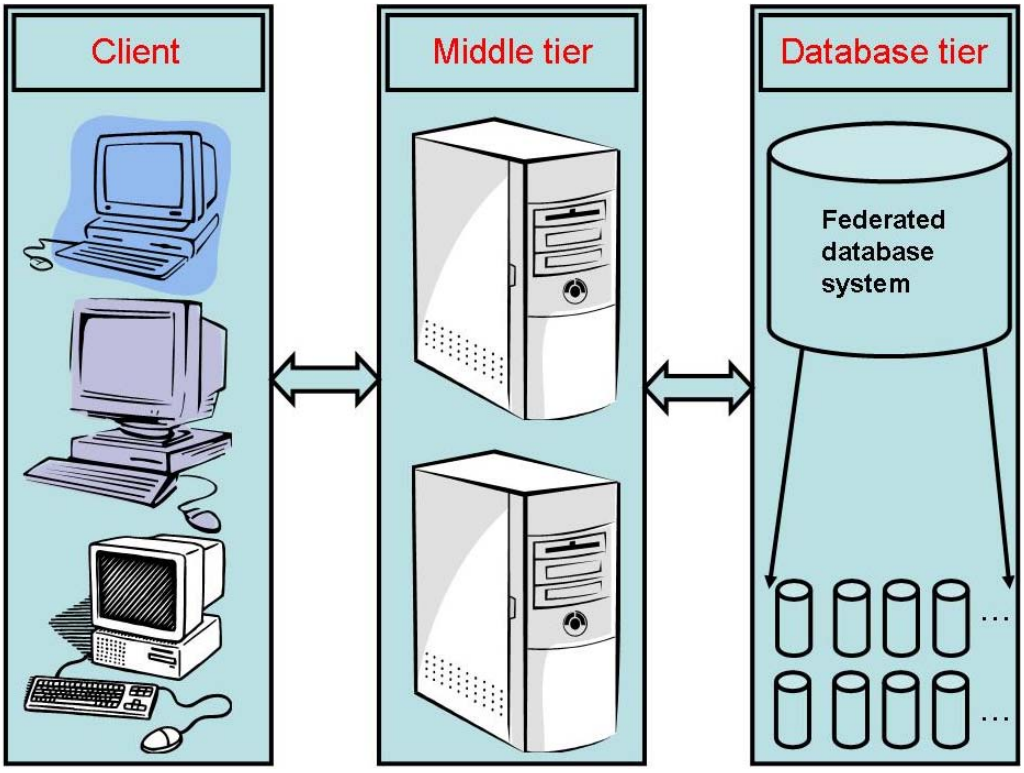


Figure 3.1. Main components of megNet.

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.

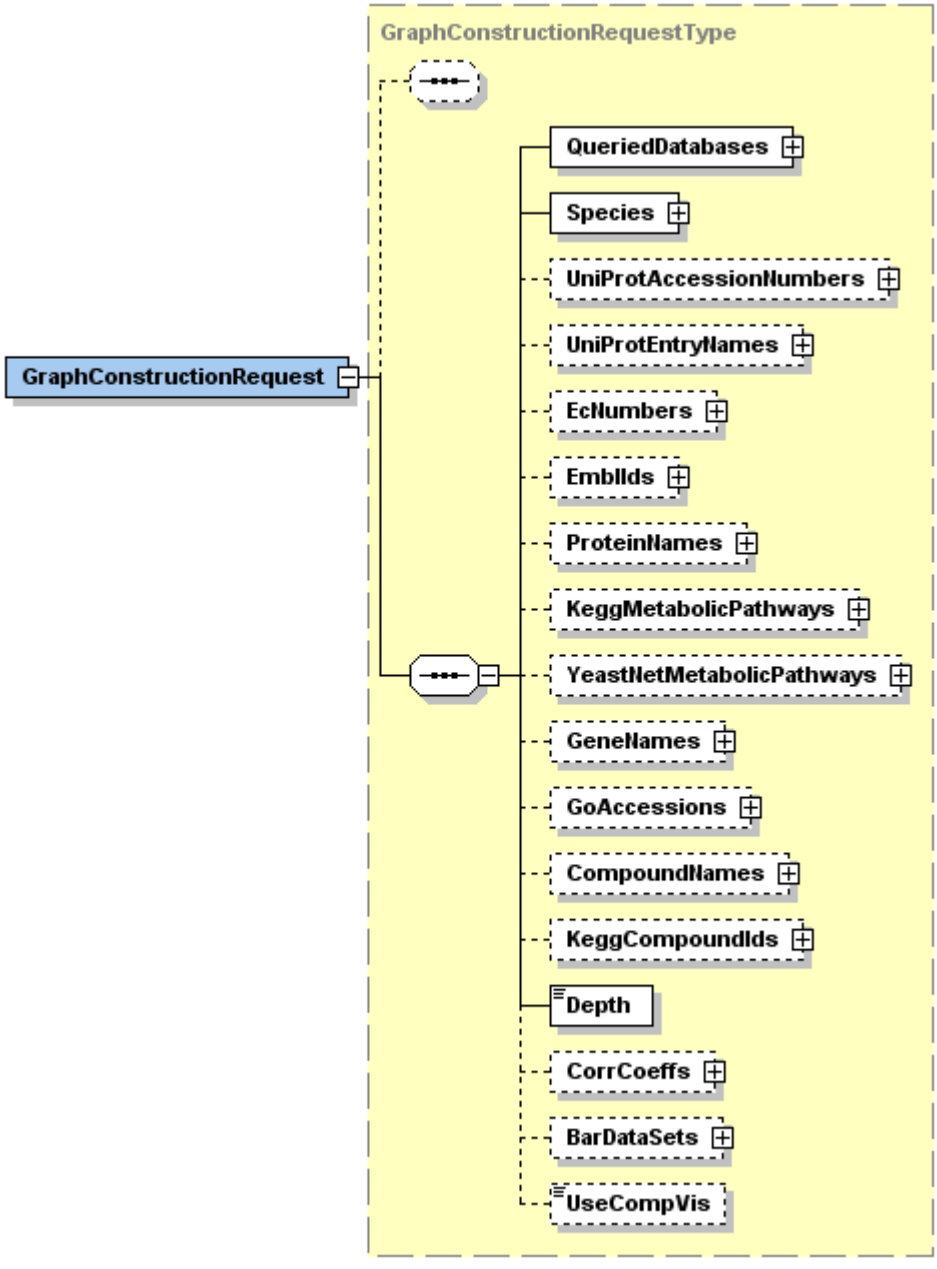


Figure 3.2. XML schema for graph construction request.

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.

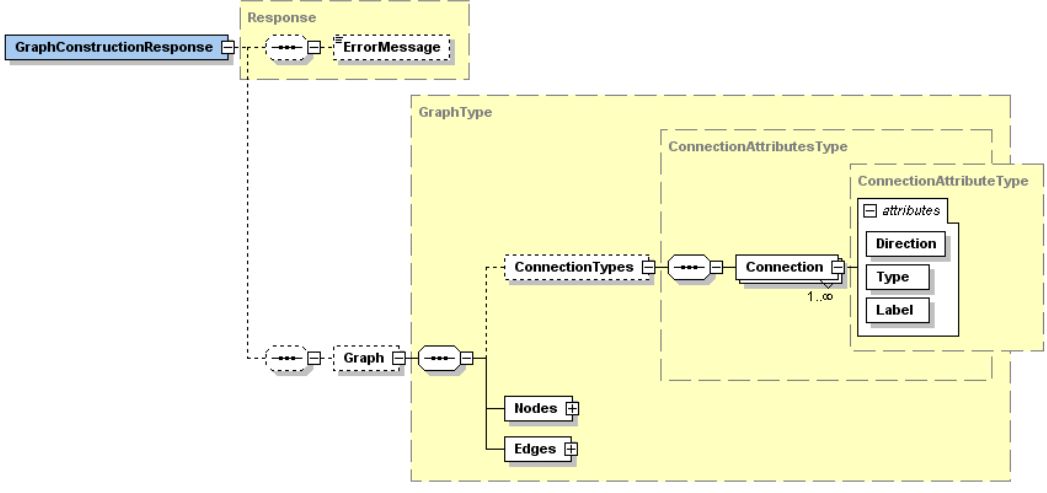


Figure 3.3. XML schema for the main elements of graph construction response. The nodes and edges elements are opened in Figures 3.4 and 3.5 respectively.

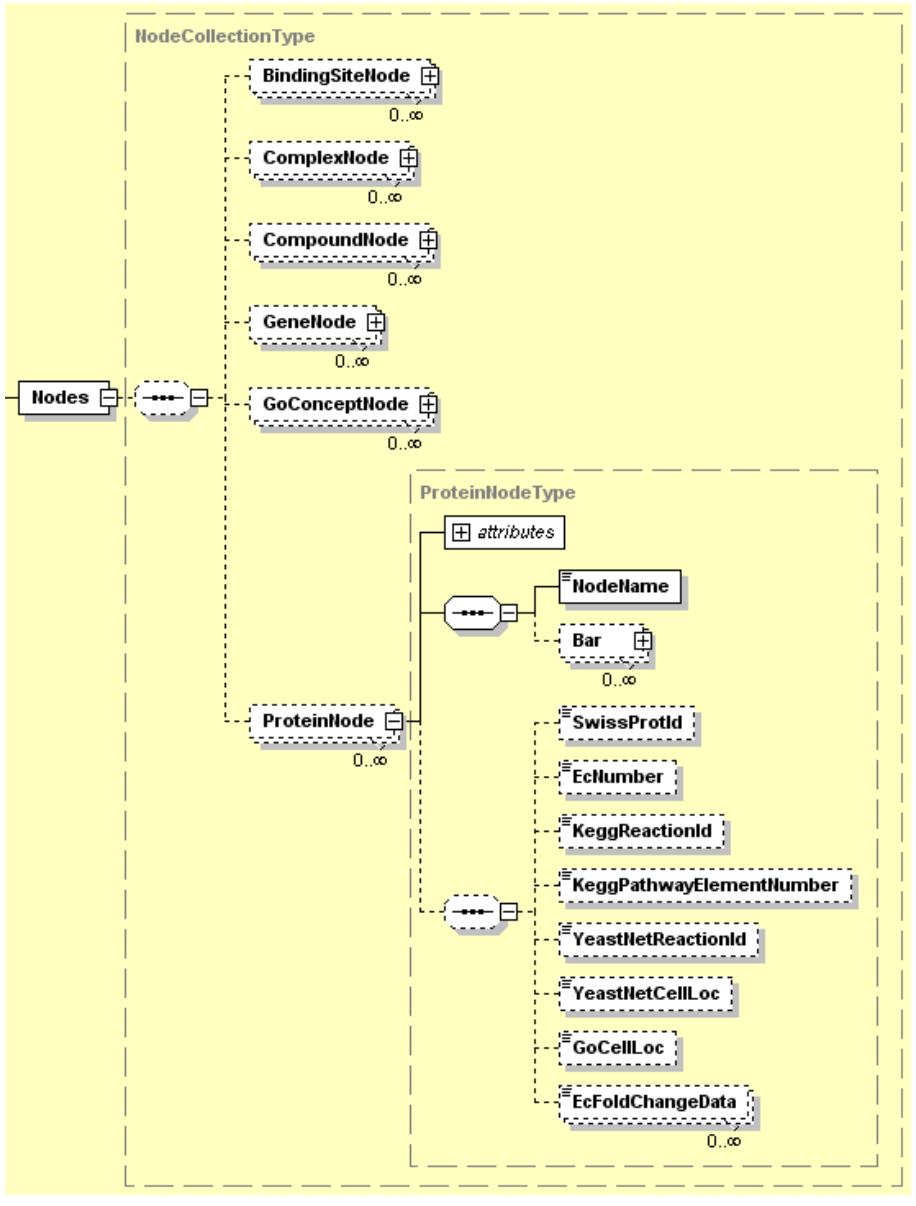


Figure 3.4. XML schema for the nodes element of graph construction response.

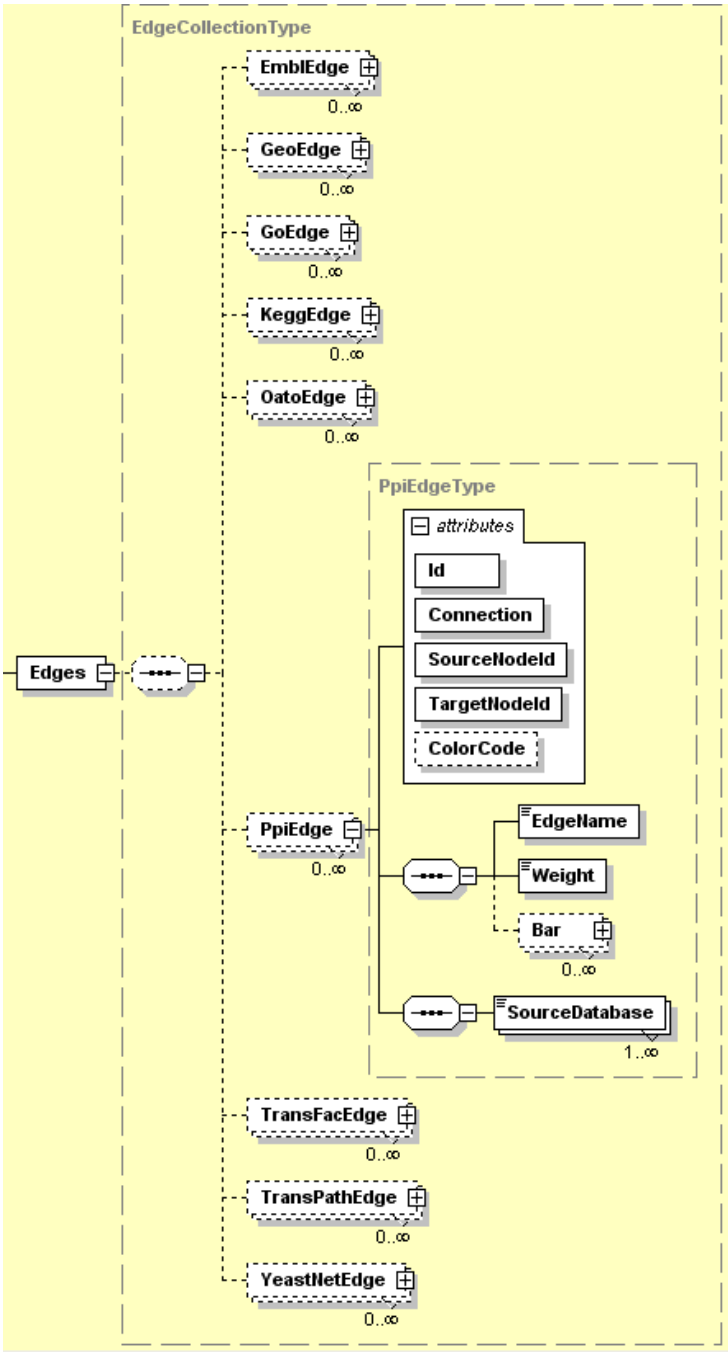


Figure 3.5. XML schema for the edges element of graph construction response.

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.

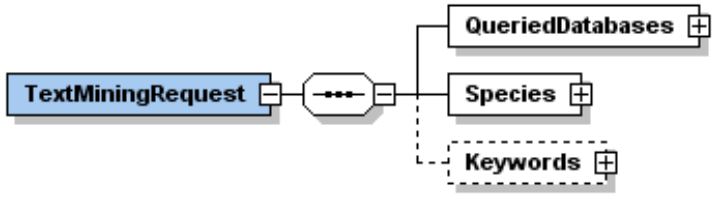


Figure 3.6. XML schema for text mining request.

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.

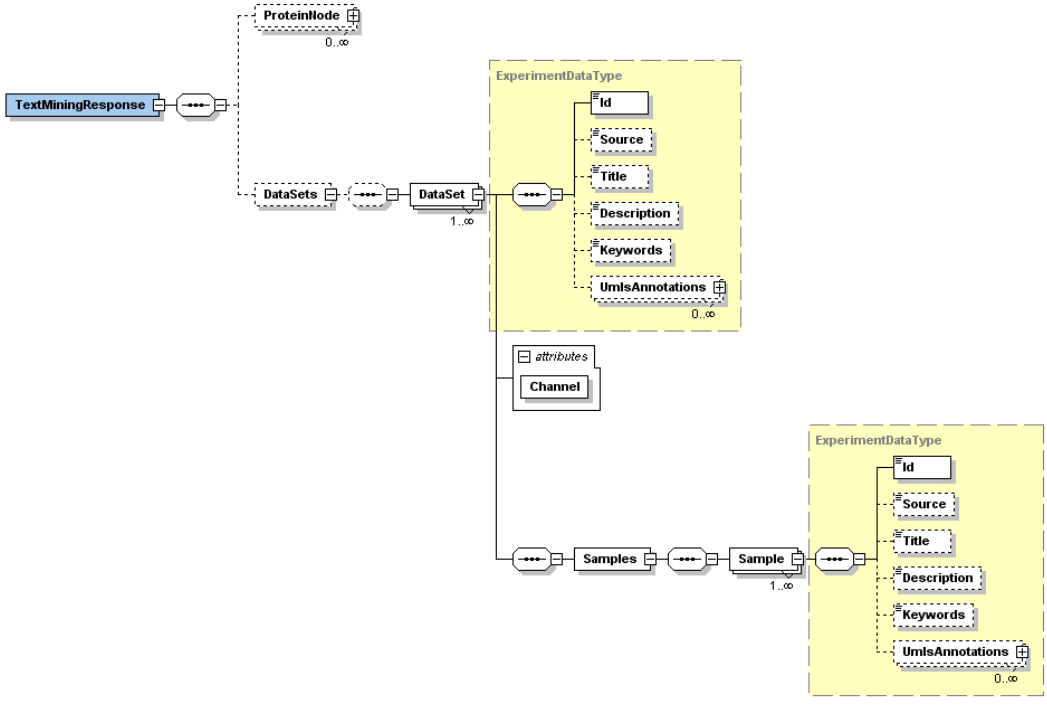


Figure 3.7. XML schema for text mining response.

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**.

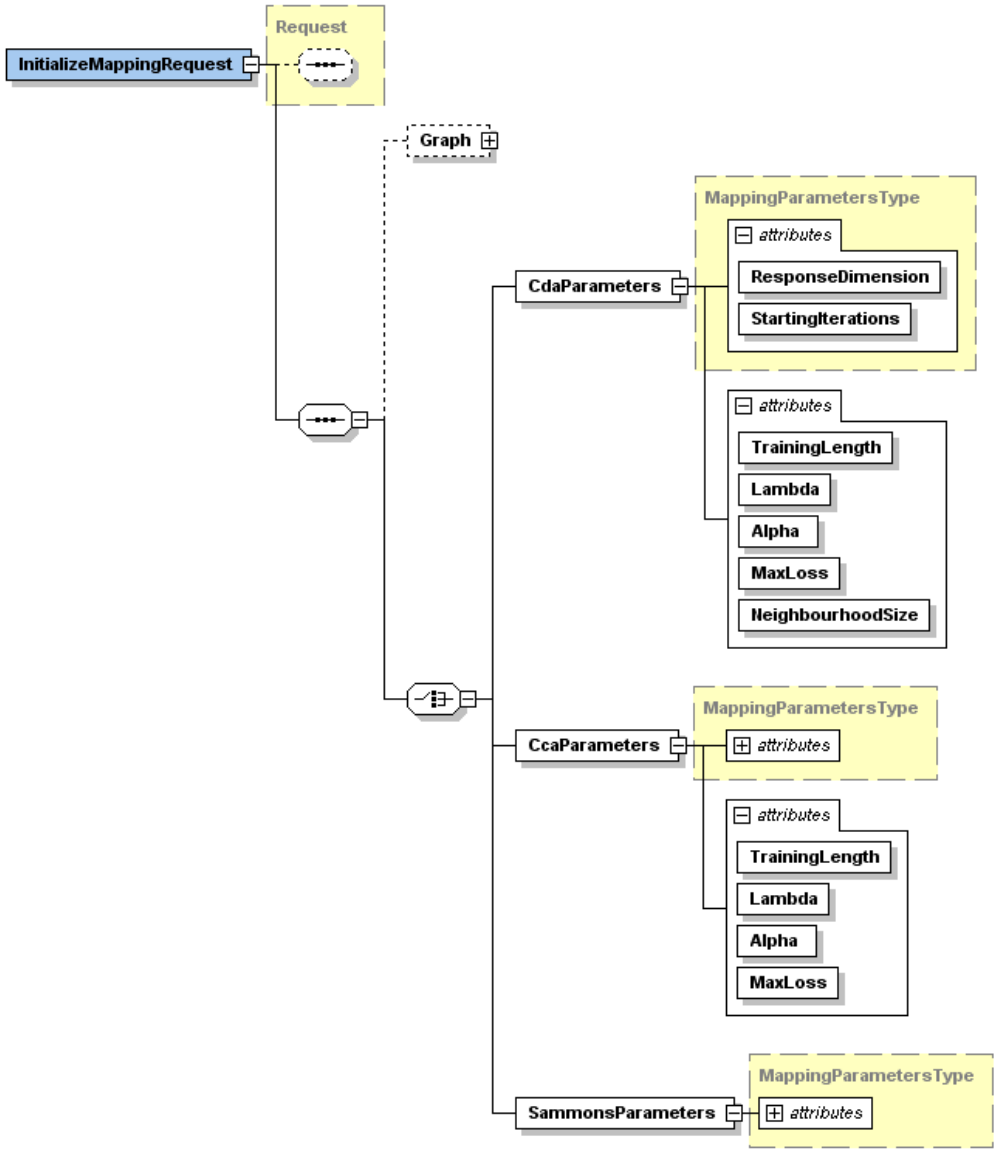


Figure 3.8. XML schema for initialize mapping request.

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).

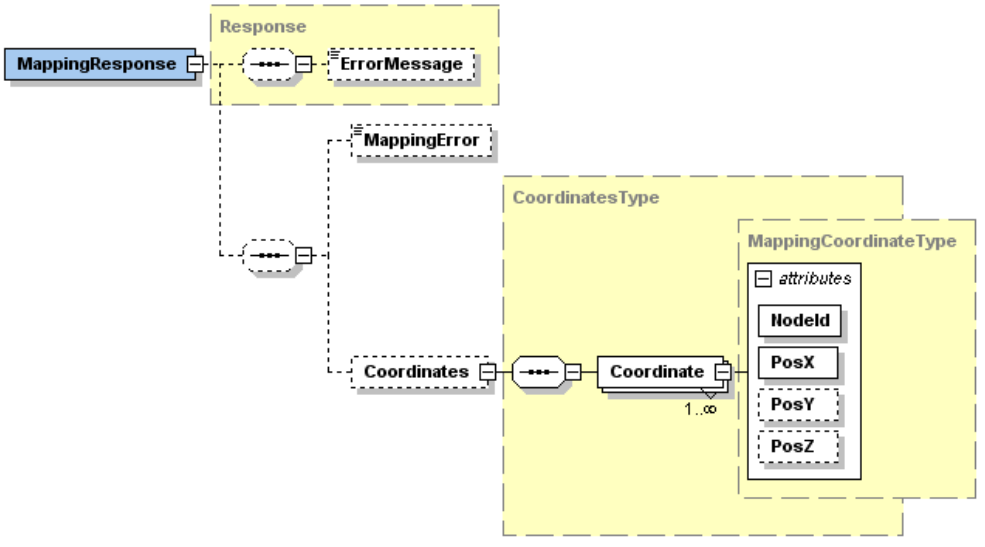


Figure 3.9. XML schema for mapping response.

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).

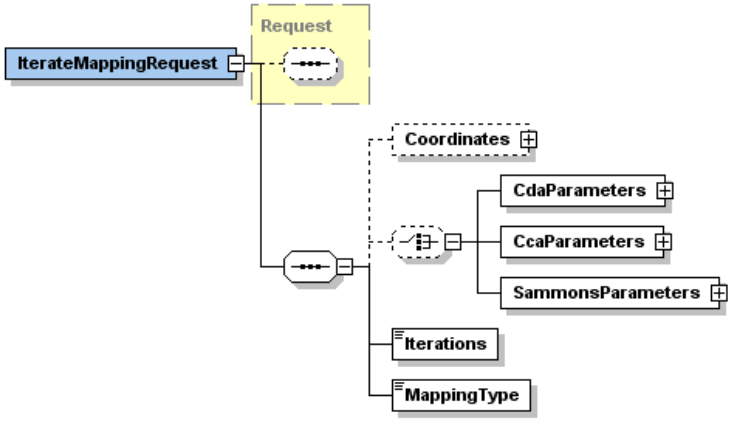


Figure 3.10. XML schema for iterate mapping request.

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.

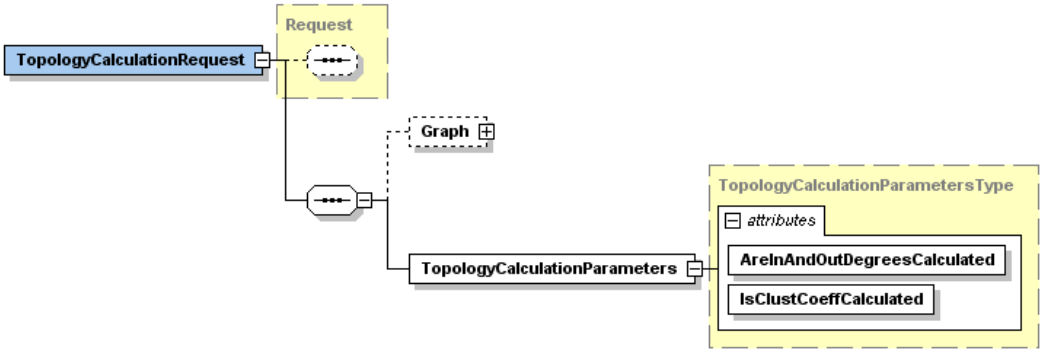


Figure 3.11. XML schema for topology calculation request.

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).

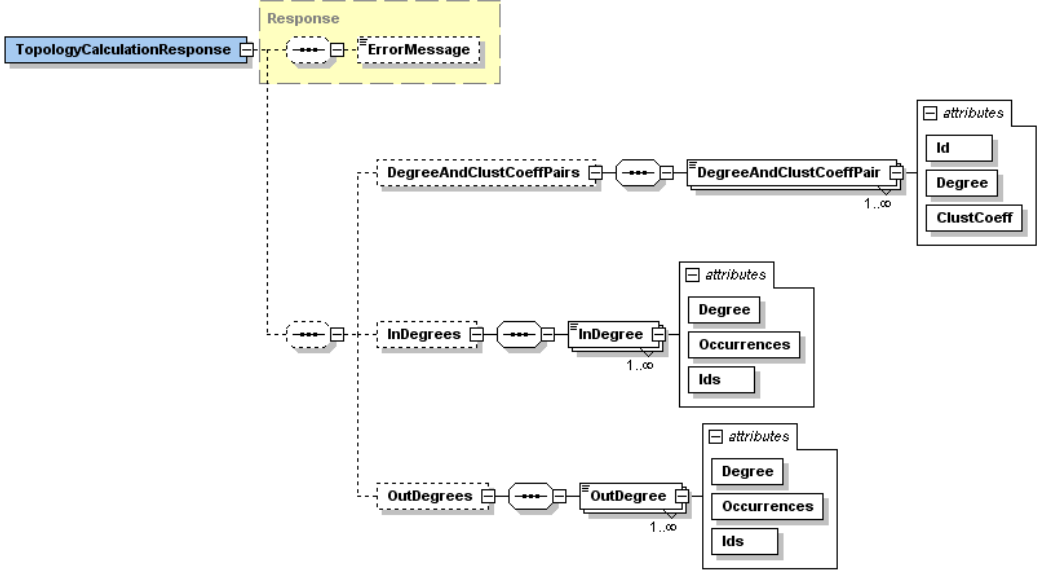


Figure 3.12. XML schema for topology calculation response.

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.

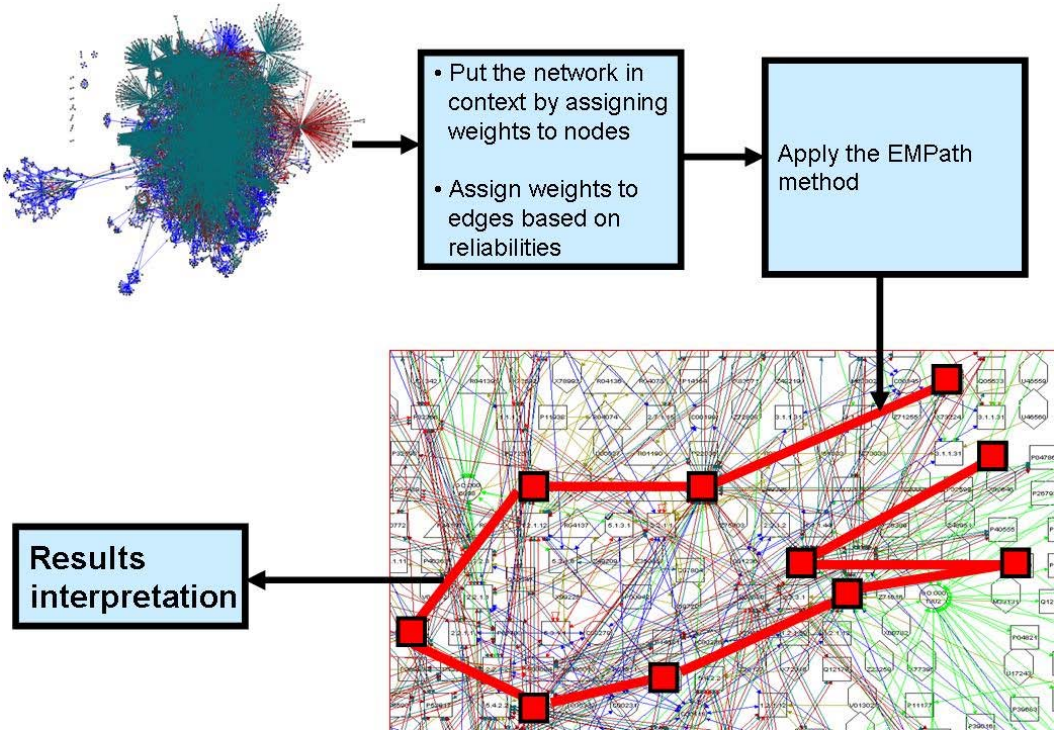


Figure 3.13. The schematic pipeline of the EMPath method.

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.

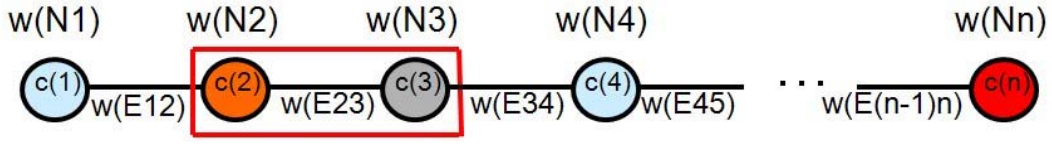


Figure 3.14. The scoring and coloring of the EMPath method.

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.

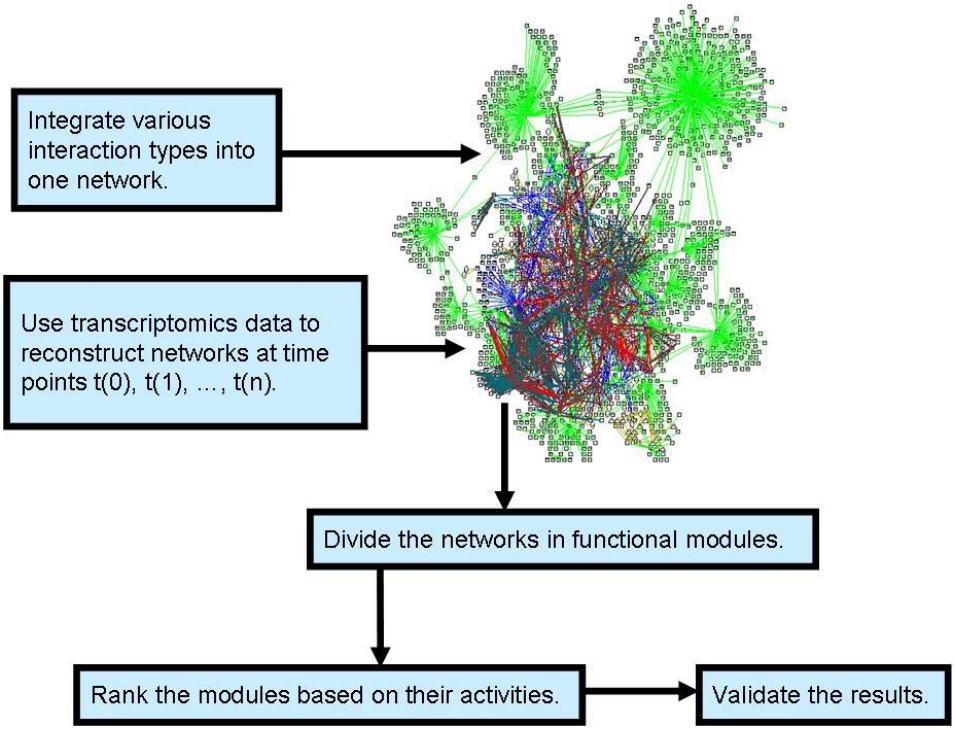


Figure 3.15. The schematic pipeline of the TEAFS method.

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.

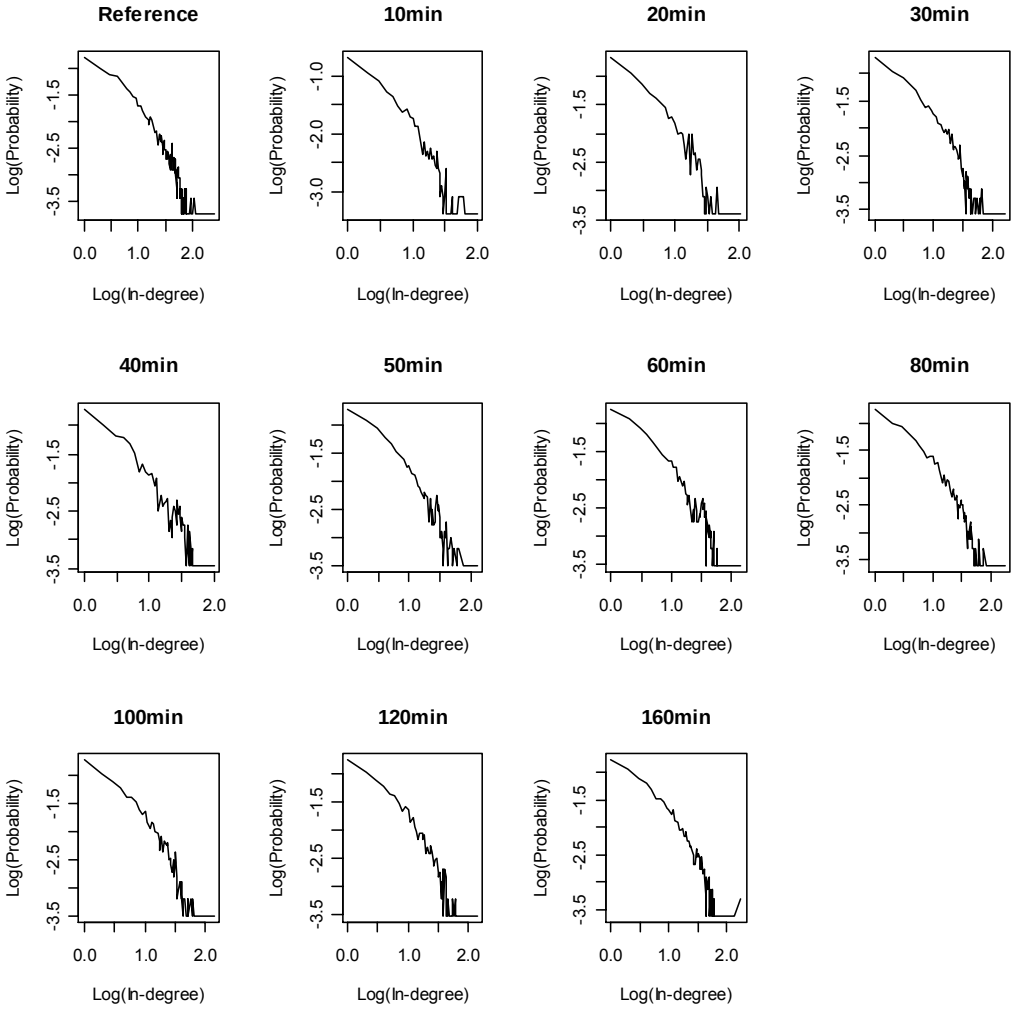


Figure 4.1. In-degree distributions for reference and networks at time points.

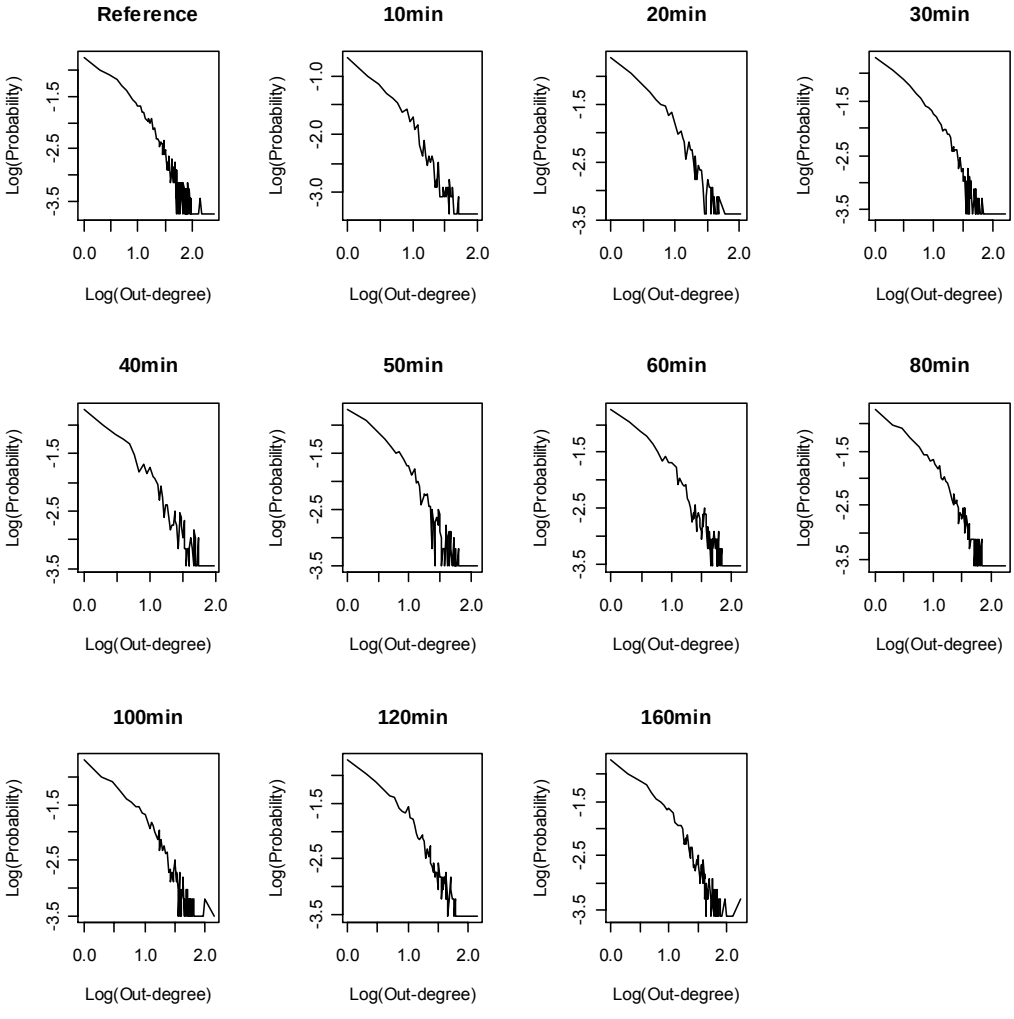


Figure 4.2. Out-degree distributions for reference and networks at time points.

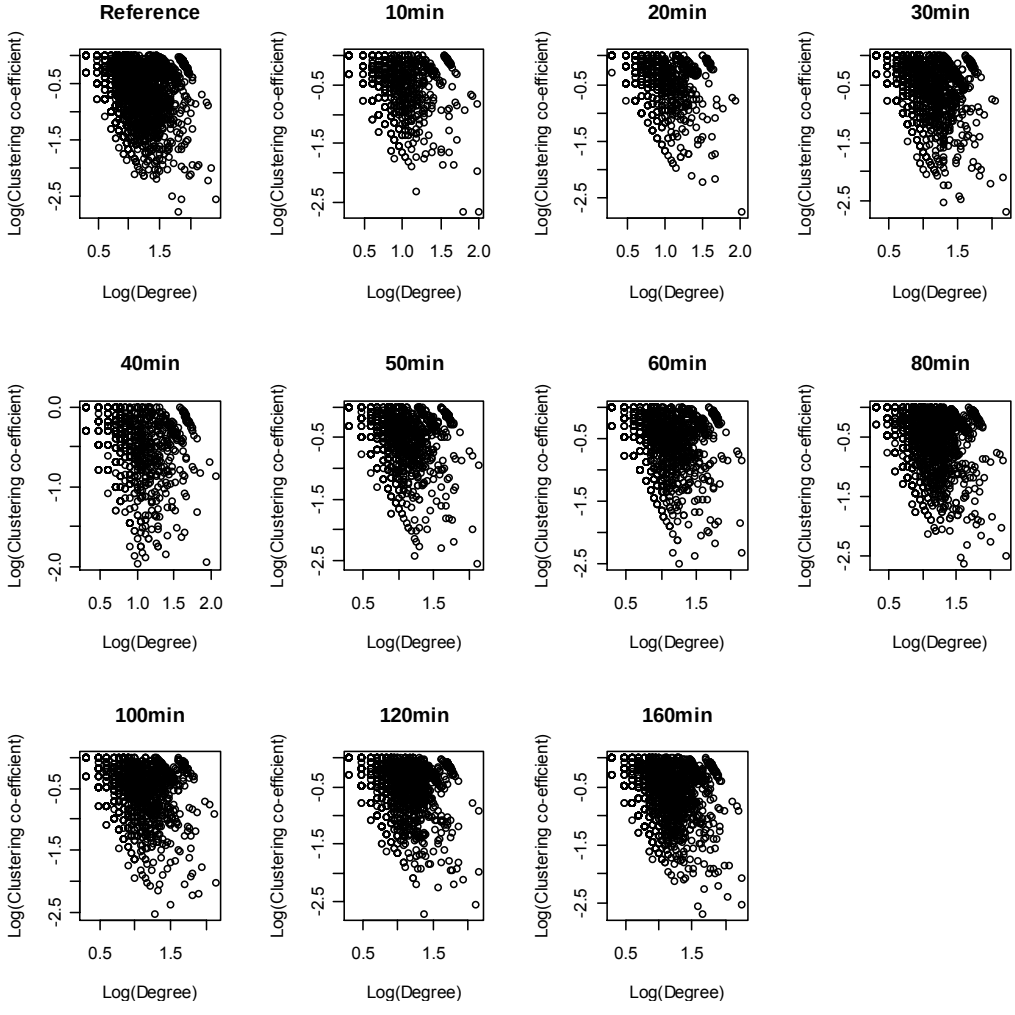


Figure 4.3. Clustering coefficient distributions for reference and networks at time points.

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.

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

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)