# Data integration with biological pathways

# Data Integration with Biological Pathways



The study presented in this thesis was performed within NUTRIM School for Nutrition, Toxicology and Metabolism 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 research was financially supported within the transnational University Limburg (tUL) by the province of Limburg.

The printing of this thesis was sponsored by the Netherlands Bioinformatics Centre (NBIC)

Cover Design: Martijn van Iersel

Layout: Martijn van Iersel

Printed by: GVO drukkers & Vormgevers B.V. | Ponsen & Looijen, Ede



ISBN 978-90-6464-423-8

# Data Integration with Biological Pathways

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op vrijdag 5 november 2010 om 12.00 uur

door

**Martijn Peter van Iersel**

geboren te Tilburg op 24 juli 1979

**Promotor**

Prof. dr. E.C.M. Mariman

**Copromotor**

Dr. C.T.A. Evelo

**Beoordelingscommissie**

Prof. dr. F-J. van Schooten (voorzitter)

Prof. dr. E. Biessen

Prof. dr. B. Mons

Prof. dr. F. Schreiber

Dr. H.J.M. Smeets

# Contents

**Chapter 1** 7
General introduction

**Chapter 2** 13
Pathways and pathway diagrams

**Chapter 3** 31
Presenting and exploring biological pathways with PathVisio

**Chapter 4** 45
WikiPathways: pathway editing for the people

**Chapter 5** 55
The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services.

**Chapter 6** 71
Pathway visualization applied to a multi-omics study of starvation in mouse intestine

**Chapter 7** 87
General discussion

**Abbreviations** 103

**Summary** 105

**Acknowledgements** 113

**Publications & Curriculum vitae** 115

Chapter 1

# General Introduction

Classical genetics has been very successful at characterizing genetic disorders such as hemophilia, sickle cell anemia and cystic fibrosis. These diseases are caused by a mutation in a single gene, and once that gene was discovered, that immediately led to a much better understanding of the disorder. But most diseases are more complicated; single-gene disorders are the exception rather than the rule. Evolutionary pressure has ensured that most cellular processes are robust, and they do not immediately fail when a single gene or protein is broken, because alternative pathways can take over. Drugs that are selected to target a single protein often turn out to be ineffective because of this redundancy [1]. For diseases such as diabetes, asthma and cancer, it has been much harder to gain a better understanding and answer important questions. These diseases are caused by a complex interplay of genes, environmental factors, diet and lifestyle. To understand them, it is not sufficient to look at a small number of genes and proteins. Instead it is necessary to study the action of multiple biological pathways as a whole and measure as many of the components of those pathways as possible. Many questions in current biological research require large amounts of data to be answered [2].

Large amounts of data are needed, and indeed large amounts of data are obtained. Technological developments have made it easy to measure a large number of RNA, proteins or metabolites at once. Large scale experiments are now so common that a whole new terminology has arisen around them. Collections of biomolecules of a certain type are designated with a word ending with –ome, such as genome for the collection of all genes, transcriptome for the collection of all (messenger RNA) transcripts, and metabolome for the collection of all metabolites. The study of the genome is then called genomics, and in similar vein the words transcriptomics, metabolomics and even interactomics, phosphoproteomics, lipidomics, localizomics and phenomics have been coined [3]. For example, microarrays let you measure the activity of all transcripts and can thus be classified as a transcriptomics technology. Equivalent technologies exist for the other types of biomolecules. All these technologies are continually improving in efficiency, sensitivity and specificity.

Each type of data forms a piece of the puzzle. Current biological questions require a full overview of all the components of the cell. For example, high glucose levels in the blood stimulate the production of insulin by the pancreas. Insulin is carried by the blood to other cells, where it activates the insulin receptor pathway. This in turn leads to the activation of pathways for the production of amino acids, lipids and glycogen and repression of gluconeogenesis and ketogenesis. In each pathway, enzymes are phosphorylated, gene expression is changed or signaling proteins are shuttled to a different location in the cell. In type II diabetes, the ineffective action of insulin can influence all these components. Integration of multiple types of omics data will be a cornerstone to achieve complete understanding of this disease.

A pathway can be loosely defined as a set of biological interactions that are functionally related. Biological pathways are little more than an abstraction, a simplification to make a complicated reality understandable. Pathways can be represented graphically as a diagram, displaying a wealth of information about a process in a single view. A pathway diagram is suitable as a point of integration of data and knowledge. Pathway diagrams may be used to visualize biological knowledge as well as a large number of data points at the same time. Software tools are needed to make this possible. Thus, the main goal of this thesis is as follows:

*Goal: develop methods to integrate multiple types of experimental data and visualize them on pathway diagrams.*

These methods do not have to be developed from scratch. There are a number of bioinformatics projects related to pathway diagrams. Chapter 2 will review the current state of pathway representation on computers, including graphical notations, databases and file formats.

The main goal can be divided in a few smaller sub goals.

*Sub goal 1: develop user-friendly software to create and view pathway diagrams.*

One of the first tools we looked at was GenMAPP [4], which allows pathway creation and visualization of datasets, but was limited in several ways: the software was written in the inflexible Visual Basic programming language, and only allowed visualization of microarray (transcriptomics) data. Chapter 3 describes PathVisio, a new pathway visualization tool that was developed as a continuation of GenMAPP, improving upon it where possible.

*Sub goal 2: annotate pathway diagrams so they can be linked to experimental data*

To be able to link experimental data to pathways, it is important that these pathways are annotated correctly, using identifiers from online databases that are standardized and globally recognized. PathVisio was developed with this goal in mind.

*Sub goal 3: maintain a curated database of pathway diagrams*

Data integration on pathways works best when there is a large collection of curated pathway diagrams available. Furthermore these diagrams should be continually checked, improved and updated when new knowledge becomes available. The GenMAPP group has created a large number of pathway diagrams but the small number of active curators could not keep up with developments in all biological research. Therefore we developed WikiPathways, an online resource for pathway diagrams that can be accessed by anyone, facilitating collaborative curation to share the burden of keeping pace with the growing body of knowledge. This work is presented in chapter 4.

*Sub goal 4: map identifiers of related biological entities*

Just using the right identifiers is not enough. Biological entities are related to each other – for example a protein is related to the gene from which it is transcribed, and a microarray probe is related to the gene that it is supposed to measure. If all these biological entities are described by identifiers, then the question is how these identifiers relate to each other. This problem and all its implications are considered in chapter 5. To encourage the re-use of code we developed a software framework called BridgeDb that provides a generalized solution that is employed by PathVisio, but could also be easily adopted by other software projects.

When the results of the work on each sub goal are taken together, we can revisit the main goal, which is to *develop methods to integrate multiple types of experimental data and visualize them on pathway diagrams.* In chapter 6 we demonstrate the validity of the approach. To make the demonstration realistic we applied our methods to a specific research problem. We focused on nutrition research, because it benefits from the multi-omics approach in particular. We have evolved over millions of years to sustain on a wide variety of diets, so our bodies have developed mechanisms to balance dietary changes that involve many genes and pathways. Eating something different does not immediately throw your body out of balance. With this in mind, it is no surprise that nutrition-related disorders such as obesity and type II diabetes are complex multi-gene diseases.

The biological question that we asked in this case was related to long-term food deprivation. Humans can survive six to eight weeks without food [5], and during that time cells in your body respond to make the most of available energy reserves. Each organ in the body responds in its own way, and complex interactions take place to make sure that vital organs get enough fuel while the non-vital organs are conserving energy as much as possible. In the study presented in chapter 6 we look in particular at the role of the intestine.

Finally, in chapter 7 conclusions are drawn and the previous chapters are discussed.

## References


1. Kitano H: **A robustness-based approach to systems-oriented drug design**. *Nat Rev Drug Discov* 2007, **6**(3):202-210.
2. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era**. *Bioessays* 2004, **26**(1):99-105.
3. Joyce AR, Palsson BO: **The model organism as a system: integrating 'omics' data sets**. *Nat Rev Mol Cell Biol* 2006, **7**(3):198-210.

4. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways**. *Nat Genet* 2002, **31**(1):19-20.
5. Collins S: **The limit of human adaptation to starvation**. *Nat Med* 1995, **1**(8):810-814.

# Chapter 2

# Pathways and Pathway Diagrams

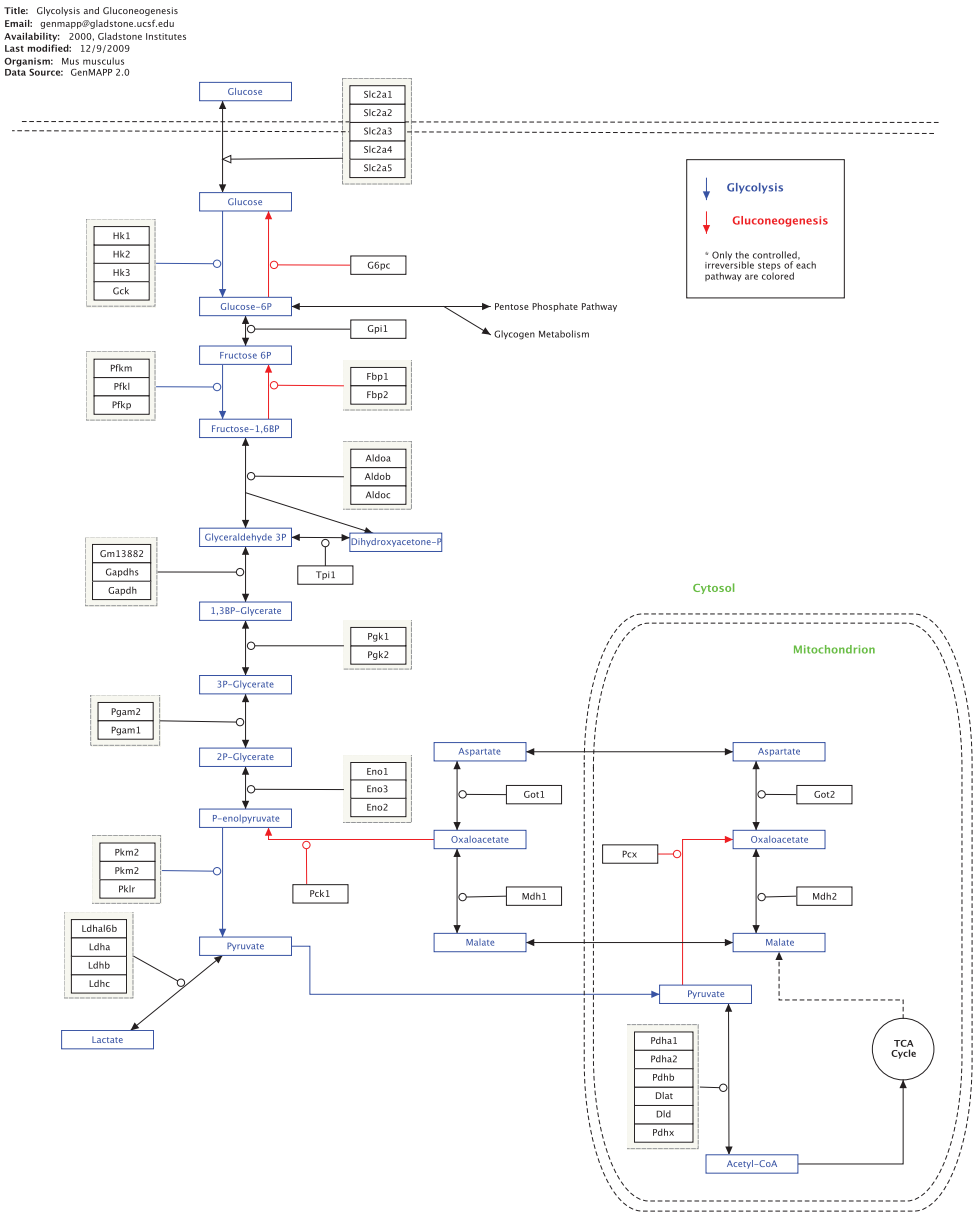


***Figure 1: Glycolysis and Gluconeogenesis pathway in* Mus musculus**. *Source: WikiPathways (WP157)[1].*

Pathway diagrams are a central concept in this thesis. In this review, some general questions about pathway diagrams will be answered. See for example Figure 1, which shows a diagram of the glycolysis and gluconeogenesis, two well-studied biochemical pathways. Upon looking at this diagram, several questions may arise. How was this diagram created? What is the meaning of each arrow and each symbol? Is the meaning of these symbols formalized?

In glycolysis, glucose is broken down to pyruvate. In gluconeogenesis, glucose is formed from smaller molecules. In this case glycolysis and gluconeogenesis are depicted together, because they share a large number of steps, but not all sources do that. How do we compare pathways and determine where they overlap? How do we determine if components on two pathways are the same? How are the endpoints of this pathway defined?

Pathways can be part of textbook knowledge or they can be highly debated as current research develops. Discussion is a natural part of science, and we can not always take pathway content for granted, as the underlying theories may be reformed at any time. Where can we find pathway diagrams to compare with? Where are pathway diagrams stored so they can be shared by the scientific community?

## What is a Pathway?

Pathways are many things to many people. A pharmaceutical researcher may look at pathways to identify key regulators that could be potential drug targets. A biotechnologist may think of kinetic modeling of enzymatic reactions in order to predict metabolite output of genetically engineered bacteria. A developmental biologist may look at signaling pathways that trigger the differentiation of embryonic stem cells to form blood or brain or skin tissues. To satisfy the wide variety of perspectives, the Biological Pathway Exchange (BioPAX) community uses the following broad definition: “A set or series of interactions, often forming a network, which biologists have found useful to group together for organizational, historic, biophysical or other reasons.” [2].

Pathway diagrams are a representation of a scientist’s knowledge in a format that is convenient, comprehensive and easy to understand. It formalizes biological entities (genes, messengers, proteins, complexes or metabolites) and the relations between them (transport, reaction, conversion). Pathway diagrams can also indicate in which cellular components the biological processes take place. Smaller pathways can be merged or large pathways can be split depending on the needs of the researcher. The glycolysis pathway of Figure 1 could easily be extended on either end with more interactions, such as the conversion of other monosaccharides to glucose, or factors that regulate the activity of the pathway.

For the sake of brevity, the term pathway is sometimes used to refer to both the biological entity, and the diagram representing it. For example, the PathVisio program is called a "pathway editor", but it is really the diagrams that are being edited, not the processes inside living cells.

## Pathway construction process

For every biological research subject, a different pathway can be drawn. Since pathway boundaries are not clearly defined, there is no objection to creating a pathway diagram that just depicts the processes that are relevant to e.g. a certain disease, or to an interesting mutated genotype. Simply the act of gathering research findings and compiling them into a comprehensive pathway is insightful (and it can be recommended to anybody who is learning about an unfamiliar research area).

Pathways can be assembled based on various sources. Viswanathan et al. [3] distinguish between knowledge driven objective (KDO) and data driven objective (DDO) pathways. DDO pathways derive from large experimental datasets (such as co-expression data or protein-protein interaction data) and are visualized in the form of large networks. KDO pathways are currently mainly derived manually, possibly with the help of text mining tools but never primarily by text mining. As a result, KDO pathways have a predefined layout. The pathway diagrams considered in this chapter fall almost exclusively in the KDO category.

DDO pathways are often large interaction networks that look like "hair balls": nodes are organized randomly with many crossing edges. This makes them hard to interpret visually. A good pathway diagram must have a layout that is easy to understand and pleasing to the eye. Automatic layout algorithms for pathway models exist and have been used successfully. [4-7]. The cellular location of proteins, if known, can provide a useful guide for layout algorithms as well [8]. But to compete with manual layouts, layout algorithms have to satisfy aesthetic criteria that are sometimes hard to codify, and manual layouts are often based on arbitrary conventions. In text book pathways, components may be duplicated to minimize edge crossing, a solution that increases computational complexity for layout algorithms. For these reasons, manual layouts can not yet be completely substituted by automatic layouts.

Examples of pathways that have been created and curated manually can be found in literature [9-14]. Information is gathered primarily from scientific literature, but also existing pathway databases and other bioinformatics databases can be used as sources. After information is organized and elements are arranged in an understandable layout, the pathway diagram should be peer-reviewed by experts. Usually this takes place in several iterations. Manual pathway construction is an iterative process, that requires a feedback loop between literature research, bioinformatics databases and experts [10].

Automated methods for pathway curation using text-mining can be problematic [3]. Natural language processing does not have high enough accuracy to derive pathways automatically. When human experts can disagree on the true facts of a pathway, an automated system can have little hope of achieving better results. Nevertheless text mining tools could serve to support manual labor.

Pathway construction can be done using pen and paper, but by using computer software the pathway diagram becomes more accessible to editing, sharing, and a variety of other uses. This software should provide easy ways to quickly try out various changes to pathways, and it should be easy to share and extend pathways with coworkers, either through public repositories, or through private communication.

## Graphical Notations

Pathway diagrams have been used in textbooks for at least 60 years [15]. However, these textbook pathways use ad hoc symbology with hardly any standardization. These ad-hoc formats often assign multiple meanings, such as stimulation, catalysis, conversion or transport, to the same arrow type. At the same time, different sources may use different symbols for the same thing, such as a line ending in a t-bar or an arrow with a minus sign to indicate inhibition. This leads to confusion and ambiguity, and hampers the interpretation of diagrams. Complex diagrams can not be readily understood without extensive explanation in text [15].

It has been pointed out [15, 16] that in other fields of endeavor, a consensus notation has been reached. This is the case in object oriented programming (UML), functional programming (flow charts) and electronic circuit diagrams. The field of molecular biology is lagging in this respect. These other diagram types are in an advantageous position because they describe a system (a program, an object model, an electronic circuit) that is designed and thus completely known. Biological pathways on the other hand can contain unknown elements that require more research. Nevertheless, a formalized notation for pathway diagrams would be useful for communicating biological knowledge, and could help forward biological research in areas of high complexity [17].

There is a clear trend towards visualization of pathways, even in places where models were historically defined purely as a list of reactions. For example Reactome, a database centered around biochemical reactions, is now planning to visualize those reactions in diagrams [18]. Also the community around the Systems Biology Markup Language (SBML), primarily concerned with models intended for computational simulation, has created the SBML layout extension to serve the need for visualization of SBML models [19]

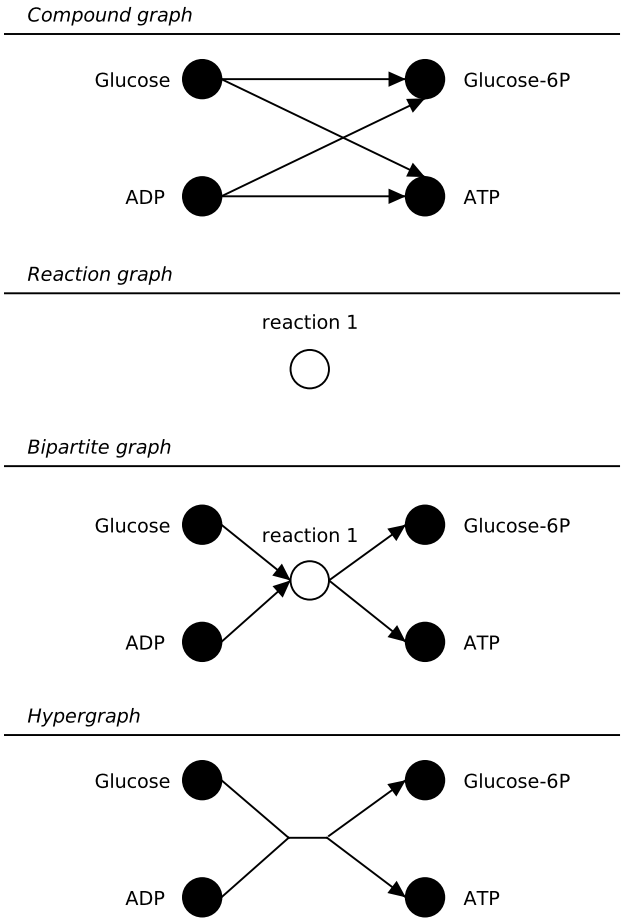


***Figure 2: Four possible graph representations of biochemical pathways.*** *These representations are based on Deville et al. [20]*

Pathways are graph-like in nature, but there exist different views on what exactly should be represented by the nodes and edges of that graph. Biochemical pathways show how metabolites are converted in steps to various end products (for example, pyruvate in the case of the glycolysis pathway). In signaling pathways on the other hand, metabolites are less important, and gene expression or protein modifications are the primary units of activity. The two are not completely separate categories however, as the bottleneck in metabolic pathways is often the concentrations of enzymes, which are regulated through gene expression. This has an important consequence for correct graphical representation of a pathway. In a compound graph, nodes are compounds and edges are reactions (See Figure 2). In a reaction graph, nodes are reactions and edges connect consecutive reac-

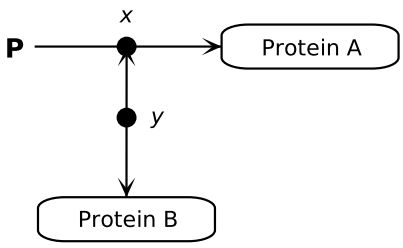


***Figure 3: A small example of MIM notation.*** *Protein A can be phosphorylated, which is represented by the arrow between P and Protein A. The dot labeled x signifies the result of that interaction, i.e. the phosphorylated state of Protein A. The double-ended arrow to Protein B represents complex formation between phosphorylated Protein A and Protein B. According to this diagram, phosphorylation of Protein A is necessary for complex formation. The resulting complex is represented by the dot labeled y.*

tions. Neither compound graphs nor reaction graphs are capable of expressing both a biochemical network and the regulation of that network in the same diagram. A graphical notation must either be a bipartite graph where nodes and edges can be of different classes or a hyper graph where an edge can connect more than 2 nodes [20]. As we will see further on, bipartite graph representations are used in current notation standards.

### *Molecular Interaction Maps*

Early efforts for defining a graphical notation [21, 22] were received with interest but were not widely adopted. Molecular Interaction Maps (MIM) was the first notation that has persisted up to date and was developed with a high level of detail [9, 16, 23, 24]. PathVisio is currently the only specialized editor for Molecular Interaction Maps (Chapter 3).

In molecular interaction maps, different arrowheads are used to represent interactions such as conversion, covalent binding, transcription, stimulation and catalysis. The result of an interaction is represented by a dot on the interaction line. See Figure 3 for an example.

Designers of graphical notations must be able to deal with diagrams that lack sufficient detail to allow computational modeling. Often there are too many components of the pathway that are simply unknown. MIM recognizes that fact and allows two possible interpretations, namely explicit and heuristic maps. A heuristic map is simply an information organizer; explicit maps are intended for computational modeling. The great visual complexity of explicit maps does little towards comprehension. If knowledge transfer is the goal, then heuristic maps are more practical.

### *Kitano Process Diagrams*

Another notation system that has emerged is Kitano process diagrams [25, 26]. Process diagrams are designed primarily to represent biochemical pathways, unlike MIM which

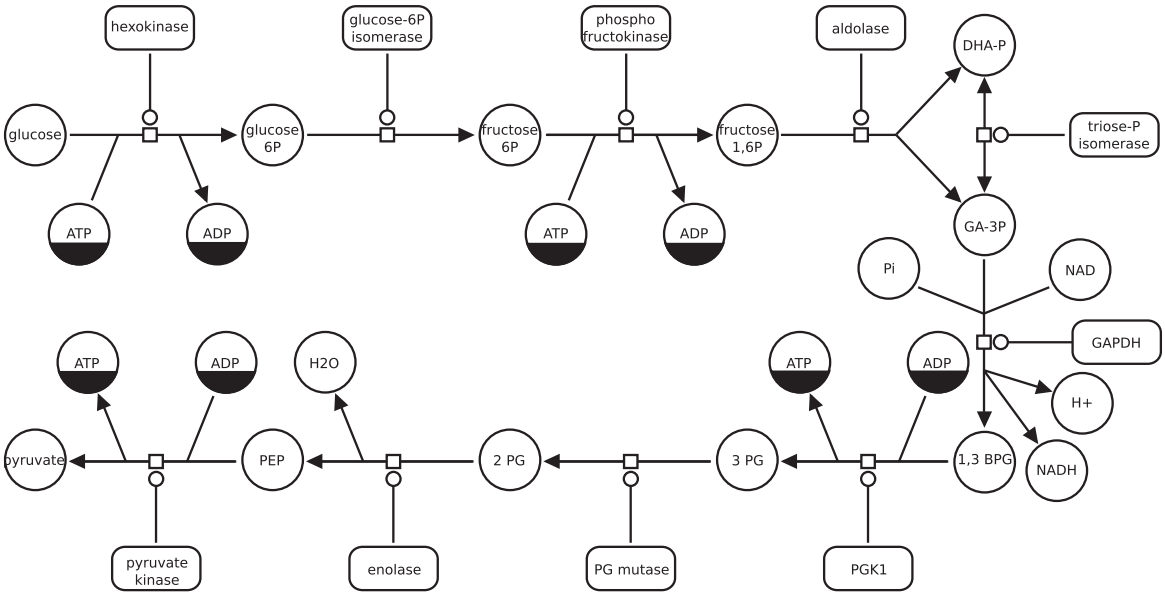


***Figure 4: Glycolysis pathway represented using SBGN Process Diagram (PD) notation.*** *In this diagram, two types of entity pool nodes can be found. Circles represent small molecular compounds, whereas rounded rectangles represent macromolecules (in this case enzymes). Some circles have the lower half colored black. This is a clone marker to indicate that these compounds occur multiple times in the diagram. Process nodes that represent the reactions are a small square. Arrows pointing to an entity pool node indicate production of that entity, lines with a open circle at the end, pointing to process nodes indicate catalysis.*

is targeted more towards signaling networks. This is reflected by the fact that process diagrams have a clear ordering of pathway steps, whereas MIM only show the relations between biomolecules but not in what order those biomolecules interact.

Another issue that must be addressed in graphical notations is the problem of combinatorial explosion, which arises when a protein can have a number of possible states. States could be covalent modifications, conformational changes or a cellular relocation. Each combination of states could lead to different reaction activities or interactions. It is not uncommon to have thousands of possible states for a single protein. In process diagrams, multiple variants of a protein are represented by multiple glyphs in the diagram, which means that they have to be duplicated. This is an important difference with MIM notation. In Molecular Interaction Maps, each entity is depicted only once, with a few exceptions. An example of this can be seen in Figure 3, in which Protein A is depicted only once even though it occurs in two states (phosphorylated and unphosphorylated).

***Systems Biology Graphical Notation (SBGN)***

Systems Biology Graphical Notation (SBGN)[15] provides a broad basis for standardization by engaging the scientific community. SBGN draws on the experience with MIM and Kitano process diagrams. To serve different uses of pathway diagrams, SBGN defines three different languages for specifying pathways at different levels of detail.

The three sublevels are Activity Flow diagrams (AF), Entity Relationship diagrams (ER) and Process Description diagrams (PD). The simplest of the three is the AF diagram, which is centered on activities, and which entities participate in those activities. An arrow simply means that one activity influences another, but does not have to imply physical contact of the biological entities. ER diagrams represent how entities relate to each other: which proteins bind to each other, which proteins are phosphorylated or otherwise modified. Entity relationship diagrams draw inspiration from Molecular Interaction Maps. A PD diagram focuses on processes, i.e. reactions, creation, degradation, phosphorylation, etc. It is a bipartite graph with one node type called Entity Pool Nodes for genes, proteins and metabolites, and a second node type called Process Nodes, to represent reactions and other processes. See Figure 4 for an example. The PD language of SBGN draws inspiration from, and improves upon, the Kitano process diagrams.

Building on the experience gained in the SBML community, the SBGN developers realized that it is impossible to develop a perfect notation system in one sitting. Instead, SBGN is released in versions called levels. This is important because it allowed SBGN to get an initial version out relatively quickly. The first version is relatively simple, using fewer different notation types than MIM.

Experience has shown that certain graphical properties are problematic in day-to-day usage. For example color is an issue when photocopying or printing on a black & white printer. Line thickness is dubious when scaling the image. For these reasons SBGN does not assign meaning to color or line thickness. SBGN also does not prescribe how the pathway should be laid out, and there is no biological meaning encoded in layout information.

## Archival and Exchange of Pathway Diagrams

Many online databases have appeared that collect and distribute pathway diagrams. Some of these databases were reviewed before [10, 27]. PathGuide [28] is an extensive directory of pathway and interaction databases. At the time of writing, PathGuide lists 324 resources related to biological pathways, of which 37 are in the category “Pathway Diagrams”. From this large number of databases we will highlight only a few, focusing on popular choices that include pathway diagrams and demonstrate the range of available features.

When comparing databases of pathway diagrams, we must also consider the file format that is used to store them. The file format used determines what the pathways can be used for, and which tools can work with them. Standard file formats that are widely used are preferable.

Most pathway databases allow export in various graphics formats. We can distinguish vector graphics formats such as SVG and PDF, and raster graphics formats such as PNG and JPG. For display on computer screens, the latter is usually sufficient, but for publication purposes a vector graphics format provides a much sharper resolution.

However, graphics file formats lack the ability to store annotations that are relevant to biology. Pathway models should contain biological annotations, in accordance with the Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) [29]. MIRIAM is an effort to define a standard for annotation of models, to enable comparison and exchange. Although MIRIAM is intended for quantitative models for use in simulations, most requirements are applicable to any pathway model. Besides simple requirements, such as title, author contact details, associated publication, time of creation and distribution terms, there is also the requirement that all biological entities in the model (genes, proteins, metabolites) are described by a suitable identifier linking to a bioinformatics database. Suitable identifiers are unique, perennial and free of semantics [30].

Thus we can identify two components of a pathway diagram that can be encoded digitally: a model component, which consists of graph information, annotations and identifiers, and a graphical component, which consists of layout, shapes and the routing of connecting lines. The ideal pathway file format must be able to handle both components together. Pure graphics formats are not entirely sufficient, neither are formats with only biological annotations.

Two relevant existing pathway standards are SBML and BioPAX. These standards are openly developed and supported by a broad community. Another notable standard is PSI-MI, but it will not be considered here because it is focused on protein interactions and less suitable for pathways in general. All features of PSI-MI are subsumed by BioPAX level 2 [31-33]. Besides BioPAX and SBML, many ad hoc pathway exchange formats may be encountered in practice, such as KGML (used by KEGG, see below) or GPML (used by WikiPathways). These formats did not undergo a standardization process because they were developed for a single database, and thus lack broad community support.

BioPAX is the most widely recognized and supported standard for pathway data exchange. BioPAX is based on the Web Ontology Language (customarily abbreviated as OWL instead of WOL). Through OWL, BioPAX defines a data model that is object oriented and set in a hierarchy of classes. Pathways are encoded as a collection of Physical

Entities and Interactions. For both, there is a tree of subclasses to allow more detailed specification. For example, a BiochemicalReaction is a subclass of Conversion, which is a subclass of Interaction. Next to Physical Entities and Interactions there is a third branch of Utility classes in the hierarchy. BioPAX is used by a wide variety of databases and tools, although the broad scope and the difficulties in defining validation criteria in OWL make it relatively difficult to work with.

BioPAX does not support storing layout information. For that reason, most pathway databases necessarily rely on custom file formats (for example KGML or GPML) that contain both annotations and layout data. In order to extract that information, custom parses must be used. Thus, the promise of BioPAX to make custom parsers for individual database formats obsolete [34] is only partially fulfilled.

Systems Biology Markup language (SBML) can encode quantitative models intended for simulation. Pathways encoded in SBML contain species (proteins or metabolites), reactions and compartments. Reactions can be described quantitatively using mathematical formulas. SBML is advanced in the usage of controlled vocabularies for biological terms. SBML can store layout data using a special extension [19], although that is not yet adopted everywhere.

The main difference between BioPAX and SBML is that the former is aimed mainly at the exchange of pathway information between databases, whereas the latter is intended as a support for modeling and simulation. Conversion between the two formats without loss of data has not yet been demonstrated.

Besides the use of file formats, there are other important points of comparison between various databases. Databases differ in the process they use to stay up-to-date. Some databases use an extensive curation process with experts, which is relatively slow but leads to high quality data. In some cases the group of experts is closed off, making it nearly impossible to influence the development of the database. In other cases interested experts are invited to join, making the process relatively open. Some databases are open to submissions from the general public to correct errors or even add completely new diagrams.

Another point of comparison is the terms of use of pathway diagrams. Commercially developed pathway databases are very restrictive, and cannot be improved and redistributed. Databases developed in the academic environment are usually more open, but the terms of use are not always clearly specified. Databases which explicitly state under what terms pathways can be used, improved and redistributed, enable scientists to make the most optimal use of pathway information.

In Table 1, pathway diagram databases are compared on curation process, supported file formats and graphical notations, and license terms.

*Table 1: Overview of the relations between databases, file formats, graphical notations and input tools*

| Database | Input process / Curation | Supported File Formats | Terms of use | Graphical Notation |
|---|---|---|---|---|
| KEGG [35, 36] | Developed in house, no procedure for external contributions. | KGML, BioPAX, Raster graphics | Free for academic use | Ad Hoc |
| Reactome [18] | Curated, Data imported using Reactome editor | BioPAX, SBML | Free and unrestricted | Ad Hoc, moving to SBGN |
| WikiPathways [37] | Post-hoc curation, Edited directly on website by any visitor | GPML, GenMAPP, Raster and Vector graphics | Free (Creative Commons Attribution License) | Ad Hoc |
| Nature-NCI Pathway Interaction Database [38] | Curated by Editors of Nature Publishing Group | XML, BioPAX, Raster and Vector graphics | Free and unrestricted | Ad Hoc |
| eMIM [39] | Developed In House, no procedure for external contributions. Software used: Illustrator and PathVisio | Raster and vector graphics | Free and unrestricted | MIM |
| BioCyc / MetaCyc [40] | External contributions are encouraged through an adoption system per species. | Downloadable PGDB format, BioPAX, raster and vector graphics | Freely available to all | Ad Hoc |

The databases in Table 1 will now be described in more detail.

***KEGG Pathway***

The database of visual pathways that is probably most popular and most widely used is KEGG (Kyoto Encyclopedia for Genes and Genomes) [35, 36]. The KEGG pathway collection is only part of a larger database that also contains other biological entities such as genes, metabolites and proteins.

KEGG pathways do not use gene or protein identifiers but rather EC numbers for indicating enzymes, which makes them independent of species and tissue. These cross-species "reference pathways" can be colored for which reactions actually occur in a species of interest.

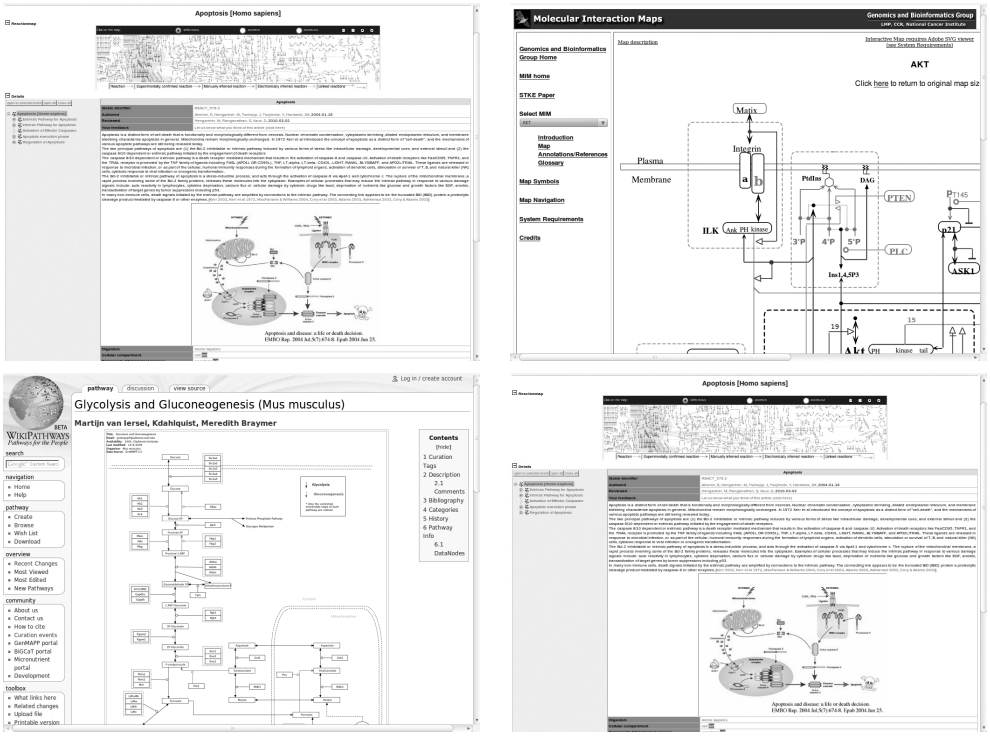


***Figure 5: Four online pathway databases.*** *Top-left: Reactome, top-right: eMIM, bottom-left: WikiPathways, bottom-right: NCI-Nature Pathway Interaction Database*

There are errors found in KEGG [41], although not necessarily problematic for visualization. The KEGG Pathway set is produced by in house curators; there is no defined process for accepting outside contributions.

Besides support for standard BioPAX, KEGG provides its own KGML format which includes both biological annotations and layout information.

### *Reactome*

Reactome is organized around biochemical reactions, and multiple reactions are organized in pathways. Although Reactome is not strictly a database of pathway diagrams, some pathways carry a diagram in the description section. More importantly, support for diagrams in SBGN graphical notation is currently under development [18]

Reactome has a well defined process for incorporating new information. Reactome relies on a wide network of expert curators, who have to take time to learn to use the Reactome Author Tool. Thus getting data into the database is a slow process, but it is in principle open to outside contributions. Reactome has grown 2.7 fold in the past three years.

Reactome is standards compliant, information can be downloaded in BioPAX and SBML formats. The Reactome Author Tool supports drawing in SBGN graphical notation. Reactome data can be used and redistributed freely.

***WikiPathways***

WikiPathways is open to contributions not only from experts but any interested party. Data in WikiPathways is provided by the public and can also be used freely by the public. Data in WikiPathways can be redistributed under the terms of the Creative Commons Attribution License.

Currently, a disadvantage of WikiPathways is that it does not support standard pathway file formats. Its main file format, GPML, is only supported by few tools (PathVisio, WikiPathways, and Eu.Gene).

WikiPathways supports more than a dozen model organisms. The pathway collection for human is the largest. Pathways of other species are partially made through automated inference from human and partially purpose built.

***NCI-Nature Pathway Interaction Database***

The NCI-Nature Pathway Interaction Database [38] is created in collaboration between the US National Cancer Institute and Nature Publishing Group. Pathway content is curated and peer-reviewed. Curation is performed by editors of the Nature Publishing Group. The main focus lies on human signaling and regulatory networks.

Pathways are principally derived from research publications, but amended with information from Reactome. Additionally, all pathways from BioCarta are incorporated. (BioCarta is an older ad-supported database of pathway diagrams, but provides fewer features and no standard pathway file formats, so it is a less attractive option).

All data in NCI-Nature PID is freely available without restriction on use. Data is provided in both a custom XML format and BioPAX Level 2 formats.

***eMIM***

eMIM forms a small database of pathways, constructed in-house in the Molecular Interaction Map notation. The focus is on signaling pathways in humans. It is hard to make an exact representation of signalling pathways because of the large number of possible states (e.g. phosphorylation states) of many important signaling proteins. The eMIM database contains pathways that are a test-bed for detailed representation of complex signaling pathways.

Pathways can be downloaded in various raster and vector graphics formats. Export to standard formats is not currently available. The online version of the diagram is linked to annotations about the various biomolecules, including literature references. These annotations can be found in text or by clicking on the diagram.

The eMIM database is small, but it is the largest collection of diagrams in the MIM format. In fact this is the only database consistently using a single graphical notation. A database of SBGN diagrams has not yet emerged.

***BioCyc/MetaCyc***

MetaCyc/BioCyc is one of the oldest comprehensive pathway databases. MetaCyc contains metabolic pathways curated from primary scientific literature. BioCyc contains organism specific annotations linked to MetaCyc pathways. Thus, pathways in MetaCyc are templates, whereas pathways in BioCyc are filled in with species-specific details.

BioCyc databases are independent but all share the same database schema, which makes them interoperable.

MetaCyc/BioCyc has its origin in the annotation of microbial genomes. As a result, the bacterium *E. coli* is the most extensively annotated BioCyc database. In higher species, *A. thaliana* and *H. sapiens* are the most expansive.

The curation model of MetaCyc is similar to Reactome in its use of a strict but public and documented curation process. A species can be adopted by an organization, meaning that they will be responsible for curation of the corresponding BioCyc database. Data is available without restrictions. However, unlike Reactome the editor software is not freely available and cannot be improved by external contributors. EcoCyc/BioCyc supports BioPAX and SBML export, but does not yet support a standard graphical notation.

## Conclusion

Projects related to pathway diagrams have undergone strong development in recent years, which has resulted in the construction of databases, development of graphical notations and standardization of exchange formats. It is to be expected that this development continues in the near future.

This review should have provided an overview of the current state of these projects. Questions to how pathway diagrams are defined, how they are formalized and how they can be stored and exchanged should have been answered. One important question was not addressed here though: what can we do with a pathway diagram? That question will be the focus of the remaining part of this thesis.

## References


1. **Glycolysis and Gluconeogenesis (Mus musculus) - WikiPathways** [http://wikipathways.org/index.php/Pathway:WP157]
2. **BioPAX Level 3 Specification** [http://www.biopax.org/release/biopax-level3-documentation.pdf]
3. Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC: **Getting started in biological pathway construction and analysis**. *PLoS Comput Biol* 2008, **4**(2):e16.
4. Demir E, Babur O, Dogrusoz U, Gursoy A, Nisanci G, Cetin-Atalay R, Ozturk M: **PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways**. *Bioinformatics* 2002, **18**(7):996-1003.
5. Babur O, Dogrusoz U, Demir E, Sander C: **ChiBE: interactive visualization and manipulation of BioPAX pathway models**. *Bioinformatics* 2010, **26**(3):429-431.
6. Becker MY, Rojas I: **A graph layout algorithm for drawing metabolic pathways**. *Bioinformatics* 2001, **17**(5):461-467.
7. Schreiber F, Dwyer T, Marriott K, Wybrow M: **A generic algorithm for layout of biological networks**. *BMC Bioinformatics* 2009, **10**:375.
8. Barsky A, Gardy JL, Hancock RE, Munzner T: **Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation**. *Bioinformatics* 2007, **23**(8):1040-1042.
9. Kohn KW: **Molecular interaction map of the mammalian cell cycle control and DNA repair systems**. *Mol Biol Cell* 1999, **10**(8):2703-2734.
10. Adriaens ME, Jaillard M, Waagmeester A, Coort SL, Pico AR, Evelo CT: **The public road to high-quality curated biological pathways**. *Drug Discov Today* 2008, **13**(19-20):856-862.
11. Oda K, Kitano H: **A comprehensive map of the toll-like receptor signaling network**. *Mol Syst Biol* 2006, **2**:2006 0015.
12. Calzone L, Gelay A, Zinovyev A, Radvanyi F, Barillot E: **A comprehensive modular map of molecular interactions in RB/E2F pathway**. *Mol Syst Biol* 2008, **4**:173.
13. Oda K, Matsuoka Y, Funahashi A, Kitano H: **A comprehensive pathway map of epidermal growth factor receptor signaling**. *Mol Syst Biol* 2005, **1**:2005 0010.
14. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C *et al*: **NetPath: a public resource of curated signal transduction pathways**. *Genome Biol*, **11**(1):R3.
15. Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM *et al*: **The Systems Biology Graphical Notation**. *Nat Biotechnol* 2009, **27**(8):735-741.
16. Kohn KW, Aladjem MI, Weinstein JN, Pommier Y: **Molecular interaction maps of bioregulatory networks: a general rubric for systems biology**. *Mol Biol Cell* 2006, **17**(1):1-13.
17. Lazebnik Y: **Can a biologist fix a radio?--Or, what I learned while studying apoptosis**. *Cancer Cell* 2002, **2**(3):179-182.

18. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al*: **Reactome knowledgebase of human biological pathways and processes**. *Nucleic Acids Res* 2009, **37**(Database issue): D619-622.
19. Gauges R, Rost U, Sahle S, Wegner K: **A model diagram layout extension for SBML**. *Bioinformatics* 2006, **22**(15):1879-1885.
20. Deville Y, Gilbert D, van Helden J, Wodak SJ: **An overview of data models for the analysis of biochemical pathways**. *Brief Bioinform* 2003, **4**(3):246-259.
21. Pirson I, Fortemaison N, Jacobs C, Dremier S, Dumont JE, Maenhaut C: **The visual display of regulatory information and networks**. *Trends Cell Biol* 2000, **10**(10):404-408.
22. Cook DL, Farley JF, Tapscott SJ: **A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems**. *Genome Biol* 2001, **2**(4):RESEARCH0012.
23. Kohn KW: **Molecular interaction maps as information organizers and simulation guides**. *Chaos* 2001, **11**(1):84-97.
24. Aladjem MI, Pasa S, Parodi S, Weinstein JN, Pommier Y, Kohn KW: **Molecular interaction maps--a diagrammatic graphical language for bioregulatory networks**. *Sci STKE* 2004, **2004**(222):pe8.
25. Kitano H, Funahashi A, Matsuoka Y, Oda K: **Using process diagrams for the graphical representation of biological networks**. *Nat Biotechnol* 2005, **23**(8):961-966.
26. Kitano H: **A graphical notation for biochemical networks**. *BIOSILICO* 2003, **1**(5):169-176.
27. Bauer-Mehren A, Furlong LI, Sanz F: **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Mol Syst Biol* 2009, **5**:290.
28. Bader GD, Cary MP, Sander C: **Pathguide: a pathway resource list**. *Nucleic Acids Res* 2006, **34**(Database issue):D504-506.
29. Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P *et al*: **Minimum information requested in the annotation of biochemical models (MIRIAM)**. *Nat Biotechnol* 2005, **23**(12):1509-1515.
30. Laibe C, Le Novere N: **MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology**. *BMC Syst Biol* 2007, **1**:58.
31. Stromback L, Jakoniene V, Tan H, Lambrix P: **Representing, storing and accessing molecular interaction data: a review of models and tools**. *Brief Bioinform* 2006, **7**(4):331-338.
32. Stromback L, Lambrix P: **Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX**. *Bioinformatics* 2005, **21**(24):4401-4407.
33. Stromback L, Hall D, Lambrix P: **A review of standards for data exchange within systems biology**. *Proteomics* 2007, **7**(6):857-867.
34. Luciano JS: **PAX of mind for pathway researchers**. *Drug Discov Today* 2005, **10**(13):937-942.

35. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
36. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Res* 2008, **36**(Database issue):D480-484.
37. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: **WikiPathways: pathway editing for the people**. *PLoS Biol* 2008, **6**(7):e184.
38. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic Acids Res* 2009, **37**(Database issue): D674-679.
39. **eMIM resource** [http://discover.nci.nih.gov/mim/index.jsp]
40. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2010, **38**(Database issue):D473-479.
41. Ott MA, Vriend G: **Correcting ligands, metabolites, and pathways**. *BMC Bioinformatics* 2006, **7**:517.

Chapter 3

# Presenting and exploring biological pathways with PathVisio

Martijn P van Iersel[1] , Thomas Kelder[1] , Alexander R Pico[2,3], Kristina Hanspers[2,3] Susan Coort[1], Bruce R Conklin[2,3], Chris Evelo[1]

[1]Department of Bioinformatics – BiGCaT, Maastricht University, the Netherlands.
[2]Gladstone Institute of Cardiovascular Disease, San Francisco, USA.
[3]Departments of Medicine, and Molecular and Cellular Pharmacology, University of California San Francisco, USA.

*BMC Bioinformatics 2008,* ***9****:399*

## Abstract

**Background:** Biological pathways are a useful abstraction of biological concepts, and software tools to deal with pathway diagrams can help biological research. PathVisio is a new visualization tool for biological pathways that mimics the popular GenMAPP tool with a completely new Java implementation that allows better integration with other open source projects. The GenMAPP MAPP file format is replaced by GPML, a new XML file format that provides seamless exchange of graphical pathway information among multiple programs.

**Results:** PathVisio can be combined with other bioinformatics tools to open up three possible uses: visual compilation of biological knowledge, interpretation of high-throughput expression datasets, and computational augmentation of pathways with interaction information. PathVisio is open source software and available at http://www.pathvisio.org.

**Conclusion:** PathVisio is a graphical editor for biological pathways, with flexibility and ease of use as primary goals.

## Background

The concerted actions of genes, proteins, and metabolites are often conceptualized as pathway diagrams. Pathways represent a familiar concept in biological research, and software designed to work with pathways can help the researcher to organize information related to research questions. Here we present PathVisio, a visualization tool for managing biological pathways, and show several ways in which this tool can facilitate the process of doing research.

It is often said that an image is worth a thousand words, and this is especially true for describing complex interactions among biomolecules. Pathways are commonly used with great effect as teaching aides in textbooks, as notes in lab journals, and as explanatory slides in research presentations. However, a pathway that is drawn in a notebook or presentation software is just a static image. The usefulness of a pathway could be increased dramatically if the software knows something about the biological entities it represents. For example, one could click on entities of a pathway to view the Ensembl page of a relevant gene in a browser window. Ideally, textbook pathways could be combined or compared with other versions of the pathway and stored in an online repository. New pathway information could be compiled from the latest experiments and discoveries. Large experimental datasets would be more understandable through pathways. Clearly, user-friendly software would be helpful for dealing with biological pathways.

Such a tool is currently available and popular with bench biologists: the GenMAPP software suite [1] (including MAPPFinder [2] for finding biologically relevant pathways). However, GenMAPP has certain limitations in its design that make it difficult to extend further.

A new program, PathVisio, is based on many design principles derived from GenMAPP, but uses more flexible programming methods that allow for completely new features and increase the possibilities for future extension. PathVisio is designed to augment the GenMAPP software suite, replacing some but not all aspects of GenMAPP. PathVisio is suitable for the creation and exploration of pathways, while relying on GenMAPP for visualization of experimental data and MAPPFinder for statistical analysis*.

PathVisio improves GenMAPP in four separate areas. First, PathVisio is written in the Java programming language as opposed to Visual Basic. Thus, PathVisio is easier to integrate with other scientific software (often written in Java), and it enables integration of PathVisio with web technologies such as Java applets and Java Webstart. PathVisio already integrates well with some other Java-based scientific tools such as Cytoscape[3] for network analysis and Eu.Gene [4] for pathway statistics. Second, PathVisio uses a newly designed file format for storing pathway information that is extensible (XML-

* Version 2.0 of PathVisio, released in 2009, does include data visualization features that improve upon those found in GenMAPP. See also chapter 6 of this thesis.

based) yet at the same time backwards compatible with the MAPP format used by GenMAPP. This means that the existing GenMAPP pathway archive can be used in PathVisio and other GPML-compliant programs. GPML has already been extended with new shapes and the capability to define relationships between nodes, allowing a network view of the pathway. Because of the novel features of GPML, it is preferable to use this format for pathway storage even if the MAPP format is used in later analysis steps. Third, the data model is conceptually separated from the rest of the application. This enables the implementation of "copy," "paste," and "undo" commands, which are expected in modern user interfaces, yet are absent in GenMAPP. Abstraction of the data model also makes it easier to support different pathway file formats and image formats. PathVisio can be used to prepare illustrations suitable for publications with its vector graphics export feature. Finally, PathVisio re-implements GenMAPP's underlying gene databases with a new optimized database schema. These four technical improvements make the software more flexible, and open the possibility of new functionality and better integration with other tools.

The most important aspect of GenMAPP that has been mimicked in PathVisio is that the software places the biologist at the center. We chose to use manual instead of automatic layout, to emulate presentation software that may be already familiar to the user. We chose locally installable synonym databases to make cross-referencing gene identifiers quick and automatic. These design choices make PathVisio very user-friendly and diminish the need for a specialized intermediate person, who often has less knowledge of the biological problem.

## Implementation

As noted above, the data model of PathVisio is completely separated from the rest of the application. It consists of two parts: the pathway data model and the synonym database model.

In the pathway data model (Figure 1), there are three main types of objects in a pathway: *DataNode* objects represent biological entities, *Line* objects represent various types of interactions, and *Shape* objects serve as graphical annotation. The term DataNode is similar to the GenMAPP term GeneProduct, but we chose to use the former to show that it can be used to refer to any type of biological entity, not just genes and proteins. DataNodes can be grouped to show that they are biologically related (e.g., for paralogous genes or proteins in a complex).

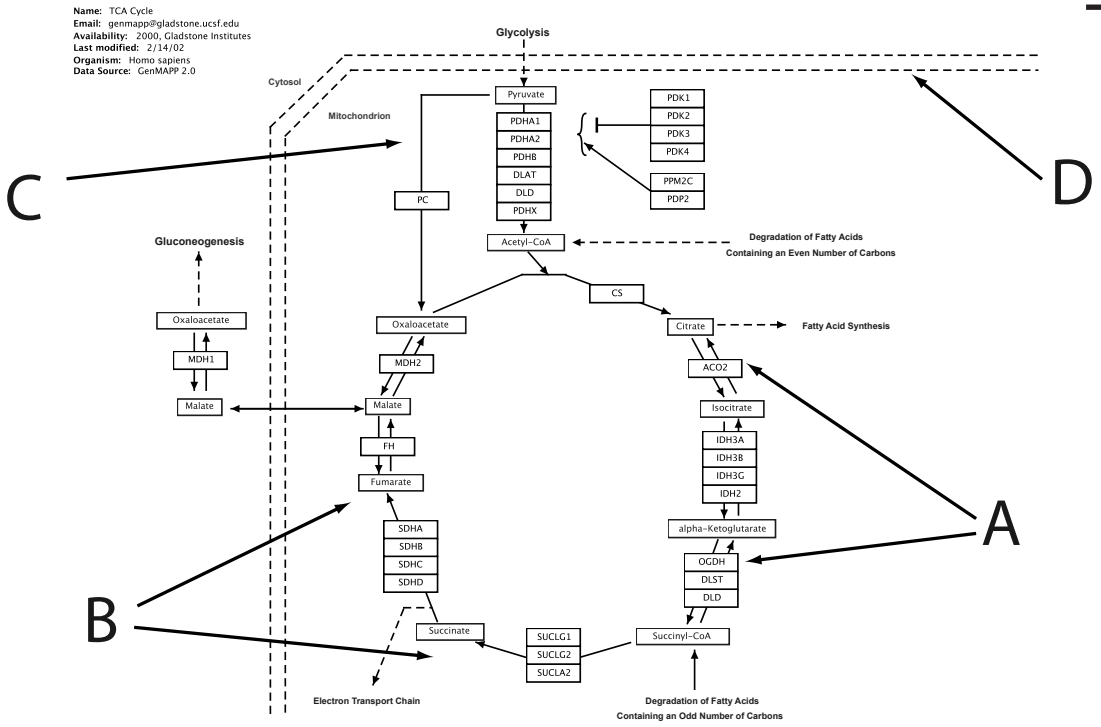


***Figure 1: Pathway data model.*** *A biological pathway as represented by PathVisio has three main classes of objects: DataNodes, Lines and Shapes. The most important are the DataNodes, represented by boxes. These data nodes can represent genes, proteins (A), or metabolites (B). DataNodes can be linked to an online database; in this example, MDH2 is linked to Entrez gene accession no 4191, and Malate is linked to HMDB identifier HDMB03256. DataNodes can be grouped to represent certain biological relationships. In this example, IDH3A, IDH3B and IDH3G are grouped to indicate that they form three subunits of a protein complex. A second class of objects is formed by lines, t-bars, and arrows that represent interactions between data nodes (C). Various shapes and text labels can be used to explain the pathway. In this example, shapes are used to distinguish the cytosol from the mitochondrion. (D). The GPML file corresponding to this pathway can be found in the supplemental data*[*]*. The PathVisio source code includes an XML Schema definition that can be used for checking the validity of GPML files.*

* Additional files are available at the BMC Bioinformatics website at http://www.biomedcentral.com/1471-2105/9/399

To store this pathway model, we developed GPML or GenMAPP Pathway Markup Language. GPML is backwards compatible with the GenMAPP MAPP format, meaning that all information stored in MAPP format can be stored in GPML as well, and pathways can be readily converted. GPML has a set of extensions on top of the initial requirement of compatibility. It allows storing relations among different elements, so that a graph can be derived from a pathway. A pathway using this facility can be converted into a network, something that is impossible with the MAPP format. The utility of this feature will be discussed below. It also can link to BioPAX [5, 6]. BioPAX is an emerging pathway standard for exchange of pathway data. The current version of BioPAX (level 2) lacks the ability to store coordinates and graphical annotations that are part of the GenMAPP format. By linking GPML elements to BioPAX elements, both are extended in XML fashion. In this way, GPML could be used as a presentation layer on top of BioPAX. To handle various pathway file formats, PathVisio makes use of a generic import /export interface.

The second part of the model used in PathVisio is the synonym database model. A variety of online genome databases are available to the bioinformatics community, leading to multiple identifiers for the same gene. One solution to this problem would have been to standardize on a specific database and let users make use of external services such as DAVID [7] to translate between ID types. This extra step for the user can be cumbersome and prone to errors. PathVisio uses another solution also implemented by GenMAPP, letting the software handle the translation through synonym databases.

Synonym databases (called gene databases in GenMAPP) can be downloaded from the website pathvisio.org. Because this type of database is potentially used intensively, we chose to create locally installable versions rather than relying on a slow web-service. The synonym database schema (Figure 2) consists of three tables: “Info”, “DataNode” and “Link”. Info provides meta-data on the database. DataNode provides per-gene information, including a short description. Link provides a many-to-many relation between entries in the DataNode table that is used to store cross-references.

Synonym databases that we produce are based on Ensembl [8] and can in principle be made for any species that is annotated in that database. They are produced based on the Derby relational database system [9], because Derby can be packaged with the PathVisio software making installation easier. However, PathVisio is not tied to a specific database back-end. Depending on speed, usability and cost requirements, a different embedded or client-server database system could be used through the Java DataBase Connectivity (JDBC) layer. The use of synonym databases is not restricted to gene information; metabolites can be used as well to unify PubChem[10], ChEBI[11] and HMDB[12].

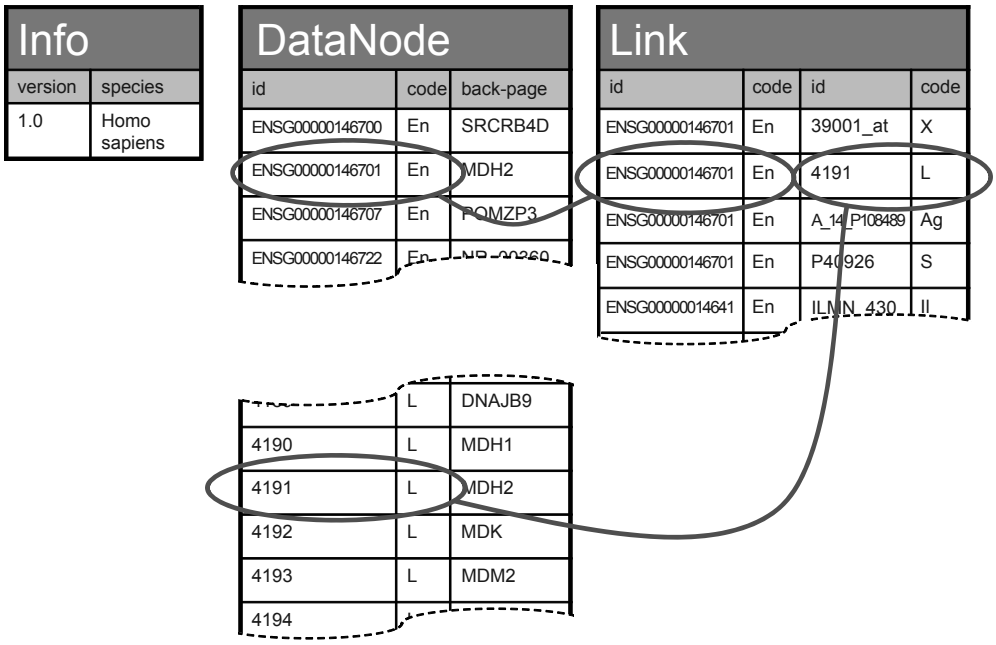


***Figure 2: Synonym database schema.*** *This figure represents the database schema for the synonym databases. There are three tables, Info, DataNode and Link. Info provides meta-data on the database. DataNode provides per-gene information, and Link provides a many-to-many relation between entries in the DataNode table that is used to store cross-references. The id column contains accession numbers from online biological databases. Systemcode is a short code (usually one or two letters) that represents the biological database that the id belongs to. A back-page is a short summary of the entity with HTML mark-up, containing at least a one-line description.*

For viewing pathways, we chose an implementation based on the Java Graphics2D API, which makes it possible to output to screen as well as various file types. The flexibility of this API makes it trivial to add Shape types in the future. The Batik SVG Toolkit [13] is used for exporting to graphic formats, including those suitable for publication.

## Results and discussion

We envision three ways in which PathVisio could aid biologists doing research.

***1: Organization of biological information***

Many research questions are related to biological pathways in some way. For example, which receptor is responsible for carrying a stress signal across the nuclear membrane? Which proteins need to be activated to lead to a choice between two possible differentiated cell types? In these cases, a question is asked about a certain unknown component of a biological pathway. Experimental results could lead to a conclusion in terms of adding new elements to a pathway, clarifying the role of an element in a pathway, or proving

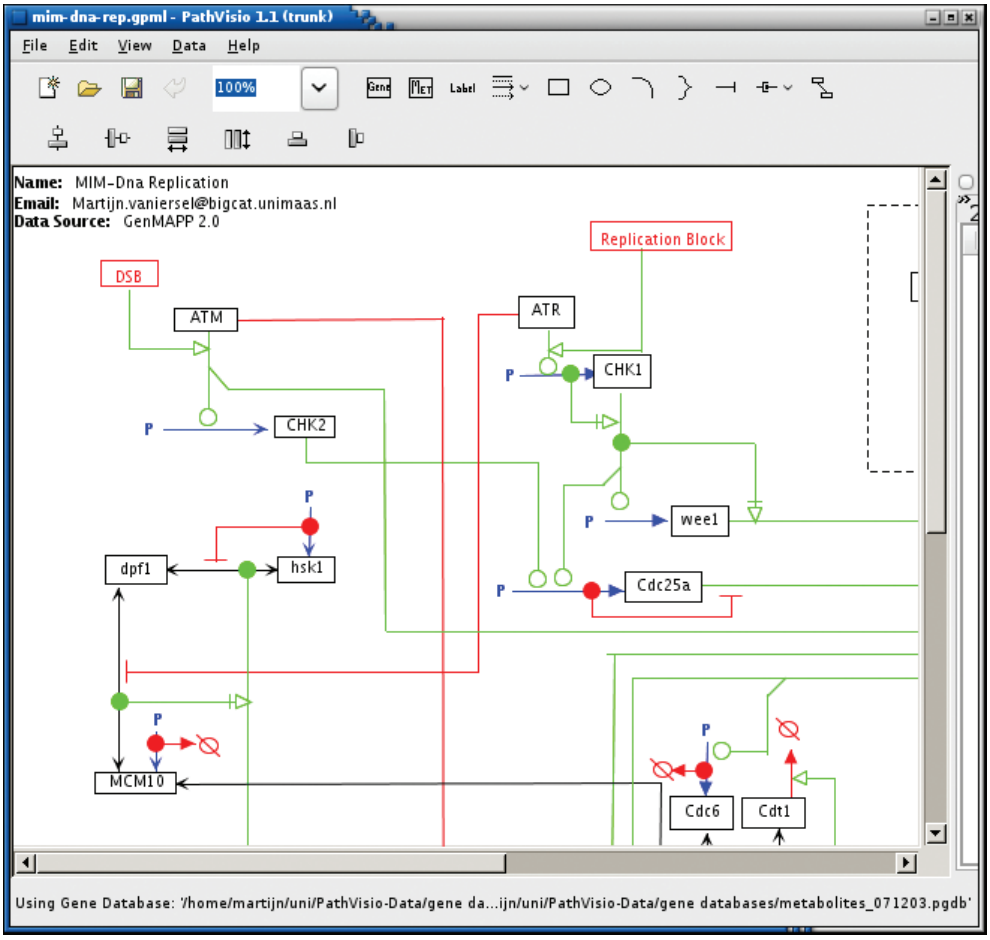


***Figure 3: Example of a Molecular Interaction Map drawn in PathVisio.*** *The pathway file itself is included in the supplemental material.*

the existence of an interaction between two elements. In all cases, biological knowledge is increased, and since pathways are a representation of this knowledge, the pathway itself is improved as a direct result of research. Expressing biological knowledge visually as a pathway can be a very powerful tool to organize disparate bits of information.

What is the best way to represent biological concepts graphically? There are certain conventions for this, as well as several published formalized symbolic languages [14-16]. The style used by GenMAPP and many textbook pathways does not pose many restrictions (e.g., an arrow can be used to mean stimulation, transport to a new compartment or simply interaction between proteins).

Molecular Interaction Maps (MIMs) [16] is one attempt to codify biological knowledge in schematics. The advantage of MIMs is that they store a large amount of information in a single diagram, and a knowledgeable person can retrieve all this information from the diagram with certainty.

PathVisio 1.0 has some support for MIMs; many of the needed graphical elements can be drawn. Work is underway to make this support complete, including making complex aspects of MIMs, such as contingency arrows, available in a user-friendly manner (Figure 3). PathVisio is the first graphical editor with support for the MIM style. We hope other editors will support MIMs in the future, so that they will be established as a standard.

Kitano [17] proposed that a pathway schematic must be semantically and visually unambiguous. The requirement of unambiguously defining a pathway is necessary for computational simulation, but is simply impossible for most biological pathways. The complexity of these data can be confusing to the biologist or teacher who is attempting to convey fundamental aspects of the pathway. For example, the combinatorial explosion that arises when dealing with a protein that can be in different phosphorylation states can increase the complexity of the diagram. A trade-off exists between clarity and completeness. At the same time, the requirement for unambiguity hampers the iterative process by which a pathway can be compiled while research is ongoing. Many components of a pathway are unknown and unspecified as long as research has not yet provided a full mechanistic explanation of the subject. Kitano resolves this by proposing a reduced notation; similarly, Kohn [16] has added a distinction between heuristic and explicit maps. GPML allows ambiguity, and doesn't enforce a particular style of notation. By being flexible, GPML allows a pathway to progress seamlessly from a rough conceptualization to a well established pathway that can be modeled. As increasing levels of complexity are understood, any of the aforementioned styles can be used depending on the intended use of the pathway.

Biological knowledge stored in a collection of pathways is most useful if it is available online and frequently updated. An applet version of PathVisio is used for the WikiPathways [18] resource, where the research community can collaborate in the task of curating and updating pathway content.

***2: Data analysis and pathway statistics***

Data from high-throughput experiments such as microarrays can be combined with pathways to achieve new insights. Using pathways one can view data in its biological context rather than in an arbitrarily ordered table. Once a set of pathways have been collected in a repository, the possibility to do pathway statistics becomes available.

PathVisio supports this workflow in combination with the GenMAPP software suite. Pathways created with PathVisio can be exported to MAPP format. Subsequently, high-throughput experimental data can be loaded into GenMAPP, and user-defined color criteria established. DataNodes on pathways in GenMAPP can then be colored accordingly, putting relevant parts of the expression dataset together in a single view.

MAPPFinder [2], another tool in the GenMAPP software suite, can be used to search for pathways that are significantly regulated under experimental conditions. MAPPFinder counts how many genes on each pathway meet user-defined criteria and compares this to the expected number of genes that meet the criteria to calculate a z-score. These z-scores can then be used to rank a set of pathways, which is very useful in hypothesis-generating experiments to identify which biological processes are affected.

Eu.Gene can provide an alternative to MAPPFinder in circumstances where something other than z-scores is required. PathVisio can export pathways to the Eu.Gene gene list format. Eu.Gene can employ Gene Set Enrichment Analysis (GSEA) or Fisher exact tests for pathway statistics.

***3: Network analysis and augmentation***

Compilation of pathway information is necessarily a manual process, but it would be very useful to augment pathway information with computational tools, including text mining, data mining and interaction information from high-throughput datasets, such as yeast-two-hybrid data.

As an improvement on GenMAPP, GPML has been extended to store node-edge relations, making it possible to store true interactions. PathVisio allows the user to define interactions by joining two items with a connector line. Fore ease of use, the connector moves together with the element to which it is connected.

The user-interface of PathVisio is optimized for the pathway model, meaning that PathVisio does not treat pathways as networks. The two concepts are closely related however, and network analysis of pathways can be useful. Other software, such as Cytoscape [3], is better suited to handle networks. Ideally, one would be able to use both programs together. Cytoscape supports a large set of plug-ins [19]. We created a GPML plug-in for Cytoscape that enables the user to transfer pathways between PathVisio and Cytoscape with copy and paste commands. This is the first step of a workflow for enhancing pathway information: 1, Create a pathway in PathVisio based on experimental results or literature research. 2, copy the pathway to Cytoscape. 3, Enhance the pathway using one of the many sources of interaction information available within Cytoscape. For example, the Agilent literature search plug-in could be used to obtain interactions from literature. The PathwayCommons plug-in makes various other sources of interaction information available. Once such interaction information is obtained, the network

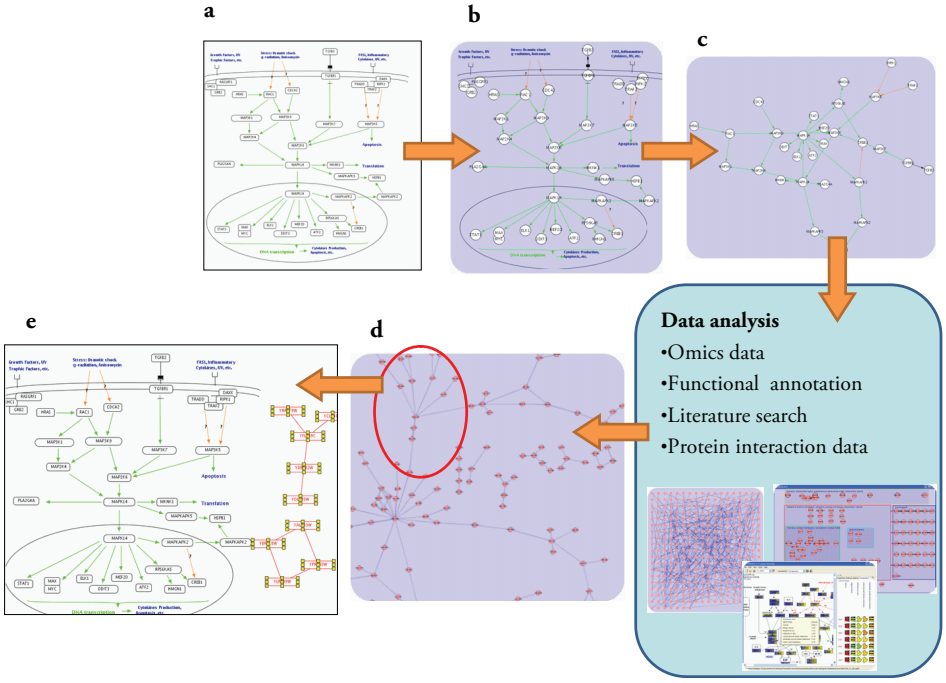


***Figure 4: Network analysis workflow.*** *This workflow suggests a way to combine manual pathway data and results from network analysis. A pathway in GPML format (a) can be copied to the Cytoscape network analysis program (b) with copy and paste commands. This program is useful for viewing the pathway as a network (c) without any extra graphical annotation. The rich plug-in set of Cytoscape can be used in various ways to extend this network. If the extended network contains new information that turns out to be instructive, it may be useful to transfer part of this network (d) back to PathVisio again (e) to add manual layout and graphical annotation.*

can be further enhanced, for example by grouping nodes by the cellular component they occur in using the BubbleRouter plug-in. 4, Copy the enhanced network back to PathVisio. 5, Re-arrange the network manually in PathVisio for presentation or publication. See figure 4 for a schematic overview of this workflow. All plug-ins mentioned in this workflow, including the GPML plug-in, can be downloaded directly using the Cytoscape plug-in manager.

### *Future*

For a specialized tool such as PathVisio to be relevant, it must be tightly integrated with other bioinformatics tools and standards. PathVisio can already be used in combination with Cytoscape through the GPML plug-in, and it is compatible with GenMAPP MAPP format and MAPPFinder for statistical analysis.

There currently exists a set of partly overlapping pathway format standards [6], and in our view, it is better to improve existing standards than to add new ones. As long as no

pathway format completely solves all problems, the second best solution is to maximize compatibility and interoperability, and a move from a binary format to an XML-based format is a step in the right direction. GPML is intended to be a backwards compatible extension of the older, less flexible MAPP format. This step makes an older format easier to convert and more interoperable with other standards.

To clarify the role of GPML, we can compare it to other existing standards related to pathway definitions. BioPAX is a standard designed for exchanging data between pathway databases. GPML can embed elements from a BioPAX document and add visual annotations to it. This capability could facilitate integration of PathVisio with other pathway sources, such as Reactome, in the future. GPML is not suitable for computational modeling. Systems Biology Markup Language (SBML) is designed specifically for that purpose, and the overlap in scope of GPML and SBML is small. SBML defines reactions and parameters necessary for computational modeling. GPML instead emphasizes links to online databases, as these are valuable for human interpretation.

Besides GenMAPP, a few other non-commercial tools for pathway visualization, such as CellDesigner [20] and VisANT [21], already exist. Like Cytoscape, VisANT is oriented towards a network view, and since it supports a plug-in interface, a similar GPML transfer mechanism might be implemented in the future.

PathVisio is fully open source, and GPML is an open format. We see open source as a necessity for this type of bioinformatics tool. Open source makes it possible for other tools to adapt to PathVisio and vice versa. With closed-source tools (e.g. CellDesigner, VisANT and commercial packages), this adaptation can only go one way, which is a strong disincentive to collaborate. In a field where integration is of utmost importance, open-source software provides an optimal solution. To further encourage cooperative development, we will add support for plug-ins, something that is facilitated by the Java programming language.

With PathVisio and GPML we have developed a framework for visual pathway analysis. This framework is very flexible with future extensions in mind. Development of PathVisio is ongoing. We wish to continue in the direction of increased flexibility and tighter integration with other bioinformatics standards and applications. This should ensure that pathway analysis and visualization can be done efficiently to improve biological research.

## Availability & Requirements

Project Name: PathVisio

Project Home Page: http://www.pathvisio.org

Operating System: tested on Windows XP. Versions for Linux and Mac OS X are under development.

Programming Language: Java

Other Requirements: Java 5 or higher

License: Free and open source under the Apache 2.0 License. There are no restrictions to use by non-academics. The source code is available at http://svn.bigcat.unimaas.nl/pathvisio.

## Author's contributions

All authors have read and agreed upon the content of this article. The PathVisio software was designed and written by MvI, TK, AP and KH. SC performed invaluable beta testing. AP and BC created the GPML concept, CE the PathVisio concept. MvI and SC drafted the paper.

## Acknowledgements

We thank Lynn Ferrante for the first work on the XML Schema definition of GPML, Gontran Zepeda for work on MAPP import/export, and Rene Besseling, Sjoerd Crijns, Margriet Palm, Erik Pelgrim and Hakim Achterberg for help in PathVisio development. This work was supported by grants from the NIH (GM080223, HG003053), the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre and funding from Transnational University Limburg (tUL).

## References

1. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR: **GenMAPP 2: new features and resources for pathway analysis**. *BMC Bioinformatics* 2007, **8**:217.
2. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data**. *Genome Biol* 2003, **4**(1):R7.
3. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.
4. Cavalieri D, Castagnini C, Toti S, Maciag K, Kelder T, Gambineri L, Angioli S, Dolara P: **Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases**. *Bioinformatics* 2007, **23**(19):2631-2632.
5. **BioPAX wiki** [http://www.biopax.org/cgi-bin/moin.cgi]
6. Cary MP, Bader GD, Sander C: **Pathway information for systems biology**. *FEBS Lett* 2005, **579**(8):1815-1820.

7. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
8. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: **Ensembl 2008**. *Nucleic Acids Res* 2008, **36**(Database issue):D707-714.
9. **Apache Derby** [http://db.apache.org/derby]
10. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**(Database issue): D13-21.
11. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic Acids Res* 2008, **36**(Database issue):D344-350.
12. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S *et al*: **HMDB: the Human Metabolome Database**. *Nucleic Acids Res* 2007, **35**(Database issue):D521-526.
13. **Batik SVG Toolkit** [http://xmlgraphics.apache.org/batik/]
14. Kitano H, Funahashi A, Matsuoka Y, Oda K: **Using process diagrams for the graphical representation of biological networks**. *Nat Biotechnol* 2005, **23**(8):961-966.
15. Pirson I, Fortemaison N, Jacobs C, Dremier S, Dumont JE, Maenhaut C: **The visual display of regulatory information and networks**. *Trends Cell Biol* 2000, **10**(10):404-408.
16. Kohn KW, Aladjem MI, Weinstein JN, Pommier Y: **Molecular interaction maps of bioregulatory networks: a general rubric for systems biology**. *Mol Biol Cell* 2006, **17**(1):1-13.
17. Kitano H: **A graphical notation for biochemical networks**. *BIOSILICO* 2003, **1**(5):169-176.
18. **WikiPathways** [http://www.wikipathways.org]
19. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B *et al*: **Integration of biological networks and gene expression data using Cytoscape**. *Nat Protoc* 2007, **2**(10):2366-2382.
20. Funahashi A, Tanimura N, Morohashi M, Kitano H: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks**. *BIOSILICO* 2003, **1**:159-162.
21. Hu Z, Ng DM, Yamada T, Chen C, Kawashima S, Mellor J, Linghu B, Kanehisa M, Stuart JM, DeLisi C: **VisANT 3.0: new modules for pathway visualization, editing, prediction and construction**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W625-632.

Chapter 4

# WikiPathways: Pathway Editing for the People

Alexander R. Pico[1,2,*], Thomas Kelder[3,*], Martijn P. van Iersel[3], Kristina Hanspers[1,2], Bruce R. Conklin[1,2], Chris Evelo[3]

[*] These authors contributed equally to this work.

[1]Gladstone Institute of Cardiovascular Disease, San Francisco, USA.
[2]Departments of Medicine, and Molecular and Cellular Pharmacology, University of California San Francisco, USA.
[3]Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands.

*Plos Biology 2008* **6***:e184*

The exponential growth of diverse types of biological data presents the research community with an unprecedented challenge and opportunity. The challenge is to stay afloat in the flood of biological data, keeping it as accessible, up-to-date, and integrated as possible. The opportunity is to cultivate new models of data curation and exchange that take advantage of direct participation by a greater portion of the community.

This combination of challenge and opportunity is especially relevant to the task of collecting biological pathway information. Pathways are critical to understanding the functions of individual genes and proteins in terms of systems and processes that contribute to normal physiology and to disease. Each biological pathway must be hewn from a mass of biological information distributed across multiple publications and databases.

The particular challenge of pathway curation is amplified because pathways are often presented as static images that are not amenable to computation, integration, or data exchange. Furthermore, pathway experts are distributed throughout the world, and most have limited time to learn about complex databases that need their expertise. This challenge can be met by taking the opportunity to develop a new community-based model for pathway curation.

One way to engage the community is with a wiki model as exemplified by Wikipedia [1]. We see the potential for a wiki-based pathway curation resource, coupled with an embedded graphical pathway editing tool, to meet the growing challenge presented by the influx of biological data and to provide an innovative example of content curation by the biology community (Figure 1).

## Facing the Challenge

The research community is experiencing massive growth in biological data, from genome and metagenome sequencing to high-throughput assays and microarray studies. This growth has created a need for models of data storage and distribution that support a continuous stream of end-user submissions, frequent updates, integrated search across databases, and access to data formats (preferably community standards) that are amenable to computational analyses. By and large, the need is being met for certain types of biological data: sequences go to GenBank and EMBL-Bank, protein structures go to the Protein Data Bank (PDB), and microarray results go to Gene Expression Omnibus (GEO) and ArrayExpress. But as the influx and complexity of biological data continues to grow, so will the challenge of organizing and maintaining these databases.

Fortunately, the biology community can provide an answer that will scale with the challenge: *community curation*. There is a growing tendency toward information exchange that supports open access, higher-order organization, community-defined data formats and collaborative online environments. This trend is most apparent with the growing

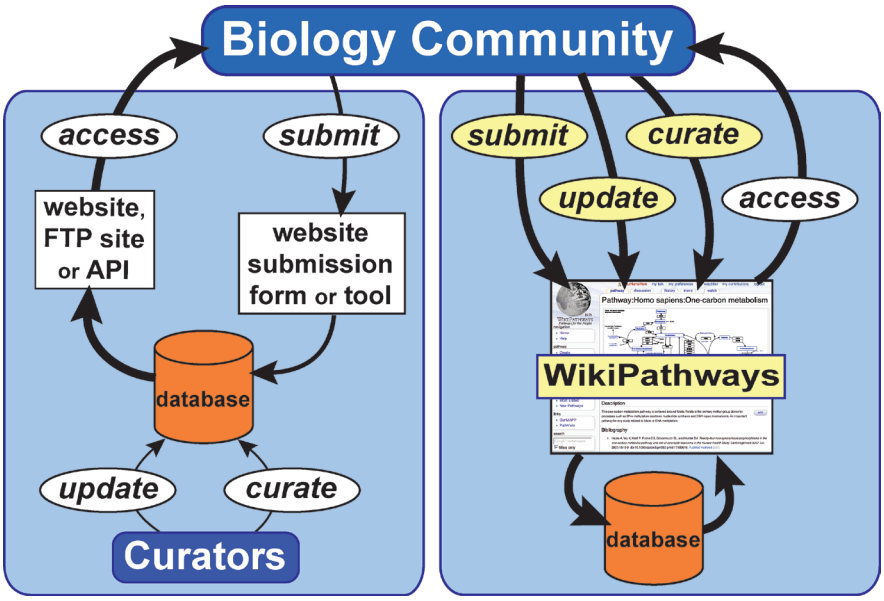


***Figure 1. Two models for managing biological data.*** *Current biological databases provide the community with data submission forms or tools and access to the compiled data via websites, FTP sites and, sometimes, programmatic interfaces (API). Internal curation teams organize and update the data. A wiki model for biological databases, such as WikiPathways, provides a single, intuitive interface for submitting, updating, organizing, and accessing data, allowing the community to participate in the curation process and keep up with the influx of new data. The widths of the arrows represent the relative capacity for data management in the two models.*

number of open access journals [2], public databases [3], and data exchange formats [4] and ontologies [5]. To promote community curation, database maintainers must be willing to relinquish some control. After removing logistical as well as technical barriers with creative support tools, the data producer will also be the data organizer. Despite initial misgivings, such projects do succeed with the right balance of infrastructure, participation and administrative principles, as demonstrated by Wikipedia [1], numerous open source software projects (e.g., Linux, Apache, MySQL, Firefox), and countless scientific collaborations, including the Internet, itself. The idea of using wiki technology for biological information has been proposed in other areas, for example, genome annotation [6,7]. The EcoliWiki provides a working example of community curation focused on *E.coli* [8]. And WikiProteins aims to combine automated text mining with community curation to annotate biomedical concepts, including protein functions, interactions and disease relationships [9].

## Representing Biological Pathways

Biological pathways present a special case in which the information is not directly coupled to data collection. One does not sequence or measure a pathway. Pathways comprise a myriad of interactions, reactions, and regulations, often identified piecemeal over extended periods and by a variety of researchers. As a result, pathway information is particularly challenging to compile and curate. Furthermore, biological pathways are often captured *only* as static images for publications or presentations. Consider the pathway illustrations common in textbooks and review articles that document any given field of biology. A typical signaling pathway, for example, represents receptor-binding events, protein complexes, phosphorylation reactions, translocations, and transcriptional regulation, with only a minimal set of symbols, lines and arrows. While these simple images are powerful visual and conceptual representations, they cannot be connected to relevant biological annotations or analyzed with respect to experimental data.

A number of groups have taken on the challenge of curating and archiving biological pathways [10]. Those efforts mainly rely on internally supported teams of biologists or contracts with volunteer experts in particular fields of biology. Their curation tools typically require download, installation, and specialized training and are not designed for broad or collaborative use. Often, the barrier is simply too high for the average biologist to consider contributing their own pathway knowledge. Even when it is contributed, pathway information can remain untouched for years in the current databases, quickly becoming outdated and out-of-sync with the continuing stream of published discoveries. Some of us (BRC, KH, ARP) have first-hand experience with maintaining the GenMAPP [11] pathway archive in this fashion over the past 8 years. The task of submitting and updating content inevitably falls on a handful of specialists who have invested significant time installing and learning how to use the curation tools. This approach is not sustainable in the face of the growing influx of biological data. Clearly, curating all of biology is a Herculean task for any single group.

## Pathway Editing for the People

To facilitate the contribution and maintenance of pathway information by the biology community, we established WikiPathways (www.wikipathways.org). WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways. WikiPathways thus presents a new model for pathway databases that enhances and complements ongoing efforts [12-14]. Building on the same MediaWiki open source software that powers Wikipedia, we added a custom graphical pathway editing tool and integrated databases covering major gene, protein, and small-molecule systems. The familiar web-based format of WikiPathways greatly reduces the barrier to participate in pathway curation. More importantly, the open, public approach of WikiPathways al-

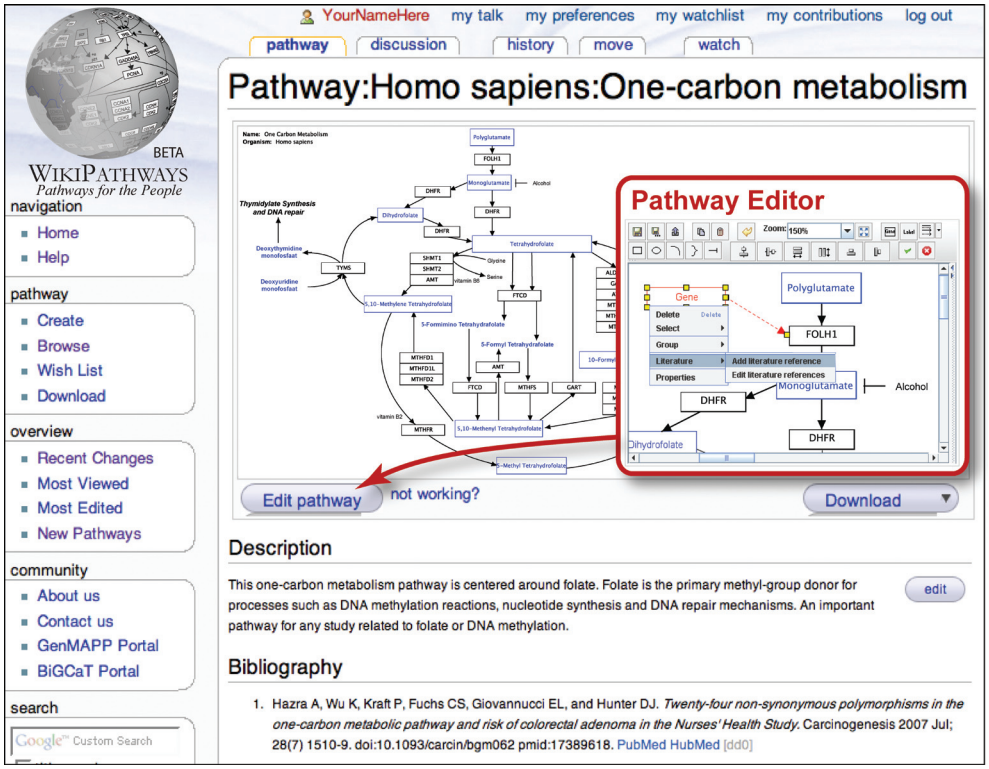


***Figure 2. Sample WikiPathways page***. *Each pathway has a dedicated page for viewing and editing content. The pathway diagram is edited with an embedded applet version of PathVisio (inset). Description and Bibliography sections can be edited in-page as well through applets that facilitate entry. Additional information about the pathway components and version history continue on the page (not shown).*

lows for broader participation by the entire community, ranging from students to senior experts in each field. This approach also shifts the bulk of peer review, editorial curation, and maintenance to the community.

Each pathway at WikiPathways has a dedicated wiki page, displaying the current diagram, description, references, version history, and component gene, protein and metabolite lists (Figure 2). Any pathway can be edited from within its wiki page by activating an embedded pathway editor. WikiPathways uses an applet version of PathVisio—a pathway-drawing tool we developed for pathway curation [15] (manuscript submitted). PathVisio provides a basic palette of objects and annotations needed to represent biological processes. Gene, protein, and metabolite objects directly map to biological annotations from multiple public databases through an extensible identifier synonym database maintained at WikiPathways. The editing tool facilitates annotation with keyword search and auto-completion. Relationships between entities can easily be drawn

using “smart connectors” that snap into place. Lines can even connect to other lines to intuitively represent catalysis or other mediated processes. Entities can be grouped to represent complexes and related collections of genes. The editing tool also makes it easy to annotate these entities and relationships with peer-reviewed literature references. The ‘Help’ section of WikiPathways provides guidelines and tutorials for how to use the editor and best represent pathway information, as well as how to download and use the pathways in GenMAPP analyses.

After editing, an updated pathway image is displayed on the wiki page along with the version history and list of components. Users can easily monitor and undo changes, compare differences and search for overlapping pathways. Any registered user can add a pathway to their “watch list” so that they receive email when the pathway is changed. All changes can be reversed, restoring the pathway to an earlier version. Different versions of pathways can be compared side-by-side using an integrated difference-viewing tool, customized for graphical pathway information. Using the search feature, one can locate particular pathways by name, by the genes and proteins they contain, or by the text displayed in their descriptions and comments. One can also browse the collection of pathways with combinations of species names and ontology-based categories. Currently, WikiPathways contains 537 species-specific pathways for human, mouse, rat, zebrafish, fruit fly, worm, and yeast. The mouse pathways, for example, contain 3741 unique genes (~13% of the mouse genome). The pathway collection was nucleated with GenMAPP pathways, which were collected over the past decade from GenMAPP users. Now at WikiPathways, the collection is growing and improving through new contributions and curation, at an unprecedented rate, which we expect to dramatically increase as community participation grows.

The pathway content at WikiPathways is freely available for download in a variety of data and image formats, including GPML, which is a custom XML format compatible with pathway visualization and analysis tools such as Cytoscape [16], GenMAPP [11] and PathVisio [14]. GPML allows researchers to draw and identify the molecular participants in a pathway, as well as the relationships among the participants. GPML is a work in progress and though it does not yet have the full expressiveness of BioPAX [17] or SBML [18], it provides the basic functionality for researchers to create appealing pathway diagrams and to perform basic statistical tests on pathways such as overrepresentation analysis. The goal of GPML is to bridge the simple elegance of a pathway drawn on a napkin by a biologist (including its rich, human interpretability) and the growing databases of gene and protein annotations, interactions and experimental data. We prioritized the development of GPML based on what is already available and what is most useful to the average biologist: connecting intuitive, human-readable graphics to standardized identifiers from popular databases. This allows users to accurately label entities on pathways and computationally map them to experimental data using pathway

analysis software. GPML also supports the representation of relationships between entities to allow network-based visualization and analysis. In a recent "community curation event" at WikiPathways we formalized network relationships in the human pathway archive. We plan to include a number of BioPAX elements into GPML to support data exchange, but the overriding goal for GPML is to lower the barrier for contributors of pathway information by keeping it simple. This approach resonates with the large portion of the biology community interested in basic statistical pathway analyses and figures for publications and presentations.

To assist pathway authors and curators, we are developing 'bots' to survey the content and identify potential inconsistencies, redundancies and incomplete data. The first of these bots identifies all the genes, proteins and metabolites in any pathway that are not connected to a synonym database identifier. These reports along with additional curator tools will help contributors to submit high-quality content and make corrections where needed. We also plan to use standard biomedical ontologies to structure the content of WikiPathways and to provide organization that can scale with rapidly growing and interrelated information.

Researchers interested in particular interactions or pathways can use WikiPathways as a resource for up-to-date pathway information and as a repository for their own findings that, in turn, are immediately available in multiple data formats for analysis as well as image formats for publication. WikiPathways can be used collaboratively to create, edit, and share pathway information with any colleague who has access to a web browser. For sensitive content that is proprietary or must first be published as an original finding, pathways can be saved locally in the GPML format, ready to be uploaded and made public at a later time. Expert curators can use WikiPathways to monitor and update pathway information associated with their fields of interest. WikiPathways is also useful to students and professors of biology, providing pathways as educational materials and the editing history of a given pathway as an example of how scientific knowledge iteratively progresses.

To encourage participation by the community we have built templates for 'User pages' and 'Portals'. User pages help users identify themselves and their work, while Portals help entire communities of users to identify themselves collectively and focus on particular pathway domains, such as diabetes-related pathways or plant pathways. By using the template you can build a site within WikiPathways dedicated to your lab, organization or area of interest within minutes. We are also organizing community curation events as a way to introduce new users to the curation tools and, at the same time, improve the quality of the pathway content. Future community curation events will focus on adding annotation, group representations, and literature references.

Even prior to this publication introducing WikiPathways, we have seen strong signs of community participation. Outside of the immediate group of developers, WikiPathways has already attracted 10 new mouse pathways, 9 new human pathways, 6 new zebrafish pathways, 3 new rat pathways and a Portal for the micro-nutrients community. There are dozens of *E. coli* and plant pathways currently being converted, and 3 new Portals under construction. The site has over 220 registered users (10% contributing users) and has attracted developers through the Google Summer of Code program.

We envision WikiPathways being part of a broader effort to extend curation capacity to larger groups and communities. This effort does not replace current approaches involving centralized teams of curators, but rather it complements and extends them. Eventually, we would like to see wiki solutions like WikiPathways utilized by current databases and curation sources. Our future directions include supporting "reference" pathways contributed by other pathway databases, and private workspaces for groups to collaboratively work on pathways before making them public. You could also imagine organizations installing local instances of WikiPathways for internal projects at research institutes or biotechnology companies. A journal, for example, could host a version of WikiPathways that only contributing authors can edit. Where the same wiki technology is used, there are opportunities for seamless integration and controlled sharing of content when it is ready to be published or released to the public site. We will continue to work towards supporting broad implementations of WikiPathways to promote contributions from established and diverse sources.

WikiPathways is an experiment. We have considerable work ahead in developing the GPML data model, implementing critical features and, most importantly, building a community of users and contributors. The success of WikiPathways will depend on the overall quality of its content, which will be a function of the infrastructure and administrative principles we employ in addition to community participation. Features such as database connectivity, automatic consistency checks, curation tools, reversible edits, the visual difference viewer, and support by literature references will assist in tracking and reverting errant contributions, stimulating curation by the community. We anticipate that lowering the entry barrier for participation will allow for a greater capacity of curation, broader consensus on content, and ultimately, higher quality control. We are confident that WikiPathways will be a powerful resource for the research community and a vital forum for pathway curation. And we are hopeful that it will serve as an example for how the continuing flood of biological data can be managed and utilized by the community to irrigate future hypotheses and discoveries.

## Source Code

We are committed to open access and open source. All content is available under a Creative Commons License (http://creativecommons.org/licenses/by-nc-sa/3.0/). All source code for WikiPathways and the PathVisio applet is available under the Apache License, Version 2.0 (http://www.apache.org/licenses/). You can download the code from:

- http://svn.bigcat.unimaas.nl/wikipathways
- http://svn.bigcat.unimaas.nl/pathvisio

## Acknowledgements

We thank Susan Coort, Nathan Salomonis, Andra Waagmeester, Grant Yang, Sam Rudy, Gary Howard, Allan Kuchinsky (Agilent Technologies), and members of the Conklin and Evelo labs for valuable advice and support.

## Funding

This work was supported by grants from the NIH (GM080223, HG003053), the Agilent Technologies Foundation, the BioRange 1.2.4 research program of the Netherlands Bioinformatics Centre, and the Google Summer of Code program.

**Competing interests statement:** The authors declare no competing financial interests.

**Abbreviations:** EMBL, European Molecular Biology Laboratory; GEO, Gene Expression Omnibus; GPML, GenMAPP Pathway Markup Language; PDB, Protein Data Bank; SBML, Systems Biology Markup Language; XML, Extensible Markup Language.

## References

1. Giles J (2005) **Internet encyclopaedias go head to head.** *Nature* **438**(7070):900-901.
2. **Directory of Open Access Journals** [http://www.doaj.org/]
3. Galperin MY (2007) **The molecular biology database collection: 2008 update.** *Nucleic Acids Res* **36**(Database issue):D2-D4.
4. Strömbäck L, Hall D, and Lambrix P (2007) **A review of standards for data exchange within systems biology.** *Proteomics* **7**(6):857-867.
5. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) **The obo foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotech* **25**(11):1251-1255.
6. Salzberg SL (2007) **Genome re-annotation: a wiki solution?** *Genome Biology* **8**:102.
7. Hu JC, Aramayo R, Bolser D, Conway T, Elsik CG, et al. (2008) **The emerging world of wikis.** *Science* **320**(5881):1289b-1290.
8. **EcoliWiki** [http://ecoliwiki.net]

9. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, et al. (2008) **Calling on a million minds for community annotation in wikiproteins.** *Genome Biology* **9**(5):R89.
10. Bader GD, Cary MP, and Sander C (2006) **Pathguide: a pathway resource list.** *Nucleic Acids Res* **34**(Database issue):D504-506.
11. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, et al. (2007). **Genmapp 2: New features and resources for pathway analysis.** *BMC Bioinformatics* **8**:217.
12. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg]
13. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, et al. (2007) **Reactome: a knowledgebase of biological pathways and processes.** *Genome Biology* **8**:R39.
14. **Pathway Commons** [http://www.pathwaycommons.org/pc/]
15. **PathVisio** [http://www.pathvisio.org]
16. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) **Integration of biological networks and gene expression data using cytoscape.** *Nat Protoc* **2**(10):2366-2382.
17. **BioPAX Wiki** [http://biopaxwiki.org/cgi-bin/moin.cgi]
18. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) **The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models.** *Bioinformatics* **19**(4):524-531.

Chapter 5

# The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services

Martijn P van Iersel[1], Alexander R Pico[2], Thomas Kelder[1], Jianjiong Gao[3], Isaac Ho[2], Kristina Hanspers[2], Bruce R Conklin[2,4] and Chris T Evelo[1]

[1]Department of Bioinformatics - BiGCaT, Maastricht University, the Netherlands.
[2]Gladstone Institute of Cardiovascular Disease, San Francisco, USA.
[3]Department of Computer Science, University of Missouri, Columbia, USA.
[4]Departments of Medicine and Cellular and Molecular Pharmacology, University of California San Francisco, USA.

*BMC Bioinformatics 2010,* **11***:5*

## Abstract

**Background:** Many complementary solutions are available for the identifier mapping problem. This creates an opportunity for bioinformatics tool developers. Tools can be made to flexibly support multiple mapping services or mapping services could be combined to get broader coverage. This approach requires an interface layer between tools and mapping services.

**Results:** Here we present BridgeDb, a software framework for gene, protein and metabolite identifier mapping. This framework provides a standardized interface layer through which bioinformatics tools can be connected to different identifier mapping services. This approach makes it easier for tool developers to support identifier mapping. Mapping services can be combined or merged to support multi-omics experiments or to integrate custom microarray annotations. BridgeDb provides its own ready-to-go mapping services, both in webservice and local database forms. However, the framework is intended for customization and adaptation to any identifier mapping service. BridgeDb has already been integrated into several bioinformatics applications.

**Conclusion:** By uncoupling bioinformatics tools from mapping services, BridgeDb improves capability and flexibility of those tools. All described software is open source and available at http://www.bridgedb.org.

## Background

Many interesting problems in bioinformatics require the integration of experimental data from different sources. Examples include merging two independently created protein-protein interaction networks in Cytoscape [1] and visualizing microarray data on a collection of biological pathways in GenMAPP [2] or PathVisio [3]. More often than not, data of different types and from different sources are annotated with different identifiers. Thus, an important step in the analysis workflow is deducing which identifiers from one set correspond to which identifiers in the other set.

This problem of identifier mapping has been recognized, and a number of resources have been developed to solve it, including DICT [4], CRONOS [5], MatchMiner [6], AliasServer [7], PICR [8], Synergizer [9] and Ensembl BioMart [10]. For the most part, these resources accurately map identifiers and provide an interface, usually a web site, to the mappings. However, each resource necessarily has a focused domain of application based on limitations in resources and in the interest of its developers. Mapping services differ in aspects, such as coverage of species, coverage of identifier types, access speed and frequency of database updates. This has created two challenges for developers of bioinformatics applications. The first challenge is to develop software that is not tied to a single identifier mapping service. Tools that are built around a single service would have to be adapted with considerable effort if a more suitable service comes along. Optimally, switching should be a simple matter of configuration. The second challenge is to combine mapping services to get the benefits of each. For example, one could combine a small mapping table of probe identifiers of a custom microarray with a large mapping resource, such as Ensembl BioMart, or one could combine metabolite mappings and gene mappings when assessing experimental data from a combination of different omics platforms.

The key to both challenges is to create a standard interface between tools and mapping services. Here we present BridgeDb, a framework that provides such a standard interface. BridgeDb is a software library intended to be used by bioinformatics tool developers. The overall architecture is described in Figure 1. BridgeDb makes is possible to write shorter and simpler code to handle identifier mapping. This framework has already been incorporated in several bioinformatics applications.

### *Concepts*

To explain the features of the BridgeDb framework, it is necessary to define some concepts. We use the term **data source** to describe a database of biological entities, indexed by unique identifiers. Usually, these identifiers are assigned and maintained internally by an independent organization. Examples of data sources include Ensembl, UniProt, ChEBI, and the Gene Ontology.

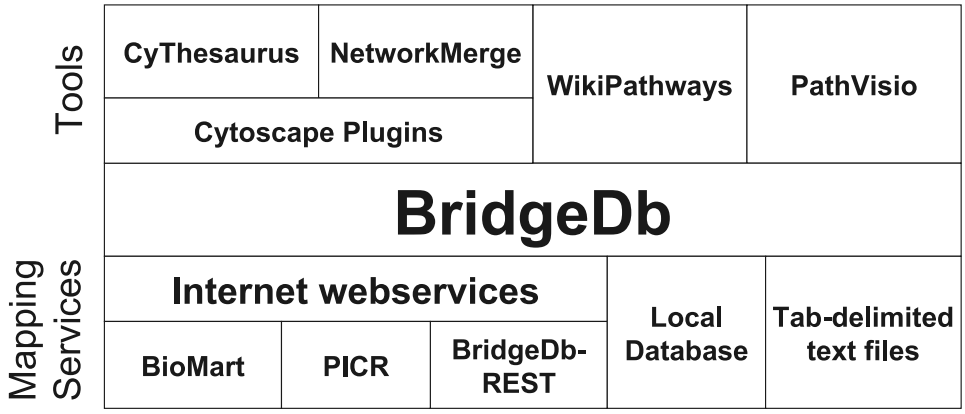


***Figure 1: The BridgeDb architecture.*** *BridgeDb provides a channel to connect multiple bioinformatics tools, such as Cytoscape or PathVisio, with online and offline identifier mapping services.*

An **internal identifier** is an identifier that is valid within the namespace of a data source. A good identifier must be unique, stable, and preferably free of semantic information. An internal identifier is not necessarily globally unique because a given identifier may be valid for two different data sources (this is especially the case for data sources that use simple integers as identifiers, such as Entrez Gene and PubChem).

In contrast with identifiers, we use the term **symbol** for a string that is a human readable representation of a biological entity. Symbols are not necessarily guaranteed to be unique or stable over time. A symbol can be a gene name such as INSR or TP53 for example. A biological entity might have several synonymous symbols or aliases. Since these symbols have a biological meaning, they are not semantic-free and do not serve as good identifiers. Semantics should be avoided because the information can be found to be "wrong" by new evidence. For example, the *Caenorhabditis elegans* gene symbol rad-5 (implying radiation sensitivity) was replaced by clk-2 (implying a function in developmental timing) after additional experimental evidence was collected and reinterpreted [11].

A **global identifier** is a globally unique identifier based on the combination of data source and internal identifier. Importantly, for the exchange of data, global identifiers must be standardized. For the representation of global identifiers, BridgeDb relies on the MIRIAM URI standard [12]. MIRIAM describes a minimal set of information to define a biological model and requires that biological identifiers be sufficiently descriptive. A valid MIRIAM URI contains both the data source and internal identifier (e.g., urn:miriam:uniprot:P62158).

We use the term **mapping service** to describe a resource for mapping information among identifiers from two or more data sources. This broad definition could include simple tables in a text file as well as complete relational databases or online web services such as Ensembl BioMart and PICR.

***Transitivity***

BridgeDb allows stacking, or combining, of different mapping sources both in transitive and non-transitive modes. Transitivity means the inference of second- or even third-degree mappings from direct mappings. Transitivity becomes an important issue especially when combining mapping services in a stack. In general it is preferable to make use of global resources such as Ensembl BioMart for annotation of microarray data, but this is not possible for custom microarrays that are not distributed widely. In those cases, custom annotation files must be used. A custom annotation file can be produced to map custom identifiers to all other biological identifiers that one might wish to consider. But BridgeDb allows an alternative approach, where one only needs a very simple mapping of custom identifiers to one gene identifier, and can rely on Ensembl BioMart to provide the rest. For example, if the custom annotation file defines relations between custom identifiers and Entrez Gene, and Ensembl BioMart provides mapping between Ensembl and Entrez Gene, then mappings between the custom identifier and Ensembl can be inferred. In this way, the task of creating custom annotations is much simpler, while at the same time enabling broader coverage of data sources. This process is depicted in Figure 2. In the same way two specialized mapping services could be chained together. For example, PICR is specialized in proteins and is not normally capable of mapping gene identifiers. However, PICR can be combined with Ensembl BioMart, using the latter to translate gene identifiers into protein identifiers that the former can understand. In principle, transitive mapping from genes to transcripts to proteins could be achieved.

## IMPLEMENTATION

The implementation of BridgeDb is described here for each of the three layers of Figure 1: tools, interface and mapping services, starting with the interface.

***Implementation of the interface layer***

The Application Programming Interface (API) of BridgeDb takes two different forms. The first form is based on Java and is suitable for Java applications. The other is based on Representational State Transfer (REST) and is suitable for all other programming languages.

***BridgeDb Java API*** The structure of the Java API is represented in Figure 3. The central BridgeDb class keeps track of all available IDMapper implementations and provides uniform access to them through the static connect method. This method takes a con-

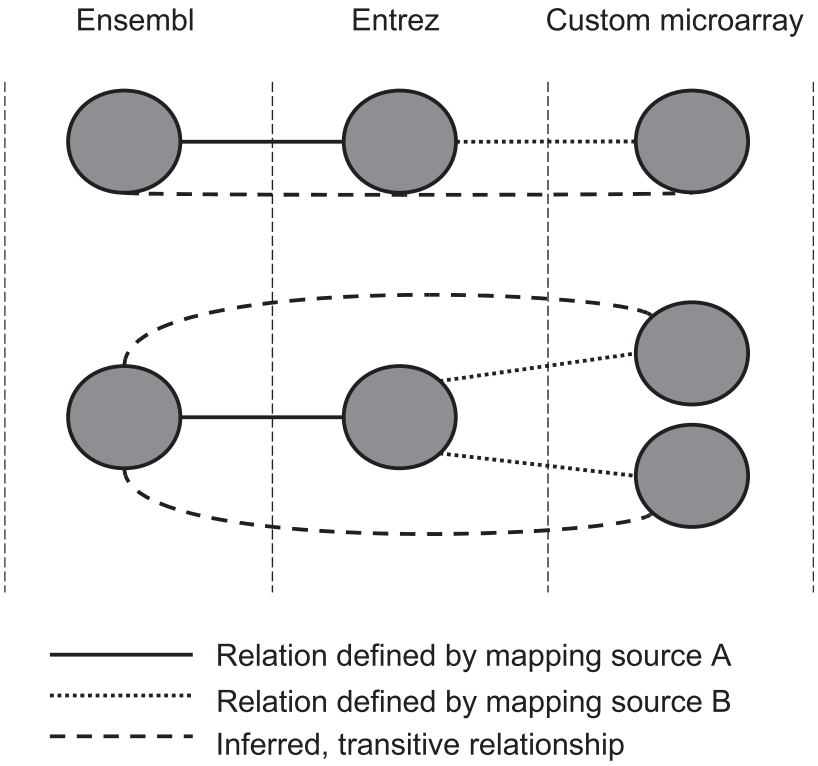


***Figure 2: Transitive relationships between identifiers.*** *In this diagram, two mapping sources (named A and B) are combined. Mapping source A defines relations between Ensembl and Entrez genes. Mapping source B is a custom microarray annotation that provides mappings between custom identifiers and Entrez. Through transitivity, the relation between the custom microarray and Ensembl can be inferred. The custom annotation file needs to define only a subset of relations.*

nection string, which contains the requested mapping service and all parameters for configuring that service. To access a completely different mapping service, only the connection string has to be modified, the rest of the program can stay the same. The connect method returns an implementation of the IDMapper interface. This interface provides methods for mapping identifiers, as well as free text search and capability introspection. The mapID method of IDMapper performs the actual mapping job. It takes an Xref object as argument. Xref is a combination of a data source and a local identifier; these two are combined to form a global identifier. The Xref class also provides methods for generating valid web links to pages describing a biological entity, and for generating a valid MIRIAM URI. Data sources are themselves represented by DataSource objects, which also hold extra information such as web links to the main page of a data source. We have created a simple example in Java that shows how to map identifiers

through webservices using BridgeDb (Additional file 1)*. To illustrate the usefulness of the standardized interface, we recreated the same functionality in a program without using BridgeDb (Additional file 2). In the latter case specialized code had to be written for each webservice, which makes the code more complex, and less flexible. For a full description of the API and more examples, see the developer documentation on the BridgeDb website [13].

***BridgeREST API*** In addition to the Java API, we provide a REST-based interface. The required software can run in the background and can be embedded in non-Java applications. Each function is accessed by a URL that specifies the address of the service and query parameters. For example, http://localhost/Human/search/ENSG00000122375 includes the following query parameters: 'search' specifies the type of query, 'Human' is the organism database being searched, and 'ENSG00000122375' is the query input. In this example, a list of identifiers and data sources corresponding to the query would be returned in the form of tab-delimited plain-text. Developer documentation for BridgeREST is also available on the BridgeDb website [14].

***Implementation of the mapping services layer***

The mapping services supported by BridgeDb can be broadly categorized in three groups: flat files, relational databases, and online web services. Flat text files are great for custom annotations. Local relational databases provide the fastest access. They are stable and make long-running analyses repeatable. Web services are not as fast as local databases but potentially provide the broadest range of data and if well maintained, are more up-to-date.

In addition to supporting a number of third-party mapping services, BridgeDb also provides its own mapping services, in the form of BridgeDerby databases for efficient mapping of genes, proteins and metabolites, and in the form of a webservice named BridgeWebservice. Thus, BridgeDb can be used as a complete identifier mapping solution out-of-the-box for common bioinformatics applications. See Table 1 for a list of supported mapping services and by whom they are provided.

* Additional files are available at the BMC Bioinformatics site at http://www.biomedcentral.com/1471-2105/11/5

*Table 1: Mapping services currently supported by BridgeDb*

| Description | Category | Provided by |
|---|---|---|
| Tab delimited text files | Flat file | Anybody |
| BridgeDerby for genes and proteins | Relational database | BridgeDb (Gladstone Institute) |
| BridgeDerby for metabolites | Relational database | BridgeDb (Gladstone Institute) |
| EnsMart | Web service | European Bioinformatics Institute |
| PICR | Web service | European Bioinformatics Institute |
| BridgeWebservice | Web service | BridgeDb (Gladstone Institute) |
| Synergizer | Web service | Harvard Medical School |
| CRONOS | Web service | Helmholtz Zentrum |

The BridgeDerby mapping service is based on the Derby [15] relational database management system. Its advantage is that it can be used to create local databases that consist of just a single file. These files can be downloaded, copied and installed easily. This system, along with the database schema, has been described before [3]. The schema can be briefly summarized as follows. Each database consists of three tables. The DataNode table contains a list of local identifiers plus a short two-letter abbreviation for the data source. The Link table contains a list of relationships between identifiers. Each link has a left part and a right part; each part refers to a unique data source and identifier in the DataNode table. Finally, the Attribute table contains symbols and other attributes for DataNode entries.

We create database files per species for genes and proteins. We also created a database file for metabolites, which is species-independent. Using the stacking mechanism, the metabolite database and a species-specific gene and protein database can easily be combined to form a complete database for biological entities of a given species. The procedures for creating these two types of database files are described below in more detail.

***Ensembl-based gene and protein database files.*** The database files for genes and proteins are based on Ensembl [16]. Relevant MySQL tables from Ensembl are locally installed and accessed via Ensembl's Perl API. Identifiers, annotations and cross-references are extracted, transformed and ultimately loaded into Apache Derby databases. The Derby databases are rebuilt and released after each Ensembl database release. Each Derby database contains information for a single species. Currently, databases are only produced for species of interest, including 36 different animal, plant, fungal, and bacterial species. However, Derby databases can readily be generated for any species supported by Ensembl or that has been effectively "ensemblized" (i.e., loaded into a database schema that is compatible with Ensembl's Perl API). Depending on the species, a selected subset of data sources are extracted from Ensembl and maintained in the Derby databases. Typically, these include Ensembl, Entrez Gene, UniProt, UniGene, RefSeq, miRBase, RFAM, PDB, TIGR, UCSC Genome Browser, and WikiGene, as well as a

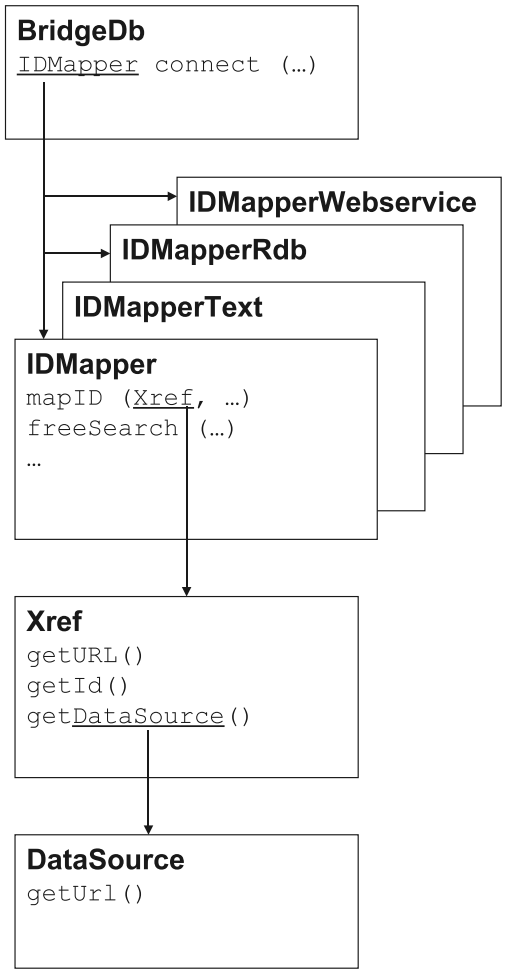


***Figure 3: Simplified UML Diagram of the BridgeDb java library.*** *The BridgeDb.connect method serves as an entry point, instantiating one of the many IDMapper implementations. The most important method of IDMapper is mapID, which takes an Xref object as argument. An Xref is a combination of an identifier String and a DataSource object, the latter representing an online biological database.*

representative model organism database, such as MGI, WormBase, ZFIN, EcoGene or TAIR, and annotations from Gene Ontology, OMIM, BioGrid, Affymetrix, Agilent, and Illumina.

The same information as contained in the Derby databases is also stored in MySQL databases, using the same schema. This form of the data is maintained for web service accessibility. Practical limitations on the size of the distributed Derby databases are not an issue for MySQL. Thus, additional identifier systems, annotations and cross-references can be stored in the MySQL databases. Furthermore, with greater capacity, this system will be able to support additional types of information such as exon, transcript, and protein domain alignments, polymorphisms, and homology, which are also available from Ensembl.

***HMDB-based metabolite database files.*** Currently, the database files for metabolites are based on HMDB Metabocards [17], because they provide free and easy-to-parse access to mapping information. Each Metabocard contains cross-references to CAS, ChEBI, PubChem and Kegg. Each metabocard is parsed by a script and identifiers are added to the DataNode table. Mappings between them are added to the Link table. This database assumes symmetry and transitivity (i.e., all identifiers on a Metabocard map to all other identifiers on the same card). The official name and all synonyms are stored in the Attribute table as symbols.

***Implementation of the tool layer***

Broadly speaking, BridgeDb can be useful whenever an application needs to make use of multiple mapping services or wants to enable the user to choose among different services. To demonstrate this, we have integrated BridgeDb into a number of bioinformatics tools. These implementations are described below illustrating four different occurrences of the identifier mapping problem.

***Use Case 1: Annotating biological pathways.*** *WikiPathways* is a collection of biological pathways open to community curation [18]. Pathways are networks of genes, proteins and metabolites that serve as a model for the actual biology of a cell. Pathway components must be properly annotated to maintain the consistency and integrity of the model for data mapping, updating and exchange. The common names of genes, proteins and metabolites are usually not suitable, because they can be ambiguous and can change over time. Identifiers from various data sources can provide exact and unambiguous references. BridgeDb is used by WikiPathways to provide integrated access to the most relevant identifiers by stacking species-specific databases for genes and proteins with a generic database for metabolites. The free search mechanism of BridgeDb helps curators find the correct identifiers for a broad range of biological entities. And BridgeDb also provides link-out URLs to all cross-references to help confirm the validity of an annotation and access more information from primary data sources.

***Use Case 2: Merging biological networks.*** The *Network Merge* plugin for Cytoscape [1] can align, compare and merge networks. A common scenario for Network Merge includes two networks that represent overlapping biological components (e.g., protein-protein interaction networks from two different yeast two-hybrid experiments). There are many valid ways to annotate such networks, but to align, compare or merge them they must be annotated with identifiers from a single data source. To solve this problem, the Cytoscape Network Merge plugin utilized BridgeDb to unify identifiers of the biological entities when merging networks, so that overlaps among networks can be recognized.

***Use Case 3: Mapping experimental data onto biological pathways.*** *PathVisio* [3] is a pathway visualization and analysis tool which has recently gained the capability to im-

port genomics data and link it to pathways. To ease the process of importing biological datasets, BridgeDb is used to map microarray reporters to the corresponding genes and proteins. The standard Ensembl-derived BridgeDerby databases contain information about a number of common chip designs. Each identifier in the experimental data is mapped to an identifier in the pathway if possible, and if a match is found, that part of the pathway can be colored depending on the measured gene expression.

***Use Case 4: Identifier translation.*** The *CyThesaurus* plugin for Cytoscape can perform large-scale identifier translation on biological entities in Cytoscape networks using BridgeDb. The plugin can be used for different purposes. For example, when multiple identifier sources are used in the networks of interest, this plugin can be used to translate different types of identifiers to a common identifier type so that identities of the biological entities in the networks will be unified. Alternatively, to export Cytoscape networks for use in other tools that require different identity types, one can utilize CyThesaurus plugin to translate the identifiers into identifier types that other tools can understand.

## Results and Discussion

Once BridgeDb has been incorporated in a bioinformatics tool, it will be possible to choose a suitable identifier mapping solution for the job at hand. There are a few considerations when choosing the right service. First, when should two identifiers map to one another? What is the basis for a mapping? An identifier is a reference to a database record about a biological entity. If two databases describe the exact same biological entity, then certainly the identifiers should map to each other. But most applications require a broader definition of identifier mapping. When aligning microarray data with a pathway model based on genes, for example, there is a need for mapping microarray reporters to genes and gene products even though these are different biological entities.

The facile switching of mapping services makes it easy to compare them programatically. In a simple test we found that most web services integrated in BridgeDb are in agreement to a high degree. Four different webservices were able to map successfully more than 80% of a set of 1000 random Affymetrix probeset identifiers. For 72% of the total set, the result was identical for each webservice. For a given set of 100 Ensembl gene identifiers, each webservice was able to map over 91%, and the identical subset was 86% (See additional file 3)[†]. The source code of BridgeDb includes the script used for this comparison (See additional file 4). In short, webservices differ in coverage of species and identifier types, but when two webservices are both able to translate the same set of identifiers, they agree to a large degree.

---

† Additional files are available at the BMC Bioinformatics site at http://www.biomedcentral.com/1471-2105/11/5

A second issue that affects the choice of mapping service is a preference for local or remote access. The first option is more efficient in the case of high-intensity usage. Another advantage of a local system is that a source of data can be frozen to make analysis reproducible. A local system will not change in the middle of a long-running analysis procedure. Alternatively, the web service approach can be more up-to-date, is centrally managed and requires fewer resources (disk space) from the end-user.

Theoretically, it would be good to do away with the multiplicity of biological databases and designate a single universal identifier for each biological entity. However, this would not eliminate the need for identifier mapping, as there would still be a need to relate different biological entities such as transcripts and reporters. That fact, combined with the current situation of multiple gene databases, means that we have to deal with the identifier mapping problem in the best possible way. We believe that standardizing on a framework used by several programs in unison provides a robust and extensible solution.

Importantly, identifier mapping should be performed as late as possible in a given workflow or annotation scheme. Gene databases change over time as more information becomes available. Experimental datasets, on the other hand, are fixed once the experiment is done. The experimental data should be annotated as closely as possible to the experimental conditions. Consider for example microarray data. Perez-Iratxeta et al.[19] have found changes to 5% of reporter annotations over a two-year timespan. The mapping of reporters to genes is not static, and the data should be linked directly to the reporter sequence that was measured rather than to a gene or genomic location. Replacing probe identifiers with gene identifiers in a microarray dataset would limit future analyses to potentially outdated associations. The importance of BridgeDb in this context is that experimental data can be annotated and linked in a consistent manner over time, ensuring that the integrity of the data is maintained while the analysis utilizes the most up-to-date information about genes and gene associations.

Several mapping services supported by BridgeDb are provided by us (Table 1). We consider these good default choices that are applicable to a broad range of situations. The local databases for genes and proteins provided by BridgeDb are based on Ensembl. We find that Ensembl defines reliable homology-based mappings that are frequently updated and available for a large number of species and, thus, provides a reasonable default choice. However, some authors [5] have noted that high thresholds of sequence similarity lead to a failure to detect all correct mappings, so extra information derived from gene nomenclature should be used. The BridgeDb interface does not dictate what constitutes a correct mapping; this is determined by the underlying mapping services. The flexible architecture of BridgeDb makes it possible to switch to a service that employs a different basis for identifier mapping if desired.

The BridgeDb concept derived from our experience as developers of different bioinformatics tools. The Application Programming Interface (API) was designed to be used by multiple tools and has proven its usability for a combination of applications with different uses.

We believe that it is better to build identifier mapping into a tool rather than requiring users to perform identifier mapping manually or with separate tools. Burdening the researcher to perform identifier mapping ignores the problem and dramatically limits the usability of the tool. Relying on external solutions also introduces unknown factors into the workflow that can lead to unreliable analysis results. By integrating a mapping service directly in the bioinformatics software tools, error-prone data preparation steps are avoided.

Different software packages solve the identifier mapping problem in different ways. We propose to modularize the identifier mapping problem into a single library. This has several advantages. Using a shared library means that developers can pool efforts, rather than investing considerable effort into maintaining isolated solutions. Tools that currently do not implement identifier mapping could do so with very little effort by just adding a module to the project.

## Conclusions

BridgeDb frees bioinformatics tools from compromising on solutions to the identifier mapping problem. By providing a standardized layer through which different mapping services can be used, BridgeDb makes it easy for tool developers to support and switch between multiple services. BridgeDb can also be used to combine or merge mapping services to support multi-omics experiments or integrate custom resources.

## Availability & Requirements

**Project name:** BridgeDb.

**Project home page:** http://www.bridgedb.org.

**Operating systems:** platform independent.

**Programming language:** implemented in Java, compatible with any programming language.

**Other requirements:** Java Runtime Environment version 1.5 or higher.

**License:** Apache 2.0 License. Source can be found at http://svn.bigcat.unimaas.nl/bridgedb. A snapshot of the code is available as additional file 4.

**Any restrictions to use by non-academics:** none.

## Authors' Contributions

MI drafted the paper. TK developed the first derby databases. AP developed the Ensembl ETL process for MySQL and derby database production. MI and JG designed the BridgeDb API. IH improved the Ensembl ETL process and developed the REST web service. JG developed both Cytoscape plugins. AP and KH contributed to the original GenMAPP identifier mapping strategy. CE and BC provided valuable feedback and support. All authors have read and approved the final manuscript.

## Acknowledgements

We acknowledge David States for useful feeback, and the Gramene project for providing plant identifier mappings for BridgeDerby.

This work was supported by the Google Summer of Code program, transnational University Limburg (tUL), the BioRange program of the Netherlands Bioinformatics Consortium (NBIC), the Netherlands Consortium for Systems Biology (NCSB), the National Institutes of Health [GM080223, HG003053], the European Nutrigenomics Organization (NuGO) and the Dutch Scientific Organisation (NWO).

## References

1. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.
2. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR: **GenMAPP 2: new features and resources for pathway analysis**. *BMC Bioinformatics* 2007, **8**:217.
3. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C: **Presenting and exploring biological pathways with PathVisio**. *BMC Bioinformatics* 2008, **9**:399.
4. Huang da W, Sherman BT, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID gene ID conversion tool**. *Bioinformation* 2008, **2**(10):428-430.
5. Waegele B, Dunger-Kaltenbach I, Fobo G, Montrone C, Mewes HW, Ruepp A: **CRONOS: the cross-reference navigation server**. *Bioinformatics* 2009, **25**(1):141-143.
6. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay W, Weinstein JN: **MatchMiner: a tool for batch navigation among gene and gene product identifiers**. *Genome Biol* 2003, **4**(4):R27.
7. Iragne F, Barre A, Goffard N, De Daruvar A: **AliasServer: a web server to handle multiple aliases used to refer to proteins**. *Bioinformatics* 2004, **20**(14):2331-2332.

8. Cote RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R, Hermjakob H: **The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases**. *BMC Bioinformatics* 2007, **8**:401.
9. Berriz GF, Roth FP: **The Synergizer service for translating gene, protein and other biological identifiers**. *Bioinformatics* 2008, **24**(19):2272-2273.
10. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data**. *Genome Res* 2004, **14**(1):160-169.
11. Ahmed S, Alpi A, Hengartner MO, Gartner A: **C. elegans RAD-5/CLK-2 defines a new DNA damage checkpoint protein**. *Curr Biol* 2001, **11**(24):1934-1944.
12. Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P *et al*: **Minimum information requested in the annotation of biochemical models (MIRIAM)**. *Nat Biotechnol* 2005, **23**(12):1509-1515.
13. **BridgeDb Java API documentation [http://www.bridgedb.org/apidoc]**. In.
14. **BridgeREST API Documentation [http://www.bridgedb.org/wiki/RestAccess]**. In.
15. **Apache Derby [http://db.apache.org/derby]**
16. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L *et al*: **Ensembl 2009**. *Nucleic Acids Res* 2009, **37**(Database issue):D690-697.
17. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S *et al*: **HMDB: the Human Metabolome Database**. *Nucleic Acids Res* 2007, **35**(Database issue):D521-526.
18. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: **WikiPathways: pathway editing for the people**. *PLoS Biol* 2008, **6**(7):e184.
19. Perez-Iratxeta C, Andrade MA: **Inconsistencies over time in 5% of NetAffx probe-to-gene annotations**. *BMC Bioinformatics* 2005, **6**:183.

Chapter 6

# Pathway visualization applied to a multi-omics study of starvation in mouse intestine

Martijn P. van Iersel[1,2,*], Milka Sokolovic[3], Kaatje Lenaerts[4], Freek Bouwman[4], Edwin C.M. Mariman[4,5], Chris T. Evelo[1,2]

[1]Department of Bioinformatics BiGCaT, Maastricht University, Maastricht, The Netherlands.
[2]Netherlands Consortium for Systems Biology (NCSB), the Netherlands.
[3]Department of Medical Biochemistry, Academic Medical Centre, University of Amsterdam, The Netherlands.
[4]Department of Human Biology, NUTRIM School for Nutrition, Toxicology, and Metabolism, Maastricht University, The Netherlands.
[5]Top Institute Food and Nutrition, Nutrigenomics Consortium, the Netherlands.

*Submitted*

## ABSTRACT

**Motivation:** We combined two earlier studies on prolonged starvation in mice, one using proteomics and one using transcriptomics technologies. We aimed to assess the feasibility and utility of pathway visualization as a means to integrate various types of omics data, and to obtain new insights in the starvation response of the gut.

**Results:** We found a low overall correlation between proteome and transcriptome data (Spearman rank correlation: 0.21). At the level of individual genes, correlation is highly variable. Several mRNA / protein pairs, such as ferritin and elongation factor 2, correlate poorly, whereas others show high correlation. At the pathway level, transcriptomics and proteomics data supported each other. To visualize the integrated datasets, we developed new plug-ins for the PathVisio program.

**Availability and implementation:** The PathVisio software, together with new plug-ins mentioned in this article, can be obtained at http://www.pathvisio.org.

## INTRODUCTION

The intestine plays an important role in the response of the body to prolonged starvation. In two previous publications, both the transcriptome [1] and the proteome [2] of the murine intestine was determined after 12, 24 and 72 hours of starvation. Here the two studies are combined and compared. The goal is twofold. First, to develop bioinformatics tools needed to make an integrated omics approach feasible. Second, to get a more comprehensive view of regulation of pathways involved in the starvation response of the gut.

### *Omics integration*

Microarray technology can be used to measure the expression of thousands of genes in a tissue at the same time. Methods for processing, analyzing and interpreting microarray data are well established, and off-the-shelf microarrays are available that have enough capacity to measure 100% of the genes in a genome.

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

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

As both proteomics and transcriptomics methods each have their own advantages, it could be beneficial to combine them. It appears that correlation at the individual gene level is much higher than correlation at the global level [3]. Therefore, we believe that integrated omics datasets must be studied on a gene-by-gene level, taking into account more detailed knowledge about the roles of those genes in biological pathways. Because the relation between genes and proteins is heterogeneous and context-dependent, we propose that integrated omics analysis must use available a-priori knowledge, for example in the form of pathway diagrams. Pathway diagrams contain biological context, such as which entities are related, what is their cellular location, which proteins are interaction partners, and what is the nature of those interactions (stimulation or repression). By visualizing the combined dataset on a pathway diagram one can interpret the biological context more easily. As part of this study we aimed to develop easy-to-use software to make this possible.

***Regulation of the intestinal response to fasting***

The present study investigates the intestinal changes that occur in response to prolonged fasting. A better understanding of fasting could lead to better understanding of weight-loss diets and better treatment of cachexia (wasting syndrome) caused by chronic disease [11].

Prolonged fasting leads to changes in gene and protein expression in different organs to sustain fuel homeostasis. A basic model for fasting response can be summarized as "sugar-fats-proteins", where the body switches to different food sources in order of preference [12]. At first, blood glucose levels are maintained mainly by glycogenolysis in the liver. When the hepatic glycogen store nears depletion, the body switches to gluconeogenesis as the main source of glucose. Gluconeogenesis is briefly fueled by muscle proteolysis, but then fatty acid oxidation and ketone body synthesis take over. Finally, the terminal phase arrives when only proteolysis can sustain glucose production. If fasting continues, this will eventually lead to death, usually by respiratory failure. Humans can survive up to 6-8 weeks without food [13]; mice can survive for 4 days.

It is clear that the various organs (liver, intestine, adipose tissue, muscle) play different roles in the fasting response, but the whole picture is far from completely elucidated.

The fasting state has a strong effect on the intestine. Transcriptomics studies have shown that two biological processes are mainly affected: cell turnover and energy metabolism. Enterocytes turn over rapidly during the digestive state, which represents a major energy expenditure. Cell turnover is progressively down-regulated during fasting, as can be seen from expression of genes in the cell cycle pathway. Cyclins and cyclin-dependent kinases are decreased, inhibitors of cyclins are increased [1]. This is accompanied with a decrease in the apoptosis pathway, the effect of which is also visible as a maturation of enterocytes [14].

Regarding energy metabolism, it has been established that cells of the intestine preferentially use glutamine and ketone bodies as energy source. During the early response, which peaks at 12h in mice, fat is catabolized to ketone bodies, and glutamate is catabolized to pyruvate. Glutamine, which the intestine can not produce sufficiently by itself, is not degraded but conserved. Pyruvate oxidation is inhibited, but is converted to lactate instead, to be excreted to the blood and used as a fuel for gluconeogenesis in the liver.

As fasting continues, a phase transition occurs, which is most clearly visible in the fatty acid oxidation pathway and electron transport chain. Gene expression in both pathways is reduced. There are indications that glucose export to the blood is increased at the expense of lactate production [1].

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

## METHODS

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

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

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

**2DE** The procedure for generating 2D protein gel images was as described before [15]. Protein samples were isolated from equal quantities of proximal and distal parts of the intestine, pooled per mouse. 1 gel was made for each mouse. 100 µg of total protein was separated by isoelectric focusing using IPG strips, and then placed onto 12.5%

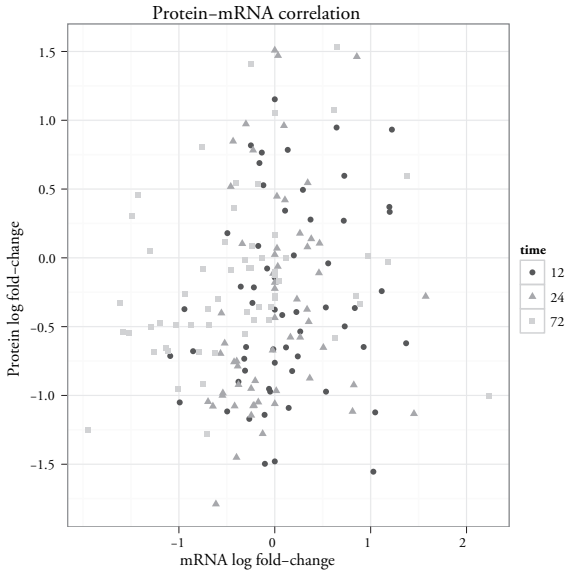


***Figure 1 Protein-mRNA correlation.*** *In this plot, the fold-change of protein and mRNA are plotted against each other. Fold-changes are calculated for each time point against t=0. In cases where multiple protein spots correspond to the same gene, the average was used. The overall correlation plot shows that there is very little agreement between protein and gene expression.*

SDS-polyacrylamide gels for protein separation in the second dimension. Gels stained with SYPRO Ruby Protein stain were scanned with the Molecular Imager FX (Bio-Rad Laboratories).

Analysis of differentially expressed proteins was performed using PDQuest 7.3 (Bio-Rad Laboratories). A number of spots were selected for identification, with a preference for spots with a significant intensity difference. Selected spots were excised and subjected to tryptic in-gel digestion and MALDI-TOF MS (Waters, Manchester, UK), generating peptide mass fingerprints which were subsequently identified.

**Microarray annotation** The microarray type used was the 60-mer Mouse Development 22k Oligo Microarray G4120A (Agilent). The microarray contains 20280 probes of 60 nucleotides each. The annotation file provided by Agilent (Version of Dec. 16 2009) associates 9616 probes to Ensembl gene identifiers.

The probes were designed based on a 5-year old genome build, which could have diverged in the intervening years. On the assumption that 60-mer probes do not require a complete sequence match to hybridize to transcripts, we investigated if a blast search with less stringent settings would increase probe annotation coverage. We selected the best blast hits against the current Ensembl (release 57) Mouse cDNA database, with a minimum e value of 1.0e-6. The blast resulted in annotation for 10696 probes, an increase of 11%. We opted to employ the blast results instead of Agilent annotations for all further analysis. Identifiers in both data sets were mapped to pathways using the BridgeDb framework [16]. None of the standard identifier mapping resources included

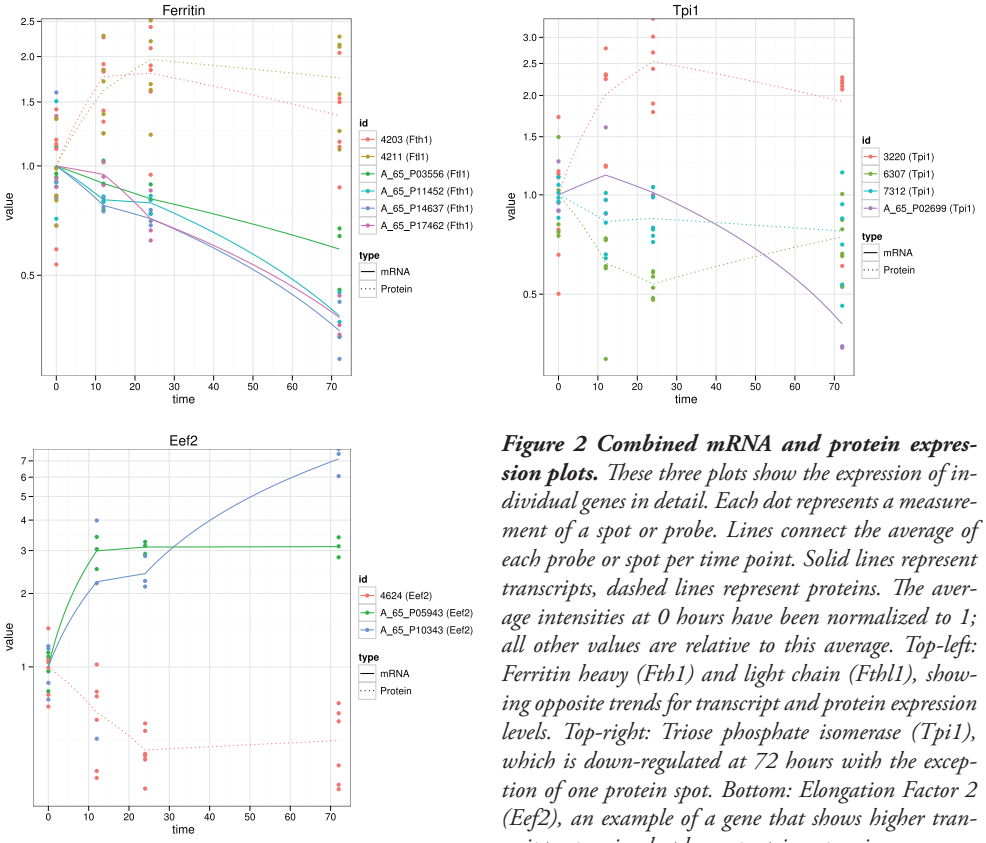


***Figure 2 Combined mRNA and protein expression plots.*** *These three plots show the expression of individual genes in detail. Each dot represents a measurement of a spot or probe. Lines connect the average of each probe or spot per time point. Solid lines represent transcripts, dashed lines represent proteins. The average intensities at 0 hours have been normalized to 1; all other values are relative to this average. Top-left: Ferritin heavy (Fth1) and light chain (Fthl1), showing opposite trends for transcript and protein expression levels. Top-right: Triose phosphate isomerase (Tpi1), which is down-regulated at 72 hours with the exception of one protein spot. Bottom: Elongation Factor 2 (Eef2), an example of a gene that shows higher transcript expression but lower protein expression.*

in BridgeDb contained mappings for this particular array type, necessitating the use of a custom mapping table. The blast results were formatted so that they could be understood by BridgeDb.

**Analysis and Pathway visualization** Correlation and expression plots were created with R/BioConductor. Pathway visualization was performed using the PathVisio program [17].

The mouse pathway set, obtained from WikiPathways [18] March 2010, was used for the analysis. This pathway set covers 3975 unique genes, or 14% out of 28063 mouse genes and pseudo-genes known in Ensembl release 57. The pathway set covers 66% (51 out of 77) of unique measured proteins and 24% (2083 out of 8648) of unique measured transcripts.

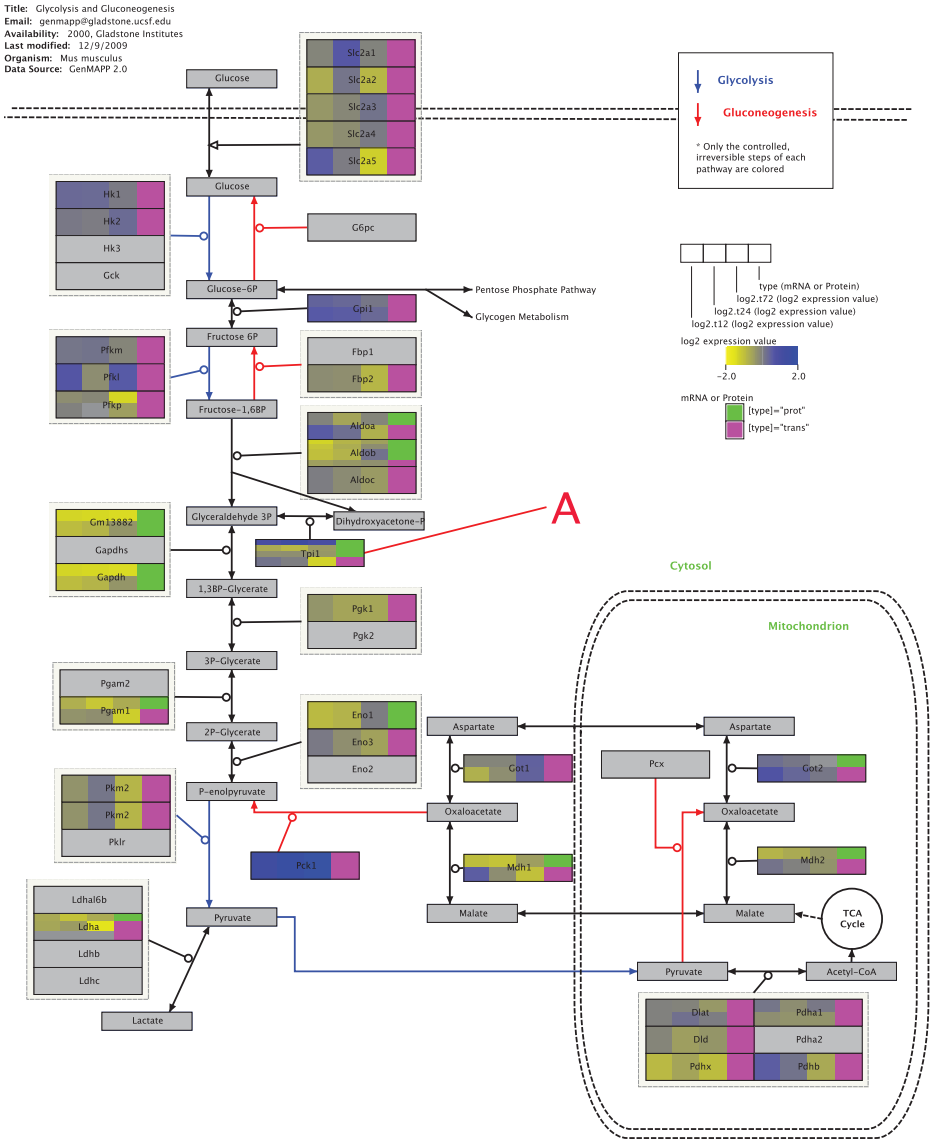


***Figure 3: Glycolysis and Gluconeogenesis pathway in PathVisio.*** *Each colored box represents a gene product. Blue indicates increased expression levels, yellow decreased. Each box is a heat map with rows representing probes or spots, and columns representing time points. The rightmost column is a flag that indicates if the given row is a protein spot (green) or microarray probe (pink). Multiple probes and/or spots can be shown in a box. Tpi1 is marked with the letter A.*

### RESULTS

The proteomics dataset contained 130 identified spots, corresponding to 77 unique protein identifiers. For only 59 out of those 77, both the mRNA and protein abundance was determined. For each of the 59 proteins and genes, the log2 ratio was calculated for 12, 24 and 72 hours. When taken together, there is very little overall correlation between the two types of data (See Figure 1). The spearman rank correlation coefficient is 0.21, which indicates a slight positive overall correlation between mRNA and protein levels.

The picture is very different for various proteins. There are highly correlating, differentially expressed transcript-protein pairs such as Vim (Vimentin), Aldob (Fructose Biphosphate Aldolase B) and Atp5h (ATP Synthase D chain).

On the other hand, some proteins show an opposite correlation, where the transcript is regulated in the opposite direction as the proteins. In most of these cases, the transcript is up-regulated but the corresponding protein is down-regulated, as is the case for Eef2 (elongation factor 2, See Figure 2), Aldh1b1, Arhgdia, Hnrnpa2b1 and Uqcrc1. For Eef2 a similar divergence between mRNA and protein levels has been reported independently in liver and muscle tissue [19].

Some proteins that are identified in multiple spots show varying expression in different isoforms of the protein. This indicates that post-translational modifications could play a role to regulate protein activity. The proteins that are potentially under post-translational regulation include metabolic enzymes (Tpi1 and Atp5h), proteins related to protein folding (Calr, Hspa8 and Hspa5) and cytoskeletal proteins (Krt19, Actb, Actg2 and Vil1).

## Software Implementation

PathVisio has a plug-in framework to allow customization to new types of pathway analysis and visualization. To visualize protein and gene expression data side-by-side in pathway context, we developed two new plug-ins for PathVisio: the Gex plug-in, which manages data visualization, and the BridgeDbConfig plug-in, which enables configuration of identifier mapping resources. The Gex plug-in is now a core plug-in that is a part of the main distribution since the release of PathVisio 2.0, the BridgeDbConfig plug-in is installed separately.

High-throughput datasets in tab-delimited text format are imported using the expression data import wizard which is part of the Gex plug-in. Data should already be normalized and preferably log-transformed before data import. During data import, the user can select which column contains identifiers for genes or proteins. A prerequisite for integration of multiple omics studies is the ability to map identifiers from various

sources [20]. In this case, each data point in the microarray dataset is identified with an Agilent probe identifier (such as A_65_P03556), and proteins were associated with a Uniprot identifier (such as P09528). PathVisio allows direct import of mixed identifiers without the need to pre-process the data.

The BridgeDb framework [16], which is incorporated into PathVisio, was used to map probe and protein identifiers to gene identifiers and integrate these two datasets.

Three proteins were annotated with erroneous protein identifiers by the identification software (in one case a GenBank accession number, in one case an identifier for the human homologue, and in one case a deprecated Uniprot identifier). Rather than fixing the original data, BridgeDb allowed us to create a separate "manual override" mapping table and combine it with the rest of the mapping resources. The advantage of doing this, rather than simply fixing up the original dataset, is that the modifications remain separate and can be more easily re-examined or reverted in the future.

Thus three sources of mapping information were used; first the results of BLAST of microarray sequences for mapping Agilent probes to genes, second a regular BridgeDerby database for mapping proteins to genes, and third a manual override table to fix errors. Using BridgeDb, the three resources were unified into a single mapping resource.

After the data import step, the user configures how the data is represented by colors. This can be done either using Boolean expressions, or by mapping numerical values to a color gradient. These colors are then displayed in the gene boxes. Multiple conditions can be displayed side-by-side.

One feature that is of particular interest for omics integration is the fact that if multiple rows of data map to the same gene box, they can be averaged and mapped to a single color, or the box can be divided horizontally and each row in the dataset gets it own row in the box. The problem of a variable number of data points per gene box occurs sometimes with microarrays that have more than one probe per gene, but it is especially important for omics integration where one is often dealing with two datasets of very unequal size (in this case 130 proteins versus 10696 transcripts), which automatically means that some boxes have more data than others. The box is divided horizontally for each corresponding row in the dataset. Thus, the box is used as a small heat map representation of a subset of the data, where each column represents a condition and each row represents a probe.

Although this subdivision means that the boxes can get cramped, we find this approach is superior to summarizing data, because that would mean throwing away data. See for example Tpi1 in Figure 3, where two rows are clearly differentially expressed. Boxes can be enlarged manually if necessary. Data can be visualized in parallel for direct comparison, but it is also possible to create separate visualizations and toggle between them

using a drop-down list. Sometimes visualizing different time points separately leads to less confusion and makes it less likely that the response of one gene in the first time point is erroneously compared to another gene in the last time point. Separating the time points in different visualizations also helps to combat the information overload in a single box.

## Discussion

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

Although proteomics data has fewer data points than transcriptomics data, there are a few instances where important transcripts are not measured, due to absence of a suitable probe on the array. In those cases, protein measurements can fill important gaps in pathways. In the urea cycle, no mRNA expression levels were measured for otc2 (ornithine carbamoyltransferase) and arg2 (arginase-2) genes (due to absence of probes for these genes on the microarray). Nevertheless, protein levels for these genes have been measured showing a clear down-regulation, in particular in the early (12 hours) response. This is consistent with glutamine conservation which is also demonstrated by the expression of genes such as Oat, PycS, Gls and GlnS.

Similarly, the down-regulation of Acaa2 (only measured as protein) is completely consistent with the reduction of fatty acid biosynthesis, also supported by the down-regulation of genes such as Hadh1. Additionally, down-regulation of the Gapdh protein, indicating a decrease in glycolysis activity, is entirely consistent with strong expression of the Pck1 gene, a gateway for gluconeogenesis (see Figure 3).

Analysis of the transcriptome has revealed strong effects on cell cycle regulation and apoptosis. In particular cyclins are decreased in expression. Cell turnover in intestine is thought to be a major source of energy expenditure. Unfortunately, it must be noted that no cyclins or other proteins related to cell cycle regulation and apoptosis have been identified, most likely because they are not abundant enough to be detected using the 2DE technique. The comparison of the two datasets clearly reveals the bias problems occurring with proteomics techniques. Proteomics analysis alone would have missed a major regulatory effect of starvation in the gut.

Different phosphorylation states lead to different spots in a 2D gel, thus a change in such a spot could mean two things: the experimental condition has led to a change in total quantity of the corresponding protein, or it has led to a change in the state of the

protein (or both). Due to the incompleteness of proteomics data, these two cases cannot be distinguished. A change at the spot level is interesting, but a decrease cannot always be straightforwardly interpreted as a decrease in functional activity.

A clear example of differentiation at the spot level is Tpi1 (Triose phosphate isomerase), an important enzyme of the gluconeogenesis pathway, which is increased in spot 3220 but decreased in spots 6307 and 7312 throughout the fasting period (See Figure 2). From the estimated mass as well as the mass spectrum it appears that the protein in spot 3220 is missing an N-terminal fragment. Possibly alternative splicing or proteolysis plays a role in the regulation of this protein. Assuming that the partial Tpi1 protein has a reduced activity, this finding is consistent with a reduced activity in the glycolysis pathway. However, to determine the exact nature of the observed change in Tpi1, follow-up experiments will be necessary.

Another protein that is affected in an interesting way is ferritin, a protein that is necessary for the storage of iron in tissues. Ferritin is down-regulated at the gene level, but up-regulated at the protein level throughout starvation (See Figure 2). Note that this is the case for both ferritin heavy chain and ferritin light chain. Because iron from food intake is drastically reduced, the decreased mRNA expression level of ferritin is unsurprising. Harder to explain is the increased protein abundance in spite of lower transcript levels. It has been shown that mRNA turnover regulated by the iron responsive element binding protein (Ireb2, not measured in this study) has a strong effect on ferritin [21], which could explain the discrepancy between protein and mRNA abundance. Another possible explanation is that the increased values are caused by a change occurring in erythrocytes, which can never be fully excluded from the protein sample. Other causes, such as the presence of other ferritin spots in the gel that have not yet been identified, can not be ruled out.

This comparison illustrates that the 2DE technique by itself does not provide enough data for a systems biology level overview. In this study, if only proteomics data had been available, important pathways such as apoptosis and cell cycle regulation would have been missed entirely. Nevertheless, protein expression data do provide interesting insights into regulation on a gene-by-gene basis. In particular those proteins that do not correlate well are interesting, at the very least for generating hypotheses for follow-up experiments.

A number of existing applications can perform visualization of high-throughput datasets, such as KEGG Atlas [22], ProMeTra [23], Vanted [24] and Reactome SkyPainter [25], and reviewed in [26, 27]. Only some of those have demonstrated the capability to perform pathway visualization with multiple omics datasets at the same time, such as ProMeTra, which is focused on combining metabolomics and transcriptomics data.

The automated mapping of mixed identifiers is a distinguishing feature of PathVisio that makes it particularly suited for integration and simultaneous visualization of datasets from different sources.

In conclusion, proteomics data sometimes reinforces the conclusions made from transcriptomics data, and sometimes poses new questions. The interpretation is not straightforward, and the correlation of gene and protein expression levels (or lack thereof) must be interpreted on a case-by-case basis. Pathway visualization can serve as a useful aid, because pathways serve as a knowledge base for a broad collection of biological information.

## Acknowledgements

This work was supported by the Dutch Ministry of Economic Affairs through the Innovation Oriented Research Program on Genomics: IOP Genomics IGE01016, and by the transnational University of Limburg (tUL). This work was carried out within the research program of the Netherlands Consortium for Systems Biology (NCSB), which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

## References

1. Sokolovic M, Wehkamp D, Sokolovic A, Vermeulen J, Gilhuijs-Pederson LA, van Haaften RI, Nikolsky Y, Evelo CT, van Kampen AH, Hakvoort TB *et al*: **Fasting induces a biphasic adaptive metabolic response in murine small intestine**. *BMC Genomics* 2007, **8**:361.
2. Lenaerts K, Sokolovic M, Bouwman FG, Lamers WH, Mariman EC, Renes J: **Starvation induces phase-specific changes in the proteome of mouse small intestine**. *J Proteome Res* 2006, **5**(9):2113-2122.
3. Nie L, Wu G, Culley DE, Scholten JC, Zhang W: **Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications**. *Crit Rev Biotechnol* 2007, **27**(2):63-75.
4. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver**. *Electrophoresis* 1997, **18**(3-4):533-537.
5. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast**. *Mol Cell Biol* 1999, **19**(3):1720-1730.
6. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**(5518):929-934.
7. Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR, 3rd: **Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 2003, **100**(6):3107-3112.

8. Pradet-Balade B, Boulme F, Beug H, Mullner EW, Garcia-Sanz JA: **Translation control: bridging the gap between genomics and proteomics?** *Trends Biochem Sci* 2001, **26**(4):225-229.
9. Chevalier F: **Highlights on the capacities of Gel-based proteomics**. *Proteome Sci* 2010, **8**(1):23.
10. Roe MR, Griffin TJ: **Gel-free mass spectrometry-based high throughput proteomics: tools for studying biological response of proteins and proteomes**. *Proteomics* 2006, **6**(17):4678-4687.
11. Argiles JM: **Cancer-associated malnutrition**. *Eur J Oncol Nurs* 2005, **9 Suppl 2**:S39-50.
12. Cahill GF, Jr.: **Fuel metabolism in starvation**. *Annu Rev Nutr* 2006, **26**:1-22.
13. Collins S: **The limit of human adaptation to starvation**. *Nat Med* 1995, **1**(8):810-814.
14. Koga A, Kimura S: **Influence of restricted diet on the cell cycle in the crypt of mouse small intestine**. *J Nutr Sci Vitaminol (Tokyo)* 1980, **26**(1):33-38.
15. Lenaerts K, Mariman E, Bouwman F, Renes J: **Glutamine regulates the expression of proteins with a potential health-promoting effect in human intestinal Caco-2 cells**. *Proteomics* 2006, **6**(8):2454-2464.
16. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, Evelo CT: **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC Bioinformatics* 2010, **11**:5.
17. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C: **Presenting and exploring biological pathways with PathVisio**. *BMC Bioinformatics* 2008, **9**:399.
18. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: **WikiPathways: pathway editing for the people**. *PLoS Biol* 2008, **6**(7):e184.
19. Yoshizawa F, Miura Y, Tsurumaru K, Kimata Y, Yagasaki K, Funabiki R: **Elongation factor 2 in the liver and skeletal muscle of mice is decreased by starvation**. *Biosci Biotechnol Biochem* 2000, **64**(11):2482-2485.
20. Waters KM, Pounds JG, Thrall BD: **Data merging for integrated microarray and proteomic analysis**. *Brief Funct Genomic Proteomic* 2006, **5**(4):261-272.
21. Hintze KJ, Theil EC: **Cellular regulation and molecular interactions of the ferritins**. *Cell Mol Life Sci* 2006, **63**(5):591-600.
22. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W423-426.
23. Neuweger H, Persicke M, Albaum SP, Bekel T, Dondrup M, Huser AT, Winnebald J, Schneider J, Kalinowski J, Goesmann A: **Visualizing post genomics data-sets on customized pathway maps by ProMeTra-aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example**. *BMC Syst Biol* 2009, **3**:82.
24. Junker BH, Klukas C, Schreiber F: **VANTED: a system for advanced data analysis and visualization in the context of biological networks**. *BMC Bioinformatics* 2006, **7**:109.

25. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al*: **Reactome knowledgebase of human biological pathways and processes**. *Nucleic Acids Res* 2009, **37**(Database issue): D619-622.
26. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D *et al*: **Visualization of omics data for systems biology**. *Nat Methods* 2010, **7**(3 Suppl):S56-68.
27. Joyce AR, Palsson BO: **The model organism as a system: integrating 'omics' data sets**. *Nat Rev Mol Cell Biol* 2006, **7**(3):198-210.

# Chapter 7

# General Discussion

The field of pathway visualization is currently undergoing strong development. Years of research have led to detailed knowledge on pathways. Online pathway databases are increasing in size; Reactome reported 2.7-fold increase in 3 years [1], and KEGG added 50 pathways in 2 years [2]. New standards and tools are being developed to get information out of those databases, and at the same time, the wiki model is starting to be used for online collaboration to get data into those databases [3]. With the efforts around Molecular Interaction Maps (MIM) [4] and the Systems Biology Graphical Notation (SBGN), pathway notations are becoming standardized in the same way they have been in electronic circuit diagrams and Unified Modeling Language (UML) diagrams for object oriented programming [5].

The goal of this thesis, as stated in chapter 1, is to *develop methods to integrate multiple types of experimental data and visualize them on pathway diagrams*. To achieve this goal, PathVisio, WikiPathways and BridgeDb were developed. As shown in chapter 3, PathVisio can be used to create pathways manually. In chapter 4 we have seen how pathways can be shared and curated collaboratively on WikiPathways. The WikiPathways resource uses the wiki model to lower the barrier to entry and attract contributing scientists. At the same time, safeguards are in place to monitor additions and ensure a minimum quality level. Chapter 5 demonstrated how BridgeDb provides a generic framework for the identifier mapping problem. Identifier mapping is crucial to the integration of diverse datasets. Finally, chapter 6 illustrated how all these pieces fit together to enable integration of multiple high-throughput datasets and visualization on pathways.

A number of new software tools are presented in this thesis, namely PathVisio, BridgeDb and WikiPathways. But the question remains: what advantages do they provide compared to similar tools? How do these tools work together with other bioinformatics software? How will they be further developed in the future?

In this chapter, we will critically evaluate these tools and compare them to other bioinformatics software, based on these aspects: the ability to aid in data interpretation, the ability to perform data integration, the role in analysis pipelines, the possibilities for code re-use, the adherence to common standards and the relation to open access and open source philosophies. For each aspect, we will identify strong points of our software framework, as well as criticism and potential future improvements. In this discussion we will also draw lessons that could benefit the field of bioinformatics as a whole.

## Data interpretation

Simply sifting through lists of differentially expressed genes in spreadsheets is a very limiting method of data analysis. Many alternative methods have been developed for the interpretation of high-throughput datasets. These approaches fall into two categories: data-driven and knowledge-driven [6]. In data-driven interpretation, the dataset is ana-

lyzed by itself, for example using various clustering methods [7] or principal component analysis (PCA). Knowledge-driven approaches use existing biological knowledge as basis for further analysis.

Different types of knowledge can be used for knowledge-driven interpretation. This thesis focused on pathway visualization, i.e. projecting the data on manually curated pathway diagrams. PathVisio is one of several tools that can do that. Biological networks are another possible source of knowledge. There are many types of biological networks, such as protein-protein interaction networks, co-citation networks and co-expression networks. Network and pathway analysis tools are described in several review articles [8-11]. Another type of knowledge that can be used are ontologies, in particular the Gene Ontology [12]. Much of the wide variety of ontological analysis tools is reviewed in [13]. Finally information about the promoter sequences of genes may be used as knowledge source. See Table 1 for a list of software options for each type of knowledge analysis.

*Table 1: A non-exhaustive list of types of knowledge-driven analysis and the bioinformatics software that can perform them.*

| **Knowledge type** | **Software options** |
|---|---|
| Pathway diagrams | PathVisio (chapter 3)<br>GenMAPP [14]<br>Reactome SkyPainter [1]<br>CellDesigner [15]<br>KEGG atlas [16] |
| Interaction, co-citation and co-expression networks | Vanted [17]<br>Visant [18]<br>GENeVis [19]<br>Cytoscape [20] |
| Gene Ontology | GO-Elite [21]<br>DAVID [22]<br>BiNGO [23]<br>Onto-Tools [24] |
| Promoter sequences | Genomatix [25] |

Data-driven approaches work best when the experimental effects are clear-cut and easily separable in clusters. Knowledge-driven approaches are better at distinguishing subtle effects [6]. For example, if multiple genes are differentially expressed consistently throughout a pathway, then the added effect is interesting, even if the individual fold-changes are very low. The disadvantage of knowledge-driven approaches is that they depend on the quality of annotation. If annotation is lacking, effects may be missed. The choice for either data-driven or knowledge-driven may depend on the study subject. For example, in many nutrition studies only subtle changes in gene expression are found, making knowledge-driven approaches more suitable.

PathVisio and WikiPathways clearly have a focus on pathway knowledge. The pathway level of knowledge is attractive because of its intuitiveness, ease of interpretation and ability to map diverse pieces of data, including not only the functional grouping of biomolecules, but also the order of interaction, complex formation, compartmentalization and localization, links to literature and online bioinformatics resources. All of these diverse pieces of information can be brought together in a single picture. Networks are harder to visualize, because of the difficulty of automatically generating a good layout (chapter 2). This makes pathway visualization programs easier to use and more user-friendly, which is a strong point of PathVisio.

The choice for pathway visualization is not exclusive. The interpretation approaches mentioned above are all complimentary, and many can be applied to the same study in addition to pathway analysis. Each method could give rise to distinct results. Re-interpretation of a published dataset using a different approach can give rise to new insights [6]. The study presented in chapter 6 is an example of this, as it is based on two datasets that were published previously.

A shortcoming of PathVisio is that it can make use of only one type of knowledge. GenMAPP and Cytoscape (through the BiNGO plug-in) can perform ontological analysis in addition to their core functionality. In the future, plug-ins could be developed for PathVisio for ontological analysis, to make it a more versatile tool for knowledge-driven interpretation.

## Data integration

There are several published studies that integrate multiple omics datasets, reviewed in [26, 27]. In addition, numerous studies using pathway visualization have been published. Studies that combine both multi-omics datasets with pathway visualization are currently not very common. Undoubtedly more will appear in the future as both the required software tools and the experimental technologies are maturing rapidly. A recent publication combines pathways with both transcriptomics and metabolomics data, using the online ProMeTra tool [28]. Although not demonstrated in this thesis, PathVisio is capable of mapping and displaying metabolomics data as well, and applying this to a real dataset will make an interesting topic for a future study. PathVisio could still be improved specifically for metabolomics data visualization. Metabolomics data is often obtained from excreted body fluids (for example blood or urine), where metabolite concentrations differ from concentrations inside the cell. Visualizing extracellular measurements on pathways as though they were measured inside the cell is a false presentation. To map metabolomics data correctly, PathVisio will have to be adapted to distinguish cellular and extra-cellular metabolites.

Merging multiple datasets often leads to a problem of mapping mixed identifiers [29]. An important advantage of PathVisio is the flexible identifier mapping enabled through BridgeDb. This makes PathVisio well suited to visualize multiple datasets directly side by side. The flexibility of BridgeDb to switch between identifier mapping services means that it will be easy to adapt PathVisio to new experimental methods in the future (chapter 5).

## Analysis work flows

How is bioinformatics software used to analyze high-throughput experiments? Using microarray data as an example, we can identify a number of steps that are fairly standard [6]. The first step is to extract raw data from the experiment. In a second step, quality control is performed, filtering out low quality data and possibly repeating parts of the experiment that were substandard. Third, data is normalized to filter out technical sources of variation, leaving only biological variation in the data. Fourth, statistical analysis is used to find meaningful differences between the conditions compared, taking into account the multiple testing problem. In the fifth step we take the complete dataset and visualize or transform it in such a way that biological interpretation can take place.

For each step in this work flow, different software tools can be used. The first and second steps are very dependent on the technology in question. In the case of microarrays, raw data is extracted by fluorescent scanning of a glass slide. Vendor software will localize spots and estimate raw intensities from the scanned images. There are a number of options for data normalization, depending on the microarray platform used, with quantile normalization commonly used for two-color arrays and Robust Multi-chip Average (RMA) for one-color arrays. A number of statistical tests, such as Student's t-test or Analysis of Variance (ANOVA), can be used in step four, followed by a multiple testing correction methods such as False Discovery Rate (FDR).

*Table 2: Examples of software used in analysis workflows*

| **Workflow step** | **Examples of software** |
|---|---|
| Raw data extraction | Vendor Software (Affymetrix GCOS, Agilent feature extraction software) |
| Quality Control | Vendor software, R/BioConductor, Spotfire |
| Normalization | R/BioConductor, Microsoft Excel |
| Statistical Analysis | R/BioConductor, Microsoft Excel |
| Interpretation | GenMAPP, Cytoscape, PathVisio, Vanted |

Some of the tools mentioned in Table 2 are comprehensive and can perform several or all steps without the need to switch to a different software package. From

the perspective of user-friendliness, it is better to do as many steps as possible within the same software environment, as each switch usually involves the need to adjust data and to familiarize oneself with new software.

A shortcoming of PathVisio is that it is insufficient by itself, to perform a complete workflow from start to end. PathVisio deals only with the final interpretation step of this workflow, and assumes statistical analysis has already been applied using different software. The need to switch between software environments during analysis is a clear downside of PathVisio, and could be improved in the future. It would be an interesting possibility to allow t-tests, analysis of variance and multiple-testing corrections to be applied within the PathVisio environment, thereby integrating statistical analysis (step four).

Comprehensive tools that integrate the whole analysis pipeline are an improvement from the viewpoint of user-friendliness. On the other hand, from the perspective of analysis power, it is important to retain the possibility to switch to a different tool between every step. Especially when dealing with new analysis methods, new experimental designs or new technologies, it is unlikely that they are supported well, and the more general the software package, the less likely it is adapted to specialized circumstances. So while it is good to strive for comprehensiveness, it is important to create exit and entry points before and after every step of the analysis workflow. At the moment, PathVisio reads expression datasets in tab-delimited text format, with very few restrictions on the formatting of rows and columns. This format was chosen because it is supported by a large number of tools. Nevertheless, tab-delimited text is often the lowest common denominator. Software packages often have different preferred formats, tab-delimited text is provided as an export option, usually with some data loss. In the future, PathVisio should be improved to support more data formats generated by tools for statistical analysis. If statistical analysis was done using R scripts, PathVisio could read the binary format of R directly. PathVisio should be able to read XLS files generated when the statistical analysis is done in Microsoft Excel. Plug-ins could be developed in the future to read these formats and make switching more user-friendly.

## Modularity and code re-use

The previous sections listed a large number of tools for analysis and interpretation of high-throughput datasets. Clearly there is an abundance of tools available in the field of bioinformatics. Generally each tool has a distinguishing feature that gives it reason

for continued development. In the crowded space of bioinformatics applications, bioinformatics developers compete for scarce resources such as funding, attraction of other developers and user attention. Thus a type of natural selection takes place. As long as there is no "killer-application" that has all important features, a stable equilibrium may exist where multiple tools can occupy a niche that serves a public need, while none of them can gain complete dominance.

From the point of view of science as a whole, is this abundance of tools a good thing? The large selection could be considered a blessing of choice and flexibility, but is also a nuisance, for several reasons [30]. One reason, as mentioned in the previous section, the transition of one tool to another often involves usability problems and requires reformatting of data. Discoverability is another problem; a good tool may exist that can perform the analysis that you need, but it is hard to find amongst all choices. Tools vary widely in quality, which means that a search for them leads to results with a low signal-to-noise ratio. Small tools may never gain the necessary mindshare, meaning that it is not known by the people who could make good use of it.

We may hope for a convergence to happen, where user attention, developer attention and funding resources come together around a small number of big projects instead a large number of small ones. But this may never happen for several reasons. First, as mentioned before, newly developed tools may be better suited for new methods or new experimental technologies. This means that there will be a continuing need for specialized tools. Second, convergence may not happen because large development projects have a huge communication overhead that creates a high risk of running late and missing goals [31]. This does not mean that large projects are impossible, but it does mean that small projects have an inherent selection advantage in the evolution of bioinformatics software applications. Third, there is no consensus around development technologies (such as e.g. choice of programming language, redistribution license, operating system, GUI toolkit, or development framework), which severely limits the possibilities for convergence. For example, BioPerl [32] and BioConductor [33] are two large, successful bioinformatics projects, the former written in Perl and the latter in R, that can not easily share code with each other or with, for instance, PathVisio, which is written in Java. Even if cooperation across the boundaries of programming languages is possible, it always requires extra work. Because each programming language has its own strengths and weaknesses, it is unlikely that we will see a convergence around one. Java for example, is good at providing cross-platform programs with a user-friendly GUI, and provides ways to structure large code bases, but is less flexible than interpreted languages.

The question should not be what we can do to overturn the existing reality of continuing development of niche bioinformatics applications, but what we can do within the given situation to achieve the most desirable outcome.

A possible solution is to focus on maximizing the rate of code re-use. It may not be realistic to expect developers to voluntarily stop developing new tools, but it might be feasible to facilitate more sharing of code, so that a larger set of features is included in even the smallest tools. Each shared module then forms a building block for large applications. Another side effect of code re-use is that it facilitates user-friendly transitions between tools, because if modules are shared then this makes transfer of data from one tool to the next easier, without the need for re-formatting data files.

Which parts of code are most amenable to re-use? A software project can be thought of as a large number of modules, some modules requiring the functionality of other modules to work. Thus, a dependency tree can be drawn, a directed graph where edges represent a dependency of one module on another. For example, see Figure 1 to see the dependency tree of BridgeDb and PathVisio combined. This dependency tree is evolving as both projects continue development. Modules at the top depend on almost everything else; modules at the bottom depend only on the programming environment. PathVisio can be split into four main modules, and a number of smaller modules (called plug-ins).

Modules can be shared by two software projects only if the modules that they depend on are shared as well. This usually means that modules with the fewest dependencies are most amenable to sharing. As can be seen from this figure, the next module that could be most easily re-used is the PathVisio core module, because it only depends on libraries external to PathVisio project. The PathVisio user-interface module can only be re-used if the PathVisio core module is re-used as well. Plug-ins in PathVisio are actually the least amenable to re-use, because they depend on all other PathVisio modules and thus carry a large dependency load.

The PathVisio core module contains support for reading and writing GPML format, thus re-use of this module could allow other applications to add GPML support easily. The Cytoscape GPML plug-in already works that way. Future efforts could be directed to streamline the core module to make it another independent code library.

The modular structure of PathVisio is a strong point. The separation of modules facilitated the separation of BridgeDb as an independent code library. Before BridgeDb started life, it already existed as one of the modules of PathVisio itself. Identifier mapping, which is a distinguishing feature of PathVisio, was abstracted out in the form of a software library component. We decided to extract the identifier mapping functionality of PathVisio into a separate project, mainly as an experiment to share code between projects, to encourage code re-use. Currently BridgeDb is being used by Cytoscape and PathVisio, but more may follow in the future. Thus the experiment can be considered successful. Hopefully, more components of bioinformatics software can be identified that are amenable to code re-use in a similar way.

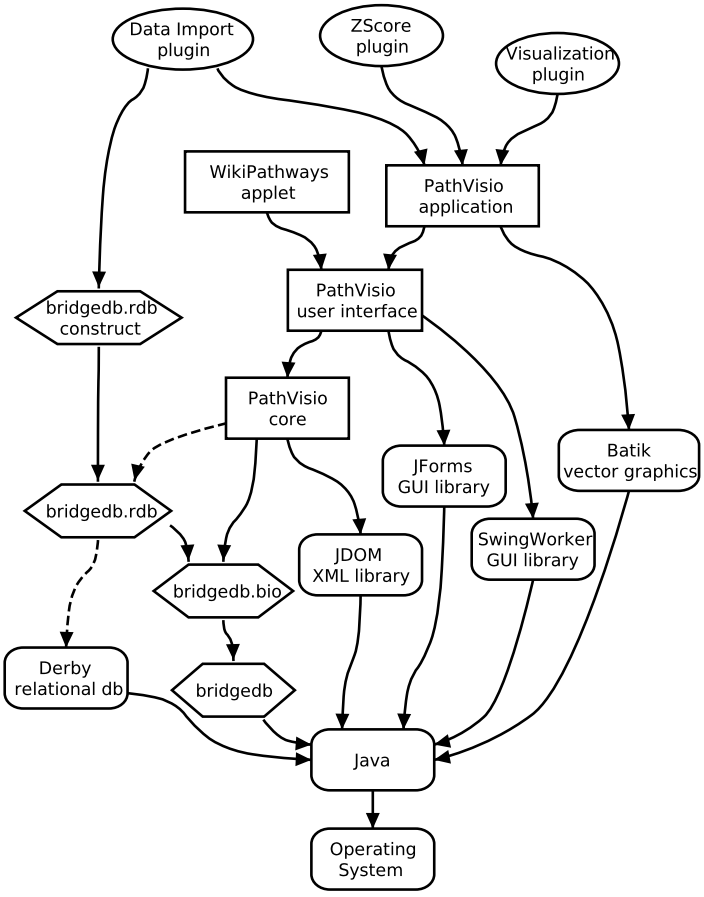


***Figure 1: dependency tree for PathVisio and BridgeDb modules.*** *Java libraries and components provided externally are represented by rounded rectangles, BridgeDb modules are represented by hexagons, main PathVisio modules are represented by rectangles and PathVisio plug-ins are represented by ovals. Absolute dependencies (i.e. at compilation time) are marked with solid arrows, and transient dependencies (i.e. at run time) are marked with dashed arrows.*

The recently started effort to create a standard library for SBGN using tools called LibSBGN can be seen as another effort in this direction. There is a large number of tools that support drawing in SBGN format, including CellDesigner [15], Vanted [17] and the Edinburgh Pathway Editor [34]. Even though these tools have an overlapping feature set, there is currently very little code shared. The hope is that a central LibSBGN module, co-developed by all tool developers, would improve this situation. LibSBGN could develop as a library for sharing diagrams using SBGN notation. A core module of LibSBGN would enable import and export to a valid SBGN file format, whereas

other modules higher up in the dependency chain could perform automated layouts, rule based validation and other interesting features that could be shared by a number of projects. When the LibSBGN library comes to fruition, incorporating it in PathVisio will be an important future improvement.

## Standards and software integration

Standardization efforts play an important role in bioinformatics. Bioinformatics studies are often based on complex workflows involving multiple databases and tools. Different software components must work well together to make workflows user-friendly and repeatable without errors. To this end, standards are being developed [35].

What is a standard? To enable cooperation, a detailed specification of something may be needed, be it an Application Programming Interface (API), a webservice, a computer file format or a programming language. When such a specification is accepted and used by many people around the world, it can be called a standard. Standards can be developed by international standards organizations, they can be developed by a community of stakeholders, or they can arise as de-facto standards through a dominant market share [36]. For example, the FASTA format was adopted by many sequence databases because of its simplicity, and therefore arose as a de-facto standard [37].

SBML and BioPAX are examples of standards that are developed by communities. SBML has been very successfully developed this way. The editorial committee of SBML has been very careful to include a large community of users. This has ensured that the imagined use cases of SBML match the actual need. SBML is successfully standardized, with support by multiple tools [38, 39]. The BioPAX standard has been developed in a similar way [38]. A disadvantage of community-based standard development is that progress is usually slow. Because of the large number of people involved, communication overhead is equally large.

PathVisio and WikiPathways integrate naturally with each other because they are based on the same code and use the same file format. Cytoscape integrates well with both projects using the GPML plug-in, which also shares part of the PathVisio code base. The GPML file format was derived from the GenMAPP MAPP format and improved in practice. But GPML is not an accepted standard outside a core group of contributors. This makes the use of GPML limited. A valid criticism of WikiPathways and PathVisio is that neither support standard pathway formats such as BioPAX and SBML. This hampers cooperation with other pathway tools. In the future, plug-ins should be developed to add support for these standards to PathVisio.

## Open source development

Open source, as defined by the Open Source Initiative (OSI), means software for which the source code is available to anyone to examine, improve and redistribute, without legal restrictions [40]. Open source enables the free exchange of ideas, encoded in software. The open sources principles are aligned with the ideals of science to "stand on the shoulders of giants". The advantages of open source software are numerous [41, 42], including:

1. *Open source is a prerequisite for maximal re-use of code.*
   As mentioned before, code re-use should be encouraged in the field of bioinformatics. An important prerequisite for code re-use is the absence of legal restrictions. Developers of bioinformatics applications will avoid the use of code libraries that cannot be redistributed together with the application. For example, the popular Gene Set Enrichment Analysis tool [43] is written in Java and freely available. Unfortunately, the associated license only permits personal use, which is not enough to qualify as Open Source under the criteria of the OSI, and would prevent redistribution in another software package. If in the future we want to add a Gene Set Enrichment Analysis feature to PathVisio, the functionality would have to be re-implemented.
2. *Open source makes it possible to share the load of maintaining and updating software.*
   If code modules are shared between developers from different groups, they can pool efforts to maintain and update that code. For example, developers from the BioPerl project [32] share the work of maintaining parsers for output of the blast program (which used to change frequently) [30].
3. *The absence of legal restrictions and hidden commercial interests encourage community formation.*
   With open source licenses, outsiders have nearly the same rights on the source code as the copyright owners. This creates a level playing field, which gives outsiders an interest in further improving the code. Thus, open source can attract developers, leading to the formation of a community with shared interests in the software.

The source code of PathVisio, BridgeDb and WikiPathways was published under the Apache 2.0 License, a permissive open source license. This is an important advantage in comparison to commercial or restricted bioinformatics software.

In recent years, the government of the Netherlands has increasingly called for valorization of academic research, by which is meant that the results of academic research should be used to create intellectual property for start-up companies [44]. This goal is mod-

eled after the successful development of companies in Silicon Valley around academic research, in particular at Stanford University. Open source development is sometimes perceived to be at odds with the goal of valorization, and this is a potential criticism of PathVisio. To counter this criticism, it must be noted that there exist valid business models around open source software, such as dual-licensing, using software as a loss-leader and selling software-as-a-service [42]. In those business models, the advantages of open source development outlined above offset the missing direct licensing revenue of shrink-wrapped software. Therefore open source software development is not at odds with valorization.

## Open access

The advantages of open access to knowledge and data are very similar to the advantages of open source. In parallel to the three advantages of open source discussed in the previous section we can identify:

1. *Open access allows knowledge to be re-used for new purposes.*
   Open access does not just mean unrestricted access to data; it also means the ability to create derivative works [45]. To maximize the usefulness of the WikiPathways collection, we want to avoid legal restrictions that form disincentives to new ideas for pathway analysis. The most interesting uses for pathways have probably not yet been imagined [46].

2. *Open access makes it possible to share the load of updating and curating data.*
   The primary reason for developing WikiPathways was the difficulty of updating the GenMAPP pathway collection to new research findings (chapter 4). WikiPathways aims to attract researchers to contribute fixes and updates. Through the creative commons license, contributors know that their work will be properly credited, but also that it will benefit everybody, not just the owners of WikiPathways.

3. *The absence of legal restrictions and hidden commercial interests encourage community formation.*
   It is beneficial to WikiPathways to have a community of curators that is as large as possible. The creative commons license ensures that everybody has the ability to copy the content of WikiPathways and use it as they see fit. If someday in the future WikiPathways takes a direction that is not approved by the community, they retain the right to create a so called “fork”, meaning to copy the data and develop in a different direction. The existence of this possibility is necessary to build trust, without which outsiders would not invest their valuable time [46].

WikiPathways can be seen as the antidote against commercial pathway databases. For WikiPathways we made a clear and explicit choice to publish all data under the Creative Commons attribution license. This license ensures that authors are credited, but creates no legal restrictions otherwise. This license was recommended by Science Commons [47], an organization that promotes unrestricted sharing of scientific data. The creative commons license affords the most freedoms (the only way to be more open is by dropping the attribution requirement, but attribution is an important motivator, especially in academia). The non-commercial clause would limit the possibility to make use of WikiPathways data in commercial tools. We believe that being able to combine WikiPathways data with commercial tools would actually encourage users of those commercial tools to contribute back to WikiPathways. The share-alike clause is less harmful but complicated and may instill confusion in users. We believe that making the terms of use explicit is an advantage of WikiPathways.

In the same spirit, all published articles in this thesis are open access publications.

## Final words

Life is incredibly complicated at the molecular level, and cannot be simplified. Through natural selection, life is optimized for fitness, not for understandability or ease of analysis. This makes it challenging to study complex nutritional disorders, such as type II diabetes, which are more and more prevalent in modern society.

The role of bioinformatics is to canalize the streams of data that exist in the post-genomic era. The toolset described in this thesis provides the means to tackle multiple aspects of this challenge. Pathway visualization with PathVisio makes it possible to see the data in an understandable context. Identifier mapping with BridgeDb can integrate different datasets. And as data turns into knowledge, it will be fed back again as a fresh new piece of pathway information in WikiPathways.

## References

1. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al*: **Reactome knowledgebase of human biological pathways and processes**. *Nucleic Acids Res* 2009, **37**(Database issue): D619-622.
2. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Res* 2008, **36**(Database issue):D480-484.
3. Waldrop M: **Big data: Wikiomics**. *Nature* 2008, **455**(7209):22-25.
4. Kohn KW, Aladjem MI, Weinstein JN, Pommier Y: **Molecular interaction maps of bioregulatory networks: a general rubric for systems biology**. *Mol Biol Cell* 2006, **17**(1):1-13.

5. Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM *et al*: **The Systems Biology Graphical Notation**. *Nat Biotechnol* 2009, **27**(8):735-741.
6. Breitling R: **Biological microarray interpretation: the rules of engagement**. *Biochim Biophys Acta* 2006, **1759**(7):319-327.
7. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
8. Curtis RK, Oresic M, Vidal-Puig A: **Pathways to the analysis of microarray data**. *Trends Biotechnol* 2005, **23**(8):429-435.
9. Suderman M, Hallett M: **Tools for visually exploring biological networks**. *Bioinformatics* 2007, **23**(20):2651-2659.
10. Cavalieri D, De Filippo C: **Bioinformatic methods for integrating whole-genome expression results into cellular networks**. *Drug Discov Today* 2005, **10**(10):727-734.
11. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D *et al*: **Visualization of omics data for systems biology**. *Nat Methods* 2010, **7**(3 Suppl):S56-68.
12. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D258-261.
13. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* 2005, **21**(18):3587-3595.
14. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways**. *Nat Genet* 2002, **31**(1):19-20.
15. Funahashi A, Tanimura N, Morohashi M, Kitano H: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks**. *BIOSILICO* 2003, **1**:159-162.
16. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W423-426.
17. Junker BH, Klukas C, Schreiber F: **VANTED: a system for advanced data analysis and visualization in the context of biological networks**. *BMC Bioinformatics* 2006, **7**:109.
18. Hu Z, Ng DM, Yamada T, Chen C, Kawashima S, Mellor J, Linghu B, Kanehisa M, Stuart JM, DeLisi C: **VisANT 3.0: new modules for pathway visualization, editing, prediction and construction**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W625-632.
19. Westenberg MA, van Hijum SAFT, Kuipers OP, Roerdink JBTM: **Visualizing genome expression and regulatory network dynamics in genomic and metabolic context**. *Comput Graph Forum* 2008, **27**:887-894.
20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

21. **GO-Elite - Software for Extended Pathway Analysis** [http://www.genmapp.org/go_elite/]
22. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
23. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**(16):3448-3449.
24. Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S: **Recent additions and improvements to the Onto-Tools**. *Nucleic Acids Res* 2005, **33**(Web Server issue): W762-765.
25. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity**. *Bioinformatics* 1999, **15**(3):180-186.
26. Nie L, Wu G, Culley DE, Scholten JC, Zhang W: **Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications**. *Crit Rev Biotechnol* 2007, **27**(2):63-75.
27. Joyce AR, Palsson BO: **The model organism as a system: integrating 'omics' data sets**. *Nat Rev Mol Cell Biol* 2006, **7**(3):198-210.
28. Neuweger H, Persicke M, Albaum SP, Bekel T, Dondrup M, Huser AT, Winnebald J, Schneider J, Kalinowski J, Goesmann A: **Visualizing post genomics data-sets on customized pathway maps by ProMeTra-aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example**. *BMC Syst Biol* 2009, **3**:82.
29. Waters KM, Pounds JG, Thrall BD: **Data merging for integrated microarray and proteomic analysis**. *Brief Funct Genomic Proteomic* 2006, **5**(4):261-272.
30. Stein L: **Creating a bioinformatics nation**. *Nature* 2002, **417**(6885):119-120.
31. Brooks FP: **The Mythical Man-Month**, Anniversary Ed. edn: Addison Wesley; 1995.
32. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H *et al*: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**(10):1611-1618.
33. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.
34. Sorokin AA, Paliy K, Selkov A, Demin OV, Dronov S, Ghazal P, Goryanin I: **The Pathway Editor: A tool for managing complex biological networks.** *IBM J Res & Dev* 2006, **50**(6):561-574.
35. Stromback L, Hall D, Lambrix P: **A review of standards for data exchange within systems biology**. *Proteomics* 2007, **7**(6):857-867.
36. Walli SR: **Chapter 8: Under the Hood: Open Source and Open Standards Business Models in Context**. In. Sebastopol: O'Reilly; 2006.
37. Stajich JE, Lapp H: **Open source tools and toolkits for bioinformatics: significance, and where are we?** *Brief Bioinform* 2006, **7**(3):287-296.

38. Stromback L, Jakoniene V, Tan H, Lambrix P: **Representing, storing and accessing molecular interaction data: a review of models and tools**. *Brief Bioinform* 2006, **7**(4):331-338.
39. Stromback L, Lambrix P: **Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX**. *Bioinformatics* 2005, **21**(24):4401-4407.
40. **The Open Source Definition | Open Source Initiative** [http://www.opensource.org/docs/osd]
41. Fogel K: **Producing Open Source Software**, 1st Ed. edn. Sebastopol: O'Reilly; 2006.
42. Raymond ES: **The Cathedral & the Bazaar**, Revised Edition edn. Sebastopol: O'Reilly; 2001.
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
44. **Valorisatie | Rijksoverheid.nl** [http://www.rijksoverheid.nl/onderwerpen/nederland-ondernemend-innovatieland/versterken-innovatief-vermogen/valorisatie]
45. MacCallum CJ: **When is open access not open access?** *PLoS Biol* 2007, **5**(10):e285.
46. Lessig L: **Free Culture: The Nature and Future of Creativity**, Reprint Edition edn: Penguin Books; 2005.
47. **Science Commons** [http://sciencecommons.org/]

# Abbreviations

## Abbreviations

| | |
|---|---|
| 2DE | Two-Dimensional Gel Electrophoresis |
| AF | Activity Flow |
| ANOVA | Analysis Of Variance |
| API | Application Programming Interface |
| BioPAX | Biological Pathway Exchange |
| DDO | Data-driven objective |
| Eef2 | Elongation factor 2 |
| eMIM | electronic Molecular Interaction Maps |
| ER | Entity Relationships |
| FDR | False Discovery Rate |
| Fth1 | Ferritin Heavy chain 1 |
| Ftl1 | Ferritin Light chain 1 |
| GenMAPP | Gene Map and Pathway Profiler |
| GPML | GenMAPP Pathway Markup Language |
| GSEA | Gene Set Enrichment Analysis |
| HMDB | Human Metabolite Database |
| HTML | Hypertext Markup Language |
| JDBC | Java Database Connectivity |
| KDO | Knowledge-driven objective |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| KGML | KEGG Markup Language |
| MIM | Molecular Interaction Map |
| MIRIAM | Minimal Information Required for the Annotation of Models |
| NCI-Nature PID | National Cancer Institute – Nature Pathway Interaction Database |
| OWL | Web Ontology Language |
| PCA | Principal Component Analysis |
| PD | Process Description |
| RMA | Robust Multichip Average |
| SBGN | Systems Biology Graphical Notation |
| SBML | Systems Biology Markup Language |
| SVG | Scalable Vector Graphics |
| Tpi1 | Triose phosphate isomerase |
| UML | Unified Modeling Language |
| URI | Uniform Resource Indicator |
| XML | eXtensible Markup Language |

# Summary

One of the reasons we study biology is to find cures for diseases. Some diseases are harder to study than others. Take for example cancer, a disease that has a big impact on public health, and consequently a lot of time and money is invested in researching it. But in spite of decades of research, we still do not understand enough of what goes wrong in a cancer cell, and we cannot always cure it.

Diseases such as cancer, and type II diabetes is another example, are what we might call *complex diseases*. The reason that they are so difficult to understand is that so many factors are involved. Inheritance, nutrition and lifestyle all play a role. There is not just one thing that goes wrong. A single inherited cancer gene is not enough to cause cancer, but it increases the risk. A single milkshake is not enough to cause diabetes, but a lifetime of overconsumption certainly does not help.

Complex diseases involve many genes, proteins and biochemicals. To better understand this, researchers have been trying, and succeeding, to measure anything and everything inside cells. In modern experiments, the activity of tens of thousands of genes can be measured at the same time. With all these measurements it is possible to completely characterize the state of a cell or tissue sample.

With so many measurements being done, it is very important to have good tools to analyze them. This is where bioinformatics comes in. Very broadly speaking, you could say that the goal of bioinformatics is to analyze biological data using computers. And when we measure different types of molecules, such as proteins and transcripts, we do not just have to analyze the data, we have to combine the different types in a single analysis. This is what we might call *data integration*.

In this thesis I have looked specifically at a set of analysis methods based on *biological pathways*. What is a pathway? Processes in the cell often form chains of reactions. For example, sugar molecules that are used as fuel for cells are not consumed all at once. This process takes place in several steps. First, a couple of enzymes prepare the molecule to be broken down, by adding chemical groups to energize it. When the sugar molecule is loosened up enough, another enzyme comes in and chops it in two. The two halves are then further processed by other enzymes still until there is nothing left. All these steps together form a pathway. Pathways are easiest to explain using *pathway diagrams*. Pathway diagrams are a little bit like the blue prints of cellular machinery. They clarify the roles of components and make their relations understandable.

By taking these two pieces, namely *pathway diagrams* and *data integration* together, we can formulate the main goal of this thesis, which is as follows:

*Goal: develop methods to integrate multiple types of experimental data and visualize them on pathway diagrams.*

## PathVisio

Pathway diagrams are easy to make. In the example given above, you could draw the sugar molecules and the enzymes as circles, and the chemical reactions as lines connecting the circles.

Pathway diagrams can be drawn with pencil and paper, but it is better to draw them using a computer. An electronic pathway diagram in a computer could be automatically linked to online databases that contain information about gene sequences, protein structures, chemical reactions and so on. Digitized pathway diagrams could be converted with the press of a button to a format suitable for a PowerPoint presentation or a journal publication. A collection of pathways could be indexed and searched quickly and easily.

To deal with pathways in software, we created a tool called PathVisio. PathVisio is first and foremost a drawing program for pathway diagrams.

Pathway drawing programs are not new. One of the first programs in this area was called GenMAPP. GenMAPP has many good features, such as a user-friendly graphical interface, and the ability to automatically link pathways to large experimental datasets.

But the biggest problem with GenMAPP was that it was written in an old fashioned programming language that is not adapted to the age of the Internet. That made it hard to implement new ideas and support new experimental methods. For example, GenMAPP was heavily focused on transcript measurements, and was not very suitable for other types of data. This was a reason to start the development of PathVisio.

Pathway diagrams have been created since the early beginnings of the field of biochemistry, often using pencil and paper. But there is no general agreement on the symbols used in pathway diagrams. Some diagrams use T-bars (a short perpendicular bar at the end of a line) to indicate negative feedback, other diagrams use an arrow with a minus sign next to it, yet others use a red color. The lack of standardization discourages researchers to make really complex and detailed diagrams, because other researchers will not understand the symbology without a lot of explanation. Because we are gaining more and more detailed knowledge about pathways, there is a need for a good, agreed upon way to put that knowledge down in a diagram.

One of the standard notations that is currently being pushed as a standard is called molecular interaction maps (MIM). PathVisio was the first bioinformatics tool to support MIM notation.

A newer system is called Systems Biology Graphical Notation (SBGN). The two are not completely separate: SBGN is a newer system that has copied several features of MIM.

A community of biologists is discussing and improving both notation standards. SBGN came out after the first version of PathVisio, so it is not yet supported. (The intention is that PathVisio will also support SBGN some day).

There is not a single bioinformatics application that is suitable for all research questions, or can handle all types of data. A well known bioinformatics application is Cytoscape, which, because of its plug-in system, has attracted a large group of bioinformatics developers. Cytoscape has features that PathVisio does not and vice versa. Therefore it is important that the two programs interact. We have enabled interaction by developing a plug-in in Cytoscape that can read pathways created in PathVisio. Transferring data from PathVisio to Cytoscape is a simple matter of copy and paste. Then Cytoscape can be used to improve the pathway, for example using a literature search plug-in, a feature that is not available in PathVisio.

## WikiPathways

We have a collection of pathways, which were originally created for GenMAPP, but which are also suitable for PathVisio. This collection would be most useful if it was always complete and up-to-date with the latest research developments. But keeping it up-to-date is a tremendous amount of work. New discoveries are being made all the time, and it is nearly impossible for a small group of people to keep up with all the new developments. Thus we faced the problem that our pathway collection was doomed to lag behind constantly.

In an attempt to solve this problem, we developed WikiPathways. WikiPathways is a website, inspired by the successful WikiPedia project, where any researcher can come and contribute pathway information. In this way the load of creating pathways and checking them against the scientific literature can be shared between a large group of people. We call this *community curation.*

On WikiPathways, each pathway has its own page. On such a page, you see a diagram of the pathway, and below that lists of genes, metabolites and literature references, plus links to information in other bioinformatics databases. Each pathway can be downloaded in several formats. But most importantly, just below the diagram there is a big “edit” button. After you click that button, you can edit the pathway directly on the website.

We expect that if more people join WikiPathways, the quantity and quality of the pathway collection will improve. Therefore we tried to make the website easy to use. For example, although we like standard graphical notations such as MIM (see above), we do not force people to use that out of fear that they would be put off by the extra complexity. Instead, pathways can be drawn in any style.

The pathway history page is another feature that is intended to lower the barrier for newcomers. WikiPathways remembers all old versions of each pathway. A newcomer may be afraid to edit a pathway out of fear of accidentally messing it up. But because it is always possible to go back to an old version, there is no danger of doing permanent damage. The absence of this danger removes an important psychological barrier to contributing. On the pathway history page, a visitor can see exactly how a pathway changed over time. Old and new versions of a pathway are shown side-by-side, and elements of the pathway are colored green if they have been added, red if they have been deleted and yellow if they have been modified.

WikiPathways is an experiment. The number of pathways, and also the number of users, is growing slowly but steadily, but it is too early to know if the wiki approach really gives good results in the long run.

## BridgeDb

I have mentioned already several times the possibility to link pathways to online databases. To make that link work, a gene on a pathway must have an identifier. That identifier then points to a record from one of the common bioinformatics databases, such as Entrez Gene from the USA or Ensembl from the UK.

Because there are so many different databases, you can choose from many identifiers. A record in database X describing gene G may be related to a record in database Y describing the same gene. Both databases contain almost the same information, but they use very different identifiers. To be able to put data from the two together, it is necessary to translate identifiers from database X to identifiers from database Y. This problem is called the *identifier mapping problem*, and this problem occurs every time two sets of data from different origins are being integrated.

The identifier mapping problem is often very messy. There are dozens of solutions out there, but they are very disorganized. For example, some work only for certain species – tough luck if you are studying an uncommon organism. You might encounter an online tool that does identifier mapping very well, but has a maximum of one thousand identifiers.

In an attempt to organize the existing identifier mapping tools, BridgeDb was created. BridgeDb is an application programming interface (API) that connects identifier mapping services on one side, to bioinformatics tools on the other side. Thus, a bioinformatics tool developer no longer has to choose which identifier mapping service to use – by incorporating BridgeDb all of them can be used together. With BridgeDb we took existing tools and organized them better so that they could be used more effectively, and applied in more situations.

## Open-source and Open access

A very important aspect that returns in all of the projects described in this thesis, PathVisio, WikiPathways and BridgeDb, are the principles of open access. Scientists never start from scratch: they always use the body of work of thousands of years of science before them. So an important academic ideal is the free sharing of information between scientists.

In this thesis, there are three types of information that are important to science: pathway diagrams, software and publications. All three types of information are important and should adhere to the basic principles of free sharing. All manuscripts are (or will be) published in open access journals. The source code of software is available under a so called open-source license. And access to the pathway data in WikiPathways is governed by creative commons licensing.

## Putting it all together

The goal of this thesis is to *develop methods integrate multiple types of experimental data and visualize them on pathway diagrams*. To meet that goal, we have created PathVisio, a tool to edit pathways. We built WikiPathways, a database for community-curated pathways. And we have developed BridgeDb, a system for mapping identifiers. Now finally, all the pieces of the puzzle can be put together and applied to a scientific study.

In this particular study, we look at the effects of long-term food deprivation in mice. Mice were not fed for several days (for a mouse, three days without food is equivalent to a whole month for humans). and the state of the cells in the small intestine was characterized using two technologies: microarrays on the one hand, which measures the amount of transcripts, and 2D Gel electrophoresis on the other hand, which measures the amount of proteins. These two datasets have been published before, but now they were combined for the first time.

Because we are integrating two diverse datasets, identifier mapping was very important to complete this study. For the protein data, Uniprot identifiers were used, for the microarray data, Agilent identifiers were used. The pathways have a mixture of identifiers that is predominantly Entrez Gene. Each type of identifiers had its own problems. For example, Agilent identifiers are not available everywhere, and we had to use sequence alignment to find the right mapping. The protein identifiers had a couple of mistakes in them that could only be fixed manually. However, because of the flexibility of BridgeDb, all these problems could be solved.

So what was the result of this analysis? Overall, starvation has an effect on two processes; cell turnover and energy metabolism. Both effects can be easily explained. The gut is normally a very active site of cell growth. Cells in the inner lining of the intestine con-

stantly divide, grow and then die off. Naturally, maintaining constant growth costs a lot of energy, and this results in the reduction of the cell cycle pathway (which is necessary for the growth of new cells), and the apoptosis pathway (which is necessary for cell death). The second process that is affected is energy metabolism. Of course, the body as a whole tries to save energy as much as possible, but there are important differences between organs. The gut for example, stores fat when it is abundant, and during a period of starvation, this fat is exported from the gut to the rest of the body. The glycolysis pathway is reduced, and lypolysis and gluconeogenesis pathways are activated.

After carefully comparing transcripts and proteins, a couple of interesting things can be noticed. Transcripts are the precursors for proteins, so we can normally assume that if the amount of transcript increases, the protein also increases. In this study, the average correlation between transcripts and proteins is not as high as you might expect (Spearman rank correlation of 0.21). Although transcripts are the precursors for proteins, there are several processes in the cell that could affect the correlation between the two.

The picture is very dependent on which gene you look at. Some genes have very good correlation between the transcript and the protein, others do not. For elongation factor 2 (Eef2), the protein expression is reduced, but the transcript clearly increases. For ferritin (heavy and light chain), the transcript is reduced but the protein is increased. Another interesting case is that of triose phosphate isomerase 1 (Tpi1), an important enzyme in the glycolysis. Analysis of the data shows that a truncated version of the enzyme becomes more abundant under starvation conditions, which would be consistent with the observation that the glycolysis pathway is reduced in activity. There is not sufficient data to provide an explanation for all these phenomena, but it is clear that these proteins are regulated by something other than transcription rate under conditions of starvation.

## Conclusion

The title of this thesis is “Data Integration with Biological Pathways”. Integration of data is currently one of the main problems in bioinformatics. In this thesis, integration of information occurs at several levels. Pathway diagrams can be used at one level, to integrate various bits of information, such as protein interactions, cellular locations, gene identifiers and literature references. At another level, identifier mapping services are used to integrate datasets from various sources. And finally, experimental data can be integrated with pathway diagrams to create visualizations that make the data easier to interpret.

# Acknowledgements

The past five years have taken me to many places far and wide, and along the way I met many helpful people without whom I would never have been able to bring this thesis to completion.

Dear Thomas, Alex and Kristina, I owe much to our fruitful cross-atlantic collaboration. From the very beginning, our work together has been natural and pleasant. Together we have been able to reach higher and further. Thank you Alex, for introducing me to so many new communities. Without those connections all the talk about open source would have been far less meaningful.

Dear colleagues from BiGCaT; over the years, I've watched BiGCaT grow in size and diversity. Thank you, Chris, Lars, Michiel, Charly, Adem, Stan, Arie, Rachel, Patrick, Jahn, Andra, Magali, Susan, Abhishek, Gontran and Tina. Thanks for the beers at the Vrijthof, the pizza nights, NuGO road trips, unreal tournaments and puma hunts. Thanks for Japan travel advice and trips to the *other* NY. Thanks also for the many lively discussions. Gontran, thank you for starting an irreversible linux & wiki revolution at BiGCaT. Thanks also to Ferry and the other interns and students who have contributed bits and pieces to PathVisio over the years. Dear Chris, thank you for giving me the freedom to try ideas and always encouraging collaborations, travel and working visits.

Dear Bruce, Nathan and the rest of the Conklin lab. Thank you for warmly welcoming me in San Francisco, showing me around, and making me feel at home. Dear Augustin, Mirit and Margot, thank you for your very kind invitation to Bethesda.

All the good open source bioinformatics tools and system biology standardization efforts are surrounded by communities of people with similar interests. Certain people from those communities have helped me tremendously along the way. I want to mention in particular Emek Demir and Igor Rodchenkov from BioPAX, Gary Bader and Allan Kuchinsky from Cytoscape, Mark Woon and Rebecca Tang from PharmGKB, and Jianjiong from the Google Summer of Code.

Dear Jildau and Suzan (Wopereis) for much useful feedback, and being my link to the metabolomics world. Dear Milka, Kaatje, Freek and Edwin, thank you for help and feedback, especially on proteomics aspects.

For proofreading (parts of) this thesis, I want to thank Alex, Susan, Augustin, Lars, Olivia, Tina, Andra and Michiel. And thank you Wouter, for helping me to get the cover design just right.

Dear Olivia, of all people I want to thank you the most. My orbit around the thesis repulsor field follows "periodic productivity", but your love and support have always put me back on track. Now I know that whatever the future may bring, together we can take on anything.

# Publications & Curriculum vitae

## Publications

1. **Martijn P van Iersel**, Milka Sokolovic, Kaatje Lenaerts, Freek Bouwman, Edwin C.M. Mariman, Chris T. Evelo **Pathway visualization applied to a multi-omics study of starvation in mouse intestine** *submitted*
2. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, **van Iersel M**, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Novère NL, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. (2010) **The BioPAX community standard for pathway data sharing.** *Nat Biotechnol.* 28(9):935-42
3. **Martijn P van Iersel**, Alexander R Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, Kristina Hanspers, Bruce R Conklin, Chris T Evelo (2010) **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services.** *BMC Bioinformatics* 11: 1
4. Thomas Kelder, Alexander R Pico, Kristina Hanspers, **Martijn P van Iersel**, Chris Evelo, Bruce R Conklin (2009) **Mining biological pathways using WikiPathways web services.** *PLoS One* 4(7): e6447
5. Susan L M Coort, **Martijn P van Iersel**, Marjan van Erk, Teake Kooistra, Robert Kleemann, Chris T A Evelo (2008) **Bioinformatics for the NuGO proof of principle study: analysis of gene expression in muscle of ApoE3*Leiden mice on a high-fat diet using PathVisio.** *Genes Nutr* 3: 3-4. 185-191
6. Baccini, Bachmaier, Biggeri, Boekschoten, Bouwman, Brennan, Caesar, Cinti, Coort, Crosley, Daniel, Drevon, Duthie, Eijssen, Elliott, van Erk, Evelo, Gibney, Heim, Horgan, Johnson, Kelder, Kleemann, Kooistra, **van Iersel**, Mariman, Mayer, McLoughlin, Müller, Mulholland, van Ommen, Polley, Pujos-Guillot, Rubio-Aliaga, Roche, de Roos, Sailer, Tonini, Williams, de Wit (2008) **The NuGO proof of principle study package: a collaborative**

**research effort of the European Nutrigenomics Organisation.** *Genes Nutr* 3: 3-4. 147-151

7. **Martijn P van Iersel**, Thomas Kelder, Alexander R Pico, Kristina Hanspers, Susan Coort, Bruce R Conklin, Chris Evelo (2008) **Presenting and exploring biological pathways with PathVisio.** *BMC Bioinformatics* 9: 09
8. Alexander R Pico, Thomas Kelder, **Martijn P van Iersel**, Kristina Hanspers, Bruce R Conklin, Chris Evelo (2008) **WikiPathways: pathway editing for the people.** *PLoS Biol* 6(7):e184
9. W B Derry, R Bierings, **M van Iersel**, T Satkunendran, V Reinke, J H Rothman (2007) **Regulation of developmental rate and germ cell proliferation in *Caenorhabditis elegans* by the p53 gene network.** *Cell Death Differ* 14: 4. 662-670
10. Markus Storvik, Pekka Tiikkainen, **Martijn van Iersel**, Garry Wong (2006) **Distinct gene expression profiles in adult rat brains after acute MK-801 and cocaine treatments.** *Eur Neuropsychopharmacol* 16: 3. 211-219

*Of interest to the contents of this thesis are also the following commentary articles:*

1. Waldrop M (2008) **Big data: Wikiomics.** Nature 455: 22-5
2. Doer A (2008) **We the curators.** Nature Methods 5 754-5

## Curriculum vitae

Martijn Peter van Iersel was born in 1979 in Tilburg, the Netherlands. He grew up in Udenhout, a small village near Tilburg, where he spent most of his time cycling back and forth to the city and reading "The Hitchhikers Guide to the Galaxy". Interest in computer programming, in particular programming of games, developed early on.

After finishing the Gymnasium in Tilburg in 1997, he went to Wageninen University to study Molecular Sciences, with Molecular Biology as specialization. As part of the MSc. program, he did a series of three large projects (these projects came under various names like internship, major and minor, but they were really broadly similar in scope and purpose). First he investigated virus-induced gene silencing in *N. benthamiana* at the laboratory for molecular biology in Wageningen starting in 2001. This project taught Martijn both the delight as well as the frustration of doing real experiments. The second project was at Kuopio University, Finland in 2002. There he spent some time in the sauna, and worked on a bioinformatics project to develop a database of microarray experiments compliant to the (then new) MIAME standard. This resulted in a brilliant piece of work that in fact completely failed to see the light of day, but it did teach him something about software project management. The third project was on a molecular biology subject again, on the analysis of the p53 homologue of *C. elegans* using RNA interference. This was done at the University of California Santa Barbara in 2003. Here Martijn discovered the beauty of *C. elegans*, an organism that is elegant, easy to experiment with, and so consistent and transparent that it just begs to be captured in a bioinformatics database.

Martijn continued to study RNA interference in *C. elegans* for one and a half years at the Hubrecht laboratory in Utrecht, but unfortunate circumstances caused him to switch back to bioinformatics again. Thus he started a PhD program at BiGCaT bioinformatics at the University of Maastricht in 2005, which ultimately culminated in this thesis. Starting from July 2010 Martijn is employed as Post-doc funded by the Netherlands Consortium for Systems Biology (NCSB).

Martijn writes about work-related interests on his blog at http://www.helixsoft.nl/blog.