Managing biological data in pathways and networks

Managing Biological Data in Pathways and Networks

The research presented in this dissertation was conducted at NUTRIM School for Nutrition, Toxicology and Metabolism of Maastricht University which participates in the Graduate School VLAG (Food Technology, Agrobiotechnology, Nutrition and Health Sciences), accredited by the Royal Netherlands Academy of Arts and Sciences.

This work was (co)financed by the Netherlands Consortium for Systems Biology (NCSB) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research (http://www.ncsb.nl).

Cover design:	Proefschriftmaken.nl || Uitgeverij BOXPress
Layout by:	Martina Summer-Kutmon
Printed by:	Proefschriftmaken.nl || Uitgeverij BOXPress
Published by:	Uitgeverij BOXPress, 's-Hertogenbosch

ISBN: 978-94-6295-044-3

Managing Biological Data in Pathways and Networks

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus, Prof. L.L.G. Soete
in accordance with the decision of the Board of Deans,
to be defended in public on
Thursday 22 January 2015, at 10:00 hrs.

by

Martina Maria Summer-Kutmon

Supervisor:
Prof. Dr. Chris T.A. Evelo

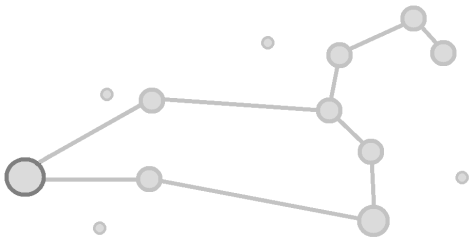
Co-supervisor:
Dr. Susan L.M. Coort

Assessment Committee:
Prof. Dr. Thomas Unger (chairman)
Prof. Dr. Ilja C.W. Arts
Prof. Dr. Jan de Boer
Prof. Dr. Bruce R. Conklin (Gladstone Institutes and UCSF, USA)
Prof. Dr. Alfonso Valencia (Spanish National Cancer Research Center, Spain)

Contents

Chapter 1

General Introduction



Studying the Complexity in Biology

In 1958 Francis Crick first presented the central dogma of molecular biology to describe the flow of sequential information within a biological system [1, 2]. It states that information passes from DNA (deoxyribonucleic acid) to proteins via RNA (ribonucleic acid), but proteins cannot pass the information back to RNA or DNA. Although the principles of the central dogma are still valid today, reality is much more complex (see Figure 1.1). Every step in this sequential information transfer is highly regulated by many different players and it is important to study the complete system in all its complexity to gain better insights into the mechanisms of living systems.

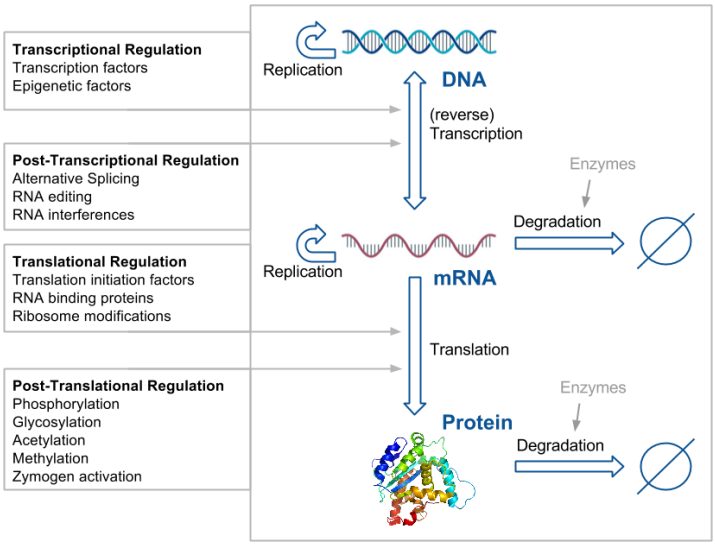


Figure 1.1: Regulation of the Central Dogma of Molecular Biology. DNA stores our genetic information and is replicated during every cell division. The process of transcribing DNA into mRNA is called transcription. This process is sometimes reversible. RNA can be replicated as well. The step from mRNA to protein is called translation and it is not reversible. Each step in the dogma is highly regulated by many different factors (shown on left side).

The central dogma of molecular biology is the foundation of biological complexity. Genes encoded in DNA, messenger RNA (mRNA) and proteins together with small molecules, often metabolites, are the functional elements in a biological system. Together they participate in metabolic and signalling pathways which are the building blocks of the large complex networks describing the logic of a biological system. Pathway diagrams have been used by researchers for many years to visually describe biological processes. With the advances in high-throughput measurement as well as computational technologies we can now look beyond single biological processes and start investigating interactions on a system-wide level. Figure 1.2 shows the complexity pyramid illustrating the different levels in biological systems [3].

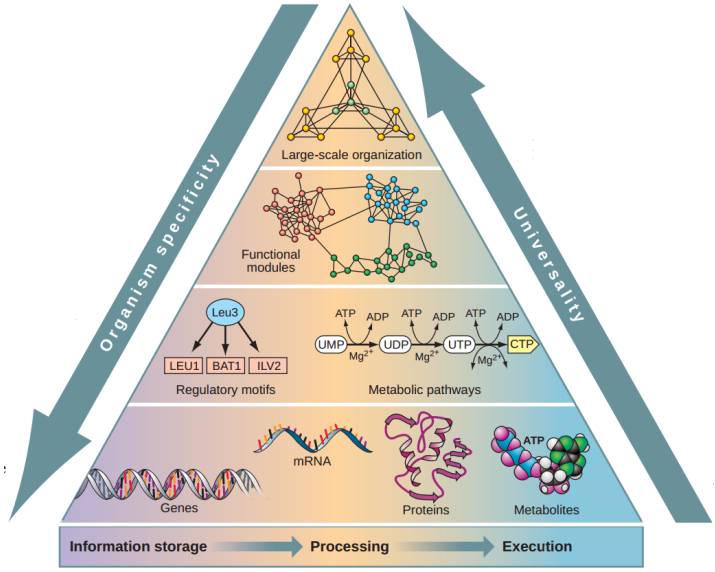


Figure 1.2: Life's Complexity Pyramid by Oltvai and Barabasi [3]. The central dogma of molecular biology represents the foundation of the complexity pyramid. Genes store the genetic information, they are processed into mRNA and further into proteins which together with metabolites execute functions in the cell. However the next layer shows that to be able to perform all their tasks it is important that they work together in metabolic or signalling pathways. The complexity is increased when we study how those different processes are interlinked and connected in larger networks.

In summary, the machinery in a living cell is very complex and the wish to better understand, model and in the end simulate a complex system like a cell gave rise to a flourishing new research field called **Systems Biology**. The goal of this emerging field is the study of interactions between the components in a system and their influence on function and behaviour of that system. Pathway models were the first use case to study specific processes in more detail. They describe the sequence of events in a visual diagram and they have become immensely useful for computational analysis. While pathways are very focused and zoomed in on one specific process, networks tend to look at the larger, holistic view of a biological system.
Biological pathways and networks are the central concepts in this thesis. In this introductory chapter, I will discuss the definition and importance of biological pathways and networks in biomedical research and how they are used to store, integrate, visualize, analyze and interpret biological data.

Biological Pathways

For many years biologists have been drawing pathway diagrams to gain a better understanding of the processes in a living cell. The diagrams are found everywhere: in textbooks, research articles, posters, lab journals or presentations and they have proven themselves as powerful tools to organize, share and discuss knowledge.

A pathway represents the current knowledge about a biological process in a visual, comprehensive and easy to understand format. Pathway diagrams drawn with state-of-the art pathway editors are much more than just images. Each biomolecule and interaction in the pathway is linked to external online databases containing additional detailed information. This also enables the automatic integration and visualization of experimental data on the pathway. Additionally every element in the pathway can be linked to scientific literature creating a comprehensive reference collection for a specific biological process.

In most cases pathway diagrams are drawn with hardly any standardization. The same graphical symbols might be used to describe different elements or relationships in the pathways which leads to confusion and misinterpretation. Although pathways are network-like in nature, there are specific expectations on how the elements in such a network should be represented. Graphical notation standards, like SBGN (Systems Biology Graphical Notation [4]) or MIM (Molecular Interaction Maps [5]), have been developed and are slowly adopted by the different online pathway databases, like WikiPathways [6], Reactome [7] and KEGG [8].

The most common types of biological pathways are metabolic and signalling pathways, which will be described in more detail below.

Metabolic Pathways

Metabolic pathways describe the biochemical reactions that are needed to build up new molecules (biosynthesis) or break down molecules (degradation). Many molecular transformations are performed in a multi-step process which are then combined in one metabolic pathway. Each step is regulated by proteins called enzymes that speed up the biochemical reactions or even invest energy to make them go in the reverse direction from normal. Glycolysis is the first metabolic pathway discovered and represents a ten-step conversion of one glucose molecule into two pyruvate molecules (see Figure 1.3). In response to changes in the environment the enzymatic activity in the cell can be controlled to balance the level of metabolites in the cell. This is important for cellular maintenance and cell survival.

Signalling Pathway

Receptors are proteins located inside a cell or on the cell surface. Their task is to receive chemical signals from outside the cell. In response to such an extracellular signal the receptor initiates an intracellular signal transduction pathway. Multiple pathways might be active and intersecting with each other at any time point. The signal can be amplified in every step of the transduction cascade, enzymes might be activated or inhibited to regulate metabolic pathways or gene transcription might be influenced by the presence or absence of transcription factors (TFs). As an example, Figure 1.4 shows the Wnt signalling pathway from KEGG.

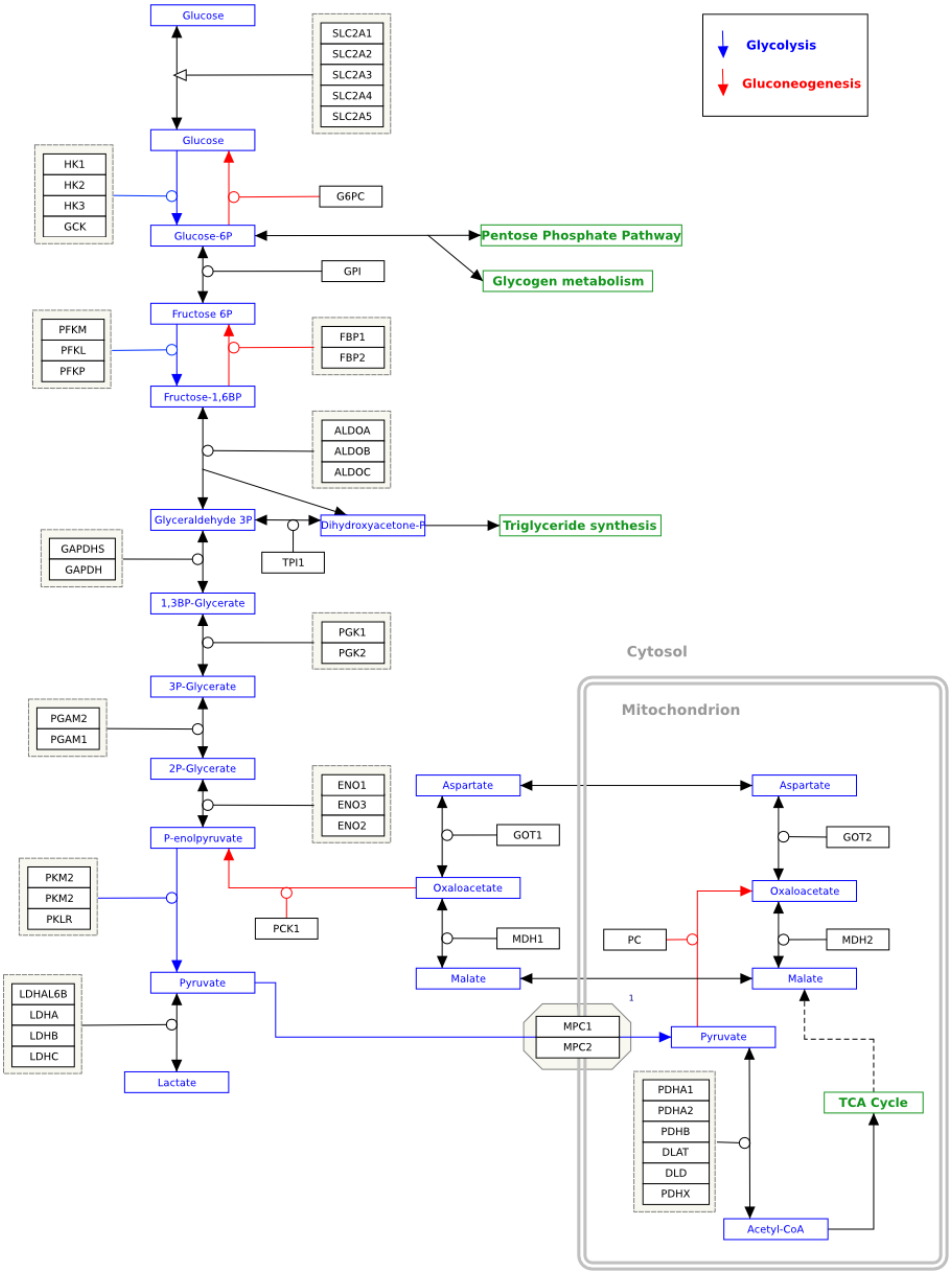


Figure 1.3: Glycolysis Pathway. This pathway from WikiPathways (http://www.wikipathways.org/instance/WP534) describes the 10-step breakdown of glucose ($C_6H_{12}O_6$) into two pyruvate (CH_3COCOO^-) molecules. This pathway contains metabolites (blue boxes) and enzymes (black boxes) as well as links to other pathways (green boxes).

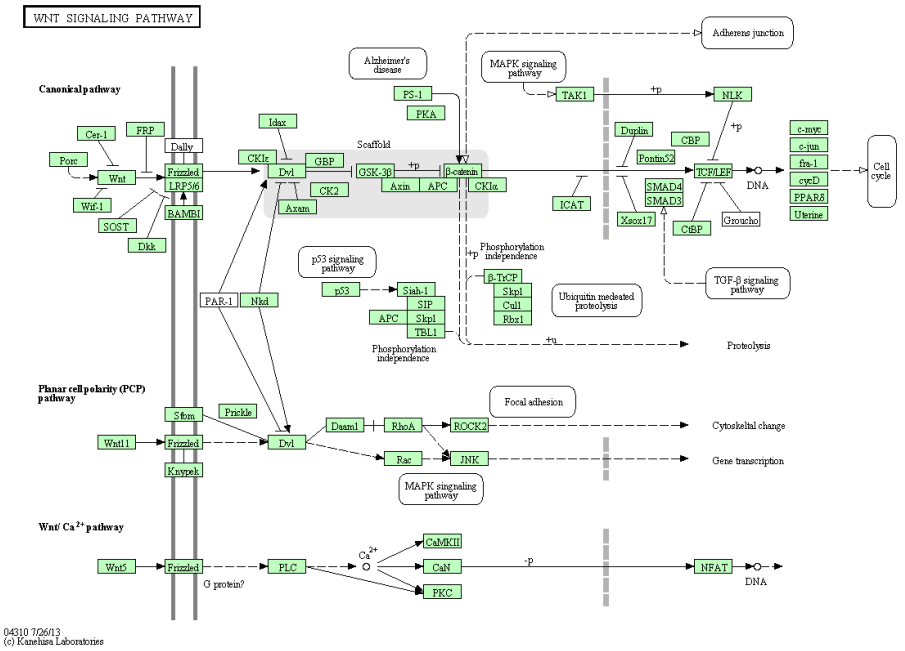


Figure 1.4: Wnt Signalling Pathway. This pathway from KEGG (http://www.genome.jp/kegg/pathway/hsa/hsa04310.html) describes the effects of the binding of Wnt proteins to their receptors. Wnt proteins are secreted lipid-modified glycoproteins.

Biological Networks

Before being applied in biology, networks have been used in many different areas. Transportation systems, social connections, the power grid systems and even the internet are represented as networks. Those are all large systems with hundreds of thousands of interacting components. The elements in a network are called "nodes" and the links between them "edges". As an example, in a protein-protein interaction network the nodes represent proteins and the edges between them show known physical interactions between two proteins. In the past, we were not able to handle the large amount of data behind such networks but with the emergence of the internet and the increase in computational power, it became possible to collect, assemble, share and analyze such large networks.

Network biology can build on several hundred years of experience and developments in graph theory, a sub-field of mathematics that focused on networks since 1736 [9]. Therefore most algorithms and approaches used in network biology are based on previously defined network properties like the shortest path between two nodes, the node degree to find hub nodes in the network or node betweenness to calculate the importance of the node for the connectivity of the network.

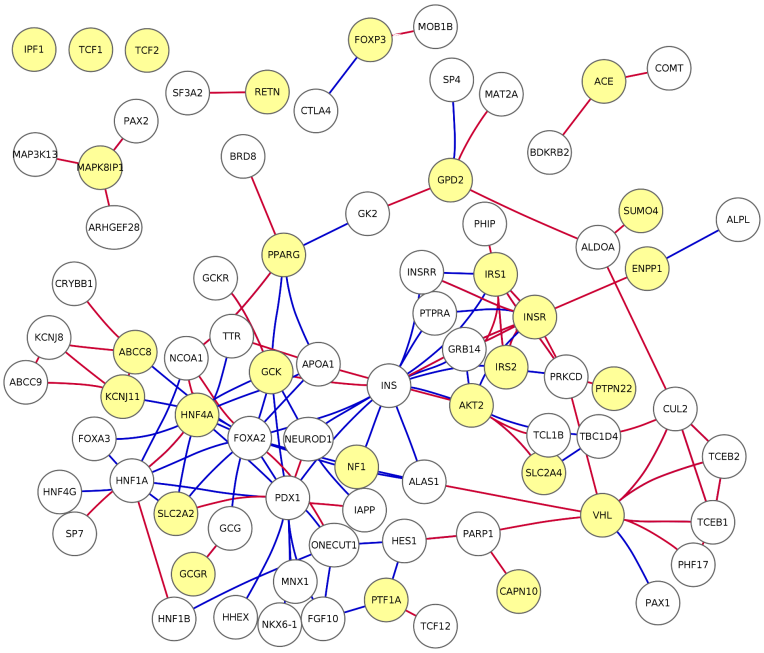


Figure 1.5: Network of 27 *Diabetes Mellitus* Related Genes. The network was created with GeneMania [10] using the 27 genes (yellow nodes) associated with *diabetes mellitus* in the Diseasome [11]. 50 related genes were added to the network (white nodes). The interactions represent physical interactions (protein-protein interactions; red edges) and pathway interactions (blue edges). No interactions were found for IPF1, TCF1 and TCF2.

In network biology we distinguish a number of different network types, for example:

Protein-protein interaction networks consist of known intentional physical contacts between pairs of proteins. The latest updates from the Human Interactome Project [12] reported a total of 17,000 unique binary interactions (HI-II-14, prepublication, http://interactome.dfci.harvard.edu/).

Regulatory networks are collections of elements that regulate the expression level of mRNA or proteins. The main players in regulatory networks are TFs that regulate the transcription of a gene into mRNA. In recent years, post-transcriptional regulators like microRNAs (miRNAs) or small interfering RNAs (siRNAs) are often also included in regulatory networks.

Metabolic networks are maps of connected metabolic pathways. Recon 2 is a global reconstruction of the known human metabolic network [13]. The current map contains 1,789 enzyme-encoding genes, 7,440 reactions and 2,626 unique metabolites.

Signalling networks show how extracellular signals are transducted in the cells. They consist of multiple signalling pathways but often also integrate regulatory and metabolic networks.

Co-expression networks are constructed based on experimental data. The edges in the network represent the correlation between the transcript abundances of pairs of genes or proteins. These networks are widely used to identify co-expression modules and hub genes.

An example network including protein-protein interactions and pathway information is shown in Figure 1.5. It is apparent that the types of pathways and networks are very similar which highlights that pathways really are the building blocks of larger complex biological systems. A metabolic pathway describes a specific process while a metabolic network links the different pathways to each other. This makes it possible to move beyond single genes, single pathways, single studies, and make full use of all information and knowledge generated thus far, providing an invaluable framework for deciphering molecular mechanisms of health and disease in their entire complexity.

Managing Biological Data

In recent years biomedical research has been experiencing a rapid growth in volume and heterogeneity of biological data. This presents an increasing challenge for biologists and bioinformaticians. Biological data covers experimental data (measurements) but also structured knowledge that has been inferred from experimental data and is usually published in literature and online databases. In the systems biology approaches described in this thesis, the goal is often to integrate existing knowledge with experimental data to verify known mechanisms, generate new hypotheses or study the molecular effects of a disease or treatment.

In this thesis we are looking at four key aspects about "*managing biological data*" from a systems biology point of view: (1) data structure and storage, (2) data integration, (3) data visualization and (4) data analysis and interpretation.

Data Structure and Storage

Experimental data. Nowadays experimental data can be made publicly available through one of the online data repositories like ArrayExpress [14] or Gene Expression Omnibus (GEO) [15] which contain nearly 50,000 experiments. Although the focus still lies on large scale functional genomics experiments, the number of available proteomics and metabolomics experiments are increasing. The European Bioinformatics Institute (EBI) also maintains repositories for proteomics (PRIDE [16]) and metabolomics (MetaboLights [17]) experiments.

Biological knowledge. There are many online databases structuring and storing information about single biological entities, like Ensembl [18] for genes and transcripts, UniProt [19] for proteins or ChEMBL [20] for metabolites. In systems biology the focus shifts from looking at one single entity to the relationships and interplay between those entities. Biological relationships are very diverse and are often stored as interactions between two entities in interaction databases. They can encompass physical protein-protein interactions (IntAct [21] or STRING [22]),

drug-target interactions (DrugBank [23] or ChEMBL [20]) or even disease-gene associations (DisGeNet [24] or OMIM [25]). Pathway databases on the other hand store pathways that contain a connected set of interactions relevant in a biological process. Commonly used pathway databases are WikiPathways [6], Reactome [7] and KEGG [8]. There is a wide variety of biological online database and the list of databases mentioned in this section is not exhaustive.

Data Integration

A crucial aspect in bioinformatics is the integration of existing knowledge and the large amount of experimental data. Biological data is diverse, complex and distributed in many different resources. Data integration allows researchers to make better informed and faster decisions about their research and enables them to also include the areas surrounding their experiments to see the bigger picture [26].

Several chapters in this thesis demonstrate and discuss the integration of data. We differ between three major types of data integration:

Integration of data from multiple online databases. Because of the large variety of biological databases it is often necessary and advantageous to combine and integrate data from multiple databases in an analysis. As an example, there are several databases storing miRNA-target gene interactions. Some of the databases contain validated interactions, like miRTarBase [27] and miRecords [28], and others provide predicted interactions, like microCosm Targets or TargetScan [29]. In practise researchers often integrate data from several of these databases or in the case of predicted interactions they only consider interactions predicted by multiple algorithms.

Integration of multiple experimental datasets. The large amount of available experimental data allows bioinformatics to integrate and combine experiments studying the same or similar conditions. If consensus between the different datasets is shown the confidence in the results is increased. The integration of different experimental datasets is also relevant when comparing different settings, like different tissues, cell types, disease states. Furthermore, the integration of different types of experimental data is crucial to investigate the biological system as a whole, for example the combination of transcriptomics and proteomics data to study the correlation between gene expression and protein abundance levels.

Integration of experimental data with biological knowledge. When analyzing experimental data the integration of existing knowledge is crucial to make sense of the data. GeneOntology (GO [30]) annotations help to investigate the functions of the genes affected by a toxin or pathway analysis allows researchers to combine expression data with pathway information to analyze changes in metabolic and signalling pathways in a disease.

Data Visualization

Nowadays biological data is almost exclusively visualized with computer-based visualization tools and the advances in computer hardware and network access make the wide range of visualization tools amenable to non-experts. However, because of the complexity and heterogeneity in biology, the visualization of biological data is still one of the biggest challenges. Friedman [31] summarizes the importance and role of data visualization:

> "*The main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesnt mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way.*"

Modern methods of data generation and the integration of many different data types make it harder to visualize data in a concise and meaningful way. Pathways are a very useful tool to reduce the complexity and visualize the data for a specific biological process. In network biology it is often necessary to apply algorithms to find the relevant parts in a hairball network and highlight those parts visually.

Data Analysis and Interpretation

The emergence of high-throughput technologies brought incredible possibilities for new discoveries but the analysis and interpretation of data became much more difficult and time-consuming. The focus in this thesis lies on the analysis and interpretation of experimental data using pathway and network analysis.

Pathway analysis groups genes, proteins and other biological molecules based on their involvement in biological pathways and therefore reduces the size of the problem. Instead of looking at thousands of genes, pathway analysis explores hundreds of pathways. As mentioned before, biological pathways are models of well studied biological processes and therefore the result of a pathway analysis has a higher explanatory power than a simple gene list.

Outline of this Thesis

The aim of the work described in this thesis is to show the power of biological pathways and networks to store, integrate, analyze, visualize and interpret biological data.

The first half of the thesis will focus on the applicability of biological pathways. **Chapter 2** introduces PathVisio, a biological pathway editor, analysis and visualization software which is widely adopted in the research community. Its third version provides a new extensible toolbox for pathway creation, data visualization and pathway analysis. **Chapter 3** acts as a proof-of-principle study, showing

the capability of pathway models to collect biological knowledge about a biological process and present it in an intuitive, visual way. In this study, we investigated the function and regulation of sterol regulatory element-binding proteins (SREBPs) by reviewing more than 50 scientific articles and integrating the information in one biological pathway. This pathway was created in PathVisio and published in the pathway database WikiPathways. **Chapter 4** demonstrates the visualization of multi-omics data on biological pathways based on a published mouse study. Transcriptomics data was combined with proteomics data to show transcript and protein levels together on a pathway diagram. In like manner, metabolomics or other biological data, numeric and nonnumeric, can be visualized on pathways.

The second half of the thesis will move from the focused, smaller pathway models to larger, more complex biological networks. **Chapter 5** will not only show the bridge between pathways and networks but also the link between two large open source communities, WikiPathways and Cytoscape [32]. Cytoscape is a popular network visualization and analysis software which can be extended by so-called apps. The WikiPathways app in Cytoscape allows users to open biological pathways as networks in Cytoscape to then perform advanced network analysis. Another Cytoscape app, CyTargetLinker, will be presented in **Chapter 6**. The integration of regulatory interactions like TF-gene, miRNA-target or drug-target interactions is crucial to study biological systems in their entire complexity. CyTargetLinker not only provides an easy-to-use interface to integrate such interactions but also allows to combination of different resources together. The final publication in this thesis, **Chapter 7**, will demonstrate the combination and application of the in previous chapters described systems biology approaches studying the biological rewiring in the diabetic liver. This study will combine pathway and network analysis in a real biological use case and emphasize the usability and immense potential of systems biology approaches to better understand disease mechanisms.

Finally, it will end with a **General Discussion** about the importance of data curation of pathway and interaction data, the challenges of data integration and the role of open data, open access and open source in biomedical research.

Bibliography

[1] F Crick. Ideas on Protein Synthesis, October 1956.

[2] F Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, August 1970.

[3] ZN Oltvai and AL Barabási. Systems biology. Life's complexity pyramid. *Science (New York, N.Y.)*, 298(5594):763–4, October 2002.

[4] N Le Novère, M Hucka, H Mi, S Moodie, F Schreiber, A Sorokin, E Demir, K Wegner, MI Aladjem, SM Wimalaratne, FT Bergman, R Gauges, P Ghazal, H Kawaji, L Li, Y Matsuoka, A Villéger, SE Boyd, L Calzone, M Courtot, U Dogrusoz, TC Freeman, A Funahashi, S Ghosh, A Jouraku, S Kim, F Kolpakov, A Luna, S Sahle, E Schmidt, S Watterson, G Wu, I Goryanin, DB Kell, C Sander, H Sauro, JL Snoep, K Kohn, and H Kitano. The Systems Biology Graphical Notation. *Nature biotechnology*, 27(8):735–41, August 2009.

[5] KW Kohn, MI Aladjem, JN Weinstein, and Y Pommier. Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Molecular biology of the cell*, 17(1):1–13, January 2006.

[6] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[7] D Croft, AF Mundo, R Haw, M Milacic, J Weiser, G Wu, M Caudy, P Garapati, M Gillespie, MR Kamdar, B Jassal, S Jupe, L Matthews, B May, S Palatnik, K Rothfels, V Shamovsky, H Song, M Williams, W Birney, H Hermjakob, L Stein, and P D'Eustachio. The Reactome pathway knowledgebase. *Nucleic acids research*, 42(Database issue):D472–7, January 2014.

[8] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000.

[9] N Biggs, EK Lloyd, and RJ Wilson. Graph Theory, 1736-1936. October 1986.

[10] J Montojo, K Zuberi, H Rodriguez, F Kazi, G Wright, SL Donaldson, Q Morris, and GD Bader. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics (Oxford, England)*, 26(22):2927–8, November 2010.

[11] KI Goh, ME Cusick, D Valle, B Childs, M Vidal, and AL Barabási. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–90, May 2007.

[12] J-F Rual, K Venkatesan, T Hao, T Hirozane-Kishikawa, A Dricot, N Li, GF Berriz, FD Gibbons, M Dreze, N Ayivi-Guedehoussou, N Klitgord, C Simon, M Boxem, S Milstein, J Rosenberg, DS Goldberg, LV Zhang, SL Wong, G Franklin, S Li, JS Albala, J Lim, C Fraughton, E Llamosas, S Cevik, C Bex, P Lamesch, RS Sikorski, J Vandenhaute, HY Zoghbi, A Smolyar, S Bosak, R Sequerra, L Doucette-Stamm, ME Cusick, DE Hill, FP Roth, and M Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8, October 2005.

[13] I Thiele, N Swainston, RMT Fleming, A Hoppe, S Sahoo, MK Aurich, H Haraldsdottir, ML Mo, O Rolfsson, MD Stobbe, SG Thorleifsson, R Agren, C Bölling, S Bordel, AK Chavali, P Dobson, WB Dunn, L Endler, D Hala, M Hucka, D Hull, D Jameson, N Jamshidi, JJ Jonsson, N Juty, S Keating, I Nookaew, N Le Novère, N Malys, A Mazein, JA Papin, ND Price, E Selkov, MI Sigurdsson, E Simeonidis, N Sonnenschein, K Smallbone, A Sorokin, JHGM van Beek, D Weichart, I Goryanin, J Nielsen, HV Westerhoff, DB Kell, P Mendes, and B ØPalsson. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–25, May 2013.

[14] G Rustici, N Kolesnikov, M Brandizi, T Burdett, M Dylag, I Emam, A Farne, E Hastings, J Ison, M Keays, N Kurbatova, J Malone, R Mani, A Mupo, R Pedro Pereira, E Pilicheva, J Rung, A Sharma, YA Tang, T Ternent, A Tikhonov, D Welter, E Williams, A Brazma, H Parkinson, and U Sarkans. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic acids research*, 41(Database issue):D987–90, January 2013.

[15] T Barrett, SE Wilhite, P Ledoux, C Evangelista, IF Kim, M Tomashevsky, KA Marshall, KH Phillippy, PM Sherman, M Holko, A Yefanov, H Lee, N Zhang, CL Robertson, N Serova, S Davis, and A Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5, January 2013.

[16] JA Vizcaíno, RG Côté, A Csordas, JA Dianes, A Fabregat, JM Foster, J Griss, E Alpi, M Birim, J Contell, G O'Kelly, A Schoenegger, D Ovelleiro, Y Pérez-Riverol, F Reisinger, D Ríos, R Wang, and H Hermjakob. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research*, 41(Database issue):D1063–9, January 2013.

[17] K Haug, RM Salek, P Conesa, J Hastings, P de Matos, M Rijnbeek, T Mahendraker, M Williams, S Neumann, P Rocca-Serra, E Maguire, A González-Beltrán, S-A Sansone, JL Griffin, and C Steinbeck. MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(Database issue):D781–6, January 2013.

[18] P Flicek, I Ahmed, MR Amode, D Barrell, K Beal, S Brent, D Carvalho-Silva, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gil, C García-Girón, L Gordon, T Hourlier, S Hunt, T Juettemann, AK Kähäri, S Keenan, M Komorowska, E Kulesha, I Longden, T Maurel, WM McLaren, M Muffato, R Nag, B Overduin, M Pignatelli, B Pritchard, E Pritchard, HS Riat, GRS Ritchie, M Ruffier, M Schuster, D Sheppard, D Sobral, K Taylor, A Thormann, S Trevanion, S White, SP Wilder, BL Aken, E Birney, F Cunningham, I Dunham, J Harrow, J Herrero, TJP Hubbard, N Johnson, R Kinsella, A Parker, G Spudich, A Yates, A Zadissa, and SMJ Searle. Ensembl 2013. *Nucleic acids research*, 41(Database issue):D48–55, January 2013.

[19] Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, 40(Database issue):D71–5, January 2012.
[20] AP Bento, A Gaulton, A Hersey, LJ Bellis, J Chambers, M Davies, FA Krüger, Y Light, L Mak, S McGlinchey, M Nowotka, G Papadatos, R Santos, and JP Overington. The ChEMBL bioactivity database: an update. *Nucleic acids research*, 42(Database issue):D1083–90, January 2014.
[21] S Orchard, M Ammari, B Aranda, L Breuza, L Briganti, F Broackes-Carter, NH Campbell, G Chavali, C Chen, N Del-Toro, M Duesbury, M Dumousseau, E Galeota, U Hinz, M Iannuccelli, S Jagannathan, R Jimenez, J Khadake, A Lagreid, L Licata, RC Lovering, B Meldal, AN Melidoni, M Milagros, D Peluso, L Perfetto, P Porras, A Raghunath, S Ricard-Blum, B Roechert, A Stutz, M Tognolli, K van Roey, G Cesareni, and H Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue):D358–63, January 2014.
[22] A Franceschini, D Szklarczyk, S Frankild, M Kuhn, M Simonovic, A Roth, J Lin, P Minguez, P Bork, C von Mering, and LJ Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue):D808–15, January 2013.
[23] C Knox, V Law, T Jewison, P Liu, S Ly, A Frolkis, A Pon, K Banco, C Mak, V Neveu, Y Djoumbou, R Eisner, AC Guo, and DS Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(Database issue):D1035–41, January 2011.
[24] A Bauer-Mehren, M Bundschus, M Rautschka, MA Mayer, F Sanz, and LI Furlong. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PloS one*, 6(6):e20284, January 2011.
[25] A Hamosh, AF Scott, JS Amberger, CA Bocchini, and VA McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(Database issue):D514–7, January 2005.
[26] M Schneider and RC Jimenez. Teaching the fundamentals of biological data integration using classroom games. *PLoS computational biology*, 8(12):e1002789, January 2012.
[27] S-D Hsu, Y-T Tseng, S Shrestha, Y-L Lin, A Khaleel, C-H Chou, C-F Chu, H-Y Huang, C-M Lin, S-Y Ho, T-Y Jian, F-M Lin, T-H Chang, S-L Weng, K-W Liao, I-E Liao, C-C Liu, and H-D Huang. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic acids research*, 42(Database issue):D78–85, January 2014.
[28] F Xiao, Z Zuo, G Cai, S Kang, X Gao, and T Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research*, 37(Database issue):D105–10, January 2009.
[29] BP Lewis, CB Burge, and DP Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, January 2005.
[30] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, May 2000.
[31] V Friedman. Data visualization and infographics. *Graphics, Monday Inspiration*, 14, 2008.
[32] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504, November 2003.

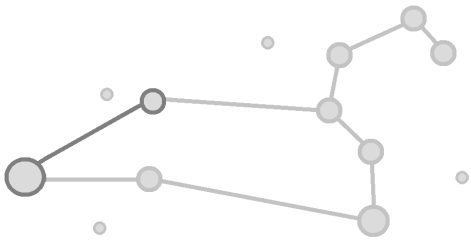
Chapter 2

PathVisio 3: An Extendable Pathway Analysis Toolbox

Martina Kutmon[1,2], Martijn P van Iersel[3], Anwesha Bohler[1], Thomas Kelder[4], Nuno Nunes[1], Alexander R Pico[5], Chris T Evelo[1]

1. Department of Bioinformatics - BiGCaT, NUTRIM School for Nutrition, Toxicology and Metabolism, Maastricht University, The Netherlands
2. Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, The Netherlands
3. General Bioinformatics, Reading, United Kingdom
4. EdgeLeap B.V., Utrecht, The Netherlands
5. Gladstone Institutes, San Francisco, CA, USA

Manuscript submitted to *PLoS Computational Biology*



Abstract

For many years biologists have been drawing pathway diagrams to gain a better understanding of the underlying biology. The diagrams are found everywhere: in textbooks, research articles, posters, lab journals or presentations and they have proven themselves as powerful tools to organize, share and discuss knowledge. In 2008, we presented the first version of our pathway visualization and analysis tool PathVisio. Since then, PathVisio has been used extensively in different studies to draw biological pathways, perform pathway statistics or visualize biological data on pathways. A light-weight applet version of PathVisio is integrated in the community curated pathway database WikiPathways to provide an online pathway editor that allows users to directly edit and curate the pathways.

In the last six years, PathVisio has been substantially extended and the core application was refactored using the OSGi framework to achieve a better, modular system that can be easily extended with so called plugins. The new plugin repository and manager bring the functionality of the plugins to all users by offering a simple and user-friendly interface for plugin installation.

PathVisio is a freely available, open-source pathway editor, visualization and analysis software that runs on all major operating systems. The focus points for this new, third version of PathVisio are modularity, extensibility and improved usability. PathVisio 3 introduces a wide variety of new features including support for different pathway drawing standards, advanced multi-omics data visualization, statistical methods, support for import and export of different file formats and the integration of data from online databases. More than 3,000 downloads between January and June 2014 show that PathVisio is widely adopted in the research community.

Introduction

A picture says more than a thousand words. For many years biologists have been drawing pathway diagrams to gain a better understanding of the underlying biology. These diagrams are found everywhere: in textbooks, research articles, posters, lab journals or presentations and they have proven themselves as powerful tools to organize, share and discuss knowledge. Pathway diagrams have also become immensely useful for computational analysis and interpretation of large-scale experimental data when properly modelled. Complex diseases like cancer or heart failure are known to be caused by malfunctioning pathways instead of individual genes, so the study and collection of biological pathways is crucial to get insights into complicated disease mechanisms. Nowadays, computers allow researchers to use tools to draw pathway diagrams that are much more than just pictures; they contain annotations, literature references and comments for each element and interaction in a pathway. These enriched pathway diagrams open the possibilities to perform advanced pathway analysis and data visualization to get a more comprehensive understanding of experimental data.

In 2008, we presented the first version of our pathway visualization and analysis tool PathVisio [1]. Since then, PathVisio has been used in numerous studies to draw biological pathways, perform pathway statistics or visualize biological data on pathways [2–10].
PathVisio has undergone active development and grown beyond a simple tool into a comprehensive and extendable pathway analysis toolbox. Besides its standalone graphical desktop version, PathVisio is often used as a library to read, write, store, convert and model pathway information. It is also used in different websites and workflows to act as a pathway editor and visualization tool. For example, a light-weight applet version of PathVisio is integrated in the community curated pathway database WikiPathways [11] and ProfileDb, a resource for proteomics and cross-omics biomarker discovery, uses PathVisio to visualize differential expression results on pathway diagrams [12].
In previous versions PathVisio provided a simple but limited interface for extensions through plugins. A plugin is a small software component that adds a specific feature to an existing application. In the case of PathVisio, a plugin could provide for example a new statistical method, a new drawing standard or additional information about elements in the pathway. The usage of available plugins enables users to refine the pathway analysis workflow in PathVisio and build an application with all the necessary modules relevant for their research.

Here we introduce the third version of the pathway visualization and analysis tool PathVisio. The aim is to present the newest additions and improvements of the application, especially the new plugin extension system, as well as the plugin repository and the integrated plugin manager. PathVisio is a freely available, open-source tool allowing independent developers to contribute plugins to provide new functionality. PathVisio is implemented in Java and thus runs on all major operating systems. The focus of this new version of PathVisio lies on modularity, extensibility and improved usability.

Design and Implementation

In the last six years, PathVisio has been substantially extended and the core application was refactored using the OSGi framework (Open Service Gateway initiative) to achieve a better, modular system that can be easily extended with so called plugins [13]. OSGi also allows plugins to depend on each other to avoid code redundancy and promote code reusability. Such modular systems keep the core of an application stable and maintainable while the functionality can be easily extended allowing users to build an application designed for their needs [14].

First, we will discuss the new modular structure of PathVisio 3, then the plugin repository will be introduced, and last the usability and advantages of the new plugin manager will be shown.

Modularisation with OSGi

PathVisio 3 consists of eight OSGi modules that build the core application, each being responsible for one crucial part of the application. As illustrated in Figure 2.1, the modules nicely separate the different parts of the application.

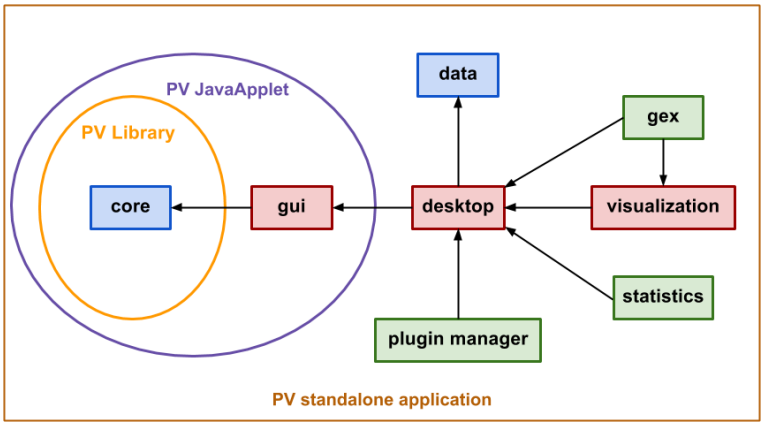


Figure 2.1: Transitive Dependency Structure of PathVisio 3. The application consists of eight modules each providing specific functionality. The modules *core* and *data* are independent modules (coloured in blue) that function as libraries that can be reused outside of PathVisio (PV). Especially the core module is often used as a PV library for reading and writing of pathway files. Other modules in red, *gui*, *desktop* and *visualization*, provide functionality that is used by other modules. Green modules, *gex*, *statistics* and *plugin manager*, are not used by other PV modules but can be used by PV plugins. The PV JavaApplet version integrated in WikiPathways uses the *core* and *gui* module.

The *core* module of PathVisio 3 contains the non-user interface backend, including the data model, import and export functionality and general settings and preferences. This module can also be used as a library by other software tools for reading, editing and writing pathway files in PathVisio's native GPML (Graphical Pathway Markup Language, http://www.pathvisio.org/gpml) format. The *gui* (graphical user interface) module implements the basic user interface which is shared between the standalone and the applet version of PathVisio. The applet version is

integrated in WikiPathways as an online pathway editor. The more advanced, full-powered graphical user interface for the standalone application is provided by the *desktop* module. It is also the central connecting point for plugins. The *plugin manager* module handles the connection to the plugin repository as well as installing and uninstalling plugins. The *gex* module contributes the functionality for importing experimental data together with the *data* module which defines the interfaces for storing and handling experimental data. The *visualization* module then provides a simple but flexible way to visualize the experimental data on the data nodes in the pathways. To identify significantly altered pathways in an experimental dataset, the *statistics* module contributes a standard over-representation analysis algorithm based on a hypergeometric test [15].

PathVisio Plugin Repository

The new PathVisio plugin repository consists of two separate parts, (i) the repository itself which stores all necessary plugin files as well as their dependencies and (ii) the PathVisio plugin database and front-end.
The PathVisio repository is located at http://repository.pathvisio.org. It contains all plugin files and third-party dependencies. The RepoIndex library (https://github.com/osgi/bindex) builds a complete dependency structure of the repository and writes it in an XML file named repository.xml.

The PathVisio plugin database is an independent mySQL database containing location information and metadata, e.g. description, authors and release notes, about each plugin. The database is integrated into the WordPress framework (http://wordpress.org) to take advantage of some of the built in functionalities of WordPress, like capabilities to tag, browse, search, comment and evaluate plugins.

Plugin Manager

To make it easier for users to find and install plugins, PathVisio 3 incorporates a plugin manager that connects to the repository and enables a one-click installation of plugins from within the application. The plugin manager allows users to browse plugins by categories and provides additional information about the plugin when selected, like description or author information.

Figure 2.2 shows the connections between the different components that are used by the plugin manager. This new plugin manager module retrieves data from two different files, the repository.xml file and the pathvisio.xml file. The repository.xml file is created by the RepoIndex library and stores the complete dependency structure of the repository. Additional metadata about the plugin, like developers, description or categories, are retrieved from the pathvisio.xml file which is created from the PathVisio plugin database.

Consequently, the new extension system takes care of the installation of plugins and all required dependencies. If a plugin depends on another plugin or a third party library, the plugin manager makes sure that all required OSGi bundles are

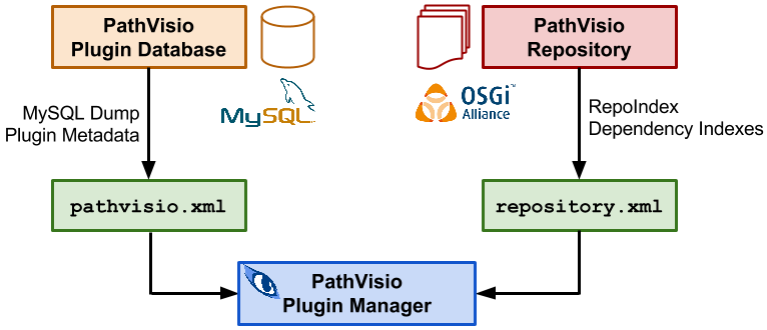


Figure 2.2: Plugin Extension and Installation System of PathVisio 3. The plugin repository stores all plugin files and their dependencies. The RepoIndex library is used to create a repository.xml file which contains the dependency indexes of all plugins. Metadata about plugins is stored in the PathVisio plugin database which is then exported into a pathvisio.xml file. The PathVisio 3 plugin manager retrieves data from both files to facilitate the installation of plugins in PathVisio 3.

downloaded, installed and started. Therefore the complex dependency structure is hidden from the user and installation is much easier and faster.

Results

PathVisio has been used in a substantial number of publications in the last six years and the analysis workflow has been further developed and improved. PathVisio 3 also provides several interfaces allowing plugins to integrate tightly into the application. The new plugin repository and manager finally bring the functionality of the plugins to all users by offering a simple and user-friendly interface for plugin installation.

In this section, we will first highlight the new features of PathVisio in an updated feature table, then the standard pathway analysis workflow in PathVisio will be demonstrated and lastly show how plugins can hook into the application and provide new functionality to the user.

Feature Table

In Table 2.1 the most important features of PathVisio 3 are summarized.

Table 2.1: PathVisio 3 Feature Table

Feature	Description
File import	Default: GPML (http://www.pathvisio.org/gpml) Plugins: MIMML ([16], MIM plugin), SBGNML ([17], SBGN plugin), SBML ([18], PathSBML), BioPAX ([19], BioPAX plugin), gene list (MAPPBuilder)
File export	Default: GPML, PNG, PDF, SVG, TIFF, Eu.Gene [20], datanode list Plugins: MIMML (MIM plugin), SBGNML (SBGN plugin), SBML (PathSBML), HTML (HTMLexporter), BioPAX (BioPAX plugin)
Pathway drawing standards	Default: Basic GPML style Plugins: SBGN, MIM
Identifier mapping	Integrated BridgeDb framework [21] for advanced identifier mapping for pathway elements and interactions in the pathways. All major database identifiers including probe ids for genes, proteins and metabolites are supported.
Pathway statistics	Default: Over-representation analysis (Z-Score) Plugins: Gene set enrichment analysis (GSEA plugin)
Data visualization	Pathway nodes: gradient-based visualization for numeric data, rule-based visualization for numeric and nonnumeric data Interactions: color and line thickness visualization (IntViz plugin)
Plugin extension system	Plugin manager allows one-click installation of plugins from central plugin repository to enable additional features.
Pathway database connection	WikiPathways: searching, browsing, updating, uploading biological pathways (WikiPathways plugin)
Workflow integration	The core module can be used as a library to read, write, store, convert and model pathway information Calling PathVisio functionality from other programming languages through XML-RPC server (PathVisioRPC)
Online data access	Several plugins provide connections to other online resources to give more information about the individual elements in the pathway, like BiomartConnect about gene products, MetInfo about metabolites or PathwayLoom about known interaction partners.

Pathway Analysis Workflow in PathVisio

The core application has three main features: (1) pathway drawing, (2) data visualization and (3) pathway statistics. The integrated identifier mapping framework BridgeDb [21] allows pathway authors to annotate the elements in their pathways with their identifier system of choice and automatically takes care of the mapping when e.g. experimental data with another identifier system is loaded.

The data visualization and pathway statistics modules have been first introduced in PathVisio 2 and further improved and extended in PathVisio 3.

(1) Pathway Drawing

Biological pathway diagrams represent the sequence of events in biological processes. They often contain different biological entities, like genes, proteins or metabolites, and interactions between them, like conversion, stimulation or inhibition. As illustrated in Figure 2.3, PathVisio is a full pathway editor which allows users to draw the biological events, add graphical elements like shapes or labels and annotate all the biological entities and interactions with external database identifiers.

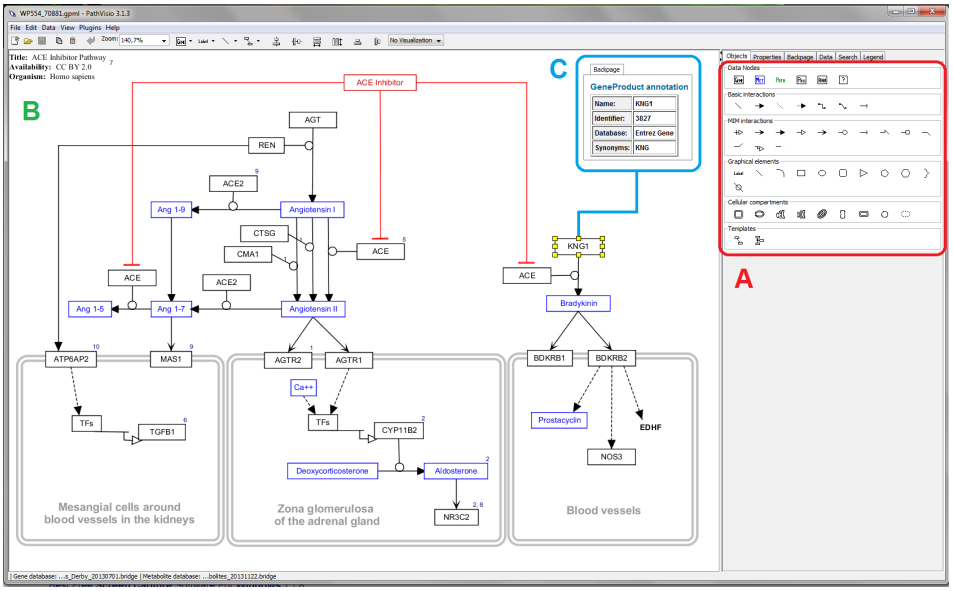


Figure 2.3: PathVisio 3, A Full-Powered Pathway Editor. (A) The basic drawing palette contains data nodes, interactions, graphical elements, cellular compartments and a few templates. Simple drag-and-drop mechanism allows users to add the elements in the pathway diagram. (B) The ACE inhibitor pathway on WikiPathways (http://www.wikipathways.org/instance/WP554) was drawn in PathVisio describing the downstream effects of angiotensin-converting-enzyme (ACE) inhibtors. (C) The entities and interactions in the pathways can be annotated with external identifiers. In this example the pathway author annotated the KNG1 gene with the Entrez Gene [22] identifier 3827. PathVisio utilizes the BridgeDb identifier mapping framework to free the user from manual identifier mapping steps.

The drag-and-drop mechanism for adding new elements is used similar as in PowerPoint and other drawing tools. Besides the external database annotation, users can also add publication references to each entity or interaction in the pathway establishing the pathway as a complete literature reference collection for the biological process described.

(2) Data Visualization

The visualization of experimental and other data is a crucial aspect in the analysis and investigation of biological pathways. PathVisio allows users to import their experimental data and visualize it on the data nodes and interactions in the pathway. The integrated identifier mapping framework takes care of mapping the data points to the intended pathway elements, therefore the user is not restricted to a specific identifier system. In integrative studies, transcriptomics, proteomics and metabolomics data can be visualized simultaneously to provide a more complete view of the underlying biology [6].

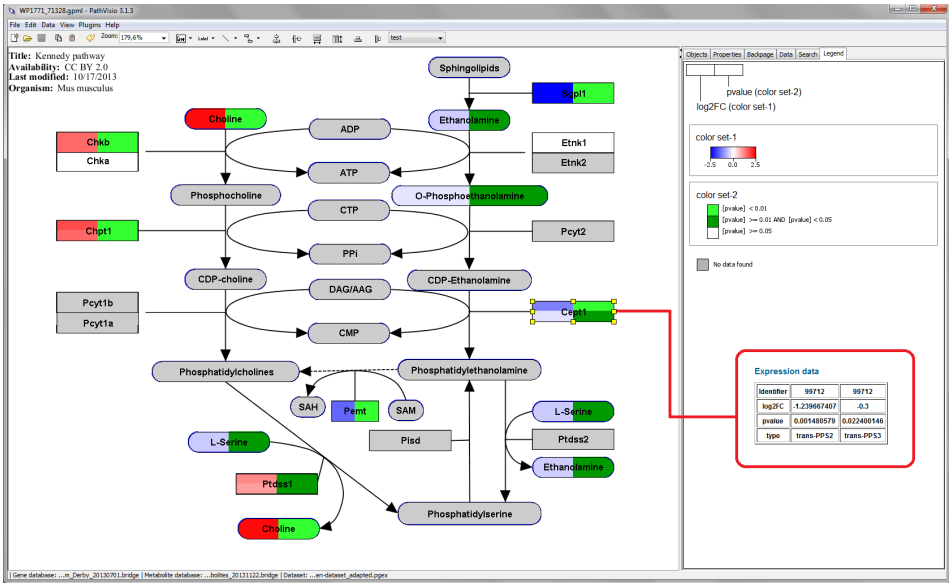


Figure 2.4: Multi-omics Visualization in PathVisio. Two transcriptomics datasets are visualized together with a metabolomics dataset on the Kennedy pathway from WikiPathways (http://www.wikipathways.org/instance/WP1771). The log2FC is visualized in the first column of the data node boxes using a gradient from blue over white to red. In the second column three levels of p-values are visualized (p-value < 0.01, < 0.05 and > 0.05). The expression data for a selected gene or metabolite is shown in the "*Data*" tab on the right side. In the red rectangle the expression data for the selected Cept1 gene is shown. There are two measurements for the gene from the two transcriptomics datasets, therefore the gene box in the pathway is split horizontally into two rows.

As detailed in Figure 2.4, the visualization interface in PathVisio enables users to visualize multiple data points on the data nodes in the diagram. The boxes are split up in separate columns and for each column the user can define a gradient or color rule visualization. A gradient is used for a continuous visualization of

numeric values like the log2FC or an activity measurement in an experiment. The color rules are used to define colors for discrete categories like p-value levels (p-value < 0.01, p-value < 0.05, p-value > 0.05) or discrete categories. The example dataset visualized in Figure 2.4 is a combined dataset of two transcriptomics and one metabolomics experiments. The first column in the datanode boxes represents the log2FC and the second column the p-value. The log2FC is visualized with a gradient from blue over white to red, while p-value is visualized with a discrete color rule. If the dataset contains multiple measurements for one data node, the box is split horizontally into separate rows each representing one measurement.

The visualization options in PathVisio 3 can be used to visualize time-series data (one column for each time point) [2], tissue expression comparisons (one column for each tissue) [23] and other complex multi-omics experiments.

(3) Pathway Statistics

The goal of pathway statistics is to find pathways that are altered in an experimental dataset. The basic pathway statistics implementation in PathVisio is an overrepresentation analysis based on the statistical methods used in the MAPPFinder tool [15].

First, the user defines a criterion to select the differentially expressed genes in the dataset. In Figure 2.5A, the criteria filters genes with an absolute log2FC $>$ 1 and a p-value < 0.05. The mouse pathway collection from WikiPathways was downloaded and selected.
The statistics module calculates the total number of genes measured in the dataset (N) and the number of genes meeting the criterion (R). All genes in N and R are present in at least one pathways. Genes that are not found in any pathway are ignored in the analysis. The Z-Score is calculated for each pathway in the collection. Therefore the statistics module counts the total number of elements in the pathway (total), the number of genes measured in the experiment (measured $\rightarrow$ n) and the number of genes meeting the criterion (positive $\rightarrow$ r) (see Figure 2.5B).
A commonly used score for overrepresentation analysis is the Z-Score. The Z-Score is the score calculated by a standard statistical test under the hypergeometric distribution. It indicates if a particular pathway shows a difference in the ratio of genes meeting the criterion as compared to the complete dataset. It is calculated by subtracting the expected number of genes meeting the criterion from the observed number divided by the standard deviation of the observed number of genes:

$$\text{Z-Score} = \frac{(r - n\frac{R}{N})}{\sqrt{n(\frac{R}{N})(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

The pathways are ranked based on their Z-Score. A positive Z-Score indicates a pathway with more genes meeting the criterion than expected by chance. A negative Z-Score indicates that less genes meet the criterion than expected by chance. In the example in Figure 2.5 pathways with a high Z-Score have more significantly

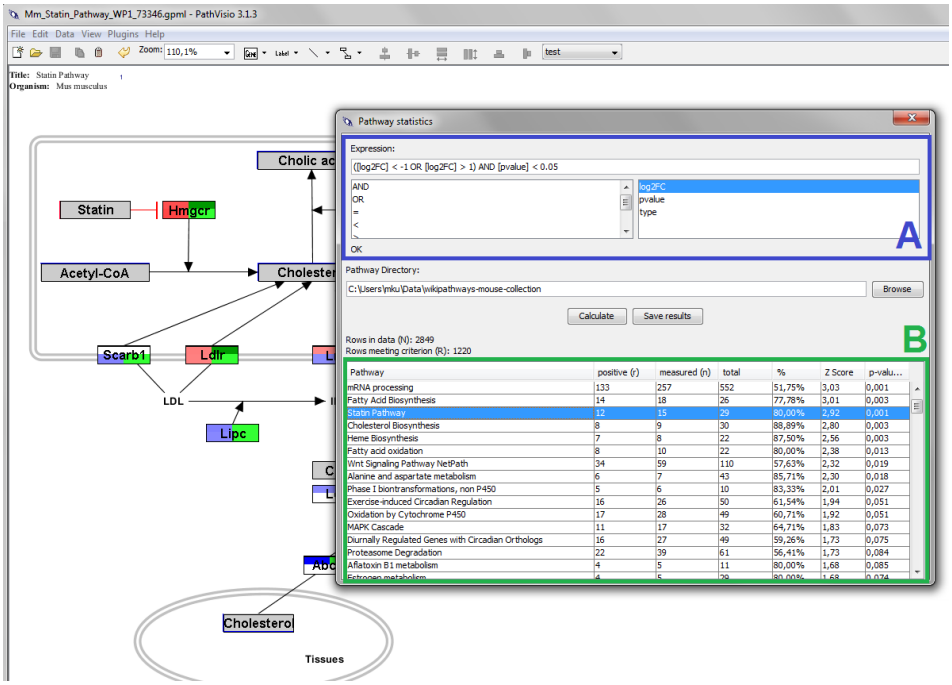


Figure 2.5: Pathway Statistics Result in PathVisio. The user defines the criterion for significantly changed genes with an absolute log2FC > 1 (A). A Z-Score is calculated for each pathway in the pathway collection and in the result table the pathways are ranked based on their Z-Score (B). A high Z-Score indicates that the pathway is more affected than expected based on the overall dataset. The user can click on each pathway to open the pathway with the data visualized on it.

up- or down-regulated genes than expected based on the complete dataset. Therefore those processes are highly affected in the experiment and should be further analyzed. Overrepresentation analysis does not take the pathway topology into account, so it is important that the users look at the pathway diagrams by clicking on the rows in the table and visualize the experimental data on the diagram to interpret the biological outcome.

Plugins in PathVisio

PathVisio provides a powerful and flexible way for plugins to integrate new functionality into the application. The variety of plugins shows that PathVisio can be extended in a lot of different ways and although initially PathVisio started as a pathway editor, it grew into an advanced and extendable pathway visualization and analysis toolbox.

The implementation of different pathway related standards is crucial to fulfil the requirements of a state-of-the-art pathway editor. BioPAX [19] is a standard language to exchange biological pathway data. The BioPAX3 plugin allows users to import and export pathways in BioPAX level 3 which is the latest release of

the BioPAX format. Furthermore, there are two plugins providing functionality to draw pathways in the commonly used SBGN [17] and MIM [16] drawing standards. The PathVisio-Validator plugin [24] assists users in creating biological pathway diagrams with the SBGN or PathVisio-MIM [25] plugins. It validates the diagrams and highlights possible warnings and errors in the pathway.

Pathway databases still only cover 47% of all human protein-coding genes. Therefore the creation and curation of biological pathways is still of high importance. Recently we released the WikiPathways plugin for PathVisio which enables users to search and browse the database directly from within PathVisio but also allows the uploading and updating of pathways through the full pathway editor. Integrating this functionality in PathVisio 3 enables pathway curators to use all the available plugins while creating new pathways or curating existing ones. Since the release of this plugin several curation related plugins have been developed to facilitate the curation of the WikiPathways pathways. Furthermore plugins focused on data integration can be used to facilitate the exploration and understanding of biological pathways. As an example, the pathway curator could use the PathVisio-Faceted Search plugin [26] to integrate experimental data and data from publicly available online resources. Another useful plugin is PathwayLoom which provides known interaction partners for a selected node in the pathway. This can help the curator to select the next element in the pathway.

Also the integration of additional data about the elements in the pathway is useful when creating and curating biological pathways. The BiomartConnect plugin queries the Ensembl database for additional information about gene products, like chromosomal position, %GC content or known variants. The MetInfo plugin provides more data about the metabolites in a pathway, like InChI key or predicted MS and NMR peaks. Plugins connecting to UniProt, PDB and interaction databases are under development.

Integration of PathVisio in Workflows and other Applications

To be able to integrate PathVisio in an automated workflow, we developed PathVisioRPC (http://projects.bigcat.unimaas.nl/pathvisiorpc) to be able to call PathVisio from other programming languages through an XML-RPC server. It enables users to programmatically draw pathways, visualize data on pathways and perform pathway statistics. This is especially convenient and time-saving when studying multiple datasets or datasets with many different comparisons.

Furthermore PathVisio is often used as a library to read, write, store, convert and model pathway information. The nice separation of the different modules in PathVisio 3 enables developers to integrate this functionality in other application simply by including the core module of PathVisio 3. This module is also used in the WikiPathways App for Cytoscape [27]. Cytoscape is a popular network analysis and visualization tool [28] and the WikiPathways app allows users to load pathways as networks in Cytoscape to perform network analysis.

Availability and Future Directions

PathVisio 3 is a freely available, open source pathway editor, visualization and analysis toolbox implemented in Java. It runs on all major operating systems as a Java webstart program or as a binary installation.

Download: http://www.pathvisio.org/downloads
Documentation and tutorials: http://www.pathvisio.org
Instructions for core and plugin developers: http://developers.pathvisio.org
Plugin repository: http://www.pathvisio.org/plugins/plugins-repo
Source code: http://svn.bigcat.unimaas.nl/pathvisio
Integrated identifier mapping framework: BridgeDb (http://www.bridgedb.org)
GPML file format: http://www.pathvisio.org/gpml

Future Directions

Future development will focus on (1) more advanced pathway analysis methods, (2) improved data integration and visualization and (3) automated update mechanisms.

(1) Advanced pathway analysis methods: The default pathway analysis method in PathVisio 3 is a simple over-representation analysis. Users can also use the *GSEA plugin* which implements a functional class scoring method which does not require a specific threshold for splitting up significant and nonsignificant measurements. This method uses all the molecular measurements and their expression levels. The next step for PathVisio is the implementation of an topology-based pathway analysis method. While over-representation analysis and functional class scoring only consider the number of genes in the pathways, topology-based methods also look at the interactions between the elements in the pathways [29].

(2) Improved data integration and visualization: PathVisio 3 supports the visualization of transcriptomics, proteomics and metabolomics data on the elements in the pathways. Recently a plugin has been developed to allow visualization of fluxomics data on the interactions in the pathways. Integration of other experimental data like genetic variation, methylation or phosphorylation states is needed to be able to study biology in all its complexity. For most of these additional data types new advanced visualization methods are needed.

(3) Automated update mechanisms: In the next major release of PathVisio, we are planning an automated update mechanism for the main application and the installed plugins. The application can be upgraded as soon as a new release is available. We will provide installers for all major operating systems that will facilitate the installation of new PathVisio versions.

Bibliography

[1] MP van Iersel, T Kelder, AR Pico, K Hanspers, S Coort, BR Conklin, and C Evelo. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics*, 9:399, January 2008.

[2] JR Tisoncik, MJ Korth, CP Simmons, J Farrar, TR Martin, and MG Katze. Into the eye of the cytokine storm. *Microbiology and molecular biology reviews : MMBR*, 76(1):16–32, March 2012.

[3] SC Baetke, ME Adriaens, R Seigneuric, CT Evelo, and LMT Eijssen. Molecular pathways involved in prostate carcinogenesis: insights from public microarray datasets. *PloS one*, 7(11):e49831, January 2012.

[4] F Spaapen, GGH van den Akker, MMJ Caron, P Prickaerts, C Rofel, VEH Dahlmans, DAM Surtel, Y Paulis, F Schweizer, TJM Welting, LM Eijssen, and JW Voncken. The immediate early gene product EGR1 and polycomb group proteins interact in epigenetic programming during chondrogenesis. *PloS one*, 8(3):e58083, January 2013.

[5] H Husi, T Van Agtmael, W Mullen, FH Bahlmann, JP Schanstra, A Vlahou, C Delles, P Perco, and H Mischak. Proteome-based systems biology analysis of the diabetic mouse aorta reveals major changes in fatty acid biosynthesis as potential hallmark in diabetes mellitus-associated vascular disease. *Circulation. Cardiovascular genetics*, 7(2):161–70, April 2014.

[6] MP van Iersel, M Sokolović, K Lenaerts, M Kutmon, FG Bouwman, WH Lamers, ECM Mariman, and CT Evelo. Integrated visualization of a multi-omics study of starvation in mouse intestine. *Journal of integrative bioinformatics*, 11(1):235, January 2014.

[7] C Jaeger, V Tellström, G Zurek, S König, S Eimer, and B Kammerer. Metabolomic changes in Caenorhabditis elegans lifespan mutants as evident from GC-EI-MS and GC-APCI-TOF-MS profiling. *Metabolomics*, pages 1–18, February 2014.

[8] J Zhong, J Sharma, R Raju, SM Palapetta, TSK Prasad, TC Huang, A Yoda, JW Tyner, D van Bodegom, DM Weinstock, SF Ziegler, and A Pandey. TSLP signaling pathway map: a platform for analysis of TSLP-mediated signaling. *Database : the journal of biological databases and curation*, 2014:bau007, January 2014.

[9] X Liu, J Huang, S Yang, Y Zhao, A Xiang, J Cao, B Fan, Z Wu, J Zhao, S Zhao, and M Zhu. Whole blood transcriptome comparison of pigs with extreme production of in vivo dsRNA-induced serum IFN-a. *Developmental and comparative immunology*, 44(1):35–43, May 2014.

[10] R Raju, SM Palapetta, VK Sandhya, A Sahu, A Alipoor, L Balakrishnan, J Advani, B George, KR Kini, NP Geetha, HS Prakash, TSK Prasad, Y-J Chang, L Chen, A Pandey, and H Gowda. A Network Map of FGF-1/FGFR Signaling System. *Journal of signal transduction*, 2014:962962, January 2014.

[11] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[12] C Bauer, A Glintschert, and J Schuchhardt. ProfileDB: a resource for proteomics and cross-omics biomarker discovery. *Biochimica et biophysica acta*, 1844(5):960–6, May 2014.

[13] R Hall, K Pauls, S McCulloch, and D Savage. *Osgi in Action: Creating Modular Applications in Java*. Manning Publications Co., April 2011.

[14] K Börner. Plug-and-play macroscopes. *Communications of the ACM*, 54(3):60, March 2011.

[15] SW Doniger, N Salomonis, KD Dahlquist, K Vranizan, SC Lawlor, and BR Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome biology*, 4(1):R7, January 2003.

[16] KW Kohn, MI Aladjem, JN Weinstein, and Y Pommier. Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Molecular biology of the cell*, 17(1):1–13, January 2006.

[17] N Le Novère, M Hucka, H Mi, S Moodie, F Schreiber, A Sorokin, E Demir, K Wegner, MI Aladjem, SM Wimalaratne, FT Bergman, R Gauges, P Ghazal, H Kawaji, L Li, Y Matsuoka, A Villéger, SE Boyd, L Calzone, M Courtot, U Dogrusoz, TC Freeman, A Funahashi, S Ghosh, A Jouraku, S Kim, F Kolpakov, A Luna, S Sahle, E Schmidt, S Watterson, G Wu, I Goryanin, DB Kell, C Sander, H Sauro, JL Snoep, K Kohn, and H Kitano. The Systems Biology Graphical Notation. *Nature biotechnology*, 27(8):735–41, August 2009.

[18] M Hucka, A Finney, HM Sauro, H Bolouri, JC Doyle, H Kitano, AP Arkin, BJ Bornstein, D Bray, A Cornish-Bowden, AA Cuellar, S Dronov, ED Gilles, M Ginkel, V Gor, II Goryanin, WJ Hedley, TC Hodgman, J-H Hofmeyr, PJ Hunter, NS Juty, JL Kasberger, A Kremling, U Kummer, N Le Novère, LM Loew, D Lucio, P Mendes, E Minch, ED Mjolsness, Y Nakayama, MR Nelson, PF Nielsen, T Sakurada, JC Schaff, BE Shapiro, TS Shimizu, HD Spence, J Stelling, K Takahashi, M Tomita, J Wagner, and J Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, 19(4):524–31, March 2003.

[19] E Demir, MP Cary, S Paley, K Fukuda, C Lemer, I Vastrik, G Wu, P D'Eustachio, C Schaefer, J Luciano, F Schacherer, I Martinez-Flores, Z Hu, V Jimenez-Jacinto, G Joshi-Tope, K Kandasamy, AC Lopez-Fuentes, H Mi, E Pichler, I Rodchenkov, A Splendiani, S Tkachev, J Zucker, G Gopinath, H Rajasimha, R Ramakrishnan, I Shah, M Syed, N Anwar, O Babur, M Blinov, E Brauner, D Corwin, S Donaldson, F Gibbons, R Goldberg, P Hornbeck, A Luna, P Murray-Rust, E Neumann, O Ruebenacker, M Samwald, M van Iersel, S Wimalaratne, K Allen, B Braun, M Whirl-Carrillo, KH Cheung, K Dahlquist, A Finney, M Gillespie, E Glass, L Gong, R Haw, M Honig, O Hubaut, D Kane, S Krupa, M Kutmon, J Leonard, D Marks, D Merberg, V Petri,

A Pico, D Ravenscroft, L Ren, N Shah, M Sunshine, R Tang, R Whaley, S Letovksy, KH Buetow, A Rzhetsky, V Schachter, BS Sobral, U Dogrusoz, S McWeeney, M Aladjem, E Birney, J Collado-Vides, S Goto, M Hucka, N Le Novère, N Maltsev, A Pandey, P Thomas, E Wingender, PD Karp, C Sander, and GD Bader. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–42, September 2010.

[20] D Cavalieri, C Castagnini, S Toti, K Maciag, T Kelder, L Gambineri, S Angioli, and P Dolara. Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. *Bioinformatics (Oxford, England)*, 23(19):2631–2, October 2007.

[21] MP van Iersel, AR Pico, T Kelder, J Gao, I Ho, K Hanspers, BR Conklin, and CT Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, 11:5, January 2010.

[22] D Maglott, J Ostell, KD Pruitt, and T Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 39(Database issue):D52–7, January 2011.

[23] DGJ Jennen, S Gaj, PJ Giesbertz, JHM van Delft, CT Evelo, and JCS Kleinjans. Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug discovery today*, 15(19-20):851–8, October 2010.

[24] K Chandan, MP van Iersel, MI Aladjem, KW Kohn, and A Luna. PathVisio-Validator: a rule-based validation plugin for graphical pathway notations. *Bioinformatics (Oxford, England)*, 28(6):889–90, March 2012.

[25] A Luna, ML Sunshine, MP van Iersel, MI Aladjem, and KW Kohn. PathVisio-MIM: PathVisio plugin for creating and editing Molecular Interaction Maps (MIMs). *Bioinformatics (Oxford, England)*, 27(15):2165–6, August 2011.

[26] JY Fried, MP van Iersel, MI Aladjem, KW Kohn, and A Luna. PathVisio-Faceted Search: an exploration tool for multi-dimensional navigation of large pathways. *Bioinformatics (Oxford, England)*, 29(11):1465–6, June 2013.

[27] M Kutmon, S Lotia, CT Evelo, and AR Pico. WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research*, 3, July 2014.

[28] ME Smoot, K Ono, J Ruscheinski, P-L Wang, and T Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–2, February 2011.

[29] P Khatri, M Sirota, and AJ Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, January 2012.

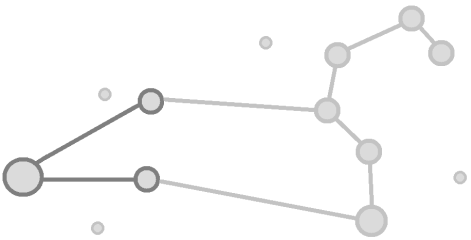
Chapter 3

A Pathway Approach to Investigate the Function and Regulation of SREBPs

Sabine Daemen, Martina Kutmon, Chris T Evelo

Department of Bioinformatics - BiGCaT, NUTRIM School for Nutrition, Toxicology and Metabolism, Maastricht University, The Netherlands

Genes & Nutrition (2013) 8:289-300



Abstract

The essential function of sterol regulatory element-binding proteins (SREBPs) in cellular lipid metabolism and homeostasis has been recognized for a long time, and the basic biological pathway involving SREBPs has been well described; however, a rapidly growing number of studies reveal the complex regulation of these SREBP transcription factors at multiple levels. This regulation allows the integration of signals of diverse pathways involving nutrients, contributing to cellular lipid and energy homeostasis. This review attempts to integrate this knowledge. The description of the SREBP pathway is web-linked as it refers to the online version of the pathway on WikiPathways (http://www.wikipathways.org), which is interactively linked to genomics databases and literature. This allows a more extensive study of the pathway through reviewing these links.

Introduction

Sterol regulatory element-binding proteins (SREBPs) play an important role in the regulation of the intracellular cholesterol concentration and in overall lipid homeostasis. Since lipids and cholesterol are important components of cellular membranes and precursors for steroid hormones, bile salts, and essential signaling molecules, a tight regulation is vital. SREBPs provide a negative feedback mechanism by sensing the intracellular levels of cholesterol. SREBPs function as transcription factors, and upon activation, by low levels of cholesterol, they stimulate the expression of genes coding for proteins involved in the synthesis of cholesterol and fatty acids and in the uptake of lipoproteins [1]. The basic signaling pathway affected by SREBPs has been elucidated in great detail. However, regulation of SREBPs themselves is proven to be very complex. In the last few years, research has brought new insights regarding this regulation and the interaction with other nutrients and hormones that play a role in energy homeostasis. Recent studies also implicated the SREBP pathway to be important in the development of a range of pathological conditions, associated with obesity and the metabolic syndrome, like liver steatosis and hyperlipidemia [2]. It has also been described that SREBP has a role in several physiological cellular processes not directly related with lipid homeostasis, like cell growth and innate immunity [3]. As the insight in the SREBP pathway becomes more and more complex, integration of the different aspects of this knowledge is vital. We will describe the SREBP protein and its isoforms, to continue with a description of the current view on the molecular basis of the SREBP pathway, its complex regulation and its physiological function. In this review, we are applying a pathway approach to investigate the function and regulation of the SREBP proteins in lipid-metabolism-related pathways.

SREBP Pathway

The description of the SREBP pathway in this review will especially focus on the role of the SREBP proteins in lipid-metabolism-related pathways. A graphical representation of the SREBP pathway (see Figure 3.1) can be found on WikiPathways, a platform for community-based curation of biological pathways [4, 5]. This pathway is a mammalian meta-pathway combining data from mouse, rat, and human studies. The description of the SREBP pathway will refer to this new pathway representation on WikiPathways. The interactive pathway viewer on WikiPathways enables the user to zoom, pan, and browse to get detailed information on pathway elements in external databases and thereby allowing a more extensive study [5]. The pathway can be found at: http://wikipathways.org/instance/WP1982.

The specific version we used for this review is:
http://wikipathways.org/instance/WP1982_r59430.

Various elements of the pathway (gene products, metabolites, interactions, and the pathway as a whole) are linked to literature references using Pubmed IDs. The gene products are among others linked to genomics databases like Ensembl [6], Entrez Gene [7] and UniProt [8] and to databases providing information on biological function and the role in diseases, including GeneOntology (GO) [9] and

OMIM [10]. The metabolites are linked to metabolite databases like HMDB [11] and ChEBI [12].

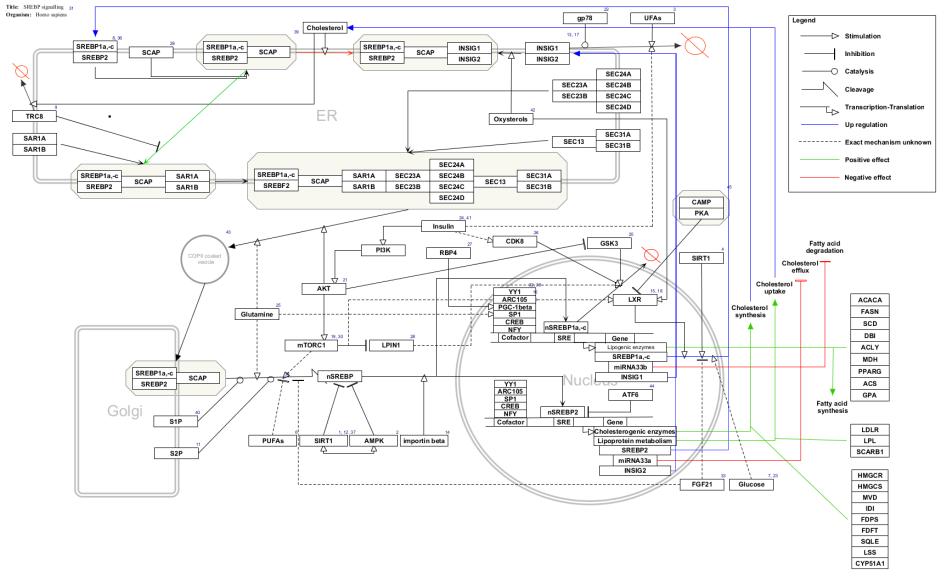


Figure 3.1: The SREBP Pathway on WikiPathways.

The SREBP Family

The SREBP family consists of three subtypes: SREBP-1a and SREBP-1c, which are the result of alternative promoter usage and transcription start sites in the SREBF1 gene, and SREBP-2. All three subtypes were identified by cDNA cloning [13, 14]. SREBPs are transcription factors that bind to the sterol regulatory element (SRE) [13]. They are synthesized as endoplasmic reticulum (ER) membrane proteins. The SREBP protein consists of three domains: a N-terminal domain which has approximately 480 amino acids, in the middle a hydrophobic region of 80 amino acids containing two membrane-spanning domains and a C-terminal regulatory domain of 590 amino acids [1]. They are oriented in a hairpin fashion in the membranes of the ER and the nuclear envelope, in which the N-terminal and C-terminal project into the cytoplasm.

The N-terminal domain is a basic-helix-loop-helix leucine zipper (bHLH-Zip). This domain is the functionally active portion of the SREBP and functions as the transcription factor. The N-terminal domain starts with an acidic domain that clusters acidic residues and functions as a transactivation domain. Deletion of this acidic domain converts SREBP-1 from an activator to an inhibitor of transcription [15]. The acidic domain is followed by a region of which the function is unknown. In SREBP-1, this region is proline and serine rich, and in SREBP-2, this region is proline, serine, glutamine, and glycine rich. This region is then followed by

the bHLH-Zip domain [1]. Figure 3.2 gives an overview of the different protein domains of the SREBP isoforms.

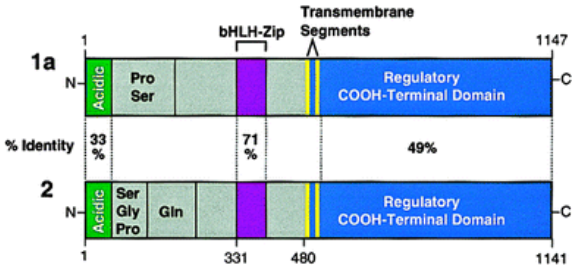


Figure 3.2: Protein Domains of the SREBP Isoforms. The structure of SREBP-1c is highly similar to SREBP-1a; SREBP-1c has a shorter transactivation domain in the N-terminus [1].

The high similarity among the N-terminal domains of the isoforms of SREBP results in the ability of all isoforms to activate all of the target genes identified so far, but with different efficiencies [16]. Given that SREBP-1c has a shorter transactivation domain, this isoform is a less potent transcription factor than SREBP-1a and SREBP-2 [17]. Several in vivo studies obtained insight in the distinct roles of the SREBP isoforms. In transgenic mice that overexpress a truncated, active nuclear form of SREBP-2 in liver and adipose tissue, it was shown that SREBP-2 is a relatively selective activator of cholesterol synthesis, as opposed to fatty acid synthesis in these tissues [18]. SREBP-1 knockout mice showed a significant decrease in mRNA coding for fatty acid synthesis enzymes. There was also a significant increase in cholesterol synthesis, but this was due to activation of SREBP-2, which compensated for the lack of SREBP-1 [19]. In general, SREBP-1 is relatively selective for lipogenic genes and SREBP-2 for cholesterogenic genes. This is due to differences among the SREBP isoforms in specificity for SREBP target promoters [20, 21].

Activation of SREBP

Since SREBP is bound to the ER membrane, the N-terminal domain must be released before SREBP can activate its target genes in the nucleus. This requires a two-step proteolytic process, which takes place in the Golgi apparatus. Therefore, the SREBP is first transported to the Golgi apparatus. Important for the regulation of the cleavage of SREBP is another ER membrane-embedded protein named SREBP cleavage-activating protein (SCAP). In mice with a SCAP-deficient liver, no nuclear form of SREBP was found, and they showed an 80% decrease in basal rates of cholesterol and fatty acid synthesis in the liver [22]. SCAP has an N-terminal domain of 730 amino acids which has eight membrane-spanning regions separated by short hydrophilic loops, which include a sterol-sensing domain (SSD). This domain is similar to the sterol-sensing domain found in other proteins which interact with sterols: 3-hydroxy-3-methylglutaryl-Coenzyme A (HMG CoA) reductase, the Niemann-Pick disease type C1 protein and Patched [23]. The C-terminal domain is a hydrophilic region of 546 amino acids containing 4 repeats

of a tryptophanaspartate repeat, WD. Both SREBP-1 and SREBP-2 form a complex with the SCAP protein on the ER membrane by binding of the WD region in SCAP to the C-terminal domain of SREBP. When there are enough sterols present in cells, cholesterol can bind directly to the sterol-sensing domain of SCAP, which then undergoes a conformational change. This conformation favors the binding of SCAP to another ER membrane protein named insulin-induced gene (Insig), which blocks translocation of the SREBPSCAP complex to the Golgi apparatus, where the proteolytic activation takes place [24]. This can be seen in the upper left corner of the pathway representation. The red arrow indicates the negative effect of cholesterol on SREBP stimulation by stimulating the binding of Insig to the SREBPSCAP complex. Metabolites, like cholesterol, are indicated in blue boxes in the pathway representation.

There are two Insig isoforms, Insig-1 and Insig-2, which are both polytopic ER membrane proteins. They play an important role in the control of lipid synthesis, not only by binding to the SCAP protein. Insigs also bind to HMG-CoA reductase, which is the rate-limiting enzyme in the synthesis of cholesterol. The binding of Insig to HMG-CoA reductase induces the ubiquitination and proteolysis of this enzyme, whereas binding of Insig to SCAP leads to ER retention [25, 26]. The dual function of Insig in cholesterol metabolism is discussed in more detail in [27]. Insig-1 and Insig-2 demonstrate an amino acid identity of 59% and are both embedded in the ER membrane by six membrane-spanning domains [28]. The regulation and the relative stability of the two isoforms differ. Insig-1 is itself a target of SREBP, whereas Insig-2a has been shown to be suppressed by insulin in hepatic cells [29, 30]. The exact mechanism of the regulation of Insig by insulin remains unclear and is therefore visualized in the WikiPathways pathway using a dashed arrow. The Insig-1 protein is quite unstable and is degraded by the ubiquitinproteasome pathway, whereas Insig-2 is a relatively stable protein, which is constitutively expressed at low levels. In transgenic mice that overexpress human Insig-1 in the liver, the levels of all nuclear SREBPs (nSREBPs) were reduced, which shows that Insig inhibits SREBP processing [31].

Upon sterol deprivation, the SREBPSCAP complex dissociates from Insig and moves to the Golgi apparatus, a process that is discussed in more detail in the next section. Insig-1 is then ubiquitinated on lysines 156 and 158 by the membrane-bound ubiquitin ligase gp78. This ubiquitin ligase has a high affinity for Insig-1, and degradation of Insig-1 in a cholesterol-rich environment is probably prevented by binding competition between gp78 and SCAP [32]. Insig-1 is subsequently degraded in proteasomes, providing a positive feedback mechanism on the activation of SREBP. nSREBPs activate the genes for cholesterol synthesis and uptake and stimulate the production of Insig-1. This upregulation can be seen in the pathway on the right in blue arrows. The new cholesterol and Insig-1 bind the SREBP-SCAP complex and the complex remains in the ER [33].

The cholesterol regulatory system is controlled not only by its end product cholesterol but also by oxysterols. Oxysterols are derivatives of cholesterol which have extra keto- or hydroxyl groups. Oxysterols were proven to regulate the interaction

of SCAP and Insig, but they do so by a different mechanism than cholesterol. Cholesterol binds to SCAP, while oxysterols bind to Insigs. This induces SCAP to bind to Insig, which inhibits the movement of the SREBPSCAP complex to the Golgi apparatus [34]. Oxysterols are also ligands of the nuclear liver X receptors (LXRs), which also play an important role in the cholesterol synthesis. Upon activation by oxysterols, LXR forms a heterodimer with the retinoid X receptors (RXRs) which binds to the LXR response element (LXRE) on target genes. An LXRE has been found in the proximal promoter region of the rat cytochrome P450 7A1 (CYP7A1) gene, which codes for an enzyme responsible for the rate-limiting step in the conversion of cholesterol to bile acids [35]. However, in the human gene promotor of CYP7A1, the LXRE appears to be not conserved. In addition, in human primary hepatocyte cultures, it has been shown that activation of the LXR represses CYP7A1 expression, indicating a species-specific difference in the regulation of cholesterol homeostasis [36]. In addition, LXRs have been implicated in the upregulation of genes involved in efflux of cholesterol from the cell, as ATP-binding cassette A1 (ABCA1). LXR/RXR can also bind the SREBP-1c promoter and induce SREBP-1 activation of fatty acid synthesis [37].

ER to Golgi Transport

If there are not enough sterols present in the cell, the SREBPSCAP complex moves to the Golgi apparatus through COPII-coated vesicles. The sorting of the complex in a COPII vesicle is depending on an amino acid sequence in the SCAP protein. SCAP has a long loop, which projects into the cytoplasm between the membrane-spanning helices 6 and 7. In this loop, the hexapeptide MELADL is found, which is required for the binding of the COPII proteins Sec23 and Sec24 to the SREBPSCAP complex. Clustering of the SREBPSCAP complex into a COPII vesicle is initiated by Sar1, a small GTPase that binds to the ER membrane GTP dependent. This binding is visualized on WikiPathways by a green arrow, which shows that this is the first step in the cascade toward activation of transcription by SREBPs. The binding of Sar1 initiates the binding of Sec23/24, which then recruits Sec13/31. This heterodimer forms the coat of the vesicle and the vesicle can bud from the ER membrane [38]. Interaction of SCAP with Insig causes a conformational change in SCAP which inhibits the interaction of MEDADL with Sec23/24.

An ER membrane protein named ring finger protein 139, also called TRC8, shown in the upper right corner on WikiPathways, was identified as a regulator in the SREBP pathway. The protein contains a SSD and a RING finger motif, which encodes for an E3 ubiquitin ligase. It is shown that the overexpressing of TRC8 inhibits SREBP-2 processing. TRC8 is capable of binding both SREBP-2 and SCAP and a TRC8SREBPSCAP complex is formed. This inhibits the binding of SCAP to Sec23/24 and blocks transport of the SREBPSCAP to the Golgi apparatus. The TRC8 protein in itself is highly unstable because of self-ubiquitination, which leads to degradation. When cells were cultured with a lipoprotein-deficient serum, the TRC8 protein became stable [39]. It is thus likely that when the SSD senses a decline in lipoprotein, it will downregulate the E3 ligase activity. It could provide

a brake on the SREBP processing in conditions of sterol depletion, preventing too much processing of SREBP [40].

Proteolytic Cleavage

After fusion of the COPII vesicle with the Golgi apparatus, the N-terminal of the SREBPs is released by intramembrane proteolysis. The processing of SREBP-nSREBP is shown in the bottom left corner on WikiPathways. The process is executed by two proteases, membrane-bound transcription factor peptidase site 1, or Site-1 protease (S1P), and Site-2 protease (S2P). The process is initiated when S1P, a membrane-bound serine protease, cleaves the leucineserine bond in the sequence RSVLS within the luminal loop of SREBP [41]. This separates the two membrane-spanning segments. The next step is cleavage by S2P, which hydrolyzes a leucinecysteine bond in the sequence DRSRILLC. This sequence lies within the N-terminal membrane-spanning domain, and cleavage occurs in three residues in this domain [42]. The result is that the N-terminal domain is released from the SREBP and functions as an active nSREBP, which migrates to the nucleus to activate target genes [43]. It has been proposed that the cleavage of S1P is required for the cleavage of S2P, because the separation of SREBP into two halves causes a conformational change in the first membrane-spanning domain which allows S2P to be exposed to its target sequence, thus favoring the cleavage of S2P [44]. In addition, it has been demonstrated that caspase 3, a cysteine protease that is involved in the induction of apoptosis, releases mature SREBP from the ER membrane, probably in a sterol-independent manner [45].

SREBP Target

The nSREBPs released during the cleavage reaction travel into the nucleus. This nuclear transport is mediated by karyopherin (importin) beta, which interacts with the bHLH-Zip motif [46] . Important genes involved in lipid metabolism that are activated by SREBP are listed individually on WikiPathways. Besides activating these target genes, SREBPs also induce transcription of the SREBP gene itself, which contains a SRE, and thus stimulate production of new SREBPs and provide a positive feedback loop. Although SREBPs mainly activate target genes, genes with a SRE sequence have been reported which are repressed by SREBPs. These genes are, for example, microsomal triglyceride transfer protein (MTTP) [47] and caveolin [48]. Since SREBP is active in cases of cholesterol depletion, it is likely that SREBPs repress these genes, which are involved in the efflux of cholesterol and the secretion of lipoproteins [16]. The inhibition of genes by SREBP could be due to an indirect effect, namely through activation of repressors. For example, in human myotubes, it has been shown that the transcriptional repressor genes BHLHB2 and BHLHB3 are SREBP-1 target genes, negatively regulating skeletal muscle development [49].

Activation of the target genes by SREBP requires several cofactors. Usually, nuclear transcription factor γ (NF-γ), Sp1 transcription factor, and CREB-binding protein (CBP) act as cofactors for SREBP. Binding sites for these factors are often found in the SREBP target gene promoters and they are involved in the

assembly of the transcription machinery [50]. The activator recruited-cofactor (ARC)-mediated co-activator complex, a large complex that associates with RNA polymerase II, has also been found to interact with SREBPs. They have been shown to use ARC105 to activate target genes [51]. The peroxisome proliferator-activated receptor-γ coactivator-1 (PGC-1) family functions as important regulators of lipid metabolism. PGC-1β has been found to interact with SREBPs and works as a transcriptional co-activator in the transcription of lipogenic genes [52].

Another transcription factor named the YY1 transcription factor seems to be negatively involved in the regulation of SREBP target gene activation. It is shown that the promoters of the HMG-CoA synthase, farnesyl diphosphate (FDP) synthase, and the low-density lipoprotein (LDL) receptor contain YY1 binding sites. YY1 seems to repress SREBP activation by the displacement of NP-Y from the promoter [53]. Other studies suggest YY1 acts by inhibiting the interaction between Sp1 and SREBP [54]. The physiological role of YY1, however, is yet to be identified. On WikiPathways, the cofactors are drawn on a cofactor-binding site in the promoter of the target genes. The green colored boxes show activators, whereas the red colored boxes represent repressors.

Interestingly, the SREBF-1 and SREBF-2 gene loci contain, respectively, miR33b in intron 17 and miR33a in intron 16. The mature microRNAs (miRNAs) differ in only two nucleotides, but are thought to have a largely overlapping target gene set [55, 56]. These miRNAs appear to work synergistically with SREBP in increasing fatty acid synthesis and cholesterol synthesis and uptake [57–59]. Interestingly, rodents lack the miR33b gene in the SREBF-1 gene [57]. Both miR33a and miRNA33b seem to inhibit the expression of genes involved in fatty acid degradation, e.g., carnitine O-octanoyltransferase (CROT), and genes that negatively regulate fat production, e.g., insulin receptor substrate 2 (IRS2) [56]. In addition, they also repress expression of ATP-binding cassette transporter A1 (ABCA1), which normally promotes the efflux of cholesterol from cells to apolipoprotein A1 (APOA1), leading to high-density lipoprotein (HDL) formation [58].

Regulation of the SREBP Pathway

Expression and processing of the isoforms of SREBP in vivo was found to be very complex. The SREBP pathway is not just regulated on cell level by the intracellular level of cholesterol, but it can be affected by the nutritional and hormonal status of the body as a whole.

Several studies provided a link between insulin, glucose, and SREBPs. It is known that glucose and insulin stimulate fatty acid synthesis through activation of hepatic lipogenic genes. It has been recognized the PI3K/AKT pathway plays an important role in the regulation of SREBP by insulin. A range of studies has been done on exploring the effect of the PI3K/AKT pathway on SREBP, and effects on transcription, activity, processing, and stability have been found [60]. A possible mechanism is that insulin increases the migration of the SREBPSCAP complex from ER to Golgi. Insulin stimulates Akt/PKB-dependent phosphory-

lation of serine and threonine residues of SREBP-1c. This leads to an increased affinity of the SREBPSCAP complex for de COPII proteins, Sar1 and Sec23/24, and a decreased affinity for Insig, which retains the SREBPSCAP complex in the ER membrane [61]. It has also been shown that insulin enhances processing of SREBP-1c in hepatic cells by stimulation of the degradation of Insig-2a mRNA, reducing Insig-2a protein levels [30]. The PI3K/AKT pathway inhibits glycogen synthase kinase 3 (GSK3) through phosphorylation. It has been proposed that this diminishes degradation of mature SREBP-1, since GSK3 has been shown to promote ubiquitination and proteasomal degradation of SREBP-1 through a phosphorylation cascade; GSK3 phosphorylates SREBP-1 at Ser-434, whereby it increases its own affinity for Ser-430 and Thr-426 in SREBP-1, leading to GSK-3-dependent phosphorylation of these sites and a binding site for the ubiquitin ligase Fbw7 [62]. One of the major downstream regulators of the PI3K/AKT pathway is the mammalian target of rapamycin (mTOR). In the past, it has been shown that the mTOR complex-1 (mTORC1) positively regulates the processing of SREBP-1. It was thought this activation was mediated by the ribosomal protein S6 kinase (RPS6K2), which is phosphorylated by mTORC1 [63]. It has recently been shown that insulin-mediated stimulation of SREBP-1c processing required mTOR, studied in a hepatic system in which the effect of insulin on SREBP-1c processing could be dissected from the effect of insulin on SREBP-1c transcription, described below. This stimulation of SREBP processing by insulin could be inhibited by using an inhibitor of p70 ribosomal S6K, leading to an increase in nSREBP-1c, which was more likely due to an increased production of nSREBP-1c then decreased degradation. The mechanisms by which S6K can lead to increase in nSREBP-1c require further investigation [64, 65].

In addition, it has been suggested that the regulation of SREBP-1 is achieved by the regulation of the nuclear entry of phosphatidate phosphatase lipin 1 by mTORC1. Lipins are involved in triacylglycerol biosynthesis and have a second function as transcriptional co-activators. Lipins are sequestered in the cytosol in a hyper-phosphorylated state, and phosphorylation is induced by mTORC1. Loss of mTORC1-mediated lipin 1 phosphorylation promotes the nuclear entry of lipin 1, and this promotes downregulation of nSREBP, of which the exact mechanism is unknown [66].

Also, insulin can increase basal transcription of the SREBP-1c gene. The liver X receptor has been reported to have a central role in this insulin-mediated activation of SREBP-1c transcription. In the mouse promoter of SREBP-1c, two LXR elements have been found. In rat primary hepatocytes, it was shown that disruption of both LXREs blunts the effect of insulin on transcription of SREBP-1c [67]. In contrast, another study did not find a major involvement of the LXREs in the response to insulin, but insulin requires the presence of SRE in the SREBF-1 promoter and enhanced the binding of SREBP-1 to its own promoter. However, it should be noticed that this study made use of a different system based on HEK293 cells [68]. cAMP, which can be activated by glucagon, and the cAMP-dependent kinase, protein kinase A (PKA) have been shown to suppress SREBP-1c transcription by phosphorylation of LXR, which inhibits the DNA binding activity

by inhibiting LXR/RXR dimerization, decreases recruitment of a coactivator, and enhances the recruitment of a corepressor [69]. In addition, it has been shown in HepG2 cells that PKA can phosphorylate SREBP-1a at Ser338, which reduces DNA binding of SREBP-1c [70]. These results indicate a role for the cAMP/PKA pathway in mediating SREBP-1 and hepatic lipogenesis.

It has been shown that the increase in SREBP-1 expression stimulated by insulin can be inhibited by wortmannin and rapamycin, indicating the PI3K-mTORC1 pathway is involved. In contrast to the stimulation of SREBP-1c processing by insulin, the increase in SREBP-1 expression by insulin could not be blocked by inhibiting S6K. This suggests that the regulation of SREBP-1c by insulin bifurcates downstream of mTORC1, with one arm controlling the processing of SREBP-1c and the other the gene expression [64, 65]. Furthermore, it has been shown that upstream of this in the liver, by using liver-specific rictor knockout mice, insulin stimulates mTOR complex-2 (mTORC2), which phosphorylated Akt at serine 473, leading to SREBP-1c activation [71]. Other studies showed that a glucose-dependent increase in SREBP-1c protein, shown in the lower right corner of the pathway, was due to an increase in SREBP-1 mRNA, suggesting that glucose regulates the expression of SREBP-1c at transcriptional level [72]. In a human renal proximal tubular cell line, it was shown the glucose-dependent activation of SREBP was potentially mediated through the PI3K/AKT pathway [73]. SREBP-2 levels remained unchanged when treated with insulin and glucose in the liver. That insulin only stimulates hepatic SREBP-1, and not SREBP-2, matches the fact that insulin and SREBP-1 have both been shown to induce lipogenesis. However, in the brain, it has been shown that in insulin-deficient diabetic mice, there is a reduction in the expression of SREBP-2, suggesting that in the brain, insulin upregulates SREBP-2 expression [74]. A complete picture of the regulation of SREBP by insulin and glucose requires additional studies. In addition, cyclin-dependent kinase 8 (CDK8) and its regulatory partner cyclin C (CycC), which are part of the coactivator mediator complexes in mammalian cells, have been identified as regulators of de novo lipogenesis in Drosophila. Site-specific phosphorylation of nuclear SREBP-1c by CDK8 results in an enhanced ubiquitination and degradation of nSREBP-1c. Insulin and feeding decreased the levels of CDK2 and CycC and enhanced the levels SREBP-1c, indicating CDK8CycC acts downstream of insulin in the regulation of de novo lipogenesis [75].

A crosstalk between SREBP and carbohydrate responsive element-binding protein (ChREBP) has been found. These transcription factors appear to work synergistically to regulate glycolytic and lipogenic gene expression. The phosphorylation of glucose to glucose-6-phosphate by hepatic glucokinase (GK) was found to be essential in the induction of glycolytic and lipogenic genes [76]. SREs have been found in the GK promoter, which is an indication that SREBP can activate GK expression after activation by insulin. In the presence of high glucose, xylulose 5-phosphate (X5P) can be formed, which can activate protein phosphatase 2A (PP2A). This phosphatase can dephosphorylate ChREBP, leading to nuclear translocation of this transcription factor, where it binds to carbohydrate response element (ChRE) in the promoter of glycolytic and lipogenic genes. In addition,

SREBP-1c can stimulate glycolytic and lipogenic gene transcription after stimulation by insulin. Thus, in the presence of high glucose and insulin, ChREBP and SREBP can work synergistically to activate glycolytic and lipogenic genes [77].

Activating transcription factor-6 (ATF6) has also been found to interact with SREBP-2. ATF6 is also an ER membrane-bound transcription factor, which upon stimulation is translocated from ER to Golgi, where proteolytic cleavage by S1P and S2P occurs [78]. ATF6 is stimulated by the accumulation of misfolded or unfolded proteins and this ER stress could be caused by glucose deprivation. The cleaved ATF6 translocates to the nucleus and binds to nSREBP-2 bound to target genes promoters. The nuclear ATF6 recruits histone deacetylase 1 (HDAC1), which downregulates SREBP-2 gene expression. The physiological relevance could be that when glucose is depleted, lipogenesis and cholesterogenesis are downregulated to save energy [79].

Alternative regulators of the SREBP pathway are polyunsaturated fatty acids (PUFAs), another example of how diet influences the activation of SREBP. PUFAs have been known as negative regulators of hepatic lipogenesis and have an inhibitory effect on the SREBP pathway. PUFAs appear to suppress the proteolytic processing of SREBP-1c. Suppression of the proteolytic processing of SREBP in turn leads to a decrease in SREBP-1c transcription through lowering SREBP-1c binding to SRE on its own promoter. The exact molecular mechanism underlying this suppression still remains unknown, which is shown by the dashed arrow in the pathway. PUFAs do not seem to affect the functioning of SREBP-2 [80]. There are several reports suggesting LXR is involved in transcriptional regulation of SREBPs by PUFA [81, 82]. However, several other studies did not find an involvement of LXR in the regulation by PUFA, which could be due to different study systems being used [80, 83]. In addition, it has been shown that unsaturated fatty acids inhibit proteosomal degradation of Insig-1. Membrane proteins of the ER can be degraded by the ubiquitination-proteasome system in a process called ER-associated degradation (ERAD). In this process, valosin-containing protein (VCP) extracts ubiquitinated proteins from the membrane making the proteins accessible for degradation in the proteasome. Another protein, named Ubxd8, recruits VCP to Insig-1. Unsaturated fatty acids (UFAs) appear to block to interaction between Ubxd8 and VCP, thereby inhibiting the extraction of Insig-1 from the membrane [84].

Recent findings suggest the amino acid glutamine is also involved the regulation of the gene expression and processing of SREBPs, suggesting another link between amino acid metabolism and lipid metabolism. Glutamine seems to increase mRNA levels of several SREBP targets. Glutamine aids in the gene expression of SREBP-1 by increasing the binding of the transcription factor Sp1 to the SREBP-1a promoter. Glutamine also increases the processing of the SREBP protein, presumably by stimulating the transport of the SREBPSCAP complex from ER to the Golgi apparatus [85].

The NAD^+-dependent deacetylase SIRT1 has been shown to directly deacetylate SREBP-1c, leading to a decreased stability of the protein and a reduced association of SREBP-1c with its target genes [86]. Furthermore, SIRT1 has been shown to downregulate target gene expression by SREBP-1c in vivo under fasting conditions [87]. Whereas these studied focuses on the liver, recently, it has been shown that SIRT1 also regulated SREBP-1c expression in skeletal muscle. Interestingly, the effect of SIRT1 on SREBP-1c expression was completely abolished when the LXR response elements in the SREBF-1 promoter were deleted, which suggest SIRT1 regulates SREBP-1c expression in muscle by deacetylation of LXR transcription factors [88]. In addition, AMP-activated protein kinase (AMPK) has been shown to directly phosphorylate SREBP-1c and thereby directly inhibit SREBP-1c processing and translocation to the nucleus in the liver [89]. Interestingly, there is evidence that AMPK and SIRT1 stimulate each other and share targets [90].

A link has also been found between fibroblast growth factor 21 (FGF21) and SREBP-1c in hepatocytes. FGF21 has been identified as a regulator of energy homeostasis, glucose, and lipid metabolism. However, little is known about the regulation or activity of this FGF. It was found that FGF21 downregulated the transcription of SREBP-1c, but the processing of SREBP-1c to its mature form was also diminished. Interestingly, it was found that SREBP-1c could also inhibit FGF21 expression. Molecular mechanisms and biological relevance of this link remain unclear for the time being [91].

Recently, the role of retinol binding protein 4 (RBP4) in lipogenesis has been explored. In HepG2 cells, human RBP4 induces an increase in mature SREBP-1 and its nuclear translocation, which was also confirmed in an in vivo experiment. In addition, treatment of HepG2 cells with RBP4 leads to a strong upregulation of the expression and protein levels of PGC-1β. This suggests that RBP4 induces SREBP-1 activation through induction of PGC-1β, leading to an increase in hepatic lipogenesis [92]. Earlier, it was already reported that retinoic acid and retinal can synergize with insulin to induce the expression of SREBP-1c in primary rat hepatocytes. This was mediated via the retinoid X receptor. This indicated a role of retinol in regulating hepatic lipogenesis [93].

Other Roles of SREBP

The important function of SREBP in lipid metabolism led to these proteins being involved in a variety of pathological conditions related to lipid metabolism, as steatosis and hyperlipidemia [2]. However, several other functions of SREBP, not directly related to lipid metabolism, have emerged recently. SREBPs have been found to regulate several cellular processes, including autophagy, phagocytosis, membrane biogenesis, immunity, hypoxia, and the cell cycle. SREBP-2 has been found to occupy promoters of genes that are involved in mediating autophagy and knockdown of SREBP-2 decreases autophagosome formation in cholesterol-depleted cells, indicating a role for SREBP-2 in autophagy [94]. Phagocytosis occurs especially within the phagocytic cells of the innate immune system to engulf exogenous particles. Phagocytosis can promote membrane biogenesis via the

activation of SREBP-1a and SREBP-2 [95]. In addition, it has been reported that bacterial pore-forming toxins can trigger cleavage and activation of SREBP-1 and SREBP-2, probably through caspase-1, which could aid in membrane repair [96]. Furthermore, it has been found that SREBP-1a can induce expression of the anti-apoptotic gene Api6 when toxin is present, which promotes cell survival [97]. In fission yeast, it was found that SREBP homologs stimulated transcription of genes that are involved in adaption to hypoxia in response to low oxygen levels [98]. Several reports found SREBP to be involved in cell cycle control. nSREBP-1 appears to be hyperphosphorylated by cyclin-dependent kinase (CDK)1/Cyclin B during mitosis, which stabilizes nSREBP. Furthermore, inactivation of SREBP-1 arrested the cells in the G1 phase of the cell cycle [99]. In addition, expression of the major CDK inhibitor p21 was found to be induced by SREBP-1 [100]. Interestingly, miR33, located in the SREBP gene locus, also appears to be involved in the regulation of the cell cycle. miR33 inhibits CDK6 and cyclin D1 and thus reduces cell cycle progression, with overactivation of miR33 even leading to a cell cycle arrest in the G1 phase [101]. The roles of SREBP beyond lipid metabolism have been reviewed in more detail in recent reviews [3, 102].

Conclusion

We previously described how literature review can be used to obtain highly curated pathways for biological processes [103, 104], which can be used for data analysis in PathVisio [105]. The new interactive browsing functionality of WikiPathways now allows the pathways themselves to be used as interactive means to study relevant literature and database information on the reactions and entities involved, their known roles in biology and disease, relevant genetic variation, chemical properties, etc. The SREBP pathway on WikiPathways described here is an example that makes full use of this functionality.
The basic pathway of SREBP signaling has been well described. When sterol levels are high, Insig retains the SREBPSCAP complex within the ER membrane. In case of sterol depletion, the SREBPSCAP complex interacts with COPII proteins and migrates in COPII vesicles to the Golgi apparatus. In the Golgi apparatus, SREBP is cleaved and active nuclear SREBP is released. This nSREBP migrates to the nucleus to activate target genes involved in lipid metabolism. However, the regulation of the pathway proves to be very complex and there are still many unanswered questions, especially regarding target genes and regulation. Increasingly, links are being found between the SREBP pathway and other regulators of lipid, protein, and carbohydrate metabolism and overall energy homeostasis: PUFAs are an example how diet influences the SREBP pathway, the link found between glutamine and SREBP suggests another link between amino acid metabolism and lipid metabolism, the interaction of ATF6 and SREBP-2 could imply that the synthesis of cholesterol is slowed in case of energy stress through SREBP-2 inhibition. Especially important to recognize are the links between insulin, glucose, and SREBP, suggesting an important role for SREBP in the pathology of current diseases, as obesity and the metabolic syndrome. Combining and integrating the growing knowledge on the SREBP pathway is essential, in which biological pathway creation and curation can play a major role.

Bibliography

[1] MS Brown and JL Goldstein. The SREBP pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell*, 89(3):331–40, May 1997.

[2] Y-A Moon, G Liang, X Xie, M Frank-Kamenetsky, K Fitzgerald, V Koteliansky, MS Brown, JL Goldstein, and JD Horton. The Scap/SREBP pathway is essential for developing diabetic fatty liver and carbohydrate-induced hypertriglyceridemia in animals. *Cell metabolism*, 15(2):240–6, February 2012.

[3] T-I Jeon and TF Osborne. SREBPs: metabolic integrators in physiology and metabolism. *Trends in endocrinology and metabolism: TEM*, 23(2):65–72, February 2012.

[4] AR Pico, T Kelder, MP van Iersel, K Hanspers, BR Conklin, and C Evelo. WikiPathways: pathway editing for the people. *PLoS biology*, 6(7):e184, July 2008.

[5] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[6] P Flicek, MR Amode, D Barrell, K Beal, S Brent, Y Chen, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gordon, M Hendrix, T Hourlier, N Johnson, A Kähäri, D Keefe, S Keenan, R Kinsella, F Kokocinski, E Kulesha, P Larsson, I Longden, W McLaren, B Overduin, B Pritchard, HS Riat, D Rios, GRS Ritchie, M Ruffier, M Schuster, D Sobral, G Spudich, YA Tang, S Trevanion, J Vandrovcova, AJ Vilella, S White, SP Wilder, A Zadissa, J Zamora, BL Aken, E Birney, F Cunningham, I Dunham, R Durbin, XM Fernández-Suarez, J Herrero, TJP Hubbard, A Parker, G Proctor, J Vogel, and SMJ Searle. Ensembl 2011. *Nucleic acids research*, 39(Database issue):D800–6, January 2011.

[7] D Maglott, J Ostell, KD Pruitt, and T Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 39(Database issue):D52–7, January 2011.

[8] Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, 40(Database issue):D71–5, January 2012.

[9] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9, May 2000.

[10] B Borate and AD Baxevanis. Searching Online Mendelian Inheritance in Man (OMIM) for information on genetic loci involved in human disease. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 1:Unit 1.2, September 2009.

[11] DS Wishart, T Jewison, AC Guo, M Wilson, C Knox, Y Liu, Y Djoumbou, R Mandal, F Aziat, E Dong, S Bouatra, I Sinelnikov, D Arndt, J Xia, P Liu, F Yallou, T Bjorndahl, R Perez-Pineiro, R Eisner, F Allen, V Neveu, R Greiner, and A Scalbert. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic acids research*, 41(Database issue):D801–7, January 2013.

[12] P de Matos, R Alcántara, A Dekker, M Ennis, J Hastings, K Haug, I Spiteri, S Turner, and C Steinbeck. Chemical Entities of Biological Interest: an update. *Nucleic acids research*, 38(Database issue):D249–54, January 2010.

[13] C Yokoyama, X Wang, MR Briggs, A Admon, J Wu, X Hua, JL Goldstein, and MS Brown. SREBP-1, a basic-helix-loop-helix-leucine zipper protein that controls transcription of the low density lipoprotein receptor gene. *Cell*, 75(1):187–97, October 1993.

[14] X Hua, C Yokoyama, J Wu, MR Briggs, MS Brown, JL Goldstein, and X Wang. SREBP-2, a second basic-helix-loop-helix-leucine zipper protein that stimulates transcription by binding to a sterol regulatory element. *Proceedings of the National Academy of Sciences of the United States of America*, 90(24):11603–7, December 1993.

[15] R Sato, J Yang, X Wang, MJ Evans, YK Ho, JL Goldstein, and MS Brown. Assignment of the membrane attachment, DNA binding, and transcriptional activation domains of sterol regulatory element-binding protein-1 (SREBP-1). *The Journal of biological chemistry*, 269(25):17267–73, June 1994.

[16] H Shimano. Sterol regulatory element-binding proteins (SREBPs): transcriptional regulators of lipid synthetic genes. *Progress in lipid research*, 40(6):439–52, November 2001.

[17] H Shimano, JD Horton, I Shimomura, RE Hammer, MS Brown, and JL Goldstein. Isoform 1c of sterol regulatory element binding protein is less active than isoform 1a in livers of transgenic mice and in cultured cells. *The Journal of clinical investigation*, 99(5):846–54, March 1997.

[18] JD Horton, I Shimomura, MS Brown, RE Hammer, JL Goldstein, and H Shimano. Activation of cholesterol synthesis in preference to fatty acid synthesis in liver and adipose tissue of transgenic mice overproducing sterol regulatory element-binding protein-2. *The Journal of clinical investigation*, 101(11):2331–9, June 1998.

[19] H Shimano, N Yahagi, M Amemiya-Kudo, AH Hasty, J Osuga, Y Tamura, F Shionoiri, Y Iizuka, K Ohashi, K Harada, T Gotoda, S Ishibashi, and N Yamada. Sterol regulatory element-binding protein-1 as a key transcription factor for nutritional induction of lipogenic enzyme genes. *The Journal of biological chemistry*, 274(50):35832–9, December 1999.

[20] M Amemiya-Kudo, H Shimano, AH Hasty, N Yahagi, T Yoshikawa, T Matsuzaka, H Okazaki, Y Tamura, Y Iizuka, K Ohashi, J Osuga, K Harada, T Gotoda, R Sato, S Kimura, S Ishibashi, and N Yamada. Transcriptional activities of nuclear SREBP-1a, -1c, and -2 to different target promoters of lipogenic and cholesterogenic genes. *Journal of lipid research*, 43(8):1220–35, August 2002.

[21] JT Pai, O Guryev, MS Brown, and JL Goldstein. Differential stimulation of cholesterol and unsaturated fatty acid biosynthesis in cells expressing individual nuclear sterol regulatory element-binding proteins. *The Journal of biological chemistry*, 273(40):26138–48, October 1998.

[22] M Matsuda, BS Korn, RE Hammer, YA Moon, R Komuro, JD Horton, JL Goldstein, MS Brown, and I Shimomura. SREBP cleavage-activating protein (SCAP) is required for increased lipid synthesis in liver induced by cholesterol deprivation and insulin elevation. *Genes & development*, 15(10):1206–16, May 2001.

[23] X Hua, A Nohturfft, JL Goldstein, and MS Brown. Sterol resistance in CHO cells traced to point mutation in SREBP cleavage-activating protein. *Cell*, 87(3):415–26, November 1996.

[24] T Yang, PJ Espenshade, ME Wright, D Yabe, Y Gong, R Aebersold, JL Goldstein, and MS Brown. Crucial step in cholesterol homeostasis: sterols promote binding of SCAP to INSIG-1, a membrane protein that facilitates retention of SREBPs in ER. *Cell*, 110(4):489–500, August 2002.

[25] J Cao, J Wang, W Qi, HH Miao, J Wang, L Ge, RA DeBose-Boyd, JJ Tang, BL Li, and BL Song. Ufd1 is a cofactor of gp78 and plays a key role in cholesterol metabolism by regulating the stability of HMG-CoA reductase. *Cell metabolism*, 6(2):115–28, August 2007.

[26] N Sever, T Yang, MS Brown, JL Goldstein, and RA DeBose-Boyd. Accelerated degradation of HMG CoA reductase mediated by binding of insig-1 to its sterol-sensing domain. *Molecular cell*, 11(1):25–33, January 2003.

[27] MT Bengoechea-Alonso and J Ericsson. SREBP in signal transduction: cholesterol metabolism and beyond. *Current opinion in cell biology*, 19(2):215–22, April 2007.

[28] D Yabe, MS Brown, and JL Goldstein. Insig-2, a second endoplasmic reticulum protein that binds SCAP and blocks export of sterol regulatory element-binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12753–8, October 2002.

[29] D Yabe, R Komuro, G Liang, JL Goldstein, and MS Brown. Liver-specific mRNA for Insig-2 down-regulated by insulin: implications for fatty acid synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3155–60, March 2003.

[30] CR Yellaturu, X Deng, EA Park, R Raghow, and MB Elam. Insulin enhances the biogenesis of nuclear sterol regulatory element-binding protein (SREBP)-1c by posttranscriptional down-regulation of Insig-2A and its dissociation from SREBP cleavage-activating protein (SCAP).SREBP-1c complex. *The Journal of biological chemistry*, 284(46):31726–34, November 2009.

[31] LJ Engelking, H Kuriyama, RE Hammer, JD Horton, MS Brown, JL Goldstein, and G Liang. Overexpression of Insig-1 in the livers of transgenic mice inhibits SREBP processing and reduces insulin-stimulated lipogenesis. *The Journal of clinical investigation*, 113(8):1168–75, April 2004.

[32] JN Lee, B Song, RA DeBose-Boyd, and J Ye. Sterol-regulated degradation of Insig-1 mediated by the membrane-bound ubiquitin ligase gp78. *The Journal of biological chemistry*, 281(51):39308–15, December 2006.

[33] Y Gong, JN Lee, PCW Lee, JL Goldstein, MS Brown, and J Ye. Sterol-regulated ubiquitination and degradation of Insig-1 creates a convergent mechanism for feedback control of cholesterol synthesis and uptake. *Cell metabolism*, 3(1):15–24, January 2006.

[34] A Radhakrishnan, Y Ikeda, HJ Kwon, MS Brown, and JL Goldstein. Sterol-regulated transport of SREBPs from endoplasmic reticulum to Golgi: oxysterols block transport by binding to Insig. *Proceedings of the National Academy of Sciences of the United States of America*, 104(16):6511–8, April 2007.

[35] JM Lehmann, SA Kliewer, LB Moore, TA Smith-Oliver, BB Oliver, JL Su, SS Sundseth, DA Winegar, DE Blanchard, TA Spencer, and TM Willson. Activation of the nuclear receptor LXR by oxysterols defines a new hormone response pathway. *The Journal of biological chemistry*, 272(6):3137–40, February 1997.

[36] B Goodwin, MA Watson, H Kim, J Miao, JK Kemper, and SA Kliewer. Differential regulation of rat and human CYP7A1 by the nuclear oxysterol receptor liver X receptor-alpha. *Molecular endocrinology (Baltimore, Md.)*, 17(3):386–94, March 2003.

[37] JR Schultz, H Tu, A Luk, JJ Repa, JC Medina, L Li, S Schwendner, S Wang, M Thoolen, DJ Mangelsdorf, KD Lustig, and B Shan. Role of LXRs in control of lipogenesis. *Genes & development*, 14(22):2831–8, November 2000.

[38] L-P Sun, L Li, JL Goldstein, and MS Brown. Insig required for sterol-mediated inhibition of Scap/SREBP binding to COPII proteins in vitro. *The Journal of biological chemistry*, 280(28):26483–90, July 2005.

[39] M Irisawa, J Inoue, N Ozawa, K Mori, and R Sato. The sterol-sensing endoplasmic reticulum (ER) membrane protein TRC8 hampers ER to Golgi transport of sterol regulatory element-binding protein-2 (SREBP-2)/SREBP cleavage-activated protein and reduces SREBP-2 cleavage. *The Journal of biological chemistry*, 284(42):28995–9004, October 2009.

[40] R Sato. Sterol metabolism and SREBP activation. *Archives of biochemistry and biophysics*, 501(2):177–81, September 2010.

[41] EA Duncan, MS Brown, JL Goldstein, and J Sakai. Cleavage site for sterol-regulated protease localized to a leu-Ser bond in the lumenal loop of sterol regulatory element-binding protein-2. *The Journal of biological chemistry*, 272(19):12778–85, May 1997.

[42] EA Duncan, UP Davé, J Sakai, JL Goldstein, and MS Brown. Second-site cleavage in sterol regulatory element-binding protein occurs at transmembrane junction as determined by cysteine panning. *The Journal of biological chemistry*, 273(28):17801–9, July 1998.

[43] MS Brown and JL Goldstein. A proteolytic pathway that controls the cholesterol content of membranes, cells, and blood. *Proceedings of the National Academy of Sciences of the United States of America*, 96(20):11041–8, September 1999.

[44] J Ye, UP Davé, NV Grishin, JL Goldstein, and MS Brown. Asparagine-proline sequence within membrane-spanning segment of SREBP triggers intramembrane cleavage by site-2 protease. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10):5123–8, May 2000.

[45] ME Higgins and YA Ioannou. Apoptosis-induced release of mature sterol regulatory element-binding proteins activates sterol-responsive genes. *Journal of lipid research*, 42(12):1939–46, December 2001.

[46] E Nagoshi, N Imamoto, R Sato, and Y Yoneda. Nuclear import of sterol regulatory element-binding protein-2, a basic helix-loop-helix-leucine zipper (bHLH-Zip)-containing transcription factor, occurs through the direct interaction of importin beta with HLH-Zip. *Molecular biology of the cell*, 10(7):2221–33, July 1999.

[47] R Sato, W Miyamoto, J Inoue, T Terada, T Imanaka, and M Maeda. Sterol regulatory element-binding protein negatively regulates microsomal triglyceride transfer protein gene transcription. *The Journal of biological chemistry*, 274(35):24714–20, August 1999.

[48] A Bist, PE Fielding, and CJ Fielding. Two sterol regulatory element-like sequences mediate up-regulation of caveolin gene transcription in response to low density lipoprotein free cholesterol. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10693–8, September 1997.

[49] V Lecomte, E Meugnier, V Euthine, C Durand, D Freyssenet, G Nemoz, S Rome, H Vidal, and E Lefai. A new role for sterol regulatory element binding protein 1 transcription factors in the regulation of muscle mass and muscle cell differentiation. *Molecular and cellular biology*, 30(5):1182–98, March 2010.

[50] MK Bennett and TF Osborne. Nutrient regulation of gene expression by the sterol regulatory element binding proteins: increased recruitment of gene-specific coregulatory factors and selective hyperacetylation of histone H3 in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):6340–4, June 2000.

[51] F Yang, BW Vought, JS Satterlee, AK Walker, Z-Y Jim Sun, JL Watts, R DeBeaumont, RM Saito, SG Hyberts, S Yang, C Macol, L Iyer, R Tjian, S van den Heuvel, AC Hart, G Wagner, and AM Näär. An ARC/Mediator subunit required for SREBP control of cholesterol and lipid homeostasis. *Nature*, 442(7103):700–4, August 2006.

[52] J Lin, R Yang, PT Tarr, P-H Wu, C Handschin, S Li, W Yang, L Pei, M Uldry, P Tontonoz, CB Newgard, and BM Spiegelman. Hyperlipidemic effects of dietary saturated fats mediated through PGC-1beta coactivation of SREBP. *Cell*, 120(2):261–73, January 2005.

[53] J Ericsson, A Usheva, and PA Edwards. YY1 is a negative regulator of transcription of three sterol regulatory element-binding protein-responsive genes. *The Journal of biological chemistry*, 274(20):14508–13, May 1999.

[54] MK Bennett, TT Ngo, JN Athanikar, JM Rosenfeld, and TF Osborne. Co-stimulation of promoter for low density lipoprotein receptor gene by sterol regulatory element-binding protein and Sp1 is specifically disrupted by the yin yang 1 protein. *The Journal of biological chemistry*, 274(19):13025–32, May 1999.

[55] A Dávalos, L Goedeke, P Smibert, CM Ramírez, NP Warrier, U Andreo, D Cirera-Salinas, K Rayner, U Suresh, JC Pastor-Pareja, E Esplugues, EA Fisher, LOF Penalva, KJ Moore, Y Suárez, EC Lai, and C Fernández-Hernando. miR-33a/b contribute to the regulation of fatty acid metabolism and insulin signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9232–7, May 2011.

[56] V Rottiers and AM Näär. MicroRNAs in metabolism and metabolic disorders. *Nature reviews. Molecular cell biology*, 13(4):239–50, April 2012.

[57] KJ Rayner, Y Suárez, A Dávalos, S Parathath, ML Fitzgerald, N Tamehiro, EA Fisher, KJ Moore, and C Fernández-Hernando. MiR-33 contributes to the regulation of cholesterol homeostasis. *Science (New York, N.Y.)*, 328(5985):1570–3, June 2010.

[58] T Horie, K Ono, M Horiguchi, H Nishi, T Nakamura, K Nagao, M Kinoshita, Y Kuwabara, H Marusawa, Y Iwanaga, K Hasegawa, M Yokode, T Kimura, and T Kita. MicroRNA-33 encoded by an intron of sterol regulatory element-binding protein 2 (Srebp2) regulates HDL in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 107(40):17321–6, October 2010.

[59] I Gerin, LA Clerbaux, O Haumont, N Lanthier, AK Das, CF Burant, IA Leclercq, OA MacDougald, and GT Bommer. Expression of miR-33 from an SREBP2 intron inhibits cholesterol export and fatty acid oxidation. *The Journal of biological chemistry*, 285(44):33652–61, October 2010.

[60] JR Krycer, LJ Sharpe, W Luu, and AJ Brown. The Akt-SREBP nexus: cell signaling meets lipid metabolism. *Trends in endocrinology and metabolism: TEM*, 21(5):268–76, May 2010.

[61] CR Yellaturu, X Deng, LM Cagen, HG Wilcox, CM Mansbach, SA Siddiqi, EA Park, R Raghow, and MB Elam. Insulin enhances post-translational processing of nascent SREBP-1c by promoting its phosphorylation and association with COPII vesicles. *The Journal of biological chemistry*, 284(12):7518–32, March 2009.

[62] MT Bengoechea-Alonso and J Ericsson. A phosphorylation cascade controls the degradation of active SREBP1. *The Journal of biological chemistry*, 284(9):5885–95, February 2009.

[63] K Düvel, JL Yecies, S Menon, P Raman, AI Lipovsky, AL Souza, E Triantafellow, Q Ma, R Gorski, S Cleaver, MG Vander Heiden, JP MacKeigan, PM Finan, CB Clish, LO Murphy, and BD Manning. Activation of a metabolic gene regulatory network downstream of mTOR complex 1. *Molecular cell*, 39(2):171–83, July 2010.

[64] JL Owen, Y Zhang, S-H Bae, MS Farooqi, G Liang, RE Hammer, JL Goldstein, and MS Brown. Insulin stimulation of SREBP-1c processing in transgenic rat hepatocytes requires p70 S6-kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16184–9, October 2012.

[65] WJ Quinn and MJ Birnbaum. Distinct mTORC1 pathways for transcription and cleavage of SREBP-1c. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):15974–5, October 2012.

[66] TR Peterson, SS Sengupta, TE Harris, AE Carmack, SA Kang, E Balderas, DA Guertin, KL Madden, AE Carpenter, BN Finck, and DM Sabatini. mTOR complex 1 regulates lipin 1 localization to control the SREBP pathway. *Cell*, 146(3):408–20, August 2011.

[67] G Chen, G Liang, J Ou, JL Goldstein, and MS Brown. Central role for liver X receptor in insulin-mediated activation of Srebp-1c transcription and stimulation of fatty acid synthesis in liver. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11245–50, August 2004.

[68] N Dif, V Euthine, E Gonnet, M Laville, H Vidal, and E Lefai. Insulin activates human sterol-regulatory-element-binding protein-1c (SREBP-1c) promoter through SRE motifs. *The Biochemical journal*, 400(1):179–88, November 2006.

[69] T Yamamoto, H Shimano, N Inoue, Y Nakagawa, T Matsuzaka, A Takahashi, N Yahagi, H Sone, H Suzuki, H Toyoshima, and N Yamada. Protein kinase A suppresses sterol regulatory element-binding protein-1C expression via phosphorylation of liver X receptor in the liver. *The Journal of biological chemistry*, 282(16):11687–95, April 2007.

[70] M Lu and J Y-J Shyy. Sterol regulatory element-binding protein 1 is negatively modulated by PKA phosphorylation. *American journal of physiology. Cell physiology*, 290(6):C1477–86, June 2006.

[71] A Hagiwara, M Cornu, N Cybulski, P Polak, C Betz, F Trapani, L Terracciano, MH Heim, MA Rüegg, and MN Hall. Hepatic mTORC2 activates glycolysis and lipogenesis through Akt, glucokinase, and SREBP1c. *Cell metabolism*, 15(5):725–38, May 2012.

[72] AH Hasty, H Shimano, N Yahagi, M Amemiya-Kudo, S Perrey, T Yoshikawa, J Osuga, H Okazaki, Y Tamura, Y Iizuka, F Shionoiri, K Ohashi, K Harada, T Gotoda, R Nagai, S Ishibashi, and N Yamada. Sterol regulatory element-binding protein-1 is regulated by glucose at the transcriptional level. *The Journal of biological chemistry*, 275(40):31069–77, October 2000.

[73] J Hao, S Liu, S Zhao, Q Liu, X Lv, H Chen, Y Niu, and H Duan. PI3K/Akt pathway mediates high glucose-induced lipogenesis and extracellular matrix accumulation in HKC cells through regulation of SREBP-1 and TGF-β1. *Histochemistry and cell biology*, 135(2):173–81, February 2011.

[74] R Suzuki, K Lee, E Jing, SB Biddinger, JG McDonald, TJ Montine, S Craft, and CR Kahn. Diabetes and insulin in regulation of brain cholesterol metabolism. *Cell metabolism*, 12(6):567–79, December 2010.

[75] X Zhao, D Feng, Q Wang, A Abdulla, X-J Xie, J Zhou, Y Sun, ES Yang, LP Liu, B Vaitheesvaran, L Bridges, IJ Kurland, R Strich, JQ Ni, C Wang, J Ericsson, JE Pessin, J-Y Ji, and F Yang. Regulation of lipogenesis by cyclin-dependent kinase 8-mediated control of SREBP-1. *The Journal of clinical investigation*, 122(7):2417–27, July 2012.

[76] R Dentin, JP Pégorier, F Benhamed, F Foufelle, P Ferré, V Fauveau, MA Magnuson, J Girard, and C Postic. Hepatic glucokinase is required for the synergistic action of ChREBP and SREBP-1c on glycolytic and lipogenic gene expression. *The Journal of biological chemistry*, 279(19):20314–26, May 2004.

[77] R Dentin, J Girard, and C Postic. Carbohydrate responsive element binding protein (ChREBP) and sterol regulatory element binding protein-1c (SREBP-1c): two key regulators of glucose metabolism and lipid synthesis in liver. *Biochimie*, 87(1):81–6, January 2005.

[78] X Chen, J Shen, and R Prywes. The luminal domain of ATF6 senses endoplasmic reticulum (ER) stress and causes translocation of ATF6 from the ER to the Golgi. *The Journal of biological chemistry*, 277(15):13045–52, April 2002.

[79] L Zeng, M Lu, K Mori, S Luo, AS Lee, Y Zhu, and JYJ Shyy. ATF6 modulates SREBP2-mediated lipogenesis. *The EMBO journal*, 23(4):950–8, February 2004.

[80] Y Takeuchi, N Yahagi, Y Izumida, M Nishi, M Kubota, Y Teraoka, T Yamamoto, T Matsuzaka, Y Nakagawa, M Sekiya, Y Iizuka, K Ohashi, J Osuga, T Gotoda, S Ishibashi, K Itaka, K Kataoka, R Nagai, N Yamada, T Kadowaki, and H Shimano. Polyunsaturated fatty acids selectively suppress sterol regulatory element-binding protein-1 through proteolytic processing and autoloop regulatory circuit. *The Journal of biological chemistry*, 285(15):11681–91, April 2010.

[81] J Ou, H Tu, B Shan, A Luk, RA DeBose-Boyd, Y Bashmakov, JL Goldstein, and MS Brown. Unsaturated fatty acids inhibit transcription of the sterol regulatory element-binding protein-1c (SREBP-1c) gene by antagonizing ligand-dependent activation of the LXR. *Proceedings of the National Academy of Sciences of the United States of America*, 98(11):6027–32, May 2001.

[82] T Yoshikawa, H Shimano, N Yahagi, T Ide, M Amemiya-Kudo, T Matsuzaka, M Nakakuki, S Tomita, H Okazaki, Y Tamura, Y Iizuka, K Ohashi, A Takahashi, H Sone, J Osuga Ji, T Gotoda, S Ishibashi, and N Yamada. Polyunsaturated fatty acids suppress sterol regulatory element-

binding protein 1c promoter activity by inhibition of liver X receptor (LXR) binding to LXR response elements. *The Journal of biological chemistry*, 277(3):1705–11, January 2002.

[83] X Deng, LM Cagen, HG Wilcox, EA Park, R Raghow, and MB Elam. Regulation of the rat SREBP-1c promoter in primary rat hepatocytes. *Biochemical and biophysical research communications*, 290(1):256–62, January 2002.

[84] JN Lee, X Zhang, JD Feramisco, Y Gong, and J Ye. Unsaturated fatty acids inhibit proteasomal degradation of Insig-1 at a postubiquitination step. *The Journal of biological chemistry*, 283(48):33772–83, November 2008.

[85] J Inoue, Y Ito, S Shimada, S-I Satoh, T Sasaki, T Hashidume, Y Kamoshida, M Shimizu, and R Sato. Glutamine stimulates the gene expression and processing of sterol regulatory element binding proteins, thereby increasing the expression of their target genes. *The FEBS journal*, 278(15):2739–50, August 2011.

[86] B Ponugoti, D-H Kim, Z Xiao, Z Smith, J Miao, M Zang, S-Y Wu, C-M Chiang, TD Veenstra, and JK Kemper. SIRT1 deacetylates and inhibits SREBP-1C activity in regulation of hepatic lipid metabolism. *The Journal of biological chemistry*, 285(44):33959–70, October 2010.

[87] AK Walker, F Yang, K Jiang, J-Y Ji, JL Watts, A Purushotham, O Boss, ML Hirsch, S Ribich, JJ Smith, K Israelian, CH Westphal, JT Rodgers, T Shioda, SL Elson, P Mulligan, H Najafi-Shoushtari, JC Black, JK Thakur, LC Kadyk, JR Whetstine, R Mostoslavsky, P Puigserver, X Li, NJ Dyson, AC Hart, and AM Näär. Conserved role of SIRT1 orthologs in fasting-dependent inhibition of the lipid/cholesterol regulator SREBP. *Genes & development*, 24(13):1403–17, July 2010.

[88] A Defour, K Dessalle, A Castro Perez, T Poyot, J Castells, YS Gallot, C Durand, V Euthine, Y Gu, D Béchet, A Peinnequin, E Lefai, and D Freyssenet. Sirtuin 1 regulates SREBP-1c expression in a LXR-dependent manner in skeletal muscle. *PloS one*, 7(9):e43490, January 2012.

[89] Y Li, S Xu, MM Mihaylova, B Zheng, X Hou, B Jiang, O Park, Z Luo, E Lefai, J Y-J Shyy, B Gao, M Wierzbicki, TJ Verbeuren, RJ Shaw, RA Cohen, and M Zang. AMPK phosphorylates and inhibits SREBP activity to attenuate hepatic steatosis and atherosclerosis in diet-induced insulin-resistant mice. *Cell metabolism*, 13(4):376–88, April 2011.

[90] NB Ruderman, XJ Xu, L Nelson, JM Cacicedo, AK Saha, F Lan, and Y Ido. AMPK and SIRT1: a long-standing partnership? *American journal of physiology. Endocrinology and metabolism*, 298(4):E751–60, April 2010.

[91] Y Zhang, T Lei, JF Huang, SB Wang, LL Zhou, ZQ Yang, and XD Chen. The link between fibroblast growth factor 21 and sterol regulatory element binding protein 1c during lipogenesis in hepatocytes. *Molecular and cellular endocrinology*, 342(1-2):41–7, August 2011.

[92] M Xia, Y Liu, H Guo, D Wang, Y Wang, and W Ling. Retinol binding protein 4 stimulates hepatic sterol regulatory element-binding protein 1 and increases lipogenesis through the peroxisome proliferator-activated receptor-γ coactivator 1β-dependent pathway. *Hepatology (Baltimore, Md.)*, 58(2):564–75, August 2013.

[93] R Li, W Chen, Y Li, Y Zhang, and G Chen. Retinoids synergized with insulin to induce Srebp-1c expression and activated its promoter via the two liver X receptor binding sites that mediate insulin action. *Biochemical and biophysical research communications*, 406(2):268–72, March 2011.

[94] Y-K Seo, T-I Jeon, HK Chong, J Biesinger, X Xie, and TF Osborne. Genome-wide localization of SREBP-2 in hepatic chromatin predicts a role in autophagy. *Cell metabolism*, 13(4):367–75, April 2011.

[95] AB Castoreno, Y Wang, W Stockinger, LA Jarzylo, H Du, JC Pagnon, EC Shieh, and A Nohturfft. Transcriptional regulation of phagocytosis-induced membrane biogenesis by sterol regulatory element binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13129–34, September 2005.

[96] L Gurcel, L Abrami, S Girardin, J Tschopp, and FG van der Goot. Caspase-1 activation of lipid metabolic pathways in response to bacterial pore-forming toxins promotes cell survival. *Cell*, 126(6):1135–45, September 2006.

[97] S-S Im and TF Osborne. Protection from bacterial-toxin-induced apoptosis in macrophages requires the lipogenic transcription factor sterol regulatory element binding protein 1a. *Molecular and cellular biology*, 32(12):2196–202, June 2012.

[98] AL Hughes, BL Todd, and PJ Espenshade. SREBP pathway responds to sterols and functions as an oxygen sensor in fission yeast. *Cell*, 120(6):831–42, March 2005.

[99] MT Bengoechea-Alonso and J Ericsson. Cdk1/cyclin B-mediated phosphorylation stabilizes SREBP1 during mitosis. *Cell cycle (Georgetown, Tex.)*, 5(15):1708–18, August 2006.

[100] N Inoue, H Shimano, M Nakakuki, T Matsuzaka, Y Nakagawa, T Yamamoto, R Sato, A Takahashi, H Sone, N Yahagi, H Suzuki, H Toyoshima, and N Yamada. Lipid synthetic transcription factor SREBP-1a activates p21WAF1/CIP1, a universal cyclin-dependent kinase inhibitor. *Molecular and cellular biology*, 25(20):8938–47, October 2005.

[101] D Cirera-Salinas, M Pauta, RM Allen, AG Salerno, CM Ramírez, A Chamorro-Jorganes, AC Wanschel, Miguel A Lasuncion, M Morales-Ruiz, Y Suarez, Á Baldan, E Esplugues, and C Fernández-Hernando. Mir-33 regulates cell proliferation and cell cycle progression. *Cell cycle (Georgetown, Tex.)*, 11(5):922–33, March 2012.

[102] W Shao and PJ Espenshade. Expanding roles for SREBP in metabolism. *Cell metabolism*, 16(4):414–9, October 2012.

[103] ME Adriaens, M Jaillard, A Waagmeester, SLM Coort, AR Pico, and CTA Evelo. The public road to high-quality curated biological pathways. *Drug discovery today*, 13(19-20):856–62, October 2008.

[104] DGJ Jennen, S Gaj, PJ Giesbertz, JHM van Delft, CT Evelo, and JCS Kleinjans. Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug discovery today*, 15(19-20):851–8, October 2010.

[105] MP van Iersel, T Kelder, AR Pico, K Hanspers, S Coort, BR Conklin, and C Evelo. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics*, 9:399, January 2008.

Chapter 4

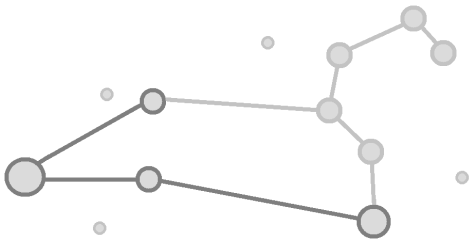
Multi-omics Data Visualization on Pathways

This chapter is based on the following publication:

Martijn P van Iersel[1,2], Milka Sokolovic[3,4], Kaatje Lenaerts[5], Martina Kutmon[1], Freek G Bouwman[6], Wouter H Lamers[7], Edwin CM Mariman[6], Chris T Evelo[1]

1. Department of Bioinformatics - BiGCaT, NUTRIM School for Nutrition, Toxicology and Metabolism, Maastricht University, The Netherlands
2. General Bioinformatics, Reading, UK
3. Department of Medical Biochemistry, Academic Medical Centre, University of Amsterdam, The Netherlands
4. European Food Information Council, Brussels, Belgium
5. Department of Surgery, Maastricht University, The Netherlands
6. Department of Human Biology, Maastricht University, The Netherlands
7. Tytgat Institute for Liver and Intestinal Research, Academic Medical Center, University of Amsterdam, The Netherlands

Journal of Integrative Bioinformatics (2014) 11(1):235



Preface

Chapter 4 is included in this thesis to show how different omics datasets can be integrated, analyzed and visualized together using pathway analysis in PathVisio. In the described study, we are using transcriptomics and proteomics data from a mouse study to validate how the integration of different kinds of experimental data can enhance our understanding of complex biological processes.

Proteomics and metabolomics technologies are improving rapidly and soon more large scale measurements of the proteome and metabolome will be available. As shown in Figure 1.1 in the General Introduction, proteins and metabolites are the functional elements in a biological system. Transcriptomics and proteomics levels do not always correlate because of post-transcriptional and translational regulation. Therefore it is crucial to include all different measurements when studying biological processes in all their complexity.

Abstract

Our understanding of complex biological processes can be enhanced by combining different kinds of high-throughput experimental data, but the use of incompatible identifiers makes data integration a challenge. We aimed to improve methods for integrating and visualizing different types of omics data. To validate these methods, we applied them to two previous studies on starvation in mice, one using proteomics and the other using transcriptomics technology.

We extended the PathVisio software with new plugins to link proteins, transcripts and pathways. A low overall correlation between proteome and transcriptome data was detected (Spearman rank correlation: 0.21). At the level of individual genes, correlation was highly variable. Many mRNA/protein pairs, such as fructose biphosphate aldolase B and ATP Synthase, show good correlation. For other pairs, such as ferritin and elongation factor 2, an interesting effect is observed, where mRNA and protein levels change in opposite directions, suggesting they are not primarily regulated at the transcriptional level. We used pathway diagrams to visualize the integrated datasets and found it encouraging that transcriptomics and proteomics data supported each other at the pathway level.

Visualization of the integrated dataset on pathways led to new observations on gene-regulation in the response of the gut to starvation. Our methods are generic and can be applied to any multi-omics study. The PathVisio software can be obtained at http://www.pathvisio.org.

Introduction

The intestine plays an important role in the response of the body to (prolonged) starvation. In two previous publications, the transcriptome [1] and the proteome [2] of the murine intestine were studied after 0, 12, 24 and 72 hours of starvation. In the present study we combine and compare data from the two earlier studies. The goal is twofold. First, to develop bioinformatics tools needed to make an integrated omics approach feasible. Second, to get a more comprehensive view of regulation of pathways involved in the starvation response of the gut.

"Omics" Integration

Microarray technology can be used to measure the expression of thousands of genes at the same time. Methods for processing, analyzing and interpreting microarray data are well established, and off-the-shelf microarrays are available with enough capacity to measure the majority of the genes in a genome.

Not transcripts but proteins are directly involved in biological activities. Microarray studies usually assume implicitly that there is a correlation between transcript and protein abundance, but only moderate levels of correlation have been reported [3–7]. A lack of correlation could have several causes such as variation in protein turnover rates and post-transcriptional regulation [8]. This lack of correlation between transcript and protein levels is an important argument for measuring protein expression directly. Two-dimensional gel electrophoresis (2DE), still one of the most common techniques for quantitative proteomics, has continued to mature in recent years and has achieved higher standards of data quality, reproducibility and protein identification [9].

Proteomic technologies have, nevertheless, a number of disadvantages. In a standard 2DE proteomics experiment in mammals, typically only 100 proteins are identified and measured, or roughly 0.5% of the genome, far less than what is typical for microarray studies. Moreover, 2DE studies suffer from problems of bias. Proteins vary more widely than mRNA molecules in physical properties such as hydrophobicity, electric charge and size. The subset of measured proteins is biased towards proteins that are easily separable by 2DE and abundant enough to be identified. Furthermore, the spots that show the clearest response to experimental conditions are picked first. Protein identification is, therefore, a bottleneck in 2DE. New gel-free proteomics techniques measure a wider range of protein chemistries and abundances and suffer less from bias problems [10]. In spite of these issues, 2DE remains commonly used for its maturity and relative simplicity [9].

As both proteomics and transcriptomics methods have their advantages, it could be beneficial to combine them. It appears that only some genes are primarily regulated at the transcriptional level, showing higher transcript/protein correlation than what would be expected from global correlation levels [3], whereas other genes are regulated post-transcriptionally, showing much lower correspondence. Therefore, we believe that integrated analysis of omics datasets must take into

account a-priori knowledge about the regulation of genes and proteins, for example in the form of pathway diagrams. Pathway diagrams contain biological context, such as which entities are related, what is their cellular location, which proteins are interaction partners, and what is the nature of those interactions (stimulation or repression). By visualizing the combined dataset on a pathway diagram one can interpret the biological context more easily. As part of this study we aimed to develop easy-to-use software to make this possible.

Regulation of the Intestinal Response to Fasting

We combined two experimental datasets to validate our data integration methodology. The experimental data relates to the intestinal changes in response to fasting. A better understanding of fasting could lead to better understanding of malnourishment and better treatment of cachexia (wasting syndrome) caused by chronic disease [11].

Based on the rate of weight loss, nitrogen excretion, concentration of plasma metabolites and resting metabolic rate, the body passes through three successive adaptive phases during fasting, often defined as postabsorptive, coping and preterminal [12]. In mammals, these phases have been associated with the primary fuel that is putatively available to the tissues [13–16]. Based on whole-body energy expenditure, the "sugars-fats-proteins" succession of energy substrates during fasting was proposed [11, 13], and this model was then extrapolated to all organs separately. However, transcriptomic studies in rodents that have prospected the adaptive response to fasting of different organs [1, 17–22] reveal a different scenario. These studies have shown that already in the postabsorptive phase organs liberally increase protein and lipid catabolism, fuelling hepatic and renal gluconeogenesis and ketogenesis. Only during prolonged fasting ($\geq$ 24h in mice) is protein catabolism minimized.

It is clear that the various organs (liver, intestine, adipose tissue, muscle, kidney, brain) play different roles in this response, but the whole picture is not yet completely elucidated. The small intestine contributes to the bodys adaptation to fasting by a biphasic response of carbohydrate metabolism, which peaks during short and again after prolonged fasting [1]. Early changes are associated with glutamine conservation, inhibition of pyruvate oxidation, stimulation of glutamate catabolism via aspartate and phosphoenolpyruvate to lactate, and enhanced fatty-acid oxidation and ketone-body synthesis. Changes upon continued fasting implied the production of glucose rather than lactate from carbohydrate backbones and downregulation of fatty-acid oxidation. In addition, cell turnover is progressively downregulated by inhibiting cell cycling and apoptosis to reduce the high energy costs of constant enterocyte turnover [1].

The question that this study tries to answer is whether protein expression data reinforce the picture that arises from transcriptomics pathway analysis, and to what extent post-transcriptional regulation plays a role in these pathways.

Methods

The experimental procedure for treatment of animals, microarray and 2DE was described before [1] [2], but is briefly summarized here.

Animals. Male FVB mice from Charles River (Maastricht, The Netherlands) were housed at 20-22°C, 50-60% humidity on a 12 hours light/dark cycle. They ingested food and water ad libitum until the age of 6 weeks. Groups of 6 mice were fasted for 0, 12, 24, and 72 hours, after which the animals were killed by cervical dislocation. The small intestine was removed and separated from adjacent tissue, and both protein and RNA were isolated. The same mice were used for both microarray and proteomics experiments.

Microarrays. Samples of the intestine were applied to 60-mer Mouse Development 22k Oligo Microarray G4120A (Agilent). Three arrays per experimental condition were used. Per microarray, 20 μg mRNA, pooled from 2 intestines, was labeled with Cy3. RNA pooled from 6 fed animals was labeled with Cy5 and used as a common reference across all arrays. After hybridization the arrays were scanned with Agilent's dual-laser microarray slide scanner and processed with Agilents Feature Extraction software 6.1.1. Quantile normalization was applied to background-subtracted intensities.

2DE. The procedure for generating 2D protein gel images was as described before [23]. Protein samples were isolated from equal quantities of proximal and distal parts of the intestine, pooled per mouse. 1 gel was made for each mouse. 100 μg of total protein was separated by isoelectric focusing using IPG strips, and then placed onto 12.5% SDS-polyacrylamide gels for protein separation in the second dimension. Gels stained with SYPRO Ruby Protein stain were scanned with the Molecular Imager FX (Bio-Rad Laboratories). Analysis of differentially expressed proteins was performed using PDQuest 7.3 (Bio-Rad Laboratories). A number of spots were selected for identification, with a preference for spots with a significant intensity difference. Selected spots were excised and subjected to tryptic in-gel digestion and MALDI-TOF MS (Waters, Manchester, UK), generating peptide mass fingerprints which were subsequently identified using the MASCOT search engine against the SwissProt database.

Microarray annotation. The microarray type used was the 60-mer Mouse Development 22k Oligo Microarray G4120A (Agilent). The microarray contains 20,280 probes of 60 nucleotides each. The annotation file provided by Agilent (Version of Dec.16 2009) associates only 9,616 probes to Ensembl gene identifiers. The probes were designed based on a 5-year old genome build, which could have diverged in the intervening years. On the assumption that 60-mer probes do not require a complete sequence match to hybridize to transcripts, we investigated if a BLAST search with less stringent settings would increase probe annotation coverage. We selected the best BLAST hits against a more recent Ensembl (release 57) Mouse cDNA database, with a minimum e value of 1.0e-6. The BLAST resulted in annotation for 10,696 probes, an increase of 11%. We opted to employ the BLAST

results instead of Agilent annotations for all further analysis. Identifiers in both data sets were mapped to pathways using the BridgeDb framework [24]. Because none of the standard identifier mapping resources included in BridgeDb contained mappings for the Agilent array, we prepared a custom mapping table that allowed proper interpretation of the BLAST results by BridgeDb.

Analysis and Pathway visualization. Correlation and expression plots were created with R/BioConductor. Pathway visualization was performed using the pathway editing and visualization tool PathVisio which was first published in 2008 [25]. We developed two new plugins for PathVisio: the Gex plugin, which manages data import and visualization, and the BridgeDb-Config plugin, which enables configuration of identifier mapping resources.

The Gex plugin handles the import of expression data in PathVisio and the mapping of the provided identifiers in the dataset to the identifiers used in the pathways. This plugin is now a core module and does not need to be installed separately. To enable more advanced options of identifier mapping, e.g. the usage of different identifier mapping tables together, the BridgeDb-Config plugin was developed. The HTML export plugin allows the export of a single pathway diagram as an HTML page as well as the export of a complete pathway statistics result including the colored pathway diagrams. Since PathVisio 3, the plugin manager provides a fast and simple way to install plugins from a central plugin repository. The BridgeDbConfig plugin and the HTML export plugin are available in the central plugin repository since March 2013.

The mouse pathway set, obtained from WikiPathways [26] March 2010, was used for visualization. This pathway set covers 3,975 unique genes, or 14% of the whole mouse genome (using all genes and pseudo genes of Ensembl release 57 as the reference). We counted the overlap between the list of measured genes and proteins and the list of genes occurring in at least one pathway. 66% of measured proteins occur in at least one pathway (51 out of the 77 unique measured proteins). 24% of measured genes occur in at least one pathway (2,083 out of 8,648 unique measured genes).

Results

Identifier Mapping

A prerequisite for integration of multiple omics studies is the ability to map identifiers from various sources [27]. Each data point in the microarray dataset was identified with an Agilent probe identifier (such as A 65 P03556), and each protein was associated with a UniProt identifier (such as P09528). One of the difficulties in identifier mapping is that there is a one-to-many relation between genes and measurements. Because microarray designs often include some redundancy, there could be more than one probe identifier per gene. Similarly, one protein might give rise to multiple spots on the 2D gel, depending on the protein modification

state [9], so there are frequently multiple spot numbers per gene.

The BridgeDb framework [24] was used to map probe and protein identifiers to gene identifiers and integrate the two datasets. This framework was used both in the PathVisio pathway visualization software [25] and in the correlation analysis in R.
Three sources of information were used for identifier mapping. By using BridgeDb, the three resources were unified into a single mapping resource. First, the results of BLAST of microarray sequences for mapping Agilent probes to genes in the transcriptomics dataset; second, a regular BridgeDerby database for mapping proteins to genes; and third, a manual override table to fix three errors in the protein dataset.

Three proteins were annotated with erroneous protein identifiers by the identification software (in one case a GenBank accession number, in another an identifier for the human homologue, and in the third a deprecated UniProt identifier). Rather than fixing the original data, BridgeDb allowed us to create a separate "manual override" mapping table and combine it with the rest of the mapping resources. The advantage of doing so, rather than simply fixing up the original dataset, was that the modifications remained separated and could be re-examined and adjusted later.

Correlation

The proteomics dataset contained 130 identified spots. Since more than one spot may arise from the same protein in different modification states, those 130 spots corresponded to only 77 unique protein identifiers. Moreover, due to limitations of the microarray used, for some of them no mRNA data was available, resulting in only 59 having both the mRNA and protein abundance determined. For each of the 59 pairs, the base-2 logarithm (log2) of the ratio relative to 0 hours was calculated for 12, 24 and 72 hours of fasting, and correlations between the two types of data were calculated. Taken together, there was very little overall correlation, with a Spearman rank correlation coefficient of 0.21 (Figure 4.1). The positive value of the coefficient nevertheless points to a slight positive overall correlation between changes in mRNA and protein levels.
The picture was very different for specific genes. There were highly correlating, differentially expressed transcript-protein pairs, such as *Aldob* and *Atp5h* (Figure 4.1) and *Vim* (not shown). Others, on the other hand, showed a negative correlation, with the transcript changing in the opposite direction of the protein. In most of these cases, the transcript was up- and the corresponding protein down-regulated. Examples are *Eef2* (elongation factor 2, Figure 4.1), *Aldh1b1*, *Arhgdia*, *Hnrnpa2b1* and *Uqcrc1*. For *Eef2*, a similar divergence of mRNA and protein levels upon fasting was reported earlier in liver and muscle [28].

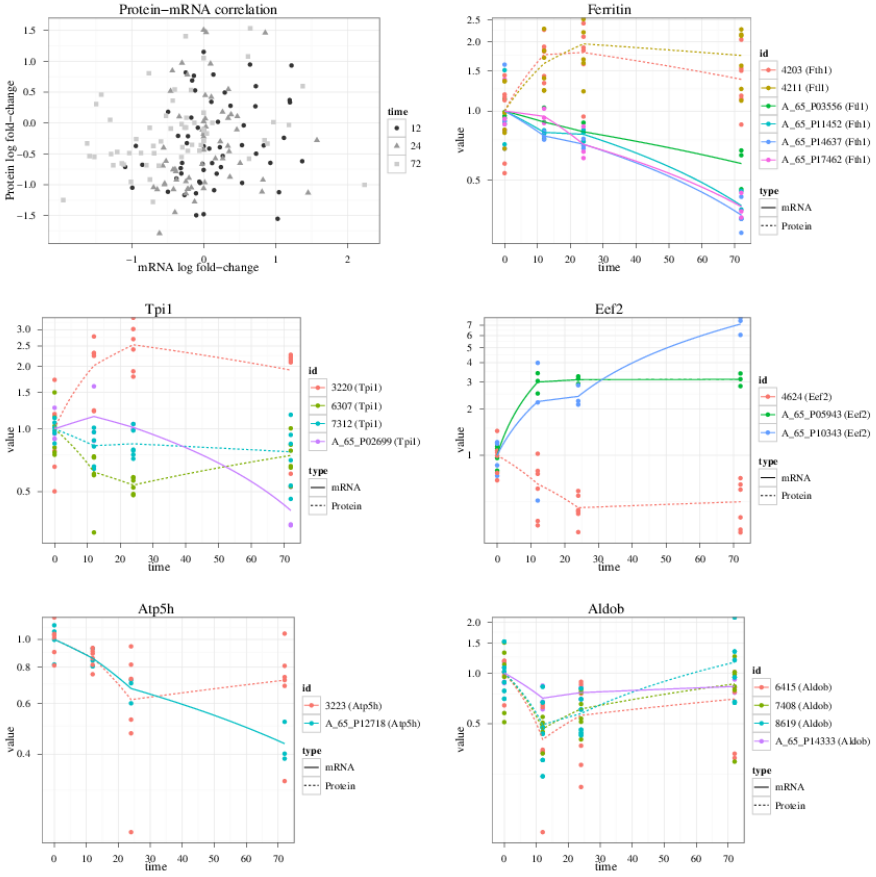


Figure 4.1: Combined mRNA and Protein Expression Plots. In the top-left plot, the fold-change of protein and mRNA expression are plotted against each other. Fold-changes are calculated for each time point against t=0. In cases where multiple protein spots correspond to the same gene, the average was used. The overall correlation plot shows that there was very little agreement between protein and gene expression. The five other plots show individual genes in detail. Each dot represents a measurement of a spot or probe. Lines connect the average of each probe or spot per time point. Solid lines represent transcripts, dashed lines represent proteins. The average intensities at 0 hours have been normalized to 1; all other values are relative to this average. Top-right: Ferritin heavy (*Fth1*) and light chain (*Ftl1*), showing opposite trends for transcript and protein expression levels. Center-left: Triose phosphate isomerase (*Tpi1*), which is down-regulated at 72 hours with the exception of one protein spot. Center-right: Elongation Factor 2 (*Eef2*), an example of a gene that shows opposite effects of fasting on transcript and protein expression. Bottom-left: ATP synthase D chain (*Atp5h*) and bottom-right: fructose biphosphate aldolase B (*Adlob*), are examples of two genes that show good correlation between protein and transcript levels.

Some of the proteins identified in multiple spots showed variable expression of different isoforms, indicating that post-translational modifications could play a role in regulation of their activity. The proteins that were potentially post-translationally modified included metabolic enzymes (TPI1 and ATP5H), proteins related to protein folding (CALR, HSPA8 and HSPA5) and cytoskeletal proteins (KRT19, ACTB, ACTG2 and VIL1).

Pathway Visualization

To visualize protein and gene expression data side-by-side in pathway context, we used the PathVisio program [25]. High-throughput datasets in tab-delimited text format are imported using the expression data import wizard plugin. Data should be normalized and preferably log-transformed before import.
During data import, the user can select a column that contains gene or protein identifiers. PathVisio allows direct import of mixed identifiers without the need to pre-process the data. After the import step, the user can configure the color representation of the data. PathVisio provides rule-based and color gradient-based visualization options. In the rule-based visualization the user defines a Boolean expression to specify a color for elements that evaluate as true. The gradient-based visualization maps numerical values to a color gradient. Colors are displayed in the gene boxes. Multiple conditions can be displayed side-by-side, and asterisks can be added to the diagram to indicate significantly changed measurements.

A feature of particular interest for omics integration is the fact that if multiple data points map to the same box, it can be divided horizontally for each corresponding row in the dataset. Thus, the box is used as a small heat map representation of a subset of the data, where each column represents a condition, and each row represents a probe. The problem of a variable number of data points per box occurs when visualizing data from microarrays that have more than one probe per gene, but it is especially important for omics integration, which often deals with two datasets of very unequal size (in this case 130 proteins versus 10,696 transcripts), which automatically means that some boxes have more data than others. Although this subdivision means that the boxes can get cramped, we find this approach superior to summarization, which inevitably means discarding data. In the example of TPI1 (Figure 4.2), the two rows are clearly differentially expressed, but this insight would be lost if the average value of these measurements was used. If needed, the boxes can be manually enlarged. Data can be visualized together for direct comparison, but it is also possible to create separate visualizations and toggle between them using a drop-down list. Visualizing different time points separately can prevent confusion and decrease the chance of an erroneous comparison of a gene at one time point with another gene at a different time point. Separating the time points into different visualizations also helps to combat the information overload in a single box.
The resulting images of pathways with overlayed data can be exported as a set of HTML image maps, which can be viewed in any browser without the need to install PathVisio.

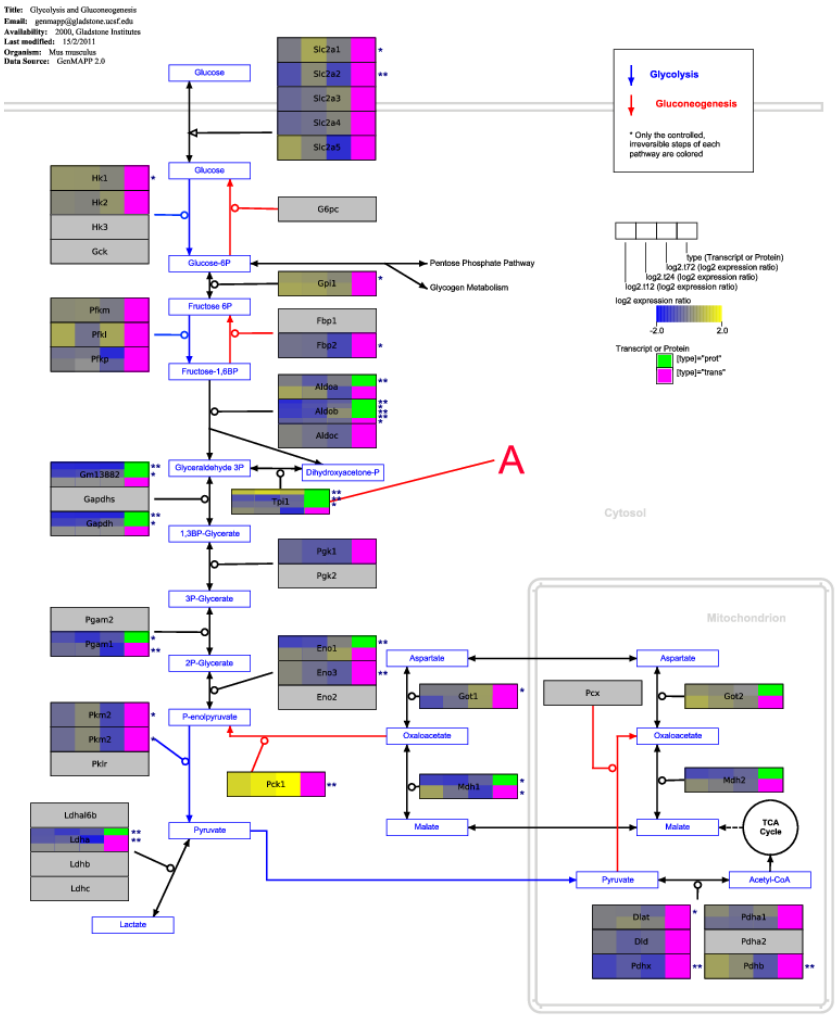


Figure 4.2: Glycolysis and Gluconeogenesis. Visualization of expression data on the glycolysis / gluconeogenesis pathway. Each colored box represents a gene product. Blue indicates decreased expression levels, yellow increased. Each box is a heat map with rows representing probes or spots, and columns representing time points. The rightmost column is a flag that indicates if the given row is a protein spot (green) or microarray probe (pink). The asterisks denote significance. Multiple probes and/or spots can be shown in a box. *Tpi1* is marked with the letter A.

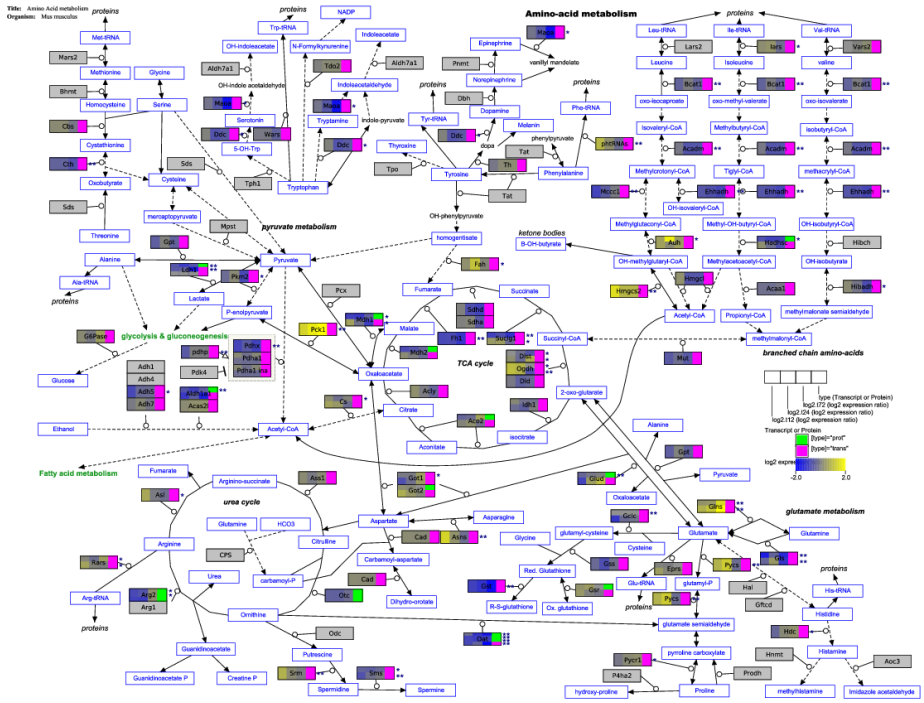


Figure 4.3: Amino Acid Metabolism. Visualization of expression data on the amino acid metabolism pathway. Coloring is identical to Figure 2. The genes *Otc2*, *Arg2* and *Oat*, mentioned in the main text, can be found in the bottom-left quadrant of the image. *Gls*, *Glns* and *Pycs* can be found in the bottom-right quadrant.

Integration at the Gene / Protein Level

The combination of proteomics and transcriptomics data provides us more information than just one of the two datasets on its own. Global correlation is not high, but on a gene-by-gene level we see a very varied picture. Gene/protein pairs that do not show high correlation present leads for investigation into post-transcriptional regulation.

Although proteomics data have fewer data points than transcriptomics data, there are a few instances where important transcripts were not measured, due to absence of a corresponding probe on the array. In those cases, protein measurements can fill important gaps in pathways. For example, no mRNA expression levels were measured for *Otc2* and *Arg2* due to absence of probes for these genes on the microarray, but their protein concentrations were measured and show a clear down-regulation, in particular in the early (12 hours) response. This is consistent with suppression of citrulline synthesis without an additional increase in arginine catabolism and with glutamine conservation via suppression of *Oat*, *Pycs* and *Gls*, and the upregulation of *Gln* (see Figure 4.3).
Similarly, the down-regulation of ACAA2 (only measured as protein) is completely consistent with the reduction of fatty acid biosynthesis, also supported by the

down-regulation of genes such as *Hadh1*. Additionally, down-regulation of the GAPDH protein, indicating an overall decrease in glycolytic and gluconeogenic activity, is entirely consistent with lower expression of other genes in the same pathway, such as *Pgam1* and *Eno3* (Figure 4.2).
Analysis of the transcriptome has revealed strong effects on cell cycle and apoptosis, by down-regulation of cyclins and upregulation of their inhibitors, cyclin-dependent kinases ([1]). Cell turnover in intestine is thought to be a major source of energy expenditure, and its (down)regulation in fasting is in good agreement with a need for energy preservation. Unfortunately, no cyclins or other proteins related to cell cycle regulation and apoptosis could be identified, most likely because their level of expression was too low for detection by the 2DE technique. The comparison of the two datasets underlines the problem of bias in proteomics techniques. Proteomics analysis alone would have missed a major regulatory effect of starvation in the gut.
The importance of glucose metabolizing pathway in fasting (Figure 4.2) is stressed by sheer number of genes differentially expressed. Out of 11 gene/protein couples found in the pathway, only in 4 both transcript and protein were significantly differentially expressed (*Pgam1*, *Ldha*, *Aldob* and *Mdh1*). In all 4, interestingly, the direction of the change was the same. The direction was also the same in case of *Got2* and *Mdh2*, in which changes occurred in both data types, but have not reached significance. For two glyceraldehyde dehydrogenases only changes in protein expression were detectable, while for *Aldoa* and *Eno1* the direction of the change differed between gene and protein. *Tpi1* had different direction of change between different protein isoforms (discussed in more detail below). Data integration and visualisation convey therefore a clear message that in fasting some of the proteins in this pathway must be regulated by other means than just transcription rate.

A difference in any of 500 known posttranslational modifications [29] (e.g. phosphorylation states) can lead to different spots in a 2D gel. Such a change could then mean either that the experimental condition has led to a change in total quantity of the corresponding protein, or to a change in the functional state of the protein (or both), but due to the incompleteness of proteomics technology, these two possibilities can hardly be distinguished. Unfortunately, a decrease at the spot level can therefore not be straightforwardly interpreted as a decrease in functional activity. A clear example of differentiation at the spot level is TPI1, an important enzyme of the gluconeogenesis pathway, which is increased in spot 3,220 but decreased in spots 6,307 and 7,312 throughout the fasting period (Figure 4.1). From the estimated mass as well as the mass spectrum it appears that the protein in spot 3,220 lacks an N-terminal fragment. Alternative splicing or proteolysis could play a role in the regulation of this protein. Assuming that the partial TPI1 protein has a reduced activity, this finding is consistent with a reduced activity in the glycolytic pathway. However, to determine the exact nature of the observed change in TPI1 and its consequence for enzyme activity, follow-up experiments are necessary.

Another protein affected in an interesting way was ferritin, a protein necessary for the storage of iron in tissues. Ferritin was down-regulated at the gene level,

but up-regulated at the protein level throughout starvation (Figure 4.1), and this was the case for both its heavy and light chain. Though dietary iron is unavailable in fasting, the prevention of iron release from the existing stores would limit its availability to the invading microorganisms, especially in the small intestine. The need for such a response is accentuated by the impressive downregulation of immune response seen in the intestine in our recent study [22]. The increased protein abundance in spite of lower transcript levels is harder to explain. It has been shown that mRNA turnover, regulated by the iron responsive element binding protein (*Ireb2*, not measured in this study), has a strong effect on ferritin [30], which could explain the discrepancy between protein and mRNA abundance. Another possible explanation is that the increased values are caused by a change occurring in erythrocytes, which can never be fully excluded from the protein sample. Other possibilities, such as the presence of other ferritin spots in the gel that have not yet been identified, cannot be ruled out.

Discussion

A number of existing applications can perform visualization of high-throughput datasets, such as KEGG Atlas [31], ProMeTra [32], Vanted [33] and Reactome SkyPainter [34], and reviewed in [35, 36]. Some of those have demonstrated the capability to perform pathway visualization with multiple omics datasets together in one pathway, such as ProMeTra, which is focused on combining metabolomics and transcriptomics data. The automated mapping of mixed identifiers is a distinguishing feature of PathVisio that makes it particularly suited for integration and simultaneous visualization of datasets from different sources.
Our examples showed that proteomics data can reinforce the conclusions deduced from transcriptomics data, and simultaneously indicate areas where post-transcriptional regulation plays a role. The 2DE technique by itself does not provide enough data for an overview on systems biology level. In this study, if only proteomics data had been available, important pathways such as apoptosis and cell cycle regulation would have been missed entirely. Nevertheless, protein expression data do provide interesting insights into regulation on a gene-by-gene basis. The proteins that do not correlate well are of particular interest, at the very least for generating hypotheses for follow-up experiments. With the maturation of proteomics (and metabolomics) technology, the issue of number of measurements will lose the impact and significance, only increasing the importance of being able to visualize the different datasets simultaneously.
The interpretation is not straightforward, and the correlation of gene and protein expression levels (or lack thereof) must be interpreted on a case-by-case basis. Pathway visualization can serve as a useful aid, given the role of pathways as a knowledge base of biological information. Flexible identifier mapping is essential for data integration. The relation between genes, microarray probes and protein spots is not straightforward, which makes software support for automated identifier mapping essential.

Bibliography

[1] M Sokolović, D Wehkamp, A Sokolović, J Vermeulen, LA Gilhuijs-Pederson, RIM van Haaften, Y Nikolsky, CTA Evelo, AHC van Kampen, TBM Hakvoort, and WH Lamers. Fasting induces a biphasic adaptive metabolic response in murine small intestine. *BMC genomics*, 8:361, January 2007.

[2] K Lenaerts, M Sokolović, FG Bouwman, WH Lamers, EC Mariman, and J Renes. Starvation induces phase-specific changes in the proteome of mouse small intestine. *Journal of proteome research*, 5(9):2113–22, September 2006.

[3] L Nie, G Wu, DE Culley, JCM Scholten, and W Zhang. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Critical reviews in biotechnology*, 27(2):63–75, 2007.

[4] L Anderson and J Seilhamer. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, 18(3-4):533–7, 1997.

[5] SP. Gygi, Y Rochon, BR Franza, and R Aebersold. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.*, 19(3):1720–1730, March 1999.

[6] T Ideker, V Thorsson, JA Ranish, R Christmas, J Buhler, JK Eng, R Bumgarner, DR Goodlett, R Aebersold, and L Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, 292(5518):929–34, May 2001.

[7] MP Washburn, A Koller, G Oshiro, RR Ulaszek, D Plouffe, C Deciu, E Winzeler, and JR Yates. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3107–12, March 2003.

[8] B Pradet-Balade, F Boulmé, H Beug, EW Müllner, and JA Garcia-Sanz. Translation control: bridging the gap between genomics and proteomics? *Trends in biochemical sciences*, 26(4):225–9, April 2001.

[9] F Chevalier. Highlights on the capacities of "Gel-based" proteomics. *Proteome science*, 8:23, January 2010.

[10] MR Roe and TJ Griffin. Gel-free mass spectrometry-based high throughput proteomics: tools for studying biological response of proteins and proteomes. *Proteomics*, 6(17):4678–87, September 2006.

[11] JM Argilés. Cancer-associated malnutrition. *European journal of oncology nursing : the official journal of European Oncology Nursing Society*, 9 Suppl 2:S39–50, January 2005.

[12] Y Le Maho, H Vu Van Kha, H Koubi, G Dewasmes, J Girard, P Ferré, and M Cagnard. Body composition, energy expenditure, and plasma metabolites in long-term fasting geese. *The American journal of physiology*, 241(5):E342–54, November 1981.

[13] M Caloin. Modeling of lipid and protein depletion during total starvation. *American journal of physiology. Endocrinology and metabolism*, 287(4):E790–8, October 2004.

[14] Y Cherel, D Attaix, D Rosolowska-Huszcz, R Belkhou, JP Robin, M Arnal, and Y Le Maho. Whole-body and tissue protein synthesis during brief and prolonged fasting in the rat. *Clinical science (London, England : 1979)*, 81(5):611–9, November 1991.

[15] Y Cherel and Y Le Maho. Refeeding after the late increase in nitrogen excretion during prolonged fasting in the rat. *Physiology & behavior*, 50(2):345–9, August 1991.

[16] C Habold, C Chevalier, S Dunel-Erb, C Foltzer-Jourdainne, Y Le Maho, and J-H Lignot. Effects of fasting and refeeding on jejunal morphology and cellular activity in rats in relation to depletion of body stores. *Scandinavian journal of gastroenterology*, 39(6):531–9, June 2004.

[17] M Bauer, AC Hamm, M Bonaus, A Jacob, J Jaekel, H Schorle, MJ Pankratz, and JD Katzenberger. Starvation response in mouse liver shows strong correlation with life-span-prolonging processes. *Physiological genomics*, 17(2):230–44, April 2004.

[18] RT Jagoe, SH Lecker, M Gomes, and AL Goldberg. Patterns of gene expression in atrophying skeletal muscles: response to food deprivation. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 16(13):1697–712, November 2002.

[19] SH Lecker, RT Jagoe, A Gilbert, M Gomes, V Baracos, J Bailey, SR Price, WE Mitch, and AL Goldberg. Multiple types of skeletal muscle atrophy involve a common program of changes in gene expression. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 18(1):39–51, January 2004.

[20] M Sokolović, A Sokolović, D Wehkamp, E Ver Loren van Themaat, DR de Waart, LA Gilhuijs-Pederson, Y Nikolsky, AHC van Kampen, TBM Hakvoort, and WH Lamers. The transcriptomic signature of fasting murine liver. *BMC genomics*, 9:528, January 2008.

[21] XQ Xiao, KL Grove, BE Grayson, and MS Smith. Inhibition of uncoupling protein expression during lactation: role of leptin. *Endocrinology*, 145(2):830–8, February 2004.

[22] TBM Hakvoort, PD Moerland, R Frijters, A Sokolović, WT Labruyère, JLM Vermeulen, E Ver Loren van Themaat, TM Breit, FRA Wittink, AHC van Kampen, AJ Verhoeven, WH Lamers, and M Sokolović. Interorgan coordination of the murine adaptive response to fasting. *The Journal of biological chemistry*, 286(18):16332–43, May 2011.

[23] K Lenaerts, E Mariman, F Bouwman, and J Renes. Glutamine regulates the expression of proteins with a potential health-promoting effect in human intestinal Caco-2 cells. *Proteomics*, 6(8):2454–64, April 2006.

[24] MP van Iersel, AR Pico, T Kelder, J Gao, I Ho, K Hanspers, BR Conklin, and CT Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, 11:5, January 2010.
[25] MP van Iersel, T Kelder, AR Pico, K Hanspers, S Coort, BR Conklin, and C Evelo. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics*, 9:399, January 2008.
[26] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.
[27] KM Waters, JG Pounds, and BD Thrall. Data merging for integrated microarray and proteomic analysis. *Briefings in functional genomics & proteomics*, 5(4):261–72, December 2006.
[28] F Yoshizawa, Y Miura, K Tsurumaru, Y Kimata, K Yagasaki, and R Funabiki. Elongation factor 2 in the liver and skeletal muscle of mice is decreased by starvation. *Bioscience, biotechnology, and biochemistry*, 64(11):2482–5, November 2000.
[29] RG Krishna and F Wold. Post-translational modification of proteins. *Advances in enzymology and related areas of molecular biology*, 67:265–98, January 1993.
[30] KJ Hintze and EC Theil. Cellular regulation and molecular interactions of the ferritins. *Cellular and molecular life sciences : CMLS*, 63(5):591–600, March 2006.
[31] S Okuda, T Yamada, M Hamajima, M Itoh, T Katayama, P Bork, S Goto, and M Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic acids research*, 36(Web Server issue):W423–6, July 2008.
[32] H Neuweger, M Persicke, SP Albaum, T Bekel, M Dondrup, AT Hüser, J Winnebald, J Schneider, J Kalinowski, and A Goesmann. Visualizing post genomics data-sets on customized pathway maps by ProMeTra-aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example. *BMC systems biology*, 3:82, January 2009.
[33] C Klukas and F Schreiber. Integration of -omics data and networks for biomedical research with VANTED. *Journal of integrative bioinformatics*, 7(2):112, January 2010.
[34] D Croft, AF Mundo, R Haw, M Milacic, J Weiser, G Wu, M Caudy, P Garapati, M Gillespie, MR Kamdar, B Jassal, S Jupe, L Matthews, B May, S Palatnik, K Rothfels, V Shamovsky, H Song, M Williams, W Birney, H Hermjakob, L Stein, and P D'Eustachio. The Reactome pathway knowledgebase. *Nucleic acids research*, 42(Database issue):D472–7, January 2014.
[35] N Gehlenborg, SI O'Donoghue, NS Baliga, A Goesmann, MA Hibbs, H Kitano, O Kohlbacher, H Neuweger, R Schneider, D Tenenbaum, and AC Gavin. Visualization of omics data for systems biology. *Nature methods*, 7(3 Suppl):S56–68, March 2010.
[36] AR Joyce and B ØPalsson. The model organism as a system: integrating 'omics' data sets. *Nature reviews. Molecular cell biology*, 7(3):198–210, March 2006.

Chapter 5

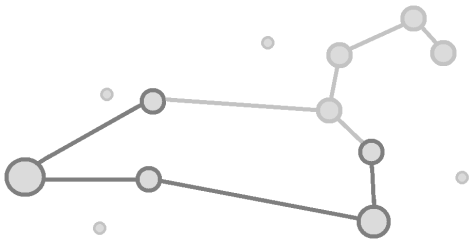
WikiPathways App for Cytoscape: Making Biological Pathways Amenable to Network Analysis and Visualization

Martina Kutmon[*,1], Samad Lotia[*,2], Chris T Evelo[1], Alexander R Pico[2]

[*] Contributed equally to this work

1. Department of Bioinformatics - BiGCaT, NUTRIM School for Nutrition, Toxicology and Metabolism, Maastricht University, The Netherlands
2. Gladstone Institutes, San Francisco, USA

F1000Research (2014) 3:152



Abstract

In this paper we present the open-source WikiPathways app for Cytoscape (http://apps.cytoscape.org/apps/wikipathways) that can be used to import biological pathways for data visualization and network analysis. WikiPathways is an open, collaborative biological pathway database that provides fully annotated pathway diagrams for manual download or through web services. The WikiPathways app allows users to load pathways in two different views: as an annotated pathway ideal for data visualization and as a simple network to perform computational analysis. An example pathway and dataset are used to demonstrate the functionality of the WikiPathways app and how they can be combined and used together with other apps. More than 2000 downloads between its first release in August 2013 and the submission of the paper in May 2014 highlight the importance and adoption of the app in the network biology field.

Introduction

Pathways are commonly used as models for understanding biological processes. WikiPathways [1] is an open, collaborative, wiki-based website for the curation of biological pathways that are more than just images. WikiPathways provides easy-to-use drawing and annotation tools to capture identities, relationships, comments and literature references for each pathway element and interaction. Contributed pathways are displayed like articles at WikiPathways and can be downloaded manually or programmatically through web services. This opens the possibility for pathway information to be accessed by other software tools for data visualization, computational analysis and the interpretation of large-scale experimental data.
Utilizing WikiPathways web services, we developed an app for Cytoscape [2], a network visualization and analysis software platform. The app queries and imports pathways from WikiPathways within the Cytoscape environment. Cytoscapes core concepts are networks (nodes and edges), tables (rows and columns) and styles, which map table values to the visual properties of networks. Cytoscape leverages a rich ecosystem of apps to provide additional domain-specific semantics and data types, as well as custom visualization and analysis capabilities. With the WikiPathways app, we implemented two ways to represent a pathway as a Cytoscape network. In the first way, pathways are loaded with the complete visual appearance of the original at WikiPathways, including graphical annotations and labels. Once in Cytoscape, experimental data can be loaded as tables and visually mapped onto these pathway-style networks to provide biological context. In the second way, pathways are loaded as simplified networks, focusing on the biological entities and their interactions without any of the graphical elements of the original pathway diagram. The basic network style is ideal for topological analyses, network merging and automatic layout.
In this paper we present the implementation and usage of the WikiPathways app for Cytoscape. By bringing pathways into Cytoscape using the WikiPathways app, it is possible to make full use of pathway models with custom visualizations and computational analyses.

Implementation

The WikiPathways app was developed for Cytoscape 3, which introduced a completely new software architecture. The new architecture is built on top of Open Service Gateway Initiative (OSGi) [3], a software framework of pluggable modules and services. To be able to take advantage of the new architecture (Cytoscape API version 3.0.0), the predecessor to the WikiPathways app, the GPML Plugin, had to be rewritten.

Pathway Import

The WikiPathways app employs the new architecture of Cytoscape in two ways. First, the app exports a user interface that can query and import pathways from the WikiPathways web service. Thanks to the service architecture in Cytoscape, this interface is seamlessly incorporated into Cytoscapes “Import from Public

Databases" dialog. Second, the app provides an API for programmatic access to the WikiPathways web services and the GPML file importer. Other apps can use the API to make queries to the WikiPathways web services and import GPML files without having to bundle the WikiPathways app. When the WikiPathways app is loaded in Cytoscape, the app registers the implementation of its API with the OSGi module system. Other apps can then request the API implementation through OSGi.

Visualization

The new architecture also posed new challenges that required us to innovate with respect to visual styles. The new architecture includes a revamped model to represent networks. This model decouples the network topology and table data from its visual style. Visual styles constitute Cytoscapes view model. When a node or edge is created in the network model, its view object is only created after a triggering of an event. Cytoscape does this to avoid redrawing of the network canvas while an app is still in process of building the network. Indeed, as the WikiPathways app reads a GPML file, it creates a series of nodes and edges in a network to represent the pathway. During this process, the app needs to assign visual styles to the nodes and edges it creates. However, as new nodes and edges are being added to the network, their view objects do not exist yet, making it impossible to assign their visual styles. To address this issue, we created a class called `DelayedVizProp` that stores our desired visual styles for nodes and edges. Once the network has been fully built, the app tells Cytoscape to create the view objects for the new nodes and edges. After that, the app looks through the `DelayedVizProp` instances and assigns nodes and edges their desired visual style.

Dependencies

The app relies on the PathVisio core library [4] to read GPML files. The PathVisio library is included in the app. In previous versions of Cytoscape, apps that included libraries often conflicted with each other. Users had to painstakingly uninstall conflicting apps for Cytoscape to become usable again. OSGi solves this problem by insulating Cytoscape modules and apps from each other. Due to OSGis architecture in Cytoscape 3, the integrated PathVisio library is hidden from other apps and modules in Cytoscape and cannot conflict with them.

The app also uses the Apache HTTP Client library to make HTTP requests to the WikiPathways REST server. We avoided the Java built-in HTTP client class (`java.net.HttpURLConnection`), which is used frequently in Cytoscape and other apps. This class does not support cancellation. Proper cancellation is important for a responsive user interface. Users behind an interrupted internet connection should be able to back out of a WikiPathways request and return to Cytoscape. Each HTTP request is wrapped in a task, a unit of work in Cytoscape. When the user clicks cancel during the task execution, the app terminates the underlying HTTP request by calling the abort method in the Apache HTTP Client library.

Results

The WikiPathways app in Cytoscape provides convenient access to the community-curated collection of biological pathways at WikiPathways. The functionality of the app is demonstrated here using the human Cardiac Hypertrophic Response pathway from WikiPathways (http://wikipathways.org/instance/WP2795) combined with an unpublished RNA-seq dataset that reflects gene expression levels during differentiation of cardiac stem cells. The logFC from *timepoint 6 hrs vs control* is visualized on the pathway. The human Cardiac Hypertrophic Response pathway contains gene products and metabolites involved in the intracellular signal-transduction pathways that coordinate Cardiac Hypertrophic Response. As described above, the WikiPathways app allows users to load pathways in two different views, as an annotated pathway and as a simple network (see Figures 5.1 and 5.2). The example dataset and pathway will be used to explain how both views can be used in Cytoscape.

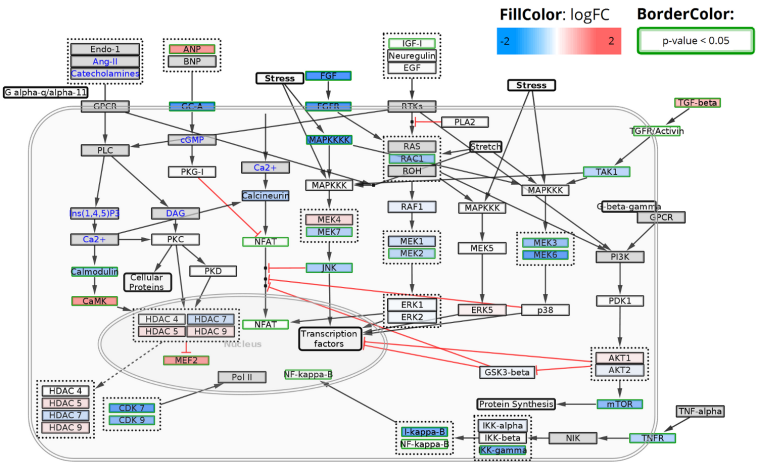


Figure 5.1: The Cardiac Hypertrophic Response Pathway Loaded as a Pathway. LogFC values are visualized as node fill color with a color gradient from blue over white to red. Significant measurements (adjusted p-value < 0.05) are highlighted with a green border color. Elements in the pathway without a measurement are colored gray.

When loaded as a pathway, the precise layout of elements is identical to its representation at WikiPathways. The graphical elements, like labels and shapes, are included in the model in Cytoscape. As a pathway diagram, the full representation of biological information is visually preserved, which is ideal for providing a meaningful context for data visualization. Figure 5.1 shows the Cardiac Hypertrophic Response pathway loaded as an annotated pathway in Cytoscape. The Entrez Gene identifiers in the pathway were mapped to Ensembl using another app called BridgeDb ([5], http://apps.cytoscape.org/apps/bridgedb) to match the identifiers used in the example dataset. The cardiac stem cell tissue development expression data can then be loaded, integrated and visualized on the pathway nodes.
When loaded as a network, all graphical annotations are removed and redundant nodes in the pathway are merged into one unique node in the network. Groups and

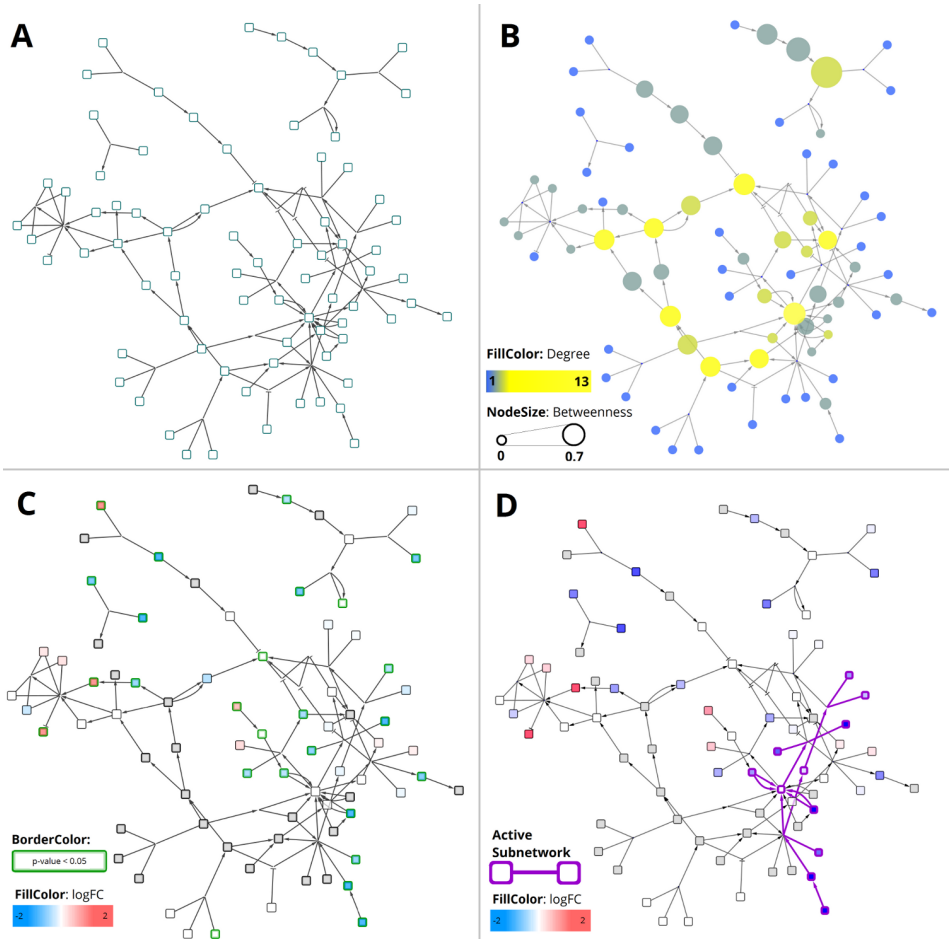


Figure 5.2: The Cardiac Hypertrophic Response Pathway Loaded as a Network. (A) The simple network does not contain graphical annotations of the pathway. (B) NetworkAnalyzer was used to visualize node degree and betweenness of the nodes in the network to identify important hub nodes. (C) The logFC of the example dataset is visualized as node fill color with a gradient (blue over white to red) and the adjusted p-value < 0.05 is highlighted with a green border color. (D) jActiveModules finds active subnetworks (highlighted in purple) that are affected by varying gene expression.

complex interactions are visualized as very small nodes and a forced directed layout is applied. As an abstracted network graph, the same molecular relationships in the pathways can be made available for network analysis and augmentation. Figure 5.2A shows the Cardiac Hypertrophic Response pathway loaded as a network in Cytoscape. This simple network structure enables researchers to use other Cytoscape features and apps to merge two pathways, apply different layouts to the network or extend the pathway, for example, with regulatory interactions (CyTargetLinker [6], http://apps.cytoscape.org/apps/cytargetlinker). It also enables users to investigate the topology of the network, like calculating degree and betweenness of the nodes with Cytoscapes built-in NetworkAnalyzer tool to identify

important hub nodes, see Figure 5.2B. Cytoscape also allows the visualization of experimental data in the network, as in Figure 5.2C which shows the cardiac stem cell tissue development expression data. There are several apps available for Cytoscape that provide methods that use experimental data to cluster nodes in the network (clusterMaker2, http://apps.cytoscape.org/apps/clustermaker2) or find subregions in the network affected by varying gene expression (jActiveModules, http://apps.cytoscape.org/apps/jactivemodules) as highlighted in Figure 5.2D.

Conclusions

In this paper we presented the WikiPathways app for Cytoscape, which allows the import of biological pathways as curated diagrams or as basic node-and-edge networks into Cytoscape. As shown in some examples, the app enables users to make full use of the pathway models by performing computational analyses and custom visualizations based on experimental data and network topology.

Software Availability

App website:
http://apps.cytoscape.org/apps/wikipathways

Source code:
https://github.com/wikipathways/cytoscape-wikipathways-app

License:
Lesser GNU Public License 3.0 (https://www.gnu.org/licenses/lgpl.html)

Bibliography

[1] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[2] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504, November 2003.

[3] Osgi Alliance. Osgi Service Platform, Release 3. December 2003.

[4] MP van Iersel, T Kelder, AR Pico, K Hanspers, S Coort, BR Conklin, and C Evelo. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics*, 9:399, January 2008.

[5] J Gao, C Zhang, M van Iersel, L Zhang, D Xu, N Schultz, and AR Pico. BridgeDb app: unifying identifier mapping services for Cytoscape. *F1000Research*, 3, July 2014.

[6] M Kutmon, T Kelder, P Mandaviya, CTA Evelo, and SL Coort. CyTargetLinker: a Cytoscape app to integrate regulatory interactions in network analysis. *PloS one*, 8(12):e82160, January 2013.

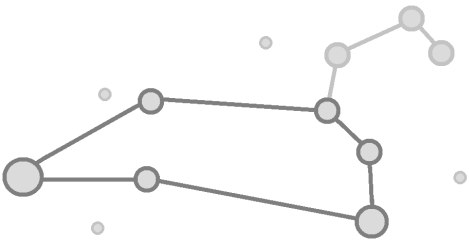
Chapter 6

CyTargetLinker: A Cytoscape App to Integrate Regulatory Interactions in Network Analysis

Martina Kutmon[1], Thomas Kelder[2], Pooja Mandaviya[1], Chris T Evelo[1], Susan L Coort[1]

1. Department of Bioinformatics - BiGCaT, NUTRIM School for Nutrition, Toxicology and Metabolism, Maastricht University, The Netherlands
2. TNO, Research Group Microbiology and Systems Biology, Zeist, The Netherlands

PLoS One (2013) 8(12):e82160



Abstract

Introduction: The high complexity and dynamic nature of the regulation of gene expression, protein synthesis, and protein activity pose a challenge to fully understand the cellular machinery. By deciphering the role of important players, including transcription factors, microRNAs, or small molecules, a better understanding of key regulatory processes can be obtained. Various databases contain information on the interactions of regulators with their targets for different organisms, data recently being extended with the results of the Encyclopedia of DNA Elements project. A systems biology approach integrating our understanding on different regulators is essential in interpreting the regulation of molecular biological processes.

Implementation: We developed CyTargetLinker, a Cytoscape app, for integrating regulatory interactions in network analysis. Recently we released CyTargetLinker as one of the first apps for Cytoscape 3. It provides a user-friendly and flexible interface to extend biological networks with regulatory interactions, such as microRNA-target, transcription factor-target and/or drug-target. Importantly, CyTargetLinker employs identifier mapping to combine various interaction data resources that use different types of identifiers.

Results: Three case studies demonstrate the strength and broad applicability of CyTargetLinker, (i) extending a mouse molecular interaction network, containing genes linked to *diabetes mellitus*, with validated and predicted microRNAs, (ii) enriching a molecular interaction network, containing DNA repair genes, with ENCODE transcription factor and (iii) building a regulatory meta-network in which a biological process is extended with information on transcription factor, microRNA and drug regulation.

Conclusions: CyTargetLinker provides a simple and extensible framework for biologists and bioinformaticians to integrate different regulatory interactions into their network analysis approaches. Visualization options enable biological interpretation of complex regulatory networks in a graphical way. Importantly the incorporation of our tool into the Cytoscape framework allows the application of CyTargetLinker in combination with a wide variety of other apps for state-of-the-art network analysis.

Introduction

Completion of the human genome project in 2003 generated a wealth of information about the human genetic code [1]. Approximately 25,000 gene coding regions were defined. However, understanding of the regulation of gene expression, protein synthesis and activity is far from complete. Recently, the ENCODE project, whose main goal was to identify all the functional elements in the human genome sequence, revealed novel insights in genetic regulation [2, 3]. It still remains a challenge to combine the new insights with existing knowledge and to understand the regulation of biological processes in detail. Many known biological processes are represented in various online repositories, like WikiPathways [4] and Reactome [5]. These processes contain genes, proteins and/or metabolites, their molecular interactions and reactions, but little regulatory information is present.

Regulation of gene expression, protein synthesis and activity occurs at different levels. Whereas gene expression is influenced by epigenetic factors and/or transcription factor (TF) binding, protein synthesis can be regulated by microRNAs (miRNAs). TFs are proteins that bind to a specific DNA sequence, i.e., the transcription factor binding site (TFBS). They either activate or repress transcription. miRNAs are small, non-coding RNA molecules that bind to miRNA-target regions in the mRNA. Upon binding miRNAs either repress translation or cleave the mRNA sequence. They are able to influence the synthesis of many proteins or even those involved in entire pathways, making them important molecules in harmonised regulation. Another group of regulatory effects are post-translational modifications which can influence protein activity. These modifications include phosphorylation, acetylation, palmitoylation and many more. In addition, small molecules such as metabolites or drugs can play a role as regulators in cellular pathways.

A systems biology approach in which interactions from different resources are combined, visualised and analysed together is an intuitive way to decipher complex biological processes. A commonly used framework to visualise and analyse biological networks is Cytoscape [6]. Its modular structure and possibilities to extend with additional functionalities through apps (formerly known as plugins) is discussed in "A travel guide to Cytoscape plugins" [7]. At the moment a few Cytoscape apps are available that either extend networks with other types of molecular interaction data or focus on one specific type of regulatory interaction. However, a user-friendly tool to combine and integrate various types of regulation is still needed. In this paper, we present a new Cytoscape app, CyTargetLinker, to automatically add regulatory interactions to biological networks to allow their inclusion in the network analysis process. CyTargetLinker is not restricted to one specific organism or regulatory interaction type and it leaves the selection of relevant and/or preferred interaction databases entirely to the user. The incorporation of our tool into the Cytoscape framework allows its application in combination with several community-contributed apps for data visualization and advanced network analysis.

Implementation

CyTargetLinker is an open source app developed for the network visualization and analysis tool Cytoscape [6] and can be installed through the app manager in Cytoscape 2.8 or 3.x. The source code is available on Github (https://github.com/mkutmon/cytargetlinker).

CyTargetLinker allows users to build regulatory networks to obtain a more complete view of biological systems. The regulatory interactions used in CyTargetLinker are derived from so called *regulatory interaction networks* (RegINs) that are either provided on the CyTargetLinker website or can be created by the user. The creation, application and content of RegINs is explained below. All functionalities of CyTargetLinker are described in the "CyTargetLinker workflow" section.

Regulatory Interaction Networks

A RegIN is a network containing regulatory interactions that are often derived from online interaction databases. The networks are stored in XGMML (the eXtensible Graph Markup and Modelling Language) format, which is supported by Cytoscape. Each regulatory interaction consists of two nodes, a source (regulatory component) and target biomolecule, connected through one directed edge. A collection of RegINs for different species and interaction types is provided on the CyTargetLinker website (http://projects.bigcat.unimaas.nl/cytargetlinker/regins), and is described in more details in Table 6.1. In addition, we provide documentation on how to create your own RegIN. The app is not restricted to the RegINs provided and the user can choose which interaction types and databases should be used in the integration process.

Table 6.1: Regulatory Interaction Files Subset of the RegINs (regulatory interaction networks) available for download on the CyTargetLinker website. All RegIN networks support the following identifier systems: (i) for genes/proteins → Ensembl, NCBI gene, UniProt, (ii) for miRNAs → miRBase accession number and ids, and (iii) for drugs → DrugBank. (* Redistribution of data not allowed, but RegIN can be created with our provided conversion script)

Database	version	Type	Human	Mouse	Rat	Zebrafish
MicroCosm	5	predicted MTI	541,039	494,822	511,057	121,992
TargetScan	6.2	predicted MTI	511,040	186,431	-	-
miRTarBase	3.5	validated MTI	3,597	712	278	104
miRecords*	4	validated MTI	1,752	395	161	48
ENCODE	2012	proximal TF-target	24,111	-	-	-
ENCODE	2012	distal TF-target	18,240	-	-	-
TFe	2012	TF-target	1,531	847	-	-
DrugBank	3	drug-target	14,070	-	-	-

A set of RegINs can be seen as a collection of online interaction databases that are formatted in the same way so they can be combined in the integration and analysis process. For the available RegINs, in order to be able to jointly use them, one unifying identifier system was used: the Ensembl gene identifier [8] was chosen for genes, the miRBase [9] accession number for miRNAs and DrugBank [10] identifiers

for drugs. The identifier mapping was performed using the BridgeDb mapping framework [11]. In addition to the main identifier system, the RegINs contain additional systems (e.g. NCBI gene [12] and UniProt [13] for genes/proteins) to give the user more freedom to choose the identifier system in the initial network which has to match to that used in the RegIN. In case the identifier system is not supported by the RegINs to be used, the user can use the BridgeDb [14] app in Cytoscape to map the identifiers to one of the supported systems.

CyTargetLinker Workflow

CyTargetLinker enables the enrichment of biological networks with regulatory information in a user-friendly and flexible manner. The CyTargetLinker workflow will now be discussed in detail and is illustrated by an example in Figure 6.1.

The first step is to load or create a biological network in Cytoscape. Starting from a protein-protein network, a biological pathway or unconnected gene nodes, the initial network that will be extended with regulatory information can be very different. In each case the elements in the network should be annotated using one of the supported identifier systems. The second step is to download or create RegINs, as described in the next section. In the third step the CyTargetLinker integration process is started in Cytoscape. In the dialog the user selects the biological network, the node identifier attribute and the local directory containing the downloaded RegIN files. Thereafter, the direction of the interaction should be selected. It is possible to only add targets, regulators or both (default). CyTargetLinker will extract only those regulatory interactions from the provided RegINs, in which one of the nodes in the initial network is a participant, either regulator or target. This reduces the amount of memory needed, speeds up the integration process, and makes CyTargetLinker scalable to large regulatory networks.

After the extension of the network the initial network nodes are visualised as grey circles whereas the added nodes are shown as pink hexagons (see Figure 6.1B). Moreover, the edge colour defines in which RegIN an interaction is present. If an interaction is supported by more than one RegIN, CyTargetLinker will add one differently coloured edge for each RegIN. In the accompanying control panel the interaction colour can be changed and the number of added interactions per RegIN is listed. In the fourth step the visualization of the regulatory network can be adapted by using the *hide/show* and/or *overlap threshold* function. The *hide/show* functionality enables the temporary removal of specific RegINs and thereby showing only the interactions from a subset of the loaded RegINs (see Figure 6.1C). The *overlap threshold* functionality makes it possible to show only the interactions that are supported by a defined number of RegINs or more (see Figure 6.1D). Both functions can be applied and restored in the same network window.

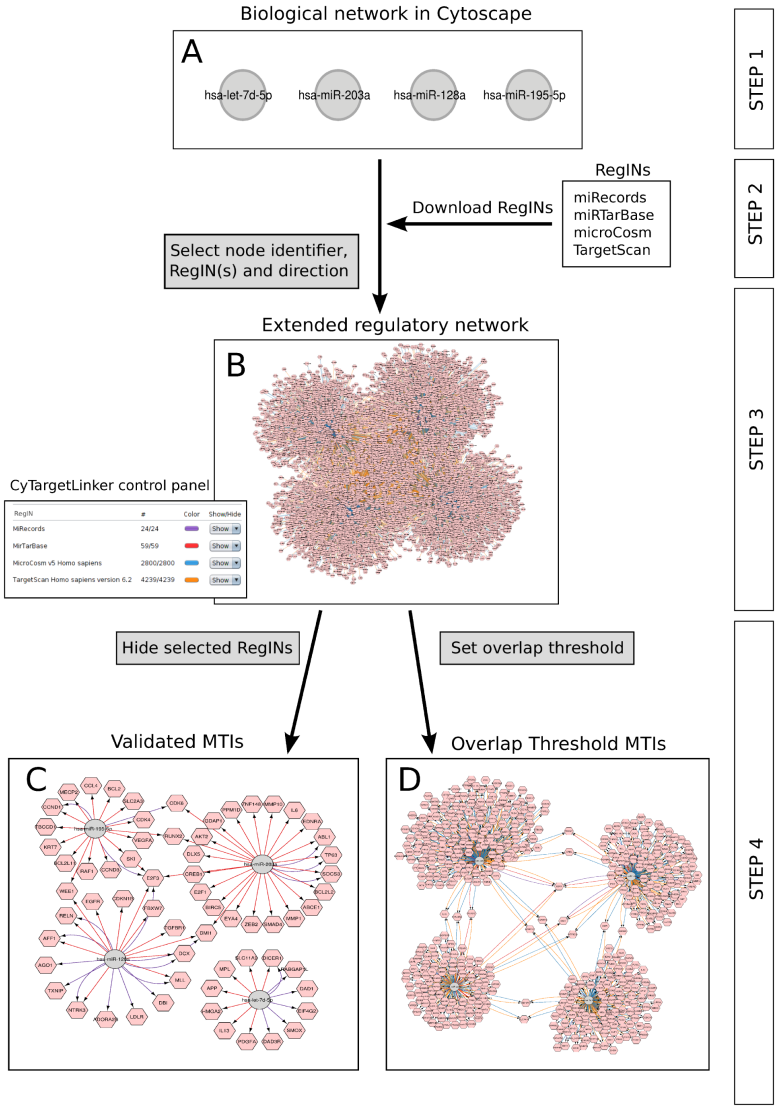


Figure 6.1: CyTargetLinker Workflow. Step 1: Four miRNAs known to be involved in prostate cancer [15] are visualised in a Cytoscape network (A). The miRNAs are annotated with miRBase accession numbers and ids. **Step 2:** The regulatory interaction networks (RegINs) harbouring miRNA-target interactions (MTIs), either validated (miRecords and miRTarBase) or predicted (microCosm and TargetScan), are downloaded from http://projects.bigcat.unimaas.nl/cytargetlinker/regins. **Step 3:** Known targets are integrated (B) after specifying the miRBase accession number column, the RegINs directory and the direction "*Add Targets*" in the CyTargetLinker dialog. In the resulting network miRNAs and target genes are defined as grey circles and pink hexagons, respectively. The predicted MTIs are visualised in orange (TargetScan: 4239) and blue (microCosm: 2800) and the validated MTIs in red (miRTarBase: 59) and purple (miRecords: 24), as shown in the control panel. **Step 4:** The hide/show and overlap threshold functions were used to visualise validated interactions exclusively or to show the overlap in MTI coverage. In the validated network only the MTIs in miRecords and miRTarBase are visualised by hiding MTIs in TargetScan and microCosm (C). In the overlap network the MTIs present in two or more RegINs are shown by setting the threshold to 2 (D).

Results

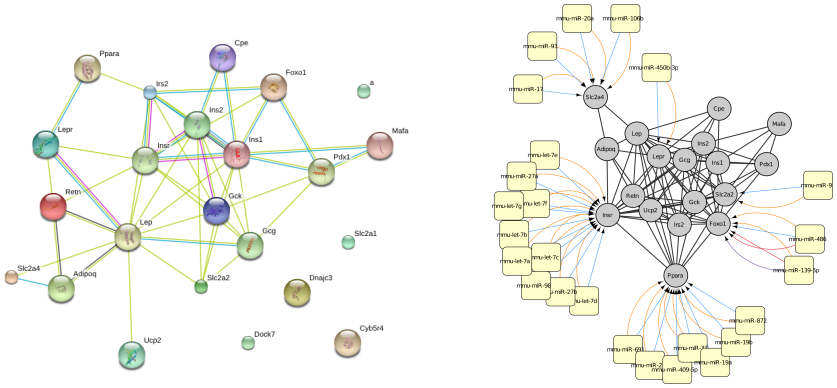
The strength and broad applicability of CyTargetLinker will be demonstrated by three different case studies. In these studies the currently available RegINs will be used for extending biological networks. In Table 6.1, a subset of the downloadable RegINs present at the CyTargetLinker website is shown. The RegINs are generated from the latest database version and will be updated once a new version is available and accessible. The older versions will stay available in an archive.

Use Cases

Case Study 1: Enrichment of a Mouse Molecular Interaction Network, Containing Genes Linked to *Diabetes Mellitus*, with miRNA Information.

The first case study demonstrates that CyTargetLinker is not limited to human networks, but can be used for other species as well. The threshold functionality is applied to show only interactions that are supported by at least two miRNA-target interaction (MTI) databases.

Diabetes mellitus is a group of metabolic diseases. The two major types are type 1 and type 2 which are characterised by impaired insulin production or insulin resistance, respectively. Worldwide the prevalence of *type 2 diabetes mellitus* (T2DM) is increasing dramatically. Although a strong environmental component is present, there is compelling evidence that genetic factors are involved in the pathogenesis of T2DM [16]. It is important to decipher the genes involved and to understand their regulation. Diabetic mouse models are often used to measure gene expression on a large scale in tissues like adipose, skeletal muscle and liver. The genes linked to *diabetes mellitus* can be functionally annotated using the terms in the disease category of MeSH (Medical Subject Headings) [17]. CyTargetLinker can be used to examine the possible role of miRNA regulation of genes known to be linked to *diabetes mellitus*.
A mouse molecular interaction network of genes linked to the MeSH term *diabetes mellitus* was obtained from Gene2MeSH [18] and the STRING database [19]. In total 18 proteins are associated with *diabetes mellitus* and are known to interact with each other. To get a better insight in how the genes are regulated by miRNAs, validated and predicted miRNAs were added and the overlap of the MTIs present in two or more databases was selected with CyTargetLinker (see Figure 6.2). After applying overlap threshold, 50 MTIs remain in the extended network originating from miRecords (1), miRTarBase (2), microCosm (24) and TargetScan (25). In the extended molecular interaction network only 6 out of 18 genes interact with mostly predicted miRNAs that were present in 2 or more interaction databases. Whereas LEPR (leptin receptor) and SLC2A2 (GLUT2) interact with only one miRNA, INSR (insulin receptor) and PPARa (peroxisomal proliferator activated receptor alpha) are highly regulated by 7 and 10 miRNAs, respectively. It is well known that the activation of the nuclear receptor, PPARalpha, has beneficial effects in T2DM. PPAR agonists are used as antidiabetic drugs to treat the symptoms of T2DM. Identifying which miRNAs interact with PPARalpha could lead to novel



(a) The Original Network from STRING.

(b) The Extended Network.

Figure 6.2: MiRNA Regulation of Genes Associated with *Diabetes Mellitus*. The genes linked to the MeSH term *diabetes mellitus* were obtained using Gene2MeSH (http://gene2mesh.ncibi.org/). The molecular interactions between the genes were obtained from the STRING database (a). In the extended network (b), genes and miRNAs are visualised as grey circles and yellow rounded rectangles, respectively. The names of the genes and miRNAs are displayed on the nodes. The MTIs originate from microCosm (486), TargetScan (207), miRTarBase (5) and miRecords (2), and are coloured in blue, orange, red and purple, respectively. The overlap threshold function was applied to show only MTIs present in at least two RegINs.

pharmacological targets. In this use case CyTargetLinker can be used to either identify miRNAs of interest or to confirm recent findings. For example, the regulation of PPARalpha by miR-21, present in the extended network, was published in 2011 in a liver study [20]. The other highly regulated gene, the insulin receptor, plays a key role in the insulin signalling pathway. Most miRNAs interacting with the insulin receptor in the extended network belong to the let-7 miRNA family (see Figure 6.2). Interestingly, Forst and Olson [21] showed that the let-7 family controls glucose homeostasis and insulin sensitivity in mice. Their study confirms that in liver and muscle the let-7 family regulates the insulin receptor as shown in the extended network.

Case Study 2: Extension with ENCODE TF Regulation Information of a Molecular Interaction Network of Human DNA Repair Genes and Its Analysis.

The second case study demonstrates how CyTargetLinker can be applied in combination with other Cytoscape apps. Moreover, it shows that published regulatory interaction data can be easily converted into a RegIN and implemented into the CyTargetLinker workflow. By using the available core functionalities of Cytoscape is it possible to adjust the colour and size of a node and to perform commonly used network analyses in the extended network.
The ENCODE project aims to delineate all functional elements encoded in the human genome [2]. Since 2003 it has generated a wealth of information on regulatory elements. Gerstein and colleagues used the recently published TF binding data to analyse differential patterns in promoter proximal and distal regulatory regions [3]. TFs bind to specific sites, TFBS, that can be proximal or distal to a

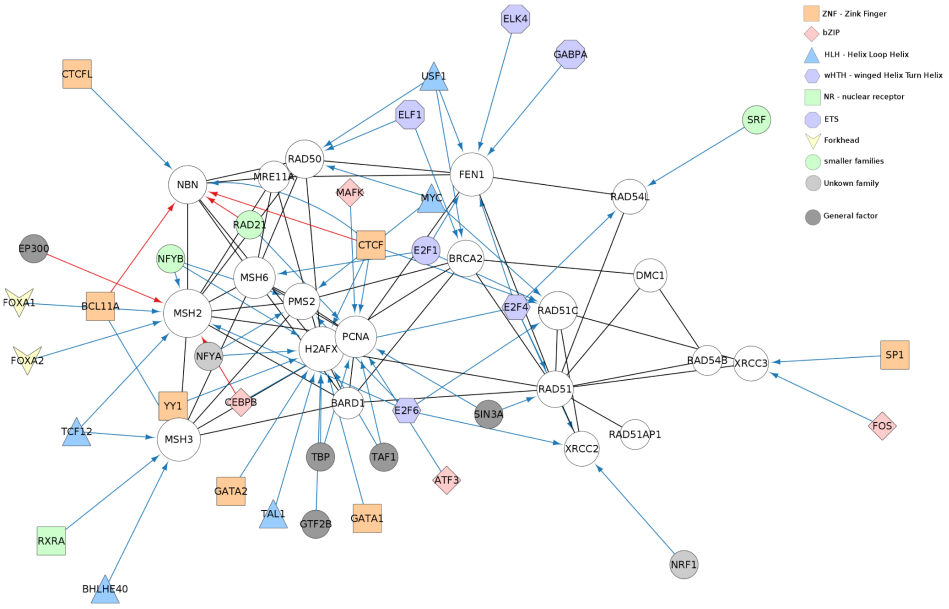


Figure 6.3: Extend a GeneMANIA Network with TF Data from the ENCODE Project. A set of known DNA repair genes (grey circled nodes) were used as input for the GeneMania app in Cytoscape. Physical interactions (black) between the query genes were added by GeneMania. Every node in a GeneMania network has an attribute "Entrez gene identifier" which was used by the CyTargetLinker app to extend the network with TF-gene interactions from the ENCODE networks. The integration direction was selected as "*Add regulators*", indicating that the app should look for interactions that target the input genes. The colours and shapes of the TFs are based on the provided TF family information in the ENCODE project [3]. The ENCODE project studied proximal and distal TF regulation which are indicated in this figure as blue and red edges.

transcription start site [22]. Distinguishing these two types of TF-regulation gives a more distinctive view on how gene expression can be influenced.
We used the proximal and distal regulatory interaction data provided by Gerstein and generated RegINs for both. A set of known DNA repair genes were used as input for the GeneMANIA app [23] in Cytoscape to create a molecular interaction network. GeneMANIA identifies the most related genes to a query gene set using a guilt-by-association approach. With CyTargetLinker the DNA repair network was enriched with the two types of regulatory data. Next, we used Cytoscape's VizMapper to shape and colour the TF nodes according to their TF family. To identify the genes highly regulated by TFs, the indegree was calculated and represented as the size of the gene nodes, see Figure 6.3. From the network it is immediately clear that the H2AFX (H2A histone family, member X) gene is highly regulated by proximal TFs. Moreover, the NBN and MSH2 genes are regulated by both proximal and distal TFs.

Case Study 3: Enrichment of a Human Pathway from WikiPathways with miRNAs, TFs and Drugs Targeting the Genes and Gene Products in the Pathway.

Case study 3 highlights the power of CyTargetLinker to build an extensive regulatory interaction network integrating a wide range of known interactions. This network can be used as a starting point for various network analysis approaches to filter out regulatory interactions that are relevant in a given context. The integration of different regulatory elements together allows the researcher to get a more complete view of possible regulatory mechanisms happening in a biological process.
The initial network represents the ErbB signaling pathway which was loaded through the WikiPathways web service provided by the WikiPathways app [24]. Insufficient ErbB signaling may cause the development of neurodegenerative diseases, such as multiple sclerosis and Alzheimer's disease. Furthermore, excessive ErbB signaling is associated with the development of various types of solid tumours [25]. The pathway contains 69 genes, proteins and metabolites (plus 13 group nodes representing grouped genes or proteins in the original pathway diagram) and 93 edges, see Figure 6.4. The network was extended with three different types of regulatory interactions, (i) drug-target interactions from DrugBank, (ii) proximal and distal TF-gene interactions from ENCODE and (iii) validated miRNA-target interactions from miRTarBase and miRecords, see Figure 6.5. In total, 558 regulatory interactions were integrated in the network, including 138 drug-target, 136 proximal TF-gene, 122 distal TF-gene and 162 validated miRNA-target interactions.

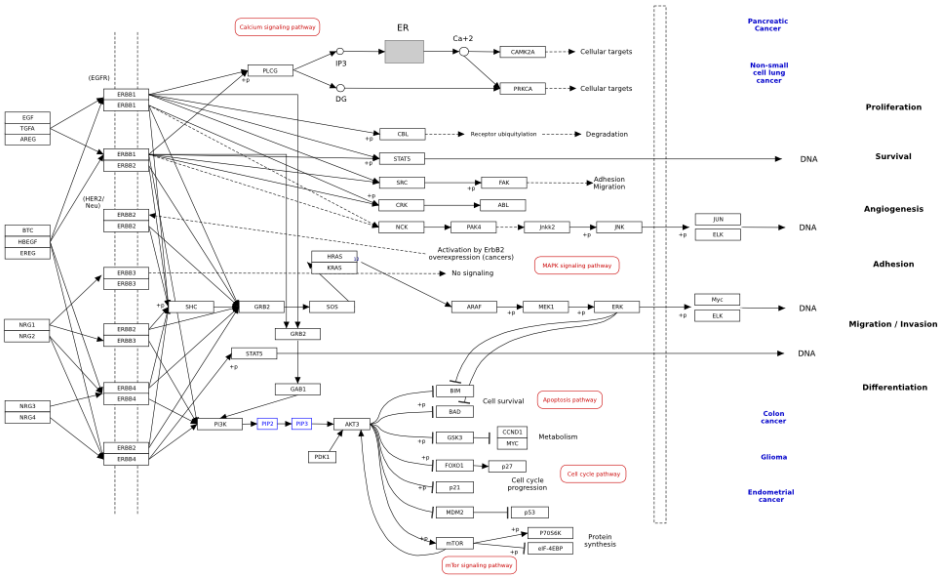


Figure 6.4: ErbB Signaling Pathway. The ErbB signaling pathway (http://wikipathways.org/instance/WP673) from WikiPathways.

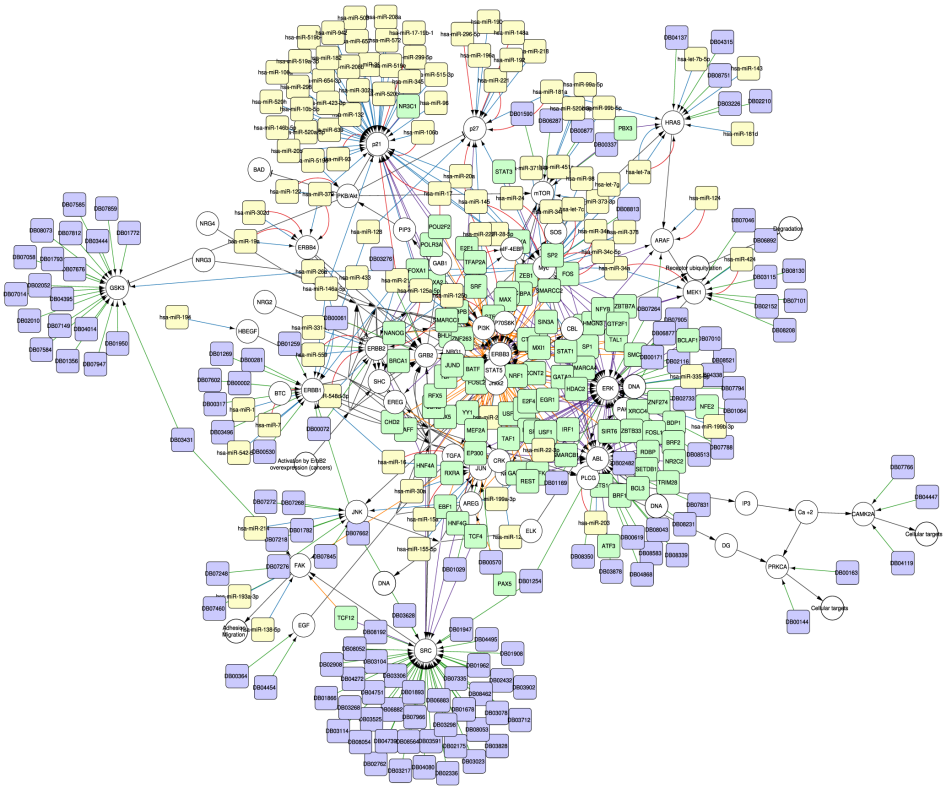


Figure 6.5: Meta Network of the ErbB Signaling Pathway Containing miRNA, TF and Drug Regulation. The ErbB signaling pathway was extended with TFs, miRNAs and drugs regulating the elements in the pathway. The nodes from the original pathway are coloured in white, TFs in green, miRNAs in yellow and drugs in purple. 258 TF-gene interactions from ENCODE, 162 miRNA-target interactions from the two validated miRNA-target RegINs (miRTarBase and miRecords) and 138 drug-target interactions from DrugBank were added to the network. The edges are coloured based on the RegIN source, purple and orange for ENCODE (proximal and distal), blue and red for miRTarBase and miRecords, and green for DrugBank.

There are a few nodes that are mostly targeted by drugs, e.g. GSK3 or SRC and other nodes that are targeted by all types of regulators, e.g. ERK or HRAS. In terms of transcription factor regulation, some nodes are regulated in a proximal setting, e.g. ERK or ABL and others mostly in a distal way, e.g. ERBB3 or STAT5. Some TFs regulate different genes in different ways, e.g. TCF12 regulates FAK distal and SRC proximal. Since a few TFs were already present in the initial network, typical network motifs can be found, e.g. four feed forward loops regulating ERBB3 over Myc by CTCF, SMC3, RAD21 and GATA2. In this integration process the direction was set to include only regulators, so the target genes of MYC and JUN, two major TFs that are present in the pathway, are not added.

Related to miRNAs our analysis shows that p21 is mostly regulated by miRNAs and all the interactions are experimentally validated (references provided in edge

attributes). Furthermore with the *overlap threshold* functionality, it can be visualised that only 30 out of 130 miRNA-target interactions are present in both miRTarbase and miRecords.

Discussion

CyTargetLinker enables quick and extensive enrichment of biological networks with regulatory information. It encapsulates all the manual integration steps which include several tasks that require advanced programming knowledge. Therefore it can be used by researchers from different fields with or without prior knowledge in programming and network analysis. CyTargetLinker is open source and freely available through the app manager for Cytoscape 2.8 and 3.x. As mentioned in the introduction, there are a few other Cytoscape apps that provide partial similar functionality. miRScape, CluePedia [26], ConReg, and BioNetBuilder [27] are most related to CyTargetLinker, see Table 6.2.

Table 6.2: Comparison of Available Cytoscape Apps. Overview of four Cytoscape apps that are most related to CyTargetLinker. This table provides a comparison of the availability and data used in the apps.

App name	Cytoscape version	Availability	Data	Organisms
miRScape	2.6	only by contacting developer	data from miRo' knowledge base [28] (last updated in 2009)	Human
CluePedia	3.0	app manager but can only be used with a license key	microCosm (miRanda) and miRecords	Human
ConReg	2.8	plugin manager	databases, text-mining and TFBS predictions	8 model organisms
BioNetBuilder	2.6	plugin manager	DIP, BIND, Prolinks, KEGG, HPRD, BioGrid, GO	>1000 organisms
CyTargetLinker	3.0	app manager	not restricted, user can provide RegINs	not restricted

miRScape identifies function, disease or process associations between genes by using miRNA-target information. CluePedia integrates experimental data to identify gene interrelations revealed by correlation weights, miRNAs regulatory aspects, protein-protein interactions as well as the functional context, in conjunction with ClueGO [29]. ConReg visualises TF-target gene networks with data from regulatory databases, text-mining approaches and TFBS predictions. It stores regulatory relations for 8 model organisms and investigates their level of conservation in related species. Lastly, BioNetBuilder focusses on the creation of biological networks using interaction data from DIP, BIND, Prolinks, KEGG, HPRD, BioGrid and GO.

BioNetBuilder and miRScape have not been maintained since version Cytoscape 2.6 (released in 2010). While CluePedia is also available as a Cytoscape 3.x app, it requires a license key. CyTargetLinker is generic and does not focus on one spe-

cific regulatory interaction type like miRScape and CluePedia. ConReg focuses only on TF-gene interactions, especially the conversion of regulatory relations in other eukaryotic model organisms and it produces a predicted conserved network. BioNetBuilder and CyTargetLinker can be used for several different species, however BioNetBuilder does not include regulatory interactions. While BioNetBuilder focusses on the creation of biological networks, CyTargetLinker extends biological networks with regulatory information. One of the advantages of the CyTargetLinker app is that it is easily expandable, a new RegIN can be added at any time and it is even possible to include a new interaction type without updating the app. Thereby, the user can use self-created RegINs in addition to the ones we provide and he can select the set of RegINs that are most suitable for his research focus. CyTargetLinker is an open source project which allows the contribution and input of other scientists to better tackle their research questions.

Conclusion

CyTargetLinker, our new Cytoscape app, enables scientists to integrate regulatory interactions into biological networks in a user-friendly and flexible manner. Various interactions, such as miRNA-target, TF-gene or drug-target, can be added, by themselves or combined. CyTargetLinker is not restricted to any organism and the commonly used identifiers for genes, proteins, and miRNAs are supported. The graphical representation in Cytoscape facilitates the identification of important regulatory interactions and can lead to new research hypotheses. The integration of CyTargetLinker into Cytoscape enables advanced network analysis and data visualization using functionality from other apps. This helps researchers to get a better understanding of the regulation of biological processes.

Future Developments

Future work, by us or other contributing groups, will include the development of new app features and conversion scripts for more publicly available databases, as well as allowing the connection to online graph databases (e.g. Neo4j) and RDF triple stores directly. This would even further simplify the integration process because the user does not need to download the RegINs beforehand.

Bibliography

[1] DR Bentley. The Human Genome Project–an overview. *Medicinal research reviews*, 20(3):189–96, May 2000.

[2] BE Bernstein, E Birney, I Dunham, ED Green, C Gunter, and M Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[3] MB Gerstein, A Kundaje, M Hariharan, SG Landt, KK Yan, C Cheng, XJ Mu, E Khurana, J Rozowsky, R Alexander, R Min, P Alves, A Abyzov, N Addleman, N Bhardwaj, AP Boyle, P Cayting, A Charos, DZ Chen, Y Cheng, D Clarke, C Eastman, G Euskirchen, S Frietze, Y Fu, J Gertz, F Grubert, A Harmanci, P Jain, M Kasowski, P Lacroute, J Leng, J Lian, H Monahan, H O'Geen, Z Ouyang, EC Partridge, D Patacsil, F Pauli, D Raha, L Ramirez, TE Reddy, B Reed, M Shi, T Slifer, J Wang, L Wu, X Yang, KY Yip, G Zilberman-Schapira, S Batzoglou, A Sidow, PJ Farnham, RM Myers, SM Weissman, and M Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, September 2012.

[4] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[5] D Croft, G O'Kelly, G Wu, R Haw, M Gillespie, L Matthews, M Caudy, P Garapati, G Gopinath, B Jassal, S Jupe, I Kalatskaya, S Mahajan, B May, N Ndegwa, E Schmidt, V Shamovsky, C Yung, E Birney, H Hermjakob, P D'Eustachio, and L Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue):D691–7, January 2011.

[6] ME Smoot, K Ono, J Ruscheinski, P-L Wang, and T Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–2, February 2011.

[7] R Saito, ME Smoot, K Ono, J Ruscheinski, P-L Wang, S Lotia, AR Pico, GD Bader, and T Ideker. A travel guide to Cytoscape plugins. *Nature methods*, 9(11):1069–76, December 2012.

[8] P Flicek, I Ahmed, MR Amode, D Barrell, K Beal, S Brent, D Carvalho-Silva, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gil, C García-Girón, L Gordon, T Hourlier, S Hunt, T Juettemann, AK Kähäri, S Keenan, M Komorowska, E Kulesha, I Longden, T Maurel, WM McLaren, M Muffato, R Nag, B Overduin, M Pignatelli, B Pritchard, E Pritchard, HS Riat, GRS Ritchie, M Ruffier, M Schuster, D Sheppard, D Sobral, K Taylor, A Thormann, S Trevanion, S White, SP Wilder, BL Aken, E Birney, F Cunningham, I Dunham, J Harrow, J Herrero, TJP Hubbard, N Johnson, R Kinsella, A Parker, G Spudich, A Yates, A Zadissa, and SMJ Searle. Ensembl 2013. *Nucleic acids research*, 41(Database issue):D48–55, January 2013.

[9] A Kozomara and S Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(Database issue):D152–7, January 2011.

[10] C Knox, V Law, T Jewison, P Liu, S Ly, A Frolkis, A Pon, K Banco, C Mak, V Neveu, Y Djoumbou, R Eisner, AC Guo, and DS Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(Database issue):D1035–41, January 2011.

[11] MP van Iersel, AR Pico, T Kelder, J Gao, I Ho, K Hanspers, BR Conklin, and CT Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, 11:5, January 2010.

[12] DL Wheeler, T Barrett, DA Benson, SH Bryant, K Canese, V Chetvernin, DM Church, M DiCuccio, R Edgar, S Federhen, LY Geer, Y Kapustin, O Khovayko, D Landsman, DJ Lipman, TL Madden, DR Maglott, J Ostell, V Miller, KD Pruitt, GD Schuler, E Sequeira, ST Sherry, K Sirotkin, A Souvorov, G Starchenko, RL Tatusov, TA Tatusova, L Wagner, and E Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 35(Database issue):D5–12, January 2007.

[13] Consortium U. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, 40(Database issue):D71–5, January 2012.

[14] J Gao, C Zhang, M van Iersel, L Zhang, D Xu, N Schultz, and AR Pico. BridgeDb app: unifying identifier mapping services for Cytoscape. *F1000Research*, 3, July 2014.

[15] HM Heneghan, N Miller, AJ Lowery, KJ Sweeney, and MJ Kerin. MicroRNAs as Novel Biomarkers for Breast Cancer. *Journal of oncology*, 2009:950201, January 2009.

[16] SA Schäfer, F Machicao, A Fritsche, H-U Häring, and K Kantartzis. New type 2 diabetes risk genes provide new insights in insulin secretion mechanisms. *Diabetes research and clinical practice*, 93 Suppl 1:S9–24, August 2011.

[17] SJ Nelson, M Schopen, AG Savage, J-L Schulman, and N Arluk. The MeSH translation maintenance system: structure, interface design, and implementation. *Studies in health technology and informatics*, 107(Pt 1):67–9, January 2004.

[18] AS Ade, ZC Wright, and DJ States. Gene2MeSH, 2007.

[19] A Franceschini, D Szklarczyk, S Frankild, M Kuhn, M Simonovic, A Roth, J Lin, P Minguez, P Bork, C von Mering, and LJ Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue):D808–15, January 2013.

[20] K Kida, M Nakajima, T Mohri, Y Oda, S Takagi, T Fukami, and T Yokoi. PPARα is regulated by miR-21 and miR-27b in human liver. *Pharmaceutical research*, 28(10):2467–76, October 2011.

[21] RJA Frost and EN Olson. Control of glucose homeostasis and insulin sensitivity by the Let-7 family of microRNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 108(52):21075–80, December 2011.

[22] WW Wasserman and A Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, 5(4):276–87, April 2004.
[23] J Montojo, K Zuberi, H Rodriguez, F Kazi, G Wright, SL Donaldson, Q Morris, and GD Bader. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics (Oxford, England)*, 26(22):2927–8, November 2010.
[24] M Kutmon, S Lotia, CT Evelo, and AR Pico. WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research*, 3, July 2014.
[25] MA Olayioye, RM Neve, HA Lane, and NE Hynes. The ErbB signaling network: receptor heterodimerization in development and cancer. *The EMBO journal*, 19(13):3159–67, July 2000.
[26] G Bindea, J Galon, and B Mlecnik. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics (Oxford, England)*, 29(5):661–3, March 2013.
[27] I Avila-Campillo, K Drew, J Lin, DJ Reiss, and R Bonneau. BioNetBuilder: automatic integration of biological networks. *Bioinformatics (Oxford, England)*, 23(3):392–3, February 2007.
[28] A Laganà, S Forte, A Giudice, MR Arena, PL Puglisi, R Giugno, A Pulvirenti, D Shasha, and A Ferro. miRò: a miRNA knowledge base. *Database : the journal of biological databases and curation*, 2009:bap008, January 2009.
[29] G Bindea, B Mlecnik, H Hackl, P Charoentong, M Tosolini, A Kirilovsky, WH Fridman, F Pagès, Z Trajanoski, and J Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)*, 25(8):1091–3, April 2009.

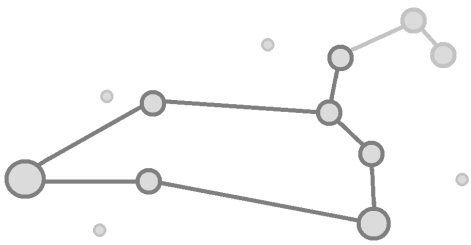
Chapter 7

A Network Biology Workflow to Study Transcriptomics Data of the Diabetic Liver

Martina Kutmon[1,2], Chris T Evelo[1], Susan L Coort[1]

1. Department of Bioinformatics - BiGCaT, NUTRIM School for Nutrition, Toxicology and Metabolism, Maastricht University, The Netherlands
2. Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, The Netherlands

BMC Genomics (2014) 15(1):971



Abstract

Background Nowadays a broad collection of transcriptomics data is publicly available in online repositories. Methods to analyse these data often aim at deciphering the influence of gene expression at process level. Biological pathways describing known processes capture the interactions of gene products and metabolites and are essential for computational analysis and interpretation of transcriptomics data.

The present study describes a comprehensive network biology workflow integrating differential gene expression in the human diabetic liver with pathway information by building a network of interconnected pathways. Worldwide the incidence of *type 2 diabetes mellitus* is increasing dramatically and to better understand this multifactorial disease more insight in the concerted action of the disease-related processes is needed. The liver is a key player in metabolic diseases and diabetic patients often develop non-alcoholic fatty liver disease.

Results A publicly available transcriptomics dataset from the liver of diabetic patients was selected after a thorough analysis. Pathway analysis revealed seven significantly altered pathways in the WikiPathways human pathway collection. These pathways were then merged into one combined network with 408 gene products, 38 metabolites and 5 pathway nodes. Further analysis highlighted 17 nodes present in multiple pathways, and therefore connecting different pathways in the network. The integration of transcription factor-gene interactions from the ENCODE project identified new links between the pathways on a regulatory level. The extension of the network with known drug-target interactions from DrugBank allows a more complete study of drug actions and the identification of other drugs targeting proteins up- or downstream that might interfere with the action or efficiency of a drug.

Conclusions The described network biology workflow uses state-of-the-art pathway and network analysis approaches to study the rewiring of the diabetic liver. The integration of experimental data and knowledge on disease affected biological pathways, including regulatory elements like transcription factors or drugs, leads to improved insights and a clearer illustration of the overall process. It also provides a resource to build new hypotheses for further follow-up studies. The approach is highly generic and can be applied in different research fields.

Background

Type 2 diabetes mellitus (T2DM) is a metabolic disorder characterized by chronic hyperglycemia with disturbances of carbohydrate, lipid and protein metabolism resulting from defects in insulin secretion, insulin resistance, or both. Obesity, the excess accumulation of lipids in the body, is a major risk factor for T2DM. Metabolism in liver, adipose tissue and skeletal muscle is of key importance for the pathogenesis of T2DM. The current study will focus on the liver. It is well known that lipid accumulation in liver contributes to insulin resistance, hyperglycemia and hyperlipidemia [1]. The hepatic lipid accumulation is the main characteristic of non-alcoholic fatty liver disease (NAFLD) and NAFLD is strongly associated with T2DM. In T2DM one of the key liver functions, the postprandial insulin-mediated uptake of glucose, is impaired [2]. Moreover, gluconeogenesis is affected because of the disturbed insulin inhibition of glucose production [3]. Published studies of gene expression in liver of patients with NAFLD suggest both increase of de novo lipogenesis and lipid oxidation [4, 5]. Although studies in NAFLD identified genes, proteins and processes that are important, not all biological mechanisms involved in the human diabetic liver are deciphered [6].

Modern technology enables global analysis of gene expression in liver tissue. Exploring published transcriptomics datasets available in online repositories revealed only one transcriptomics study investigating the human diabetic fatty liver. Pihlajamäki *et al* [7] measured gene expression with microarray technology in liver biopsies of lean, obese and obese, diabetic subjects.

Instead of only investigating the individual gene expression, nowadays analysis methods often aim at deciphering the biological function at process level. Generally, gene set enrichment analysis and pathway analysis are applied to find biological processes of interest [8]. In the original study, these approaches identified a relationship between thyroid hormone action and the altered gene expression pattern. The present study uses state of the art pathway and network analysis methods to integrate differential gene expression with pathway information by building a network of interconnected pathways.

Pathway analysis was used to find significantly altered biological processes. Pathway databases like WikiPathways [9] provide pathway collections commonly used in pathway analysis. Next, network analysis was applied to i) identify genes linking pathways relevant in the diabetic fatty liver, ii) investigate the transcriptional regulation within the pathways and iii) identify known drugs targeting genes in the pathways and their effects. The open-source and popular pathway and network visualization and analysis tools PathVisio [10] and Cytoscape [11] were used in a comprehensive network biology approach.

Although biological pathways are usually seen as independent processes, they do interact with and depend on each other. Our workflow combines and integrates relevant pathways into one biological network which allows researchers to study the effects of a disease or treatment in the pathway of interest but also in downstream

or related pathways. Further extending the network with additional information, like transcription factor (TF) regulation, provides a more complete picture of the complex biological processes and will help to better understand the mechanisms of diseases.

Methods

Transcriptomics Dataset

In this study a published and publicly available transcriptomics dataset generated by Pihlajamäki [7] was used. The dataset is available from the Gene Expression Omnibus (accession number GSE15653). 18 individuals, 5 lean and 13 obese, undergoing elective surgery for obesity or gallstones participated in the study. Based on a preoperative oral glucose tolerance test the obese subjects were diagnosed for T2DM. The percentage of liver fat content was significantly ($p<0.05$) increased in obese, diabetic subjects compared to lean subjects indicating the development of NAFLD in these subjects. Gene expression was measured in surgical liver biopsies from 4 obese subjects, 9 obese subjects with T2DM and 5 lean control subjects during fasting using Affymetrix Human Genome U133A microarrays. We selected the 5 lean and 9 obese, diabetic subjects to study the molecular changes in the diabetic fatty liver.

Affymetrix Microarray Analysis

The raw data for 9 obese, diabetic subjects and 5 lean control subjects was re-analyzed with ArrayAnalysis.org, an online microarray quality control and pre-processing pipeline [12]. The data was normalized using the GC-RMA method and further evaluated.
The ArrayAnalysis.org statistics module uses the Limma package [13] of Bioconductor in R which applies linear regression models to make the statistical comparison between obese, diabetic subjects vs. control lean group. Genes were considered to be differentially expressed when their (1) absolute log2 fold change (FC) > 1 and (2) p-value < 0.05.

Gene Ontology Analysis

Gene Ontology (GO) analysis was performed using the GO-Elite web-interface [14] to identify biological processes for the differentially expressed genes in the dataset. The following settings were used: (1) 2000 permutations, (2) Z-score pruning algorithm, (3) Z-score threshold > 1.96, (4) p-value threshold < 0.05 and (5) minimum number of changed genes is 3.

Pathway Analysis

Pathway analysis was performed in PathVisio 3.1.3 (http://www.pathvisio.org) to interpret and visualize the molecular changes on a pathway level. The human pathway collection containing 262 pathways was obtained from WikiPathways (http://www.wikipathways.org). An overrepresentation analysis was performed

using differentially expressed genes. The pathways are ranked based on a standardized difference score (Z-score) using the expected value and standard deviation of the number of differentially expressed genes in a pathway under a hypergeometric distribution. A positive Z-score indicates pathways with a greater number of significantly changed genes than is expected by chance [15]. Pathways were considered significantly changed when (1) Z-score > 1.96, (2) permutated p-value < 0.05 and (3) minimum number of changed genes is 3. Additionally, the log2FC and p-value were visualized on pathways with the visualization module in PathVisio.

Network Analysis

First, all significantly changed pathways were combined and visualized with Cytoscape 3.1.1. The pathways were loaded as networks using the WikiPathways app [16] and an identifier mapping step was performed with the BridgeDb App [17] to unify the identifiers to Ensembl for gene products and HMDB for metabolites. By applying the network merge functionality in Cytoscape the pathways were combined into one integrated network.
Second, the gene expression data was visualized on the nodes of the network. Third, the integrated network was extended with information on transcriptional regulation derived from the ENCODE project [18] using the CyTargetLinker app [19]. Finally, drug-target interactions from DrugBank version 4 [20] were integrated in the network.
The comprehensive visualization functionality in Cytoscape was applied to further explore the complex regulatory mechanisms.

Results

A network biology workflow was developed to decipher the biological processes involved in the human diabetic liver. The results obtained with pathway and network analysis will be explained in more detail.

Differential Expression

In the selected human diabetic fatty liver dataset 11,878 genes were measured and annotated in both lean and obese, diabetic subjects. Statistical analysis showed that 181 genes were differentially expressed (absolute logFC > 1 and p-value < 0.05), of these were 118 up-regulated and 63 down-regulated in the obese, diabetic subjects compared to the lean control group.

The GO analysis was performed with GO-Elite and showed relevant processes for T2DM being over-represented, e.g. triglyceride metabolic process (GO:0006641), cholesterol metabolic process (GO:0008203), glucose metabolic process (GO:0006006), response to glucose stimulus (GO:0009749), cholesterol homeostasis (GO:0042632) and complement activation, classical pathway (GO:0006958). Furthermore, processes related to one-carbon metabolism, humoral immune response, protein-lipid complex subunit organization and organic anion transport were found.

Pathway Analysis

Biological processes in which differentially expressed genes are enriched were identified by performing pathway analysis in PathVisio. The statistical analysis resulted in seven significantly changed pathways (Z-score > 1.96, p-value < 0.05, minimum of three changed genes) (see Table 7.1). Most of these pathways are processes relevant for T2DM but there are also some bigger pathways included, like Proteasome Degradation or Adipogenesis. To illustrate the visualization of the analyzed gene expression data two pathways known to be important in drug treatment of T2DM were selected, i.e. the AMPK Signaling and the Statin pathway (Figure 7.1A and 7.1B).

Table 7.1: Seven Pathways Changed in the Diabetic Liver. Pathway statistics in PathVisio revealed seven significantly altered pathways (Z-score > 1.96, P-value < 0.05, minimum of 3 changed genes). The number of genes (# Genes) represent the number of differentially expressed genes in the pathway compared to the total number of measured genes in the pathway. The arrows indicate up (↑) and down-regulation (↓).

Pathway	Z-score	P-value	# Genes	Genes
Triacylglyceride Synthesis	3.78	0.001	3 / 19	↑ AGPAT2, GPD1, DGAT1
Proteasome Degradation	3.32	0.006	5 / 53	↑RPN1, PSMB3, HLA-B, HLA-E, HLA-J
Statin Pathway	3.10	0.006	3 / 25	↑DGAT1, APOA4, CYP7A1
Fluoropyrimidine Activity	2.84	0.013	3 / 28	↑SLC22A7 ↓ABCG2, DPYD
Pathogenic Escherichia coli infection	2.76	0.011	4 / 46	↑ ARPC1A, ARPC1B, ACTB ↓ ROCK1
Adipogenesis	2.41	0.016	7 / 121	↑ SREBF1, CDKN1A, NR1H3, PNPLA3, AGPAT2 ↓ CISD1, ZMPSTE24
AMPK Signaling	2.38	0.029	4 / 54	↑ SREBF1, P21 ↓ LEPR, PFKFB3

In the AMPK Signaling pathway the gene expression of the upstream regulating kinases of AMPK, i.e, CAMKK, LKB1, MO25 and STRADA, were significantly up-regulated. Moreover, the expression of the glucose transport protein 4 (GLUT4) is significantly increased together with an increase in the GLUT4 enhancer factor (GEF; p-value = 0.057). Most downstream AMPK targets were up-regulated, i.e., HNF4A, SREBF1, eEF2, TSC2, p21 and some are down-regulated, like PFK2 and TSC1.
In the Statin pathway the inhibitory action of statin, a cholesterol-lowering drug, on HMG-CoA reductase (HMGCR) is depicted. The expression of HMGCR remains unaltered in the diabetic fatty liver compared to lean controls. Moreover, cholesterol synthesis is described in the pathway and almost all differentially expressed genes are up-regulated, like DGAT1, CYP7A1, SCARB1, LCAT and APOA4.

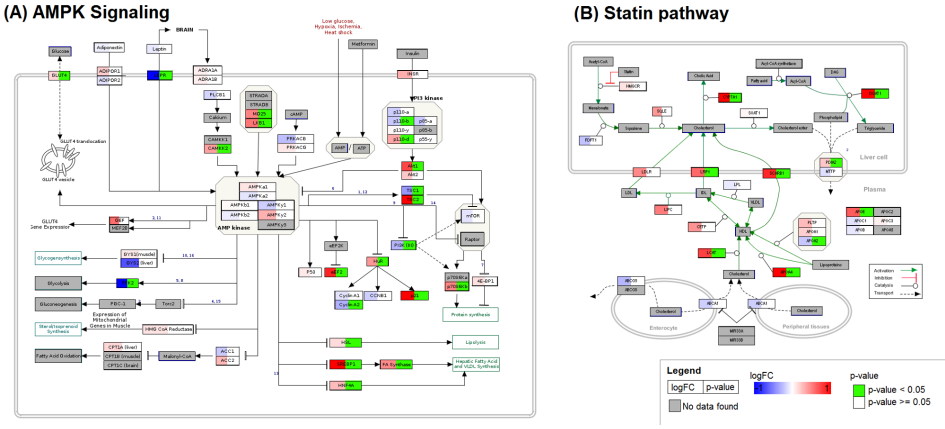


Figure 7.1: Visualization of Two Pathways Relevant for Drug Treatment of T2DM. Gene expression is visualized on (A) AMPK Signaling pathway, http://www.wikipathways.org/instance/WP1403 and (B) Statin pathway, http://www.wikipathways.org/instance/WP430 from WikiPathways. The visualization of the gene product boxes in the pathways is split into two parts, (1) the log2 FC in the left part of the box (blue is down-regulated over white is not changed to red is up-regulated) and (2) the p-value in the right part of the box (green when significant). Pathway elements including metabolites that have not been measured in the selected dataset are gray.

Network Analysis

Pathway integration.
Pathway analysis revealed seven pathways with a Z-score > 1.96 (see Table 7.1) which were then combined into one biological network and analyzed in the network visualization and analysis tool Cytoscape, see Figure 7.2. The created network contains 642 edges connecting 580 nodes, consisting of 408 gene products, 38 metabolites and 5 pathway nodes. 129 nodes are visualized as very small nodes to represent groups and complexes as well as complex interactions in the pathways. Pathways from WikiPathways can contain pathway nodes that link to other pathways. The created network has therefore links to five other pathways: Glycolysis (WP534), DNA Repair (WP1805), Fatty Acid Oxidation (WP143), Fatty Acid Synthesis (WP357) and Apoptosis (WP254).
Fourteen genes in the network are significantly up regulated in obese, diabetic subjects including three genes linking two or more pathways such as AGPAT2, CDKN1A and SREBF1. Seven genes, all present in only one pathway, are significantly down regulated.

Figure 7.3 shows how fourteen genes and three metabolites are linking two or more of the selected pathways to each other. The transcriptomics dataset comparing obese, diabetic subjects with lean subjects is visualized in the network. The log2FC is indicated by a color gradient on the nodes and significance (p-value $<$ 0.05) is represented by a light-green border. Nodes including metabolites without a measurement in the dataset are colored gray.
In the liver of obese, diabetic subjects, the gene expression of three of the linker genes, i.e. AGPAT2, CDKN1A and SREBF1, is significantly ($p < 0.05$) up-

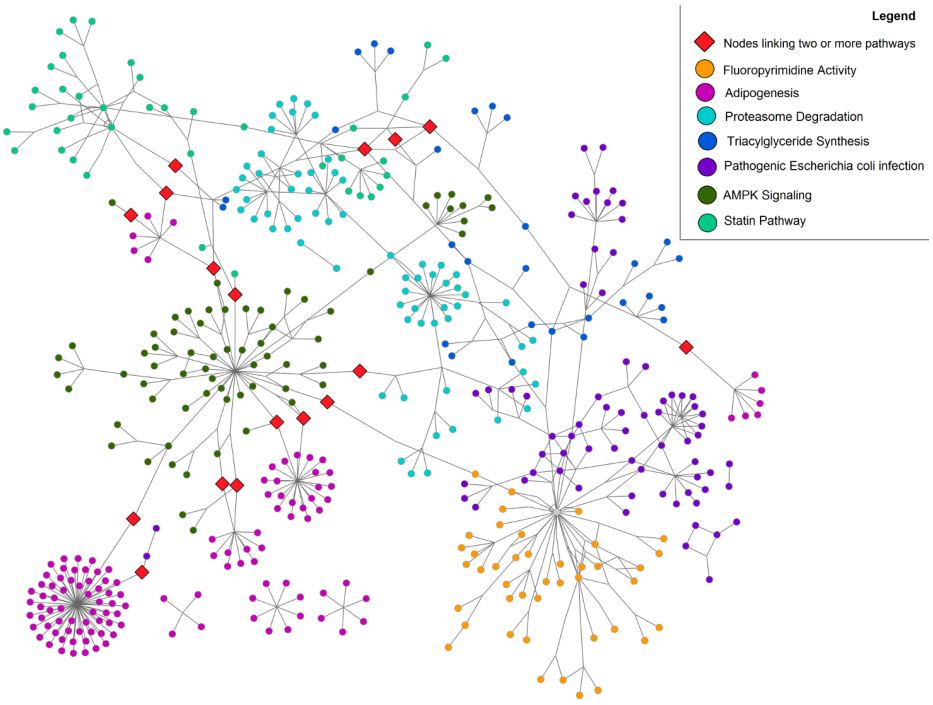


Figure 7.2: Integrated Network of Seven Interconnected Pathways that are Changed in the Diabetic Fatty Liver.

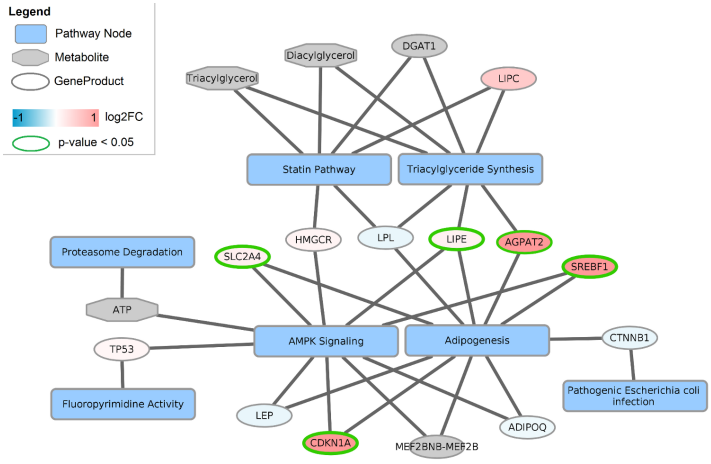


Figure 7.3: Nodes Linking the Seven Significantly Changed Pathways. Each pathway is represented as a yellow rounded rectangle. Gene products and metabolites are visualized as ellipses and octagons, respectively. The transcription dataset is visualized on the gene nodes in the network using a color gradient from blue (down-regulated) over white (not changed) to red (up-regulated). Nodes with a significant p-value (< 0.05) have a light-green border color. Most nodes linking multiple pathways are either up-regulated (e.g. SREBF1, CDKN1A, AGPAT2) or not altered significantly (e.g. LPL, HMGCR).

regulated. AGPAT2 is an enzyme that plays an important role in the production of glycerophospholipids and triacylglycerols. It is known to be relevant to the liver and development of hepatic steatosis [21]. CDKN1A is a potent cell cycle inhibitor important for the induction and maintenance of cellular senescence. SREBF1 is a TF regulating genes required for glucose and fatty acids metabolism and lipid production. Studies showed a clear link between mutations in CDKN1A and SREBF1 and the risk of developing NAFLD [22, 23] strengthening the involvement of these genes in diabetic fatty liver.

Extension with transcriptional regulation.
The integrated network was extended with transcriptional regulation to obtain a better insight in how biological processes affected in the human fatty liver are regulated. The CyTargetLinker app in Cytoscape was used to extend the network with proximal TF-target interactions from the ENCODE project [24]. The app identified sixteen nodes in the network as TFs, most of which are nodes present in only one pathway, except for SREBF1. All TFs are present in either the AMPK Signaling pathway or the Adipogenesis pathway.

Figure 7.4 shows a network containing the sixteen TFs as diamonds and 90 of their targets which are present in one of the selected seven pathways. The interactions in this network are not present in the pathways but have been reported in the EN-

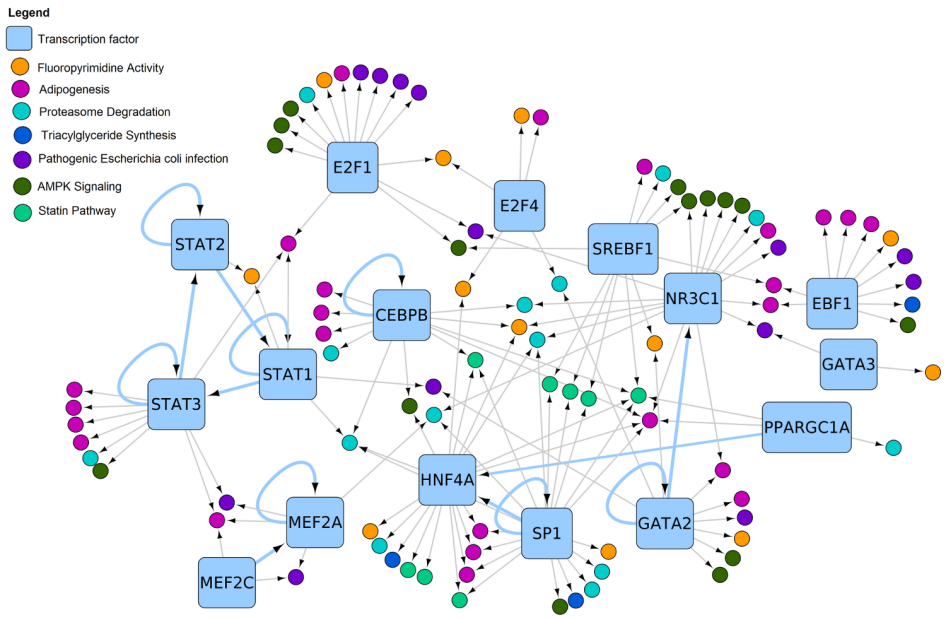


Figure 7.4: TF Regulation in the Diabetic Fatty Liver Pathways. Using CyTargetLinker, sixteen TFs have been identified in the seven pathways. TFs are visualized as rounded rectangles and their target genes as circles colored based on their presence in different pathways. 56 genes are targeted by only one TF and 33 genes are targeted by 2 or more TFs. Light-blue edges indicate regulation of TFs by other TFs.

CODE regulatory network derived by Gerstein *et al.* [18]. In the initial network, pathway elements were not present in more than three out of seven pathways, however with the CyTargetLinker app, TFs were identified that target genes in up to six out of seven different pathways, like SP1 or HNF4A. This approach also found some TFs regulating other TFs (highlighted as light-blue edges) and discovered typical network motifs like feed-forward loops (e.g. STAT1 → STAT3 → STAT2 → STAT1) or self-regulation (e.g. SP1, GATA2 or CEBPB). Three of the TFs in the network are up-regulated in obese, diabetic subjects, SREBF1 (log2FC: 1.05, p-value: 0.03), STAT3 (log2FC: 0.88, p-value: 0.008) and CEBPB (log2FC: 0.41, p-value: 0.01). All three have been reported to play a role in the development of NAFLD [25–27]

Extension with drug-target information.
CyTargetLinker provides a regulatory interaction network for drug-target interactions from DrugBank. The combined network was extended only with approved drugs leaving out the ones that are withdrawn or experimental. In total 280 drugs were added targeting 76 gene products in the pathways, see Figure 7.5. Based on the categories used in DrugBank the drugs associated with the treatment of diabetes (= antidiabetic and hypoglycemic agents; colored in red), dietary supplements/micronutrients (colored in green), immune response related (colored in orange) and anticholesteremic agents (colored in purple) are highlighted in Figure 7.5. In the extended network, the amount of drug targets related to diabetes is significantly higher than expected by random.

The insulin receptor (INSR) is targeted by nine drugs of which eight are categorized in DrugBank as antidiabetic and/or hypoglycemic agents and the ninth, Mecasermin, is an insulin-like growth factor used for long-term treatment of growth failure in children with severe primary IGF-1 deficiency [28]. The INSR is activated by insulin binding and after activation it phosphorylates and thereby activates insulin receptor substrate 1 (IRS1) which in turn activates the PI3-kinase and AKT signalling pathways. Insulin is a natural hormone produced by beta cells in the pancreas and has many functions, e.g. promotion of cellular uptake of glucose and energy storage via glycogenesis. T2DM patients who are unable to control their glucose levels can be treated with insulin analogues, like Insulin Aspart, Detemir or Glargine and these are indeed drugs targeting the INSR in the drug-target network. Insulin Aspart is a fast-acting analogue simulating the insulin spikes following meals in non-diabetic subjects. Insulin Detemir, on the other hand, is a long-acting analogue used to maintain the basal insulin level in diabetes patients. Insulin Glargine is released in small doses from microprecipitates to achieve a long duration of action.
Furthermore, two diabetes related drugs, Phenformin and Metformin, target the two subunits of AMPK, PRKAA1 and PRKAB1. Both subunits are also activated by the dietary supplement, adenosine monophosphate (AMP). The network shows that several hypoglycemic agents, like Rosiglitazone, Glipizide or Pioglitazone, are known to target PPARG, a receptor which regulates fatty acid storage and glucose metabolism. All these drugs are known to be prescribed to T2DM patients depending on the state of the disease.

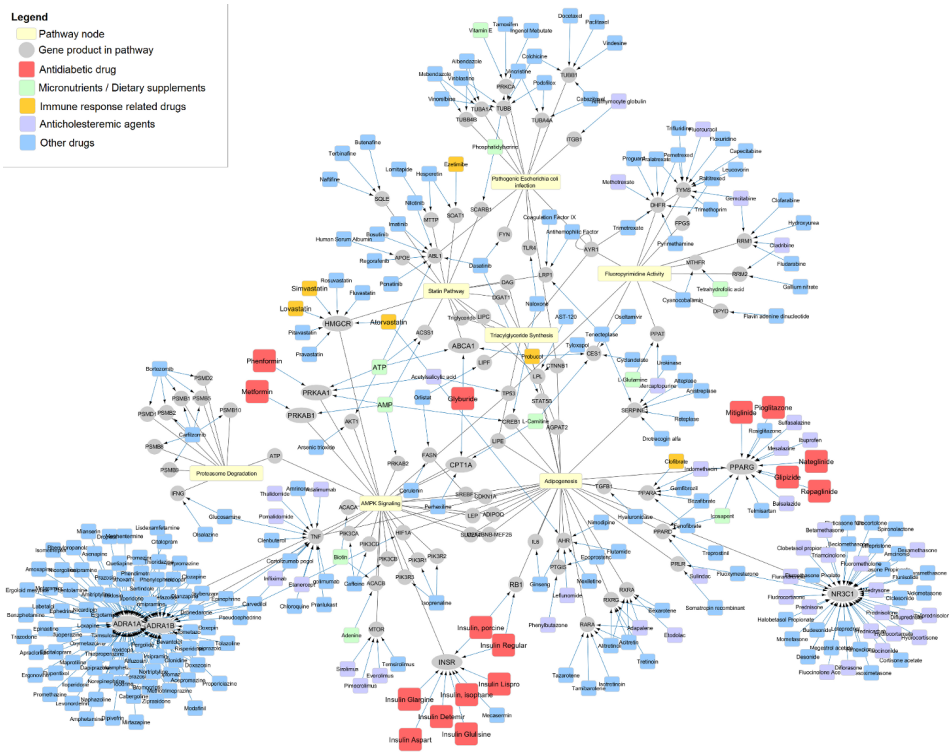


Figure 7.5: Significant Amount of Antidiabetic Drugs Targeting Gene Products in the Network. The network has been extended with drug-target interactions from DrugBank 4 using CyTargetLinker. Nodes present in only one pathway and not targeted by any drugs have been grouped in pathway nodes (yellow rounded rectangles). Drugs targeting genes in the network are indicated as blue rectangles, drugs associated with diabetes are colored in red, micronutrients/dietary supplements in green, drugs related to immune response in orange and anticholesteremic agents in purple. Diabetes related drugs target 7 gene products in the network: INSR (8 drugs), PPARG (5 drugs), RB1 (2 drugs), ABCA1 (1 drug), CPT1A (1 drug), PRKAA1 (1 drug), PRKAB1 (1 drug).

Discussion

It has been shown that insulin resistance conditions like obesity and T2DM are strongly associated with accumulation of lipids in the liver which is the main characteristics of NAFLD [29, 30]. Although it is known that impaired substrate metabolism is involved in the development of the fatty liver in T2DM, the exact mechanisms remain unclear. In this study we applied a network biology approach to investigate the molecular mechanisms in the diabetic fatty liver.

Biological pathways are useful to better understand the mechanism affected in a disease state. As illustrated in this study researchers can investigate pathways in diseased subjects compared to control subjects to gain more insights into the causes of a disease and the mechanisms of disease progression. Although the pathway collections are covering many well described biological mechanisms they are

incomplete resulting in a bias towards well studied processes. Nevertheless, pathway databases are growing and pathway analysis has proven itself as a valid and intuitive first step in the analysis process. Wiki-based community curated pathway databases like WikiPathways reduce the barrier to participate in pathway curation and allow experts to add new findings immediately to the pathway diagrams.
While standard pathway analysis investigates each pathway individually, biological processes are not independent but interact and influence each other. Therefore it is relevant to investigate the links between them as well as shared regulatory mechanisms. This study describes an workflow to further explore the interplay between pathways involved in the human fatty liver by combining them in a biological network and extending them with additional knowledge on TF regulation and drug targeting.

Pathway analysis demonstrated that Triacylglyceride Synthesis and Adipogenesis are significantly altered in obese, diabetic subjects with a fatty liver compared to the lean control group. Interestingly, two pathways related to drug treatment in T2DM, i.e., the AMPK signaling pathway and the Statin pathway, are among the significantly altered pathways.
AMPK is known as the metabolic regulator and its activation influences many metabolic processes [31]. Under catabolic situations, like in a fasted state, AMPK is activated thereby increasing glycolysis and FA beta-oxidation and decreasing FA synthesis [32]. The liver biopsies measured in the selected dataset were collected after overnight fasting. Phosphorylation of AMPK is activated by the upstream regulators, CAMKK2 and LKB1 in complex with MO25 and STRADA [33]. The gene expression of all upstream regulators of AMPK are significantly up-regulated strongly indicating that AMPK is activated in the human diabetic liver.
In the Statin pathway the expression of genes involved in the cholesterol and triacylglycerol production, DGAT1, CYP7A1, SCARB1, LCAT and APOA4, are all significantly up-regulated in the diabetic liver in humans. These findings indicate that hepatic lipid accumulation is facilitated by increased expression of these genes.

Regulatory elements like TFs are often not included in pathway diagrams to keep them comprehensible. Nevertheless, as this study confirms, TFs play a crucial role in the understanding of complex diseases since they often regulate multiple pathways simultaneously. Our analysis showed that TFs can be considered additional links between pathways and adding the regulatory interactions increases the overall connectivity of the network significantly. Typical TF network motifs, e.g. feed-forward loops, single input module or self-regulation can be identified using standard network algorithms.
Hepatocyte nuclear factor 4-alpha (HNF4A) is an essential TF in the extended regulatory network (see Figure 7.4) which is regulated by two TFs, i.e., PPARGC1A and SP1 and regulates genes in 6 out of 7 pathways (not in the E.coli infection pathway). HNF4A is a nuclear receptor (NR) which is a key regulator of the liver cell function, a sensor of inflammation and known to regulate genes in lipid and glucose metabolism [34].
Furthermore, CyTargetLinker revealed 93 additional TFs, not yet present in one of the pathways, that target 212 nodes in the network. TFs generally regulate

multiple targets and so more than 800 regulatory interactions are added. Two hub TFs targeting more than 35 genes in the pathways are CTCF and EP300. CTCF is a general TF which has been reported to mediate the effect of insulin on glucagon expression and therefore is a possible new target for diabetes treatment [35]. EP300 is a general TF regulating cell growth and division and has been reported as a key participant in hepatic steatosis [27].

The extension of the network with drug-target information resulted in a higher number of known drugs used to treat T2DM than expected which confirms the validity of the approach described in this study. The network shows where the drug targets and what effect can be expected in a pathway. It might be possible to identify other drugs with similar pharmacological effects or advantageous drug combinations when targeting at two positions in the network to get a stronger effect of the treatment, e.g. combination of Glipizide (PPARG) and Metformin (AMPK). It might suggest new drugs targeting pathway elements upstream, downstream or even parallel to the currently used targets. When studying the effects of drugs on a whole pathway the grouped pathway nodes in Figure 7.5 can be expanded again to see the complete pathway and its interactions.
Besides the diabetes related drugs, there are many other drugs known to target genes in the selected pathways. HMGCR, a highly targeted gene product, is part of the Statin pathway depicted in Figure 7.1A and targeted by seven drugs, all belonging to the statin family. This drug family has a cholesterol lowering effect and is often used in combination with anti-diabetic drugs [36]. Furthermore three gene products in the network are highly targeted by more than 40 drugs, ADRA1A, ADRA1B and NR3C1. ADRA1A and ADRA1B are members of a subfamily of the G protein-coupled receptors (GPCRs) and there are several studies investigating the potential of GPCRs for treatment of T2DM [37, 38]. NR3C1 is part of the NR superfamily of TFs that are known to play a role in development and adaptations to liver diseases. NRs are also suggested as potential drug targets for treatment of diabetes and NAFLD [39].

In general, the described workflow can be applied for different diseases and pathway sets. Integrating the pathway information into the network allows researchers to investigate downstream effects of drugs and contributes to the identification of other treatment possibilities. Also the link with other pathways, affected or not affected by the disease state, is of importance to predict possible side-effects of a drug. Including the information about the effects of other drugs in linked or related pathways can help to identify interferences with the action and efficiency of the drug.

Conclusions

In this study we demonstrated how pathway analysis results, which are often considered a final step in the biological interpretation of transcriptomics data, can be used and combined in a biological network to gain more insights in the interplay and relation between processes. Instead of starting with a large protein-protein interaction network and finding the important parts in it, we believe that building

the networks based on relevant pathways can be another very useful approach to start the investigation. Also the inclusion of all elements present in the pathway, e.g. metabolites, provides a framework which can integrate different types of omics-data. The biological interpretation might be more straight-forward because it builds on the pathway diagrams which are usually intuitive and well studied.
Regulation by TFs or drugs does not only have effects on one pathway but also has effects on downstream processes. This integration leads to improved insight and also a much clearer illustration of the overall process, and the most important elements. Inclusion of information about drugs and micronutrients and their targets makes the mode of action of currently used compounds more understandable and can be useful to suggest drug repositioning and new drugs or micronutrient related lifestyle interventions.
The tools used in this study, especially PathVisio, Cytoscape and CyTargetLinker, facilitated the data integration, visualization and interpretation immensely.

Bibliography

[1] T Takamura, H Misu, T Ota, and S Kaneko. Fatty liver as a consequence and cause of insulin resistance: lessons from type 2 diabetic liver. *Endocrine journal*, 59(9):745–63, September 2012.

[2] P Iozzo, K Hallsten, V Oikonen, KA Virtanen, J Kemppainen, O Solin, E Ferrannini, J Knuuti, and P Nuutila. Insulin-mediated hepatic glucose uptake is impaired in type 2 diabetes: evidence for a relationship with glycemic control. *The Journal of clinical endocrinology and metabolism*, 88(5):2055–60, May 2003.

[3] R Basu, V Chandramouli, B Dicke, B Landau, and R Rizza. Obesity and type 2 diabetes impair insulin-induced suppression of glycogenolysis as well as gluconeogenesis. *Diabetes*, 54(7):1942–8, July 2005.

[4] M Kohjima, M Enjoji, N Higuchi, M Kato, K Kotoh, T Yoshimoto, T Fujino, M Yada, R Yada, N Harada, R Takayanagi, and M Nakamuta. Re-evaluation of fatty acid metabolism-related gene expression in nonalcoholic fatty liver disease. *International journal of molecular medicine*, 20(3):351–8, September 2007.

[5] H Misu, T Takamura, N Matsuzawa, A Shimizu, T Ota, M Sakurai, H Ando, K Arai, T Yamashita, M Honda, and S Kaneko. Genes involved in oxidative phosphorylation are coordinately upregulated with fasting hyperglycaemia in livers of patients with type 2 diabetes. *Diabetologia*, 50(2):268–77, February 2007.

[6] T Takamura, H Misu, N Matsuzawa-Nagata, M Sakurai, T Ota, A Shimizu, S Kurita, Y Takeshita, H Ando, M Honda, and S Kaneko. Obesity upregulates genes involved in oxidative phosphorylation in livers of diabetic patients. *Obesity (Silver Spring, Md.)*, 16(12):2601–9, December 2008.

[7] J Pihlajamäki, T Boes, E-Y Kim, F Dearie, BW Kim, J Schroeder, E Mun, I Nasser, PJ Park, AC Bianco, AB Goldfine, and ME Patti. Thyroid hormone-related regulation of gene expression in human fatty liver. *The Journal of clinical endocrinology and metabolism*, 94(9):3521–9, September 2009.

[8] P Khatri, M Sirota, and AJ Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, January 2012.

[9] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[10] MP van Iersel, T Kelder, AR Pico, K Hanspers, S Coort, BR Conklin, and C Evelo. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics*, 9:399, January 2008.

[11] ME Smoot, K Ono, J Ruscheinski, P-L Wang, and T Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–2, February 2011.

[12] LMT Eijssen, M Jaillard, ME Adriaens, S Gaj, PJ de Groot, M Müller, and CT Evelo. User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. *Nucleic acids research*, 41(Web Server issue):W71–6, July 2013.

[13] GK Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3, January 2004.

[14] AC Zambon, S Gaj, I Ho, K Hanspers, K Vranizan, CT Evelo, BR Conklin, AR Pico, and N Salomonis. GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics (Oxford, England)*, 28(16):2209–10, August 2012.

[15] SW Doniger, N Salomonis, KD Dahlquist, K Vranizan, SC Lawlor, and BR Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome biology*, 4(1):R7, January 2003.

[16] M Kutmon, S Lotia, CT Evelo, and AR Pico. WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research*, 3, July 2014.

[17] J Gao, C Zhang, M van Iersel, L Zhang, D Xu, N Schultz, and AR Pico. BridgeDb app: unifying identifier mapping services for Cytoscape. *F1000Research*, 3, July 2014.

[18] MB Gerstein, A Kundaje, M Hariharan, SG Landt, KK Yan, C Cheng, XJ Mu, E Khurana, J Rozowsky, R Alexander, R Min, P Alves, A Abyzov, N Addleman, N Bhardwaj, AP Boyle, P Cayting, A Charos, DZ Chen, Y Cheng, D Clarke, C Eastman, G Euskirchen, S Frietze, Y Fu, J Gertz, F Grubert, A Harmanci, P Jain, M Kasowski, P Lacroute, J Leng, J Lian, H Monahan, H O'Geen, Z Ouyang, EC Partridge, D Patacsil, F Pauli, D Raha, L Ramirez, TE Reddy, B Reed, M Shi, T Slifer, J Wang, L Wu, X Yang, KY Yip, G Zilberman-Schapira, S Batzoglou, A Sidow, PJ Farnham, RM Myers, SM Weissman, and M Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, September 2012.

[19] M Kutmon, T Kelder, P Mandaviya, CTA Evelo, and SL Coort. CyTargetLinker: a Cytoscape app to integrate regulatory interactions in network analysis. *PloS one*, 8(12):e82160, January 2013.

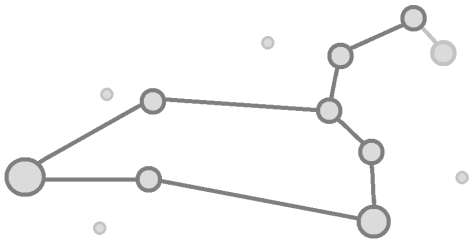
[20] C Knox, V Law, T Jewison, P Liu, S Ly, A Frolkis, A Pon, K Banco, C Mak, V Neveu, Y Djoumbou, R Eisner, AC Guo, and DS Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(Database issue):D1035–41, January 2011.

[21] GG Schweitzer and BN Finck. Targeting Hepatic Glycerolipid Synthesis and Turnover to Treat Fatty Liver Disease. *Advances in Hepatology*, 2014.

[22] G Musso, M Cassader, S Bo, F De Michieli, and R Gambino. Sterol regulatory element-binding factor 2 (SREBF-2) predicts 7-year NAFLD incidence and severity of liver disease and lipoprotein and glucose dysmetabolism. *Diabetes*, 62(4):1109–20, April 2013.
[23] A Aravinthan, G Mells, M Allison, J Leathart, A Kotronen, H Yki-Jarvinen, AK Daly, CP Day, QM Anstee, and G Alexander. Gene polymorphisms of cellular senescence marker p21 and disease progression in non-alcohol-related fatty liver disease. *Cell cycle (Georgetown, Tex.)*, 13(9):1489–94, May 2014.
[24] The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–40, October 2004.
[25] HY Quan, DY Kim, SJ Kim, HK Jo, GW Kim, and SH Chung. Betulinic acid alleviates non-alcoholic fatty liver by inhibiting SREBP1 activity via the AMPK-mTOR-SREBP signaling pathway. *Biochemical pharmacology*, 85(9):1330–40, May 2013.
[26] S Sookoian, G Castaño, TF Gianotti, C Gemma, MS Rosselli, and CJ Pirola. Genetic variants in STAT3 are associated with nonalcoholic fatty liver disease. *Cytokine*, 44(1):201–6, October 2008.
[27] J Jin, P Iakova, M Breaux, E Sullivan, N Jawanmardi, D Chen, Y Jiang, EM Medrano, and NA Timchenko. Increased expression of enzymes of triglyceride synthesis is essential for the development of hepatic steatosis. *Cell reports*, 3(3):831–43, March 2013.
[28] D Fintini, C Brufani, and M Cappa. Profile of mecasermin for the long-term treatment of growth failure in children and adolescents with severe primary IGF-1 deficiency. *Therapeutics and clinical risk management*, 5(3):553–9, June 2009.
[29] JK Dowman, JW Tomlinson, and PN Newsome. Pathogenesis of non-alcoholic fatty liver disease. *QJM : monthly journal of the Association of Physicians*, 103(2):71–83, February 2010.
[30] B Fruci, S Giuliano, A Mazza, R Malaguarnera, and A Belfiore. Nonalcoholic Fatty liver: a possible new target for type 2 diabetes prevention and treatment. *International journal of molecular sciences*, 14(11):22933–66, January 2013.
[31] DG Hardie. AMPK: a key regulator of energy balance in the single cell and the whole organism. *International journal of obesity (2005)*, 32 Suppl 4:S7–12, September 2008.
[32] B Viollet, B Guigas, J Leclerc, S Hébrard, L Lantier, R Mounier, F Andreelli, and M Foretz. AMP-activated protein kinase in the regulation of hepatic energy metabolism: from physiology to therapeutic perspectives. *Acta physiologica (Oxford, England)*, 196(1):81–98, May 2009.
[33] SA Hawley, J Boudeau, JL Reid, KJ Mustard, L Udd, TP Mäkelä, DR Alessi, and DG Hardie. Complexes between the LKB1 tumor suppressor, STRAD alpha/beta and MO25 alpha/beta are upstream kinases in the AMP-activated protein kinase cascade. *Journal of biology*, 2(4):28, January 2003.
[34] JP Babeu and F Boudreau. Hepatocyte nuclear factor 4-alpha involvement in liver and intestinal inflammatory networks. *World journal of gastroenterology : WJG*, 20(1):22–30, January 2014.
[35] S Tsui, J Gao, C Wang, and L Lu. CTCF mediates effect of insulin on glucagon expression. *Experimental cell research*, 318(8):887–95, May 2012.
[36] J Pang, DC Chan, and GF Watts. Origin and therapy for hypertriglyceridaemia in type 2 diabetes. *World journal of diabetes*, 5(2):165–175, April 2014.
[37] MS Winzell and B Ahrén. G-protein-coupled receptors and islet function-implications for treatment of type 2 diabetes. *Pharmacology & therapeutics*, 116(3):437–48, December 2007.
[38] B Ahrén. Islet G protein-coupled receptors as potential targets for treatment of type 2 diabetes. *Nature reviews. Drug discovery*, 8(5):369–85, May 2009.
[39] M Arrese and SJ Karpen. Nuclear receptors, inflammation, and liver disease: insights for cholestatic and fatty liver diseases. *Clinical pharmacology and therapeutics*, 87(4):473–8, April 2010.

Chapter 8

General Discussion



In this thesis we demonstrated the power of pathway and networks to store and visualize knowledge and to analyze and interpret biomedical experiments. The approaches and tools presented provide small but relevant pieces to the systems and network biology fields to reach the final goal of completely understanding complex biological systems.

Biological Data Curation

Nowadays, advances in measuring technologies lead to the production of huge amounts of biological data. Bioinformatics tries to develop tools and automated pipelines to facilitate the analysis of such data. However, the interpretation of the results and the translation into new knowledge and applications is often still a manual process performed by experts. This process is time-consuming, focused on the researchers field of expertise and it is not possible to use all the available resources.

Pathway diagrams have been used by biologists to organize, share and discuss knowledge about biological processes for many years. Instead of looking at thousands of genes, pathway analysis reduced the problem to hundreds of pathways studying biology on a process level instead of on a gene level. Because of the advances in measuring techniques and computational technologies, we are now able to study complex biological systems as large networks. One of the most crucial limitations of pathway but also network analysis is the bias towards what we already know. As mentioned in Chapter 2, pathway databases only cover less than 50% of the known human protein-coding genes and are often missing relevant regulatory elements like microRNAs (miRNAs) or other non-coding RNAs (ncRNAs). Pathway interactions can be combined with interaction databases which focus on binary interactions like miRNA-target or protein-protein interactions and even though the number of interactions is much higher, we are still missing a lot of information.

PathVisio is a pathway editor, analysis and visualization software [1] which allows researchers to create new pathways, curate existing ones and analyze and visualize experimental data on biological pathways. A biological pathway can summarize and visualize the current knowledge about a biological process and all the elements can be fully annotated and referenced. Pathways can be immediately published on WikiPathways [2] which allows other researchers to contribute, discuss and use the pathway in their analysis.

All the chapters in this thesis rely on the community curated pathway database WikiPathways. This database applies the same concepts as used in Wikipedia, like collaboratively editing, collection and curation of knowledge, free content and open access. The main focus of the database is the collection and curation of biological pathways in a form that is both human readable and amenable to computational analysis. This project highly depends on the number of participating experts that contribute pathways to the database. Recent years showed that this approach has been hugely successful. Since its beginning in January 2008, the number of human pathways in WikiPathways has increased from around 100 to over 600 human pathways and WikiPathways now supports 29 species. Figure 8.1 shows the growth of the pathway collections for human, mouse, rat and zebrafish since the

start of WikiPathways in 2008. While the human pathway collection still has the most active curators, other species are growing and the collections become more extensive. The recent integration of the human Reactome [3] pathway collection into WikiPathways to enable the community to contribute to the curation efforts has added more than 200 pathways to the collection.

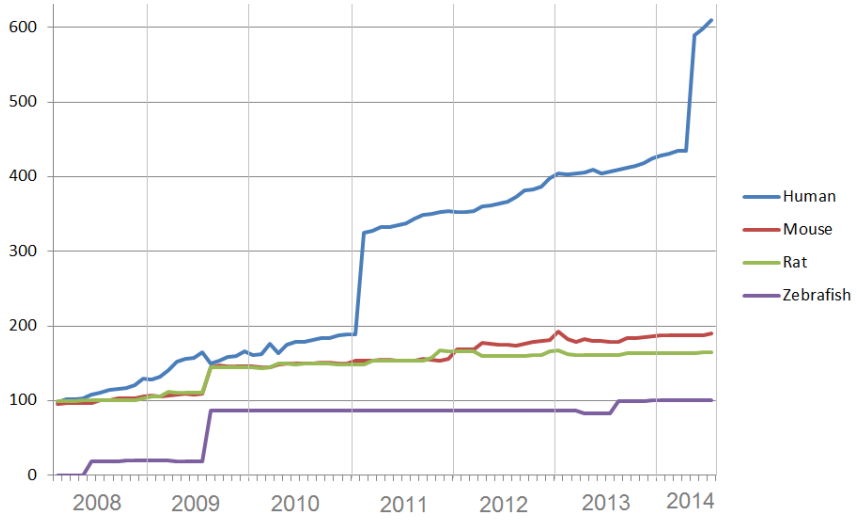


Figure 8.1: Growth of WikiPathways Pathway Collections for Human, Mouse, Rat and Zebrafish. In the last 6 years, the human collection grew from 100 to 600 pathways, the mouse collection now reached 190, the rat collection 165 and the zebrafish collection 100 pathways. The statistics for all collections can be found on http://wikipathways.org/index.php/WikiPathways:Statistics

Chapter 3 demonstrated how powerful pathway diagrams are to organize and publish biological knowledge. It allowed us to go beyond the basic SREBP signalling pathway to study the regulatory mechanisms of this pathway. We were able to identify many links between SREBP and other regulators of lipid, protein and carbohydrate metabolism as well as overall energy homeostasis. Submitting the pathway to WikiPathways allows other researchers to discuss, adapt and further extend the pathway. The pathway serves not only as an overview but also as a dashboard to relevant publications and entries in genomics databases. Additionally it can be used for advanced data analysis and visualization in PathVisio.
In 2006, Bader *et al.* started to collect all pathway and interaction related resources on http://www.pathguide.org [4]. At present, PathGuide lists nearly 550 resources containing biological pathways as well as protein-protein, gene regulatory and protein-compound interactions. Biological pathways are network-like in nature, making them a valuable resource for network analysis. Other typical interaction resources contain mostly binary interactions and they are much larger than pathway collections. The interaction data is often produced by combining literature research and prediction algorithms. Data extracted from literature is extremely noisy and prediction algorithms often include assumptions that might not reflect the actual mechanisms. Therefore much more validation is needed to improve the quality of interaction data. As an example, the Interactome Project

[5] at the Center for Cancer Systems Biology at the Dana-Farber Cancer Institute released an experimentally validated human interactome (protein-protein interactions) consisting of 3,882 binary interactions. This year they are planning to increase the number to 17,000 experimentally validated protein-protein interactions (HI-II-14) with the goal to create a complete human protein-protein reference network. Similar projects are running for plants like *Arabidopsis thaliana* [6], the worm interactome of *C. elegans* [7] or the yeast interactome [8]. Such efforts are of high importance to improve the quality of the data.

Data Integration

Data integration has been a major challenge in bioinformatics for many years. In 2014, the database issue of Nucleic Acids Research (NAR) reported a growing number of data resources now reaching 1552 databases [9]. Especially the diversity and distribution of biological data resources creates a problem for bioinformaticians to combine and integrate data. While it is important to collect and publish biological knowledge there is also a danger. The web makes is very easy to publish a resource and being a resource provider often increases ones reputation, however because of the diversity of the discipline there is a large number of very specialized resources and only a few centralized data centers. Maintaining and updating a resource is costly and time-consuming and sometimes small specialized resources end up as orphaned databases that are not supported anymore.
Since every database has its own data model and data access methods, bioinformaticians need to learn many different data models and use many different data access methods to integrate data from various resources. This also resulted in a large number of resources that try to integrate several primary resources into one database to facilitate the integration process [10]. Standardized formats to share data from different databases, like BioPAX [11] or SBML [12], are envisioned to simplify the data integration process.
Two years ago we developed CyTargetLinker a tool that facilitates the integration of several regulatory interaction resources in Cytoscape [13] providing a visual way to study the overlap between the different resources. Interactions that are present in multiple resources are shown as individual coloured edges, so the researcher immediately sees the origin of the edge. When studying miRNA-gene interactions the number of validated interactions is still limited. However there are prediction algorithms that produce a wealth of possible miRNA-target interactions. Because of different assumptions in the prediction algorithms the different databases have only very little overlap as demonstrated in Chapter 6 (Figure 6.1). Although the algorithms produce non-overlapping results, they may still give (conditionally) valid results. Therefore choosing only one or a many votes approach might lead to wrong conclusions. This shows how important it is to integrate data from different resources and that there is a need for more experimental validation to get higher quality interaction data.
Because of the wealth of biological databases it is often difficult to decide which should be used in an analysis. This decision might also influence the results found and conclusions drawn from an experiment. Chapter 6 presents a tool that provides a simple and visual way to integrate different resources. CyTargetLinker

provides regulatory interaction networks (RegINs) for several different regulatory interaction resources, including transcription factor (TF)-gene interactions from ENCODE or miRNA-target interactions from TargetScan [14] or miRTarBase [15]. The process of generating these RegINs showed how challenging data integration of biological databases is. Often the databases are not up-to-date, no stable identifiers are used or the data can not be easily accessed. There are many efforts to integrate different resources, like GeneMANIA [16], OpenPhacts [17] or Bio2RDF [18]. These will facilitate the development of new RegINs in the future.
In Chapter 7, CyTargetLinker is used to extend a network built from diabetic, fatty liver pathways with TF-target and drug-target interactions. The analysis enabled us to identify additional links between the pathways and key regulators of liver function, inflammation and lipid and glucose metabolism were identified. The extension of the network with drug-target interactions from DrugBank [19] showed a significantly higher number of known diabetes related drugs compared to randomly generated and extended networks which confirms the validity of the approach.
Besides the integration of different resources, the integration of different experimental data is crucial when studying a biological process in all its complexity. Chapter 4 shows how transcriptomics and proteomics data can be combined and integrated in pathway analysis. Although proteomics and metabolomics measuring techniques develop rapidly, many dataset do not contain enough data for an overview on systems biology level. Nevertheless, the combination with large scale transcriptomics data already enables the identification of relevant biological processes which can then be studied in detail. Figure 1.1 in the introduction shows nicely that transcript level (mRNA) is only an intermediate step and in the future, with better proteomics and metabolomics technologies, the number of measurements in proteomics and metabolomics data will increase and it will be more important to be able to analyze and visualize different datasets together.

Open Data, Open Access and Open Source Development

With the rise of the internet it became easy to publish and share data with a wide audience at virtually no cost. Early on several organizations started "*open*" movements like open source, open content, open access and open data. The *open* definition says that a piece of data, content, software is open if anyone is free to use, reuse and redistribute it, sometimes with defined regulations on attribution of authors or document modifications.
Such open movements play a role in a lot of different fields, like finance, environment, cultural but also science. In 2004, twenty five Nobel prize winners sent an open letter to the US Congress stating that

> "*Open access truly expands shared knowledge across scientific fields - it is the best path for accelerating multi-disciplinary breakthroughs in research.*"

The vision of open research is to increase the pace of scientific discovery and encourage innovation. Nowadays, many biomedical journals require the submission of experimental data to one of the publicly available data repositories, when

publishing scientific software it mostly has to be developed under an open-source license and large grant organizations like Welcome Trust, National Institutes of Health (NIH) or the European Union require open-access publications. The results of publicly funded research should also be available to the public.
As an example, Chapter 7 describes a study solely conducted with open source software and open data. The dataset was chosen from the publicly available data repository Gene Expression Omnibus (GEO [20]). There is still a huge amount of information in these data that can be extracted with further analysis. The fact that the data is publicly available allows researchers to apply new and state-of-the-art analysis methods to find new insights about the underlying biology.
Cytoscape and WikiPathways are two major open-source communities in Bioinformatics. They provide public resources to enable the exchange and use of biological networks and pathways. The three software tools described in this thesis, PathVisio, the pathway editor software of WikiPathways and a powerful analysis tool, the Cytoscape WikiPathways App [21] and the CyTargetLinker App [22], are all developed under an open-source license and are freely available.
Cytoscape and PathVisio are two standalone tools that have an extension system to facilitate the integration of new functionality. Modularity is an important aspect in software development to allow others to reuse and integrate existing code instead of reimplementing it anew. Code can only be reused and integrated in another software tool if the code is available under an open-source license. Furthermore it enables developers to share the load of maintaining the software application. If a developer is not able to continue on the project, other developers can take over and make sure the project progresses further. Community building is an important aspect of open source development and for Cytoscape and PathVisio new collaborations emerge easily with groups that develop new extensions for the software applications.
All publications in this thesis are published in open-access journals, results are openly available as supplementary data and all software tools developed are available under an open source license.

Conclusion

Pathways and networks are very useful tools for the analysis of experimental data. The data is put into a biological context and the visual representation in pathway diagrams but also larger networks facilitates the interpretation step.
Pathway analysis is biased towards what is already known and well studied, so pathway curation is still a crucial point. WikiPathways is a community pathway database that allows researchers to create and curate biological pathways themselves and new findings can be added instantaneously. Therefore, the pathway content is updated continuously and pathway analysis will become more and more powerful. Another strength of pathway analysis is the possibility to visualize complex data on the pathway diagrams.
Also networks are very powerful visual and mathematical representations of complex processes and systems. The simplification of such systems as nodes and edges enables the discovery of new mechanisms and can give further insights into the underlying biology. Networks can represent known biological interactions but also

interaction purely derived from experimental data, like in a co-expression networks. Interaction data is still very noisy and more experimental validation is needed to improve the quality. Nevertheless network biology has already been shown to improve our ability to understand complex molecular mechanisms underlying health and how they fail in disease.
In scientific research it is important to collaborate with others and to build on each others work. Sharing, validating and building upon other work is only possible if the data, tools and results are openly available. This thesis follows this principle by creating open source software, using open data to analyse and publishing the results in open access journals.

Bibliography

[1] MP van Iersel, T Kelder, AR Pico, K Hanspers, S Coort, BR Conklin, and C Evelo. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics*, 9:399, January 2008.

[2] T Kelder, MP van Iersel, K Hanspers, M Kutmon, BR Conklin, CT Evelo, and AR Pico. WikiPathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, January 2012.

[3] D Croft, AF Mundo, R Haw, M Milacic, J Weiser, G Wu, M Caudy, P Garapati, M Gillespie, MR Kamdar, B Jassal, S Jupe, L Matthews, B May, S Palatnik, K Rothfels, V Shamovsky, H Song, M Williams, W Birney, H Hermjakob, L Stein, and P D'Eustachio. The Reactome pathway knowledgebase. *Nucleic acids research*, 42(Database issue):D472–7, January 2014.

[4] GD Bader, MP Cary, and C Sander. Pathguide: a pathway resource list. *Nucleic acids research*, 34(Database issue):D504–6, January 2006.

[5] J-F Rual, K Venkatesan, T Hao, T Hirozane-Kishikawa, A Dricot, N Li, GF Berriz, FD Gibbons, M Dreze, N Ayivi-Guedehoussou, N Klitgord, C Simon, M Boxem, S Milstein, J Rosenberg, DS Goldberg, LV Zhang, SL Wong, G Franklin, S Li, JS Albala, J Lim, C Fraughton, E Llamosas, S Cevik, C Bex, P Lamesch, RS Sikorski, J Vandenhaute, HY Zoghbi, A Smolyar, S Bosak, R Sequerra, L Doucette-Stamm, ME Cusick, DE Hill, FP Roth, and M Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8, October 2005.

[6] Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science (New York, N.Y.)*, 333(6042):601–7, July 2011.

[7] N Simonis, J-F Rual, A-R Carvunis, M Tasan, I Lemmens, T Hirozane-Kishikawa, T Hao, JM Sahalie, K Venkatesan, F Gebreab, S Cevik, N Klitgord, C Fan, P Braun, N Li, N Ayivi-Guedehoussou, E Dann, N Bertin, D Szeto, A Dricot, MA Yildirim, C Lin, A-S de Smet, H-L Kao, C Simon, A Smolyar, JS Ahn, M Tewari, M Boxem, S Milstein, H Yu, M Dreze, J Vandenhaute, KC Gunsalus, ME Cusick, DE Hill, J Tavernier, FP Roth, and M Vidal. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nature methods*, 6(1):47–54, January 2009.

[8] H Yu, P Braun, MA Yildirim, I Lemmens, K Venkatesan, J Sahalie, T Hirozane-Kishikawa, F Gebreab, N Li, N Simonis, T Hao, JF Rual, A Dricot, A Vazquez, RR Murray, C Simon, L Tardivo, S Tam, N Svrzikapa, C Fan, AS de Smet, A Motyl, ME Hudson, J Park, X Xin, ME Cusick, T Moore, C Boone, M Snyder, FP Roth, AL Barabási, J Tavernier, DE Hill, and M Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science (New York, N.Y.)*, 322(5898):104–10, October 2008.

[9] XM Fernández-Suárez, DJ Rigden, and MY Galperin. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic acids research*, 42(Database issue):D1–6, January 2014.

[10] C Goble and R Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–93, October 2008.

[11] E Demir, MP Cary, S Paley, K Fukuda, C Lemer, I Vastrik, G Wu, P D'Eustachio, C Schaefer, J Luciano, F Schacherer, I Martinez-Flores, Z Hu, V Jimenez-Jacinto, G Joshi-Tope, K Kandasamy, AC Lopez-Fuentes, H Mi, E Pichler, I Rodchenkov, A Splendiani, S Tkachev, J Zucker, G Gopinath, H Rajasimha, R Ramakrishnan, I Shah, M Syed, N Anwar, O Babur, M Blinov, E Brauner, D Corwin, S Donaldson, F Gibbons, R Goldberg, P Hornbeck, A Luna, P Murray-Rust, E Neumann, O Ruebenacker, M Samwald, M van Iersel, S Wimalaratne, K Allen, B Braun, M Whirl-Carrillo, KH Cheung, K Dahlquist, A Finney, M Gillespie, E Glass, L Gong, R Haw, M Honig, O Hubaut, D Kane, S Krupa, M Kutmon, J Leonard, D Marks, D Merberg, V Petri, A Pico, D Ravenscroft, L Ren, N Shah, M Sunshine, R Tang, R Whaley, S Letovksy, KH Buetow, A Rzhetsky, V Schachter, BS Sobral, U Dogrusoz, S McWeeney, M Aladjem, E Birney, J Collado-Vides, S Goto, M Hucka, N Le Novère, N Maltsev, A Pandey, P Thomas, E Wingender, PD Karp, C Sander, and GD Bader. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–42, September 2010.

[12] M Hucka, A Finney, HM Sauro, H Bolouri, JC Doyle, H Kitano, AP Arkin, BJ Bornstein, D Bray, A Cornish-Bowden, AA Cuellar, S Dronov, ED Gilles, M Ginkel, V Gor, II Goryanin, WJ Hedley, TC Hodgman, J-H Hofmeyr, PJ Hunter, NS Juty, JL Kasberger, A Kremling, U Kummer, N Le Novère, LM Loew, D Lucio, P Mendes, E Minch, ED Mjolsness, Y Nakayama, MR Nelson, PF Nielsen, T Sakurada, JC Schaff, BE Shapiro, TS Shimizu, HD Spence, J Stelling, K Takahashi, M Tomita, J Wagner, and J Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, 19(4):524–31, March 2003.

[13] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504, November 2003.

[14] BP Lewis, CB Burge, and DP Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, January 2005.

[15] S-D Hsu, Y-T Tseng, S Shrestha, Y-L Lin, A Khaleel, C-H Chou, C-F Chu, H-Y Huang, C-M Lin, S-Y Ho, T-Y Jian, F-M Lin, T-H Chang, S-L Weng, K-W Liao, I-E Liao, C-C Liu, and

H-D Huang. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic acids research*, 42(Database issue):D78–85, January 2014.

[16] J Montojo, K Zuberi, H Rodriguez, F Kazi, G Wright, SL Donaldson, Q Morris, and GD Bader. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics (Oxford, England)*, 26(22):2927–8, November 2010.

[17] AJ Williams, L Harland, P Groth, S Pettifer, C Chichester, EL Willighagen, CT Evelo, N Blomberg, G Ecker, C Goble, and B Mons. Open PHACTS: semantic interoperability for drug discovery. *Drug discovery today*, 17(21-22):1188–98, November 2012.

[18] F Belleau, M-A Nolin, N Tourigny, P Rigault, and J Morissette. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–16, October 2008.

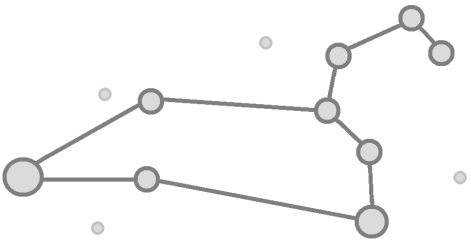
[19] C Knox, V Law, T Jewison, P Liu, S Ly, A Frolkis, A Pon, K Banco, C Mak, V Neveu, Y Djoumbou, R Eisner, AC Guo, and DS Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(Database issue):D1035–41, January 2011.

[20] T Barrett, SE Wilhite, P Ledoux, C Evangelista, IF Kim, M Tomashevsky, KA Marshall, KH Phillippy, PM Sherman, M Holko, A Yefanov, H Lee, N Zhang, CL Robertson, N Serova, S Davis, and A Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5, January 2013.

[21] M Kutmon, S Lotia, CT Evelo, and AR Pico. WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research*, 3, July 2014.

[22] M Kutmon, T Kelder, P Mandaviya, CTA Evelo, and SL Coort. CyTargetLinker: a Cytoscape app to integrate regulatory interactions in network analysis. *PloS one*, 8(12):e82160, January 2013.

Summary



In 2000, Stephen Hawking stated that the 21st century will be the century of complexity. Even though Hawking's field of expertise is physics, this statement is also very true when looking at biomedical research. In the past, researchers often looked at single aspects like how does one element influence the other. But of course there are many different elements in biological systems and the goal to better understand the interactions between the components in a system and their influence on function and behaviour of that system gave rise to a flourishing new research field called Systems Biology. The advances in measuring technologies to determine the abundances of different types of molecules in parallel (e.g. transcriptomics, proteomics, metabolomics) and the developments in computer technology, enable us to start looking at the interactions between the different elements on a system-wide level. In this project, I have studied how pathways and networks can be used to store, integrate, visualize and analyze biological data. The wealth of biological data being produced also requires the development of tools to manage and analyze it. I introduce three bioinformatics tools and show how they can help researchers to better understand and visualize their own data, to integrate data with existing knowledge and therefore to develop better strategies to improve health as well as understand diseases.

Pathway Analysis

Biological pathways are intuitive and graphical representations of the processes that occur in living systems. They describe how genes, proteins and metabolites are all working together. Such pathways are responsible for controlling a cells activities. A series of signals, from one molecule to the next, triggers specific actions like the production of a protein, increased sugar uptake or even cell division. Pathways are really the functional units of any living system and defects in any of the pathways might cause disease. There are diseases that are caused by mutations in one single gene, like Haemophilia which impairs the bodys ability to control blood clotting or coagulation. Diseases like cancer or diabetes, however, are multifactoral and much more complex. Multiple disruption in a single but also in several different pathways, might be required for a disease to develop.

The first half of the thesis focuses on biological pathways and pathway analysis. In **chapter 2**, I introduce the third version of the pathway editor, visualization and analysis software PathVisio. This tool is already widely adopted in the research community and it allows users to create new pathways, visualize experimental data on such pathways as well as identify which pathways are affected in a specific experiment using pathway statistics.
Pathways are very useful for collecting and organizing knowledge. A pathway diagram is much more than just an image, it contains identifiers and literature references for every element in the pathway. The review studying the function and regulation of sterol regulatory element-binding proteins (SREBPS) in **chapter 3** uses a pathway to structure and organize the knowledge from over 50 research articles. The reader can use the interactive pathway viewer on WikiPathways, a collaborative pathway database, to zoom and browse to get detailed information on pathway elements in external databases and thereby allowing a more extensive

study.
The number of studies measuring multi-omics data is increasing and more advanced methods to analyse the data together are required. PathVisio provides functionality to integrate, visualize and analyse multi-omics datasets. In **chapter 4**, two studies on starvation in mice have been combined and analysed using PathVisio, one using proteomics and the other using transcriptomics technology. The integrated identifier mapping solution of PathVisio makes it particularly suited for integration and simultaneous visualization of datasets from different sources. In this example, we showed that proteomics data can reinforce the conclusions deduced from transcriptomics data, and simultaneously indicate areas where post-transcriptional regulation plays a role. The interpretation of such datasets is not straightforward, and pathway visualization can serve as a useful aid, given the role of pathways as a knowledgebase of biological information.

Network Biology

Networks are the universal structure that allows us to describe the relationships and interactions between multiple things. In biology, the regulation of a gene by a transcription factor, the building of a complex of two proteins or a conversion of a metabolite are examples for biological interactions. Such interactions are also often described in pathways which can then be represented as small networks. If we have all the pieces and can put them together, we will be able to for example trace every step of the mechanisms of action of a specific drug or we will be able to find out which nutrient is causing an allergic reaction and why.

The second half of my thesis is mainly focused on biological networks. The widely adopted network analysis and visualization tool Cytoscape, allows developers to provide additional features as so called apps.
Pathways are network-like in nature and the WikiPathways app for Cytoscape, described in **chapter 5**, nicely links the pathway part of my thesis with the network part. The app allows users to load pathways from WikiPathways and PathVisio in Cytoscape and to make full use of the pathway models by performing computational analyses and custom visualizations based on experimental data and network topology.
Pathways usually have a very specific layout to keep them structured and easy to understand and it is therefore hard to automatically extend them with additional interactions. Networks, however, do not have a fixed layout and new interactions can be added easily. Therefore we developed the CyTargetLinker app for Cytoscape to integrate regulatory interactions in network analysis (**chapter 6**). There are many different online resources containing information about regulatory elements like transcription factors, microRNAs or drugs and CyTargetLinker provides an easy and fast way to integrate this knowledge in an existing network (which could also be a pathway). Especially the visualization options enable the biological interpretation of complex regulatory networks in a graphical way. This app can be used in many different studies, like identifying validated and predicted microRNAs targeting genes related to diabetes mellitus, enriching a network of DNA repair genes with transcription factors from ENCODE or even building a

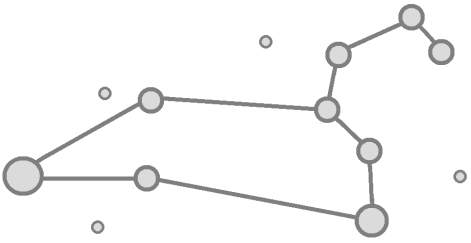
regulatory meta-network by adding a whole regulatory level (transcription factors, microRNAs and drugs) to a biological process.
The incorporation of both extensions into the Cytoscape framework enables their usage in combination with a wide variety of other apps for state-of-the-art network analysis.

In **chapter 7**, it is clearly demonstrated how the tools and approaches described in this thesis can be combined in one bioinformatics workflow. We selected a publicly available transcriptomics dataset to study the obese, diabetic liver. The workflow described uses PathVisio to identify disease affected biological pathways which are then loaded in Cytoscape using the WikiPathways app. The CyTargetLinker app enabled us to incorporate transcription factors and drugs to study the regulatory mechanisms. The whole approach leads to improved insights and a clearer understanding of the overall disease processes.

Conclusion

Pathways and networks are powerful tools to store, organize, integrate, analyze and visualize biological data. It is important to mention that pathway analysis is biased towards what is already known and well studied, and interaction data can be very noisy and more experimental validation is needed to improve the quality. The availability of accessible, high quality experimental data, well organized and structured knowledge and freely available, user-friendly software tools is very important for biomedical research. In this thesis, I presented a number of commonly used bioinformatics tools for pathway and network analysis and demonstrated how they can be applied in different research fields.

Samenvatting



In het jaar 2000 heeft Stephen Hawking opgemerkt dat de 21e eeuw de eeuw van de complexiteit zal worden. Ook al is Hawking een expert op het gebied van de natuurkunde, deze bewering is zonder meer ook van toepassing op het biomedische onderzoeksdomein. In het verleden hebben onderzoekers zich vaak bezig gehouden met kleine deelvragen, zoals hoe een bepaald element invloed heeft op een ander. Maar er zijn vanzelfsprekend heel veel verschillende elementen in biologische systemen, en het doel om de samenwerking tussen al deze componenten in het systeem en hun invloed op de functie en het gedrag ervan beter te begrijpen, hebben geleid tot het bloeiende nieuwe onderzoeksveld van de systeembiologie. De ontwikkelingen in meettechnologieën om verschillende gegevenstypes naast elkaar te verkrijgen (bijvoorbeeld transcriptomics, proteomics, metabolomics) en de technologische ontwikkelingen in de informatica, stellen ons in staat om een begin te maken met het verkennen van de interacties tussen de verschillende elementen op het niveau van het gehele systeem.
In dit project heb ik bestudeerd hoe pathways (cellulaire moleculaire processen) en netwerken gebruikt kunnen worden voor de opslag, integratie, visualisatie, en analyse van biologische gegevens en metingen. De rijkdom aan biologische data die gegenereerd worden, vereist ook de ontwikkeling van gereedschap om deze te beheren en te analyseren. Ik introduceer drie bioinformatica tools en laat zien hoe deze onderzoekers kunnen helpen om hun eigen meetgegevens beter te kunnen weergeven en begrijpen, deze met bestaande kennis te integreren, en daarmee betere strategieën te ontwikkelen om gezondheid te bevorderen en ziekteprocessen te begrijpen.

Pathway analyse

Biologische pathways vormen een intuïtieve en grafische weergave van de processen die optreden in levende systemen. Ze beschrijven hoe genen, eiwitten, en metabolieten allemaal met elkaar samenwerken. Dergelijke pathways zijn verantwoordelijk voor het aansturen van de activiteiten van een cel. Een serie signalen, van één molecule naar de volgende, leidt tot specifieke acties zoals de aanmaak van een eiwit, verhoogde opname van suikers, of zelfs celdeling. Pathways zijn de werkelijke functionele eenheden van elk levend systeem en fouten in elk van de pathways kunnen tot het ontstaan van ziekte leiden. Sommige ziekten worden veroorzaakt door mutaties in één gen, zoals hemofilie, een ziektebeeld waarbij de bloedstolling is aangedaan. Ziekten zoals kanker of diabetes (suikerziekte), zijn daarentegen veel complexer. Meervoudige verstoringen in een enkele of zelfs in meerdere verschillende pathways kunnen nodig zijn om tot de ontwikkeling van de ziekte te leiden.

De eerste helft van deze thesis heeft betrekking op biologische pathways en pathway analyse. In **hoofdstuk 2** introduceer ik de derde versie van de PathVisio software, waarmee pathways bewerkt, weergegeven en geanalyseerd kunnen worden. PathVisio is al veelgebruikt in de onderzoeksgemeenschap en stelt gebruikers in staat om nieuwe pathways te creëren, experimentele meetgegevens op pathways te visualiseren en met pathway statistiek te bepalen welke pathways aangedaan zijn in een bepaald experiment.

Pathways zijn erg nuttig voor het verzamelen en organiseren van kennis. Een pathway diagram is veel meer dan enkel een plaatje, het bevat identificatiecodes en literatuurverwijzingen voor elk element in de pathway. Het overzichtsartikel over de functie en regulatie van sterol regulatory element-binding proteins (SREBPS) in **hoofdstuk 3** gebruikt een pathway om de kennis van meer dan 50 wetenschappelijke artikelen te structureren en organiseren. De lezer kan de interactieve pathway weergave op WikiPathways, een collaboratieve pathway databank, gebruiken om in te zoomen en de pathway te verkennen om gedetailleerde informatie over de elementen te verkrijgen uit externe databronnen, en daarmee een meer uitgebreide bestudering mogelijk te maken.
Het aantal experimentele studies waarin multi-omics data gegenereerd worden neemt toe, en meer geavanceerde methodes om deze data gezamenlijk te verwerken zijn nodig. PathVisio voorziet in functionaliteit om multi-omics datasets te integreren, visualiseren en analyseren. In **hoofdstuk 4**, worden twee studies over uithongering bij muizen gecombineerd en geanalyseerd met PathVisio, de ene met proteomics metingen en de andere met transcriptomics metingen. De ingebouwde omzetting van identificatiecodes in PathVisio maakt het bijzonder geschikt voor de integratie en gelijktijdige weergave van datasets van verschillende origine. In dit voorbeeld toonden we aan dat proteomics data conclusies op grond van transcriptomics data kunnen versterken, en tegelijkertijd gebieden kunnen aanduiden waar posttranscriptionele regulatie een rol speelt. De interpretatie van dergelijke datasets is niet gemakkelijk en pathway visualisatie kan als een nuttig hulpmiddel dienen, gebruik makend van de rol van pathways als kennisbank van biologische informatie.

Netwerkbiologie

Netwerken vormen de universele structuur die ons in staat stelt om relaties en interacties tussen meerdere zaken weer te geven. Voorbeelden van interacties in de biologie zijn de regulatie van een gen door een transcriptiefactor, het bouwen van een complex van twee eiwitten of de omzetting van een metaboliet. Dergelijke interacties worden ook vaak beschreven in pathways, die vervolgens gerepresenteerd kunnen worden als kleine netwerken. Als we alle stukjes van de puzzle verzameld hebben en deze samen kunnen brengen, zijn we in staat om bijvoorbeeld alle stappen in het werkingsmechanisme van een bepaald medicijn te volgen, of uit te vinden welke voedingsstof een allergische reactie oproept en waarom.

De tweede helft van mijn thesis is voornamelijk gericht op biologische netwerken. Het veelgebruikte netwerk analyse en visualisatie programma Cytoscape biedt ontwikkelaars de mogelijkheid om extra functionaliteiten toe te voegen in de vorm van zogenaamde apps.
Pathways hebben ook veel karakteristieken van netwerken en de in **hoofdstuk 5** beschreven WikiPathways app voor Cytoscape vormt een mooie brug tussen het pathway en het netwerk deel van mijn thesis. Deze app stelt gebruikers in staat om pathways van WikiPathways en PathVisio in te laden in Cytoscape en daarmee volledig gebruik te maken van de pathway modellen middels het uitvoeren van berekeningen en op maat aangepaste visualisaties gebaseerd op experimentele

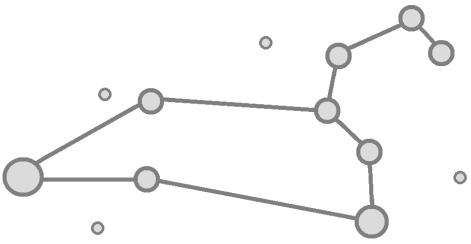
meetgegevens en netwerk topologie.
Pathways hebben meestal een bewust gekozen opmaak om ze gestructureerd en gemakkelijk te begrijpen te houden. Daardoor is het lastig om ze geautomatiseerd uit te breiden met aanvullende interacties. Netwerken daarentegen, hebben geen vastgestelde opmaak en nieuwe interacties kunnen gemakkelijk toegevoegd worden. Om die reden hebben we de CyTargetLinker app voor Cytoscape ontwikkeld om regulatoire interacties te integreren in netwerkanalyse (**hoofdstuk 6**). Er zijn veel verschillende online kennisbronnen die informatie bevatten over regulatoire elementen zoals transcriptiefactoren, microRNAs of geneesmiddelen. CyTargetLinker voorziet een makkelijke en snelle manier om deze kennis te integreren in een bestaand netwerk (dat ook een pathway kan zijn). In het bijzonder de visualisatie opties maken biologische interpretatie van gecompliceerde netwerken op een grafische manier mogelijk. De app kan bij velerlei studies gebruikt worden, zoals het identificeren van gevalideerde en voorspelde microRNAs die diabetes gerelateerde genen reguleren, het verrijken van een netwerk van DNA herstelgenen met transcriptiefactoren van ENCODE of zelfs het bouwen van een regulatoir meta-netwerk door het toevoegen van een geheel extra regulatoir niveau (transcriptiefactoren, miRNAs en geneesmiddelen) aan een biologische proces.
Het toevoegen van beide uitbreidingen aan Cytoscape laat hun gebruik toe in combinatie met een grote verscheidenheid aan andere apps voor *state-of-the-art* netwerkanalyse.

In **hoofdstuk 7** wordt duidelijk getoond hoe de tools en methodieken die in deze thesis beschreven worden, gecombineerd kunnen worden tot één bioinformatica workflow. We hebben een publiek beschikbare transcriptomics dataset geselecteerd om de obese, diabetische lever te bestuderen. De beschreven workflow gebruikt PathVisio om de pathways op te sporen die door het ziekteproces beïnvloed worden, om die vervolgens met de WikiPathways app in te laden in Cytoscape. De CyTargetLinker app stelde ons in staat om transcriptiefactoren en geneesmiddelen aan het netwerk toe te voegen om de regulatoire mechanismen te bestuderen. De totaalaanpak leidt tot verbeterde inzichten en een duidelijker begrip van het gehele ziekteproces.

Conclusie

Pathways en netwerken zijn krachtige middelen voor de opslag, organisatie, integratie, analyse en visualisatie van biologische data. Het is van belang aan te geven dat pathway analyse gekleurd is door wat al bekend en goed bestudeerd is en dat interactie data veel ruis kunnen bevatten, en dus meer experimentele validatie nodig is om de kwaliteit te verbeteren.
De beschikbaarheid van toegankelijke hoog-kwalitatieve experimentele data, goed georganiseerde en gestructureerde kennis en vrij beschikbare gebruiksvriendelijke software is uitermate belangrijk voor het biomedisch onderzoek. In deze thesis heb ik een aantal veelgebruikte bioinformatica tools voor pathway en netwerk analyse besproken en gedemonstreerd hoe deze toegepast kunnen worden in verschillende onderzoeksvelden.

Valorization



Introduction

Biological systems are highly complex and bioinformaticians have to deal with vastly increasing amount of data. In this PhD thesis on "Managing biological data in pathways and networks", I present new network-based approaches to organize, analyze and visualize biological data. The developed methods and tools help other researchers to make sense of their large and complex datasets and put them into a biological context. Because of the huge amount of diverse biological data produced everyday the integration, analysis and interpretation of data is one of the most difficult aspects in biomedical research projects nowadays. The developed approaches will help researchers to better understand and visualize their own data, to integrate it with existing knowledge and therefore to develop better strategies to improve health as well as understand and cure diseases.

In this thesis, I demonstrate the power of pathway and network analysis in different examples, studying the diabetic liver (chapter 7), combining knowledge about regulation of cholesterol biosynthesis and uptake by SREBP (chapter 3) or investigating the effect of starvation in a mouse animal study (chapter 4). One of the developed tools enables the study of drug effects on biological processes or networks (chapter 6). In the diabetic liver study (chapter 7), it is shown how this could be used to find new drug-targets, advantageous drug combinations or possible interferences with the action or efficiency of a drug caused by other drugs.

Open Source Software

Three chapters (2,5,6) in this thesis present new software tools and their applications. To ensure that others can benefit from the tools and methods developed, everything is published under an open source license. Because of the collaborative nature of open source development, effective, scalable and adaptive software can be produced more quickly. We deliberately choose an open source license with very few restrictions to maximize reusability. Although I am leading the development of all three tools, there are many other developers around the world who contribute source code or give input. The methods described in this thesis are already used by many scientists in their research and the numbers of downloads as well as publications confirm the wide adoption of the tools in the research community.

The methods used are highly generic and therefore they can be used in many different research fields. While PathVisio has a more biological focus, Cytoscape can be used for any network visualization or analysis project including social networks, transportation maps or business process modeling. Therefore the apps developed for Cytoscape can reach a much wider audience also outside the scientific community.

As part of the National Resource for Network Biology (NRNB), we also participate in large Open Source events which reach a very broad audience. For example, in the Google Summer of Code in which Google supports Open Source organizations

by providing money for students around the world to work within such an organization, we yearly get between two and five students who are paid to work on our tools during the summer. Although this money does not directly reach the department or university, it definitely results in an improvement and therefore an increased value of the tools we develop and consequently an increased visibility and reputation of the university.

Sharing Knowledge and Data

Besides the tools and methods developed, there is another key result of the research in this PhD thesis, the organization of data. Here we are not only talking about experimental data but also knowledge that has to be structure and visualized. Pathway diagrams, like the SREBP pathway described in chapter 2, are intuitive visual representations of the processes happening in our bodies and in nature around us. In some complex diseases, like Parkinson's Disease, those diagrams might be used by clinicians to explain the disease to their patients and their families. A visual illustration is much more powerful than saying that gene X is not functional. With pathway tools like PathVisio, we can go one step further towards personalized medicine and visualize the patients data on the pathway. This provides the clinician with a tool to give a unique, personalized view of the disease to the specific patient. Although high-quality pathway diagrams for many complex diseases are not yet available and further research is necessary, for some diseases this could possibly already be used in the near future.

In this thesis, we are using publicly available datasets from different online repositories. Therefore, I would like to shortly mention the value of open data and how all parties involved can benefit when data is findable, accessible, interoperable and re-usable (guide to FAIRness of data, www.datafairport.org). Repositories like ArrayExpress or GEO enable researchers to use and combine existing datasets that might lead to new insights and generate new hypotheses. The findings can also be organized and structured for example in pathways which further increases the value of data.

Spreading / Workshops

One important, probably the most important, aspect of valorization is the spreading of the new tools and methods developed. The shown statistics and number of citations already indicate that the tools are downloaded and used by many researchers around the world but we envision to further increase those numbers and make sure that researchers are using the tools in the correct, intended ways. In a first step it is important to present the tools and methods at different conferences, not only bioinformatics focused but also more biological focused ones, to inform scientists about how these tools can be used in their research.

During this 4-year research project, the tools and methods developed have been presented at numerous Bioinformatics and Systems Biology conferences in the Netherlands and abroad. Many of the approaches to analyse biological data with

the tools we developed can become very complex. This leads to a market for tutorials, courses and such. We already developed several workshop and tutorial sessions on 'How to draw a biological pathway?', 'How to use pathway and network analysis to interpret biological data' or more focused on the different tools like 'How to use PathVisio/Cytoscape to understand biological processes?'. These workshops can be further extended and improved. Then we can offer them, if possible even as online web seminars, to biomedical research groups, pharmaceutical companies and/or as educational tools to students. This could definitely result in the acquisition of money to further develop the tools and carry out new research projects.

Conclusion

In conclusion, I believe that in scientific research, collaborative approaches allow us to build on each others work and expertise and move forward faster. This is only possible with an open attitude towards sharing data and knowledge. This thesis really follows this principle by creating open source software, using open data to analyse and publishing the results in open access journals. As a final statement, I would also like to mention that having the tools and the expertise to apply them in different research fields can also lead to more traditional valorization in the future for example by participation in research directed towards health improvement or drug development.

Abbreviations

2DE	two-dimensional gel electrophoresis
API	Application programming interface
ARC	activator recruited-cofactor
ATP	Adenosine triphosphate
bHLH-Zip	basic helix-loop-helix leucine zipper
BIND	Biomolecular Interaction Network Database
BioPAX	Biological Pathway Exchange
BLAST	Basic Local Alignment Search Tool
cAMP	Cyclic adenosine monophosphate
cDNA	complementary DNA
ChEBI	Chemical Entities of Biological Interest
ChEMBL	Chemical database at European Molecular Biology Laboratory
ChRE	carbohydrate response element
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
ENCODE	Encyclopedia pf DNA Elements
ER	endoplasmic reticulum
ERAD	ER-associated degradation
FC	fold change
FDP	farnesyl diphosphate
GC-RMA	GeneChip Robust Multiarray Averaging
GEO	Gene expression omnibus
GO	Gene Ontology
GPCR	G protein-coupled receptor
GPML	Graphical Pathway Markup Language
GSEA	Gene set enrichment analysis
GTP	Guanosine-5'-triphosphate
gui	graphical user interface
HDL	high density lipoprotein
HI-II-14	Human Interactome, Space II, 2014
HMDB	Human Metabolite Database
HMG CoA	3-hydroxy-3-methylglutaryl-Coenzyme A
HPRD	Human Protein Reference Database
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
InChI	International Chemical Identifier
Insig	insulin-induced gene
KEGG	Kyoto Encyclopedia of Genes and Genomes
LDL	low-density lipoprotein
log2FC	base 2 log fold change
LXR	liver X receptor
LXRE	LXR response element
MeSH	Medical Subject Headings
MIM	Molecular Interaction Maps
MIMML	Molecular Interaction Maps Markup Language
miRNA	microRNA
mRNA	messenger RNA
MS	Mass spectrometry
MTI	microRNA-target interaction
MTTP	microsomal triglyceride transfer protein
NAD	Nicotinamide adenine dinucleotide
NAFDL	non-alcoholic fatty liver disease
NAR	Nucleic Acids Research
NCBI	National Center for Biotechnology Information
ncRNA	non coding RNA
NIH	National Institute of Health
NMR	Nuclear magnetic resonance
NR	nuclear receptor
nSREBP	nuclear SREBP
OMIM	Online Mendelian Inheritance in Man
OSGi	Open Service Gateway Initiative
PDB	Protein Data Bank
PDF	Portable Document Format
PNG	Portable Network Graphics
PRIDE	Proteomics Identifications
PUFA	polyunsaturated fatty acid
PV	PathVisio
RDF	Resource Description Framework
Recon	Reconstruction of human metabolism
RegIN	Regulatory Interaction Network
REST	Representational state transfer
RING	Really Interesting New Gene
RNA	Ribonucleic acid
RXR	retinoid X receptor
SBGN	Systems Biology Graphical Notation
SBGNML	Systems biology graphical notation markup language
SBML	Systems Biology Markup Language
SCAP	SREBP cleavage-activating protein
siRNA	small interfering RNA
SQL	Structured Query Language
SREBP	Sterol regulatory element-binding protein
SSD	sterol-sensing domain
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVG	Scalable Vector Graphics
T2DM	type 2 diabetes mellitus
TF	transcription factor
TFBS	transcription factor binding site
TIFF	Tag Image File Format
UFA	unsaturated fatty acid
WP	WikiPathways
XML	Extensible Markup Language
XGMML	extensible graph markup and modelling language
XML-RPC	XML remote procedure call

Acknowledgements

The last four years have been exciting, challenging and sometimes overwhelming and there are a number of people who really helped me along the way.

First of all, I would like to express my gratitude to my PhD supervisor. Chris, you always supported me and gave me the freedom to learn, develop and grow into the researcher I am today. Thank you for your advice, support and encouragement throughout the project. Thank you for giving me the possibility to present and discuss my work at so many places around the world. This was not only challenging and educational but also very motivating.

I would also like to give a heartfelt, special thanks to my co-supervisor. Susan, I really appreciate our collaboration and your assistance in keeping my progress on schedule. Being able to talk to you about the ups and downs in the life of a PhD student always helped and encouraged me. Thank you for being my supervisor, colleague and friend.

The completion of this PhD project could not have been accomplished without the support of all my current and former colleagues at BiGCaT. Thanks to everyone for productive collaborations and so many great discussions but also for not forgetting to laugh and having a bit of fun.
Lars, there are so many (hilarious, serious, funny, happy, sad, etc) moments that come to mind when thinking about you that I could fill a whole book with stories. Here I will simply say thank you for your support and your friendship. Anwesha, my honorary paranymph, I am very grateful for all your valuable input on my thesis project. Being able to discuss my ideas, thoughts or difficulties (work related and personal) with you was always incredible helpful. Most of all, thank you for being my friend. Martijn and Thomas, working together with you in the first year gave me a clear vision for my PhD project and your collaboration (also after you left Maastricht) really helped me to get this thesis done. I hope I have been taking good care of your creations :-).
Nuno, I appreciate your dedication and support. I dont think I would have been able to get everything done without your help! Andra, thanks for the collaboration and many discussions over coffee about work and other topics :-).

Sabine, thank you for the great work during your *Jaarwerkstuk*! Bart, Cristian, Egon, Jonathan and Magali, thanks for your help and collaboration throughout the years. Elisa, even in this short time at the end of my project, you really helped and motivated me immensely. Michiel, thanks for all your help (especially fixing ice cream machines). Jahn, thanks for always welcoming us in Berlin for a quick getaway. Jos, thank you for all your help and support during the last four years.

In addition, I have been very privileged to get to know and to collaborate with many other great people who became friends over the last several years.
I would like to offer my special thanks to our collaborators at the Gladstone Institutes in San Francisco. Alex, I learned so much from you and I want to thank you for all your help and the many invaluable discussions about everything and anything. Anders, Kristina and Samad, it has always been a pleasure to work with you and I would like to thank you all for making my visits to San Francisco such a great experience.
To all the (inter)national communities I was able to be part of in the last four years (National Resource of Network Biology, COMBINE, Google Summer of Code, Netherlands Consortium for Systems Biology, Netherlands Bioinformatics Centre), thank you all for your help, valuable feedback and educational experiences. I hope to be part of these great initiatives for many years to come. Special thanks to Sravanthi, my GSoC student and co-mentor, for her dedication and help in the PathVisio project.
Marijana, thanks for your motivation and support, and I wish you all the best with EdgeLeap!

As a tool developer, I would also like to thank all the users for their bug reports and feature requests. I always do my best to fix them as fast as possible :-).

It is not possible to get through the stressful times as a PhD student without some really good friends. Kathrin and Olaf, even though we dont see each other often enough, I always know that you are there for me when I need you. I am very grateful for your support and friendship. Verena and Falk, thank you for making my start in Maastricht such a fantastic time and for staying my friends even after I moved out after half a year. Dank aan de meisjes van de Nederlands cursus, Juliane, Kathrin en Marieke, voor jullie vriendschap en ondersteuning. Lisanne and Dennis, thank you for the game nights, spa days and 'lekkere' Tokyoto dinners. Lauren, thanks for making our lunches so entertaining and I hope I can live up to your expectations :-).

To my family and in-laws, especially my mum and dad, thank you for giving me the support and freedom to follow my dreams. It doesn't matter where in the world I am, I always know that you love me and that I can count on you.

Georg, thank you for all your love. You have been standing beside me throughout this journey, you have been standing in front of me when I needed protection and you have been standing behind me to always show that you support me and believe in me. Thank you for being my best friend, my partner and my husband!

Curriculum Vitae

Martina Maria Summer-Kutmon was born on August 1st in 1986 in Rum in Austria (which amuses many police and border control officers). She spent most of her youth jumping around in the gym as an artistic gymnast and after finishing her high school at the BORG in Linz in 2005, she started her bachelor study in *Bioinformatics* at the Upper Austrian University of Applied Sciences Hagenberg in Austria. From that moment on, she spent a lot more time in front of the computer than in the gym. Martina loves to travel and during her bachelor, she took the opportunity to do a research visit at the Institute of Agrobiotechnology at CERTH in Thessaloniki in Greece and an internship at the Centre of Integrative Bioinformatics in Vienna in Austria.

In 2008, she started her two-year master in *Bio- and Medical Informatics* in Hagenberg. Martina spent the second year of this master in Canada writing her master thesis in the Department of Cellular and Molecular Medicine at the University of Ottawa on "StemMaze: An integrative bioinformatics platform for annotating, visualizing and analyzing cell regulatory processes and networks". She graduated with honours in July 2010.

One month later in August 2010, Martina moved to the Netherlands and joined the Department of Bioinformatics (BiGCaT) at Maastricht University as a PhD student. Here she was involved in several national and international communities related to her work on pathway and network analysis.

In August 2014, Martina continued her career at BiGCaT and joined the new Maastricht Centre of Systems Biology (MaCSBio).

Publications

A network biology workflow to study transcriptomics data of the diabetic liver
BMC Genomics (2014) 15(1):971
Kutmon M, Evelo CT, Coort SL

WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization
F1000Research (2014) 3:152
Kutmon M, Lotia S, Evelo CT, Pico AR

Integrated visualization of a multi-omics study of starvation in mouse intestine
Journal of Integrative Bioinformatics (2014) 11(1):235
van Iersel MP, Sokolovic M, Lenaerts K, Kutmon M, Bouwman FG, Lamers WH, Mariman ECM, Evelo CT

CyTargetLinker: A Cytoscape app to integrate regulatory interactions in network analysis
PLoS One 8.12 (2013): e82160
Kutmon M, Kelder T, Mandaviya P, Evelo CT, Coort SL

A pathway approach to investigate the function and regulation of SREBPs
Genes & Nutrition (2013): 1-12
Daemen S, Kutmon M, Evelo CT

WikiPathways: building research communities on biological pathways
Nucleic acids research 40.D1 (2012): D1301-D1307
Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR

The BioPAX community standard for pathway data sharing
Nature Biotechnology (2010) 935-942
Demir E, Cary MP, 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 AC, 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, Ruebenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, 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 KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novere N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD.

Submitted
PathVisio 3: An extendable pathway analysis toolbox
Submitted to *PLoS Computational Biology*
Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, Evelo CT