Towards the complete picture

Citation for published version (APA):

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.