Towards the complete picture

Towards the Complete Picture

Combining Modelling and Experimental Data in a Systems Biology Approach

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, the Agilent Thought Leader Grant awarded to Chris Evelo, and the Google Summer of Code program.

ISBN/EAN 9789462955776

Towards the Complete Picture

Combining Modelling and Experimental Data in a Systems Biology Approach

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. Rianne Letschert
in accordance with the decision of the Board of Deans,
to be defended in public on

16th of February 2017, Thursday at 16:00 hrs.

by

Anwesha Bohler

Supervisor

Prof. Dr. Chris T. A. Evelo

Co-supervisor

Dr. Martina M. Summer-Kutmon

Assessment Comittee

Prof. Dr. Wout Lamers (chairman)

Dr. Michael Lenz

Prof. Dr. Natal van Riel, AMC Amsterdam

Prof. Dr. Julio Saez-Rodriguez, RWTH Aachen

Dr. Zita Soons

Contents

CHAPTER 1

General Introduction

Metabolism

Metabolism comes from the Greek word *metabole* meaning "a change" [1]. It is the set of biochemical reactions by which energy is either created or used in the cell [2]. Biochemical reactions transform a set of chemical substances into another [3]. Reaction rates are constant at a constant temperature and chemical concentration. With an increase in temperature reaction rates typically increase due to excess availability of thermal energy to reach the necessary activation energy for breaking bonds between atoms. Reactions may proceed in the forward or reverse direction until they go to completion or reach equilibrium. In spontaneous reactions the Gibbs free energy (ΔG) associated with the reactants exceeds that associated with the products allowing these reactions to proceed without the input of free energy, e.g. cellular respiration:

Simplified reaction:

$$C6H12O6\ (s) + 6\ O2\ (g) \rightarrow 6\ CO2\ (g) + 6\ H2O\ (l) + \text{heat}$$

$$\Delta G = -2880\ \text{kJ per mol of C6H12O6}$$

The negative ΔG indicates that the reaction can occur spontaneously.

In the case of non-spontaneous reactions the input of free energy is required to proceed, e.g. photosynthesis driven by absorption of electromagnetic radiation in the form of sunlight, energy-use step in glycolysis preceding the energy generation step

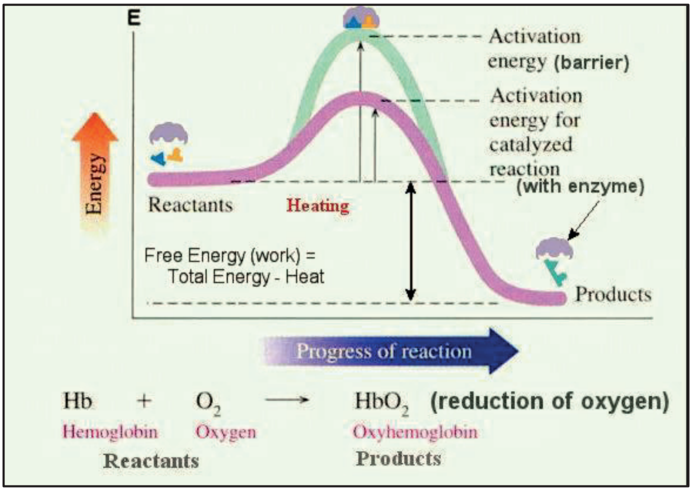

E
Energy
Activation energy (barrier)
Activation energy for catalyzed reaction (with enzyme)
Reactants
Heating
Free Energy (work) = Total Energy - Heat
Products
Progress of reaction
Hb + O2 → HbO2 (reduction of oxygen)
Hemoglobin
Oxygen
Oxyhemoglobin
Reactants
Products

Figure 1.1 Enzymatic reactions. The rate of a reaction can be affected by the temperature, the amounts of reactants in the container, and enzymatic catalysts. Figure taken from [4]

Enzymes, which are catalytic proteins, often catalyse biochemical reactions. Enzymes increase the rates of biochemical reactions so that metabolic syntheses and decompositions, impossible under ordinary conditions, can occur at the otherwise sub-optimal temperatures and concentrations present within a cell (Figure 1.1). A series of chemical reactions occurring consecutively, where the product of one reaction is the substrate of the next reaction, forms a metabolic pathway.

Pathway Diagrams

Defining pathways and understanding their physiological roles have been among the most fruitful pursuits in biological research. During the "golden age of biochemistry", from the 1920s to 1960s, most of the metabolic network that utilizes nutrients to produce energy in humans and other organisms was defined. These included the core activities such as glycolysis by Embden, Meyerhof, and Parnas, respiration by Warburg, the tricarboxylic acid and urea cycles by Krebs, glycogen catabolism by Cori and Cori, oxidative phosphorylation by Mitchell, and the supremacy of ATP in energy transfer reactions by Lipmann [2]. . Biochemistry and metabolic pathway analysis was the focus of basic and medically oriented research during these decades, with some 15 Nobel Prizes in either Physiology/Medicine or Chemistry awarded for work related to energy balance or core metabolic pathways [2]. By the end of this period, it was possible to understand which enzymes control complex matters such as the temporal and organ-specific regulation of fuel preferences [5]. Recent studies suggest that the development of diseases with complex genotype–phenotype relationships, such as cancer, neurological disorders, obesity, metabolic syndrome, and Type 2 diabetes, is primarily the consequence of an intricate interplay between genetic, environmental, and nutritional factors. It has thus become clear that this is a multifaceted problem that requires a systemic approach to explore a system of causes.

Advancements in high-throughput technologies allow system-wide measurements of transcripts, proteins, and metabolites. The current knowledge about a biological process, summarised in a diagram, is commonly used as an intuitive framework to integrate and co-analyse system-wide transcriptomics, proteomics and metabolomics measurements. Regulatory mechanisms are present at every stage of the process, namely transcription, post-transcriptional modifications, translation, post-translational modifications. For instance, organisms often silence genes by employing methylation as a regulatory mechanism.

Glycolysis is the metabolic pathway that converts glucose into pyruvate (Figure 1.2). The pathway, as it is known today, took almost 100 years to discover fully. It is the initial step of cellular respiration and proceeds in both aerobic and anaerobic conditions. In glycolysis, a six-carbon glucose molecule is ultimately split into two three-carbon molecules called pyruvate. These carbon molecules are utilized to produce NADH and ATP. For the glucose molecule to be oxidized to pyruvate, an input of ATP molecules is required. Therefore, although the total yield of ATP is four molecules, but the net gain is two ATP molecules.

Metabolism is regulated by signal transduction and it is becoming increasingly clear that cellular signalling and metabolism are tightly linked [6, 7], e.g. insulin binding to its receptor on fat and muscle cells to stimulate the uptake of glucose, thereby establishing metabolic stores of triglycerides and glycogen [8, 9].

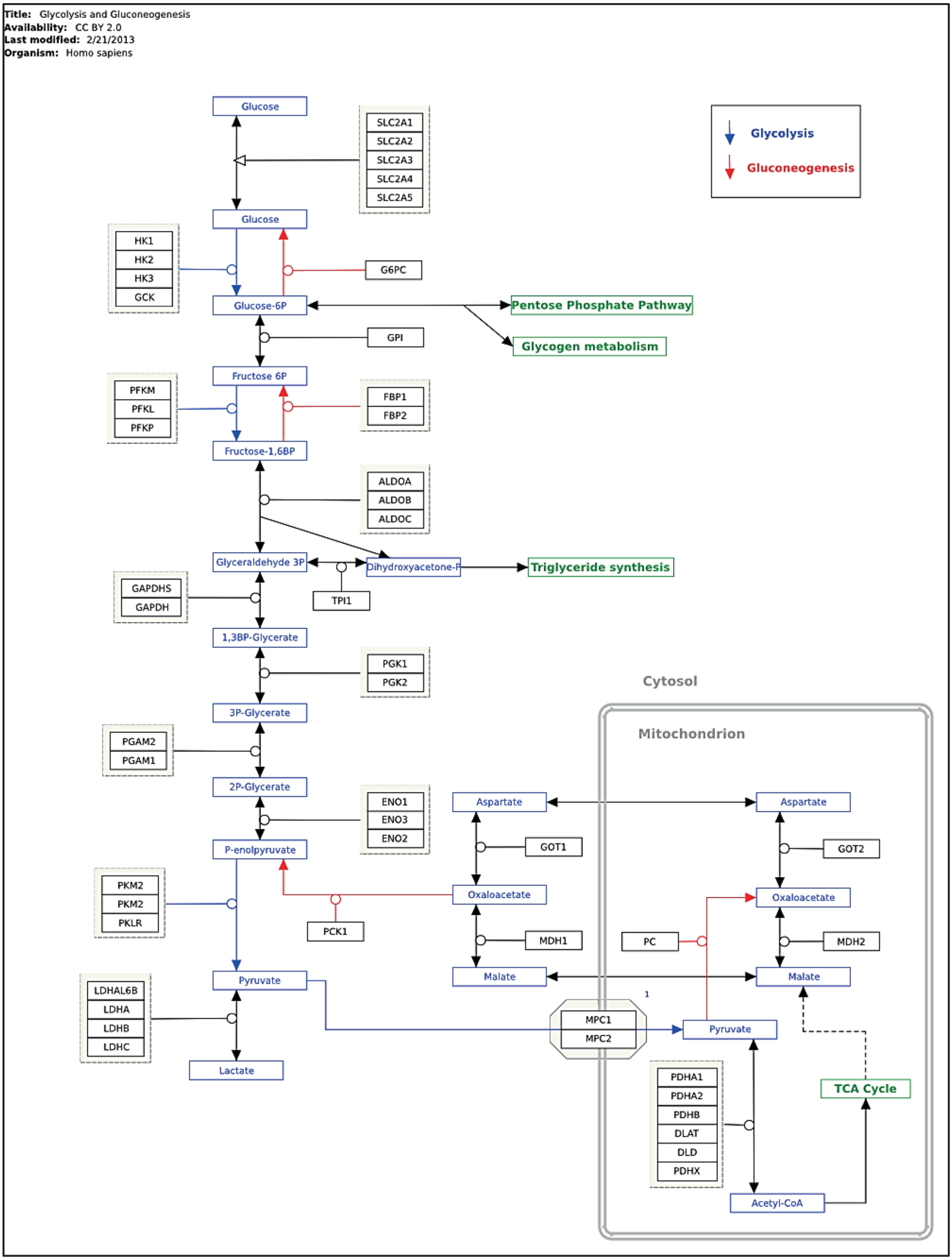


Figure 1.2 The WikiPathways pathway diagram for Glycolysis and gluconeogenesis pathway (http://wikipathways.org/instance/WP534)

Computational Data Analysis and Visualization

The mathematician John Tukey argues in one of the most classic books on data visualization and analysis, Exploratory Data Analysis, that the

> *"greatest value of a picture is when it forces us to notice what we never expected to see"* [10].

Computational tools are imperative in trying to create such a picture from large scale heterogeneous biological data. Several methods have been developed that allow systematic analysis of large datasets to generate biologically meaningful interpretations, some of which will be discussed in the following paragraphs.

Pathway Analysis

Pathway analysis has become the first choice for detecting which signalling or metabolic pathways are activated under certain experimental conditions or disease states. Pathway analysis is preferred as it reduces the complexity to just several hundred pathways by grouping thousands of genes, proteins, and other biological molecules measured and identifies active pathways that differ between two conditions instead of simply a list of different genes or proteins [11]. Pathway analysis lends context to experimental measurements and can help in understanding the molecular mechanisms of a disease state, identifying genes and proteins associated with the aetiology of a specific disease, and predicting drug targets.

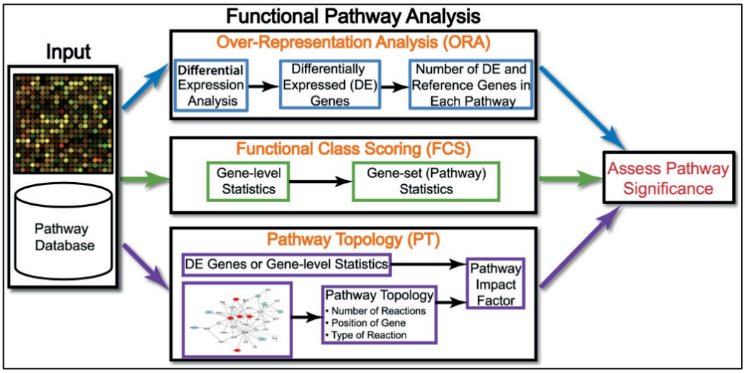

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Impact Factor
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Assess Pathway Significance

Figure 1.3 Overview of existing pathway analysis methods using gene expression data as an example by Khatri et al. [12].

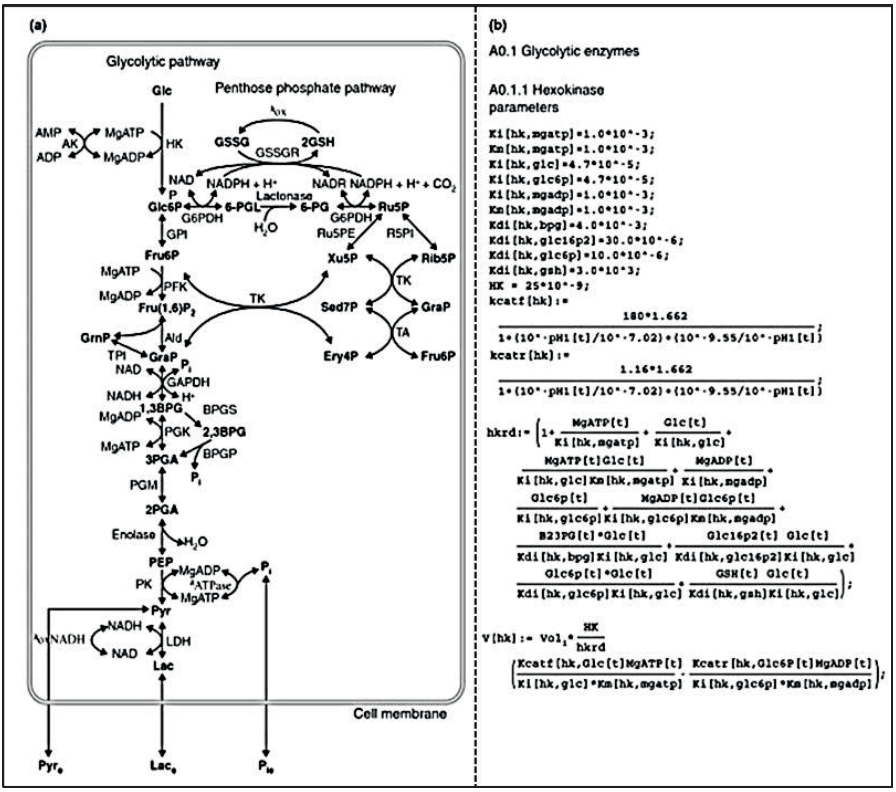
Several approaches to exploit pathway knowledge in analysing high-throughput measurements of genes and proteins exist (Figure 1.3). However, the power of a pathway-level statistic can depend on a number of factors, namely the proportion of genes in a pathway that are differentially expressed, the pathway size, and the amount of correlation between genes in the pathway. Interestingly, although multivariate statistics are generally preferred as they are expected to have a higher statistical power. When applied to real biological data, univariate statistics show more statistical power at stringent cut-offs ($p \leq 0.001$), and both show equal statistical power at less stringent cut-offs ($p \leq 0.05$) [13].

However, available pathway analysis techniques do not take into account cell specific pathway rewiring. True pathway topology depends on several factors, such as cell type, disease state. But, such condition-specific pathways are rarely available, and the knowledge is fragmented over multiple pathway databases [14]. Other limitations of pathway analysis methods are the inability to predict dynamic states of a system and the failure to consider interactions between pathways.

Kinetic Modelling

The dynamic states of a biological process can be modelled using differential equations that define the rate of change of metabolites, *i.e.* metabolic flux. Metabolic fluxes are a function of gene expression, translation, protein post-translational modifications and protein-metabolite interactions [15]. Regulation of flux within cells is vital for all metabolic pathways to adapt the pathway's activity under different conditions [16]. Numerous mathematical models based on differential equations for the simulation of metabolic pathways have been developed, as an example, the human glycolytic pathway and the rate equation devised for the first step of the pathway are shown in Figure 1.4.

Kinetic models aim to verify existing hypotheses and make new quantitative predictions. If the mathematical model of a biological system based on current knowledge is capable of qualitatively reproducing time-resolved data, then this is a good indication that the assumptions made about the molecular mechanisms underlying the biological process are correct. The system's behaviour when subjected to various perturbations, such as gene knock-out or overexpression, application of inhibitors or other drugs can then be predicted using this model. Agreement with experiments further consolidates the existing knowledge and falsification leading to the development of new hypotheses which again feedback to improve the model description. An illustrative example of how this mutual stepwise improvement of models and experiments leads to new discoveries is presented by Locke *et al.* [17], who utilize this approach to discover new genes involved in the plant circadian clock.



A0.1 Glycolytic enzymes

A0.1.1 Hexokinase
parameters

```
Ki[hk,mgatp]=1.0*10^-3;
Km[hk,mgatp]=1.0*10^-3;
Ki[hk,glc]=4.7*10^-5;
Ki[hk,glc6p]=4.7*10^-5;
Ki[hk,mgadp]=1.0*10^-3;
Km[hk,mgadp]=1.0*10^-3;
Kdi[hk,bpg]=4.0*10^-3;
Kdi[hk,glc16p2]=30.0*10^-6;
Kdi[hk,glc6p]=10.0*10^-6;
Kdi[hk,gsh]=3.0*10^3;
HK = 25*10^-9;
```

$$\mathrm{kcatf[hk]:=}\ \frac{180*1.662}{1+(10^{\wedge}\text{-}\mathrm{pHi[t]}/10^{\wedge}\text{-}7.02)+(10^{\wedge}\text{-}9.55/10^{\wedge}\text{-}\mathrm{pHi[t]})};$$

$$\mathrm{kcatr[hk]:=}\ \frac{1.16*1.662}{1+(10^{\wedge}\text{-}\mathrm{pHi[t]}/10^{\wedge}\text{-}7.02)+(10^{\wedge}\text{-}9.55/10^{\wedge}\text{-}\mathrm{pHi[t]})};$$

$$\mathrm{hkrd:=}\left(1+\frac{\mathrm{MgATP[t]}}{\mathrm{Ki[hk,mgatp]}}+\frac{\mathrm{Glc[t]}}{\mathrm{Ki[hk,glc]}}+\frac{\mathrm{MgATP[t]Glc[t]}}{\mathrm{Ki[hk,glc]Km[hk,mgatp]}}+\frac{\mathrm{MgADP[t]}}{\mathrm{Ki[hk,mgadp]}}+\frac{\mathrm{Glc6p[t]}}{\mathrm{Ki[hk,glc6p]}}+\frac{\mathrm{MgADP[t]Glc6p[t]}}{\mathrm{Ki[hk,glc6p]Km[hk,mgadp]}}+\frac{\mathrm{B23PG[t]*Glc[t]}}{\mathrm{Kdi[hk,bpg]Ki[hk,glc]}}+\frac{\mathrm{Glc16p2[t]\ Glc[t]}}{\mathrm{Kdi[hk,glc16p2]Ki[hk,glc]}}+\frac{\mathrm{Glc6p[t]*Glc[t]}}{\mathrm{Kdi[hk,glc6p]Ki[hk,glc]}}+\frac{\mathrm{GSH[t]\ Glc[t]}}{\mathrm{Kdi[hk,gsh]Ki[hk,glc]}}\right);$$

$$\mathrm{v[hk]:=\ Vol_1*}\frac{\mathrm{HK}}{\mathrm{hkrd}}\left(\frac{\mathrm{Kcatf[hk,Glc[t]MgATP[t]}}{\mathrm{Ki[hk,glc]*Km[hk,mgatp]}}-\frac{\mathrm{Kcatr[hk,Glc6P[t]MgADP[t]}}{\mathrm{Ki[hk,glc6p]*Km[hk,mgadp]}}\right);$$

Figure 1.4 Dynamic model of the human glycolytic pathway shown in Figure 1.2. (a) the reaction model for the glycolytic pathway, and (b) the first rate equation used in the model of the glycolytic pathway for hexokinase. The figure is taken from Kuchel [18].

However, a simple up-scaling of kinetic modelling approach to genome-scale is difficult due to the lack of knowledge of most enzymatic parameters and uncertainty in enzyme concentrations and post-translational modifications. Since high-throughput data is noisy, it is difficult to ensure that the model is fitted to the actual data and not to the noise. Also, thorough analysis of such a model is hardly possible considering the multitude of degrees of freedom.

Constraint-based modelling

Genome-scale metabolic network reconstruction has become an indispensable tool for studying the systems biology of metabolism [19-25]. They are reconstructed using the genome annotation of the target organism and biochemical databases [26-28].

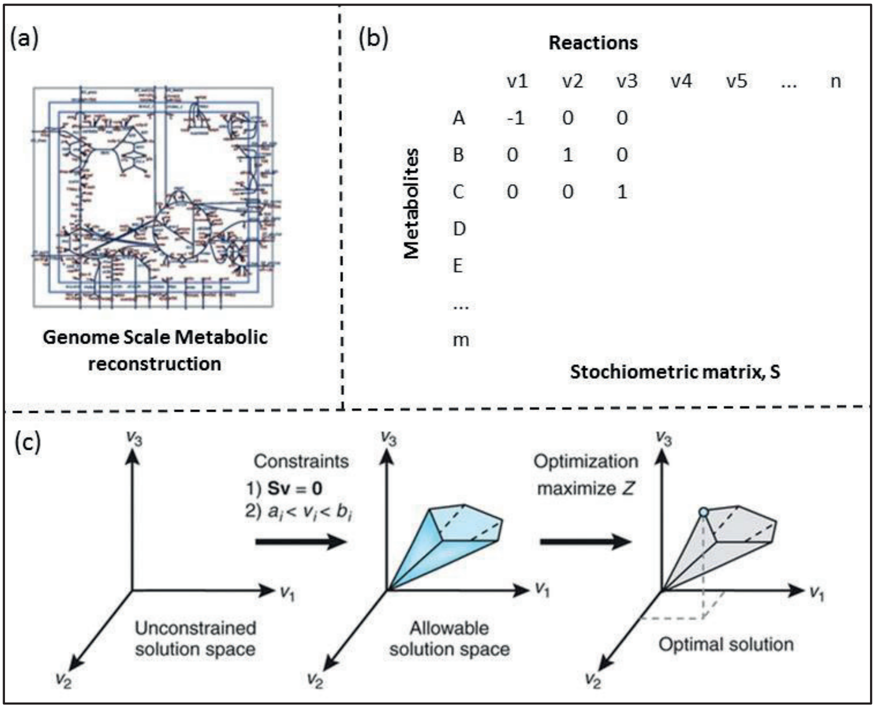


Figure 1.5 Overview of Flux Balance Analysis. (a) A genome-scale metabolic reconstruction can be curated from literature and online data sources, (b) Mathematical description of the stoichiometry of this reaction network in a stoichiometric matrix, where each column represents the stoichiometry of a reaction. Negative values indicate the molecule is consumed, a reactant and positive values indicate that the molecule is produced, a product. (c) Reaction constraints (e.g., mass balance, the assumption of steady-state, and measured rates of metabolite consumption) reduce the space of feasible flux distributions. An objective function, e.g. maximizing biomass is defined, and the optimal flux distributions are obtained through linear programming. Figure adapted from **[29]**.

Reconstructed genome-scale metabolic networks are converted to a stoichiometric matrix in which each row represents a metabolite, and each column represents a reaction If a metabolite is consumed in a reaction it is assigned the value -1, if it is produced then 1, otherwise, 0 is assigned. Flux balance analysis is widely used for studying genome-scale metabolic network reconstructions [25, 30-32] (Figure 1.5). A multi-dimensional "solution space" for the flux of a reaction is defined by imposing known upper and lower bounds on each reaction. The actual flux carried through the reaction falls within this solution space. Additional constraints such as enzyme capacity, spatial localization, metabolite sequestration, and multiple levels of regulation at the gene, transcript, and protein levels can further shrink the solution space to focus in on the actual flux state of the network [33].

An objective function is then defined, *e.g.* increase in biomass, and optimized through linear programming to identify a single optimal flux distribution [29].

The first genome-scale metabolic model was generated in 1995 for *Haemophilus influenza* [26]. In 1998 the first multicellular organism, *C. elegans*, was reconstructed [34]. Since then, many reconstructions have been created, converted into models, and experimentally validated. A list of these are available at http://sbrg.ucsd.edu/InSilicoOrganisms/OtherOrganisms. Another milestone in systems biology was reached in 2007 when a genome-scale metabolic model for humans, Recon 1, was presented [30]. This model could be utilised to direct hypotheses of novel metabolic functions in human metabolism. Underlying the effectiveness of a systems approach for the discovery of novel biochemical pathways [35].

Multi-omics Data Visualization

Obtaining up-to-date graphical representations of metabolic models simplifies model evaluation and correction. The graphical representations of the mathematical models can be easily compared with diagrams from other pathway databases to improve the models.

In addition, metabolic fluxes modelled by mathematical models can then be visualized alongside transcriptomics, proteomics, and metabolomics data on pathway diagrams or graphical representations of the mathematical models themselves (Figure 1.6). Visualization of multi-omics data helps in confirming whether the other levels of experimental measurements match up to what would be expected from the model.

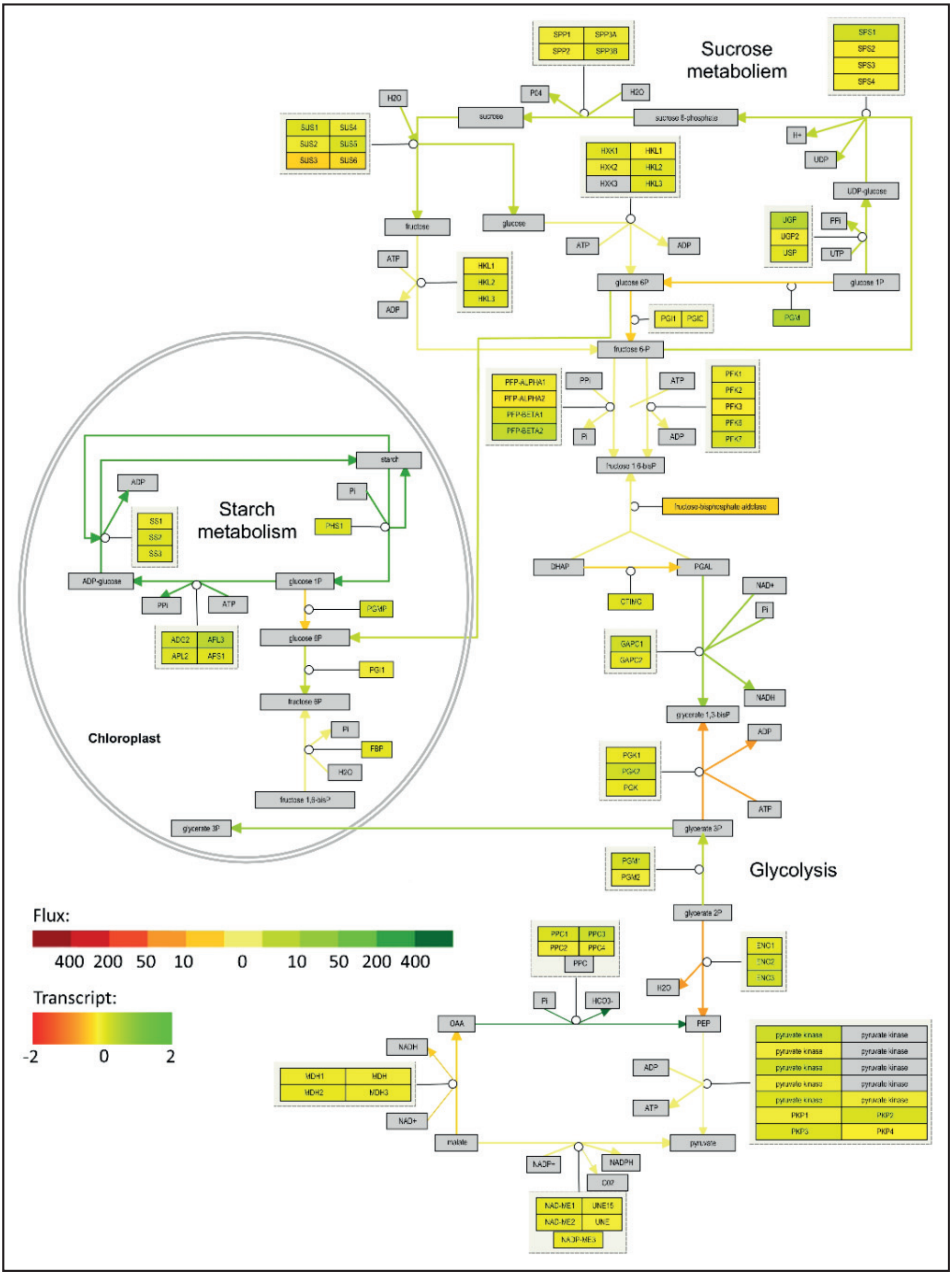


Figure 1.6 Sugar metabolism in Photosynthesis. Combined visualisation of transcriptomics [36] and flux data from the AraGEM model on the sucrose metabolism (WP2623), starch metabolism (WP2622), and glycolysis (WP2621) pathways from WikiPathways. Log fold changes for the transcripts are visualised on the nodes using a colour gradient green, yellow, and red corresponding to the values 2, 0, and -2. Metabolic fluxes are visualised on the interactions. Positive fluxes are represented in shades of green, light to dark. Similarly, the negative values are represented in shades of red. The higher the flux, the darker the colour, and vice versa. Negative values indicate that the reaction proceeds in the direction opposite to what is denoted in the diagram. Zero fluxes are represented in yellow.

Networks as tools for biological discovery

Signal transduction pathways are known to regulate metabolism. However, our knowledge of all biological processes is not complete. In the post-genomic era, data-driven approaches can be used to infer causal signalling networks from high- throughput gene/protein data. In the case of signal transduction proteomics data is used [37]. In such network constructions, proteins are the nodes and directed edges represent interactions in which the child undergoes a biochemical conformation by the action on the parent (e.g. mediated by phosphorylation, ubiquitination, methylation) (Figure 1.7).

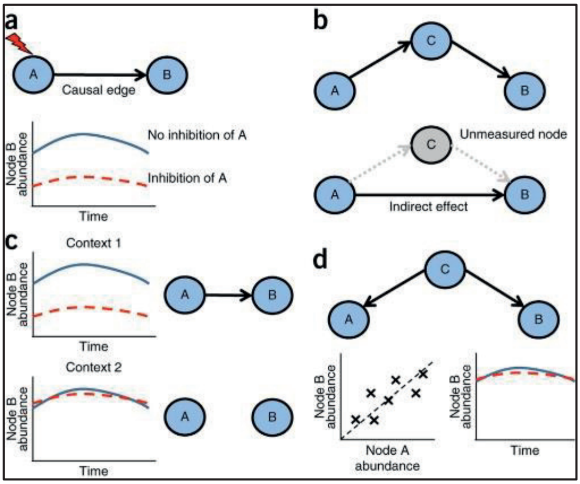


Figure 1.7 Inferring causal networks. (a) A directed edge denotes that the inhibition of the parent node A can change the abundance of the child node B. (b)If node A causally influences the node B via the measured node C, then the causal network should contain edges connecting A to C and C to B, but A to B. However, if node C is not measured, then the causal network should contain an edge connecting A to B. (c) Causal edges may depend on the biological context; for example, a causal edge from A to B appears in context 1, but not in context 2 (lines in graphs are as defined in a). (d) Nodes A and B are correlated due to being regulated by the same node C, but in this example no sequence of mechanistic events links A to B, and thus inhibition of A does not change the abundance of B (lines in the bottom right graph are as defined in a). Figure is taken from [38].

Context-specific metabolic networks can be derived from genome-scale metabolic models for different disease states or cell types using high-throughput genomic data. These networks can be compared using network analysis methods to study the network rewiring [39]. These networks can also be extended with interactions between transcription factors (TF) and their target proteins, as well as post-translational regulation by other factors such as microRNAs. MicroRNAs and transcription factors are known to be major metabolic regulators, e.g., the presence of a binding element upstream of gene A for a TF, which gene B codes for may induce a regulation of gene A by gene B.

Gene-disease associations could be studied to detect which diseases might have pathophysiological connections to each other. Drug target analysis could be done to detect drugs which might be candidates for repurposing or might target genes upstream and downstream to detect which drugs might interfere with or enhance each other’s effects (Figure 1.8).

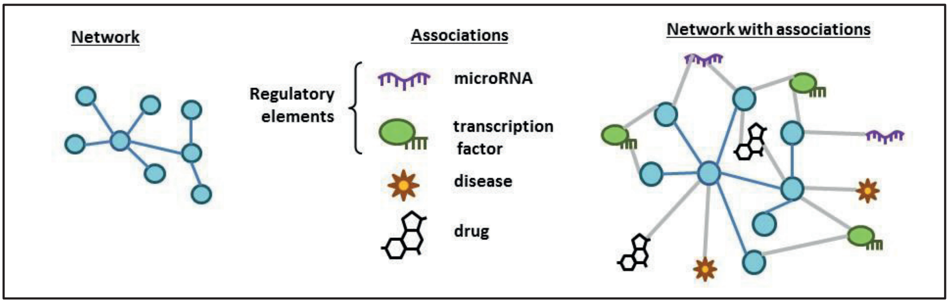


Figure 1.8 A metabolic network extended with regulatory information, drug-target interactions, and diseases associations.

Outline of the thesis

This thesis aims to integrate, visualise, interpret, and explore large-scale genomic data using pathway, network, and modelling based approaches.

The research community has widely adopted the PathVisio biological pathway editor, analysis, and visualization software for the analysis and visualization of transcriptomics, proteomics, and metabolomics data. As part of this thesis, a toolset was developed to enable the visualization of fluxes modelled by various mathematical approaches or measured by C^{13} based experiments to be visualized on the interactions of pathway diagrams enabling a holistic view of the biological process.

To link data to interactions, it is important that the interactions in pathways be annotated correctly, using identifiers from online databases that are standardized and globally recognized. **Chapter 2** describes the third release of the PathVisio software in which annotation of interactions was enabled. Furthermore, the software was redesigned to simplify the extension of the core software using add-ons enabling additional features for pathway creation, data visualization, and pathway analysis.

Primary data analysis, such as quality control, normalisation, and statistical analysis, is often performed in scripting languages like R, Perl, and Python. Subsequently, pathway analysis is carried out using dedicated external applications. Workflows involving the manual use of multiple environments are time-consuming and error-prone. Therefore, tools that enable pathway analysis directly, within the same scripting languages used for primary data analyses, are needed. **Chapter 3** describes PathVisioRPC, an XMLRPC interface for PathVisio, which enables pathway analysis in the same environment used for primary data analyses. Since R is the gold standard statistical environment, an R package for PathVisioRPC was developed. In addition, a pathway module is introduced for the microarray data analysis portal ArrayAnalysis.org that was developed to exemplify how the PathVisioRPC interface can be used by data analysis pipelines for functional analysis of processed genomics data.

There are two main approaches to creating biological pathways: knowledge driven or data driven. As the names suggest, knowledge-driven pathway diagrams are drawn based on the mechanistic understanding of the biological process, while data-driven pathways are inferred from high-throughput datasets. **Chapter 4** links two large open source knowledge-based pathway databases, WikiPathways and Reactome. WikiPathways benefits from the dedicated focus and attention provided to the content converted from Reactome. Reactome, in turn, benefits from the continuous community curation available on WikiPathways. The research community at large benefits from the availability of a larger set of pathways for analysis in PathVisio and Cytoscape.

Chapter 5 describes a toolset for integrating and visualizing modelled and measured flux data alongside transcript, protein, and metabolite data on pathways. The toolset comprises an identifier mapping database for interactions and three PathVisio plugins. The identifier mapping database enables data mapping on interactions;

the FindYourInteraction plugin simplifies the annotation of interactions on pathways; the IntViz plugin enables visualization of data on interactions, and the PathSBML plugin enables the import of mathematical models of biological processes as graphical diagrams providing more knowledge frameworks for data integration. To illustrate the functionality of the toolset we combine fluxes modelled by a publicly available genome-scale metabolic model and a published transcriptomics dataset about photosynthesis in *Arabidopsis thaliana* on the glycolysis, starch metabolism, and sucrose metabolism pathways. As an example for multi-omics visualization, we visualize the modelled fluxes on the interactions and the transcriptomics data on the genes, in like manner, data from all omics levels and other biological data, numeric and non-numeric, can be visualized on pathways.

A dysregulation of metabolism leads to metabolic diseases. Metabolic diseases are found to be most common in obese individuals with a metabolically unfavourable profile are prone to developing associated comorbidities. In **Chapter 6**, we combine publicly available modelled and measured genomic data on pathways to explore what happens during weight gain in adipose tissue of obese individuals with an unfavourable metabolic profile using pathway and network approaches. Furthermore, we integrated TF-gene and miRNA-target interactions and identified a significantly downregulated network of fourteen pathways in metabolically unhealthy obese individuals. In addition, we investigated disease-gene associations for the significantly changed network. Finally, we investigated drugs that target genes in the significantly changed network and identify drugs that can be repurposed and new drug targets.

In conclusion, we discuss big data in healthcare, how we can organise and leverage current knowledge to analyse big data, and tools and analyses methods. Metabolism is tightly regulated by signal transduction. Our knowledge about metabolic and signal transduction pathways summarized in various pathway databases is not complete. Data-driven network inference methods can help us complete the picture by de-novo biological pathway generation. As signal transduction occurs at the level of proteins, signal transduction pathways are inferred from proteomics data is also illustrated using time-course proteomics data from the HPN- DREAM Breast cancer network inference challenge. The inferred biological networks can help complete our knowledge about the underlying biology of metabolism. Furthermore, we discuss the importance of automating analyses, data integration and open science in the **General Discussion.**

References

1. Harper, D., **Online etymology dictionary**. 2001.
2. DeBerardinis, R.J. and C.B. Thompson, **Cellular metabolism and disease: what do metabolic outliers teach us?** *Cell*, 2012. **148**(6): p. 1132-1144.
3. McNaught, A.D. and A.D. McNaught, **Compendium of chemical terminology**. Vol. 1669. 1997: Blackwell Science Oxford.
4. **Molecules**. 2016; Available from: https://universe-review.ca/F12-molecule13.htm.
5. Krebs, H., **Some aspects of the regulation of fuel supply in omnivorous animals**. *Advances in enzyme regulation*, 1972. **10**: p. 397-420.
6. Wellen, K.E. and C.B. Thompson, **A two-way street: reciprocal regulation of metabolism and signalling**. *Nature reviews Molecular cell biology*, 2012. **13**(4): p. 270-276.
7. Shimobayashi, M. and M.N. Hall, **Making new contacts: the mTOR network in metabolism and signalling crosstalk**. *Nature reviews Molecular cell biology*, 2014. **15**(3): p. 155-162.
8. Chang, L., S.-H. Chiang, and A. Saltlel, **Insulin signaling and the regulation of glucose transport**. *Molecular Medicine*, 2004. **10**(7/12): p. 65.
9. Herman, M.A. and B.B. Kahn, **Glucose transport and sensing in the maintenance of glucose homeostasis and metabolic harmony**. *Journal of Clinical Investigation*, 2006. **116**(7): p. 1767.
10. Tukey, J.W., **Exploratory data analysis**. 1977.
11. Glazko, G.V. and F. Emmert-Streib, **Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets**. *Bioinformatics*, 2009. **25**(18): p. 2348-2354.
12. Khatri, P., M. Sirota, and A.J. Butte, **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS Comput Biol*, 2012. **8**(2): p. e1002375.
13. Efron, B. and R. Tibshirani, **On testing the significance of sets of genes**. *The annals of applied statistics*, 2007: p. 107-129.
14. Bauer ‐ Mehren, A., L.I. Furlong, and F. Sanz, **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Molecular systems biology*, 2009. **5**(1).
15. Nielsen, J., **It is all about metabolic fluxes**. *Journal of Bacteriology*, 2003. **185**(24): p. 7031-7035.
16. Voet, D. and J.G. Voet, **Biochemistry, 4-th Edition**. *NewYork: John Wiley& SonsInc*, 2011: p. 492-496.
17. Locke, J.C., L. Kozma ‐ Bognár, P.D. Gould, B. Fehér, E. Kevei, F. Nagy, M.S. Turner, A. Hall, and A.J. Millar, **Experimental validation of a predicted feedback loop in the multi ‐ oscillator clock of Arabidopsis thaliana**. *Molecular systems biology*, 2006. **2**(1).
18. Kuchel, P.W., **Models of the human metabolic network: aiming to reconcile metabolomics and genomics**. *Genome medicine*, 2010. **2**(7): p. 46.
19. Almaas, E., B. Kovacs, T. Vicsek, Z. Oltvai, and A.-L. Barabási, **Global organization of metabolic fluxes in the bacterium Escherichia coli**. *Nature*, 2004. **427**(6977): p. 839-843.
20. Thiele, I., N.D. Price, T.D. Vo, and B.Ø. Palsson, **Candidate metabolic network states in human mitochondria impact of diabetes, ischemia, and diet**. *Journal of Biological Chemistry*, 2005. **280**(12): p. 11683-11695.
21. Pál, C., B. Papp, M.J. Lercher, P. Csermely, S.G. Oliver, and L.D. Hurst, **Chance and necessity in the evolution of minimal metabolic networks**. *Nature*, 2006. **440**(7084): p. 667-670.
22. Barrett, C.L., C.D. Herring, J.L. Reed, and B.O. Palsson, **The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states**. *Proceedings of the National Academy of Sciences of the United States of America*, 2005. **102**(52): p. 19103-19108.
23. Covert, M.W., E.M. Knight, J.L. Reed, M.J. Herrgard, and B.O. Palsson, **Integrating high-throughput and computational data elucidates bacterial networks**. *Nature*, 2004. **429**(6987): p. 92-96.
24. Segre, D., D. Vitkup, and G.M. Church, **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences*, 2002. **99**(23): p. 15112-15117.
25. Feist, A.M. and B.Ø. Palsson, **The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli**. *Nature biotechnology*, 2008. **26**(6): p. 659-667.
26. Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, and J.M. Merrick, **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science*, 1995. **269**(5223): p. 496-512.
27. Overbeek, R., D. Bartels, V. Vonstein, and F. Meyer, **Annotation of bacterial and archaeal genomes: improving accuracy and consistency**. *Chemical reviews*, 2007. **107**(8): p. 3431-3447.
28. Manichaikul, A., L. Ghamsari, E.F. Hom, C. Lin, R.R. Murray, R.L. Chang, S. Balaji, T. Hao, Y. Shen, and A.K. Chavali, **Metabolic network analysis integrated with transcript verification for sequenced genomes**. *Nature methods*, 2009. **6**(8): p. 589.
29. Orth, J.D., I. Thiele, and B.Ø. Palsson, **What is flux balance analysis?** *Nature biotechnology*, 2010. **28**(3): p. 245-248.
30. Duarte, N.C., S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, and B.Ø. Palsson, **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proceedings of the National Academy of Sciences*, 2007. **104**(6): p. 1777-1782.
31. Feist, A.M., M.J. Herrgård, I. Thiele, J.L. Reed, and B.Ø. Palsson, **Reconstruction of biochemical networks in microorganisms**. *Nature Reviews Microbiology*, 2009. **7**(2): p. 129-143.
32. Oberhardt, M.A., B.Ø. Palsson, and J.A. Papin, **Applications of genome ‐ scale metabolic reconstructions**. *Molecular systems biology*, 2009. **5**(1): p. 320.
33. Lewis, N.E., H. Nagarajan, and B.O. Palsson, **Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods**. *Nature Reviews Microbiology*, 2012. **10**(4): p. 291-305.
34. The C. elegans Sequencing Consortium, **Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology**. *Science*, 1998. **282**: p. 2012-2018.
35. Rolfsson, O., B.Ø. Palsson, and I. Thiele, **The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions**. *BMC systems biology*, 2011. **5**(1): p. 155.

36. Fujimoto, R., J.M. Taylor, S. Shirasawa, W.J. Peacock, and E.S. Dennis, **Heterosis of Arabidopsis hybrids between C24 and Col is associated with increased photosynthesis capacity**. *Proc Natl Acad Sci U S A*, 2012. **109**(18): p. 7109-14.

37. Terfve, C. and J. Saez-Rodriguez, **Modeling signaling networks using high-throughput phospho-proteomics**, in *Advances in Systems Biology*. 2012, Springer. p. 19-57.

38. Hill, S.M., L.M. Heiser, T. Cokelaer, M. Unger, N.K. Nesser, D.E. Carlin, Y. Zhang, A. Sokolov, E.O. Paull, and C.K. Wong, **Inferring causal molecular networks: empirical assessment through a community-based effort**. *Nature methods*, 2016.

39. Kutmon, M., C.T. Evelo, and S.L. Coort, **A network biology workflow to study transcriptomics data of the diabetic liver**. *BMC genomics*, 2014. **15**(1): p. 971.

CHAPTER 2

PathVisio 3: An Extendable Pathway Analysis Toolbox

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

1. Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands
2. Maastricht Center for Systems Biology (MaCSBio), Maastricht University, The Netherlands
3. General Bioinformatics, Reading, UK
4. EdgeLeap B.V., Utrecht, The Netherlands
5. Gladstone Institutes, San Francisco, USA

Published in **PLoS Comput Biol** 11.2 (2015): e1004085.

Abstract

PathVisio is a commonly used pathway editor, visualization and analysis software. Biological pathways have been used by biologists for many years to describe the detailed steps in biological processes. Those powerful, visual representations help researchers to better understand, share and discuss knowledge. Since the first publication of PathVisio in 2008, the original paper was cited more than 170 times and PathVisio was used in many different biological studies. As an online editor PathVisio is also integrated in the community curated pathway database WikiPathways.

Here we present the third version of PathVisio with the newest additions and improvements of the application. The core features of PathVisio are pathway drawing, advanced data visualization and pathway statistics. Additionally, PathVisio 3 introduces a new powerful extension systems that allows other developers to contribute additional functionality in form of plugins without changing the core application.

PathVisio can be downloaded from http://www.pathvisio.org and from January to September 2014 PathVisio 3 has been downloaded over 3,000 times. There are already more than 15 plugins available in the central plugin repository. PathVisio is a freely available, open-source tool published under the Apache 2.0 license (http://www.apache.org/licenses/LICENSE-2.0). It is implemented in Java and thus runs on all major operating systems. The code repository is available at http://svn.bigcat.unimaas.nl/pathvisio. The support mailing list for users is available on https://groups.google.com/forum/#!forum/wikipathways-discuss and for developers on https://groups.google.com/forum/#!forum/wikipathways-devel.

Author Summary

Pathway 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. We introduce the third version of our freely available pathway editor, visualization and analysis software, PathVisio. Our tool enables researchers to draw and annotate new pathways, so they can be used for data visualization or analysis. The integrated visualization and analysis of experimental data (transcriptomics, metabolomics, and proteomics) in well described pathway diagrams helps researchers to get a more comprehensive understanding of the experimental data.

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 and [11] 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.

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].

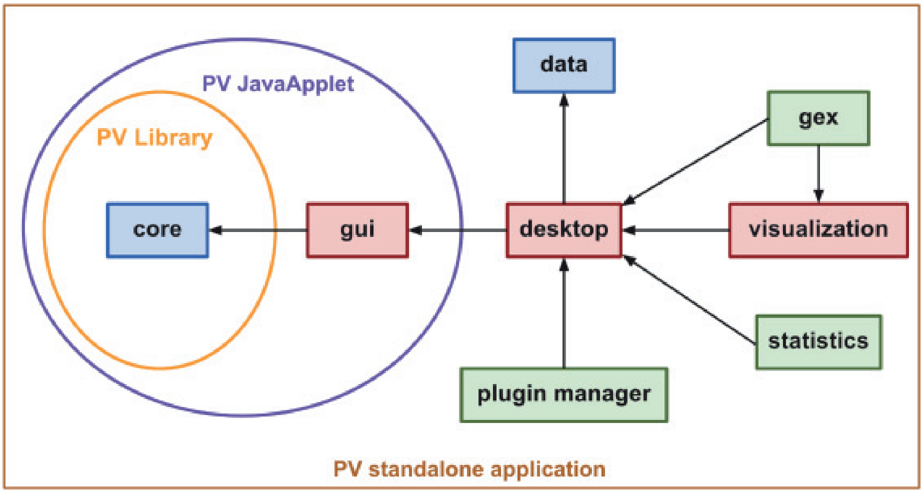


Figure 2.1 Transitive Dependency Structure of PathVisio 3. The application consists of eight modules each providing specific functionality. The modules core and data are independent modules (colored in blue) that function as libraries that can be reused outside of PathVisio (PV). Especially the core 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.

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.

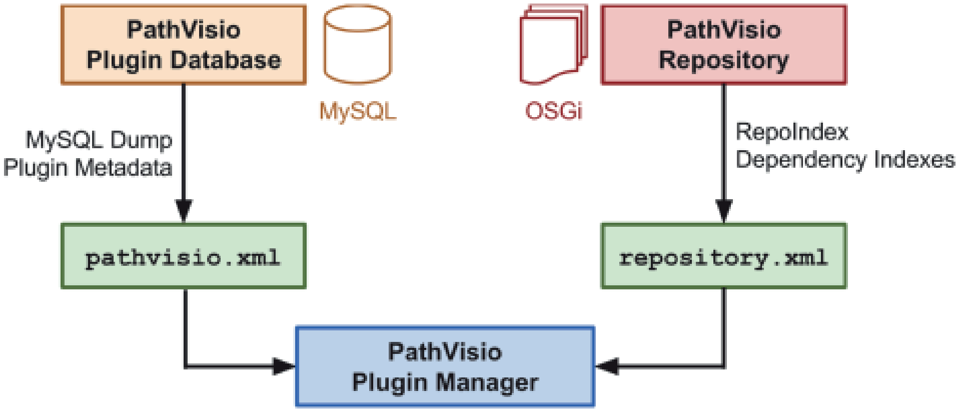


Figure 2.2 Plugin Extension and Installation System of PathVisio 3. The plugin repository stores all plugin files and their dependencies. The RepoIndex library is used to create a repository.xml file which contains the dependency indexes of all plugins. Metadata about plugins is stored in the PathVisio 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.

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 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. A feature comparison table with other tools is available as supplementary data.

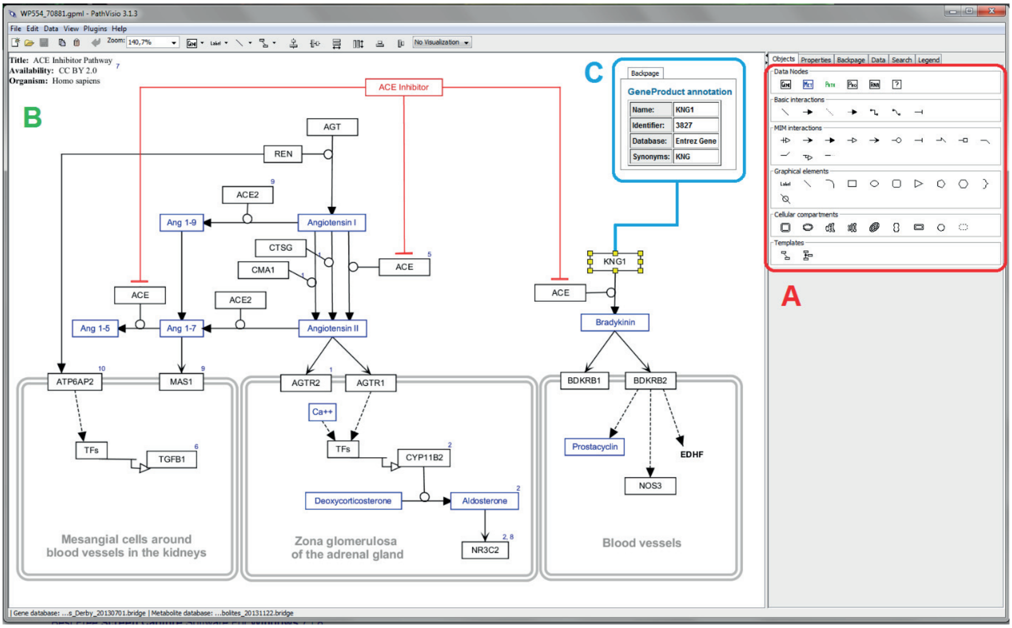
Table 2.1 PathVisio 3 Feature Table

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, 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.



WP554_70881.gpml - PathVisio 3.1.3

Zoom: 140,7%

No Visualization

Title: ACE Inhibitor Pathway
Availability: CC BY 2.0
Organism: Homo sapiens

GeneProduct annotation

Name:	KNG1
Identifier:	3827
Database:	Entrez Gene
Synonyms:	KNG

| Gene database: ...s_Derby_20130701.bridge | Metabolite database: ...bolites_20131122.bridge

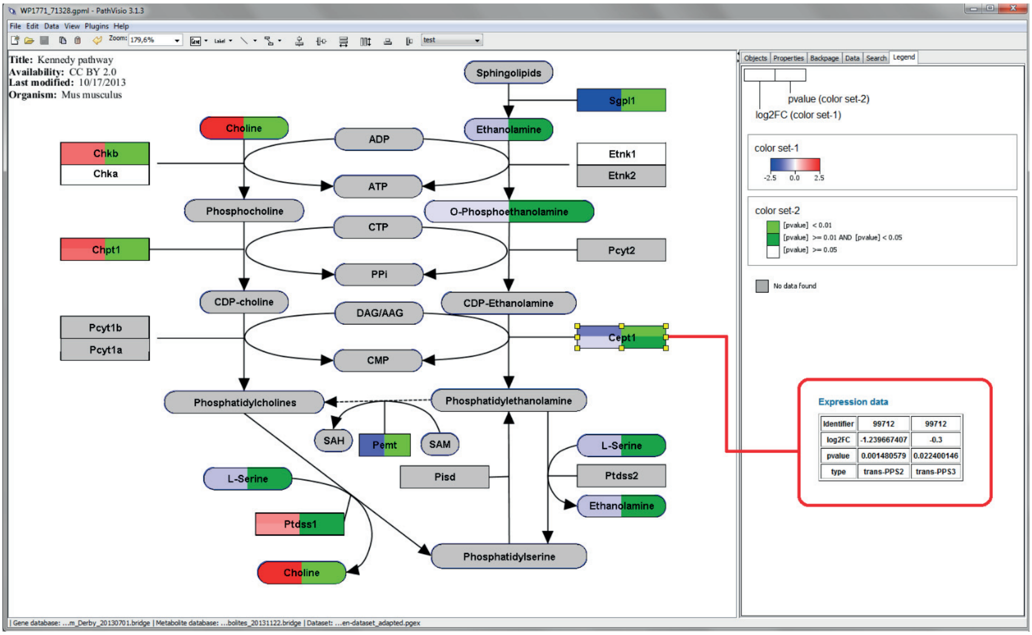
Figure 2.3 PathVisio 3, A Full-Powered Pathway Editor. (A) The basic drawing palette contains data nodes, interactions, graphical elements, cellular compartments and a few templates. Simple drag-and-drop mechanism allows users to add the elements in the pathway diagram. (B) The ACE inhibitor pathway 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.

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. 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.

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].



Title: Kennedy pathway
Availability: CC BY 2.0
Last modified: 10/17/2013
Organism: Mus musculus

Expression data

Identifier	99712	99712
log2FC	-1.239667407	-0.3
pvalue	0.001480579	0.022400146
type	trans-PPS2	trans-PPS3

| Gene database: ...m_Derby_20130701.bridge | Metabolite database: ...bolites_20131122.bridge | Dataset: ...en-dataset_adapted.pgex

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). The example dataset visualized in Figure 2.5 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.

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 over-representation analysis based on the statistical methods used in the MAPPFinder tool [15].

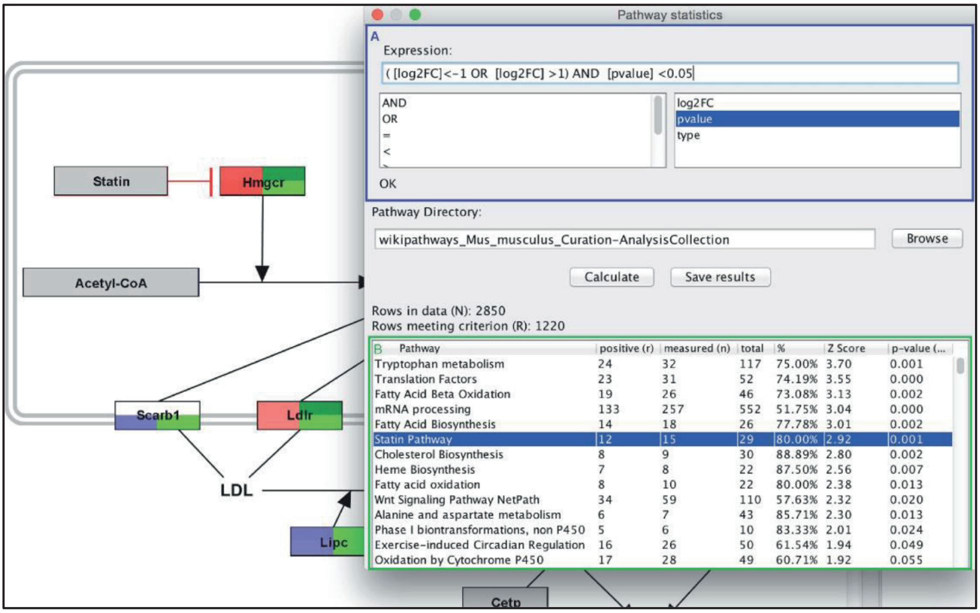
First, the user defines a criterion to select the differentially expressed genes in the dataset. In Figure 2.5A, the criteria filters genes with an absolute log2FC > 1 and a p-value < 0.05. The mouse pathway collection from WikiPathways was downloaded and selected.

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 → n) and the number of genes meeting the criterion (positive → r) (see Figure 2.5 B).

A commonly used score for over-representation 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:

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

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 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 analysed. Over-representation 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.



Expression: ([log2FC]<-1 OR [log2FC] >1) AND [pvalue] <0.05

Pathway Directory: wikipathways_Mus_musculus_Curation-AnalysisCollection

Rows in data (N): 2850
Rows meeting criterion (R): 1220

Pathway	positive (r)	measured (n)	total	%	Z Score	p-value (...
Tryptophan metabolism	24	32	117	75.00%	3.70	0.001
Translation Factors	23	31	52	74.19%	3.55	0.000
Fatty Acid Beta Oxidation	19	26	46	73.08%	3.13	0.002
mRNA processing	133	257	552	51.75%	3.04	0.000
Fatty Acid Biosynthesis	14	18	26	77.78%	3.01	0.002
Statin Pathway	12	15	29	80.00%	2.92	0.001
Cholesterol Biosynthesis	8	9	30	88.89%	2.80	0.002
Heme Biosynthesis	7	8	22	87.50%	2.56	0.007
Fatty acid oxidation	8	10	22	80.00%	2.38	0.013
Wnt Signaling Pathway NetPath	34	59	110	57.63%	2.32	0.020
Alanine and aspartate metabolism	6	7	43	85.71%	2.30	0.013
Phase I biontransformations, non P450	5	6	10	83.33%	2.01	0.024
Exercise-induced Circadian Regulation	16	26	50	61.54%	1.94	0.049
Oxidation by Cytochrome P450	17	28	49	60.71%	1.92	0.055

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.

Plugins in PathVisio

PathVisio 3 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 is a standard language to exchange biological pathway data [19]. 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 (Systems Biology Graphical Notation [17]) and MIM (Molecular Interaction Maps [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 48% of all human protein-coding genes (see supplementary material). 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 focussed 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 3 [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)

Future Directions

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

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].

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.

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.

Supplementary Files

Supplementary files are available from:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004085#sec005

My contributions:

- Extension of the GPML (Graphical Pathway Markup Language) data model used by PathVisio and WikiPathways to save interaction annotations
- Implementation of the user interface to annotate interactions in the PathVisio core software
- Development of the PathVisioRPC interface for PathVisio enabling automating pathway analysis workflows in many programming languages including R, Python and Perl.
- Contributions to the new PathVisio website, various tutorials, user and developer support, and writing of this manuscript

References

1. van Iersel, M., T. Kelder, A. Pico, K. Hanspers, S. Coort, B. Conklin, and C. Evelo, **Presenting and exploring biological pathways with PathVisio**. *BMC Bioinformatics*, 2008. **9**(1): p. 399.
2. Tisoncik, J.R., M.J. Korth, C.P. Simmons, J. Farrar, T.R. Martin, and M.G. Katze, **Into the eye of the cytokine storm**. *Microbiology and Molecular Biology Reviews*, 2012. **76**(1): p. 16-32.
3. Baetke, S.C., M.E. Adriaens, R. Seigneuric, C.T. Evelo, and L.M. Eijssen, **Molecular pathways involved in prostate carcinogenesis: insights from public microarray datasets**. *PloS one*, 2012. **7**(11): p. e49831.
4. Spaapen, F., G.G. van den Akker, M.M. Caron, P. Prickaerts, C. Rofel, V.E. Dahlmans, D.A. Surtel, Y. Paulis, F. Schweizer, and T.J. Welting, **The immediate early gene product EGR1 and polycomb group proteins interact in epigenetic programming during chondrogenesis**. *PloS one*, 2013. **8**(3): p. e58083.
5. Husi, H., T. Van Agtmael, W. Mullen, F.H. Bahlmann, J.P. 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 associated vascular disease**. *Circulation: Cardiovascular Genetics*, 2014: p. CIRCGENETICS. 113.000196.
6. van Iersel, M.P., M. Sokolovic, K. Lenaerts, M. Kutmon, F.G. Bouwman, W.H. Lamers, E. Mariman, and C.T. Evelo, **Integrated visualization of a multi-omics study of starvation in mouse intestine**. *J Integr Bioinform*, 2014. **11**: p. 235.
7. Jaeger, C., 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*, 2014. **10**(5): p. 859-876.
8. Zhong, J., J. Sharma, R. Raju, S.M. Palapetta, T.K. Prasad, T.-C. Huang, A. Yoda, J.W. Tyner, D. van Bodegom, and D.M. Weinstock, **TSLP signaling pathway map: a platform for analysis of TSLP-mediated signaling**. *Database*, 2014. **2014**: p. bau007.
9. Liu, X., J. Huang, S. Yang, Y. Zhao, A. Xiang, J. Cao, B. Fan, Z. Wu, J. Zhao, and S. Zhao, **Whole blood transcriptome comparison of pigs with extreme production of in vivo dsRNA-induced serum IFN-a**. *Developmental & Comparative Immunology*, 2014. **44**(1): p. 35-43.
10. Raju, R., S.M. Palapetta, V.K. Sandhya, A. Sahu, A. Alipoor, L. Balakrishnan, J. Advani, B. George, K.R. Kini, and N. Geetha, **A network map of FGF-1/FGFR signaling system**. *Journal of signal transduction*, 2014. **2014**.
11. Kelder, T., M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C.T. Evelo, and A.R. Pico, **WikiPathways: building research communities on biological pathways**. *Nucleic acids research*, 2012. **40**(D1): p. D1301-D1307.
12. Bauer, C., A. Glintschert, and J. Schuchhardt, **ProfileDB: A resource for proteomics and cross-omics biomarker discovery**. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 2014. **1844**(5): p. 960-966.
13. Hall, R., K. Pauls, S. McCulloch, and D. Savage, **OSGi in action: Creating modular applications in Java**. 2011: Manning Publications Co.
14. Börner, K., **Plug-and-play macroscopes**. *Communications of the ACM*, 2011. **54**(3): p. 60-69.
15. Doniger, S.W., N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor, and B.R. Conklin, **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biol*, 2003. **4**(1): p. R7.
16. Kohn, K.W., M.I. Aladjem, J.N. Weinstein, and Y. Pommier, **Molecular interaction maps of bioregulatory networks: a general rubric for systems biology**. *Molecular biology of the cell*, 2006. **17**(1): p. 1-13.
17. Le Novere, N., M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M.I. Aladjem, and S.M. Wimalaratne, **The systems biology graphical notation**. *Nature biotechnology*, 2009. **27**(8): p. 735-741.
18. Hucka, M., A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, and A. Cornish-Bowden, **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics*, 2003. **19**(4): p. 524-531.
19. Demir, E., M.P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, and J. Luciano, **The BioPAX community standard for pathway data sharing**. *Nature Biotechnology*, 2010. **28**(9): p. 935-942.
20. Cavalieri, D., 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*, 2007. **23**(19): p. 2631-2632.
21. van Iersel, M.P., A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo, **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC Bioinformatics*, 2010. **11**(1): p. 1.
22. Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova, **Entrez Gene: gene-centered information at NCBI**. *Nucleic acids research*, 2005. **33**(suppl 1): p. D54-D58.
23. Jennen, D.G., S. Gaj, P.J. Giesbertz, J.H. van Delft, C.T. Evelo, and J.C. Kleinjans, **Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes**. *Drug discovery today*, 2010. **15**(19): p. 851-858.
24. Chandan, K., M.P. van Iersel, M.I. Aladjem, K.W. Kohn, and A. Luna, **PathVisio-Validator: a rule-based validation plugin for graphical pathway notations**. *Bioinformatics*, 2012. **28**(6): p. 889-890.
25. Luna, A., M.L. Sunshine, M.P. van Iersel, M.I. Aladjem, and K.W. Kohn, **PathVisio-MIM: PathVisio plugin for creating and editing molecular interaction maps (MIMs)**. *Bioinformatics*, 2011. **27**(15): p. 2165-2166.
26. Fried, J.Y., M.P. van Iersel, M.I. Aladjem, K.W. Kohn, and A. Luna, **PathVisio-Faceted Search: an exploration tool for multi-dimensional navigation of large pathways**. *Bioinformatics*, 2013: p. btt146.
27. Kutmon, M., S. Lotia, C.T. Evelo, and A.R. Pico, **WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization**. *F1000Research*, 2014. **3**.
28. Smoot, M.E., K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics*, 2011. **27**(3): p. 431-432.
29. Khatri, P., M. Sirota, and A.J. Butte, **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS Comput Biol*, 2012. **8**(2): p. e1002375.

CHAPTER 3

Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment

Anwesha Bohler[1, 2], *Lars M T Eijssen*[1], *Martijn P van Iersel* [3], *Christ Leemans*[1], *Egon L Willighagen*[1], *Martina Kutmon*[1, 2, 4], *Magali Jaillard*[5], *and Chris T Evelo*[1, 2, 4]

1. Department of Bioinformatics - BiGCaT, Maastricht University, P.O. Box 616, UNS 50 Box 19, 6200 MD Maastricht, The Netherlands
2. Netherlands Consortium for Systems Biology (NCSB), The Netherlands
3. General Bioinformatics, Reading, Berkshire, RG4 7RT, United Kingdom
4. Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, P.O. Box 616, UNS 50 Box 19, 6200 MD Maastricht, The Netherlands
5. Bioinformatics Research Department, bioMérieux S.A, 69280 Marcy l'Etoile, France

Published in **BMC Bioinformatics** (2015) 16(267):1

Abstract

Background

Biological pathways are descriptive diagrams of biological processes widely used for functional analysis of differentially expressed genes or proteins. Primary data analysis, such as quality control, normalisation, and statistical analysis, is often performed in scripting languages like R, Perl, and Python. Subsequent pathway analysis is usually performed using dedicated external applications. Workflows involving manual use of multiple environments are time consuming and error prone. Therefore, tools are needed that enable pathway analysis directly within the same scripting languages used for primary data analyses. Existing tools have limited capability in terms of available pathway content, pathway editing and visualisation options, and export file formats. Consequently, making the full-fledged pathway analysis tool PathVisio available from various scripting languages will benefit researchers.

Results

We developed PathVisioRPC, an XMLRPC interface for the pathway analysis software PathVisio. PathVisioRPC enables creating and editing biological pathways, visualising data on pathways, performing pathway statistics, and exporting results in several image formats in multiple programming environments. We demonstrate PathVisioRPC functionalities using examples in Python. Subsequently, we analyse a publicly available NCBI GEO gene expression dataset studying tumour bearing mice treated with cyclophosphamide in R. The R scripts demonstrate how calls to existing R packages for data processing and calls to PathVisioRPC can directly work together. To further support R users, we have created RPathVisio simplifying the use of PathVisioRPC in this environment. We have also created a pathway module for the microarray data analysis portal ArrayAnalysis.org that calls the PathVisioRPC interface to perform pathway analysis. This module allows users to use PathVisio functionality online without having to download and install the software and exemplifies how the PathVisioRPC interface can be used by data analysis pipelines for functional analysis of processed genomics data.

Conclusions

PathVisioRPC enables data visualisation and pathway analysis directly from within various analytical environments used for preliminary analyses. It supports the use of existing pathways from WikiPathways or pathways created using the RPC itself. It also enables automation of tasks performed using PathVisio, making it useful to PathVisio users performing repeated visualisation and analysis tasks. PathVisioRPC is freely available for academic and commercial use at http://projects.bigcat.unimaas.nl/pathvisiorpc.

Background

Biological pathways are descriptive diagrams used to depict complex cellular processes such as metabolism, gene regulation, and signal transduction. Disturbances in such processes can cause disease. Pathways thus can help understand the functions of individual genes and proteins in terms of the systems and processes that contribute to normal physiology and to disease. Pathway analysis integrates and visualises global high throughput measurements and is a widely applied method among researchers to gain functional understanding from biological quantifications, such as gene expression and (relative) metabolite and protein abundances [1].

Pathways can be obtained from pathway databases. The Pathguide website [2, 3], at the time of writing, listed more than 570 pathway-related databases. A recent review found that among them Reactome [4, 5], KEGG [6], WikiPathways [7], Nature Pathway Interaction Database [8], and Pathway Commons [9] are most used [10]. Similarly, a number of tools exist for working with pathways, such as Ingenuity Pathway Analysis software [11], Pathway Studio [12], Metacore, PathVisio [13], Vanted [14], Reactome [4, 5], and Pathway Tools [15].

PathVisio is an open source tool for pathway editing and analysis that has seen rapid adoption by the scientific community since its release in 2008, counting 171 citations. A selection of the citing papers demonstrates the visualisation possibilities [16-20]. In addition, PathVisio is extendable using plugins such as PathVisio-Faceted Search, PathVisio-MIM, and PathVisio-Validator, to name a few [20-23]. In PathVisio, pathways are drawn as graphical diagrams containing nodes and edges, referred to as datanodes and interactions respectively. Datanodes represent biological entities such as genes, transcripts, proteins, and metabolites. Interactions can for instance indicate activation, inhibition, conversion, and catalysis. Both datanodes and interactions can be annotated using external database identifiers and all entities on a pathway can be annotated with literature references. The BridgeDb identifier mapping framework [24] is integrated into PathVisio for resolving identifiers from different databases. This allows mapping experimental measurements, such as gene expression, metabolite concentration, and fluxes, to pathway elements irrespective of the identifier formats used in the datasets or the pathways [25]. Pathways drawn in PathVisio are stored in GPML (Graphical Pathway Mark-up Language) files. The GPML file format is also used by WikiPathways to store pathway files. Users can create their own pathways, use pathways from WikiPathways, or pathways converted to GPML from other sources, such as KEGG, Reactome, and BioPAX [26]. Data can be visually represented on the pathways using colours, colour gradients, and textual labels. Furthermore, the core PathVisio program allows calculation of pathway statistics using an over-representation analysis algorithm (Z score) [27].

However, when researchers perform the same analysis using multiple sets of data and/or pathways using the graphical user interface of PathVisio they have to click the same set of buttons repeatedly for uploading the data, creating visualisations, performing pathway statistics, and exporting results, making the analyses error prone and time consuming on the user side. In this paper we describe the newly developed XMLRPC interface PathVisioRPC, which enables direct programmatic access to PathVisio from scripting languages. It enables automation of these steps by calling the defined functions, e.g. from inside a loop. An additional advantage is that a script written to automate PathVisio usage can be stored to enable later reference for validation and sharing purposes.

RPC stands for Remote Procedure Call. An RPC is initiated by a client that sends a request to a local or remote server to execute a task with supplied parameters. The server sends a response back to the client. XMLRPC uses XML to encode its calls and uses HTTP as a transport mechanism [28]. The PathVisioRPC interface wraps PathVisio functionality into XMLRPC functions that can be called from different programming languages to execute tasks. The interface serves as a communication channel between PathVisio and the script. Implementing the interface using XMLRPC makes the layer independent of the client's programming language or operating system. Additionally it allows the server to be run on one machine and to be called by users from other machines. Among the code bases most used for genomics data analysis are Bioconductor for R [29], BioPerl [30], and Biopython [31]. Each of these offers numerous data retrieval, normalisation, and statistical analysis packages that are often used prior to performing pathway analysis in PathVisio. XMLRPC functionality can be accessed from R and Perl using available add-on libraries; for Python no such library is necessary. The Bioconductor package repository in R provides an important toolset for bioinformatics data analysis. Therefore, we developed an R package RPathVisio for simplifying PathVisioRPC use in R [32]. For identifier mapping, RPathVisio uses the BridgeDbR package [33]. In addition, a module for pathway analysis was developed for ArrayAnalysis.org, an online microarray data analysis platform. This pathway module uses PathVisioRPC in the background [34].

Implementation

The PathVisioRPC interface (Figure 3.1) allows users to access functionality of the pathway analysis program PathVisio from R, Perl, Python, Java, C, C++, PHP, and many other programming languages. It allows access to all core functionalities of PathVisio including creation and modification of pathway diagrams, data visualisation, overrepresentation statistics, and various data exporters. The interface is available both as a PathVisio plugin [35] and as a standalone program. The latest versions can be obtained from the project website [36]. PathVisio's plugin manager provides an easy two-click installation of the PathVisioRPC plugin that automatically installs the HTML export plugin [37] as well. The latter provides the underlying functionality to create HTML exports of pathways with or without data visualisation. The standalone version is an executable Java archive (jar) file containing PathVisioRPC, the core PathVisio program, and the HTML export plugin.

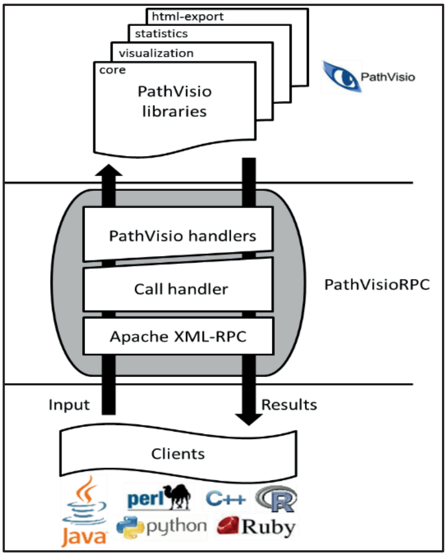


Figure 3.1 Interaction diagram of PathVisioRPC

PathVisioRPC is a client-server application that handles client requests to execute tasks after the server is launched. Running the PathVisioRPC.jar file launches the server on a port that is passed as an argument or on the default port 7777 of the local computer. The PathVisioRPC plugin for PathVisio adds a new "XMLRPC server setup" sub-menu item to the PathVisio "Plugins" menu. This is used to define the port for the server and to start or stop the server. A detailed API documentation along with code snippets from different languages for calling the implemented functions is available from the online documentation [38].

Results and Discussion

Example use of PathVisioRPC functionality

The following three examples use PathVisioRPC in Python to illustrate some of the functionalities implemented in PathVisioRPC.

Using same set of functions on multiple files

The first example demonstrates how the same functionalities can be applied to multiple files. The example1.py script loops over three files with gene expression data to create a pathway with the genes in each one of them and visualise the measurements. The script opens and reads each file, creates a new empty pathway, adds the genes in the file as nodes to the pathway, and visualises the gene expression data on the nodes [39]. In the example we generate a separate pathway for each list, it would also be possible to create a single pathway containing all the genes from the three lists.

Using different functionalities on different files

The second example demonstrates how files of different types can be combined/processed together. A pathway is created with a list of genes in a file and the data from a second file is displayed on the pathway created. The same data is then also visualised on a pathway from WikiPathways. The example2.py script reads a file with a list of genes, creates a new empty pathway, and then adds the genes as nodes to the newly created pathway. Subsequently, the script visualises the data from the second file on the pathway. The gene expression data is also visualised on the Statin pathway [40] from WikiPathways by passing its WikiPathways identifier (WP1) [39] to a dedicated function that has been implemented in PathVisioRPC for visualising data directly on WikiPathways pathways without first having to download those first. This requires an active internet connection. Alternatively, a previously downloaded pathway can be called upon using the local file path.

Looping over multiple files with different settings

The third example shows how multiple data files can be processed with different settings based on a configuration file. The script reads a settings file that contains the names of data files, associated species (human or mouse), and numbers of the data columns to be visualised. For each data file name present, the script loads the file, uses the correct species identifier mapping database, and visualises the data in the desired column(s) using a gradient on a WikiPathways pathway (cholesterol biosynthesis) from the correct species.

Multi-tissue time-series gene expression data analysis in R, for tumour-bearing mice treated with cyclophosphamide

As an illustration of the use of PathVisioRPC functionality to analyse a real genomics dataset we performed data visualisation and pathway analysis with gene expression data from a study by Moschella et al [41]. They performed gene expression profiling at several time points to study the effects of cyclophosphamide in bone marrow, spleen, and PBMCs of tumour-bearing mice, to gain insights into the core mechanisms of chemoimmunotherapy. We retrieved and analysed the microarray data sets for each of the three tissues and all time points in R as further described below. In short, we automated the entire analysis starting from retrieval of the data sets and the pathway collection, moving on to the pre-processing of the data sets, followed by differential expression analysis of the genes, pathway analysis, and visualisation of the data sets on WikiPathways pathways, and finally Gene Ontology Analysis and visualisation. For quality control and normalisation we used arrayQC, which is included in Additional File 4. All other scripts used and provided in Additional File 4 were custom made using existing R libraries.

Data Retrieval and Pre-processing of data

The gene expression datasets [41] (accession GSE27421, GSE27422, and GSE27423) were retrieved from the NCBI GEO database [42] using GEOquery [43]. The dataset contains 40 samples for bone marrow, 30 samples for peripheral blood leukocytes, and 38 samples for spleen tissue type, where each sample has been hybridised to microarray slides spotted with 13,443 70-mer oligonucleotides (Operon version 1.1; CRIBI Microarray Service, University of Padua, Italy). Next, the curated collection of mouse (*Mus musculus*) pathways was downloaded from WikiPathways using the WikiPathways web service [44].

The arrayQC workflow was used for quality control and normalisation of the retrieved datasets as they were obtained using the Genepix scanning platform. This workflow is similar to the published affyQC workflow for Affymetrix arrays [34]. The up-to-date version of the R scripts for the arrayQC workflow can be downloaded from the tab "Download Sources" at the ArrayAnalysis.org website [45]. Based on the quality control results, all arrays were of sufficient quality for inclusion in further analysis (see Additional File 5). The data were normalised using Lowess [46]. Based on the different quality indicators and flags from the quality control report of the arrayQC workflow, spots of insufficient quality were discarded. Furthermore, since each probe was spotted twice on the slides, the intensities of duplicate spots of sufficient quality were averaged. The GEOquery package was used to obtain the platform annotation file (GPL13209) to annotate the normalised data with NCBI gene identifiers.

Differential expression analysis

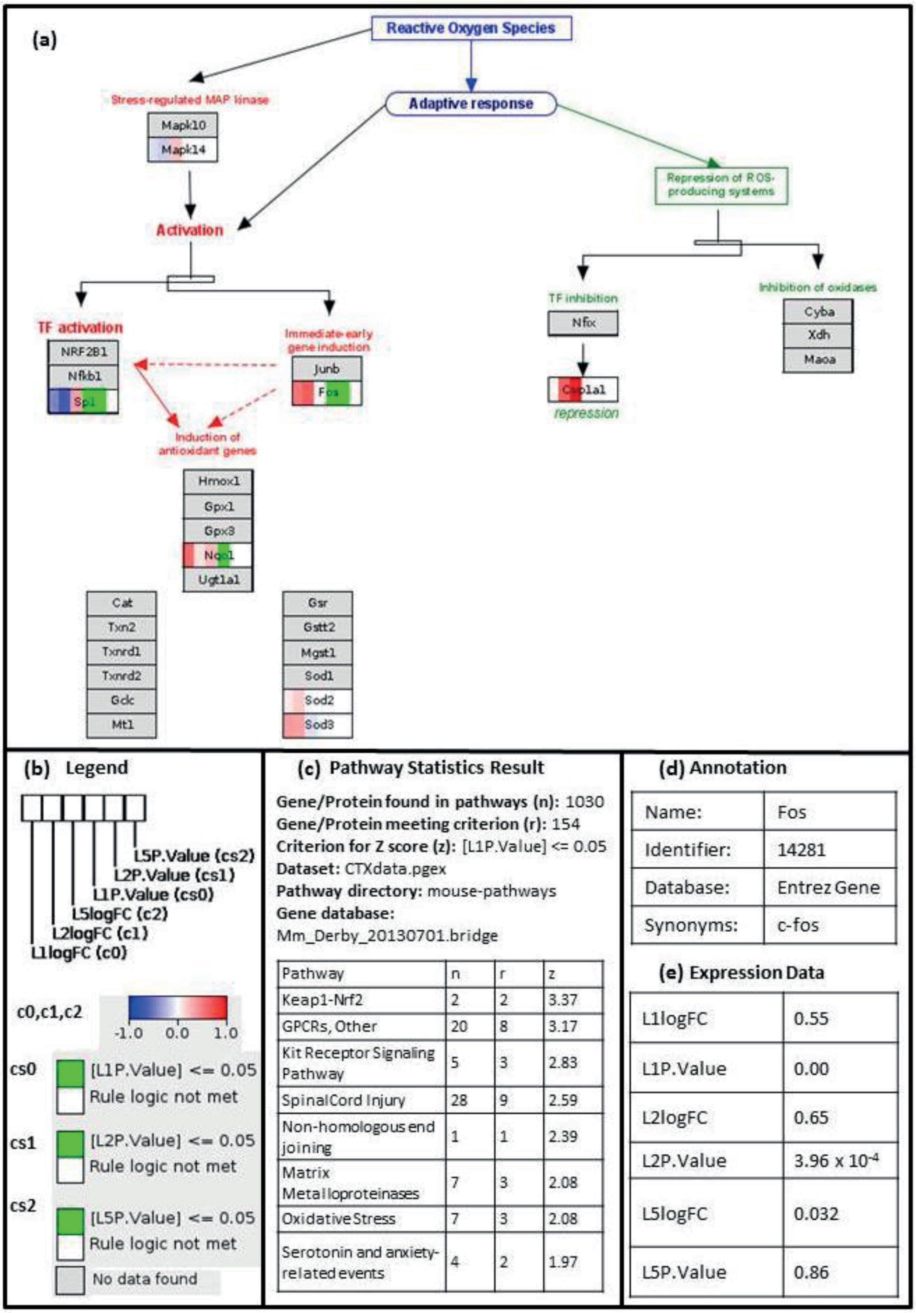
Differential expression analysis of the pre-processed normalised dataset was performed using the R package limma [47]. A 3x4 factorial design comparing every tissue with every treatment including the proper pairing arrangement regarding samples from the same individuals was used for fitting the data. Student's t-test comparisons were made between day 1, day 2, and day 5 vs. the control state (day 0), for each cell type: PBMCs, bone marrow cells, and splenocytes. The lists of statistical results for each of the nine comparisons were collated into a single file, which was used for further analysis (see Additional File 6).

Pathway over-representation analysis and data visualisation on WikiPathways pathways using PathVisioRPC

Pathway over-representation analysis was performed with the table of gene level statistics calculated in the previous step and the curated collection of WikiPathways mouse pathways, in order to identify the biological processes that were most affected in each tissue at each time point. Nine Z score tests were conducted, one for each cell type (PBMCs, bone marrow cells, and splenocytes) at each time point (day 1, day 2, and day 5). The Z score for over-representation analysis is calculated as a score for each pathway in the pathway collection, by subtracting the expected number of genes in a pathway meeting the criterion from the observed number of genes and dividing by the standard deviation of the observed number of genes [27]. The criterion for the calculation of the Z-score was the P-value for the gene level statistics being < 0.05.

The dataset was then visualised on the pathways. Log2 fold-changes of the genes were visualised using colour gradients (with blue, white, and red corresponding to the values -1, 0, 1) and significant genes were visualised using colour rules (P-value ≤ 0.05 represented by green and white otherwise) in each tissue for each time point. Genes in the pathways absent in the dataset were coloured light grey.

Next, the results of the analyses were exported as hyperlinked HTML pages containing an overview of the settings used to calculate the Z scores, a clickable list of pathways ranked according to their Z scores, a legend, and a frame to display the pathway when clicked [48]. The pathway images are PNG files embedded in the HTML and use image maps to link each component to its corresponding back page containing the data uploaded for that gene as well as links to relevant database entries. The legend contains the colour codes for the gene visualisation. Clicking on a gene in the pathway shows the gene identification and data uploaded for it in a new tab.



(c) Pathway Statistics Result

Gene/Protein found in pathways (n): 1030
Gene/Protein meeting criterion (r): 154
Criterion for Z score (z): [L1P.Value] <= 0.05
Dataset: CTXdata.pgex
Pathway directory: mouse-pathways
Gene database: Mm_Derby_20130701.bridge

Pathway	n	r	z
Keap1-Nrf2	2	2	3.37
GPCRs, Other	20	8	3.17
Kit Receptor Signaling Pathway	5	3	2.83
Spinal Cord Injury	28	9	2.59
Non-homologous end joining	1	1	2.39
Matrix Metalloproteinases	7	3	2.08
Oxidative Stress	7	3	2.08
Serotonin and anxiety-related events	4	2	1.97

(d) Annotation

Name:	Fos
Identifier:	14281
Database:	Entrez Gene
Synonyms:	c-fos

(e) Expression Data

L1logFC	0.55
L1P.Value	0.00
L2logFC	0.65
L2P.Value	3.96 x 10^{-4}
L5logFC	0.032
L5P.Value	0.86

Figure 3.2 Pathway Statistics Results for PBMCs. (a) Oxidative stress pathway [49] for PBMCs showing the logFC and P-value for day 1, day 2 and day 5, (b) Legend showing the colour gradients and rules used to visualise logFC and P-value of the genes, every data node is divided into 6 columns, 3 for the logFCs and 3 for the P-values of PBMCs (L) at the three time points, day 1 (1), day 2 (2) and day 5 (5) (c) Parameters used to calculate the Z score and ranked list of pathways, and (d) Back page showing the annotation for the gene Fos, (e) Back page showing the gene expression data for PBMCs for the gene Fos for day 1, day 2, and day 5.

Gene Ontology (GO) enrichment analysis and data visualisation on GO terms using PathVisioRPC

Furthermore, Gene Ontology (GO) enrichment analysis was performed using the topGO Bioconductor package [50] separately for the three cell types PBMCs, bone marrow cells, and splenocytes. The significant genes for each tissue type (all time points combined) were selected as the numerator list and the entire dataset was taken as the denominator list. A list of the top 50 terms for each cell type was obtained as a result (see Additional file 7). PathVisioRPC was then used to create one pathway per cell type, with the highly enriched GO terms as nodes. Subsequently, the gene expression data for each cell type was visualised on the respective pathway. The log2 fold-changes of the genes were visualised in rows on the GO terms they belong to, using the same colour gradient as used before (blue, white, and red, corresponding to the values -1, 0, 1). For each cell type, the pathway created using enriched GO terms along with the visualised results, was exported in HTML format [51]. Clicking on the GO nodes of the pathway opens a new tab in the browser, containing columns of data for each of the genes in the dataset belonging to that GO term.

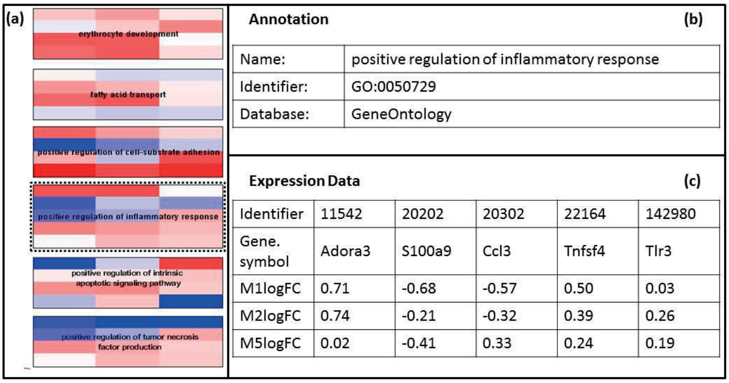
Interpreting the pathway and gene ontology (GO) analysis results

Overall, the Z score indicates whether the number of genes meeting the criterion is higher or lower than what is expected based on the complete dataset. A positive score indicates that in that pathway more genes are changed than expected; a negative score means fewer genes are changed than expected. Thus, it can be inferred that the highest-ranking pathways are potentially the most interesting ones for the given conditions and need to be evaluated further, since they were detected to contain an overabundance of differentially expressed genes. Subsequently, the involvement of the pathways in cellular functions has to be determined for a biological interpretation of the results.

We used the Oxidative stress pathway [49] to illustrate the visualisation produced using PathVisioRPC. Every data node on the pathway is divided into 6 columns; the first 3 columns are used to visualise the log2 fold change for the transcript at each time point (day 1, day 2, and day 5) and the last 3 columns are used to visualise the P-value for that time point (day 1, day 2, and day 5). In PBMCs, the transcripts for FOS and NQO1 are significantly upregulated for day 1 and/or day 2. Moreover, the transcript for SP1 is down regulated for day 1 and day 2 (Figure 3.2). In splenocytes, the transcripts for MAPK14, SP1, and CYP1A1 are all significantly down regulated for either day 1 or both day 1 and day 2 (see Additional File 8). However, in bone marrow cells, the transcripts for MAPK14 and SP1 are both significantly upregulated for day 1 and 2, while the gene FOS is significantly down regulated (see Additional File 9). In all the three cell types, the expressions of the genes seem to return to baseline levels by day 5 as also observed in the original publication.

In addition to the pathway analysis, a GO term enrichment analysis can shed additional light on the processes that are targeted by the experimental conditions. GO terms describe functions that can span over several pathways or be involved in multiple cellular processes or are not known in large enough detail to be represented in pathway collections. Thus, the discovery of GO terms over-represented in the data sets can lead to information that is potentially not available from pathways. From the GO term enrichment analysis it can be deduced which functions are highly involved or targeted by the experimental conditions. In addition, the visualisation of the gene expression data on the enriched GO terms can allow determining whether these functions are impacted positively or negatively by the treatment, or whether they display a mixed variation. PathVisioRPC can be used for creating such a visualisation. For the dataset presented here, transcripts with decreased abundance are visualised in blue and transcripts with increased abundance are visualised in red.

In bone marrow cells, processes related to defence response, developmental processes, and response to external/chemical stimuli are highly affected, as was also observed in the original publication [51]. For instance, the GO term “positive regulation of inflammatory response” is highly stimulated and the transcripts in the dataset related to that term are mostly upregulated (Figure 3.3). For splenocytes, transcripts increased in abundance are mostly involved in cell cycle arrest and regulation, whereas the original publication reported these to be down regulated. Even though the original publication reports that transcripts involved in immune response are mostly upregulated, it is clear from the present analysis that these transcripts show a 50/50 ratio between up- and down-regulation. Whether that is caused by real differences in remaining data after quality control and statistical evaluation or just differences in how data is communicated is not clear since the processed data from the original publication is not available. The GO terms RNA processing and proteolysis contain mostly transcripts down regulated under the experimental conditions [51]. In PBMCs, similar to the original publication, only few GO terms were enriched for up-regulated transcripts. For down-regulated transcripts, regulation of gene expression is the most notable term that is enriched. Other enriched GO terms for PBMCs contain approximately as many up- and down-regulated transcripts [51].


(a)
erythrocyte development
fatty acid transport
positive regulation of cell-substrate adhesion
positive regulation of inflammatory response
positive regulation of intrinsic apoptotic signaling pathway
positive regulation of tumor necrosis factor production
(b)
(c)

Annotation

Name:	positive regulation of inflammatory response
Identifier:	GO:0050729
Database:	GeneOntology

Expression Data

Identifier	11542	20202	20302	22164	142980
Gene. symbol	Adora3	S100a9	Ccl3	Tnfsf4	Tlr3
M1logFC	0.71	-0.68	-0.57	0.50	0.03
M2logFC	0.74	-0.21	-0.32	0.39	0.26
M5logFC	0.02	-0.41	0.33	0.24	0.19

Figure 3.3 Gene Ontology Enrichment analysis for Bone Marrow Cells. (a) Gene Expression Data visualised on Gene Ontology Terms, (b) Back page showing the Gene Ontology Annotation for GO term GO:0050729, positive regulation of inflammatory response, (c) Back page showing the Gene expression data for the five genes (Adora3, S100a9,Ccl3, Tnfsf4, and Tlr3) found in the dataset which map to the GO class positive regulation of inflammatory response.

RPathVisio: A package for performing pathway analysis and data visualisation in the R statistical environment

RPathVisio (see Additional File 10, submitted to the Bioconductor repository) makes use of the PathVisioRPC interface for providing direct access to PathVisio functionality and the wealth of WikiPathways pathways within the R statistical environment. RPathVisio not only hides the technicalities of an XMLRPC call, but also verifies the user input to check whether the correct system code has been used or not. The package depends on the XMLRPC and BridgeDbR packages. The BridgeDbR package is also new and has been submitted to the Bioconductor repository (see Additional File 10). It facilitates identifier mapping in R. When called from RPathVisio, it downloads the necessary identifier mapping databases from the BridgeDb website [52] and uses them to check whether known database identifiers have been used to annotate the data.

RPathVisio provides simple R commands to perform the PathVisioRPC API calls. For instance, the PathVisioRPC call in R for creating a pathway using the pure XMLRPC formalism is:

```
xml.rpc(server_address,"PathVisio.createPathway","Glycolysis")
```

where server_address refers to the address at which the PathVisioRPC server has been launched. However, when using RPathVisio the call is simplified to:

```
createPathway(name="Glycolysis").
```

Several other packages are available in R to access pathways as gene sets and perform pathway analyses. Notable among them are PathView [53], KEGGGraph [54], ReactomePA [55] and sigPathway [56], which provide access to the KEGG [57], Reactome [4, 5], BioCarta [58], and BioCyc [59] pathway collections. The KEGGGraph and PathView packages offer pathway visualisation on KEGG pathways. The RCytoscape package offers data visualisation styles as well, but the data can be visualised solely on networks created in Cytoscape. RPathVisio provides users access to the entire WikiPathways collection, including pathways from the Reactome [4, 5], NetPath [60], and WormBase [61] collections. Pathways from other pathway resources can also be used as long as these can be exported in BioPAX format, which can then be transformed into GPML used by WikiPathways. Results from RPathVisio can be exported as GPML files, images (PNG, SVG, PDF), and HTML files. The HTML export report is navigable as a mini self-contained website containing a list of all pathways used for the analysis sorted by their Z scores, the data visualised pathway images, and the corresponding measurements per gene, protein, or metabolite. A comparison of RPathVisio and these other packages is available in Table 3.1. There are alternatives available for BridgeDbR as an identifier-mapping tool. However, RPathVisio uses BridgeDbR to verify whether the annotations used in the data are suitable for use with PathVisio, which uses the same BridgeDb approach.

Table 3.1 Comparison of RPathVisio with other similar packages available in R.

Software	**SigPathway**	**ReactomePA**	**KEGGGraph**	**PathView**	**RPathVisio**
Feature					
Pathway sources available	GO, KEGG, BioCarta, BioCyc, SuperArray	Reactome	KEGG	KEGG	WikiPathways, Reactome, NetPath, WormBase
Pathway building	—	—	+	—	+
Multi-omics support	—	—	+	+	+
Plots	—	+	—	+	—
Pathway visualisation	—	—	+	+	+
Multiple data visualisation	—	—	+	+	+
Pathway statistics	Gene set statistics	EA, GSEA FM detection	—	—	EA
Export	Text, HTML	Text	Images	Images	Text, GPML, Images, HTML

GO Gene Ontology, EA Enrichment Analysis, GSEA Gene Set Enrichment Analysis, FM Functional Module +Present, —Absent

Pathway module of ArrayAnalysis.org

As an example of workflow integration, we have developed a module for pathway analysis for ArrayAnalysis.org, a web-based platform for microarray data analysis. It has a modular setup with quality control and normalisation modules for several types of microarray platforms. The workflow has been extended with modules for statistical analysis and pathway analysis that accept data either from the built-in data normalisation modules or from user uploads. PathVisioRPC powers the pathway module of ArrayAnalysis.org. This workflow is available at the ArrayAnalysis.org website [45].

Online Form Description

The online pathway module (Figure 3.4) in ArrayAnalysis.org accepts delimited text files of different kinds, for example a statistics results file such as produced by limma [47]. The genes, proteins, and metabolites should be annotated using database identifiers in order to be recognised by the internal BridgeDB identifier mapping support of PathVisio. When identifiers from a single database are used, the database can be selected from a drop down list, otherwise the file should have a column containing the system codes of the databases [62]. The identifier mapping database and pathway collections are determined based on the species selected by the user. The pathway collection used by default is the curated analysis collection of pathways from WikiPathways for the species selected and the default identifier-mapping file used is the gene/protein database from BridgeDb for that species. The user can choose a different pathway collection or identifier-mapping file at the end of the form.

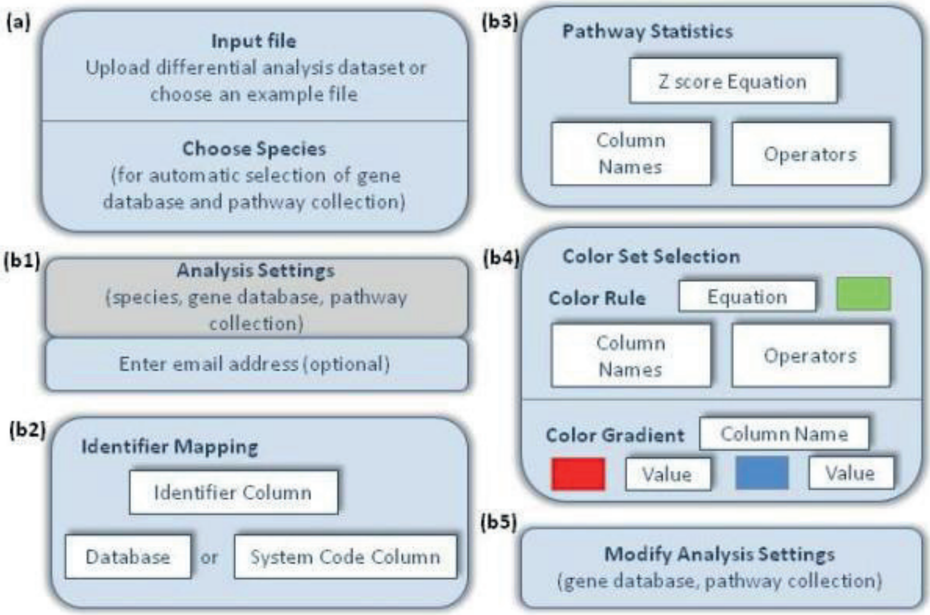


Figure 3.4 Schematic representation of the input wizard of the Pathway Analysis module of ArrayAnalysis.org. (a) Allows upload of a dataset (e.g. differential analysis data, metabolite concentration data) and to select species; (b1) Shows the selected species, gene identifier mapping database, pathway collection, and asks for an optional email address; (b2) Selects Identifier Column and a Database or a System Code Column; (b3) Specifies criterion for Z score calculation; (b4) Chooses colour rules and/or gradients; and (b5) Modifies gene database and pathway collection.

In the next form, the user can specify the colour gradients and/or colour rules to visualise the data on the pathways. To choose a colour gradient for a variable, the user selects the variable and then two or three colours and corresponding values to create a gradient. In order to choose a colour rule, the user enters a criterion and a colour to be used if the criterion is met. For performing pathway statistics, a criterion for calculating the Z score needs to be set. After setting the options, on clicking Run, the pathway module is launched. The form inputs are converted into R commands which in turn call the PathVisioRPC server for executing the tasks. After completion of the run, results display on screen.

Conclusions

PathVisioRPC enables automated calls to PathVisio to support users in performing repeated analyses and building workflows. We have demonstrated the advantages of using PathVisioRPC with scripts written in Python and R, commonly used scripting languages for bioinformatics analysis. PathVisioRPC simplifies the task of creating pathways and visualising data using multiple files as demonstrated by our example use cases in Python. It also allows users to perform multiple pathway statistics comparisons as shown by the biological analysis in R that compares the transcriptomics profile of three different tissues in three different time-points. We have also shown that the functionality provided by the interface can be easily combined with existing analysis packages from Bioconductor, as existing R packages can be used for statistical analysis and GO analysis and then RPathVisio can be used for pathway building, pathway analysis, data visualisation on pathways, and export. The same approach could be followed in other programming languages, example code snippets are available in R, Perl, and Python from the project website [37]. The pathway analysis and data visualisation results are presented as hyperlinked HTML pages. This makes it possible to manoeuvre through the results and explore the interesting pathways. In the case of a pathway diagram created with enriched GO terms, clicking on the nodes displays the data for all the genes annotated with that GO term. This gives a quick overview of how each enriched GO term is affected.

Writing scripts to create pathways, visualise data, perform pathway statistics comparisons, and export data visualised pathway images makes these tasks less error- prone and time- consuming for the user as compared to doing this manually. Using a script also allows the user to retain method provenance. This ensures reproducibility, reference at a later time, and sharing the method used with others. Therefore, PathVisioRPC simplifies workflows involving the visualisation of experimental data on multiple pathway sets from WikiPathways and allows repeated over-representation analyses in various programming languages and for different gene classification approaches such as GO. This allows researchers to integrate PathVisio as a visualisation and pathway analysis tool in workflows, facilitating automated downstream analysis of (multiple) datasets. These pathways can be exported with or without included data visualisation in various image formats for publication, uploaded to WikiPathways for community curation, or used in Cytoscape [63] for network analysis using the WikiPathways app [64].

Furthermore, the pathway module of ArrayAnalysis.org illustrates how the data visualisation and over-representation analysis of PathVisio can be incorporated into an existing data analysis pipeline using PathVisioRPC. This online module can also be particularly useful for researchers who only need pathway analysis once for their current dataset, allowing them to use the core functionalities of PathVisio over the internet without having to download and install the software itself.

Reactome pathways can be used in analysis already as they have been converted to GPML. Similarly other pathway sets (e.g. KEGG and BioPAX) can be used in analysis when these are converted into GPML. The functionality can also be easily extended by registering functionalities of other PathVisio plugins like the GSEA plugin [65] in the PathVisioRPC interface. There are some ongoing initiatives to use PathVisioRPC in larger analysis approaches. We are aware of pathway visualisation modules planned for a project regarding gene-based clinical assays to guide therapy in leukaemia patients and one for the Open PHACTS [66] project that will use the PathVisioRPC interface as well. Finally, the ArrayAnalysis.org pipeline is part of the larger dbNP project [67] that would contain other data analysis pipelines that are meant to make use of the pathway module provided at ArrayAnalysis.org as well. PathVisioRPC thus offers versatile solutions to integrate pathway analysis and visualisation approaches in many workflows.

Availability and Requirements

Project name: PathVisioRPC

Project home page: http://projects.bigcat.unimaas.nl/pathvisiorpc

Operating system(s): Platform-independent

Programming language: Java

Other requirements: None

License: Apache 2 License

Any restrictions to use by non-academics: None other than those defined by the license.

Additional Files:

Additional files are available from:
http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0708-8

References

1. Khatri, P., M. Sirota, and A.J. Butte, **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS Comput Biol*, 2012. **8**(2): p. e1002375.
2. Bader, G.D., M.P. Cary, and C. Sander, **Pathguide: a pathway resource list**. *Nucleic acids research*, 2006. **34**(suppl 1): p. D504-D506.
3. **Pathguide: the pathway resource list** Available from: http://www.pathguide.org/.
4. Croft, D., **Building models using Reactome pathways as templates**. *In Silico Systems Biology*, 2013: p. 273-283.
5. Milacic, M., R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, and L. Stein, **Annotating cancer variants and anti-cancer therapeutics in reactome**. *Cancers*, 2012. **4**(4): p. 1180-1211.
6. Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic Acids Research*, 2014. **42**(D1): p. D199-D205.
7. Kelder, T., M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C.T. Evelo, and A.R. Pico, **WikiPathways: building research communities on biological pathways**. *Nucleic acids research*, 2012. **40**(D1): p. D1301-D1307.
8. Schaefer, C.F., K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K.H. Buetow, **PID: the pathway interaction database**. *Nucleic Acids Research*, 2009. **37**(suppl 1): p. D674-D679.
9. Cerami, E.G., B.E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G.D. Bader, and C. Sander, **Pathway Commons, a web resource for biological pathway data**. *Nucleic Acids Research*, 2011. **39**(suppl 1): p. D685-D690.
10. Bauer ‐ Mehren, A., L.I. Furlong, and F. Sanz, **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Molecular systems biology*, 2009. **5**(1).
11. **Ingenuity Pathway Analysis**. Available from: www.qiagen.com/ingenuity.
12. **Pathway Studio**. Available from: http://www.ariadnegenomics.com/products/pathway-studio/.
13. van Iersel, M.P., T. Kelder, A.R. Pico, K. Hanspers, S. Coort, B.R. Conklin, and C. Evelo, **Presenting and exploring biological pathways with PathVisio**. *BMC bioinformatics*, 2008. **9**(1): p. 399.
14. Junker, B.H., C. Klukas, and F. Schreiber, **VANTED: a system for advanced data analysis and visualization in the context of biological networks**. *BMC Bioinformatics*, 2006. **7**(1): p. 109.
15. Karp, P.D., S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T.J. Lee, P. Kaipa, F. Gilham, A. Spaulding, and L. Popescu, **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology**. *Briefings in bioinformatics*, 2009: p. bbp043.
16. Kursawe, R., M. Eszlinger, D. Narayan, T. Liu, M. Bazuine, A.M. Cali, E. D'Adamo, M. Shaw, B. Pierpont, and G.I. Shulman, **Cellularity and adipogenic profile of the abdominal subcutaneous adipose tissue from obese adolescents: association with insulin resistance and hepatic steatosis**. *Diabetes*, 2010. **59**(9): p. 2288-2296.
17. Tisoncik, J.R., M.J. Korth, C.P. Simmons, J. Farrar, T.R. Martin, and M.G. Katze, **Into the eye of the cytokine storm**. *Microbiology and Molecular Biology Reviews*, 2012. **76**(1): p. 16-32.
18. Jitendra, S., A. Nanda, S. Kaur, and M. Singh, **A comprehensive molecular interaction map for Hepatitis B virus and drug designing of a novel inhibitor for Hepatitis BX protein**. *Bioinformation*, 2011. **7**(1): p. 9.
19. Zhou, C., Q. Zhong, L.V. Rhodes, I. Townley, M.R. Bratton, Q. Zhang, E.C. Martin, S. Elliott, B.M. Collins-Burow, and M.E. Burow, **Proteomic analysis of acquired tamoxifen resistance in MCF-7 cells reveals expression signatures associated with enhanced migration**. *Breast Cancer Res*, 2012. **14**(2): p. R45.
20. Rubio-Aliaga, I., B. de Roos, M. Sailer, G.A. McLoughlin, M.V. Boekschoten, M. van Erk, E.-M. Bachmair, E.M. Van Schothorst, J. Keijer, and S.L. Coort, **Alterations in hepatic one-carbon metabolism and related pathways following a high-fat dietary intervention**. *Physiological genomics*, 2011. **43**(8): p. 408-416.
21. Fried, J.Y., M.P. van Iersel, M.I. Aladjem, K.W. Kohn, and A. Luna, **PathVisio-Faceted Search: an exploration tool for multi-dimensional navigation of large pathways**. *Bioinformatics*, 2013: p. btt146.
22. Chandan, K., M.P. van Iersel, M.I. Aladjem, K.W. Kohn, and A. Luna, **PathVisio-Validator: a rule-based validation plugin for graphical pathway notations**. *Bioinformatics*, 2012. **28**(6): p. 889-890.
23. Luna, A., M.L. Sunshine, M.P. van Iersel, M.I. Aladjem, and K.W. Kohn, **PathVisio-MIM: PathVisio plugin for creating and editing molecular interaction maps (MIMs)**. *Bioinformatics*, 2011. **27**(15): p. 2165-2166.
24. van Iersel, M.P., A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo, **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC Bioinformatics*, 2010. **11**(1): p. 1.
25. van Iersel, M.P., M. Sokolovic, K. Lenaerts, M. Kutmon, F.G. Bouwman, W.H. Lamers, E. Mariman, and C.T. Evelo, **Integrated visualization of a multi-omics study of starvation in mouse intestine**. *J Integr Bioinform*, 2014. **11**: p. 235.
26. Demir, E., M.P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, and J. Luciano, **The BioPAX community standard for pathway data sharing**. *Nature Biotechnology*, 2010. **28**(9): p. 935-942.
27. Doniger, S.W., N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor, and B.R. Conklin, **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biol*, 2003. **4**(1): p. R7.
28. Laurent, S.S., J. Johnston, E. Dumbill, and D. Winer, **Programming web services with XML-RPC**. 2001: " O'Reilly Media, Inc.".
29. Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, and J. Gentry, **Bioconductor: open software development for computational biology and bioinformatics**. *Genome biology*, 2004. **5**(10): p. R80.
30. Stajich, J.E., D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, I. Korf, and H. Lapp, **The Bioperl toolkit: Perl modules for the life sciences**. *Genome research*, 2002. **12**(10): p. 1611-1618.
31. Cock, P.J., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, and B. Wilczynski, **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics*, 2009. **25**(11): p. 1422-1423.
32. Leemans, C., A. Bohler, and E.L. Willighagen. **RPathVisio**. Available from: https://github.com/BiGCAT-UM/RPathVisio.
33. Leemans, C., A. Bohler, and E.L. Willighagen. **BridgeDbR**. Available from: https://github.com/BiGCAT-UM/bridgedb-r.
34. Eijssen, L.M., M. Jaillard, M.E. Adriaens, S. Gaj, P.J. de Groot, M. Müller, and C.T. Evelo, **User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis. org**. *Nucleic Acids Research*, 2013. **41**(W1): p. W71-W76.
35. A, B. **PathVisioRPC plugin**. Available from: http://www.pathvisio.org/plugin/pathvisiorpc-2/.

36. **PathVisioRPC project website**. Available from: http://projects.bigcat.unimaas.nl/pathvisiorpc/.

37. **HTML Export Plugin**.

38. Bohler, A. **PathVisioRPC Documentation** Available from: [http://projects.bigcat.unimaas.nl/pathvisiorpc/documentation].

39. Bohler, A. **Python Examples Results**. Available from: http://projects.bigcat.unimaas.nl/data/pathvisiorpc/data/Python_Results/.

40. Pico, A., K. Hanspers, B. Conklin, and N. Salomonis. **Statin Pathway for *Mus musculus***. Available from: http://wikipathways.org/index.php/Pathway:WP1.

41. Moschella, F., M. Valentini, E. Aricò, I. Macchia, P. Sestili, M.T. D'Urso, C. Alessandri, F. Belardelli, and E. Proietti, **Unraveling cancer chemoimmunotherapy mechanisms by gene and protein expression profiling of responses to cyclophosphamide**. *Cancer research*, 2011. **71**(10): p. 3528-3539.

42. Barrett, T., S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, and M. Holko, **NCBI GEO: archive for functional genomics data sets—update**. *Nucleic Acids Research*, 2013. **41**(D1): p. D991-D995.

43. Davis, S. and P.S. Meltzer, **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor**. *Bioinformatics*, 2007. **23**(14): p. 1846-1847.

44. Kelder, T., A.R. Pico, K. Hanspers, M.P. Van Iersel, C. Evelo, and B.R. Conklin, **Mining biological pathways using WikiPathways web services**. *PloS one*, 2009. **4**(7): p. e6447.

45. Bohler, A., M. Jaillard, and L. Eijssen, **ArrayAnalysis pathway module**.

46. Cleveland, W.S., **LOWESS: A program for smoothing scatterplots by robust locally weighted regression**. *The American Statistician*, 1981. **35**(1): p. 54.

47. Smyth, G.K., **Limma: linear models for microarray data**, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 397-420.

48. Bohler, A. **Pathway Analysis Results**. Available from: http://projects.bigcat.unimaas.nl/data/pathvisiorpc/data/PathwayAnalysis/.

49. Evelo, C., D. Digles, R. I, and N. Reyes, **Oxidative Stress Pathway for Mus musculus**.

50. Alexa, A., J. Rahnenfuhrer, M. Alexa, and A. Suggests, **biocViews Microarray B: Package 'topGO'**.

51. Bohler, A. **Gene Ontology Analysis**. Available from: http://projects.bigcat.unimaas.nl/data/pathvisiorpc/data/GOanalysis/.

52. **BridgeDb Identifier Mapping Databases**.

53. Luo, W. and C. Brouwer, **Pathview: an R/Bioconductor package for pathway-based data integration and visualization**. *Bioinformatics*, 2013: p. btt285.

54. Zhang, J.D. and S. Wiemann, **KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor**. *Bioinformatics*, 2009. **25**(11): p. 1470-1471.

55. Yu, G., **Reactome Pathway Analysis**. *Homo*, 2012: p. 1266738:1266738.

56. Lai, W., L. Tian, and P. Park, **SigPathway: pathway analysis with microarray data**. 2013.

57. Kotera, M., M. Hirakawa, T. Tokimatsu, S. Goto, and M. Kanehisa, **The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals**. *Next Generation Microarray Bioinformatics: Methods and Protocols*, 2012: p. 19-39.

58. Nishimura, D., **BioCarta**. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2001. **2**(3): p. 117-120.

59. Karp, P.D., C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, S. Tsoka, N. Darzentas, V. Kunin, and N. López-Bigas, **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes**. *Nucleic Acids Research*, 2005. **33**(19): p. 6083-6089.

60. Kandasamy, K., S.S. Mohan, R. Raju, S. Keerthikumar, G.S.S. Kumar, A.K. Venugopal, D. Telikicherla, D.J. Navarro, S. Mathivanan, and C. Pecquet, **NetPath: a public resource of curated signal transduction pathways**. *Genome biology*, 2010. **11**(1): p. 1-9.

61. Harris, T.W., J. Baran, T. Bieri, A. Cabunoc, J. Chan, W.J. Chen, P. Davis, J. Done, C. Grove, and K. Howe, **WormBase 2014: new views of curated biology**. *Nucleic acids research*, 2014. **42**(D1): p. D789-D793.

62. **System Codes List**. Available from: http://www.pathvisio.org/documentation/system-codes/.

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

64. Kutmon, M., S. Lotia, C.T. Evelo, and A.R. Pico, **WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization**. *F1000Research*, 2014. **3**.

65. Bilican, A. and M. Kutmon. **GSEA plugin**. Available from: http://www.pathvisio.org/plugin/gene-set-enrichtment-analysis-plugin/.

66. **OpenPHACTS**. Available from: http://www.openphacts.org/.

67. van Ommen, B., J. Bouwman, L.O. Dragsted, C.A. Drevon, R. Elliott, P. de Groot, J. Kaput, J.C. Mathers, M. Müller, and F. Pepping, **Challenges of molecular nutrition research 6: the nutritional phenotype database to store, share and evaluate nutritional systems biology studies**. *Genes & nutrition*, 2010. **5**(3): p. 189-203.

Chapter 4

Reactome from a WikiPathways perspective

Anwesha Bohler [1], *Guanming Wu* [2], *Martina Kutmon* [1,3], *Leontius Adhika Pradhana* [4], *Susan L Coort* [1], *Kristina Hanspers* [5], *Robin Haw* [2], *Alexander R Pico* [5], *Chris T Evelo* [1,3]

1. Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands
2. Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada
3. Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands
4. Department of Pharmacy, Faculty of Science, National University of Singapore, Singapore, Republic of Singapore
5. Gladstone Institutes, San Francisco, California, United States of America

Published in **PLoS Comput Biol** (2016) 12(5): e1004941.

Abstract

Reactome and WikiPathways are two of the most popular freely available databases for biological pathways. Reactome pathways are centrally curated with periodic input from selected domain experts. WikiPathways is a community-based platform where pathways are created and continually curated by any interested party. The nascent collaboration between WikiPathways and Reactome illustrates the mutual benefits of combining these two approaches. We created a format converter that converts Reactome pathways to the GPML format used in WikiPathways. In addition, we developed the ComplexViz plugin for PathVisio which simplifies looking up complex components. The plugin can also score the complexes on a pathway based on a user defined criterion. This score can then be visualized on the complex nodes using the visualization options provided by the plugin.

We analyzed a published publicly available gene expression dataset about polycystic ovary syndrome using the merged collection of curated and converted Reactome pathways to illustrate the improved coverage of biological processes. This conversion allows researchers to visualize their data on Reactome pathways using PathVisio's advanced data visualization functionalities.

WikiPathways benefits from the dedicated focus and attention provided to the content converted from Reactome and the wealth of semantic information about interactions. Reactome in turn benefits from the continuous community curation available on WikiPathways. The research community at large benefits from the availability of a larger set of pathways for analysis in PathVisio and Cytoscape. The pathway statistics results obtained from PathVisio are significantly better when using a larger set of candidate pathways for analysis. The conversion serves as a general model for integration of multiple pathway resources developed using different approaches.

Author Summary

Biological pathways are descriptive diagrams that describe biological processes, i.e. interactions between genes, proteins, and metabolites. Pathways can therefore be used to integrate and visualize molecular measurements of genes, proteins, and metabolites in different biological conditions, e.g. healthy state *vs.* diseased state. This helps researchers investigate a disease. For instance the low expression of a certain gene might in turn lead to the low abundance of a certain protein which might prevent the breakdown of a certain metabolite, the accumulation of which might be causing the disease. Modern experimental equipment produce vast quantities of numerical data. Biological pathways provide an intuitive knowledge-based scaffold for integrating these data. WikiPathways and Reactome are two commonly used pathway databases. Reactome pathways are centrally curated with periodic input by domain experts, while WikiPathways is a community-based platform where pathways are created and continually curated by any interested party. As part of an ongoing collaboration between Reactome and WikiPathways, we have added the Reactome pathways to WikiPathways and made them available from the Reactome portal on WikiPathways. Here, we demonstrate how such an integration is advantageous to both the Reactome and WikiPathways communities and to the general research community at large.

Introduction

Pathway diagrams are a common way to represent a wealth of information about biological molecules, interactions and processes. Currently, the Pathguide collection lists 45 freely available pathway databases with human data, out of which only 14 provide the data in a machine readable format [1, 2]. Even fewer of these provide a pathway diagram that can be used for data visualization and downloaded for further analysis and conversion into other formats (Supplementary File S1). Notable among them are WikiPathways and Reactome, each with its unique user base, contributors, and curation cycle [3].

WikiPathways is an open, collaborative platform for drawing, curating, and sharing biological pathways, built using the same MediaWiki software underlying Wikipedia. WikiPathways leverages community curation to grow and maintain its pathway collection beyond the capabilities of an internal curation team. Anybody can register at WikiPathways to create new pathways and curate existing ones. WikiPathways provides a JavaScript-based viewer for interactively navigating and highlighting pathway elements and a Java based editor for creating and curating pathways. It makes use of BridgeDb web services [4] to provide identifier resolution and links to primary data sources. Pathways can be tagged for classification and quality control, e.g. pathways with the tag "curated" are regularly checked by a dedicated curation team and are deemed suitably annotated for analysis [5]. Pathways can also be tagged with various ontology tags from various pre-existing established ontologies, such as the Pathway ontology [6] and Disease Ontology [7]. Pathways from WikiPathways can be used to integrate, visualize, and analyze system-wide transcriptomics, proteomics, and metabolomics measurements using the open source pathway analysis tool PathVisio [8]. Pathways can also be analyzed as networks in Cytoscape [9], using the WikiPathways app to convert the pathways into networks [10]. WikiPathways pathways are also used by several other tools, such as GO-Elite [11] and SNPLogic [12]. Domain experts often curate specific subsets of pathways in WikiPathways, which are made available in portals e.g. the plant portal [13, 14], CIRM portal [15], exRNA portal [16]. In addition, WikiPathways data is available in RDF (Resource Description Framework) format, which is incorporated into the Open PHACTS Discovery platform, which integrates pharmacological data from a variety of information resources and provides tools and services to question this integrated data to support pharmacological research [17, 18].

Like WikiPathways, Reactome is an open-source, open access pathway database with a substantial collection of diverse pathway models [19, 20]. However, it differs from WikiPathways as the pathway annotations are annotations are curated by the Reactome editorial staff in collaboration with external experts in the research community. Reactome provides an intuitive website to navigate pathway knowledge and a suite of data analysis tools to support the pathway-based analysis of complex experimental and computational data sets. Similar to WikiPathways, visualization of Reactome pathways is facilitated by the Pathway Browser that supports zooming, scrolling, and highlighting, and can show detailed information about entities in the pathway. It makes use of PSICQUIC web services [21] to overlay molecular interaction data from the Reactome Functional Interaction Network [22] and external interaction databases , including IntAct [23], and ChEBI [24]. Pathways in Reactome are explicitly constructed in terms of biochemical reactions and drawn in accordance with the community standard Systems Biology Graphical Notation (SBGN) [25]. Reactome also provides pathway analysis tools which can be used to perform ID mapping, pathway assignment, and over-representation or enrichment analysis with user-supplied datasets.

The integration of Reactome content in WikiPathways provides Reactome with the power of community curation and broader format availability, including the semantic format using the WikiPathways RDF generator [26]. At the same time, WikiPathways benefits from the additional content and curation attention from the Reactome team. A connection between Reactome and WikiPathways was first proposed in 2008, using either the EBI created CSV format and a novel converter, or the BioPAX format and Cytoscape [27]. However, neither of these routes was very successful in preventing loss of data. Therefore, these generic methods of conversion were abandoned for a more specific format conversion. Pathways in WikiPathways are stored using the Graphical

Pathway Markup Language (GPML) format, while pathways in Reactome are stored in a relational database organized by the Reactome data model with their diagrams stored in the database as XML strings with other related information [28, 29]. We created a converter to convert pathways directly from the relational database into the GPML format.

In this manuscript, we describe the newly developed format converter to convert Reactome content for inclusion in WikiPathways. The addition of the Reactome pathways to the analysis collection of pathways available from WikiPathways improves the coverage of gene ontology biological process terms of the analysis collection to 90%. The converted Reactome pathways can be analyzed with several new analysis tools, such as the pathway analysis tool PathVisio and network analysis tool Cytoscape. As a pedagogic example, we perform pathway analysis using a publicly available transcriptomics dataset and the combined collection of pathways from WikiPathways and Reactome.

Results

We developed a Java based format converter to convert Reactome pathways into the WikiPathways format (see Methods). The converter was used to convert the human pathways from Reactome. 431 pathways were converted from version 54 of Reactome ,tagged as “reactome_approved”, and made available from the Reactome portal [30]. The same converter was also used to convert pathways from Plant Reactome. 102 pathways were converted and added to the Plant portal in WikiPathways [14].

Pathway View: WikiPathways vs. Reactome

A pathway in WikiPathways consists of data nodes, interactions, and graphical elements, e.g. cellular compartments. Data nodes can be of the following types: *gene product*, *protein*, *RNA*, *metabolite*, *pathway*, *complex*, and *unknown*. *Gene product* is the default data node that can be used for all products of genes such as transcripts, proteins, RNAs, and genes. By default, these are represented as open rectangular boxes with black labels and borders. The more specific data node types such as *protein* and *RNA* can be used in the specific cases instead of a generic *gene product* node. The *protein* node is visually the same as the *gene product* node while RNAs are represented in purple. The *metabolite* node represents metabolites, drugs, or other small molecules; it is represented in blue. The *pathway* data node is used to denote a connection to another pathway, and represented in green without a border. The *complex* data node represents two types of complexes either a set of proteins represented as a brown rounded rectangle or a set of interacting proteins represented by a brown hexagon. Data nodes of type *unknown* are represented the same as the generic *gene product* node. Interactions describe the relationship between two data nodes. Currently, two collections of interactions are available in the drawing palette: basic interactions and Molecular Interaction Map (MIM) interactions [31]. Arrows can be used to describe basic interactions like conversion, translocation, activation, binding, and modification. T-bars denote inhibition. The MIM interaction palette can be used for more formal and easier machine-readable descriptions of Binding, Conversion, Catalysis, Stimulation and Necessary Stimulation, and Transcription/Translation. Graphical elements can be used to provide contextual meaning to the pathways. Graphical Shapes, lines, and labels can for instance be used to annotate a biological process and generally to make things visually clearer to biologists. Similarly, graphical cellular compartments such as mitochondria, endoplasmic reticulum, nucleus and cell walls can be also added to the pathway as predefined shapes for a richer diagram.

Reactome uses a comparable but graphically slightly different method to describe pathway content. In Reactome, the core unit of the data model is the reaction. Entities (nucleic acids, proteins, complexes, and small molecules) participate in reactions. These reactions form a network of biological interactions and are grouped into pathways. Reactome uses the SBGN Process Description format [17] to draw pathway diagrams. Small Molecules are represented by a green oval, proteins by a green rounded rectangle, and complexes by blue hexagons. A group of entities playing the same roles in a reaction is annotated as EntitySet in Reactome, which is displayed as rounded rectangle with a double line border. Organelles are represented by orange rectangles with double line borders for membranes or single line borders for non-membrane organelles. The following reaction types can be represented: Transition/Process, Association/Binding, Catalysis, Inhibition, Dissociation, Omitted, and Uncertain. Stoichiometry, catalysis, positive and negative regulation, and other types of reaction attributes can also be represented in pathway diagrams based on SGBN. Figure 4.1 shows how the pathway elements from the Reactome pathways were represented in the converted pathway.

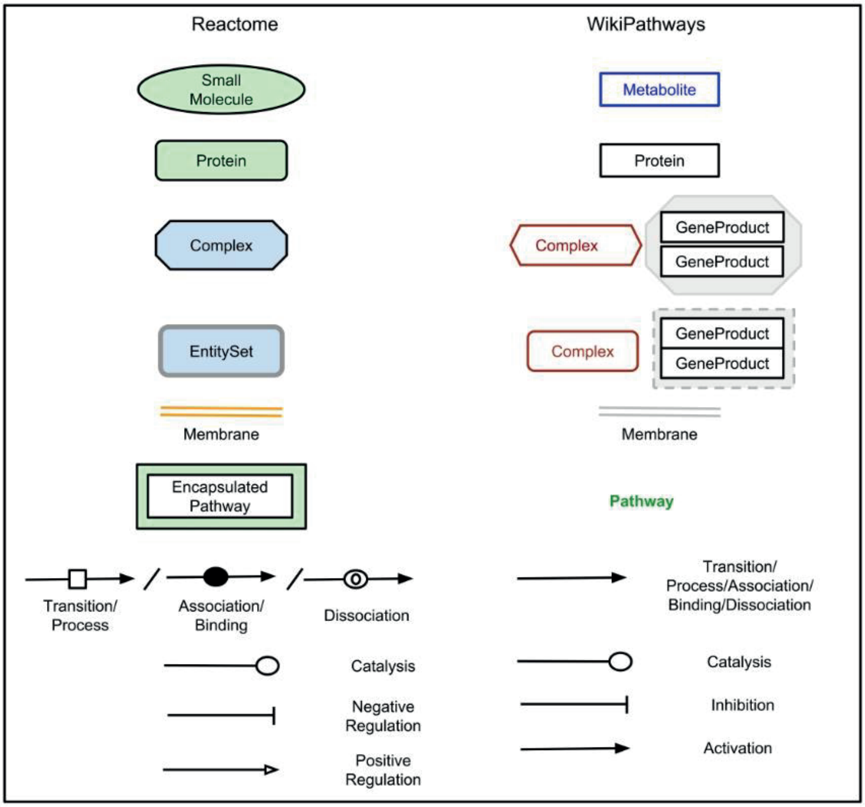


Figure 4.1 Mapping Reactome pathways elements to WikiPathways pathway elements. This diagram shows the symbols used to represent different biological entities in Reactome and the corresponding symbol used to represent the same biological entity in WikiPathways.

Each element of the pathway can be annotated using database identifiers for data analysis and also annotated with literature references. As an example of the conversion, the Abacavir transport and metabolism pathway is shown here (Figure 4.2).

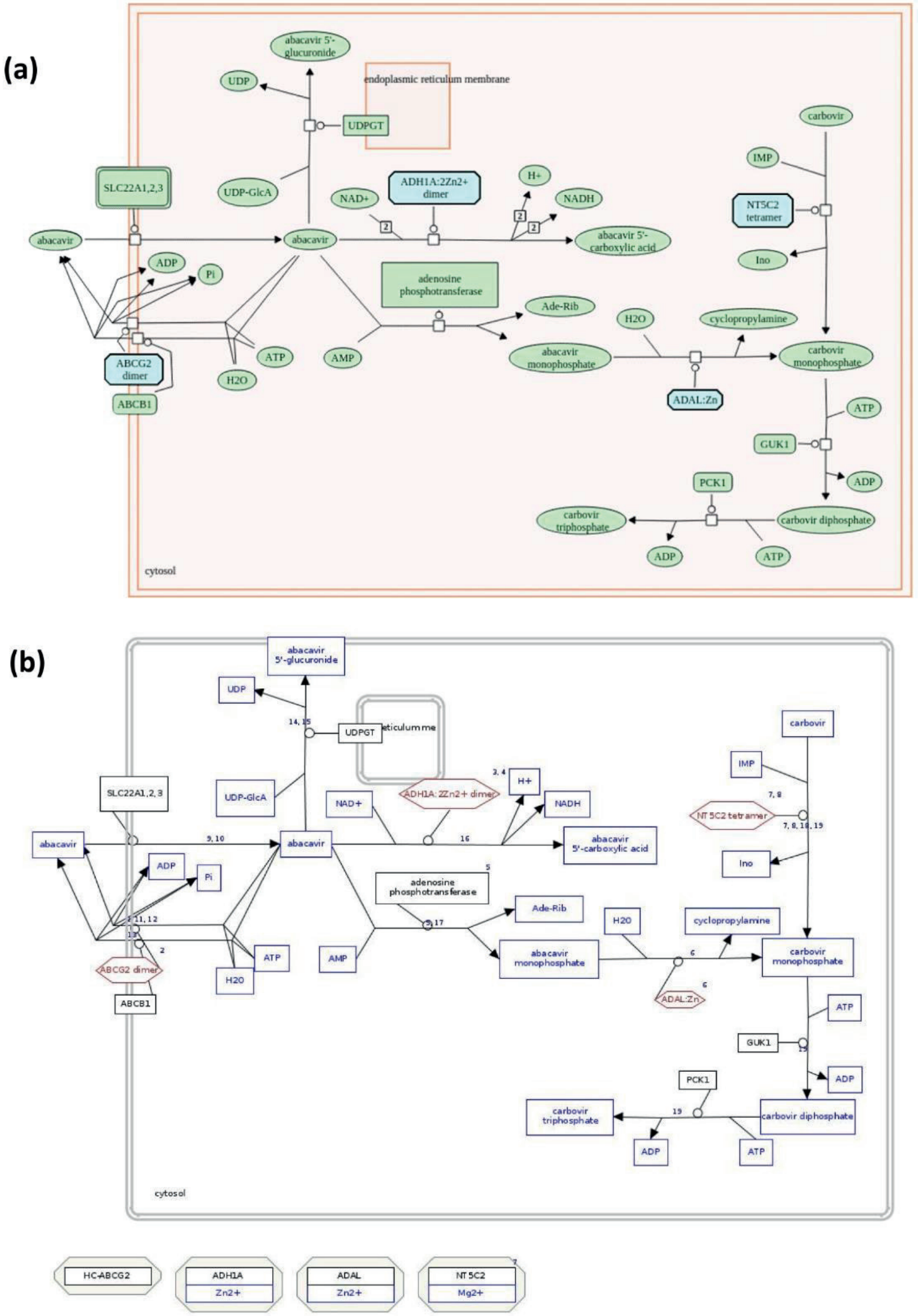


Figure 4.2 A comparative view of the Abacavir transport and metabolism pathway for Homo sapiens from Reactome Database (version 54). (a) Reactome View of Abacavir transport and metabolism (Homo sapiens) [32] and (b) Pathway view on WikiPathways(WP2712_r83598) [33].

In addition to converting the elements of the Reactome pathway diagram, the converter also draws the components of the complexes and entity sets at the bottom of the pathway. This helps with data visualization and gives better results for pathway analysis. Because all the complex and entity set members are also present in the same pathway diagram, they are also taken into consideration by the pathway statistics algorithm for determining the importance of the pathway for the given dataset in the given condition.

The ComplexViz plugin enables the user to highlight the components on the bottom of the pathway belonging to the complex selected in the pathway diagram or vice versa.

Reactome content improves human biological entity coverage in WikiPathways

Coverage of Gene Ontology terms

Gene Ontology (GO) terms provide structured vocabularies for annotating the molecular function, biological process, and cellular location of gene products in a highly systematic way [34]. Here, we analyze the coverage of all GO terms together and the biological process, molecular function, and cellular compartment GO classes separately by the genes and proteins of the curated and reactome_approved collection pathways from WikiPathways.

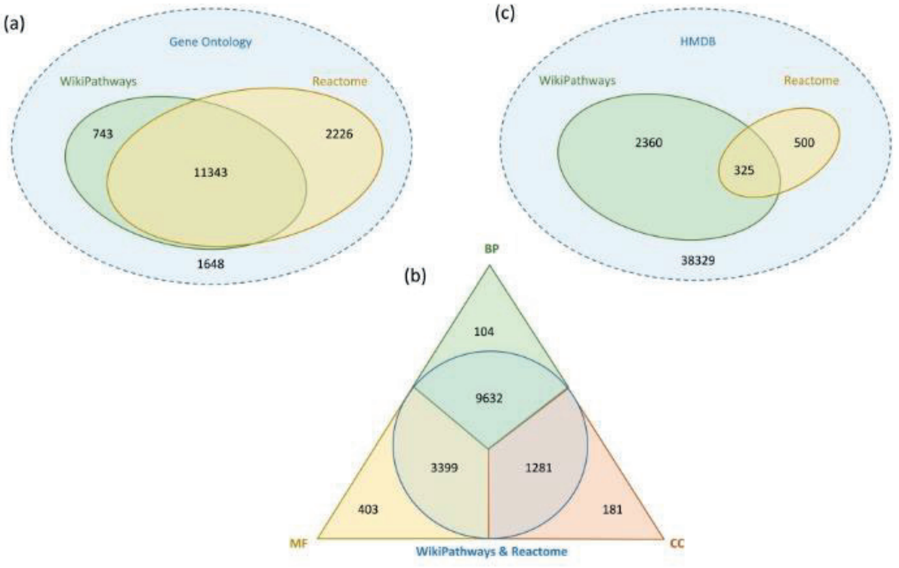


Figure 4.3 Venn Diagrams showing coverage of other external databases by WikiPathways and Reactome. (a) Venn Diagram showing coverage of Gene Ontology Terms by Gene Products of WikiPathways and Reactome, (b) Venn Diagram showing coverage of Biological Process (BP), Molecular Function (MF), and Cellular Compartment (CC) Gene Ontology Terms by Gene Products of curated and reactome_approved collections of WikiPathways pathways, and (c) Venn Diagram showing coverage of the Human Metabolome Database (HMDB) by metabolites curated and reactome_approved collections of WikiPathways pathways.

Out of the 15960 human GO terms 90% are now covered by the combined curated and reactome_approved collection pathways available for analysis from WikiPathways (*i.e.* at least one gene in one of the pathways is annotated with the term). 11343 (71%) GO terms are covered by both collections. The curated collection of WikiPathways includes an additional 5% and the reactome_approved collection an additional 14% to bring the total coverage of GO terms up to 90% (Figure 4.3a). The coverage of gene ontology terms by the two pathway collections is shown in Table 4.1. The coverage of each gene ontology branch, biological process, molecular function, and cellular compartment, by the combined set of pathways from the curated and Reactome approved collections is shown in Figure4.3a. Venn Diagrams showing coverage of each gene ontology branch by curated collection, the reactome approved collection, their overlap, and terms not yet covered is provided in supplementary file S3.

Table 4.1 Summary of the Gene Ontology Term Coverage.

Category	Total terms	Coverage by the different collections			
		Only WikiPathways	Only Reactome	Combined	Overlap
All GO terms	15960	743 (5%)	2226 (14%)	14312 (90%)	11343 (71%)
BP terms	10696	555 (5%)	1191 (11%)	9632 (90%)	7886 (74%)
MF terms	3802	147 (4%)	825 (22%)	3399 (89%)	2427 (64%)
CC terms	1462	41 (3%)	210 (14%)	1281 (88%)	1030 (70%)

GO, Gene Ontology; BP, Biological Process; MF, Molecular Function; CC, Cellular Compartment

Metabolome Coverage

The Human Metabolome Database (HMDB) is currently the most complete and comprehensive curated collection of human metabolites [35]. Here, we analyze the coverage of HMDB by the curated and reactome_approved collection pathways from WikiPathways. (Figure 4.3c). 41515 unique metabolites have been reported in the current version (3.6) of HMDB. 2685 of these metabolites are covered by WikiPathways, of which 325 metabolites are covered by Reactome as well. The inclusion of Reactome pathways contributed 500 new metabolites to the WikiPathways collection.

Plant Pathways converted from Plant Reactome

The Reactome converter was also used to convert plant pathways from the Plant Reactome database freely available at http://plantreactome.gramene.org/. Pathways for the species *Oryza sativa*, *Zea mays*, and *Arabidopsis thaliana* were converted. The pathways for rice are manually curated, the pathways for the other species are computationally inferred from the rice pathways. These pathways have been made available in the plant portal at WikiPathways [14].

WikiPathways infrastructure to analyze Reactome data

Pathway analysis and visualization: The conversion of pathways from Reactome has contributed 431 new manually curated pathways to the WikiPathways analysis set. In addition to the 293 pathways originally available in the curated collection. This combined set of 724 pathways are now available for analysis from WikiPathways, essentially doubling the pathway quantity. To evaluate the effect of this content enrichment on pathway analysis of genomics datasets we performed pathway analysis with the combined set.

Polycystic ovary syndrome (PCOS) is a common heterogeneous endocrine disorder characterized by irregular menses, hyperandrogenism, and polycystic ovaries [36]. Its clinical manifestations may include: menstrual irregularities, signs of androgen excess, and obesity. Insulin resistance and elevated serum Luteinizing Hormone levels are also common features in PCOS. PCOS is associated with an increased risk of type 2 diabetes and cardiovascular events [37]. A study by Kaur *et al* investigates PCOS using granulosa cells of 40 women discordant for PCOS undergoing in vitro fertilization [38]. Granulosa cell gene expression profiling was accomplished using Affymetrix Human Genome-U133 arrays. The samples were analyzed for differences in their transcript profile between PCOS and normal ovulatory women. Here, we obtained the raw data from GEO (GSE34526) and performed quality control and normalization using the Affymetrix quality control and pre-processing module of arrayanalysis.org[39]. All arrays were determined to be of sufficient quality for inclusion in further analysis. The quality control report generated by arrayanalysis.org has been provided (Supplementary File S3). Statistical analysis was also performed in arrayanalysis.org using the statistical analysis module [40]. The gene level statistics have been provided (Supplementary File S4). Over-representation analysis of the gene level statistics was performed in PathVisio using the combined collection of curated and reactome_approved pathways from WikiPathways. The Z score was calculated using the criterion absolute log fold change > 1 and P. value < 0.05. This criterion is commonly used for detecting differentially expressed genes in microarray datasets [41-43].

Table 4.2 shows the top ten pathways obtained through pathway analysis. 2 pathways from WikiPathways and 8 from Reactome show up in the list. The Toll-Like Receptors Cascades pathway (Figure 4.3) from Reactome shows up as the most affected pathways with a Z score of 7.25. Subsequently, the gene statistics were visualized on the pathway. The logFC values were visualized using a color gradient: blue to yellow, corresponding to the value -2 to 2. The P.value was visualized using a color rule, the genes with "P.value < 0.05" were marked in green and the rest red. The ComplexViz plugin was used to score the complexes on the pathway to highlight the complexes of interest. The same criteria "P.value < 0.05" was used to calculate the percentage scores for the complexes. The scores were then visualized on the pathway. Complexes with a percent score higher than 25 were marked in orange and the rest were colored dark grey.

Table 4.2. Table showing the top ten pathways obtained performing over-representation analysis in PathVisio.

Pathway	Z score	Source
Toll-Like Receptors Cascades	7.25	Reactome
Cell surface interactions at the vascular wall	7.09	Reactome
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	7.00	Reactome
IL1 and megakaryotyces in obesity	6.71	WikiPathways
Interferon gamma signalling	6.30	Reactome
Signal regulatory protein (SIRP) family interactions	5.17	Reactome
Costimulation by the CD28 family	4.82	Reactome
IL-4 Signaling Pathway	4.33	WikiPathways
Interleukin-3, 5 and GM-CSF signalling	4.25	Reactome
Syndecan interactions	4.07	Reactome

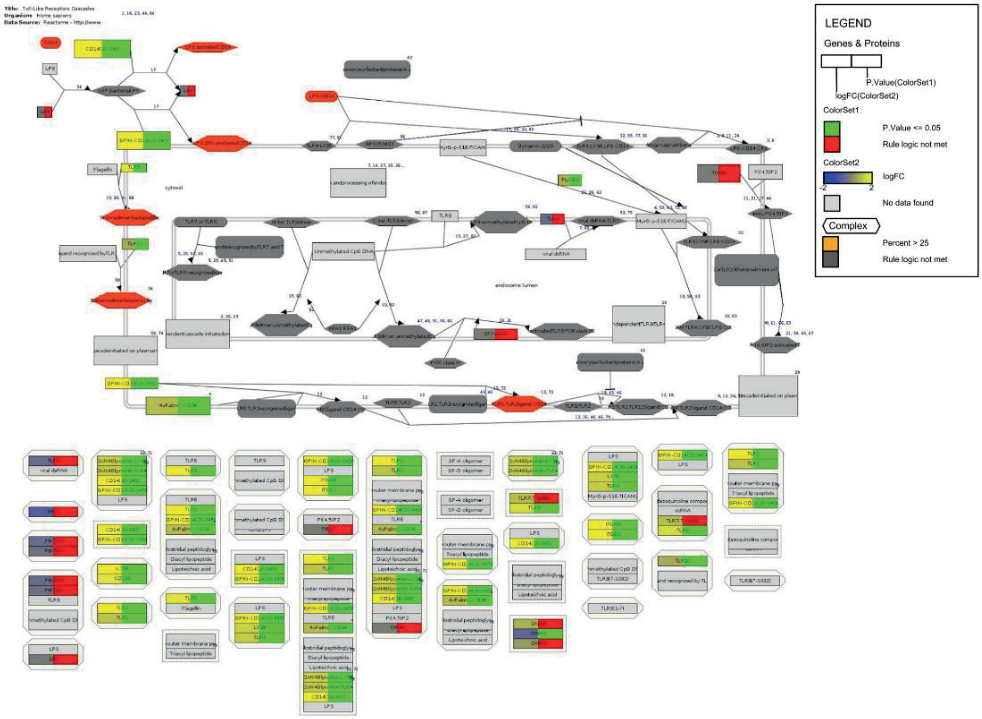


Figure 4.4 Human granulosa cells gene expression in normal ovulatory versus PCOS women visualized on the Toll-Like Receptors Cascades pathway (WP2775_r83597) in PathVisio [44]. Human granulosa cells were isolated from ovarian aspirates of normal ovulatory and PCOS women undergoing IVF. For each sample, RNA was extracted and hybridized to an Affymetrix Gene Chip. Genes not measured appear in gray. The log fold change (logFC) is depicted with a blue to yellow color gradient corresponding to the values -2 to 2. Significant genes with a P.value < 0.05 are marked in green and the rest in red. Significant complexes with a score > 25 are marked in orange and the rest in dark gray.

Network analysis and visualization in Cytoscape

The biological entities in pathways and their relationships can be represented as nodes and edges in abstract biological networks. This opens up a large variety of network analysis methods to further extend, analyze and visualize biological pathways.

The incorporation of the WikiPathways and ReactomeFIViz apps in the Cytoscape framework allows further investigation of biological pathways using a wide variety of Cytoscape apps for network analysis and visualization. The visualization of an example Reactome pathway with both apps is provided (Supplementary file S5).

Discussion

The WikiPathways project has developed a suite of pathway visualization and editing tools for users to view and edit pathways, and established a dynamic community to continuously crowd source updates and novel pathway content. The contents in Reactome are created by select domain experts in target fields of research with Reactome editorial staff. Including Reactome content has significantly expanded the coverage of pathway information at WikiPathways. Likewise, incorporating community edits from the WikiPathways versions of Reactome content significantly expands their pool of contributors, helping them produce more frequent updates and create links to outside databases. In the current implementation, we use a notification mechanism developed in WikiPathways to send edits from WikiPathways to Reactome. However, such an approach cannot be scaled up if many edits occur in the WikiPathways web site. We plan to develop a robust round-trip software approach in the Reactome curator tool so that edits in WikiPathways for Reactome pathways can be imported into the Reactome database easily. Such a tool will find new edits, and then present them to Reactome curators in graphical user interfaces so that curators can decide whether or not these edits should be committed into the Reactome database. We believe a true round-trip approach between Reactome and WikiPathways will benefit both projects, and set an example for other projects to collaborate with each other.

The conversion of Reactome pathways to the GPML format enables the analysis and data visualization of Reactome pathways in PathVisio. PathVisio is a widely used pathway analysis software, preferred due to its excellent data visualization capabilities as demonstrated by its use in numerous academic publications [45-49]. PathVisio allows multiple data points to be visualized on one node using colors and color gradients permitting easy visualization of time series data. The data visualized images can then be exported as images for further publication or in html format as a mini-website to easily maneuver the uploaded data on the pathway image. The new ComplexViz plugin simplifies analysis of the converted Reactome pathways. As Reactome pathways typically contain numerous complexes, the plugin enables highlighting complex components on the bottom of the pathway diagram and the other way around. It also enables browsing complex components in a side panel and visualizing data uploaded for the complex components on the parent complex node. This highlights the complexes of interest, which can then be further studied. The complex component diagram on the side panel also displays the data uploaded thereby making it simpler to look at them without having to look for them on the bottom of the pathway.

The pathway analysis case study presented here with a transcriptomics dataset comparing women with normal ovulatory physiology with those with PCOS shows that addition of the Reactome pathway set clearly improves pathway analysis results. The list of top ten most affected pathways feature pathways from both the curated and reactome_approved collection of pathways from WikiPathways. More pathways appear from the reactome_approved collection, which is expected since the collection is manually curated. The reactome_approved collection adds 4417 new gene products and 500 new metabolites. However, the curated collection still contains 1414 unique gene products and 2360 unique metabolites. There are 3438 gene products and 325 metabolites in common between the two collections. Therefore, the conversion adds content without much overlap.

The toll like receptor (TLR) cascades pathway, which shows up as the most changed pathway in this condition is from the Reactome collection. TLRs are an important family of pattern recognition receptors (PRR) involved in innate immunity. The innate immune system initiates an inflammatory response after recognizing pathogens by PRRs [50]. Emerging evidence suggests that PCOS is associated with systemic inflammation [51, 52]. Furthermore, various studies have reported that TLRs are expressed in the female reproductive tract [53]. Therefore, this pathway is clearly interesting for PCOS. In addition, the Cell surface interactions at the vascular wall pathway, which is the second most highly affected pathway is also from the Reactome collection. This pathway is annotated with the Gene Ontology biology process term, leukocyte migration. Since PCOS is associated with elevated levels of circulating leukocytes [54], this pathway is clearly of interest. Both pathways, are from the newly converted collection of pathways from Reactome and clearly add biological knowledge, as illustrated with the case study for PCOS.

The availability of Reactome pathways in WikiPathways allows the analysis of the pathways with several new analysis tools. Besides the analysis of pathways in PathVisio, users can also use the WikiPathways app for Cytoscape to analyze the pathways as biological networks. While the ReactomeFIViz app focuses on functional gene interaction networks, the WikiPathways app creates a representative network of the pathway including metabolites and other pathway elements. Consequently the created network provides a new analysis tool that allows the integrated analysis of different omics datasets.

Methods

Reactome Converter

A Java based format converter has been developed to convert pathways from the Reactome database to the WikiPathways format. Pathways in Reactome are stored in a relational database with their diagrams encoded in XML strings, while pathways in WikiPathways are stored as GPML, which is an XML based file format. In addition, both repositories have Java based APIs according to which the pathway files can be read and written. These internal data models of the two databases are used to read and write the pathways obtained from them. This allows the converter to remain flexible and backwards compatible as long as the data models themselves are. This also makes the converter stable through version updates of pathways as long the pathways are organized according to the same model. The conversion is done in the following steps: (i) Creating a GPML pathway and adding pathway attributes, (ii) Converting the pathway elements, and (iii) Annotating the pathway and pathway elements. These steps are described further below. The converter is open source and the code is available from the GitHub repository [55].

Step (i): Creating a GPML pathway

A Reactome pathway combined with its rendering information is read from the database via the Reactome Java API [56]. A new GPML pathway is created by instantiating the pathway class and pertinent information is added to the GPML. This information consists of the data source (Reactome), the Reactome version, the organism for which the pathway has been drawn, e.g. Homo sapiens. Biologists who have drawn the pathway are added as authors, the Reactome team members who have edited the pathway are added as maintainers, along with their email addresses.

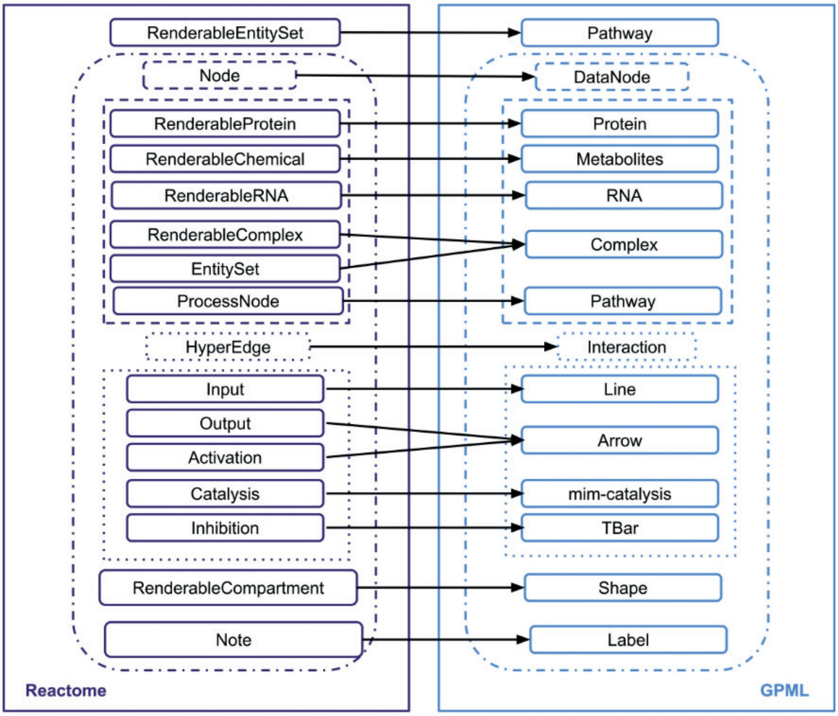


Figure 4.5 Conversion of Reactome Java data classes to corresponding WikiPathways Java data classes

Step (ii): Converting the pathway elements

Each entity in the Reactome pathway is converted to the corresponding WikiPathways element. Figure 4.5 presents an overview of mapping of the elements of a Reactome pathway to the corresponding WikiPathways pathway elements. Nodes are converted to DataNodes. Individual node types, such as Proteins, Small molecules, RNAs, Process Nodes are converted to the corresponding WikiPathways elements, namely Protein, RNA, Metabolite, and Pathway. Complexes and entity sets in Reactome are converted to Group in WikiPathways, with the styles "complex" and "group" respectively. The components of the complexes and entity sets are obtained and these are added to the bottom of the pathway diagram. Duplicates are not displayed on the pathway diagram for keeping the pathway diagram concise but are maintained in the GPML and showed in the "Properties" side panel of PathVisio. Compartments from the Reactome Pathway are converted into a group in WikiPathways. Notes in the Reactome pathway are converted to Labels. In Reactome, the reactions known as hyper edges are modeled such that there is a backbone reaction to which the inputs, outputs, catalysts, activators, and inhibitors are connected (Figure 4.6).

Each branch of the hyper edge (inputs, outputs, catalysts, inhibitors, activators) is converted into a GPML interaction and connected to a backbone interaction using anchors; this achieves the same SBGN compliant reaction view for all substrates, products, enzymes, activators, and inhibitors in GPML.

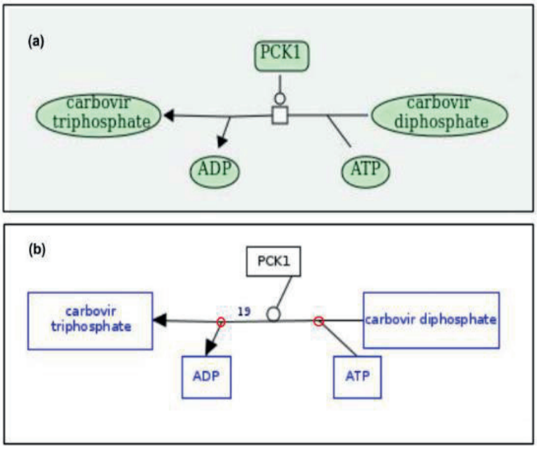


Figure 4.6 A comparative view of hyperedges in a Reactome pathways and how they are converted in WikiPathways. A hyperedge from the Abacavir transport and metabolism pathway is shown. (a)Reactome View (b) WikiPathways view, the anchors are highlighted.

Step (iii): Annotating the pathway and pathway elements.

Subsequently, the Reactome pathway object is mined for annotations for the different elements. Preferably, the proteins are annotated with UniProt identifiers and the metabolites with ChEBI identifiers, in absence of the preferred annotation Reactome identifiers are used. Interactions, Complexes, and Pathways are annotated with Reactome identifiers. All pathway elements are also annotated with literature references using PubMed identifiers.

Calculating coverage of biological entities

Human GO Terms were downloaded from UniProt-GOA [57]. Scripts in Java were written to parse the document to obtain the GO identifiers and identifiers of the terms for the three structured ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions. The current release version 3.6 of HMDB was downloaded to obtain the superset of all human metabolites. Additional scripts in Java were written to map all gene products in the two pathway collections to Ensembl and all metabolites to HMDB. All the scripts used are available from GitHub [58]. The Ensembl gene identifiers were mapped to GO terms using Ensembl BioMart [59], to obtain the total GO term coverage of the two pathway collections and also individual coverage of each GO category. The R package gplots was then used to create Venn diagrams showing GO and HMDB coverage [60]. The Venn diagrams were manually updated in PowerPoint.

ComplexViz Plugin

The newly developed plugin improves visualization of data on complexes and their components. The plugin can be installed in PathVisio using the plugin manager and adds a side panel “Components”. The top half of this panel displays the components of the complex that is clicked as a mini pathway diagram. Imported data is visualized both on the main pathway diagram and on the “Components” side panel containing the complex component diagram. Clicking on the buttons in the side panel next to the mini pathway diagram, displays the cross-references and expression data available for that data node on the bottom half of the panel. The plugin also adds the submenu item “Complex Visualization” to the Data menu. Clicking it opens a dialog box for setting visualization options for complexes. Three visualization options have been implemented. These methods allow changing the border color of complex and components, coloring complex nodes according to a calculated ratio, and drawing the complex label. Users can select a border color for complexes and their components to indicate which complex and components belong together. Complexes can be colored based on the percentage of complex components that qualify the user defined criterion. This percentage is calculated for all complexes on the pathway. Color gradients or rules can be used to visualize the score on the complexes. Text labels can be drawn on the Complexes after data has been visualized, the font and size of text of the label can be changed. The plugin is open source and the code is available from the GitHub repository [61]. A detailed user guide is provided (Supplementary File S6). An up-to-date copy will be maintained at the GitHub wiki [62].

Supplementary Files

Supplementary files are available from:
http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004941#sec019

References

1. Bader, G.D., M.P. Cary, and C. Sander, **Pathguide: a pathway resource list**. *Nucleic acids research*, 2006. **34**(suppl 1): p. D504-D506.
2. **PathGuide the pathway resource list**. 18-07-2015]; Available from: http://pathguide.org/.
3. Bauer‐Mehren, A., L.I. Furlong, and F. Sanz, **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Molecular systems biology*, 2009. **5**(1): p. 290.
4. van Iersel, M.P., A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo, **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC bioinformatics*, 2010. **11**(1): p. 5.
5. **WikiPathways Curation Protocol**. Available from: http://wikipathways.org/index.php/Help:Curation_Protocol.
6. Petri, V., P. Jayaraman, M. Tutaj, G.T. Hayman, J.R. Smith, J. De Pons, S.J. Laulederkind, T. Lowry, R. Nigam, and S.-J. Wang, **The pathway ontology-updates and applications**. *J. Biomedical Semantics*, 2014. **5**: p. 7.
7. Schriml, L.M., C. Arze, S. Nadendla, Y.-W.W. Chang, M. Mazaitis, V. Felix, G. Feng, and W.A. Kibbe, **Disease Ontology: a backbone for disease semantic integration**. *Nucleic acids research*, 2012. **40**(D1): p. D940-D946.
8. Kutmon, M., M.P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A.R. Pico, and C.T. Evelo, **PathVisio 3: an extendable pathway analysis toolbox**. *PLoS computational biology*, 2015. **11**(2).
9. Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome research*, 2003. **13**(11): p. 2498-2504.
10. Kutmon, M., S. Lotia, C.T. Evelo, and A.R. Pico, **WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization**. *F1000Research*, 2014. **3**.
11. Zambon, A.C., S. Gaj, I. Ho, K. Hanspers, K. Vranizan, C.T. Evelo, B.R. Conklin, A.R. Pico, and N. Salomonis, **GO-Elite: a flexible solution for pathway and ontology over-representation**. *Bioinformatics*, 2012. **28**(16): p. 2209-2210.
12. Pico, A.R., I.V. Smirnov, J.S. Chang, R.-F. Yeh, J.L. Wiemels, J.K. Wiencke, T. Tihan, B.R. Conklin, and M. Wrensch, **SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system**. *Nucleic acids research*, 2009. **37**(suppl 1): p. D803-D809.
13. Hanumappa, M., J. Preece, J. Elser, D. Nemeth, G. Bono, K. Wu, and P. Jaiswal, **WikiPathways for plants: a community pathway curation portal and a case study in rice and arabidopsis seed development networks**. *Rice (N. Y)*, 2013. **6**: p. 14.
14. **WikiPathways plant portal**. Available from: http://wikipathways.org/index.php/Portal:Plants.
15. **CIRM Stem Cell Pathways**. Available from: http://www.wikipathways.org/index.php/Portal:CIRM.
16. **extracellular RNA research community**. Available from: http://www.wikipathways.org/index.php/Portal:ExRNA.
17. Waagmeester, A., H. Deus, and C.T. Evelo, **Exposing WikiPathways as Linked Open Data**. 2011.
18. **WikiPathways Sparql queries**. Available from: http://www.wikipathways.org/index.php/Help:WikiPathways_Sparql_queries.
19. Croft, D., **Building models using Reactome pathways as templates**, in *In Silico Systems Biology*. 2013, Springer. p. 273-283.
20. Milacic, M., R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, and L. Stein, **Annotating cancer variants and anti-cancer therapeutics in reactome**. *Cancers*, 2012. **4**(4): p. 1180-1211.
21. Aranda, B., H. Blankenburg, S. Kerrien, F.S. Brinkman, A. Ceol, E. Chautard, J.M. Dana, J. De Las Rivas, M. Dumousseau, and E. Galeota, **PSICQUIC and PSISCORE: accessing and scoring molecular interactions**. *Nature methods*, 2011. **8**(7): p. 528-529.
22. Wu, G., X. Feng, and L. Stein, **Research a human functional protein interaction network and its application to cancer data analysis**. *Genome Biol*, 2010. **11**: p. R53.
23. Orchard, S., M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N.H. Campbell, G. Chavali, C. Chen, and N. Del-Toro, **The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases**. *Nucleic acids research*, 2013: p. gkt1115.
24. Hastings, J., P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, and M. Williams, **The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013**. *Nucleic acids research*, 2013. **41**(D1): p. D456-D463.
25. Le Novere, N., M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M.I. Aladjem, and S.M. Wimalaratne, **The systems biology graphical notation**. *Nature biotechnology*, 2009. **27**(8): p. 735-741.
26. **WikiPathways RDF**. Available from: http://wikipathways.org/index.php/Help:WikiPathways_RDF.
27. Adriaens, M.E., M. Jaillard, A. Waagmeester, S.L. Coort, A.R. Pico, and C.T. Evelo, **The public road to high-quality curated biological pathways**. *Drug discovery today*, 2008. **13**(19): p. 856-862.
28. **GPML Description**. Available from: http://www.pathvisio.org/gpml/.
29. **Reactome Data Model**. Available from: http://www.reactome.org/pages/documentation/data-model/.
30. **Portal:Reactome - WikiPathways** 2015 [cited 2015 02-09-2015]; Available from: http://wikipathways.org/index.php/Portal:Reactome.
31. Kohn, K.W., M.I. Aladjem, J.N. Weinstein, and Y. Pommier, **Molecular interaction maps of bioregulatory networks: a general rubric for systems biology**. *Molecular biology of the cell*, 2006. **17**(1): p. 1-13.
32. D'Eustachio, P. **Abacavir transport and metabolism [Homo sapiens]**.
33. Reactome Team, A. Bohler, and **Abacavir transport and metabolism (Homo sapiens)**. Available from: http://wikipathways.org/index.php/Pathway:WP2712.
34. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, and J.T. Eppig, **Gene Ontology: tool for the unification of biology**. *Nature genetics*, 2000. **25**(1): p. 25-29.
35. Wishart, D.S., D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, and S. Sawhney, **HMDB: the human metabolome database**. *Nucleic acids research*, 2007. **35**(suppl 1): p. D521-D526.
36. Sirmans, S.M. and K.A. Pate, **Epidemiology, diagnosis, and management of polycystic ovary syndrome**. *Clinical epidemiology*, 2014. **6**: p. 1.
37. Rotterdam, E. and P. ASRM-Sponsored, **Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS)**. *Human Reproduction (Oxford, England)*, 2004. **19**(1): p. 41.

38. Kaur, S., K.J. Archer, M.G. Devi, A. Kriplani, J.F. Strauss III, and R. Singh, **Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis**. *The Journal of Clinical Endocrinology & Metabolism*, 2012.

39. Eijssen, L.M., M. Jaillard, M.E. Adriaens, S. Gaj, P.J. de Groot, M. Müller, and C.T. Evelo, **User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis. org**. *Nucleic acids research*, 2013. **41**(W1): p. W71-W76.

40. Dutta, A., **Adding automated Statistical Analysis and Biological Evaluation modules to www. arrayanalysis. org**. 2011, Maastricht University.

41. Zhu, H., Q. Wang, Y. Yao, J. Fang, F. Sun, Y. Ni, Y. Shen, H. Wang, and S. Shao, **Microarray analysis of Long non-coding RNA expression profiles in human gastric cells and tissues with Helicobacter pylori Infection**. *BMC medical genomics*, 2015. **8**(1): p. 1.

42. Shim, U., H.-N. Kim, H. Lee, J.-Y. Oh, Y.-A. Sung, and H.-L. Kim, **Pathway Analysis Based on a Genome-Wide Association Study of Polycystic Ovary Syndrome**. *PloS one*, 2015. **10**(8): p. e0136609.

43. Chang, Y.-H., C.-M. Chen, H.-Y. Chen, and P.-C. Yang, **Pathway-based gene signatures predicting clinical outcome of lung adenocarcinoma**. *Scientific reports*, 2015. **5**.

44. Reactome Team, A. Bohler, and E. Willighagen. **Toll-Like Receptors Cascades (Homo sapiens)**. Available from: http://wikipathways.org/index.php/Pathway:WP2775.

45. Tisoncik, J.R., M.J. Korth, C.P. Simmons, J. Farrar, T.R. Martin, and M.G. Katze, **Into the eye of the cytokine storm**. *Microbiology and Molecular Biology Reviews*, 2012. **76**(1): p. 16-32.

46. Kursawe, R., M. Eszlinger, D. Narayan, T. Liu, M. Bazuine, A.M. Cali, E. D'Adamo, M. Shaw, B. Pierpont, and G.I. Shulman, **Cellularity and adipogenic profile of the abdominal subcutaneous adipose tissue from obese adolescents: association with insulin resistance and hepatic steatosis**. *Diabetes*, 2010. **59**(9): p. 2288-2296.

47. Jitendra, S., A. Nanda, S. Kaur, and M. Singh, **A comprehensive molecular interaction map for Hepatitis B virus and drug designing of a novel inhibitor for Hepatitis BX protein**. *Bioinformation*, 2011. **7**(1): p. 9.

48. Zhou, C., Q. Zhong, L.V. Rhodes, I. Townley, M.R. Bratton, Q. Zhang, E.C. Martin, S. Elliott, B.M. Collins-Burow, and M.E. Burow, **Proteomic analysis of acquired tamoxifen resistance in MCF-7 cells reveals expression signatures associated with enhanced migration**. *Breast Cancer Res*, 2012. **14**(2): p. R45.

49. Rubio-Aliaga, I., B. de Roos, M. Sailer, G.A. McLoughlin, M.V. Boekschoten, M. van Erk, E.-M. Bachmair, E.M. Van Schothorst, J. Keijer, and S.L. Coort, **Alterations in hepatic one-carbon metabolism and related pathways following a high-fat dietary intervention**. *Physiological genomics*, 2011. **43**(8): p. 408-416.

50. Nasu, K. and H. Narahara, **Pattern recognition via the toll-like receptor system in the human female genital tract**. *Mediators of inflammation*, 2010. **2010**.

51. Rojas, J., M. Chávez, L. Olivar, M. Rojas, J. Morillo, J. Mejías, M. Calvo, and V. Bermúdez, **Polycystic ovary syndrome, insulin resistance, and obesity: navigating the pathophysiologic labyrinth**. *International journal of reproductive medicine*, 2014. **2014**.

52. Duleba, A.J. and A. Dokras, **Is PCOS an inflammatory process?** *Fertility and sterility*, 2012. **97**(1): p. 7-12.

53. Aflatoonian, R. and A. Fazeli, **Toll-like receptors in female reproductive tract and their menstrual cycle dependent expression**. *Journal of reproductive immunology*, 2008. **77**(1): p. 7-13.

54. Covington, J.D., C.S. Tam, M. Pasarica, and L.M. Redman, **Higher circulating leukocytes in women with PCOS is reversed by aerobic exercise**. *Biochimie*, 2014.

55. Bohler, A., G. Wu, and L.A. Pradhana. **Reactome converter source code**. Available from: https://github.com/wikipathways/reactome2gpml-converter.

56. The Developement team of Reactome. **Reactome Curator Tool**. 2015; Available from: https://github.com/reactome/CuratorTool.

57. Barrell, D., E. Dimmer, R.P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, **The GOA database in 2009—an integrated Gene Ontology Annotation resource**. *Nucleic acids research*, 2009. **37**(suppl 1): p. D396-D403.

58. **Java Scripts** Available from: https://github.com/pennatula/Utilities.

59. Cunningham, F., M.R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, and S. Fitzgerald, **Ensembl 2015**. *Nucleic acids research*, 2015. **43**(D1): p. D662-D669.

60. Warnes, G.R., B. Bolker, L. Bonebakker, R. Gentleman, W. Huber, A. Liaw, T. Lumley, M. Maechler, A. Magnusson, and S. Moeller, **gplots: Various R programming tools for plotting data**. *R package version*, 2009. **2**(4).

61. Bohler, A. **ComplexViz Plugin source code**. 2015; Available from: , https://github.com/pennatula/ComplexViz.

62. Bohler, A. **User Guide for the ComplexViz Plugin**. 2015; Available from: https://github.com/pennatula/ComplexViz/wiki/User-Guide-for-the-ComplexViz-Plugin.

CHAPTER 5

A toolset for flux data integration in pathway analysis

Anwesha Bohler *[1], Susan Coort [1], Sacha Bohler [2], Jonathan Melius [1], Rhizhou Guo [3], Sri Harsha Pamu [4], Martina Kutmon [1,5], Chris Evelo [1,5]*

1. Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands
2. Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands
3. Department of Biomedical Engineering - CBio Group, Technische Universiteit Eindhoven, Eindhoven, The Netherlands
4. College of Technology, Purdue University, Indiana, United States of America
5. Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands

Manuscript under review in **Oxford Database**.

Abstract

Several online databases storing information about biomolecular interactions are available. Each database has its own unique identifier. This results in multiple identifiers from different databases that refer to the same interaction. Therefore, we developed an identifier mapping database for interactions using the open source BridgeDb framework and the identifier mappings from EBI-Rhea. This new mapping database allows users to annotate and map data onto interactions in any software which uses the BridgeDb framework for identifier mapping (e.g. WikiPathways, PathVisio, and Cytoscape). It is accessible programmatically through the BridgeDb web sernvices for identifier mapping and in the R environment for analysis in Bioconductor/R via the BridgeDbR package. It has been incorporated into WikiPathways to provide cross references for annotated interactions. In PathVisio, the mapping database maps the experimental data onto pathways irrespective of the identifiers used in the data file. Three additional tools have been developed as plugins for PathVisio. PathSBML for importing metabolic models encoded in SBML stored locally or directly from the BioModels database, FYI for simplifying interaction annotation by automatically querying the Rhea database and suitable identifiers, and IntViz for visualizing single or multiple data points on interactions by changing the colour or thickness of the representative line. The newly developed toolset enables interaction annotation, linking interactions through identifiers to online data sources for more information, and visualizing data on interactions.

Database URL: http://projects.bigcat.unimaas.nl/interactiondatatoolset

Introduction

Pathways are critical for a systems level understanding of how the functions of individual genes and proteins contribute to normal physiology and pathophysiology [1]. Pathway analysis has thus become an analysis of choice for integrating different "-omics" levels of experimental measurements, such as transcriptomics, proteomics, and metabolomics, for a holistic understanding of biological processes [2] . These multi-omics measurements are visualised on genes, proteins, and metabolites, i.e. nodes of a biological pathway diagram, which lends the experimental measurements a biological context. Using colours to visualise these data ease biological interpretation for combinatorial analyses.

As the functional behaviour of genes, proteins, and metabolites emerge only through interactions across multiple metabolic and regulatory layers , many mechanisms of complex biological processes cannot be understood from monitoring the components alone [3]. Interactions in pathway diagrams represent the relationship between the connected nodes (e.g. activation, inhibition, catalysis). The molecular reaction rates (fluxes) are the ultimate outcome of the activities of different biomolecules (DNA, gene products and small molecules) can be quantified by experimental techniques such as 13C flux analysis [4]. Fluxes can also be simulated by mathematical models. Flux balance analysis is a widely used mathematical modelling approach for studying genome scale reconstructions of biochemical networks [5]. Fluxes can then be visualised along with the other genomics data on the pathway diagrams themselves.

A number of pathway analysis tools are developed by the scientific community to analyse and visualise genomics data such as PathVisio [6], Cytoscape [7] , Biological Network Analyzer (BINA) [8], VANTED [9] , and Pathway Tools [10]. Some commercial tools also exist in this field, e.g. QIAGEN's Ingenuity Pathway Analysis (www.qiagen.com/ingenuity), GeneGo MetaCore from Thomson Reuters (http://www.genego.com/metacore.php) , Omix [11], Ariadne Pathway Studio [12], and Advaita Bio's iPathwayGuide (http://www.advaitabio.com/ipathwayguide). PathVisio is an open source pathway analysis tool widely adopted by the scientific community for data visualisation on pathways [13-17]. The pathway analysis and visualization capabilities of PathVisio are extensive and because of the incorporated identifier mapping simplifies data integration. Pathways drawn in PathVisio can be shared on WikiPathways for community curation [18]. Conversely, pathways from WikiPathways can be used in PathVisio for data analysis and visualisation. PathVisio uses the BridgeDb framework for mapping experimental data to the elements on the pathways [19].

Identifier mapping databases for genes, proteins, and metabolites are already available allowing analysis and visualisation of transcriptomics, proteomics, and metabolomics data. For enabling the mapping of data on the interactions in the pathway, we developed an identifier mapping database using mappings provided by the Rhea database [20]. Rhea identifiers are mapped manually to IntEnz (E.C numbers) [21, 22], MetaCyc reactions [23], KEGG reactions [24], UniPathway enzymatic and chemical reactions [25] and automatically to MACiE (Mechanism, Annotation and Classification in Enzymes) [26], Reactome , and UniProt [27].

Additionally, we developed three plugins for PathVisio, PathSBML, FYI, and IntViz. PathSBML imports and creates pathway diagrams for metabolic models saved in SBML (Systems Biology Markup Language). This allows visualization of modelled and measured data on models using PathVisio functionalities. The FYI (FindYourInteractions) plugin queries the web service of the Rhea database for suggesting suitable identifiers for interactions, facilitating and speeding up the proper annotation of interactions. The IntViz plugin provides visualisation styles for visualizing data on interactions, thereby adding another level of data integration for multi-omics studies. Multiple data points can be visualised together as a bar chart or dynamically using a slider.

In this manuscript, we describe the newly developed toolset for mapping and visualizing data on interactions. We also demonstrate its use in a case study combining a publicly available transcriptomics dataset with metabolic fluxes derived from a genome scale mathematical model for the model organism *A. thaliana*.

Interaction identifier mapping database

BridgeDb is an open source identifier mapping framework for bioinformatics applications. BridgeDb needs actual mapping databases or link services to function. A Java script is developed to retrieve the mappings from the Rhea database and create an interaction identifier mapping database based on the BridgeDb framework. The Rhea Identifier for each interaction is assigned as its primary identifier and the identifiers from other databases are mapped to this primary identifier and saved in the derby database. The direction of the reaction is added as an attribute. The code is available from GitHub (https://github.com/BiGCAT-UM/InteractionDB).

The current mapping database for interactions provides mapping for 8693 unique interactions connecting 7554 unique compounds (http://www.rhea-db.org/statistics). It can be freely downloaded from http://bridgedb.org/data/gene_database/ and is regularly updated following Rhea releases.

Programmatic access

BridgeDb REST web services provide programmatic access to the identifier mapping database (http://developers.bridgedb.org/wiki/BridgeWebservice). Scripts in any programming language can be written to use the web services. In R, a language and environment for statistical computing and graphics, the database can be used for identifier mapping with the BridgeDbR package (http://www.bioconductor.org/packages/release/bioc/html/BridgeDbR.html)[28].

Annotating interactions in WikiPathways and PathVisio

In WikiPathways and PathVisio interactions in pathway diagrams can be annotated using the annotation tab of the properties dialog box which is shown on double clicking. All interactions in the pathways of the "reactome_approved" collection of WikiPathways, automatically converted from Reactome [29, 30], are already annotated with Reactome identifiers (http://wikipathways.org/index.php/Portal:Reactome). There are 724 pathways available from WikiPathways for analysis including the "curated" and the "reactome_approved" collections (http://wikipathways.org/index.php/Download_Pathways). They contain 46222 interactions, of which 81% are annotated.

An extension to simplify interaction annotation in PathVisio

The FYI plugin (http://www.pathvisio.org/plugin/findyourinteraction-plugin/) for PathVisio uses the Rhea web service to find potential identifiers for the interactions (http://www.rhea-db.org/webservice). ChEBI/UniProt identifiers or text labels of the two nodes that the selected interaction connects are used in the query. The results can be selected from a clickable table that is then displayed. If no hits are found with both partners, hits with only one of the partners are shown. A Tutorial is available online at http://plugins.pathvisio.org/fyi/tutorial/.

An extension to import SBML models as pathways in PathVisio

The PathSBML plugin (http://www.pathvisio.org/plugin/pathsbml/) for PathVisio enables the import of metabolic models encoded in SBML [31]. The SBML files are converted to SBGN-PD [32] based pathway diagrams on which the model simulation or any other data can be visualised. The plugin provides direct access to the BioModels models database, a free public repository of biochemical models [33]. Models can be browsed or searched using model name, PubMed ID of publication etc. The plugin also enables checking SBML files for errors with the validate function. In addition, it provides an option to apply force directed layout to the pathway diagram. Tutorials are available online at http://plugins.pathvisio.org/pathsbml/tutorials/.

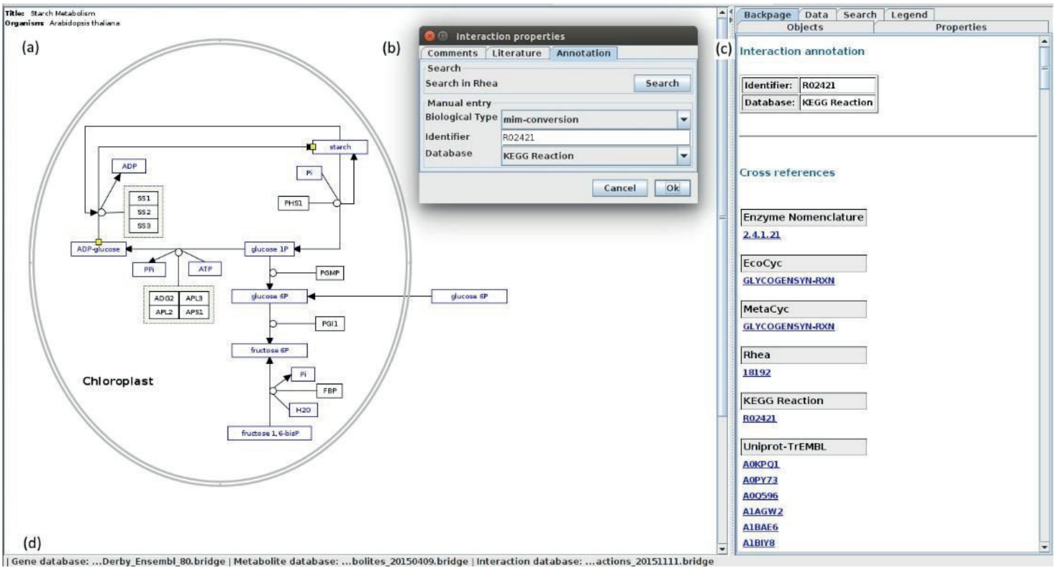
An extension to visualise flux data on pathways using PathVisio

The IntViz plugin (http://www.pathvisio.org/plugin/intviz/) for PathVisio enables the visualisation of single or multiple columns of data on the interactions of a pathway. Three visualisation styles have been implemented: (i) visualizing data on the interactions of the pathway using colours or colour gradients, (ii) dynamically visualizing time series data on interactions using a slider, and (iii) visualizing multiple columns of data on the interactions using a bar chart. Tutorials are available online at http://plugins.pathvisio.org/intviz/tutorials/ .

Case study: Studying photosynthesis in *A. thaliana* by combining modelled metabolic fluxes with gene expression data

A. thaliana is a model organism in plant science, therefore many experimental data sets of different omics levels and metabolic models are already available. We drew pathways in PathVisio describing glycolysis (WP2621), starch metabolism (WP2622), and sucrose metabolism (WP2623) and shared them on WikiPathways. These three pathways are chosen because they represent the three paths in the primary metabolism that carbon fixated during photosynthesis can take. The direct conversion of the photosynthetate to pyruvate through Glycolysis used in the Krebs cycle to produce energy during the daytime, the long term storage of the photosynthetate as sucrose, and the short term storage of the photosynthetate as starch, to be used as an energy source at night.

To illustrate that interactions can be annotated in a pathway, we annotate the interactions using the FindYourInteractions (FYI) plugin (http://plugins.pathvisio.org/fyi/tutorial/). Genes and metabolites are annotated using existing PathVisio functionality (http://www.pathvisio.org/documentation/tutorials/tutorial-1/). This links the interactions, genes, and metabolites through BridgeDb to online data sources that provide more information about them. The side tab "Backpage" in PathVisio displays the annotation for the pathway element selected in the pathway diagram and its cross references to other databases as hyperlinks. In WikiPathways, interaction annotations are shown on the pathway page with all other annotation information. Clicking on the interaction in the pathway or on the information icon in the annotation list opens a popup with cross references to other databases shown as hyperlinks (Figure 5.1).



Title: Starch Metabolism
Organism: Arabidopsis thaliana

Interaction properties

Search
Search in Rhea

Manual entry
Biological Type: mim-conversion
Identifier: R02421
Database: KEGG Reaction

Interaction annotation

Identifier:	R02421
Database:	KEGG Reaction

Cross references

Enzyme Nomenclature
2.4.1.21

EcoCyc
GLYCOGENSYN-RXN

MetaCyc
GLYCOGENSYN-RXN

Rhea
18192

KEGG Reaction
R02421

Uniprot-TrEMBL
A0KPQ1
A0PY73
A0Q596
A1AGW2
A1BAE6
A1BIY8

| Gene database: ...Derby_Ensembl_80.bridge | Metabolite database: ...bolites_20150409.bridge | Interaction database: ...actions_20151111.bridge

Figure 5.1 Interaction annotation in PathVisio. (a) Pathway editing pane showing the pathway and the selected interaction, (b) interaction annotation dialog box, (c) Backpage panel showing interaction annotation and cross references for the interaction, (d) Bottom bar showing loaded gene, metabolite, and Interaction identifier mapping databases.

As an example for dynamic visualisation of fluxes, metabolic fluxes modelled by AraGEM are visualised on the glycolysis pathway (WP2621) from WikiPathways. The AraGEM model is a genome-scale metabolic model describing the functional primary metabolism of *A. thaliana* [34]. AraGEM simulates metabolic fluxes for three conditions; photosynthesis (i.e. photosynthesis without photorespiration), photorespiration (photosynthesis and photorespiration), and respiration/nitrogen assimilation in a non- photosynthetic cell (night time metabolism). Flux values for all three conditions are visualized using the slider feature of the IntViz plugin. Negative fluxes indicate that the reaction proceeds in the direction opposite to what is denoted in the diagram.

We integrate and visualise metabolic flux data simulated by the AraGEM model for the photosynthesis condition and gene expression data for photosynthesis comparing two wild type *A. thaliana* ecotypes Columbia-0 and C24 with their offspring. Ecotypes describe genetically distinct geographic varieties within a species that exhibit different phenotypes [35, 36]. Fujimoto *et al* observed increased photosynthetic capability by the offspring of a cross between the two wild type *A. thaliana* ecotypes Columbia-0 and C24 [37]. Transcriptome analysis is performed at 10 days after sowing by RNA extraction from the aerial tissues of the plants and hybridization to the cDNA Arabidopsis ATH1 Genome Array (Affymetrix). We downloaded the raw data (GSE32281) from GEO and performed quality control and normalization on ArrayAnalysis.org using the default criteria [38]. All arrays are found to be of sufficient quality for further analysis. Moderated t- statistics is then applied to the normalized dataset comparing the F1 hybrids with the parents, Columbia-0 and C24, using the statistical analysis module of ArrayAnalysis.org [39]. Benjamini and Hochberg's method is used for multiple correction. The log fold changes from the gene level statistics are represented on gene products using a colour gradient green, yellow, red corresponding to the values 2, 0, and -2. The metabolic fluxes for the photosynthesis condition are represented on the interactions (Figure 5.2).

The highest fluxes in the sugar metabolism of photosynthetic cells under normal daytime conditions are observed in the reactions for storage of carbon as starch. This is expected, since the products of photosynthesis are stored as starch during the day, to be degraded and exported for use in glycolysis and Krebs cycle at night. Low fluxes are observed in reactions leading to sucrose formation. Transcript levels vary little between the two genotypes. Low positive variations can be seen in the transcripts involved in starch storage, showing that increased photosynthesis would lead to more storage as starch. Contrarily, transcripts involved in sucrose metabolism decrease, decreasing sucrose formation. Sucrose being an osmotically active molecule, could cause osmotic imbalance upon accumulation. Since increased photosynthetic activity leads to increased substrate availability for sucrose formation, the decrease in associated transcripts could be the result of a regulatory process to keep sucrose concentrations stable. In Glycolysis, transcript levels change the least, and most changes appear to equilibrate between homologous transcripts. It is therefore understandable that the flux does not change much, as is found by the modelling approach, and the products keep channelling through 3-phosphoglycerate into the chloroplast for starch formation. The latter is also confirmed by the modelling results.

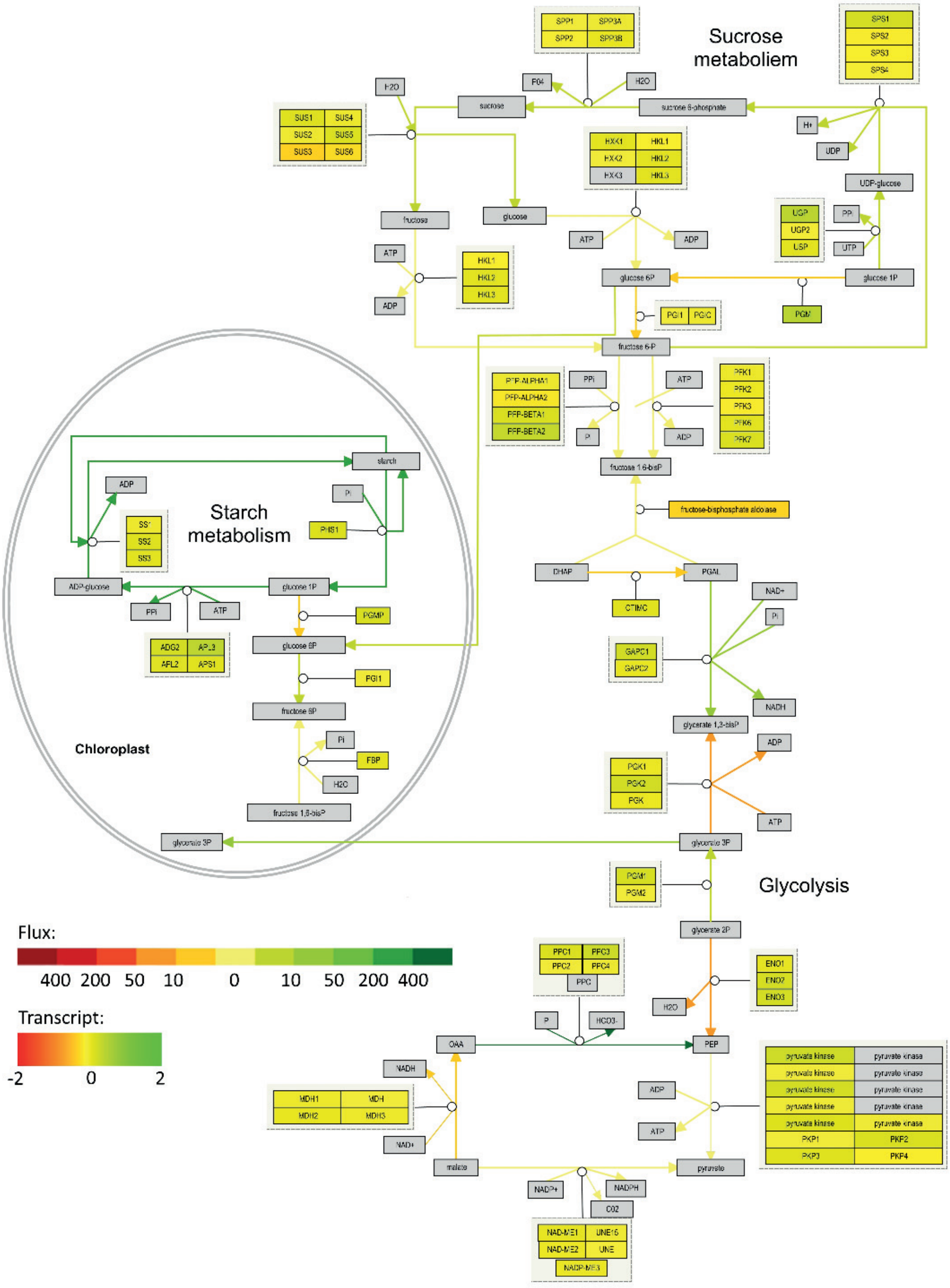


Figure 5.2 Sugar metabolism in Photosynthesis. Combined visualisation of transcriptomics and in-silico fluxomics data on the sucrose metabolism (WP2623) [a], starch metabolism (WP2622) [b], and glycolysis (WP2621) [c] pathways from WikiPathways. Log fold changes for the transcripts are visualised on the nodes using a colour gradient green, yellow, and red corresponding to the values 2, 0, and -2. Metabolic fluxes are visualised on the interactions. Positive fluxes are represented in shades of green, light to dark. Similarly, the negative values are represented in shades of red. Darker the colour, higher the flux and vice versa. Negative values indicate that the reaction proceeds in the direction opposite to what is denoted in the diagram. Zero fluxes are represenfted in yellow.

Discussion

To obtain a full picture of biological processes, measurements from different omics levels must be combined. The newly developed toolset allows pathway integration and visualisation of fluxomics data with transcriptomics, proteomics, and metabolomics data for a systemic overview of the biological process.

The interaction identifier mapping database allows users to annotate and map data to interactions in any analysis tool that uses BridgeDb as the identifier mapping framework. Therefore, in PathVisio this database and associated software extensions enable interactions in pathway diagrams to be annotated, cross referenced with hyperlinks to online data sources, and mapped to experimental or modelled data. In case an interaction has not yet been assigned an identifier by a public database, a mock identifier can be used for data mapping. Pathways drawn in PathVisio can be shared through WikiPathways. Interactions can be annotated online on WikiPathways directly. The 724 pathways available for analysis from WikiPathways cumulatively contain 46222 interactions, of which 81% are already annotated and can be used for data mapping.

The database is hosted online as part of the BridgeDb web services and can be queried programmatically as an identifier mapping service for interactions or to retrieve the directionality of interactions. For use in R, the interaction database can be downloaded and queried offline using the BridgeDbR package. In the future, we plan to query the database for curation purposes in WikiPathways. An automatic bot will systematically search the interaction database for each source/target node pair of each interaction. The annotations obtained by the automatic search will be listed on the pathway page as suggestions for human curators to confirm.

We developed three new plugins for PathVisio: PathSBML, FYI, and IntViz. The PathSBML plugin allows import, validation, and layout of SBML models. The FYI plugin directly queries the Rhea database for suitable identifiers based on the identifiers of the nodes the interaction connects, thus simplifying annotating interactions. The IntViz plugin for PathVisio provides visualisation styles for the visualisation of data on interactions, similar to the in-built visualisation features for nodes in PathVisio. Data can be visualised using colours, colour gradients, and the thickness of the line representing the interactions.

To illustrate the power of the combined visualisation we performed a cross-omics study in *A. thaliana* combining a published transcriptomics dataset comparing the two *A. thaliana* wild type genotypes with their hybrid, with increased photosynthetic activity, and in-silico fluxomics data for photosynthesis from AraGEM. This shows the value of combining flux data with other genomics data for pathway analysis.

Some other open source tools provide comparable visualisation features as the current toolset, but offer these for other environments. E.g. the FluxViz plugin [40] for Cytoscape, the FluxMap plugin [41] for VANTED, the fa-BINA plugin for BINA [42]. Most tools allow visualisation of data on pathways from KEGG and models encoded in SBML. FluxMap is the only one that like our IntViz allows combination of different conditions or time points it also allows data visualisation on pathways created in Pajek [43]. In most cases these tools could benefit from the extra identifier mapping opportunities that the BridgeDb approach and the new interaction identifier mapping database provide. The current toolset complements the existing flux visualisation tools by allowing analysis of flux data on WikiPathways and Reactome pathways. All tools allow users to create network diagrams for visualizing the flux data. An advantage of the developed toolset is that the interaction annotated pathway diagrams drawn in PathVisio can be shared on WikiPathways for community curation.

References

1. Kelder, T., M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C.T. Evelo, and A.R. Pico, **WikiPathways: building research communities on biological pathways**. *Nucleic acids research*, 2012. **40**(D1): p. D1301-D1307.
2. Khatri, P., M. Sirota, and A.J. Butte, **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS Comput Biol*, 2012. **8**(2): p. e1002375.
3. Sauer, U., **Metabolic networks in motion: 13C ‐ based flux analysis**. *Molecular systems biology*, 2006. **2**(1): p. 62.
4. Christensen, B. and J. Nielsen, **Isotopomer analysis using GC-MS**. *Metabolic engineering*, 1999. **1**(4): p. 282-290.
5. Orth, J.D., I. Thiele, and B.Ø. Palsson, **What is flux balance analysis?** *Nature biotechnology*, 2010. **28**(3): p. 245-248.
6. Kutmon, M., M.P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A.R. Pico, and C.T. Evelo, **PathVisio 3: an extendable pathway analysis toolbox**. *PLoS computational biology*, 2015. **11**(2).
7. Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome research*, 2003. **13**(11): p. 2498-2504.
8. Küntzer, J., T. Blum, A. Gerasch, C. Backes, A. Hildebrandt, M. Kaufmann, O. Kohlbacher, and H.-P. Lenhof, **BN++-a biological information system**. *J Integr Bioinformatics*, 2006. **3**(2): p. 34.
9. Rohn, H., A. Junker, A. Hartmann, E. Grafahrend-Belau, H. Treutler, M. Klapperstuck, T. Czauderna, C. Klukas, and F. Schreiber, **VANTED v2: a framework for systems biology applications**. *BMC Syst Biol*, 2012. **6**: p. 139.
10. Karp, P.D., S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T.J. Lee, P. Kaipa, F. Gilham, A. Spaulding, and L. Popescu, **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology**. *Briefings in bioinformatics*, 2009: p. bbp043.
11. Droste, P., K. Nöh, and W. Wiechert, **Omix–A Visualization Tool for Metabolic Networks with Highest Usability and Customizability in Focus**. *Chemie Ingenieur Technik*, 2013. **85**(6): p. 849-862.
12. Nikitin, A., S. Egorov, N. Daraselia, and I. Mazo, **Pathway studio—the analysis and navigation of molecular networks**. *Bioinformatics*, 2003. **19**(16): p. 2155-2157.
13. Kursawe, R., M. Eszlinger, D. Narayan, T. Liu, M. Bazuine, A.M. Cali, E. D'Adamo, M. Shaw, B. Pierpont, and G.I. Shulman, **Cellularity and adipogenic profile of the abdominal subcutaneous adipose tissue from obese adolescents: association with insulin resistance and hepatic steatosis**. *Diabetes*, 2010. **59**(9): p. 2288-2296.
14. Tisoncik, J.R., M.J. Korth, C.P. Simmons, J. Farrar, T.R. Martin, and M.G. Katze, **Into the eye of the cytokine storm**. *Microbiology and Molecular Biology Reviews*, 2012. **76**(1): p. 16-32.
15. Jitendra, S., A. Nanda, S. Kaur, and M. Singh, **A comprehensive molecular interaction map for Hepatitis B virus and drug designing of a novel inhibitor for Hepatitis BX protein**. *Bioinformation*, 2011. **7**(1): p. 9.
16. Zhou, C., Q. Zhong, L.V. Rhodes, I. Townley, M.R. Bratton, Q. Zhang, E.C. Martin, S. Elliott, B.M. Collins-Burow, and M.E. Burow, **Proteomic analysis of acquired tamoxifen resistance in MCF-7 cells reveals expression signatures associated with enhanced migration**. *Breast Cancer Res*, 2012. **14**(2): p. R45.
17. Rubio-Aliaga, I., B. de Roos, M. Sailer, G.A. McLoughlin, M.V. Boekschoten, M. van Erk, E.-M. Bachmair, E.M. Van Schothorst, J. Keijer, and S.L. Coort, **Alterations in hepatic one-carbon metabolism and related pathways following a high-fat dietary intervention**. *Physiological genomics*, 2011. **43**(8): p. 408-416.
18. Kutmon, M., A. Riutta, N. Nunes, K. Hanspers, E.L. Willighagen, A. Bohler, J. Melius, A. Waagmeester, S.R. Sinha, R. Miller, S.L. Coort, E. Cirillo, B. Smeets, C.T. Evelo, and A.R. Pico, **WikiPathways: capturing the full diversity of pathway knowledge**. *Nucleic Acids Res*, 2015.
19. van Iersel, M.P., A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo, **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC bioinformatics*, 2010. **11**(1): p. 5.
20. Morgat, A., K.B. Axelsen, T. Lombardot, R. Alcantara, L. Aimo, M. Zerara, A. Niknejad, E. Belda, N. Hyka-Nouspikel, E. Coudert, N. Redaschi, L. Bougueleret, C. Steinbeck, I. Xenarios, and A. Bridge, **Updates in Rhea--a manually curated resource of biochemical reactions**. *Nucleic Acids Res*, 2015. **43**(Database issue): p. D459-64.
21. Bairoch, A., **The ENZYME database in 2000**. *Nucleic acids research*, 2000. **28**(1): p. 304-305.
22. Fleischmann, A., M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K.B. Axelsen, A. Bairoch, D. Schomburg, K.F. Tipton, and R. Apweiler, **IntEnz, the integrated relational enzyme database**. *Nucleic acids research*, 2004. **32**(suppl 1): p. D434-D437.
23. Caspi, R., R. Billington, L. Ferrer, H. Foerster, C.A. Fulcher, I.M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D.S. Weaver, and P.D. Karp, **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res*, 2015.
24. Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic Acids Res*, 2014. **42**(Database issue): p. D199-205.
25. Morgat, A., E. Coissac, E. Coudert, K.B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari, **UniPathway: a resource for the exploration and annotation of metabolic pathways**. *Nucleic acids research*, 2011: p. gkr1023.
26. Holliday, G.L., D.E. Almonacid, G.J. Bartlett, N.M. O'Boyle, J.W. Torrance, P. Murray-Rust, J.B. Mitchell, and J.M. Thornton, **MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms**. *Nucleic acids research*, 2007. **35**(suppl 1): p. D515-D520.
27. Consortium, U., **UniProt: a hub for protein information**. *Nucleic acids research*, 2014: p. gku989.
28. Bohler, A., L.M. Eijssen, M.P. van Iersel, C. Leemans, E.L. Willighagen, M. Kutmon, M. Jaillard, and C.T. Evelo, **Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment**. *BMC bioinformatics*, 2015. **16**: p. 267.
29. Croft, D., A.F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, and M.R. Kamdar, **The Reactome pathway knowledgebase**. *Nucleic acids research*, 2014. **42**(D1): p. D472-D477.
30. Milacic, M., R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, and L. Stein, **Annotating cancer variants and anti-cancer therapeutics in reactome**. *Cancers*, 2012. **4**(4): p. 1180-1211.
31. Hucka, M., A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, and A. Cornish-Bowden, **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics*, 2003. **19**(4): p. 524-531.

32. Le Novere, N., S. Moodie, A. Sorokin, M. Hucka, F. Schreiber, E. Demir, H. Mi, Y. Matsuoka, K. Wegner, and H. Kitano, **Systems biology graphical notation: process diagram level 1**. *Nature Precedings*, 2008.

33. Le Novere, N., B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, and B. Shapiro, **BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems**. *Nucleic acids research*, 2006. **34**(suppl 1): p. D689-D691.

34. de Oliveira Dal'Molin, C.G., L.-E. Quek, R.W. Palfreyman, S.M. Brumbley, and L.K. Nielsen, **AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis**. *Plant physiology*, 2010. **152**(2): p. 579-589.

35. Begon, M., C.R. Townsend, and J.L. Harper, **Ecology: from individuals to ecosystems**. 2006.

36. Turesson, G., **The genotypical response of the plant species to the habitat**. *Hereditas*, 1922. **3**(3): p. 211-350.

37. Fujimoto, R., J.M. Taylor, S. Shirasawa, W.J. Peacock, and E.S. Dennis, **Heterosis of Arabidopsis hybrids between C24 and Col is associated with increased photosynthesis capacity**. *Proc Natl Acad Sci U S A*, 2012. **109**(18): p. 7109-14.

38. Eijssen, L.M., M. Jaillard, M.E. Adriaens, S. Gaj, P.J. de Groot, M. Müller, and C.T. Evelo, **User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis. org**. *Nucleic acids research*, 2013. **41**(W1): p. W71-W76.

39. Dutta, A., **Adding automated Statistical Analysis and Biological Evaluation modules to www. arrayanalysis. org**. 2011, Maastricht University.

40. König, M. and H.-G. Holzhütter, **Fluxviz—cytoscape plug-in for visualization of flux distributions in networks**. *Genome Informatics*, 2010: p. 96.

41. Rohn, H., A. Hartmann, A. Junker, B.H. Junker, and F. Schreiber, **FluxMap: a VANTED add-on for the visual exploration of flux distributions in biological networks**. *BMC Syst Biol*, 2012. **6**: p. 33.

42. Hoppe, A., S. Hoffmann, A. Gerasch, C. Gille, and H.-G. Holzhütter, **FASIMU: flexible software for flux-balance computation series in large metabolic networks**. *BMC bioinformatics*, 2011. **12**(1): p. 28.

43. Batagelj, V. and A. Mrvar, **Pajek-program for large network analysis**. *Connections*, 1998. **21**(2): p. 47-57.

CHAPTER 6

Computational exploration of metabolically abnormal obesity

Manuscript in preparation

Abstract

Obesity is defined as an excess of adipose tissue, which is the source of various hormones and other active compounds that play a pivotal role in body weight regulation and metabolic homeostasis. It is therefore of key interest to study adipose tissue dysfunction in the context of metabolic diseases. In the current study, we explore the molecular mechanisms in the absorptive state of adipose tissue after moderate weight gain in metabolically abnormal obese individuals compared to metabolically normal obese individuals combining publicly available transcriptomics and flux data.

Pathway analysis, using the WikiPathways and Reactome human pathway collections, revealed fourteen significantly altered pathways between metabolically abnormal vs normal obese individuals. These pathways were then merged into a network, revealing the connections between the significantly altered pathways. Flux data for reactions occurring in the adipose tissue were obtained by flux balance analysis of a publicly available genome scale metabolic model of adipocytes. The model predicted flux values were used as an additional verification of the activity of genes involved in gaining weight.

Fatty Acid Synthase (FASN), Acetyl-CoA Carboxylase Alpha (ACACA), and Acetyl-CoA Carboxylase Beta (ACACB), showed up as key metabolic regulators downregulated in the transcriptomics dataset, central in the network connecting five pathways related to metabolism and its regulation, and found to catalyse reactions that have fluxes in the absorptive state of adipose tissue according to the metabolic model.

The integration of transcription factor-gene interactions and microRNAs from mirTarBase, identified new links between the pathways on a regulatory level, linking FASN, ACACA, and ACACB to all fourteen significantly changed pathways by shared regulatory elements. Extension of the network with gene-disease associations from DisGeNET revealed strong associations with cardiomyopathy, mental retardation, obesity, and insulin resistance. Further extension with known drug-target interactions from DrugBank highlights several approved drugs targeting genes in the network, amongst which are Cerulenin, Fomepizole, Mecasermin, Mefloquine, Nedocromil and Quercetin, as drugs which could be further investigated for treatment of metabolic syndrome and obesity.

Introduction

Obesity or excessive body adiposity is commonly measured using Body Mass Index (BMI) and adults with a BMI > 30 are considered obese [1]. The number of obese people worldwide doubled since the 1980s and the global nature of the obesity epidemic were formally recognized by a World Health Organization consultation in 1997 [2]. The outbreak of obesity also highlights that only a subgroup of obese people develop associated comorbidities such as type 2 diabetes [3, 4], cardiovascular disease [4, 5], depression [6, 7], neurodegenerative disorders [8, 9], infectious diseases [10, 11], cancer [12, 13], and chronic kidney diseases [14]. These are the obese individuals with other metabolic abnormalities such as dyslipidaemia, insulin resistance, hypertension, and an unfavourable inflammatory profile [15-17], together referred to as metabolic syndrome [18]. The subgroup of obese individuals with metabolic syndrome are referred to as Metabolically Abnormal Obese and the obese individuals with a favourable metabolic profile are referred to as Metabolically Normal Obese.

White adipose tissue stores excess energy as triglycerides in the lipid droplets of adipocytes [19]. Interestingly, it also produces hormones playing an important role in body weight regulation and metabolic homeostasis, such as leptin, visfatin, apelin, resistin, and adiponectin [20]. It also produces other proinflammatory cytokines, complement factors and components of the coagulation/fibrinolytic cascade that may mediate the metabolic and cardiovascular complications associated with obesity [21]. Sex steroids and glucocorticoids, which are important determinants of body fat distribution and cardiovascular risk are metabolised in the adipose tissue as well [22, 23]. The adipose tissue is therefore now considered a major endocrine organ, at the heart of a complex network that influences energy homeostasis, glucose and lipid metabolism, vascular homeostasis, immune response and even reproduction [24]. In addition, Wagner *et al.* recently stated that the adipose transcriptome was even more highly correlated with total body weight than the liver transcriptome in a mouse study [25]. This indicates that the molecular mechanisms of adipose tissue are affected due to changes in body weight.

Other studies have found gaining weight worsens the health of MAO individuals [5]. Combined with the importance of transcriptomic changes in adipose tissue in obesity as described above, it is therefore of key interest to study the metabolism of adipose tissue after weight gain, to explore the effect of gaining weight on the metabolism. In the current study we used a transcriptomics dataset (GSE62832) published by Fabbrini *et al.* [26] and performed pathways analysis to detect which biological process are significantly changed in MAO individuals compared to MNO individuals, before and after they gain a moderate amount of weight. We map predicted fluxes from a genome-scale metabolic model of adipocytes simulating the absorptive state onto the metabolic genes in the biological pathways as an extra verification of the activity of the genes involved in gaining weight. Subsequently, we merge the significantly changed pathways into a network to study how they are connected. Following which we explore regulation of the network by transcription factors and microRNAs. Finally, we identify potential drug targets and drugs which might be candidates for repurposing.

Results

Overview of approach to explore the significantly affected biological processes in metabolically abnormal obese

Our analyses aimed to identify processes that are significantly altered in metabolically abnormal obese (MAO) individuals compared to metabolically normal obese (MNO) individuals and the effect of gaining weight on both groups by integrating multiple levels of omics data. Obese individuals with intrahepatic triglyceride levels higher than 10% were classified as MAO and lower than 5.6% were classified as MNO.

We identify a network of active pathways in the absorptive state in adipose tissue by performing pathway analysis with publicly available transcriptomic data. Next we map the flux data simulated by a genome-scale metabolic model to find genes of interest from the tail ends of the molecular process. Microarray studies use transcriptional abundance as a stand-in/representative of protein/enzyme abundance, which optimally requires a linear relationship between the two, which is often not the case. However, a recent study has demonstrated that differentially expressed mRNAs correlate significantly better with their protein product than non-differentially expressed mRNAs [27], confirming the relevance of using transcriptomics data for the study of molecular mechanisms. Metabolic fluxes on the other hand, are regulated by substrate availability and enzyme activity. The active state of an enzyme is a result of the total of transcriptomic, post transcriptomic, translational, and post-translational changes. Combined evidence from both transcriptomics and flux data highlights the key regulators of the network.

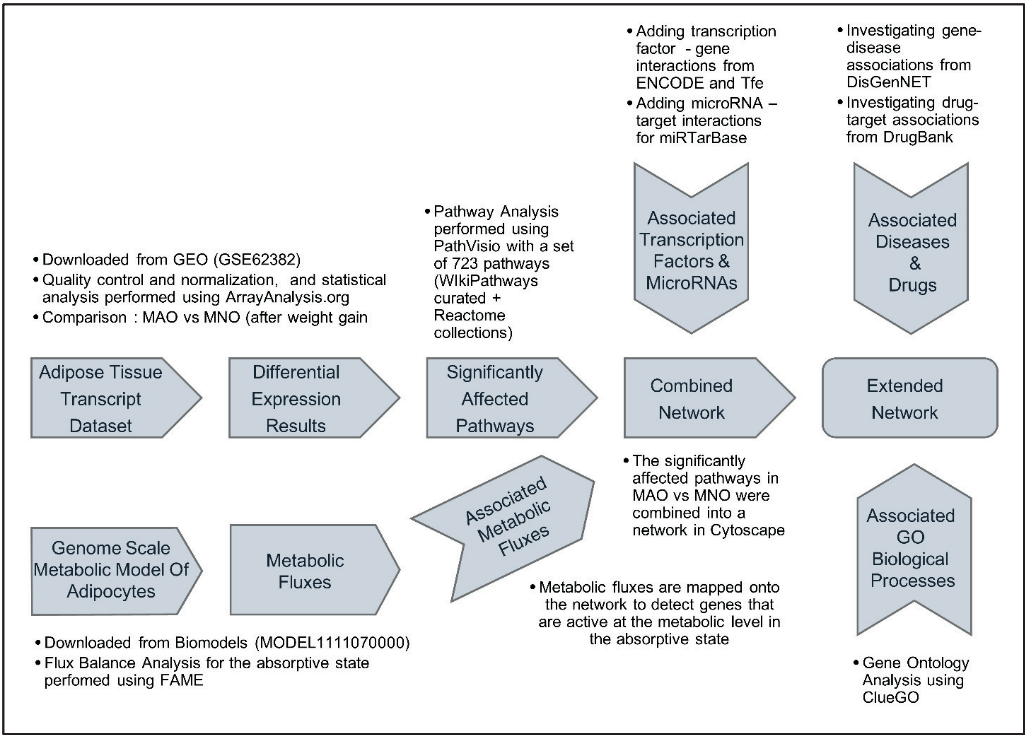

• Adding transcription factor - gene interactions from ENCODE and Tfe
• Adding microRNA – target interactions for miRTarBase
• Investigating gene-disease associations from DisGenNET
• Investigating drug-target associations from DrugBank
Associated Transcription Factors & MicroRNAs
Associated Diseases & Drugs
• Pathway Analysis performed using PathVisio with a set of 723 pathways (WIkiPathways curated + Reactome collections)
• Downloaded from GEO (GSE62382)
• Quality control and normalization, and statistical analysis performed using ArrayAnalysis.org
• Comparison : MAO vs MNO (after weight gain
Adipose Tissue Transcript Dataset
Differential Expression Results
Significantly Affected Pathways
Combined Network
Extended Network
Associated Metabolic Fluxes
• The significantly affected pathways in MAO vs MNO were combined into a network in Cytoscape
Associated GO Biological Processes
Genome Scale Metabolic Model Of Adipocytes
Metabolic Fluxes
• Metabolic fluxes are mapped onto the network to detect genes that are active at the metabolic level in the absorptive state
• Downloaded from Biomodels (MODEL1111070000)
• Flux Balance Analysis for the absorptive state perfomed using FAME
• Gene Ontology Analysis using ClueGO

Figure 6.1 Flow chart illustrating the analytical design of this study. We analysed a publicly available transcriptomics dataset (GSE62832) comparing metabolically abnormal obese to metabolically normal obese individuals to obtain differentially expressed genes. We performed Pathway Analysis to identify significantly changed pathways in metabolically abnormal obese. The Fold Changes and P values of the genes were visualized on the pathways. We considered the most significantly changed pathways to be candidates for further investigation and combined them into a network to identify how the biological pathways are interconnected. Next, we used a publicly available genome scale metabolic model for adipocytes to estimate metabolic flux values for reactions that would be active in adipocytes during the absorptive state using Flux Balance Analysis. Genes coding for enzymes catalysing reactions with metabolic fluxes are considered active during the absorptive state. Next, we extended this network with regulatory information about transcription factors and microRNAs to study regulatory effects. For additional investigation of the extended network. We performed gene ontology analysis to investigate processes significantly affected in metabolically abnormal obese individuals. We also investigated gene-disease associations and drugs for the genes in the network.

We analysed a publicly available transcriptomics dataset (GSE62832) comparing MAO to MNO individuals before and after weight gain to obtain differentially expressed genes in each case. Pathway analysis was performed with the differentially expressed datasets and pathways from WikiPathways and Reactome to identify biological processes that are significantly altered in MAO individuals compared to MNO individuals. Pathway analysis is performed using a publicly available gene expression dataset to identify biological processes that are significantly altered between MAO and MNO individuals. Biological pathways are not disparate entities. They interact with other pathways to form networks and thus operate in living systems. We combined the significantly affected pathways into a network to study the links between significantly affected biological processes (Figure 6.1).

A publicly available genome-scale metabolic model for adipocytes was then used to estimate metabolic flux values for reactions that are active in adipocytes during the absorptive state using Flux Balance Analysis. The transcriptomics dataset is also generated from adipose tissue in the absorptive state, as the sampling was carried out one hour after the start of glucose administration. This makes the transcript data and the model particularly compatible for the study of the absorptive state of the adipose tissue. The flux data from the metabolic model is mapped onto the network to highlight genes catalysing reactions in the absorptive state. Next the network is extended with transcription factors from TFe [28] and ENCODE [29] and microRNAs from mirTarBase [30] to study additional regulatory links between the pathways. Subsequently, we investigated the extended network further by (1) performing Gene Ontology analysis using the ClueGO app in Cytoscape to identify biological processes affected by the integrated network, (2) investigating gene-disease associations from DisGeNET [31], and (3) drugs from DrugBank [32] that target genes in the extended network using the Cytargetlinker app in Cytoscape (Figure 6.1).

Identifying significantly affected biological processes using the transcriptomics dataset

Fourteen significantly changed pathways were obtained by pathway analysis for the comparison between MAO and MNO individuals after weight gain (Table 6.1). Six pathways are related to metabolism and its regulation (Activation of gene expression by SREBF, Fatty Acid Biosynthesis, Fatty acid, triacylglycerol, and ketone body metabolism, Integration of energy metabolism, Phase 1 - Functionalization of compounds, AMPK signalling), six are related to transcription/translation, one to disease, and one was a transport pathway.

The Activation of gene expression by SREBF (SREBP) pathway (WP2706, converted from Reactome) and Fatty Acid Biosynthesis Pathway (WP357, from the WikiPathways curated collection) show up as the top 2 affected pathways (Figure 6.2). Both pathways are significantly downregulated in the current dataset. Significantly changed genes Acetyl-CoA Carboxylase Alpha (ACACA; FC: -1.47), Acetyl-CoA Carboxylase Beta (ACACB; FC:-1.31), and Fatty Acid synthase (FASN; FC: -1.87) are present in both pathways. Furthermore, flux balance analysis data shows that these three enzymes also catalyse reactions that have flux during the absorptive state in absorptive tissue, highlighting them as clear genes of interest for the absorptive state in adipose tissue from both transcriptomics and flux data. Other significantly changed genes in the Activation of gene expression by SREBF (SREBP) pathway, include ELOVL6 (FC: -2.13), TM7PSF2 (FC: -1.47), and LSS (FC: -1.31). Besides FASN, ACACA, and ACACB, ACLY (FC: -1.41) and HADHSC (FC: -1.31), are also significantly changed in the Fatty Acid Biosynthesis pathway.

Table 6.1 Fourteen pathways significantly changed in MAO vs MNO after the two groups gained weight. The pathways are ranked according to Z score.

Pathway	Pathway Class	Collection	Z Score	p-value (permuted)
Activation of gene expression by SREBF (SREBP)	Metabolism	Reactome	9.55	0.001
Fatty Acid Biosynthesis	Metabolism	WikiPathways	7.43	0.001
Eukaryotic Translation Elongation	Translation	Reactome	5.27	0.001
Binding and Uptake of Ligands by Scavenger Receptors	Vesicle mediated transport	Reactome	5.15	0.001
SRP-dependent cotranslational protein targeting to membrane	Translation	Reactome	5.1	0.001
Eukaryotic Translation Termination	Translation	Reactome	4.58	0.002
Cytoplasmic Ribosomal Proteins	Translation	WikiPathways	4.43	0.001
Nonsense-Mediated Decay (NMD)	Transcription	Reactome	3.81	0.003
AMPK Signalling	Metabolism	WikiPathways	3.62	0.007
Eukaryotic Translation Initiation	Translation	Reactome	3.59	0.002
Fatty acid, triacylglycerol, and ketone body metabolism	Metabolism	Reactome	3.59	0.001
Influenza Life Cycle	Disease	Reactome	3.48	0.001
Integration of energy metabolism	Metabolism	Reactome	2.85	0.008
Phase 1 - Functionalization of compounds	Metabolism	Reactome	2.08	0.032

MAO: Metabolically Abnormal Obese; MNO: Metabolically Normal Obese

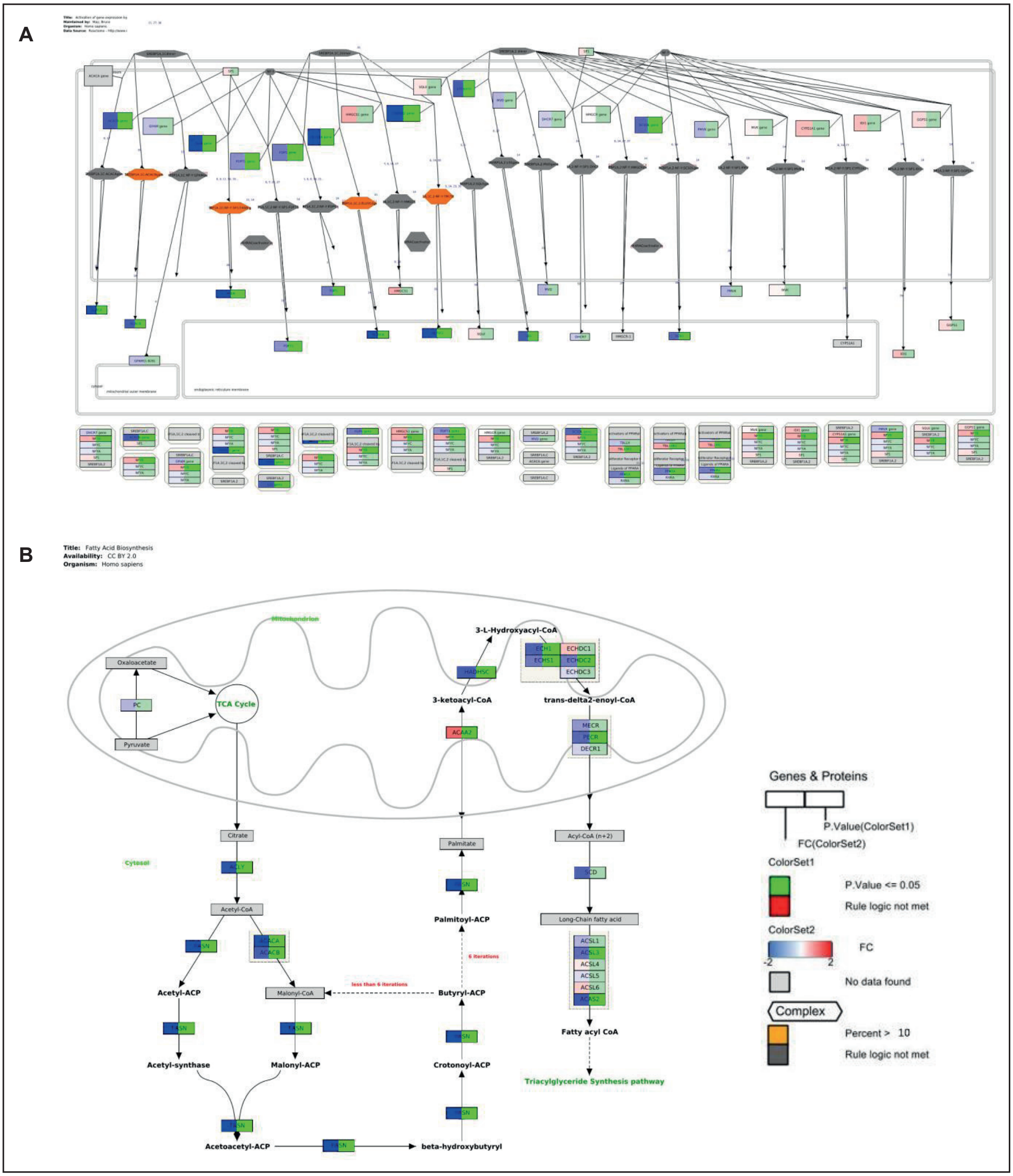


Figure 6.2 Fold changes, P values, and flux activity of genes are visualized on the two top ranking pathways. Genes are represented as rectangles. Complexes are represented as hexagons and their components are listed on the bottom of the pathway. The Fold Change, P value, and Flux activity is visualized on the genes of the network. Downregulated genes are visualized in blue, upregulated in red and unchanged in white. Significant genes (P value < 0.05) are highlighted in green and genes with an active flux are highlighted in dark green. (A)Activation of gene expression by SREBF pathway, (B) Fatty acid Biosynthesis pathway.

Significantly affected network of biological processes in metabolically abnormal obese

We combined the fourteen significantly changed pathways into a network to study the connections between the pathways and detect which genes connect a number of processes and could potentially be a key regulator in the network. A network of 584 gene products and 310 metabolites is obtained, in which thirty five genes and thirty one metabolites link two or more pathways. Thirteen of these genes, namely ELOVL6, FASN, ACACA, ACACB, ACLY, RPS5, RPS9, RPS15, ACSS2, RPL10A, RPL9, RPL12, and RPL29, are downregulated. RPS4Y1 is the only gene connecting multiple pathways that is upregulated. The pathways related to similar processes group together in the network (Figure 6.3). It is furthermore notable that most of the proteins present in multiple pathways are involved in translation and transcription.

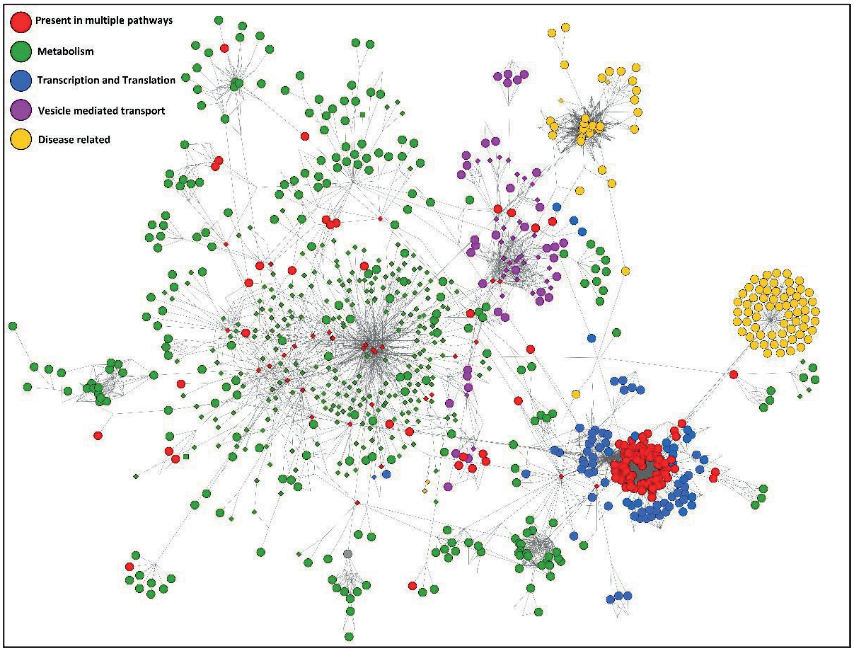

Present in multiple pathways
Metabolism
Transcription and Translation
Vesicle mediated transport
Disease related

Figure 6.3 Network of all significantly affected pathways in MAO vs MNO after weight gain. The fourteen pathways have been classified into four groups. Nodes are coloured according to the group in which the pathway they belong to falls. Red nodes are present in multiple pathways.

Adding regulatory information

The network of all fourteen pathways was then extended with regulatory information about microRNAs from mirTarBase [30] and transcription factors from Transcription Factor Encyclopaedia (TFe) [28] and ENCODE [29]. In total 124 transcription factors were added. 74 transcription factor-gene interactions were found in TFe, 530 distal (i.e. regulatory element distant from the gene) and 1702 proximal (i.e. regulatory element adjacent to the gene) transcription-gene connections were found in the ENCODE dataset. 93 transcription factors showed both proximal and distal control. 141 validated microRNAs targeting genes in the network were added from mirTarBase with 787 regulatory connections.

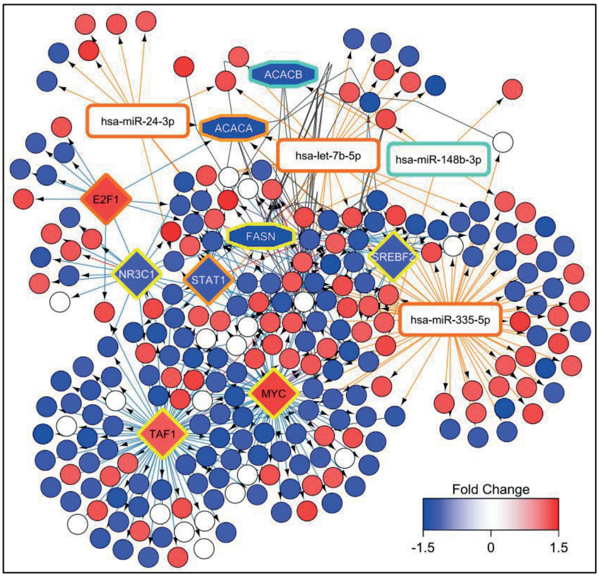


Figure 6.4 The network of fourteen significant pathways extended with transcription factors from Encode and TFe and microRNAs from mirTarBase. All fourteen significant pathways were connected through shared regulatory elements to FASN, ACACA, and ACACB, which were all downregulated in the transcriptomics dataset and the reactions catalzyed by them had changed fluxes according to the metabolic model. The fold change values are visualized on the nodes. Gene products are represented as ellipses, transcription factors are represented as diamonds, and microRNAs are represented as rounded rectangles. FASN, ACACA, and ACACB and the transcription factors and miRNAs that target them are highlighted using borders of the same colour. Only transcription factors and microRNAs targeting FASN, ACACA, or ACACB are shown.

Subsequently, a publicly available genome-scale metabolic model of the adipose tissue simulating the absorptive state (MODEL1111070000) published by Bordbar et al. [33] was used to estimate fluxes by flux balance analysis. Reactions catalysed by three genes, FASN, ACACA, and ACACB, had flux values associated with them in the modelling results. They were also significantly changed in the transcriptomics dataset and hold a central position in the created network as they are linked to all fourteen significantly altered pathways by shared transcription factors and microRNAs. FASN is regulated by four transcription factors NR3C1, SREBF2, MYC, and TAF1. ACACA is regulated by two transcription factors E2F1 and STAT1 and three microRNAs hsa-mir-335-5p, has-mir-24-3p, and has-let-7b-5p. ACACB is regulated by has-mir-148b-3p (Figure 6.4).

Overrepresented Biological processes

We identified important functions of the network of the fourteen significantly affected pathways extended with regulatory information using the Cytoscape app ClueGO [34]. The app creates a network of interconnected GO biological processes based on the similarity of their associated differentially expressed genes (Figure 6.5).

The genes in the network were grouped into thirty GO groups, including those represented by small molecule metabolic process (GO: 0044281), lipid localization (GO: 0010876), fatty acid metabolic process (GO: 0006631), sterol metabolic process (GO: 0016125), response to hormone (GO: 0009725), response to lipid (GO: 0033993), response to nutrient (GO: 0007584), and response to wounding (GO: 0009611).

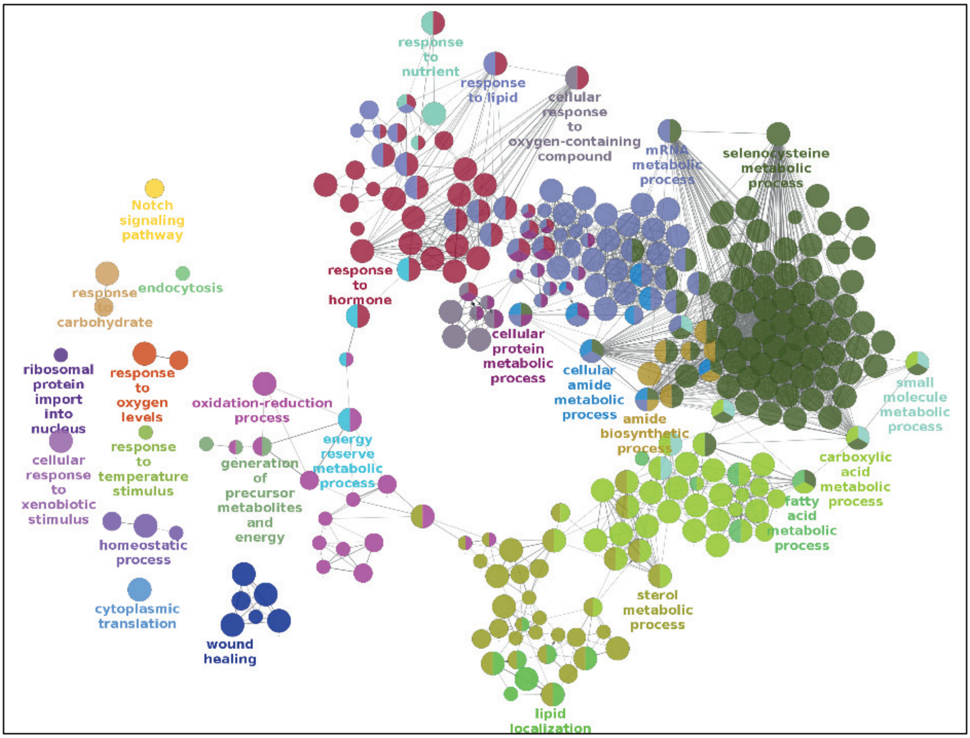

response to nutrient
response to lipid
cellular response to oxygen-containing compound
mRNA metabolic process
selenocysteine metabolic process
Notch signaling pathway
response to carbohydrate
endocytosis
response to hormone
cellular protein metabolic process
cellular amide metabolic process
amide biosynthetic process
small molecule metabolic process
ribosomal protein import into nucleus
response to oxygen levels
oxidation-reduction process
energy reserve metabolic process
cellular response to xenobiotic stimulus
response to temperature stimulus
generation of precursor metabolites and energy
carboxylic acid metabolic process
fatty acid metabolic process
homeostatic process
cytoplasmic translation
wound healing
sterol metabolic process
lipid localization

Figure 6.5 ClueGO network for the gene products in the extended network. The ClueGO app was used to find overrepresented GO processes and a network of connected GO terms was created. Each node represents a GO biological process, and the colours represent the GO group. Thirty GO groups are present in the network, one representing GO biological process per group is named in the figure. The edges reflect the relationships between the terms based on the similarity of their associated genes.

Associated comorbidities

78 diseases were found to be associated with genes in the network Diseases were classified into 11 groups according to the MeSH disease classification: Cardiovascular Diseases, Nervous System Diseases, Mental Disorders, Nutritional and Metabolic Diseases, Musculoskeletal Diseases, Kidney Diseases (3), Eye Diseases, Neoplasms, Pathological Conditions, Signs and Symptoms, Skin and Connective Tissue Diseases, and Virus Diseases (Figure 6.6). Most populous groups were, neoplasms: 29, diseases of the nervous system: 12, nutritional and metabolic diseases: 10, and cardiovascular diseases: 10.

Referring back to the three genes mentioned earlier, ACACA is linked to Hepatocellular Carcinoma, which targets is linked to four other genes in the network MTOR, ACSL4, SCD, RPS6KA3. ACACB is linked to Insulin resistance and Obesity. In the network, insulin resistance also links to PRKAA1 and PRKAA2, and Obesity also targets PFKFB3 and ACSL1. ACSL1 in turn links with Arsenic Poisoning, Myocardial Ischemia, and Skin Diseases. FASN is not found known to be associated with any diseases according to DisGeNET. Other than ACACA and ACACB, only PFKFB3 is downregulated (FC: -1.43, P-value: 0.02), all other genes with known disease associations are not significantly changed.

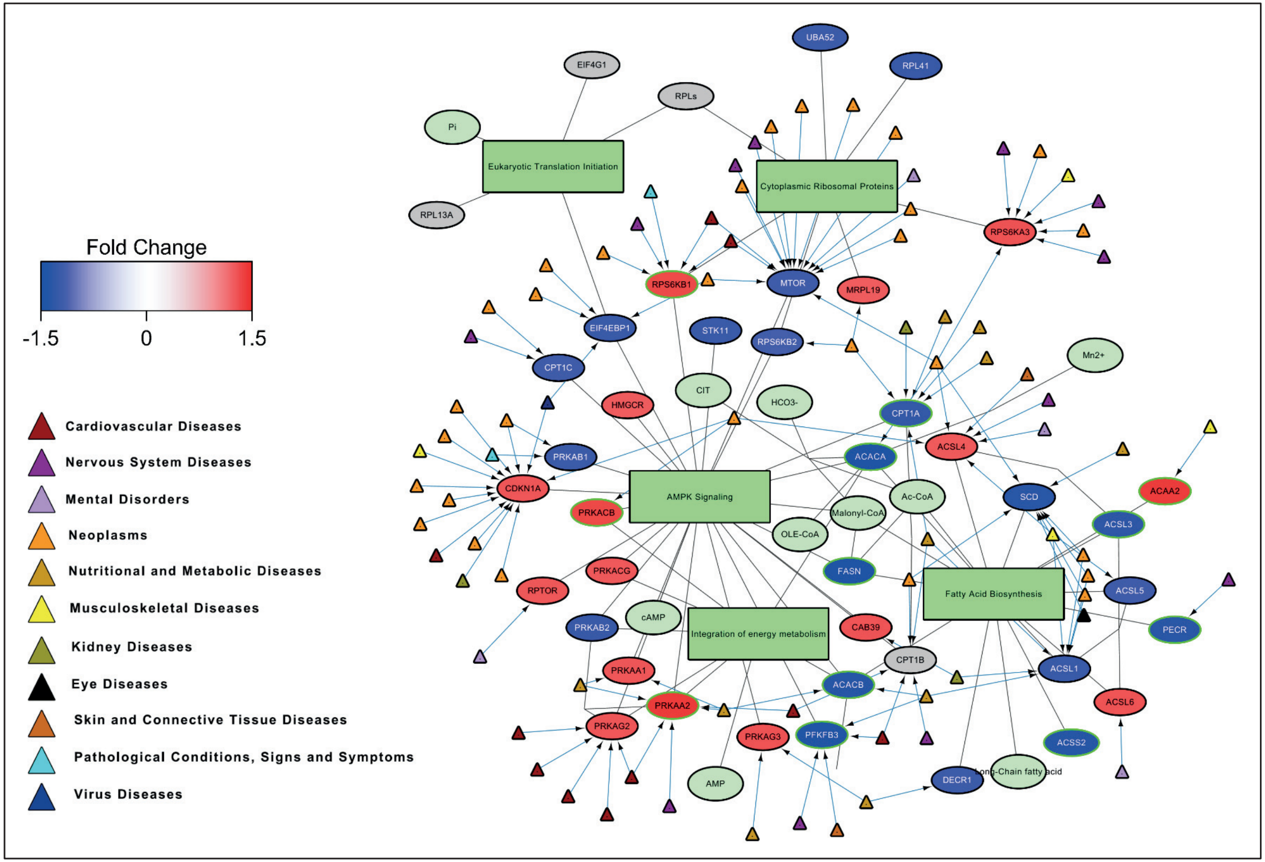


Figure 6.6 The network was extended with gene disease associations from DisGeNET to identify diseases associated with metabolic syndrome. Each pathway is represented as a light green rectangle. Genes are represented as rounded rectangles. Diseases are represented as triangles. The transcription data is visualized on the genes using a colour gradient from blue (downregulated) over white (not changed) to red (up-regulated). Grey nodes have no data. Metabolites are represented in light green. Nodes with a significant p-value (< 0.05) have a light-green border colour. 79 disease associations were found, most of which were for genes in AMPK signalling pathway (http://wikipathways.org/instance/WP1403) or genes linking multiple pathways. The diseases have been classified into twelve classes using MeSH: Cardiovascular Diseases, Nervous System Diseases represented, Mental Disorders, Neoplasms, Nutritional and Metabolic Diseases, Musculoskeletal Diseases, Kidney Diseases, Eye Diseases, Skin and Connective Tissue Diseases, Pathological Conditions, Signs and Symptoms, and Virus Diseases.

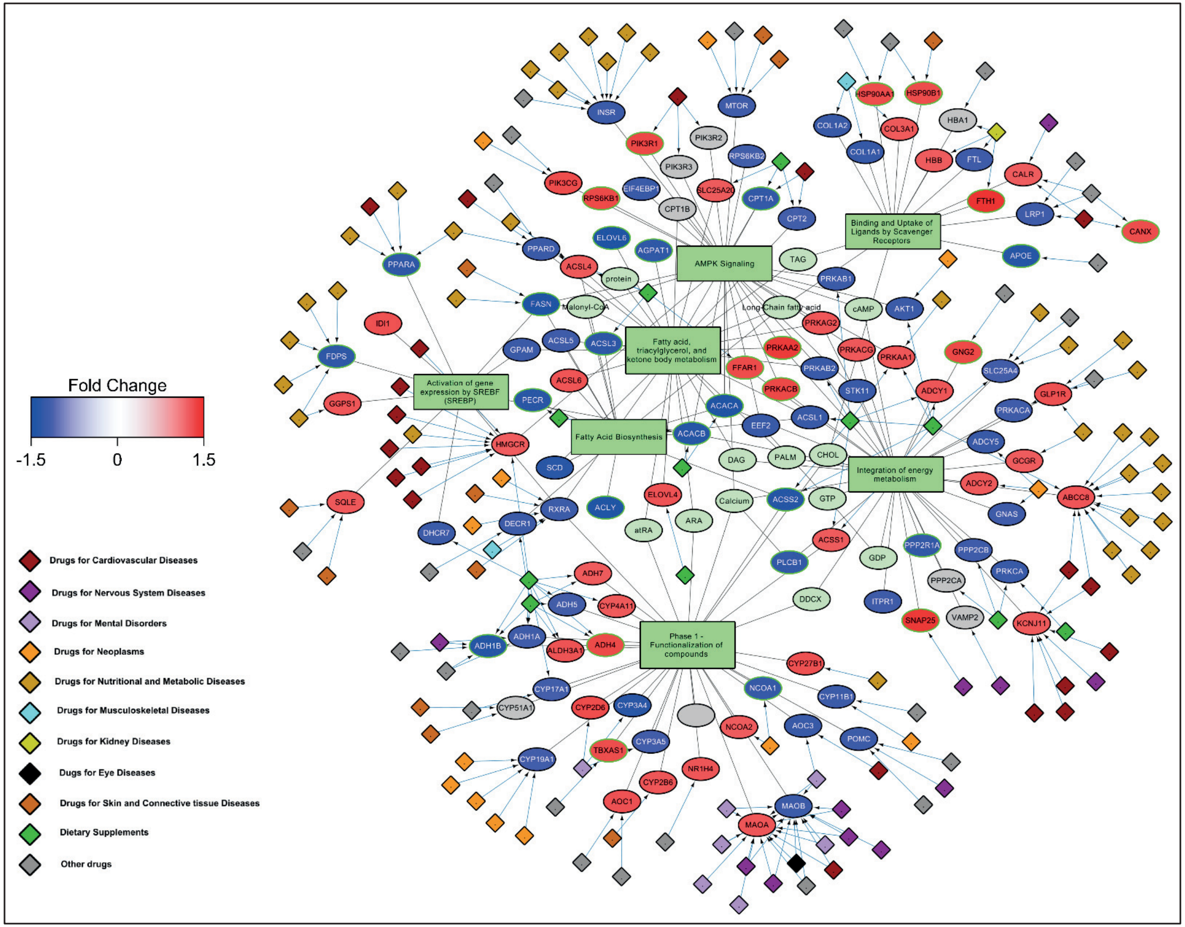


Figure 6.7 The network was extended with drug associations from Drug Bank to identify drugs targeting genes associated with metabolic syndrome. Each pathway is represented as a light green rectangle. Genes are represented as elypses. Drugs are represented as diamonds. The transcription dataset is visualized on the genes using a colour gradient from blue (downregulated) over white (not changed) to red (up-regulated). Grey nodes have no data. Metabolites are represented in light green. Nodes with a significant p-value (< 0.05) have a green border colour. 139 drugs were found to target genes in the network The drugs were classified into 11 categories of diseases found to be associated with the network.

Potential drugs and drug targets

Drug target analysis revealed that 139 drugs target genes in the network (Figure 6.7). The drugs were classified according to the 11 disease classes obtained by the MeSH classifications: drugs for Cardiovascular Diseases, Nervous System Diseases, Mental Disorders, Nutritional and Metabolic Diseases, Musculoskeletal Diseases, Kidney Diseases, Eye Diseases, Neoplasms, Skin and Connective Tissue Diseases, dietary supplements, and other drugs. Numerous drugs target genes are involved in more than one disease, e.g. glyburide which is an anti-arrhythmia, antidiabetic, and hypoglycaemic agents.

FASN is targeted by Cerulin and Orlistat. The nutritional supplement Adenine targets ACACB. Another nutritional supplement, Biotin, targets both ACACA and ACACB. Quercetin and Staurosporine (investigational cancer drugs) target PIK3CG. Zonisamide and pargyline target MAOA and MAOB, both known targets for drugs for nervous system diseases and mental diseases. Another nutritional supplement, Iron Dextran, targets FTH1 (FC: 1.36). ELOVL6 (FC: -2.14) and ACLY (FC: -1.41), both significantly changed in the network, were targeted by no drugs. APOE, FTH1, and SNAP25, other significantly changed genes in the network, were each targeted by a single drug not known for use in a disease associated with the metabolically active network. [35]. Etorphine, Fomepizole, Halothane, Iron Dextran, Mecasermin, Mefloquine, Metyrapone, MMDA, Nedocromil, Progesterone, Quercetin, Rifabutin, and Diminazene also target genes in the network, but these are not known to be used in the treatment of any of the diseases associated with the network.

Discussion

In the wake of the global obesity epidemic, it has come to light that not all obese people develop related comorbidities. Obese individuals with an unfavourable metabolic profile (for example elevated IHTG content) are more at risk for developing other diseases related to obesity compared to obese individuals with a favourable metabolic profile. In this paper, we performed an exploratory analysis to investigate which biological processes are significantly affected in metabolically abnormal obesity compared to metabolically normal obesity, how these biological processes are linked to each other, and which of the genes in the network could be key regulators.

Transcriptomics level data measure the expression of all the genes in the tissue. In pathway analysis, transcriptomics datasets are commonly mapped to the enzymes catalysing biological reactions and the set of pathways are then analysed to detect the significantly affected ones. However, genes undergo post transcriptional, translational, and post translational modifications prior to becoming the active enzymes that catalyse biological reactions. A reaction is accompanied by the flow of metabolic flux, therefore the flux value for a reaction indicates that the reaction occurs in the condition under investigation. Therefore, metabolic fluxes [26]provide an additional layer of information about the activity of the gene at the process level.

In this study we performed pathway analysis with a publicly available transcriptomics dataset measuring gene expression in adipose tissue during absorptive state for MAO and MNO individuals, after a moderate amount of weight gain.

Activation of gene expression by SREBPs is most affected in MAO. SREBPs or sterol regulatory element-binding proteins control intracellular levels of cholesterol and fatty acids through a feedback regulatory system [36]. Downregulation of this process indicates that the ability to regulate intracellular levels of cholesterol and fatty acids is affected in obese individuals with metabolic abnormality. Fatty Acid Biosynthesis pathway is downregulated contrary to an increase of de novo FA synthesis as would be expected due to the development of obesity, however this has also stated by other authors [37-39]. Interestingly, the AMPK signalling pathway is also downregulated; as this might indicate problems with energy utilization in metabolically abnormal obese. AMPK inactivates lipid synthesis by phosphorylating amongst others ACACA and ACACB, both of which are downregulated in the present data. ACACB is involved in the regulation of mitochondrial fatty acid oxidation [40]. In addition, AMPKa2 inhibits PFK2, which is also downregulated in the present data, and leads to the downregulation of glycolysis. Downregulation of biological pathways and genes associated with adipose tissue fatty acid metabolism is observed in agreement with the findings of Fabbrini et al [26]. This would mean that in MNO individuals, the stored fatty acids in the adipose tissue are metabolized, but not in MAO individuals. This would lead to a more permanent accumulation of adipose tissue in MAO individuals, thereby aggravating the negative effects of obesity. Our findings support that MAO individuals are predisposed to the adverse metabolic effects of moderate weight gain compared to MNO individuals.

We combined the significantly affected pathways into a network to study how the affected biological processes link to each other. The network was layout with the Force directed algorithm taking edge betweenness values into account. The force directed layout algorithm lays the nodes in the network based only on information contained within the structure of the network itself. It calculates the location of one node based on repulsive forces between all the nodes in the network and attractive forces between adjacent nodes. The edge betweenness centrality gives the number of shortest paths that go through an edge in a network, an edge with a high edge betweenness centrality score represents a bridge-like connector between two parts of a network. The layout used effectively grouping together the nodes that are highly connected to each other. Pathways relating to similar biological processes grouped together in the network confirming the process level knowledge in the biological pathways.

Upon combining the significantly changed pathways into a network FASN, ACACA, and ACACB show up as central genes significantly downregulated in the transcript dataset and connecting five pathways relating to the metabolism and regulation of fatty acids. Subsequently, metabolic fluxes for the reactions occurring in the adipose tissue during the absorptive state, obtained by flux balance analysis of a publicly available genome-scale metabolic model of adipocytes is mapped to the genes that code for the enzymes that catalyse those reactions. The importance of FASN, ACACA, and ACACB is confirmed further, as the reactions catalysed by the enzymes they code for have flux values in the modelling results. However, it is noteworthy that only 77 reactions out of the 305 active reactions from the metabolic model were mapped to KEGG Reactions, allowing us to combine the information with transcript data. Better annotation of the metabolic model would allow better data integration.

A recent study stated that downregulation of the ACACA metabolic network, that is genes within two reaction steps of ACACA, is an important feature in the pathophysiology of type 2 diabetes in obese individuals [41]. In addition, other studies have stated that chronically activated AMPK suppresses the synthesis of ACACA, FASN (downregulated in current dataset) and other lipogenic enzymes [42-44]. Adolescents with higher quantities of visceral adipose tissue as compared to subcutaneous adipose tissue have down-regulation of genes related to insulin sensitivity (ADIPOQ, GLUT4, PPARG2, and SIRT1) and lipogenesis (SREBP1c, ACC, LPL, and FASN) [45], downregulation of both GLUT4 and FASN is also observed in the current dataset. Downregulation of ELOVL6 (Fold Change: -2.13, P value: 0.008) causes accumulation of Palmitic Acids [46], also indicates lipid dysregulation.

Transcription factors and microRNAs are known to be major regulators of metabolic processes, so we extended the metabolic network with them. The central genes FASN, ACACA, and ACACB were connected to all the fourteen significantly changed pathways by shared regulatory elements. FASN is regulated by four transcription factors NR3C1, SREBF2, MYC, and TAF1. Glucocorticoid receptor gene (NR3C1) is implicated in metabolic syndrome in middle-aged men [47]. SREBP-2 is a comparatively selective activator of cholesterol synthesis, but not fatty acid synthesis, in liver and adipose tissue of mice [48]. MYC is also found to be induced in many post-mitotic tissues, such as adult myocardium where growth proceeds predominantly by an increase in cell size (hypertrophy) rather than number (hyperplasia), in response to stress [49]. This is in accordance with the fact that that adipose tissue in obese individuals stores excess energy as triglycerides in the lipid droplets of adipocytes mainly through hypertrophy to meet the need for additional fat storage capacity in the progression of obesity [50]. Enlarged adipocytes increase lipolysis [51], and increased delivery of free fatty acids to liver has been found to exacerbate insulin resistance promoting dyslipidaemia. Expansion of visceral adipose tissue has also been associated with a lower capacity to transfer cholesteryl esters in reverse cholesterol transport and predicts atherosclerosis [52]. TAF1 possesses an intrinsic protein kinase activity that is essential for cell G1 progression and apoptosis [53], a downregulation as observed in the current dataset indicates the accumulation of damaged cells.

ACACA is regulated by 2 transcription factors E2F1 and STAT1 and 3 microRNAs hsa-mir-335-5p, has-mir-24-3p, and has-let-7b-5p. Tumour progression is associated with downregulation of STAT1 and upregulation of hsa-miR-24-3p is implicated in nasopharyngeal cancer [54]. hsa-mir-335-5p putative role in osteoarthritis [55] and rectal cancer [56]. hsa-miR-24-3p was found positively associated with a hypoxia gene signature in breast cancer [57] and nasopharyngeal carcinoma [54]. hsa-let-7b-5p is implicated in Crohn's disease [58], colorectal cancer [59], and nasopharyngeal carcinoma [60]. ACACB is regulated by has-mir-148b-3p, which is one of the main coordinators of breast cancer progression [61]. Since the microRNAs regulating ACACA and ACACB have been associated with many types of cancer, and the core metabolic network connecting ACACA and ACACB is significantly changed in MAO, this suggests that metabolic dysfunction as a result of obesity and as a result of cancer share some common mechanisms. ACACA is regulated by 2 transcription factors E2F1 and STAT1 and 3 microRNAs hsa-mir-335-5p, has-mir-24-3p, and has-let-7b-5p. Tumour progression is associated with downregulation of STAT1 and upregulation of hsa-miR-24-3p is implicated in nasopharyngeal cancer [54]. hsa-mir-335-5p putative role in osteoarthritis [55] and rectal cancer [56]. hsa-miR-24-3p was found positively associated with a hypoxia gene signature in breast cancer [57] and nasopharyngeal carcinoma [54]. hsa-let-7b-5p is implicated in Crohn's disease [58], colorectal cancer [59], and nasopharyngeal carcinoma [60]. ACACB is regulated by has-mir-148b-3p, which is one of the main coordinators of breast cancer progression [61]. Since the microRNAs regulating ACACA and ACACB have been associated with many types of cancer, and the core metabolic network connecting ACACA and ACACB is significantly changed in MAO, this points to a link between metabolic dysfunction and cancer. Suggests that metabolic dysfunction as a result of obesity and as a result of cancer share some common mechanisms.

Gene ontology analysis of genes in the extended network highlighted processes involved in metabolism, immune response, and transport of lipids as also observed by Fabbrini et al [62]. Immune response and metabolic regulation are highly integrated and the proper function of each is dependent on the other [63]. The close association between immune response and metabolism can be observed by studying common ancestral structures. One such structure is the *Drosophila* fat body, which incorporates the mammalian homologues of the liver and the haematopoietic and immune systems [64, 65], and adipose tissue, sharing similar developmental and functional pathways [66, 67]. Therefore, it is possible to imagine a situation in which common or overlapping pathways regulate both metabolic and immune functions through common key regulatory molecules and signalling systems. This might allow nutrients to act through pathogen-sensing systems such as Toll-like receptors (TLRs), giving rise to metabolically or nutritionally induced inflammatory responses [64, 68-70].

Furthermore, we investigated disease associations for genes in the network of significantly affected processes in metabolically abnormal obese. A large number of Cardiovascular Diseases, Nervous System Diseases, Mental Disorders, Nutritional and Metabolic Diseases, Musculoskeletal Diseases, Neoplasms were associated with genes in the affected metabolic network. Strong associations were found for cardiomyopathy, mental retardation, obesity, and insulin resistance. Mental retardation could be caused due to a dysfunction in the metabolism of the brain [71, 72]. Cardiomyopathy is caused in part by elevations in blood glucose and lipids [73, 74]. The links between cardiomyopathy and insulin resistance [75] and obesity [76] have also been reported. The association of the diseases with the detected significantly altered network for MAO individuals compared to MNO individuals highlights the importance of these genes in context of metabolic abnormalities.

In addition, we investigated drugs that target the genes in the network. FASN is targeted by two drugs, Cerulenin and Orlisat. The latter is an anti-obesity drug, which would be expected for a gene linked to metabolic abnormal obese. However, Cerulenin is currently used as an antifungal agent, although it has been shown to cause dramatic weight loss in rodents by reducing food intake [77]. Our investigation points to Cerulenin being a potential drug for treating metabolic syndrome. Interestingly, the effect of Cerulenin on hepatic function in steatotic obese mice has been recently investigated. Cheng at al observed that Cerulenin treatment markedly improved hepatic function in obese mice, including the increase in hepatic ATP levels, and concurrent decrease in fat content in the hepatocyte [78]. This is interesting as accumulating evidence supports a pathophysiological association Non Alcoholic Fatty Liver Disease and metabolic syndrome as reviewed in [79]. Adenine and biotin, both nutritional supplements, target ACACB and could be of potential therapeutic interest. MAOB is targeted by Zonisamide and pargyline among other drugs mostly for nervous system or mental disorders, indicating the potential use of both drugs for Nervous Systems disorders or mental diseases. The potential benefit of Zonisamide, currently used for partial-onset seizures, or Parkinson's disease has also been proposed in a recent study by Iwata et al [80]. Pargyline, currently used as a cardiovascular drug, is also of therapeutic interest for Parkinson's disease and under current investigation as reviewed in [81].

In addition, numerous drugs not yet known to be associated with metabolic health were found to be targeting genes in the network. Background knowledge about these drugs indicates that they could be candidates for treatment of metabolic syndrome. Fomepizole is a competitive alcohol dehydrogenase inhibitor, part of the catabolism of ethylene glycol and methanol. Mecasermin is a biosynthetic form of human insulin growth factor 1 active in the liver. Mefloquine is an antimalarial agent, overdoses of which may lead to weight loss [82]. Nedocromil inhibits a variety of inflammatory cell types associated with asthma. It prevents activation and release of a variety of inflammatory mediators (e.g. histamine, prostaglandin D2 and leukotrienes c4). Quercetin improves cardiac health. In addition, Quercetin is a non-selective PI 3-kinase inhibitor whose cancer treating capacities have been reported in epidemiological studies[83]. All of these drugs are potential candidates of drugs that could be repurposed for treating metabolic syndrome. Drug repurposing, as opposed to drug discovery, has a significant advantage in having gone through the process of clinical testing and approval already. Repurposed drugs can be accepted for use much quicker than new molecules, once they have been shown to treat a certain ailment. In this case, the above listed drugs, after appropriate testing, could be re-purposed to treat metabolic syndrome.

ELOVL6, ACLY, APOE, FTH1, and SNAP25 were all significantly changed in expression and could be potential drug targets. ACLY has been proposed as a drug target for cancer [35], and APOE has been investigated as a drug target for brain related diseases in other studies But APOE may be too central a regulator and targeting it could have many side effects.

Materials and Methods

Transcriptomics data

Description of the Experiment:

In this study, we used a published and publicly available transcriptomics dataset generated by Fabbrini et al. [26]. Global transcriptional profile in adipose tissue was evaluated before and after moderate (~6%) weight gain using for 18 obese individuals with a mean BMI 36 ± 4 kg/m2. The subjects were classified as metabolically abnormal obese (MAO) if their intrahepatic triglyceride (IHTG) levels were higher than 10% and metabolically normal obese (MNO) if their IHTG content was <5.6%. Increased IHTG content is considered a robust marker of inappropriate fat distribution and metabolic dysfunction [84-86].

Microarray Analysis

The raw data for seven metabolically abnormal obese (MAO) and eleven metabolically normal obese (MNO) individuals, after weight gain was downloaded from the Gene Expression Omnibus (accession number GSE62832). The data were reanalysed with the online microarray quality control and pre-processing pipeline ArrayAnalysis.org [87]. All samples were found to be of good quality and were included for further analysis. Data was normalized using the RMA method.

Statistical analysis was then performed using the statistics module of ArrayAnalysis.org [88] to identify differentially expressed genes between the MAO and MNO individuals after weight gain. The module performs a statistical analysis using the limma package from R/Bioconductor [89], which provides a substantial improvement compared to the standard t-test and is advised for statistical analysis of microarray data [90]. Genes with an absolute fold change (FC) > 1.3 and a p-value < 0.05 were considered to be differentially expressed. Using a combination of cut-offs for fold changes and p-values is shown to give biologically relevant results [91].

Flux data

Description of the Genome Scale Metabolic Model

Bordbar and colleagues have created cell specific models for adipocytes, hepatocytes, and myocytes from Recon1 [33]. In this study, we used the adipocyte model downloaded from the BioModels database (accession: MODEL1111070000). The model contains 649 reactions and 614 species (i.e. metabolites). We used the flux data simulated for the absorptive state by the adipocyte model. The absorptive state is an anabolic process during which absorbed glucose is utilized by the human body to produce glycogen, triacylglycerol, and amino acids [92]. Energy is produced by nutrient utilization to meet immediate needs and excess energy is stored causing weight gain. Therefore during weight gain, the absorptive state must be prolonged, since the increased uptake of nutrients needs to be addressed. Therefore, the study of adipose tissue metabolism during the absorptive state is of key interest.

Flux Balance Analysis

Flux balance analysis was performed using the FAME web-interface [93] to simulate the flux distribution in the adipocyte model during absorptive state. The results from FAME show 305 reactions that were active (absolute flux > 0), of which 77 were identified as KEGG Reactions by the FAME interface. Scripts in R were written for identifier mapping using the interaction identifier mapping and gene identifier mapping databases. The enzyme catalysing each reaction was identified by mapping the reaction identifiers from KEGG to protein identifiers from UniProt. The UniProt enzymes were then mapped to Ensembl genes to get the list of genes active during the absorptive state or weight gain in adipose tissue. 107 genes were identified, 101 of which were also measured in the transcriptomics dataset. These 101 genes were marked active according to the model.

Pathway analysis

Pathway analysis was performed to interpret and visualize the molecular changes on a pathway level using PathVisio [94]. The combined collection of 723 biological pathways from the curated collection of WikiPathways [95] and pathways converted from Reactome [96, 97] was used for overrepresentation analysis. The pathways were ranked based on a standardized difference score (Z-score) calculated using the expected number of differentially expressed genes in a pathway and the standard deviation from that number under a hypergeometric distribution. A Z-score is positive when a pathway contains a greater number of significantly changed genes (in our case absolute fold change > 1.3, p-value < 0.05) than is expected by chance [98]. Pathways were considered significantly changed when (1) Z-score > 1.96, (2) permutated p-value < 0.05 and (3) number of changed genes ≥ 5. Furthermore, the fold changes and p-values of the genes were visualized on the pathway diagrams.

Integrative Network Analysis

All fourteen pathways significantly altered between MAO and MNO after weight gain were combined into a network and visualized with Cytoscape 3.2.1 [99]. The pathways were merged using a Java program, which maps the gene products and metabolites in the pathways to Ensembl and HMDB respectively and merges the pathways into a network. Transcript and flux data was imported and visualized. The network was analysed using inbuilt Cytoscape functionality to calculate edge betweenness values which were then used with the force directed layout method, also part of inbuilt Cytoscape functionality, to layout the network. The CyTargetLinker app [100] was then used to extend the network with miRNAs from mirTarBase[30] and transcription factors from ENCODE [29] and TFe [28]. Next, the ClueGO app was used for Gene Ontology analysis of the integrated network to detect overrepresented biological processes [34]. Subsequently, using the CyTargetLinker app the network was extended further with gene-disease associations from DisGeNET [31] to confirm that the genes in the network are indeed associated with diseases known to affect MAO individuals. Finally, drugs from DrugBank [32] targeting the genes in the network were added using the CyTargetLinker app to detect candidates for drug repurposing. and TFe [28]. Next, the ClueGO app was used for Gene Ontology analysis of the integrated network to detect overrepresented biological processes Subsequently, using the CyTargetLinker app the network was extended further with gene-disease associations from DisGeNET [31] to confirm that the genes in the network are indeed associated with diseases known to affect MAO individuals. Finally, drugs from DrugBank [32] targeting the genes in the network were added using the CyTargetLinker app to detect candidates for drug repurposing.

References

1. Caballero, B., *The global epidemic of obesity: an overview.* Epidemiologic reviews, 2007. **29**(1): p. 1-5.
2. Organization, W.H., *Obesity: preventing and managing the global epidemic*. 2000: World Health Organization.
3. Eckel, R.H., et al., *Obesity and type 2 diabetes: what can be unified and what needs to be individualized?* The Journal of Clinical Endocrinology & Metabolism, 2011. **96**(6): p. 1654-1663.
4. Guo, F. and W.T. Garvey, *Cardiometabolic disease risk in metabolically healthy and unhealthy obesity: Stability of metabolic health status in adults.* Obesity (Silver Spring), 2016. **24**(2): p. 516-25.
5. Dobson, R., et al., *Metabolically healthy and unhealthy obesity: differential effects on myocardial function according to metabolic syndrome, rather than obesity.* International Journal of Obesity, 2015.
6. Hamer, M., G.D. Batty, and M. Kivimaki, *Risk of future depression in people who are obese but metabolically healthy: the English longitudinal study of ageing.* Molecular psychiatry, 2012. **17**(9): p. 940-945.
7. Jokela, M., et al., *Association of metabolically healthy obesity with depressive symptoms: pooled analysis of eight studies.* Molecular psychiatry, 2014. **19**(8): p. 910-914.
8. Kaidanovich-Beilin, O., D.S. Cha, and R.S. McIntyre, *Crosstalk between metabolic and neuropsychiatric disorders.* F1000 Biol Rep, 2012. **4**: p. 14.
9. Talbot, K., et al., *Demonstrated brain insulin resistance in Alzheimer's disease patients is associated with IGF-1 resistance, IRS-1 dysregulation, and cognitive decline.* The Journal of clinical investigation, 2012. **122**(4): p. 1316-1338.
10. Jeon, C.Y., et al., *Helicobacter pylori infection is associated with an increased rate of diabetes.* Diabetes care, 2012. **35**(3): p. 520-525.
11. Witsø, E., *The role of infection-associated risk factors in prosthetic surgery.* Hip International, 2012.
12. Byers, T. and R. Sedjo, *Does intentional weight loss reduce cancer risk?* Diabetes, Obesity and Metabolism, 2011. **13**(12): p. 1063-1072.
13. Spyridopoulos, T.N., et al., *Insulin resistance and risk of renal cell cancer: a case-control study.* Hormones (Athens), 2012. **11**(3): p. 308-315.
14. Chen, S., et al., *Association between metabolically unhealthy overweight/obesity and chronic kidney disease: the role of inflammation.* Diabetes & metabolism, 2014. **40**(6): p. 423-430.
15. Karelis, A. and R. Rabasa-Lhoret, *Inclusion of C-reactive protein in the identification of metabolically healthy but obese (MHO) individuals.* Diabetes & metabolism, 2008. **34**(2): p. 183-184.
16. Stefan, N., et al., *Identification and characterization of metabolically benign obesity in humans.* Archives of internal medicine, 2008. **168**(15): p. 1609-1616.
17. Wildman, R.P., et al., *The obese without cardiometabolic risk factor clustering and the normal weight with cardiometabolic risk factor clustering: prevalence and correlates of 2 phenotypes among the US population (NHANES 1999-2004).* Archives of internal medicine, 2008. **168**(15): p. 1617-1624.
18. Grundy, S.M., et al., *Definition of metabolic syndrome report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on scientific issues related to definition.* Circulation, 2004. **109**(3): p. 433-438.
19. Hausman, D., et al., *The biology of white adipocyte proliferation.* Obesity reviews, 2001. **2**(4): p. 239-254.
20. Singla, P., A. Bardoloi, and A.A. Parkash, *Metabolic effects of obesity: a review.* World J Diabetes, 2010. **1**(3): p. 76-88.
21. Kershaw, E.E. and J.S. Flier, *Adipose tissue as an endocrine organ.* The Journal of Clinical Endocrinology & Metabolism, 2004. **89**(6): p. 2548-2556.
22. Björntorp, P., *The regulation of adipose tissue distribution in humans.* International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity, 1996. **20**(4): p. 291-302.
23. Siiteri, P.K., *Adipose tissue as a source of hormones.* The American journal of clinical nutrition, 1987. **45**(1): p. 277-282.
24. Guerre-Millo, M., *Adipose tissue hormones.* Journal of endocrinological investigation, 2002. **25**(10): p. 855-861.
25. Wagner, A., et al., *Drugs that reverse disease transcriptomic signatures are more effective in a mouse model of dyslipidemia.* Molecular systems biology, 2015. **11**(3): p. 791.
26. Fabbrini, E., et al., *Metabolically normal obese people are protected from adverse effects following weight gain.* The Journal of clinical investigation, 2015. **125**(2): p. 787.
27. Koussounadis, A., et al., *Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system.* Scientific reports, 2015. **5**.
28. Yusuf, D., et al., *The transcription factor encyclopedia.* Genome biology, 2012. **13**(3): p. 1-25.
29. Consortium, E.P., *The ENCODE (ENCyclopedia of DNA elements) project.* Science, 2004. **306**(5696): p. 636-640.
30. Chou, C.-H., et al., *miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database.* Nucleic acids research, 2015: p. gkv1258.
31. Piñero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.* Database, 2015. **2015**: p. bav028.
32. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic acids research, 2008. **36**(suppl 1): p. D901-D906.
33. Bordbar, A., et al. *A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology.* BMC systems biology, 2011. **5**, 180 DOI: 10.1186/1752-0509-5-180.
34. Bindea, G., et al., *ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.* Bioinformatics, 2009. **25**(8): p. 1091-1093.
35. Vander Heiden, M.G., *Targeting cancer metabolism: a therapeutic window opens.* Nature reviews Drug discovery, 2011. **10**(9): p. 671-684.
36. Ye, J. and R.A. DeBose-Boyd, *Regulation of cholesterol and fatty acid synthesis.* Cold Spring Harbor perspectives in biology, 2011. **3**(7): p. a004754.
37. Guiu-Jurado, E., et al., *Downregulation of de Novo Fatty Acid Synthesis in Subcutaneous Adipose Tissue of Moderately Obese Women.* International journal of molecular sciences, 2015. **16**(12): p. 29911-29922.
38. Ortega, F.J., et al., *The gene expression of the main lipogenic enzymes is downregulated in visceral adipose tissue of obese subjects.* Obesity, 2010. **18**(1): p. 13-20.
39. Fernández ‐ Galilea, M., et al., *α ‐ lipoic acid reduces fatty acid esterification and lipogenesis in adipocytes from overweight/obese subjects.* Obesity, 2014. **22**(10): p. 2210-2215.

40. Abu-Elheiga, L., et al., *The subcellular localization of acetyl-CoA carboxylase 2.* Proceedings of the National Academy of Sciences, 2000. **97**(4): p. 1444-1449.
41. Dharuri, H., et al., *Downregulation of the acetyl-CoA metabolic network in adipose tissue of obese diabetic individuals and recovery after weight loss.* Diabetologia, 2014. **57**(11): p. 2384-2392.
42. Hardie, D.G., *Minireview: the AMP-activated protein kinase cascade: the key sensor of cellular energy status.* Endocrinology, 2003. **144**(12): p. 5179-5183.
43. Long, Y.C. and J.R. Zierath, *AMP-activated protein kinase signaling in metabolic regulation.* The Journal of clinical investigation, 2006. **116**(7): p. 1776-1783.
44. Luo, Z., et al., *AMPK, the metabolic syndrome and cancer.* Trends in pharmacological sciences, 2005. **26**(2): p. 69-76.
45. Kursawe, R., et al., *A role of the Inflammasome in the low storage capacity of the abdominal subcutaneous adipose tissue in obese adolescents.* Diabetes, 2015: p. db151478.
46. Chu, X., et al., *Sterol Regulatory Element–Binding Protein-1c Mediates Increase of Postprandial Stearic Acid, a Potential Target for Improving Insulin Resistance, in Hyperlipidemia.* Diabetes, 2013. **62**(2): p. 561-571.
47. Buemann, B., et al., *The N363S polymorphism of the glucocorticoid receptor and metabolic syndrome factors in men.* Obesity research, 2005. **13**(5): p. 862-867.
48. Horton, J.D., et al., *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.* Journal of Clinical Investigation, 1998. **101**(11): p. 2331.
49. Zhong, W., et al., *Hypertrophic growth in cardiac myocytes is mediated by Myc through a Cyclin D2 ‐ dependent pathway.* The EMBO journal, 2006. **25**(16): p. 3869-3879.
50. Faust, I.M., et al., *Diet-induced adipocyte number increase in adult rats: a new model of obesity.* Am J Physiol, 1978. **235**(3): p. E279-86.
51. Engfeldt, P. and P. Arner, *Lipolysis in human adipocytes, effects of cell size, age and of regional differences.* Hormone and metabolic research. Supplement series, 1987. **19**: p. 26-29.
52. Sam, S., et al., *Relationship of abdominal visceral and subcutaneous adipose tissue with lipoprotein particle number and size in type 2 diabetes.* Diabetes, 2008. **57**(8): p. 2022-2027.
53. Li, H.-H., et al., *Phosphorylation on Thr-55 by TAF1 mediates degradation of p53: a role for TAF1 in cell G1 progression.* Molecular cell, 2004. **13**(6): p. 867-878.
54. Ye, S.-b., et al., *Tumor-derived exosomes promote tumor progression and T-cell dysfunction through the regulation of enriched exosomal microRNAs in human nasopharyngeal carcinoma.* Oncotarget, 2014. **5**(14): p. 5439-5452.
55. Tornero-Esteban, P., et al., *Signature of microRNA expression during osteogenic differentiation of bone marrow MSCs reveals a putative role of miR-335-5p in osteoarthritis.* BMC musculoskeletal disorders, 2015. **16**(1): p. 182.
56. Slattery, M.L., et al., *An evaluation and replication of miRNAs with disease stage and colorectal cancer ‐ specific mortality.* International Journal of Cancer, 2015. **137**(2): p. 428-438.
57. Camps, C., et al., *Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia.* Molecular cancer, 2014. **13**(1): p. 1.
58. Guo, Z., et al., *Altered microRNA expression in inflamed and non ‐ inflamed terminal ileal mucosa of adult patients with active Crohn's disease.* Journal of gastroenterology and hepatology, 2015. **30**(1): p. 109-116.
59. Vishnubalaji, R., et al., *Genome-wide mRNA and miRNA expression profiling reveal multiple regulatory networks in colorectal cancer.* Cell death & disease, 2015. **6**(1): p. e1614.
60. Alajez, N., et al., *Enhancer of Zeste homolog 2 (EZH2) is overexpressed in recurrent nasopharyngeal carcinoma and is regulated by miR-26a, miR-101, and miR-98.* Cell death & disease, 2010. **1**(10): p. e85.
61. Cimino, D., et al., *miR148b is a major coordinator of breast cancer progression in a relapse-associated microRNA signature by targeting ITGA5, ROCK1, PIK3CA, NRAS, and CSF1.* The FASEB Journal, 2013. **27**(3): p. 1223-1235.
62. Fabbrini, E., et al., *Metabolically normal obese people are protected from adverse effects following weight gain.* The Journal of Clinical Investigation, 2015. **125**(2): p. 787-795.
63. Hotamisligil, G.S., *Inflammation and metabolic disorders.* Nature, 2006. **444**(7121): p. 860-867.
64. Søndergaard, L., *Homology between the mammalian liver and the Drosophila fat body.* Trends in Genetics, 1993. **9**(6): p. 193.
65. Leclerc, V. and J.M. Reichhart, *The immune response of Drosophila melanogaster.* Immunological reviews, 2004. **198**(1): p. 59-71.
66. Tong, Q., et al., *Function of GATA transcription factors in preadipocyte-adipocyte transition.* Science, 2000. **290**(5489): p. 134-138.
67. Rusten, T.E., et al., *Programmed autophagy in the Drosophila fat body is induced by ecdysone through regulation of the PI3K pathway.* Developmental cell, 2004. **7**(2): p. 179-192.
68. Beutler, B., *Innate immunity: an overview.* Molecular immunology, 2004. **40**(12): p. 845-859.
69. Song, M.J., et al., *Activation of Toll-like receptor 4 is associated with insulin resistance in adipocytes.* Biochemical and biophysical research communications, 2006. **346**(3): p. 739-745.
70. Shi, H., et al., *TLR4 links innate immunity and fatty acid–induced insulin resistance.* The Journal of clinical investigation, 2006. **116**(11): p. 3015-3025.
71. Kahler, S.G. and M.C. Fahey. *Metabolic disorders and mental retardation.* in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*. 2003. Wiley Online Library.
72. Control, C.f.D. and Prevention, *Mental retardation following diagnosis of a metabolic disorder in children aged 3-10 years--metropolitan Atlanta, Georgia, 1991-1994.* MMWR. Morbidity and mortality weekly report, 1999. **48**(17): p. 353.
73. Turan, B. and N.S. Dhalla, *Diabetic Cardiomyopathy: Biochemical and Molecular Mechanisms*. Vol. 9. 2014: Springer Science & Business Media.
74. Banerjee, S. and L.R. Peterson, *Myocardial metabolism and cardiac performance in obesity and insulin resistance.* Current cardiology reports, 2007. **9**(2): p. 143-149.
75. Witteles, R.M. and M.B. Fowler, *Insulin-resistant cardiomyopathy: clinical evidence, mechanisms, and treatment options.* Journal of the American College of Cardiology, 2008. **51**(2): p. 93-102.
76. Alpert, M.A., *Obesity Cardiomyopathy:: Pathophysiology and Evolution of the Clinical Syndrome.* The American journal of the medical sciences, 2001. **321**(4): p. 225-236.
77. Loftus, T.M., et al., *Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors.* Science, 2000. **288**(5475): p. 2379-2381.
78. Cheng, G., et al., *Cerulenin blockade of fatty acid synthase reverses hepatic steatosis in ob/ob mice.* PloS one, 2013. **8**(9): p. e75980.
79. Paschos, P. and K. Paletas, *Non alcoholic fatty liver disease and metabolic syndrome.* Hippokratia, 2009. **13**(1): p. 9-19.

80. Iwata, Y., et al., *Effects of zonisamide on tardive dyskinesia: a preliminary open-label trial.* Journal of the neurological sciences, 2012. **315**(1): p. 137-140.

81. Points, E., *Effects of tocopherol and deprenyl on the progression of disability in early Parkinson's disease.* N Engl j Med, 1993. **1993**(328): p. 176-183.

82. Dow, G., et al., *Mefloquine induces dose-related neurological effects in a rat model.* Antimicrobial agents and chemotherapy, 2006. **50**(3): p. 1045-1053.

83. Russo, G.L., et al., *Quercetin: a pleiotropic kinase inhibitor against cancer*, in *Advances in nutrition and cancer*. 2014, Springer. p. 185-205.

84. Fabbrini, E., et al., *Intrahepatic fat, not visceral fat, is linked with metabolic complications of obesity.* Proceedings of the National Academy of Sciences, 2009. **106**(36): p. 15430-15435.

85. Korenblat, K.M., et al., *Liver, muscle, and adipose tissue insulin action is directly related to intrahepatic triglyceride content in obese subjects.* Gastroenterology, 2008. **134**(5): p. 1369-1375.

86. Deivanayagam, S., et al., *Nonalcoholic fatty liver disease is associated with hepatic and skeletal muscle insulin resistance in overweight adolescents.* The American journal of clinical nutrition, 2008. **88**(2): p. 257-262.

87. Eijssen, L.M., et al., *User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis. org.* Nucleic acids research, 2013. **41**(W1): p. W71-W76.

88. Dutta, A., *Adding automated Statistical Analysis and Biological Evaluation modules to www. arrayanalysis. org*. 2011, Maastricht University.

89. Berkeley, C., *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.* E-book available at http://www. bepress. com/sagmb/vol3/iss1/art3.[PubMed], 2004.

90. Jeanmougin, M., et al., *Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies.* PloS one, 2010. **5**(9): p. e12336.

91. Grigoryev, D.N., et al., *Orthologous gene-expression profiling in multi-species models: search for candidate genes.* Genome Biol, 2004. **5**(5): p. R34.

92. Marieb, E.N. and K. Hoehn, *Human anatomy & physiology.* 2007: Pearson Education.

93. Boele, J., B.G. Olivier, and B. Teusink, *FAME, the flux analysis and modeling environment.* BMC systems biology, 2012. **6**(1): p. 1.

94. Kutmon, M., et al., *PathVisio 3: an extendable pathway analysis toolbox.* PLoS computational biology, 2015. **11**(2).

95. Kutmon, M., et al., *WikiPathways: capturing the full diversity of pathway knowledge.* Nucleic Acids Res, 2015.

96. Croft, D., et al., *The Reactome pathway knowledgebase.* Nucleic Acids Res, 2014. **42**(Database issue): p. D472-7.

97. Milacic, M., et al., *Annotating cancer variants and anti-cancer therapeutics in reactome.* Cancers, 2012. **4**(4): p. 1180-1211.

98. Doniger, S.W., et al., *MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.* Genome biol, 2003. **4**(1): p. R7.

99. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-2504.

100. Kutmon, M., et al., *CyTargetLinker: a cytoscape app to integrate regulatory interactions in network analysis.* PloS one, 2013. **8**(12): p. e82160.

CHAPTER 7

General Discussion

Cells have evolved the impressive capability to modify organic molecules to make other useful organic molecules. Series of such metabolic modifications, largely mediated by enzymes, constitute cell metabolism. Cell metabolism is highly regulated, and many other gene products are involved in regulation of the metabolic activity. Fora systems-level description of metabolism, one needs to know

1. the identity and the nature of the components that constitute the biological system,
2. the dynamic behaviour of these components (i.e., how their abundance or activity changes over time in various conditions), and
3. the interactions among these components [1].

Ultimately, this information can be combined into *in silico* models that are based on current knowledge of the molecular mechanisms of metabolism. Using these models can provide new insights and predictions [2]. Such *in silico* model could then, for instance, be used to evaluate the molecular mechanisms of human disease development [3, 4].

This thesis describes the integration of gene and metabolite measurements with modelled and measured metabolic flux measurements to obtain a more complete picture of the molecular mechanisms inside a cell. The tools developed allow the visualization of metabolic fluxes on the interactions of pathway diagrams created in PathVisio, imported from WikiPathways, or on the graphical representation of the mathematical model itself. Alongside the gene, protein, and transcript level information visualized on the nodes.

Big data in health research

There is much attention to data problems in modern health research, problems that are often coined the term "big data". Big data in biology is not just about size. Biological data can be complex, noisy, heterogeneous, and longitudinal, in addition to being voluminous [5]. Size is a problem, although the European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 10^{15} bytes) of data and back-ups about genes, proteins, and small molecules. Genomic data account for two petabytes of that, a number that more than doubles every year [6]. The challenge in healthcare is poignantly illustrated in a quote from Elaine Grant, writing online for the Harvard School of Public Health [7]:

> ***"Petabytes of raw information could provide clues for everything from preventing TB to shrinking health care costs—if we can figure out how to use them."***

Microarray technologies, recent advances in protein mass spectrometry, and high-throughput metabolite analyses provide detailed information on genes, proteins, and metabolites in a sample. However, to elucidate the functional state of the organism, these heterogeneous knowledge bases must be combined. Genes, proteins, and metabolites interact with each other, forming biological pathways, e.g. glycolysis, which converts glucose to pyruvate. The TCA cycle uses pyruvate to produce reducing power which ultimately drives the mitochondrial electron transport chain to produce energy. Biological pathways help us break down the vast amount of data provided by the various high throughput technologies into chunks of related information. These pathways can then serve as a launch pad for combination with other information sources, e.g. the epigenome, genetic variations, transcription factors, and microRNAs which have been shown to be important metabolic regulators in many recent studies.

The combination of all these data is only possible by correct annotation of the elements with identifiers from online databases. For instance, all available information about a certain gene or protein can be obtained by mapping their identifiers to each other, which. This also enables the longitudinal combination of gene expression, protein abundances, or other experimental data from different experiments. Identifiers can also be used to combine the information about a gene, the enzyme it transcribes, and the metabolic reaction that enzyme catalyses, to obtain a functional interpretation of individual gene, protein, and metabolite measurements. When combining such information for thousands of genes, the many thousands more proteins they transcribe, and the metabolic reactions they catalyse, identifiers are indispensable. Effective data integration is only possible if every element is annotated using database identifiers, and if mappings between these identifiers are available.

Identifiers are important in unifying data about the same elements from different sources. They are indispensable for a holistic approach, such as the one often taken in systems biologyin systems biology approaches, going beyond the individual gene to instead analyse pathways and networks. This approach is fundamental to elucidate the underlying molecular mechanisms of human health and disease. Biological pathways are the diagrammatic representation of life processes describing the complex interplay between genes, proteins, metabolites, in a series of reactions. The proper functioning of biological pathways leads to a healthy state and a dysfunction in these reaction chains leads to the development of diseases.

Organizing and leveraging current knowledge

Biologists often describe things like complex series of metabolic reactions and regulatory processes in biological pathways that can be represented graphically. In fact, the literature and even textbooks are full of these. Pathways are precious in their ability to organise biological knowledge which can be leveraged as frameworks for integration, analysis, and visualization of high-throughput genomic data, such as transcriptomics, proteomics, metabolomics, and more recently metabolite fluxes. Metabolic fluxes represent the rate of turnover of metabolites through chemical reactions. They are difficult to measure but are commonly predicted by mathematical models. A metabolic flux is regulated by one or more enzymes catalysing the reaction. Enzyme activities are the result of the transcriptional, post—transcriptional, translational, and post-translational changes of the protein involved. Visualizing metabolic fluxes alongside transcript, protein and metabolite abundances allows further evaluation of the pathway.

As a first step to link data to the interactions in a pathway, the data and the entities in the pathway must both be annotated. In **Chapter 2**, I described the third version of our pathway analysis software tool PathVisio, which now allows annotation of interactions in pathway diagrams. As WikiPathways and PathVisio shared the same core software to edit pathway diagrams, the developments described in this Chapter enable interactions to be annotated online with the WikiPathways pathway editor as well. Annotating interactions enable linking them to online databases containing more information about these interactions. In this way, the pathway diagrams are enriched further as annotations for interactions can be added alongside annotations for genes, proteins, and metabolites.

Furthermore, **Chapter 2** describes the reorganization of the PathVisio software architecture which simplifies the introduction of new features as an extension to the core software. These extensions are known as “plugins.” Many plugins have already been made freely available from the PathVisio plugin repository. Allowing functionalities for simplifying the creation and annotation of pathways (e.g. MappBuilder: for creating pathways from gene lists), visualizing data on them (e.g. IntViz: that was introduced in this thesis, for visualizing data on interactions), and performing pathway statistics (e.g. GSEA: for performing gene-set enrichment statistics).

Additionally, a new plugin FindYourInteractions was developed to simplify annotating interactions in PathVisio (described in **Chapter 5**). The plugin automatically searches the Rhea database for identifiers for the interaction selected in the pathway diagram. The Rhea database is a freely accessible resource of manually curated reactions [8]. All reactions in Rhea are manually annotated and chemically balanced where possible. The FindYourInteractions plugin recognises the interactions based on the ChEBI/UniProt identifiers or text labels of the nodes connected by the interaction.

Automating analyses

In **Chapter 3,** I introduce PathVisioRPC, the XMLRPC interface for PathVisio, which enables remote access to PathVisio from all the major programming languages. Allowing pathway analysis using WikiPathways and Reactome pathways, in the same environment used for primary data analysis. Primary data analysis is often performed in R, so we developed an additional R package, RPathVisio, for simplifying the utilization of the PathVisioRPC interface in R. The access to PathVisio functionality through scripting languages opens up a world of possibilities in the context of automation. Users can script their pathway analysis workflow as described in the chapter, automating the analysis and visualization processes. Helping to maintain an exact record of the analysis performed, and it ensures reproducibility, in addition to saving time. Furthermore, functionalities of PathVisio can be easily integrated into other current data analysis pipelines allowing re-use of code, as illustrated by developing a pathway analysis module for the online microarray data analysis pipeline arrayanalysis.org.

In addition, using PathVisioRPC analysis done in PathVisio can be easily combined with data analysis packages already available in the programming environment of choice. In systems biology approaches, it is vital to combine various analytical techniques to get the right answer. For instance, in R there are numerous bioinformatics data analysis packages available from Bioconductor, which can now be coupled with PathVisio analysis. For example in **Chapter 3**, we combine topGO analysis with pathways analysis. Gene Ontology enrichment using the topGO package in R is performed on the transcriptomics dataset, the overrepresented Gene Ontology terms are represented in a pathway like format, and the data of the genes is visualized on the data nodes representing the Gene Ontology term. Highlighting this can not only point out the most highly affected genes in one glance but also show in which direction they are affected.

Connecting the fragmented wealth of biological connections

Pathway knowledge is fragmented over numerous pathway databases. A recent review has named Reactome, Kyoto Encyclopaedia of Genes and Genomes (KEGG), WikiPathways, Nature Pathway Interaction Database (PID) and Pathway Commons as the most preferred sources [9]. We develop and maintain WikiPathways in collaboration with Alex Pico's group at the Gladstone Institutes in California. WikiPathways operates on an open wiki-based philosophy. Registration is free to all, and all registered users can contribute and curate biological pathway knowledge. Reactome is an expert-curated pathway database maintained at EBI. Reactome mainly focusses on the organization of reactions into pathway diagrams. These diagrams are related hierarchically, simplifying the study of the connections between biological pathways. Both interactions (i.e. enzymatic reactions, transport, signalling) and nodes (i.e. genes, proteins, and metabolites), are annotated in Reactome pathway diagrams. **Chapter 4** describes a collaborative effort by the WikiPathways and Reactome teams to convert Reactome pathways to the latest GPML format. Annotations of all data nodes and interactions as well as all literature references are preserved. The converted pathways are made available through the Reactome portal on WikiPathways and can be used for analysis. The conversion of Reactome pathways added 430 pathways to the already 273 pathways of the WikiPathways analysis collection of human pathways, and can now be used for pathway analysis and data visualization in PathVisio. The interactions on the Reactome pathways are annotated with database identifiers allowing the mapping of modelled and measured data onto them. Furthermore, Reactome pathways are rich in protein information, and tend to contain a large number of complexes; we developed a PathVisio plugin for PathVisio, ComplexViz to allow visualization of data on complexes. The data associated with the components of a complex is summarised into a score for the complex based on a user defined criterion. This score can then be visualized on the complex nodes. Thus the complexes with many components significantly changed are highlighted. The ComplexViz plugin also allows easy navigation through complex components and associated data by listing the elements of the selected complex in a side tab as a mini-map on which the data uploaded for each element is visualized.

Mathematical models describe the biological processes which are described in pathways and networks, but as a series of mathematical equations. However, it is often difficult to modify such models. To simplify model correction or extension, and to enable biologists and modellers to leverage each other's knowledge, it is imperative to represent the models in aa form that people can understand better, and that allows data integration. Graphical, visual formats have proven themselves as the best choice for such applications. Various standard formats for representing models in a graphical format exist; notable among these is Systems Biology Graphical Notation (SBGN). We developed a plugin PathSBML for our pathway analysis software PathVisio, to import mathematical models as SBGN compliant graphical diagrams, described in **Chapter 5**. Enabling modellers to obtain up-to-date diagrams of their mathematical models based on the standardised Systems Biology Graphical Notation [10]. The graphical representation facilitates model correction by highlighting bottlenecks and unreachable reactions and so forth by making parts of pathways that are lumped into "black boxes" in models explicitly visible. This representation can then be compared to existing pathways for the same process, which will facilitate both pathway improvement and critical assessment of model implementation aspects like lumping of reaction steps and reduction of parallel routes.

The graphical representations of the mathematical models can also be used as frameworks for multi-omics data integration. Many mathematical models are available freely through online databases, notably the BioModels database. The BioModels Database is an online repository of mathematical models [11]. The models available are chosen from literature and manually curated with database references. The PathSBML plugin enables direct import of mathematical models from BioModels Database as pathway diagrams providing additional knowledge frameworks for data integration and visualization. Furthermore, imported models can then be shared with the community for distribution and further curation through WikiPathways. As a proof of principle, we converted a few mathematical models and uploaded them to WikiPathways as listed below in Table 7.1.

Table 7.1 Mathematical models from Biomodels.org converted into pathway diagrams and uploaded into WikiPathways.

Name	Type	Publication	BioModels ID	WikiPathways ID
Brännmark2013 - Insulin signalling in human adipocytes (normal condition)	Signalling pathway	Brannmark, *et al*[12]	BIOMD0000000448	WP3634
Brännmark2013 - Insulin signalling in human adipocytes (diabetic condition)	Signalling pathway	Brannmark, *et al* [12]	BIOMD0000000449	WP3635
Teusink1998_Glycolysis_TurboDesign	Metabolic pathway	Teusink *et al*[13]	BIOMD0000000253	WP3636

The full picture

As we introduced before, the fragmented biological knowledge over numerous online biochemical databases must be combined to enable exploratory data analysis. Each online database typically uses its unique identifier to annotate the genes, proteins, metabolites, and interactions it records. To efficiently combine the data about a single entity the identifiers from various online data sources must be mapped to each other. BridgeDb is an open source identifier mapping framework; it can be used with identifier mapping databases. Such databases can be created for instance by harvesting other meta resources that provide identifiers between different database systems [14]. PathVisio uses BridgeDb to map data onto the pathway elements.

Chapter 5 also describes an identifier mapping database for interactions, based on the BridgeDb framework, created using the identifier mappings freely provided by the Rhea database [8]. The Rhea database is a freely accessible resource of manually curated reactions. The database also provides identifier mapping files for interactions, mapping the identifier from the Rhea database to EC numbers (Enzyme Classification: IntEnz) [15], MetaCyc reactions [16], KEGG reactions [17], UniPathway enzymatic reactions and chemical reactions [18], MACiE (Mechanism, Annotation and Classification in Enzymes) [19], Reactome [20, 21], and UniProt protein entries [22]. Using the identifier mapping database annotated interactions in pathway diagrams can be connected to various online databases (listed above) that provide more information about them.

The interaction identifier mapping database also enables the mapping of data onto interactions. Another PathVisio plugin, IntViz, was developed and has been described in **Chapter 5**, which allows the visualization of data on interactions using colours and line thickness. Time-series data can be dynamically visualized using a slider. The identifier mapping database and the IntViz plugin enables the visualization modelled and measured fluxes on pathway diagrams in PathVisio. The pathway diagrams can be obtained from WikiPathways. Alternatively, the model could be converted using the PathSBML plugin for data visualization. The newly developed set of tools, therefore, enables the dynamic visualization of the model outputs on the models themselves. Visualisation of fluxomics results in combination with transcriptomics and metabolomics data allows a complete overview of all key players in a metabolic pathway. As a pedagogic example, measured transcript data and fluxes modelled by a genome-scale metabolic model for photosynthesis are visualized together on the nodes and interactions of the primary metabolism related pathways of *Arabidopsis thaliana* in **Chapter 5**.

Active metabolic networks

Pathways significantly affected in a certain condition are combined into networks to study the link between them. Molecular networks are invaluable in highlighting links between processes and their individual elements and are commonly used to integrate and interpret large-scale datasets. In **Chapter 6**, we study the differences in metabolism of the adipose tissue in the absorptive state, after a moderate amount of weight gain, in metabolically abnormal obese individuals compared to metabolically normal obese individuals. For this purpose, obese people with intra-hepatocyte triglyceride levels > 10 % were considered metabolically abnormal and those with intra-hepatocyte triglyceride levels < 5.6 % were categorised as metabolically normal, the same categorisation that was done by Fabbrini *et al.* [23]., the original authors. We combine publicly available transcriptomics data and fluxes predicted by a publicly available metabolic model to highlight genes of interest.

Fourteen pathways were found to be significantly affected in metabolically abnormal obese compared to metabolically normal obese after weight gain. Most of the pathways were related to metabolism and its regulation; one was related to transport and one to disease. Significantly changed pathways were combined into a network. An active subnetwork of five pathways regarding fat metabolism and its regulation, connected by FASN, ACACA, and ACACB, were found to be downregulated in the transcriptomics data, and these reactions are predicted by the model to be carrying flux during weight gain. Upon extending the network with regulatory information about transcription factors and micro RNAs, all fourteen significantly affected pathways were found to be linked to *FASN* (Fatty Acid Synthase), *ACACA* (Acetyl CoA Carboxylase, alpha), and ACACB (Acetyl CoA Carboxylase, beta). Furthermore, upon extension with disease associations, genes in the network were found to be associated with mainly cardiovascular, nervous system, mental, and musculoskeletal diseases. Furthermore, drug target analysis showed Cerulenin and Nedocromil, as potential drugs for metabolic illness and obesity.

Data driven generation of networks

Networks can be generated based on prior knowledge as described in **Chapter 6** or by data driven procedures. The current boom in big data makes it especially interesting to study molecular mechanisms from a data driven perspective. Signal transduction pathways are known to be metabolic regulators. Cells regulate life processes such as growth, survival, apoptosis (cell death), and migration in response to their environment by activating and deactivating the relevant pathways. As signalling happens functionally at the level of proteins, data driven approaches to infer signalling pathways, use proteome level data.

Post-translational modifications, notably phosphorylation, play a vital role in the regulation of metabolism by altering signalling cascades. In cancer cells, signalling networks frequently become compromised, leading to abnormal behaviours and responses to external stimuli. Data-driven learning of regulatory connections in molecular networks has long been a key topic in computational biology [24-29]. An emerging notion is that networks describing a certain biological process (e.g., signal transduction or gene regulation) may depend on biological contexts such as cell type, tissue type and disease state [30, 31], motivating efforts to elucidate context-specific molecular networks [32-37]. Therefore, many current and emerging cancer treatments are designed to block nodes in signalling networks. Although there is a wealth of literature describing canonical cell signalling networks, little is known about exactly how these networks operate in different cancer cells. Advancing our understanding of how these networks are deregulated across cancer cells will ultimately lead to more efficient treatment strategies for patients. We used datasets provided by the Heritage-DREAM Breast cancer network inference challenge to infer causal signalling networks in cancer. This work is presented as an intermezzo below. The resulting network could be used to apply the approaches described in this thesis, which was in fact done by my colleagues when they contributed to a next DREAM challenge on drug treatment combinations in prostate cancer.

The next section is based on our submission to the HPN Dream Breast Cancer inference challenge. A manuscript summarizing the challenge results was published in Nature Methods [38].

Inferring causal relationships

A causal link is expected when the change in expression of a parent node over time causes a concomitant shift in the expression of the child node (Figure 7.1).

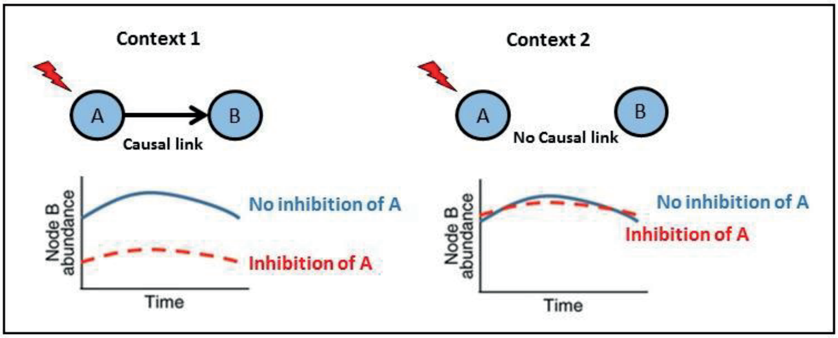


Figure 7.1 Causal relationships in different contexts. A directed edge denotes that inhibition of the parent node A can change the abundance of the child node B. Figure adapted from Hill et al. [38].

The training datasets, containing normalized protein abundance measurements for *in silico* and experimental data were read in R as data lists. The *in silico* relationship between stimuli, inhibitors, and their target proteins provided in the Synapse Wiki was used to create another R data list. For experimental data, we obtained a prior network of relationships from the literature [39-41]. We imputed missing data points in the experimental data set by averaging neighbouring data points. We did not perform imputation for cases where missing data were at the first or the last position in the time series, or there was more than one data point absent in a row. Replicates in experimental data were averaged.

The protein signalling networks are inferred by comparing the protein expression levels in all parent-child node combinations, upon inhibition of the parent node compared to no inhibition. Nodes targeted by the treatments were defined as the parent nodes. Area Between Curves (ABC) integrated over all time points was obtained for every parent node in combination with all other nodes in the network. We normalize the ABC values for each child on a scale of 0 to 1, by obtaining a ratio between its ABC value and the maximum ABC value obtained for any child in the network. In the final network, only those edges were included which scored higher than 0.1. This arbitrary threshold score was decided upon observation of the expression graphs plotted for each of the two conditions for a larger number of parent-child edges. Thresholds greater than 0.1, appeared to describe visually different curves.

This method was implemented in R using the "simp" function from the StreamMetabolism R package [42]. All R scripts are available from GitHub at https://github.com/gungorbudak/netinf-bigcat/

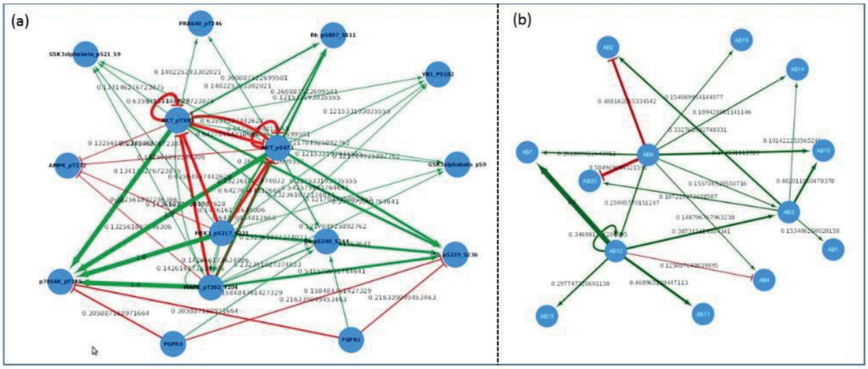


Figure 7.2 Causal networks identified. (a) from experimental data, (b) from in-silico data

The causal signalling networks inferred from the provided experimental and *in silico* datasets are shown in Figure 7.2.

Open Science

Open data, open source, and collaborative development are indispensable for biological discovery. Microarray studies are widely shared through portals such as ArrayExpress [43] and Gene Expression Omnibus [44]. Below I describe how open science has been used in this thesis.

In **Chapters 3 and 4**, we analyse publicly available transcriptomics datasets from GEO to explore effects of cyclophosphamide on tumour bearing mice [45] and discover significantly affected biological processes in women with polycystic ovary syndrome [46]. In **Chapter 5** we visualize a transcriptomics dataset [47] along with metabolite fluxes predicted by a genome-scale metabolic model of Arabidopsis (AraGEM) [48] to visualize what happens during photosynthesis. In **Chapter 6** we investigate what happens during weight gain in the metabolically unhealthy obese, also combining publicly available transcriptomics data [23] and metabolic fluxes predicted by a publicly available genome-scale metabolic model of the adipose tissue [49]. In addition, all analyses were performed using open source and freely available software. Statistical analysis was performed in R with available Bioconductor packages. Following which pathway analysis was performed using RPathVisio [50], while tools like RCytoscape [51] could be used to perform network analysis in R, and gene ontology analysis can be conducted using many Bioconductor packages such as topGO [52], as illustrated in **Chapter 3**. Data from various public databases about microRNAs and transcription factors were combined in Chapter 6. Gene-disease associations are investigated by combining data from DisGeNET [53] followed by drug-target interactions obtained from DrugBank [54] all freely available data sources.

Pathway analysis was performed with the freely available tool PathVisio [55], the three plugins for visualizing flux data on pathway diagrams and pathway representations of metabolic models were developed as part of the Google Summer of Code Programme, an open source software development community. The PathSBML plugin connects to yet another freely available database, BioModels Database, which houses mathematical models. These models can be downloaded and converted into pathway diagrams to ease model correction and comparison. These diagrams can also be used for data integration. The identifier mapping framework, for mapping data onto interactions, was developed based on BridgeDb yet another open source identifier mapping framework, using freely available interaction mapping data from the Rhea database [8]. Network analysis was performed in Cytoscape [56], another freely available network analysis tool, using various plugins contributed by other researchers. Flux Balance Analysis was performed using the online Flux Analysis and Modelling Environment (FAME) [57], also freely available.

We thereby demonstrate the value for open access data and open source software development. Open access data allow researchers to re-analyse experimental data. Combining large-scale datasets has been widely accepted as the way leading to biological discovery. Open access data publishing is crucial to that cause. In addition, open source software development allows developers worldwide to contribute to a software, leading to the creation of a vast arsenal of knowledge and tools in a short span of time. Premier examples include the pathway database WikiPathways and the network analysis tool Cytoscape. WikiPathways, since its launch in 2008, has grown by leaps and bounds. It currently contains 2471 pathways from about 30 species. Specialized communities have contributed knowledge in their fields as comprehensive sets of pathway diagrams which are available from "portals". **Chapter 4** describes the new Reactome portal at WikiPathways providing Reactome human pathways. It also describes a significant addition to the Plant portal by converting pathways from Plant Reactome. Other notable examples include the open source tools PathVisio and Cytoscape. Many plugins have been contributed to each of these tools by global participation. The Cytoscape wall of apps boasts an impressive suite of 270 apps contributed by people worldwide (Figure 7.3).

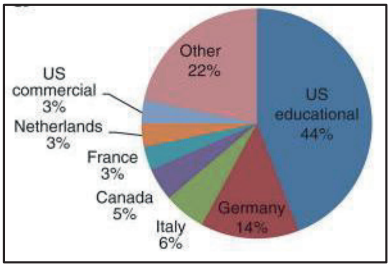


Figure 7.3 Pie Chart showing the Percentage of Cytoscape apps developed by individuals from different countries. The figure is taken from [58].

Conclusion

The tools and approaches described in this thesis help us to gain a better understanding of metabolic processes by annotating the involved reactions, connecting them to external online databases, and visualizing data on them. They also help to bridge the gap between mathematical models and biological pathways and thereby allow better collaboration between biologists and mathematical modellers, each of whom have their own view on biological data and knowledge

References

1. Kitano, H., **Systems biology: a brief overview**. *Science*, 2002. **295**(5560): p. 1662-1664.
2. Albert, R., **Network inference, analysis, and modeling in systems biology**. *The Plant Cell*, 2007. **19**(11): p. 3327-3338.
3. Mardinoglu, A., R. Agren, C. Kampf, A. Asplund, M. Uhlen, and J. Nielsen, **Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease**. *Nature communications*, 2014. **5**.
4. Yizhak, K., S.E. Le Dévédec, V.M. Rogkoti, F. Baenke, V.C. de Boer, C. Frezza, A. Schulze, B. van de Water, and E. Ruppin, **A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration**. *Molecular systems biology*, 2014. **10**(8): p. 744.
5. Mattmann, C.A., **Computing: A vision for data science**. *Nature*, 2013. **493**(7433): p. 473-475.
6. Marx, V., **Biology: The big challenges of big data**. *Nature*, 2013. **498**(7453): p. 255-260.
7. Grant, E. **The promise of big data**. [cited 2016; Available from: http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/.
8. Morgat, A., K.B. Axelsen, T. Lombardot, R. Alcantara, L. Aimo, M. Zerara, A. Niknejad, E. Belda, N. Hyka-Nouspikel, E. Coudert, N. Redaschi, L. Bougueleret, C. Steinbeck, I. Xenarios, and A. Bridge, **Updates in Rhea--a manually curated resource of biochemical reactions**. *Nucleic Acids Res*, 2015. **43**(Database issue): p. D459-64.
9. Bauer-Mehren, A., L.I. Furlong, and F. Sanz, **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Molecular systems biology*, 2009. **5**(1): p. 290.
10. Le Novere, N., S. Moodie, A. Sorokin, M. Hucka, F. Schreiber, E. Demir, H. Mi, Y. Matsuoka, K. Wegner, and H. Kitano, **Systems biology graphical notation: process diagram level 1**. *Nature Precedings*, 2008.
11. Le Novere, N., B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, and B. Shapiro, **BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems**. *Nucleic acids research*, 2006. **34**(suppl 1): p. D689-D691.
12. Brannmark, C., E. Nyman, S. Fagerholm, L. Bergenholm, E.M. Ekstrand, G. Cedersund, and P. Stralfors, **Insulin signaling in type 2 diabetes: experimental and modeling analyses reveal mechanisms of insulin resistance in human adipocytes**. *J Biol Chem*, 2013. **288**(14): p. 9867-80.
13. Teusink, B., M.C. Walsh, K. van Dam, and H.V. Westerhoff, **The danger of metabolic pathways with turbo design**. *Trends Biochem Sci*, 1998. **23**(5): p. 162-9.
14. van Iersel, M.P., A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo, **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC bioinformatics*, 2010. **11**(1): p. 5.
15. Fleischmann, A., M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K.B. Axelsen, A. Bairoch, D. Schomburg, K.F. Tipton, and R. Apweiler, **IntEnz, the integrated relational enzyme database**. *Nucleic acids research*, 2004. **32**(suppl 1): p. D434-D437.
16. Caspi, R., R. Billington, L. Ferrer, H. Foerster, C.A. Fulcher, I.M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D.S. Weaver, and P.D. Karp, **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res*, 2015.
17. Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, **KEGG as a reference resource for gene and protein annotation**. *Nucleic acids research*, 2016. **44**(D1): p. D457-D462.
18. Morgat, A., E. Coissac, E. Coudert, K.B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari, **UniPathway: a resource for the exploration and annotation of metabolic pathways**. *Nucleic acids research*, 2011: p. gkr1023.
19. Holliday, G.L., D.E. Almonacid, G.J. Bartlett, N.M. O'Boyle, J.W. Torrance, P. Murray-Rust, J.B. Mitchell, and J.M. Thornton, **MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms**. *Nucleic acids research*, 2007. **35**(suppl 1): p. D515-D520.
20. Milacic, M., R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, and L. Stein, **Annotating cancer variants and anti-cancer therapeutics in reactome**. *Cancers*, 2012. **4**(4): p. 1180-1211.
21. Croft, D., A.F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, and M.R. Kamdar, **The Reactome pathway knowledgebase**. *Nucleic acids research*, 2014. **42**(D1): p. D472-D477.
22. Consortium, U., **UniProt: a hub for protein information**. *Nucleic acids research*, 2014: p. gku989.
23. Fabbrini, E., J. Yoshino, M. Yoshino, F. Magkos, C.T. Luecking, D. Samovski, G. Fraterrigo, A.L. Okunade, B.W. Patterson, and S. Klein, **Metabolically normal obese people are protected from adverse effects following weight gain**. *The Journal of clinical investigation*, 2015. **125**(2): p. 787.
24. Bansal, M., V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo, **How to infer gene networks from expression profiles**. *Molecular systems biology*, 2007. **3**(1): p. 78.
25. Markowetz, F. and R. Spang, **Inferring cellular networks–a review**. *BMC bioinformatics*, 2007. **8**(Suppl 6): p. S5.
26. Hecker, M., S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, **Gene regulatory network inference: data integration in dynamic models—a review**. *Biosystems*, 2009. **96**(1): p. 86-103.
27. De Smet, R. and K. Marchal, **Advantages and limitations of current network inference methods**. *Nature Reviews Microbiology*, 2010. **8**(10): p. 717-729.

28. Marbach, D., R.J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, **Revealing strengths and weaknesses of methods for gene network inference**. *Proceedings of the National Academy of Sciences*, 2010. **107**(14): p. 6286-6291.
29. Maetschke, S.R., P.B. Madhamshettiwar, M.J. Davis, and M.A. Ragan, **Supervised, semi-supervised and unsupervised inference of gene regulatory networks**. *Briefings in bioinformatics*, 2013: p. bbt034.
30. Ideker, T. and N.J. Krogan, **Differential network biology**. *Molecular systems biology*, 2012. **8**(1): p. 565.
31. de la Fuente, A., **From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases**. *Trends in genetics*, 2010. **26**(7): p. 326-333.
32. Hill, S.M., Y. Lu, J. Molina, L.M. Heiser, P.T. Spellman, T.P. Speed, J.W. Gray, G.B. Mills, and S. Mukherjee, **Bayesian inference of signaling network topology in a cancer cell line**. *Bioinformatics*, 2012. **28**(21): p. 2804-2810.
33. Saez-Rodriguez, J., L.G. Alexopoulos, M. Zhang, M.K. Morris, D.A. Lauffenburger, and P.K. Sorger, **Comparing signaling networks between normal and transformed hepatocytes using discrete logical models**. *Cancer research*, 2011. **71**(16): p. 5400-5411.
34. Akbani, R., P.K.S. Ng, H.M. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.-Y. Yang, K. Yoshihara, and J. Li, **A pan-cancer proteomic perspective on The Cancer Genome Atlas**. *Nature communications*, 2014. **5**.
35. Chen, W.W., B. Schoeberl, P.J. Jasper, M. Niepel, U.B. Nielsen, D.A. Lauffenburger, and P.K. Sorger, **Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data**. *Molecular systems biology*, 2009. **5**(1).
36. Molinelli, E.J., A. Korkut, W. Wang, M.L. Miller, N.P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, and D.B. Solit, **Perturbation biology: inferring signaling networks in cellular systems**. *PLoS Comput Biol*, 2013. **9**(12): p. e1003290.
37. Eduati, F., J. De Las Rivas, B. Di Camillo, G. Toffolo, and J. Saez-Rodriguez, **Integrating literature-constrained and data-driven inference of signalling networks**. *Bioinformatics*, 2012. **28**(18): p. 2311-2317.
38. Hill, S.M., L.M. Heiser, T. Cokelaer, M. Unger, N.K. Nesser, D.E. Carlin, Y. Zhang, A. Sokolov, E.O. Paull, and C.K. Wong, **Inferring causal molecular networks: empirical assessment through a community-based effort**. *Nature methods*, 2016.
39. Rhodes, N., D.A. Heerding, D.R. Duckett, D.J. Eberwein, V.B. Knick, T.J. Lansing, R.T. McConnell, T.M. Gilmer, S.-Y. Zhang, and K. Robell, **Characterization of an Akt kinase inhibitor with potent pharmacodynamic and antitumor activity**. *Cancer research*, 2008. **68**(7): p. 2366-2374.
40. Gilmartin, A.G., M.R. Bleam, A. Groy, K.G. Moss, E.A. Minthorn, S.G. Kulkarni, C.M. Rominger, S. Erskine, K.E. Fisher, and J. Yang, **GSK1120212 (JTP-74057) is an inhibitor of MEK activity and activation with favorable pharmacokinetic properties for sustained in vivo pathway inhibition**. *Clinical Cancer Research*, 2011: p. clincanres. 2200.2010.
41. Pardo, O.E., J. Latigo, R.E. Jeffery, E. Nye, R. Poulsom, B. Spencer-Dene, N.R. Lemoine, G.W. Stamp, E.O. Aboagye, and M.J. Seckl, **The fibroblast growth factor receptor inhibitor PD173074 blocks small cell lung cancer growth in vitro and in vivo**. *Cancer research*, 2009. **69**(22): p. 8645-8651.
42. Sefick Jr, S., **Stream Metabolism-A package for calculating single station metabolism from diurnal Oxygen curves**. *R package version 0.03-3*, 2009.
43. Kolesnikov, N., E. Hastings, M. Keays, O. Melnichuk, Y.A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, and T. Burdett, **ArrayExpress update—simplifying data submissions**. *Nucleic acids research*, 2014: p. gku1057.
44. Barrett, T., S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, and M. Holko, **NCBI GEO: archive for functional genomics data sets—update**. *Nucleic acids research*, 2013. **41**(D1): p. D991-D995.
45. Moschella, F., M. Valentini, E. Aricò, I. Macchia, P. Sestili, M.T. D'Urso, C. Alessandri, F. Belardelli, and E. Proietti, **Unraveling cancer chemoimmunotherapy mechanisms by gene and protein expression profiling of responses to cyclophosphamide**. *Cancer research*, 2011. **71**(10): p. 3528-3539.
46. Kaur, S., K.J. Archer, M.G. Devi, A. Kriplani, J.F. Strauss III, and R. Singh, **Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis**. *The Journal of Clinical Endocrinology & Metabolism*, 2012.
47. Fujimoto, R., J.M. Taylor, S. Shirasawa, W.J. Peacock, and E.S. Dennis, **Heterosis of Arabidopsis hybrids between C24 and Col is associated with increased photosynthesis capacity**. *Proc Natl Acad Sci U S A*, 2012. **109**(18): p. 7109-14.
48. de Oliveira Dal'Molin, C.G., L.-E. Quek, R.W. Palfreyman, S.M. Brumbley, and L.K. Nielsen, **AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis**. *Plant physiology*, 2010. **152**(2): p. 579-589.
49. Bordbar, A., A.M. Feist, R. Usaite-Black, J. Woodcock, B.O. Palsson, and I. Famili, **A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology**. *BMC systems biology*, 2011. **5**(1): p. 1.
50. Bohler, A., L.M. Eijssen, M.P. van Iersel, C. Leemans, E.L. Willighagen, M. Kutmon, M. Jaillard, and C.T. Evelo, **Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment**. *BMC bioinformatics*, 2015. **16**: p. 267.
51. Shannon, P.T., M. Grimes, B. Kutlu, J.J. Bot, and D.J. Galas, **RCytoscape: tools for exploratory network analysis**. *BMC bioinformatics*, 2013. **14**(1): p. 217.
52. Alexa, A., J. Rahnenführer, and T. Lengauer, **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure**. *Bioinformatics*, 2006. **22**(13): p. 1600-1607.

53. Piñero, J., N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L.I. Furlong, **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. *Database*, 2015. **2015**: p. bav028.

54. Wishart, D.S., C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic acids research*, 2008. **36**(suppl 1): p. D901-D906.

55. Kutmon, M., M.P. van Iersel, A. Bohler, T. Kelder, N. Nunes, A.R. Pico, and C.T. Evelo, **PathVisio 3: an extendable pathway analysis toolbox**. *PLoS computational biology*, 2015. **11**(2).

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

57. Boele, J., B.G. Olivier, and B. Teusink, **FAME, the flux analysis and modeling environment**. *BMC systems biology*, 2012. **6**(1): p. 1.

58. Saito, R., M.E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, S. Lotia, A.R. Pico, G.D. Bader, and T. Ideker, **A travel guide to Cytoscape plugins**. *Nature methods*, 2012. **9**(11): p. 1069-1076.

Summary

Systems biology focuses on complex interactions within biological systems, using a holistic approach. Why is the whole greater than the sum of it's parts? Because the parts interact, making the whole an emergent characteristic of the parts and their interactions with each other. High-throughput studies of biological systems are rapidly accumulating a wealth of 'omics'-scale data. Visualization is a key aspect of both the analysis and understanding of these data.

It is common to describe biological processes as pathway diagrams. The pathway nodes represent the participating molecules in the biological process (genes, proteins, metabolites etc.) and edges connecting the nodes describe the relationship between the participants (reactions, interactions etc.). In this thesis, I have focused on metabolic pathways describing the metabolic processes of an organism.

Metabolic pathways are series of chemical reactions occurring within a cell. Although all chemical reactions are technically reversible, conditions in the cell are often such that it is thermodynamically more favorable for a reaction to flow in one direction.

High throughput technologies exist for measuring expression of genes, and abundances of proteins and metabolites. Transcriptomics datasets are freely available from online databases, notably ArrayExpress and GEO where datasets can be searched based on tissue of interest, disease of interest, organism of interest etc. Pathway diagrams are also available from various online databases; among them are WikiPathways and Reactome. Pathway diagrams can be used to integrate and co-analyze the different layers of data to have a complete overview of the biological process.

Pathway analysis softwares are available for performing such analyses. PathVisio is a widely adopted pathway editing, visualization, and statistics tool. PathVisio can furthermore be used for drawing pathway diagrams. The genes, proteins, metabolites in the pathway diagrams can be annotated with unique identifiers from online databases. To visualize data onto the diagrams the data uploaded must also be annotated with database identifiers. There are various online gene, protein, and metabolite databases. Identifiers from almost any of them can be used to annotate diagrams and datasets. PathVisio works together with BridgeDB to make the mapping between database identifiers easier, and identifier mapping databases are available which can map the gene or gene product related identifiers of one gene product from many online databases to each other. Such a mapping database is also available for metabolites. These identifier mapping databases are what allows mapping data onto the diagram, and visualizing it using colours.

However, by visualizing data about nodes alone, we are missing a key component to complete the picture: the data about interactions. Not many experimental techniques exist to measure metabolic fluxes; i.e. the reactions that actually occur in the cell as an end result of the transcriptional, translational, and regulatory effects in a cell. Metabolic fluxes are therefore often estimated through modelling. Mathematical models are created in which equations represent the reactions in the in-silico cell. There are various techniques of analysing these mathematical models to obtain metabolic flux values through the different reactions in the model.

Even though, mathematical models are an excellent tool for simulating the dynamic reactions occurring within cells, they are notoriously difficult to correct, share, and update. Pathway diagrams, on the other hand, are widely considered useful for representing a process, while maintaining the knowledge about the topology of the process. Creating pathway diagrams of mathematical models would not only allow modellers to better understand and update their models; it would also enable modellers and biologists to collaborate better and share knowledge. In this thesis we describe a software plugin for PathVisio that makes this workflow possible. The PathSBML plugin was developed in collaboration with Sriharsha Pamu as part of the Google Summer of Code 2013 program where I served as a mentor. It converts computational models commonly encoded in the Systems Biology Markup Language (SBML) to

pathway diagrams encoded in the Graphical Pathway Markup Language (GPML) format used by WikiPathways and PathVisio. The plugin also allows a direct import of models available from the open access database Biomodels.org. This enables visualizing a model as a pathway diagram, running that model on the online Biomodels website or in other modelling software and visualizing the model output on the model's diagram as described in this thesis.

However, enabling flux visualization required development of more components. In order to visualize flux data on the reactions and interactions of a pathway in PathVisio, the possibility to annotate the lines signifying such interactions in a pathway was created. Changes to the core of the PathVisio software and the data model for saving a pathway diagram were made in order to allow that. This enabled storing the annotation information about reactions/interactions, similar to how that was already possible for the nodes of the pathway diagrams i.e the gene, proteins, and metabolites.

For mapping uploaded data onto the diagram an identifier mapping database is needed as described above, which is why a new BridgeDb derby database was created for mapping reaction and interaction identifiers from the different online data sources. The mappings were obtained from the Open Access Database Rhea.

Additionally, the IntViz plugin for PathVisio was developed in collaboration with Rhizhou Guo from the Eindhoven University of Technology as part of his Master's thesis. This plugin adds Visualization options for interactions. Rule based and gradient based visualization options are now available for visualizing data on the reactions and interactions in a pathway. This plugin also has a slider feature that allows visualizing time series data by sliding through time.

However, in order to include flux data in pathway analysis and perform a meaningful analysis on a genome scale level, a large number of pathways with annotated interactions are necessary. Most interactions in WikiPathways pathways are not annotated yet, but the pathways in Reactome are. A Java based converter was created that converts Reactome pathways to the GPML format. This allows Reactome to take advantage of the community curation model of the WikiPathways community, in addition to performing pathway analysis using PathVisio, and newly including flux data, additionally to transcriptomics, proteomics, and metabolomics data. This allows combined statistical pathway analysis (combined enrichment scores) and the results to be quantitatively visualized using PathVisio. This integration will give a more complete overview of key players in a given biological process.

This thesis has extended the pathway analysis software PathVisio's capabilities by a complete toolset, enabling the integrations and visualization of interaction data. It has added to the wealth of knowledge available through WikiPathways by adding the human and plant collections of Reactome pathways. This improves pathway analysis capabilities by adding new genes, and new proteins to WikiPathways' already large collection of genes, proteins and metabolites, in addition to interaction annotations. These interaction annotations could be mined to automatically annotate other interactions between the same participants in other pathways in WikiPathways. It has also opened the PathVisio software to the modelling community allowing them to visualize their models and results dynamically. The best way to make an analysis reliable and repeatable is to automate it. In this thesis I also developed PathVisioRPC, an XML based Remote Procedure Call interface for PathVisio, that allows users to directly call PathVisio functions to draw and annotate biological pathways, visualize data on them and perform pathway statistics from within different programming languages. The entire analysis workflow can be automated by writing a script calling the relevant PathVisio functions, creating the possibility for easy integration of Pathway Analysis into Data Analysis Pipelines. This is further demonstrated in this thesis, by creating a pathway analysis module for the existing microarray analysis pipeline ArrayAnalysis.org.

The final chapter of this theses applies the principle of combining flux and gene expression data to investigate differences in the metabolism of metabolically unhealthy obese adults in comparison to metabolically healthy obese adults. The flux data originated from flux balance analysis of a model describing the flux in adipose tissue in the absorptive state, whereas pre-existing array data sets comparing the adipose tissue of metabolically unhealthy obese adults with metabolically healthy obese adults was used for gene expression. Pathway analysis was performed to identify the pathways that were significantly affected in metabolically unhealthy obese adults in comparison to metabolically healthy obese adults. Fourteen pathways were found to be significantly different. These fourteen pathways were merged into a network and all the pathways were found to be connected through three central genes FASN, ACACA, and ACACB and microRNAs and transcription factors that target these genes. All these three genes were downregulated. The flux data confirms that FASN, ACACA, ACACB might be important regulators as non-zero fluxes were obtained for the reactions catalysed by the enzymes encoded by these genes, by performing flux balance analysis using the metabolic model for the adipose tissue. This indicates that the reactions catalyzed by these genes are active in the adipose tissue, since the metabolic reactions catalyzed by these genes carry fluxes. The networks were further enriched with drugs and diseases. The disease associations helped to identify other diseases that people with metabolic syndrome will be prone to develop, such as cardiomyopathy, mental retardation, obesity, and insulin resistance. The drug associations helped to identify drugs currently in use for other diseases, amongst which are Cerulenin, Fomepizole, Mecasermin, Mefloquine, Nedocromil and Quercetin, which have clinical effects that would be desirable in treating metabolic syndrome.

This content described in this thesis is a step towards the complete picture of a biological process and enables integration and visualization of metabolic fluxes from mathematical modelling on interactions alongside experimental measurements of genes, proteins, and metabolites on nodes of pathway diagrams or pathway representations of the models themselves.

Valorization

Introduction

Cells are the basic unit of life. Inside every cell in our body, multiple reactions occur that produce energy from the food we consume, store excess energy, and so on. These biological processes can be represented graphically as pathway diagrams, in which the nodes represent the various genes, proteins, and metabolites, and the edges connecting the nodes define how they interact. High throughput technologies have created a big data explosion in biology by making it possible to measure the expression or abundances of thousands of genes, proteins, and metabolites. Visualizing these together on a pathway diagram eases the understanding of the complex biological process involving many players.

Metabolic fluxes are the movement of matter through the reactions in pathways. Within cells, regulation of flux is vital to regulate the activity under different conditions. The presence or absence of metabolic fluxes can therefore indicate whether a metabolic pathway is active under a certain condition. Metabolic fluxes are often modelled using computational models that, using mathematical equations, describe the same processes that are described by pathway diagrams. The resulting metabolic flux values can be negative or positive. Negative results would indicate that the reaction proceeds in a direction opposite to what is assumed in the model. These modelling results could be visualized together with measured transcriptomics, proteomics, and metabolomics data to obtain a more complete picture of the inner workings of a biological process.

In this thesis "Towards the Complete Picture: Combining Modelling and Experimental Data in a Systems Biology Approach", I have developed an extensive toolset, complementing our widely adopted Pathway analysis software PathVisio, enabling the integration and visualization of measured or modelled metabolic flux data alongside transcriptomics, proteomics, and metabolomics data. To include flux data in pathway analysis, annotated interactions are needed. A Java based format converter was developed to convert Reactome pathways, that have annotated interactions, to the PathVisio native format also used by WikiPathways for use in pathway analysis. The toolset also enables modellers to obtain up-to-date graphical representations of their models, facilitating the correction of the models as well as sharing and collaborating with others. Automating repetitive analyses is key to producing reliable, reproducible research. One of the tools developed as part of this thesis enables automating the PathVisio pathway analysis workflow by writing scripts. Notably, this allows performing Pathway analysis in R, a platform commonly used for quality control, normalization, and statistical analysis of data prior to pathway and network analysis.

Various studies were performed using each of the tools developed to provide a pedagogic example. In the final study all the tools developed have been used in one of the possible workflows enabled by the new developments. In the study we use modelled metabolic fluxes as an additional layer of confirmation alongside measured transcriptomics data to detect which pathways are crucial to study in metabolically unhealthy obese individuals.

Open Science

Science is broadly understood as collecting, analyzing, publishing, re-analyzing, critiquing, and reusing data. This can be done best in an open environment. Open science encompasses open data, so others can reanalyse and reuse data and confirm or disprove results. Analyses nowadays require software, hence open source software is critical, as scientific analyses tend to be quite unique. Therefore, publishing the software open source enables others to reproduce the analysis using the same software and verify the results. Following a modular approach to open source software development allow researchers to leverage each other's work. They can create "plugins" that can be used with a core software platform adding the various needed functionalities while reusing the functionalities of the core platform. Open source softwares have been very successful in attracting community participation enabling complex analytical approaches to be available free of cost. Research should then be published open access allowing dissemination of the knowledge as much as possible to enable wider participation in the scientific process.

In this thesis all the tenets of open science have been followed. All the studies have been performed with open data available through online databases. The developed tools are all open source allowing anybody to modify them as required. One economic opportunity here could be to start a support company that maintains, updates, and creates new plugins and provides trainings and workshops for using the plugins. The developed tools themselves could be applied broadly in other informatics infrastructures beyond infrastructures for analysing biological data. The PathVisio software is commonly used for drawing biological pathways that represent biological processes. However, any process can be represented using such a pathway diagram, therefore this free software could be of interest to business owners for example to map out their business processes and save all related information together or to physicians wishing to explain a care pathway to patients and other medical professionals. The pathway created using PathVisio can be merged to create a network in Cytoscape, another popular open source software for network analysis, to study the critical processes and prioritize them for optimisation. The Bridgedb identifier mapping software could be similarly used in corporate IT environments, a mapping database can also be used in facilitating a 360° view of the account, mapping Account IDs in SAP to Account IDs in Salesforce, this will allow data from different systems such as Inventory Management and Customer Relationship management to be combined easily.

The publications that are part of this thesis have all been published in open access journals. This allows general public access to the document. Anybody trying to use any of our tools and learn bioinformatics analysis will have access to the pedagogic examples of pathway and network analyses performed. This can be used by high-school and university teachers to create educational material, such as practical hands on sessions. For example, in Chapter 3 of this thesis, scripts in R are provided for the entire biological discovery process, starting from obtaining microarray data, cleaning, checking for data quality, normalization, statistical analysis, pathway analysis, and gene ontology analysis.

As part of the National Resource for Network Biology (NRNB), I also participated 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.

Public Adoption

The greatest value of research is its wide adoption. The wide adoption of the tools and methods developed as part of this thesis is evident by the statistics of tool downloads and citations of the related publications. For example the BridgeDbR package allowing the use of the BridgeDb identifier mapping platform in R is in the top 20% of all downloads from the Bioconductor package repository. To encourage public use of the tools we have organized various workshops to educate the scientific community of its use. We have also gone beyond the systems biology community and reached out to a broader audience such as the Bio-IT World conference that is targeted at the clinical, pharmaceutical and biomedical research community at universities, research institutes as well as companies.

Conclusion

In conclusion, I believe that in scientific research, collaborative approaches allow us to build on each other's work and expertise and move forward faster. This is accelerated by open science. This thesis, as stated before, follows this principle; all tools developed are open source, all analyses performed are using open data, and all results have been published in open access journals. In my opinion a systems approach is applicable not only to biology but to many other fields as they are facing the challenges of big data that have been faced by biologists for decades. The Internet of things will provide more and more data about every connected device, a full picture of each of which can only be obtained by combining all relevant data. Bioinformatics analyses are simply advanced analytical techniques and not dependant on the biological origin of the data. Hence any industry facing the challenges of data integration and visualization can leverage these analytical techniques.

Acknowledgements

I would like to begin by thanking the rector Prof. Rianne Letschert for allowing me to defend this thesis, the assessment committee chair Prof. Wout Lamers, and members Prof. Natal van Riel, Prof. Julio Saez-Rodriguez, Dr. Zita Soons, and Dr. Michael Lenz for taking the time to read and approve this thesis for defense.

I thank my supervisor Prof. Dr. Chris Evelo for his guidance and trust in me and for letting me work independently and supporting me where necessary. I first met Chris in Manipal for the inaugural ceremony of the Maastricht Manipal Exchange programme and followed him to Maastricht for my Master's internship. Following which I also started a PhD in his lab, BiGCaT. Needless to say I enjoyed working at BiGCaT.

My heartfelt gratitude goes to Dr. Martina Summer-Kutmon, my co-supervisor, who at the beginning of this journey was a senior PhD student, and my guide in all things PathVisio, WikiPathways and Cytoscape. In the past 5 years we have created lectures and practicals together, gone to conferences and trainings, organized workshops, maintained our software applications, and participated in Google Summer of Code programs. Through all of that I could always count on Tina's help. As I have told you before, Tina, I have always relied on you. I truly thank you. We have also grown together in our personal lives. We got married within a month of each other. Coming back from our honeymoon in Thailand, we drove for ten hours to Austria for Tina's wedding, carrying the BiGCaT Iceland Surprise for Tina and Georg. You made a beautiful bride and it was a great party, totally worth the twenty hours there and back. We also both became mothers for the first time. It has been a delight, Tina, to have you as my friend.

Thank you Martijn for your supervision during the PathVisioRPC development, for introducing me to the PathVisio and BridgeDb codebases and good programming practices. I thank you for initiating the SBML plugin and your help on programming the other projects that I have worked on.

Thank you Lars for your supervision during the development of modules of the ArrayAnalysis.org workflow, for your help and support with microarray data analysis, R scripting in general, writing the manuscripts of this thesis, and getting all the supplementary material and use cases ready for submission. My heartfelt thanks for introducing me to Maastricht, to its delightful beer and lively Carnaval,

Thank you Alex for giving me the chance to work with Reactome pathways and creating the Reactome Converter, one of the crucial components of this thesis. You have an infectiously positive attitude to work and a very pleasant demeanour. It is truly a pleasure to work with you and I sincerely hope to be able to work together in the future again.

Thank you Egon for introducing me to Jenkins, blogging, and being more social about science. Thank you for your supervision during the Google Summer of Code program and for your help with creating the R packages. You are truly inspiring in your passion for Open Science. I hope the movement becomes the only way in which science is done in the coming years.

Thank you Susan for helpful discussions during the entire course of this PhD and for your help in developing and delivering lectures and practicals.

Thank you Thomas for creating WikiPathways, for initiating Cytargetlinker, presenting which brought me my first application showcase prize. How you balanced a young family and a PhD degree inspired me to do the same.

Thank you Andra for the interesting conversations about your very adventurous experiences. From whale watching in the North Sea to helping out in a ladies hostel, you've done it all. I'm glad to have met you and gotten to hear of your experiences. Thank you for introducing me to the world of online sites that teach programming interactively and supervising the Python practicals together and our very interesting discussions on marriage and parenthood and associated pros and cons. Bollywood! And also our interesting discussions about various business propositions, notably Belgian artisanal beer pubs in India.

Thomas and Andra, I wish you both all the very best with your analytics companies. You have both inspired me to perhaps someday give that a go myself.

Thank you Nuno for your invaluable help with shell scripting, databases, websites and general ICTS administration. Nuno, you are the go-to-guy for checking any ICTS issue. You are by far the best system administrator I have had the pleasure to work with. Your help was invaluable in getting the Arrayanalysis pipeline working, setting up the Reactome Database and working on the Reactome Converter. The PathVisio websites and websites for all the various plugins and Bridgedb components would not have been the same without you. I am very much missing your presence in my current job. I am also glad to have met Vera. Good luck with your career and good luck to Vera with her music.

Thank you Jonathan for your help with Bridgedb. Jonathan, it was a pleasure to work with you. You have been one of the best Master's students I had the pleasure of working with during my time at BiGCaT. You have expertly taken over the programming duties associated with the maintenance and upkeep of our various tools. It was very easy and nice to work with you. One regret I have is not seeing the Salsa dancing that I heard so much about.

Thank you Elisa for being the shining bundle of happiness that you are. It was lovely sharing an office with you. And such a weird coincidence that we both got pregnant at the same time with the same due date. It is so lovely to see our kids playing together, it is a shame it does not happen more often. I would like to wish you all the best for the rest of your PhD and to your little family a wonderful life ahead.

Michiel, Ryan, Mirella, Linda, and Freddie, it was a pleasure working with you all, albeit we didn't get a lot of time to work together. Wishing you all the best with your careers.

Thanks also go to all the students I worked with, for allowing me to gain supervising skills and for help with the tools.

Sriharsha, it was a pleasure working with you. You are a very good programmer, and a fast learner. You allowed me to get my first mentorship experience, and I am very proud that you are doing so well. Thanks for helping me with one of the very crucial components of this thesis.

Rhizhou, it was a pleasure working with you as well. You were able to work independently with a complex code base with minimum help. Thank you for helping me complete the IntViz plugin, one of the crucial components of this thesis.

Gungor, it was a pleasure working with you and getting the chance to participate in a DREAM consortium project about inferring signalling networks thanks to you. I wish you all the best with your career.

Christ, it was great collaborating with you. You are one of the most motivated bachelor students I have had a chance to work with. Thanks for creating the BridgeDbR package, which is one of the most downloaded Bioconductor packages!

Thank you Jos, Rhea, and Myrtle for always being there to take care of administrative difficulties I had.

Thank you Magali for your help with PHP and in completing the ArrayAnalysis.org pipeline front end modules for statistical analysis and pathway analysis modules.

Thank you Kristina for giving me the chance to work with Plant pathways from Reactome and for personally checking whether the files were of good quality or not.

Rianne, Bart, Stan, Rachel, thanks for your help and company while supervising practicals together and the fun conversations we had. Rianne I wish you all the best with the rest of your PhD as well.

Lauren, Monica and Nicola thanks for your friendship. Our get-togethers were always fun. Hope to have more of them soon. Lauren and Monica best of luck with your PhDs too!

To the Reactome Team, especially Guanming Wu, thanks for your help with the existing code of the Reactome Converter.

To all the researchers who I met at all the conferences I went to to present my work. For all the fruitful discussions and collaborations we had and for all the fun (read drunk) social evenings. Notable among them are the two long courses I attended in the beginning of my PhD. The *in silico* systems biology course in Cambridge, it was really fun meeting all of you and I wish you all the best in your careers. Hope to stay in contact through our Facebook group and hopefully meet again. The Algorithms for Biological Networks course in Delft, where every evening we tried a different cuisine, Greek, Mexican and many more. It was a gastronomic delight.

I would also like to thank all the other PhD students whom I met in the various courses arranged by the University itself. It was great meeting you all, especially in the Dutch class. It was great to meet so many international faces from very diverse backgrounds and fields of expertise. Talking with students of Law, English, History etc was a fresh change in my regular conversations. Hope you all are progressing well in your PhD journeys and hope to meet you again someday.

I would like to thank all my teachers from my high-school, bachelor's and master's degree Thank you for all that you have taught me.

I would like to thank the people in my personal life. My parents, who have always supported me and expected nothing less than excellence. It has been difficult sometimes but it has pushed me to achieve all that I have and pushes me to strive forward and achieve all that I want. Dearest Ma and Baba, thank you for all my talents that I have inherited from you and all the opportunities that you have provided me. I am most thankful for instilling in me the love of reading. My dearest Ma, you have been the guiding light of my life. Almost all of my childhood memories are about things that I did with my mother. My earliest teacher, you studied all my books with me, helped me with all my homework. You made me love history and become interested in Greek and Roman mythology. I also owe my love for Biology to you. My love for Mathematics comes from Baba. I always found it truly impressive the way you could calculate long multiplications in your head.

I would also like to thank my grandparents. My Dida who has always encouraged me to be a scientist, something she had always wanted to be herself, and my love for Chemistry, which was her favourite subject. My Dadu for his wonderfully catalogued library and many wonderful afternoons of stories he told me that he just spun out on the go. Stories that broadened my creativity and imagination and that I soon contributed to. My uncle and aunt (Mama & Mami) for all your love and support towards this entire journey and for introducing me to computers. My other aunt Pishi, for being the most hard-working person I have met and for inspiring me to do the same. The fact that she had multiple jobs encouraged me to start teaching, and later to work while studying, providing valuable experience. I thank my cousins, Neel, Deeya, and Sunny for the fun times we had together, that made this journey that much better. The journey of a PhD may end at the culmination of it but it does not start at its beginning. It starts with the very first day that I started my journey of education. All of you have always shown great confidence in my studies and encouraged me to pursue a scientific career. Thank you all for all your support and love.

Most of all, I thank you Sacha. My partner, friend, guide, and confidante. I met you within a month of starting my PhD. You have truly been there for me throughout this entire time. Lending a patient ear to my PhD woes and sorrows, discussing through my ideas, proofreading every paper, poster, and letter and helping me make visualizations. You have helped me and supported me every step of the process. Joining each other on conferences made the conferences we went to so much more fun. The ESOF conference that you went to in Dublin, deserves a special mention as in the vacation that followed we got engaged while visiting Scotland on the Edinburgh castle. We got married in Belgium and in India, went for a long honeymoon to Thailand, welcomed our son Selvyn to this world and bought a house. All this while pursuing a PhD would not have been possible without you. Thank you Sacha for the joy you have brought to my life.

Last but by no means the least, my dearest Selvyn, my biggest thanks go to you for patiently tolerating Mummy spending endless weekends typing away at the computer. I am delighted to inform you that shall happen no more and thank you for being such an angel and brightening my stressful days with your carefree giggles.

Brief Curriculum Vitae

Anwesha Bohler (née Dutta) was born on the 25th of January, 1988 in the "city of joy" Kolkata, the capital of the Indian state of West Bengal. She attended a newly opened girl's school in her locality, G.D Birla, for her schooling. Even though her father shortly moved to a considerable distance away from her beloved school, she made the 2.5 hour journey there and back again daily, in almost every imaginable form of transportation. After, she completed her undergraduate degree in Science from the Lady Brabourne College, affiliated to Calcutta University, with Honours; majoring in Zoology and with Chemistry and Botany as minor subjects in 2009. She moved to the very south of India, to Manipal, to pursue her postgraduate degree in Bioinformatics from the Manipal University. She did a three month internship in the Mass Spectrometry Lab in the Manipal Life Sciences Centre, developing a Perl application for metabolite identification, following which she came to Maastricht as part of the student exchange programme. At BiGCaT, the Department of Bioinformatics at Maastricht University, she performed a six month internship working on the XMLRPC interface of PathVisio in Java and developing statistical analysis and pathway analysis modules in R for the Arrayanalysis.org workflow. After the successful completion of her Master's degree in 2011 she was offered a PhD position at the same department to work on integrating and visualizing modelled metabolic fluxes with measured multi-omics data, i.e. transcriptomics, proteomics, metabolomics. On the 1st of June 2016 she started on her next venture as a Data Scientist in the European Operations Center of Medtronic. At her current role she is leading the Advanced Analytics program within the IT Department.

Publication List

Reactome from a WikiPathways Perspective
PLoS Computational Biology 12 (5), e1004941 : (2016)
Anwesha Bohler, Guanming Wu, Martina Kutmon, Leontius Adhika Pradhana, Susan L. Coort, Kristina Hanspers, Robin Haw, Alexander R. Pico, and Chris T. Evelo.

Inferring causal molecular networks: empirical assessment through a community-based effort.
Nature methods 13 (4), 310-318 : (2016)
Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, Kiley Graim, Adrian Bivol, Haizhou Wang, Fan Zhu, Bahman Afsari, Ludmila V Danilova, Alexander V Favorov, Wai Shing Lee, Dane Taylor, Chenyue W Hu, Byron L Long, David P Noren, Alexander J Bisberg, Bahman Afsari, Rami Al-Ouran, Bernat Anton, Tomasz Arodz, Omid Askari Sichani, Neda Bagheri, Noah Berlow, Alexander J Bisberg, Adrian Bivol, Anwesha Bohler, Jaume Bonet, Richard Bonneau, Gungor Budak, Razvan Bunescu, Mehmet Caglar, Binghuang Cai, Chunhui Cai, Daniel E Carlin, Azzurra Carlon, Lujia Chen, Mark F Ciaccio, Thomas Cokelaer, Gregory Cooper, Chad J Creighton, Seyed-Mohammad-Hadi Daneshmand, Alberto de la Fuente, Barbara Di Camillo, Ludmila V Danilova, Joyeeta Dutta-Moscato, Kevin Emmett, Chris Evelo, Mohammad-Kasim H Fassia, Alexander V Favorov, Elana J Fertig, Justin D Finkle, Francesca Finotello, Stephen Friend, Xi Gao, Jean Gao, Javier Garcia-Garcia, Samik Ghosh, Alberto Giaretta, Kiley Graim, Joe W Gray, Ruth Großeholz, Yuanfang Guan, Justin Guinney, Christoph Hafemeister, Oliver Hahn, Saad Haider, Takeshi Hase, Laura M Heiser, Steven M Hill, Jay Hodgson, Bruce Hoff, Chih Hao Hsu, Chenyue W Hu, Ying Hu, Xun Huang, Mahdi Jalili, Xia Jiang, Tim Kacprowski, Lars Kaderali, Mingon Kang, Venkateshan Kannan, Michael Kellen, Kaito Kikuchi, Dong-Chul Kim, Hiroaki Kitano, Bettina Knapp, George Komatsoulis, Heinz Koeppl, Andreas Krämer, Miron Bartosz Kursa, Martina Kutmon, Wai Shing Lee, Yichao Li, Xiaoyu Liang, Zhaoqi Liu, Yu Liu, Byron L Long, Songjian Lu, Xinghua Lu, Marco Manfrini, Marta R A Matos, Daoud Meerzaman, Gordon B Mills, Wenwen Min, Sach Mukherjee, Christian Lorenz Müller, Richard E Neapolitan, Nicole K Nesser, David P Noren, Thea Norman, Baldo Oliva, Stephen Obol Opiyo, Ranadip Pal, Aljoscha Palinkas, Evan O Paull, Joan Planas-Iglesias, Daniel Poglayen, Amina A Qutub, Julio Saez-Rodriguez, Francesco Sambo, Tiziana Sanavia, Ali Sharifi-Zarchi, Janusz Slawek, Artem Sokolov, Mingzhou Song, Paul T Spellman, Adam Streck, Gustavo Stolovitzky, Sonja Strunz, Joshua M Stuart, Dane Taylor, Jesper Tegnér, Kirste Thobe, Gianna Maria Toffolo, Emanuele Trifoglio, Michael Unger, Qian Wan, Haizhou Wang, Lonnie Welch, Chris K Wong, Jia J Wu, Albert Y Xue, Ryota Yamanaka, Chunhua Yan, Sakellarios Zairis, Michael Zengerling, Hector Zenil, Shihua Zhang, Yang Zhang, Fan Zhu & Zhike Zi, Gordon B Mills, Joe W Gray, Michael Kellen, Thea Norman, Stephen Friend, Amina A Qutub, Elana J Fertig, Yuanfang Guan, Mingzhou Song, Joshua M Stuart, Paul T Spellman, Heinz Koeppl, Gustavo Stolovitzky, Julio Saez-Rodriguez, and Sach Mukherjee.

WikiPathways: capturing the full diversity of pathway knowledge
Nucleic acids research gkv1024: (2015)
Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi Sinha, Ryan Miller, and Susan L. Coort.

Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment
BMC bioinformatics 16(1), 267 : (2015)
Anwesha Bohler, Lars MT Eijssen, Martijn P. van Iersel, Christ Leemans, Egon L. Willighagen, Martina Kutmon, Magali Jaillard, and Chris T. Evelo.

PathVisio 3: an extendable pathway analysis toolbox
PLoS Computational Biology 11(2), e1004085 ; (2015)
Martina Kutmon, Martijn P. van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, Alexander R. Pico, and Chris T. Evelo.

A Mass Spectrometric Study for Comparative Analysis and Evaluation of Metabolite Recovery from Plasma by Various Solvent Systems
Journal of biomolecular techniques 23 (4), 128 ; (2012)
Anwesha Dutta, Premalatha Shetty, Smitha Bhat, Yeshaswini Ramachandra, and Shrinidhi Hegde.

A toolset for flux data integration in pathway analysis
Under Review
Anwesha Bohler, Susan L. Coort, Sacha Bohler, Rhizhou Guo, Sriharsha Pamu, Jonathan Melius, Chris Evelo.

Exploring metabolic health combining measured and modelled genomic data
In Preparation
Anwesha Bohler, Martina Kutmon, Chris Evelo.