Towards the complete picture

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.