Flexible Integration and Visualisation of *Drosophila melanogaster* Datasets

Martijn P. van Iersel[1], Julian A. T. Dow[2], Nadia Anwar[1]

[1]General Bioinformatics Ltd., Reading, UK
[2]Institute of Molecular Cell and Systems Biology, University of Glasgow, Glasgow UK

Abstract. The challenge in bioinformatics is not integration by itself, which could be achieved with ad-hoc scripting, but to do so in a manner that is repeatable, customisable, and enables powerful queries and visualisations. Here we present a case study which should provide some valuable insights. We set out to integrate public datasets from *Drosophila melanogaster*, including pathway data from BioCyc and tissue expression profiles from FlyAtlas. Cytoscape and PathVisio were used for visualisation of networks and pathways. We used existing ontologies, such as BioPAX, where possible. We found situations where the existing standard had to be amended, for example we inferred new Xref classes using Identifiers.org IRI's. For the purposes of repeatability and provenance, we made every aspect of our system fully scripted. We discovered that Cytoscape visualisations work best to answer questions about the whole dataset, such as "Which submodules of the data are primarily active in one tissue?". The PathVisio tool is more adept at answering detail-oriented questions, such as "Which enzymes are gatekeepers to the glycolysis, and in which tissues are they active?". Our position is that semantic technologies provide more flexibility; repeatability through scripting is of utmost importance; and multiple visualisation tools can be combined to address a broader range of biological questions.
Availability: The integrated data is available as RDF dumps, cytoscape session and a SPARQL endpoint at http://fly.cloud.generalbioinformatics.com

Introduction

Integration of diverse data is a recurring problem in life sciences. The challenge is not integration by itself, which could be achieved with ad-hoc scripting, but to do so in a manner that is repeatable, customisable, and enables powerful queries and visualisations. In short, data integration should be flexible. We believe that semantic integration provides this flexibility although it is not completely sufficient to deliver a "cradle to grave" solution. Here we describe a case study, which gives insights into the roads toward the desired flexibility.

General Bioinformatics performs bespoke data integration and visualisation for customers, using customer data combined with public domain data and tools. Semantic web technologies allow integration and management of multifarious biological data. Biological questions are often diffuse in nature and require the

ability to explore datasets. Thus, integration and knowledge management must be complemented with powerful visualisations. Semantic integration of data delivers the basis for generation of those visualisations.

Integration and Visualisation of Fly Data

We aimed to integrate and visualise public datasets concerning the fruit fly *Drosophila melanogaster*. We employed only open source software, namely Virtuoso for the triple store, Apache Jena, Rdflib and R for data processing, and Cytoscape [4] and PathVisio [10] for visualisation. We used public ontologies where possible, in particular BioPAX [3] for biological concepts and Identifiers.org [5] for identifier resolution. An overview of our architecture is shown in the Appendix in Figure 1.

The system is a collection of linked data, with the graphs stored in Virtuoso's triple store. We used the following datasets: pathways from BioCyc [6] and WikiPathways[8], tissue expression profiles from FlyAtlas, and gene and protein descriptions from FlyBase. In general, our approach is as follows: first, data is converted into triple format if not already available as such and loaded into Virtuoso. Each datasource is imported into a separate graph to allow independent updating. The second step is to create a layer of inferences to "loosely" integrate the data. After loading the data, it is available for querying, though the data is not yet "linked". Linking depends on two aspects: the use of common predicates or ontologies, and the use of common IRI's. Both can be inferred using SPARQL `CONSTRUCT` queries. We relied on Identifiers.org to provide common resolvable IRI's for biological concepts.

We consider repeatability of primary importance, and therefore all data preparation steps are performed using scripts, either in Python using Rdflib, or Groovy using Jena.

Integration

What follows is a more detailed description of some of the integration steps and inferences we used.

FlyAtlas [9] is a collection of tissue specific gene expression profiles of fruit flies. We obtained this data from openflydata.org in RDF form. (The RDF is shown in the Appendix as the black parts of Figure 2) For visualisation of FlyAtlas, both in PathVisio and in Cytoscape, it was necessary to calculate log fold-changes. Untransformed fold-changes map poorly to a linear range of colours since fold-changes between 0 and 1 correspond to the same dynamic range as fold-changes between 1 and infinity. We calculated log fold-changes using a separate script in the R statistical programming language and augmented the graph with these triples. The FlyAtlas data was combined with probe-to-gene mappings, using gene identifiers from FlyBase[7]. Similarly, linking to BioCyc, was achieved through creating triples using Identifiers.org IRI's for each flybase gene, using `biopax:xref` as predicate.

BioCyc is a collection of organism-specific databases of gene annotations, metabolic reactions, and pathways. For Drosophila this combined database is called FlyCyc and is available as BioPAX [3]. FlyCyc uses it's own identifier system for most fly proteins (with standard Uniprot being used for only a handful of proteins). However, we found that it was possible to create standard IRI's by performing text transformation.

WikiPathways [1] is a pathway database which accepts contributions from anyone. Although WikiPathways is less accurate, it has much broader coverage of topics, including signalling and regulatory pathways not found in BioCyc. To incorporate WikiPathways into the system we had to ensure the availability of sufficient fly-specific pathways. Therefore, we created translations of pathways based on ortholog mappings provided by Inparanoid [2].

Visualisation

The integrated data can be visualised differently depending on the biological question to be addressed.

To get a global overview of the data, it is possible to load the whole network of enzymes and metabolites into Cytoscape, and project the expression profiles on top. Network visualisation in Cytoscape can give answers to the following type of question: which submodules of the metabolic network are tissue dependent? Cytoscape also allows one to apply different layout algorithms, and to see the connections between pathways.

The transfer from triple store to Cytoscape takes two steps. First, the backbone of the Cytoscape graph is generated using a SPARQL `SELECT` query with three output columns, representing source node, edge type and target node. The result is stored in Simple Interaction Format (SIF), which is imported into Cytoscape. In the second step, any number of node and edge attributes can be extracted with further SPARQL `SELECT` queries.

We used Groovy scripts to automate the procedure of querying our triple store and saving the data to SIF format. We then load this data into Cytoscape from within the same script, and generate network views, by using the Cytoscape RPC plug-in. (Figure 4 in the Appendix shows what the result looks like in Cytoscape and this Cytoscape session is also available for download from our website.)

Cytoscape provides a complete overview of a network, and it can be hard to get a detailed view of just a single pathway. PathVisio is better at displaying pathways, and better at answering pathway-oriented questions, such as: in which tissue is a given pathway active?

To answer this question, we also need to know which tissue the constituent genes are primarily active in. Simply assigning a gene to the tissue where it is most highly expressed is not statistically sound, since a gene could be just a fraction higher than average in one tissue due to random chance. Therefore we used an unsupervised clustering to group genes with similar expression patterns. We performed k-means clustering in R to assign genes to clusters. One characteristic of k-means clustering is that the number of clusters has to be chosen a priori.

After several attempts we settled on 16 clusters, as that appeared the optimal number where clusters appeared to have a distinct, biologically interesting profile. Cluster numbers were added to `flyatlas:CalculatedValues` for the whole fly (See also Figure 2, pink node in the Appendix below).

For example, cluster number `1` contains genes that are over expressed in the fatbody tissue. To see the effect of this, we ranked pathways by enrichment of this cluster. Indeed, the resulting ranking shows a clear pattern, with many pathways related to fat metabolism (For further details please see Table1 in the Appendix). We created similar rankings for each of the 16 clusters, which are available for download from our website.

PathVisio allows for a more detailed view of what is happening in pathways. Examination of the glycolysis pathway image quickly leads us to find tissue specific-effects occurring in CG10924, the first step of the gluconeogenesis, which is highly overexpressed in fatbody. This is shown in the Appendix in Figure 1. Also SLC2A1, a glucose transporter with a gatekeeper function, appears to be regulated in a highly tissue-specific manner.

Discussion

During the course of this project we encountered some pitfalls, and it may be instructive to mention them here. For starters, we heavily relied on existing ontologies and actively encourage the use of existing standards such as BioPAX. BioPAX is a mature standard for pathway exchange. However, we found that we needed to add to the BioPAX standard. For visualisation in Cytoscape, we wished to create a bipartite graph representation of metabolic reactions, where each node corresponds either to an entity (metabolite or enzyme) or to a process (reaction or transport). Transformation of the BioPAX graph to a bipartite representation is made difficult by the fact that each connection between biochemical reactions and enzymes is mediated by a `catalysis` object. A `catalysis` object ties the enzyme (`BP:controller`) and the reaction (`BP:controlled`), plus a few more properties. Therefore, we inferred a third relationship directly between BioPAX `BiochemicalReaction` and enzyme named `BP:controlledBy` (See Figure 3, red parts). With this extra axiom it is now possible to extract the complete bipartite metabolic network in a single query.

We faced a generic problem with BioPAX `EntityReferences`.
`EntityReference` is a class with two datatype properties, one to refer to a bioinformatics database, e.g. "FlyBase", and one for a local identifier of that database, such as "FBgn0013682". The local identifier by itself is not enough to uniquely identify a biomolecule. Links across graphs based on EntityReference are complex operations that involve comparing two literals at the same time. Matching in SPARQL is more efficient when the nodes use an IRI. We inferred new bp:xref's for each existing bp:xref, using a standard IRI from Identifiers.org. (This is depicted in the Appendix in Figure 3, purple parts). Identifiers.org enables convenient identifier resolution, as well as direct URI-based linking. These are customisations for BioPAX that we find useful, and we believe that well-

written tools that consume BioPAX ought to be tolerant to this layer of inferences and adopt the fully flexible open world assumption that is offered by using semantic technologies. Although most BioPAX tools assume a closed world, we suggest that a more flexible approach in these tools would be more appropriate and will encourage consumer use of BioPAX data.

In this study we had difficulties using blank nodes in RDF. The problem we faced was to add additional calculated data to the flyatlas:TissueVals nodes, which were all blank nodes. We were generating a graph with calculated values. We wanted to combine this with the existing graph, but it is close to impossible to merge two graphs on blank nodes - on import, new blank nodes get generated, leading to all kinds of problems in querying the data. The solution in our case was to construct new nodes with assigned IRI, and re-assert triples for each triple asserted for that blank node. The generated IRI is not resolvable, but it is identifiable. And that makes it still preferable over blank nodes, which are neither resolvable nor identifiable. We believe the use of blank nodes should be avoided for resources where new triples might be asserted in the future. Given that it is hard to predict the future use of data, our general advice is to avoid blank nodes altogether when possible.

Ideally, a fully flexible system would achieve complete decoupling between data and tools, using only the semantic layer as intermediate. In essence, none of the output tools depend on any of the input tools: such that we could swap input data and each tool can be easily adapted to use fresh data. Thus, WikiPathways could replace BioCyc as a source of pathway data, and Ensembl could replace Flybase as a gene database. However, this is not always possible to achieve because of data dependencies of tools. For example, although PathVisio can read high-throughput data from our system, it can not yet read pathway data in triple form and relies on a direct link to WikiPathways. In this respect our system could be improved.

Conclusion

The flexibility that we strive for is difficult to define, we believe that integration should be fit for purpose and yet it is impossible to predict how data can be used and reused as technologies progress. Given this, integration should be as flexible as current technologies allow. We feel strongly that semantic web technologies provide this flexibility. In semantic reasoning, it is important to pragmatically focus on the problem at hand. We recommend to use simple identifier schemes such as Identifiers.org, and we readily expand the boundaries of BioPAX with new inferences when it serves our purpose. Semantic integration relies on common IRI's used by the subjects in the triples and the ontologies used in the predicates. In this project we used a loose integration approach and added layers of inferences on top of FlyAtlas and other datasets. The beauty of semantic technologies is that new assertions can be easily added in the RDF or OWL as needs dictate, giving a fully flexible data representation.

In our opinion any analysis must be easily reproducible. Therefore, every part of the exercise should be scripted in a manner that allows not just for repeatability but allow tweaks. For example, If triple loading and construct queries were done manually, we would quickly lose track of the source of a given triple, and lose the ability to make small improvements to the process. In this respect, our choice of software was not arbitrary: we picked tools for their scriptability (and this requirement incidentally favours open source software). We created a custom API to enable scripted access to Virtuoso and scripted all Cytoscape visualisations using its RPC interface.

Finally, we want to emphasize the usefulness of combining visualisation tools with semantic technologies. In our experience biological questions require the transformation of data, re-purposing and re-spinning data into specific contexts and delivering it in a format that can be visually understood. Data that has been integrated is by definition more complex than the individual sources. The sum is greater than the parts. We found that data integration is not sufficient when answering biological questions and that being able to visualise integrated data in the contexts of the biological questions is what makes the integration valuable. Furthermore, we believe that the use of multiple independent visualisation tools helps to address a broad range of biological questions. Too heavy reliance on a single piece of software would hurt this capability, on the other hand, being able to take the data and visualise for specific purposes is a considerable asset. The Cytoscape network visualisation tool shines when used for whole-dataset questions. For example, it can quickly generate visualisations to answer questions such as "what fraction of enzymes have been measured in FlyAtlas?", "which submodules of the data are primarily active in one tissue?". PathVisio, on the other hand, is more adept at answering detail-oriented questions, such as "Which enzymes are gatekeepers to the glycolysis, and in which tissues are they active?".

For our clients we find the semantic web approach to data integration is not only effective, it is more efficient and more scalable. However, biological questions rarely come in the form that can be delivered in a single SPARQL query, and it is the visualisation of integrated data that has benefited our clients in particular.

References

1. Alexander R. Pico et al. Wikipathways: pathway editing for the people. *PLoS Biol*, 6(7):e184, Jul 2008.
2. Ann-Charlotte Berglund et al. Inparanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, 36(Database issue):D263–D266, Jan 2008.
3. Emek Demir et al. The biopax community standard for pathway data sharing. *Nat Biotechnol*, 28(9):935–942, Sep 2010.
4. Michael E. Smoot et al. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, Feb 2011.
5. Nick Juty et al. Identifiers.org and miriam registry: community resources to provide persistent identification. *Nucleic Acids Res*, 40(Database issue):D580–D586, Jan 2012.

6. Ron Caspi et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 38(Database issue):D473–D479, Jan 2010.
7. Susan Tweedie et al. Flybase: enhancing drosophila gene ontology annotations. *Nucleic Acids Res*, 37(Database issue):D555–D559, Jan 2009.
8. Thomas Kelder et al. Wikipathways: building research communities on biological pathways. *Nucleic Acids Res*, 40(Database issue):D1301–D1307, Jan 2012.
9. Venkateswara R. Chintapalli et al. Using flyatlas to identify better drosophila melanogaster models of human disease. *Nat Genet*, 39(6):715–720, Jun 2007.
10. Martijn P. van Iersel et al. Presenting and exploring biological pathways with pathvisio. *BMC Bioinformatics*, 9:399, 2008.

Appendix

Further details of the data, integration and visualisation are provided below. Our results are available on our public website. There is a SPARQL endpoint and downloadable RDF of the integrated data, a Cytoscape session containing the network visualisation, and a collection of pathway visualisations created with PathVisio. This and more details can be found at
http://fly.cloud.generalbioinformatics.com

For visualisation of tissue enriched pathways we created pathway rankings for each of the 16 clusters. An example of the ranking for Cluster 1 is shown below:

Table 1. Ranking of pathways that are enriched with cluster 1.

Pathway	positive genes (r)	measured genes (n)	%	Z score
Glucuronidation	9	27	33.33%	7.67
Catalytic cycle of Flavin-containing mono-oxygenases	2	2	100.00%	6.79
Vitamin D synthesis	5	12	41.67%	6.54
Synthesis and Degradation of Ketone Bodies	3	5	60.00%	6.27
Amino acid conjugation of benzoic acid	3	5	60.00%	6.27
Oxidative Stress	9	41	21.95%	5.79
Mitochondrial LC-Fatty Acid Beta-Oxidation	4	15	26.67%	4.39
Sulfation	2	5	40.00%	4.02
Steroid Biosynthesis	1	2	50.00%	3.25
Pentose Phosphate Pathway	2	8	25.00%	2.96
Glycogen Metabolism	3	18	16.67%	2.68
Fatty Acid Beta Oxidation	4	36	11.11%	2.12
Fatty Acid Biosynthesis	2	15	13.33%	1.79
Signal Transduction of S1P Receptor	1	7	14.29%	1.35
Notch Signaling Pathway	2	22	9.09%	1.17
Heart Development	2	23	8.70%	1.1
Fluoropyrimidine Activity	2	26	7.69%	0.91
Glycolysis and Gluconeogenesis	2	27	7.41%	0.85
Triacylglyceride Synthesis	2	27	7.41%	0.85

For demonstration, the graphs and visualisations are exemplified in Figures below.

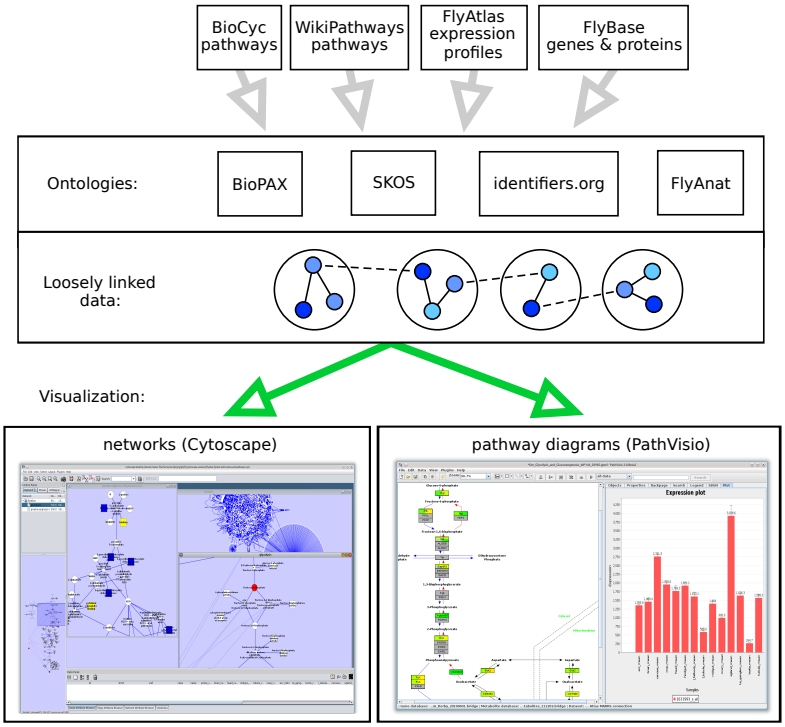

Fig. 1. Architecture overview

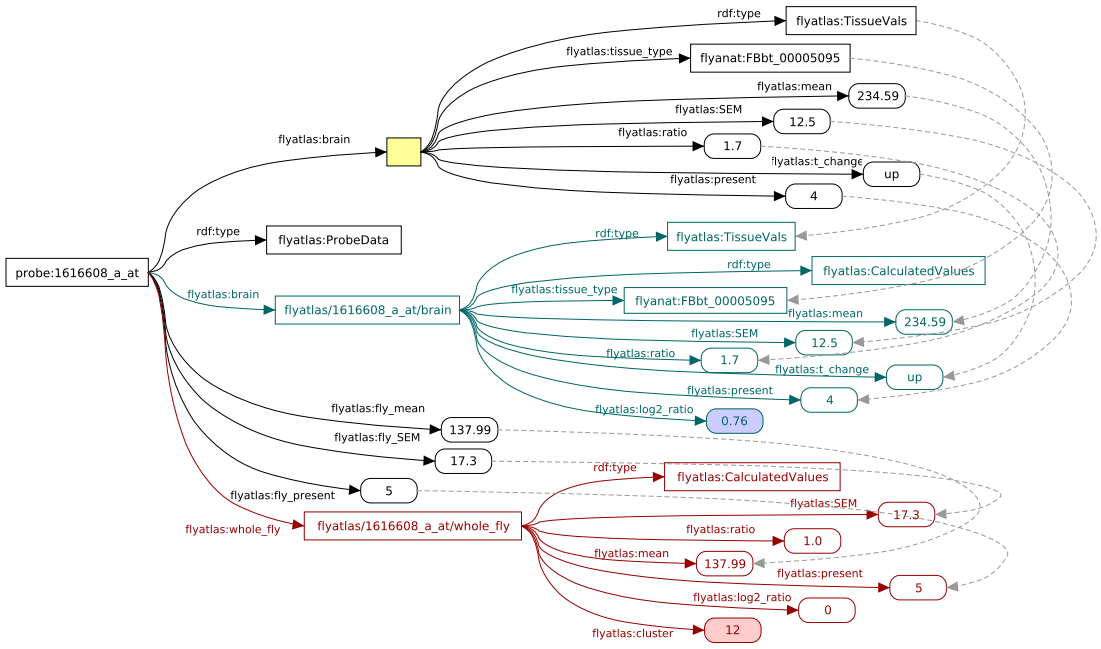

Fig. 2. Sample of triples from the FlyAtlas graph model, with inferences. This figures shows the triples asserted for probe 1616608_a_at. Only data for two tissues, namely brain and whole fly, are shown. Literals are represented by rounded nodes. Black: triples from the original data source. Teal: inferred CalculatedValues to work around the presence of blank nodes (yellow). The blue node contains the log-transformed fold-change. Red: whole fly values. The pink node contains the cluster designation. For each probe/tissue combination, there is a `flyatlas:TissueData` resource with associated statistics for the average expression of that probe, standard deviation and the fold-change ratio compared to whole fly. Additionally, there is data for the number of replicates in which a probe was deemed "present" by Affymetrix software, and a textual label containing either "up", "down" or "none" (indicating positive fold-change, negative fold-change, or statistically insignificant change according to a standard t-test). Each `flyatlas:TissueData` resource also has a reference to the relevant fly anatomy ontology term.

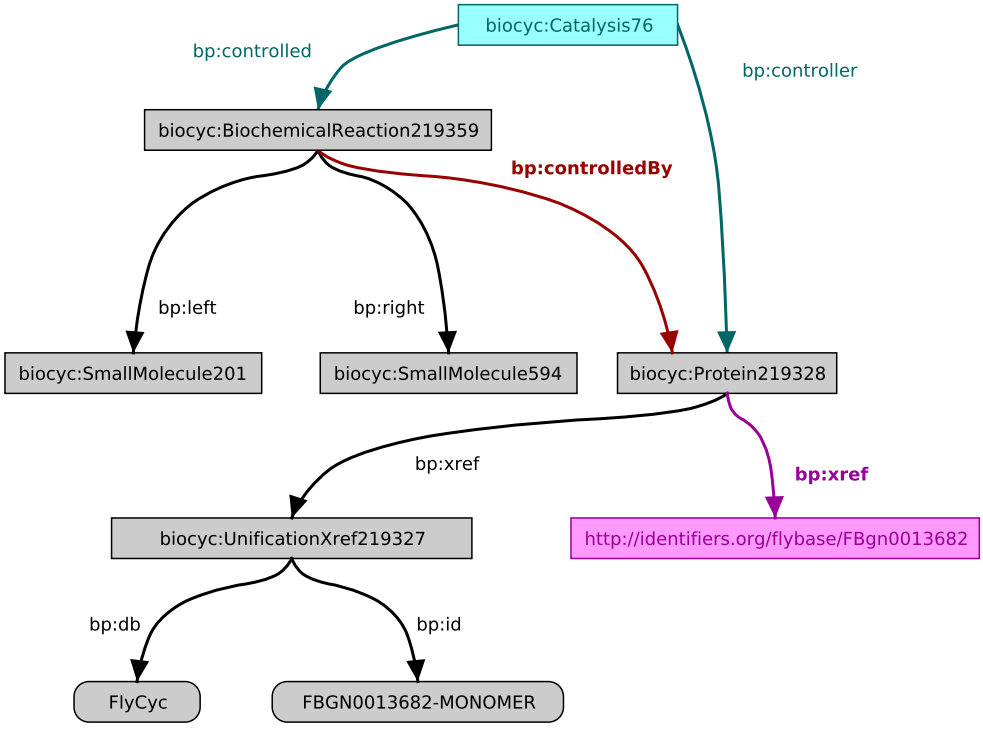

Fig. 3. fragment of a BioPAX graph, illustrating the inference of bp:controlledBy relations, and the inference of identifiers.org URI's for bp:xref. Catalysis objects (blue) link enzymes with reactions, but they have no place in a bipartite graph representation of metabolic networks. We replaced them with controlledBy relations (red). Furthermore, we inferred identifiers.org IRI's (purple) for existing Xref's to make it easier to merge two graphs.

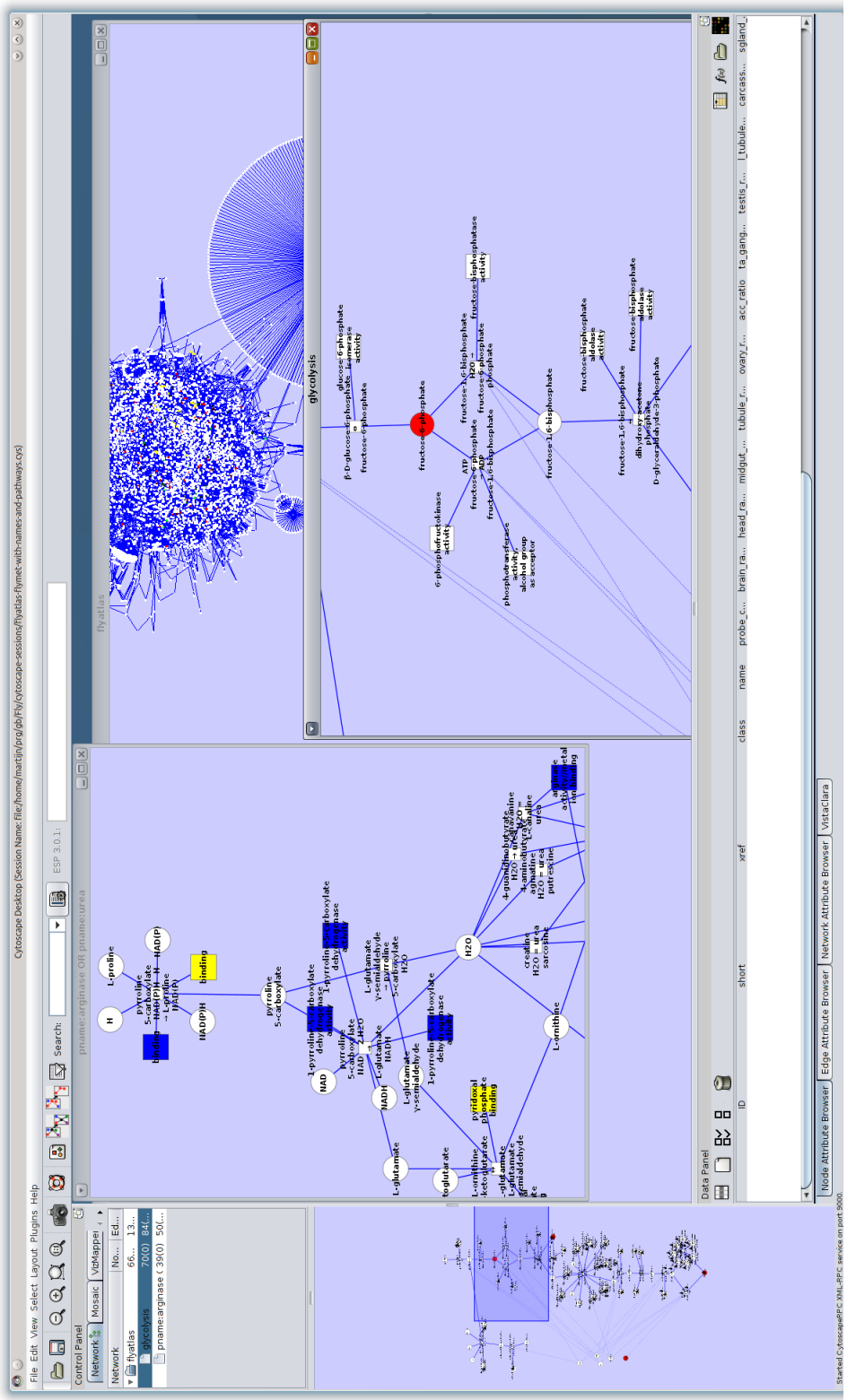

Fig. 4. Visualisation of a bipartite metabolic network in Cytoscape. The main graph is shown in the background. The other windows highlight parts of the network.

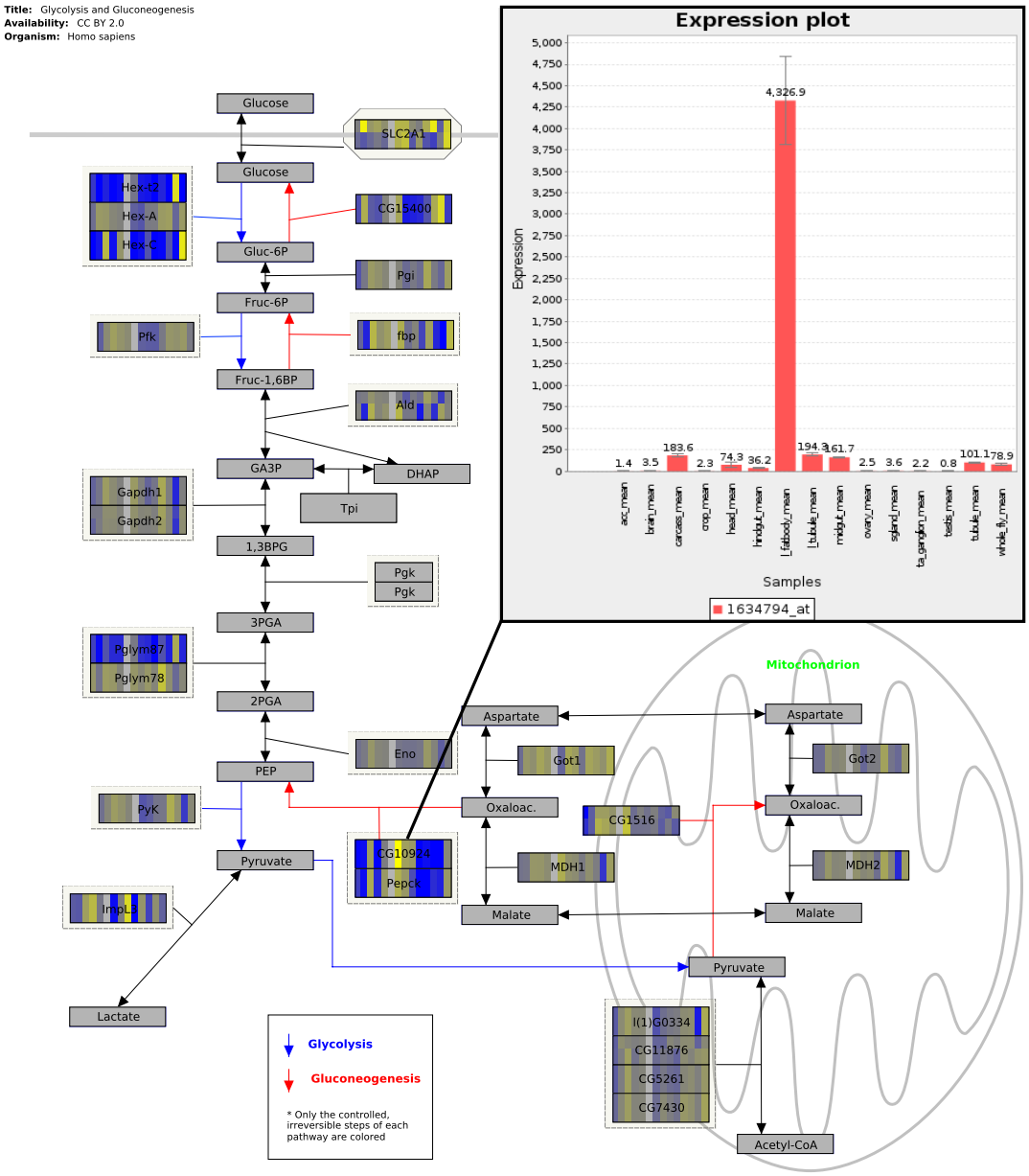

Fig. 5. Illustration of visualisation of the glycolysis and gluconeogenesis pathways.. Each box represents a small heat map representation of the expression profile of that gene. Examination of the bright yellow sport in this image can help find the tissue specific-effects occurring in CG10924, the first step of the gluconeogenesis, and SLC2A1, a glucose transporter. Inset contains a bar chart of the expression level of CG10924 in each tissue.