Comparison of Metabolic Pathways using Constraint Graph Drawing

Falk Schreiber

School of Information Technologies
University of Sydney
Madsen F09, Sydney, NSW 2006, Australia
Email: falk@it.usyd.edu.au

Abstract

Databases contain a large amount of data about metabolic pathways, in particular about similar pathways in different species. Biologists are familiar with visual representations of metabolic pathways. Visualizations help them to understand the complex relationships between the components of the pathways, to extract information from the data, and to compare pathways between different species. However, visual interfaces to metabolic pathway databases cannot cope well with the visual comparison of similar pathways in different species.

This paper presents a new approach for the visual comparison of metabolic pathways using a constraint graph drawing algorithm. Using layout constraints the same parts of similar pathways in different species are placed side by side, thereby highlighting similarities and differences between these pathways. This visualization method can be used as a visual interface to databases and has been tested with data obtained from the BioPath system and from the KEGG database.

Keywords: Metabolic pathways, visualization, graph drawing, constraints

1 Introduction

Metabolic reactions form a large and complex network consisting of substances, co-substances, and enzymes with multiple interconnections representing reactions and regulation. A *metabolic pathway* is a part of this network. An example for the complexity of metabolic pathways is given by the *Boehringer Biochemical Pathways* poster (Michal 1993); see Figure 1 for a cutting of this poster. A large amount of data about metabolic processes in different species is accessible from freely available databases. Biologists are familiar with visual representations of metabolic pathways; they help them to understand the complex relationships between the components of these pathways, to compare similar pathways across different species or organism classes, and to find differences in pathways. Such different parts in pathways highlight different metabolic processes in these species and may be target points for new drugs.

A *pathway diagram* is a visual representation of a metabolic pathway. Pathway diagrams, as in the poster (Michal 1993), in textbooks on biochemistry, and in information systems such as *KEGG* (Kanehisa & Goto 2000) are made manually. These diagrams represent the knowledge at the time of generation. They are created once, used very often, and they are usually hardcopies. This type of pathway visualization is called *static visualization* (Brandenburg, Gruber, Himsolt & Schreiber 1998).

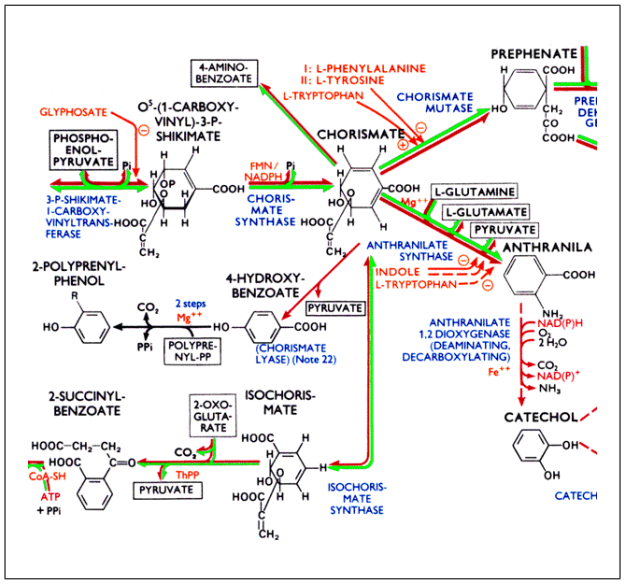

Figure 1: A small part of the *Boehringer Biochemical Pathways* poster (Michal 1993). Different colours are used for metabolic reactions in different organism classes.

However, for visual interfaces of biological databases and for interactive comparison of metabolic pathways, dynamic visualizations are necessary. *Dynamic visualization* is the generation of a pathway diagram on demand at the time the drawing is needed. For this kind of visualization metabolic pathways are typically modelled as graphs or hyper-graphs in which nodes represent the substances, co-substances, and enzymes within a pathway, and edges represent the reactions. The placement of nodes (e. g., substances) and the routing of their connections is a typical graph drawing problem. Special graph drawing algorithms have been developed for the dynamic visualization of metabolic pathways (Karp & Paley 1994, Mendes 2000, Becker & Rojas 2001, Schreiber 2002).

These algorithms focus on the visualization of simple pathways, not on their visual comparison. A common approach to compare similar pathways in different species is to draw a general pathway map and to distinguish between different species by colours. For example, in the visual interface of the KEGG database all enzymes found in the gene catalog of a specific species are marked green in the reference pathway map in order to identify the species-specific pathways (see Figure 2). To compare pathways in two species, two diagrams are needed. Some approaches show pathways from different species in one diagram. The *Boehringer Biochemical Pathways* poster uses four different colours for general pathways (black), animals (blue), higher plants (green),

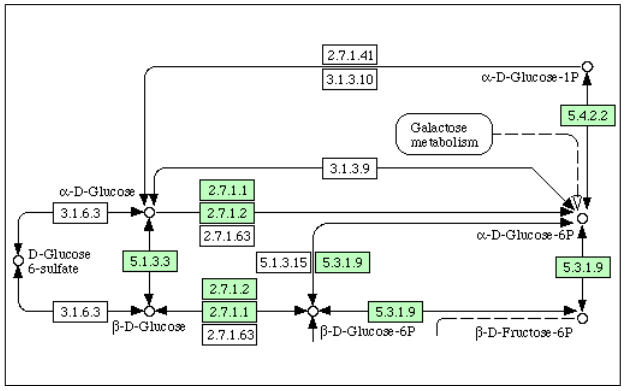

Figure 2: The visual interface of KEGG (Kanehisa & Goto 2000) highlights enzymes which occur in a specific species in the reference pathway diagram. As an example, this image shows a pathway map of *Saccharomyces cerevisiae*.

and unicellular organisms (red) as shown in Figure 1. Figure 3, a dynamic visualization obtained from the *BioPath* system (Forster, Pick, Raitner, Schreiber & Brandenburg 2002), shows a comparison of a metabolic pathway in different organism classes. The colour-code is the same as that used on the poster. Mixed colours are used for pathways occurring in more than one organism group, for example reactions in unicellular organisms and higher plants are brown.

The main drawback of this approach is that comparing pathways is quite difficult. For example, in Figure 3 the similarities and differences between the pathways in animals, higher plants and unicellular organisms are not obvious. Figure 4 shows a better solution where the pathways for these organism classes are visualized separately. This drawing shows the similarities and differences much better than the visualization in Figure 3, however, it can be improved. The differences and common parts of these pathways are not highlighted. For example, it is not obvious from the picture that there is a reaction *D-fructose* $\leftrightarrow$ *β-D-fructose-6-phosphate* catalyzed by the enzyme *hexokinase* in all three pathways.

This paper presents a new approach for the visual comparison of metabolic pathways which uses a constraints graph drawing algorithm. Using layout constraints, the same parts of similar pathways in different species are placed side by side, therefore highlighting similarities and differences between these pathways as shown in Figure 7.

2 Visualization background

Our visualization method extends the graph drawing algorithm in (Schreiber 2002). This algorithm computes general layered (hierarchical) drawings of graphs taking into consideration node sizes and layout constraints. It is used for the dynamic visualization of metabolic pathways and has the following properties:

1. Local placement: Components of the reactions are placed in the established drawing style of biochemistry as in textbooks (e. g. enzymes and co-substances beside the reaction arrow; co-substances in the correct order).
2. Global placement: All reactions are placed according to their temporal order and distinguished paths such as open and closed cycles (e. g. citrate cycle, fatty acid synthesis) are placed such that the cyclic structure is clearly visible.

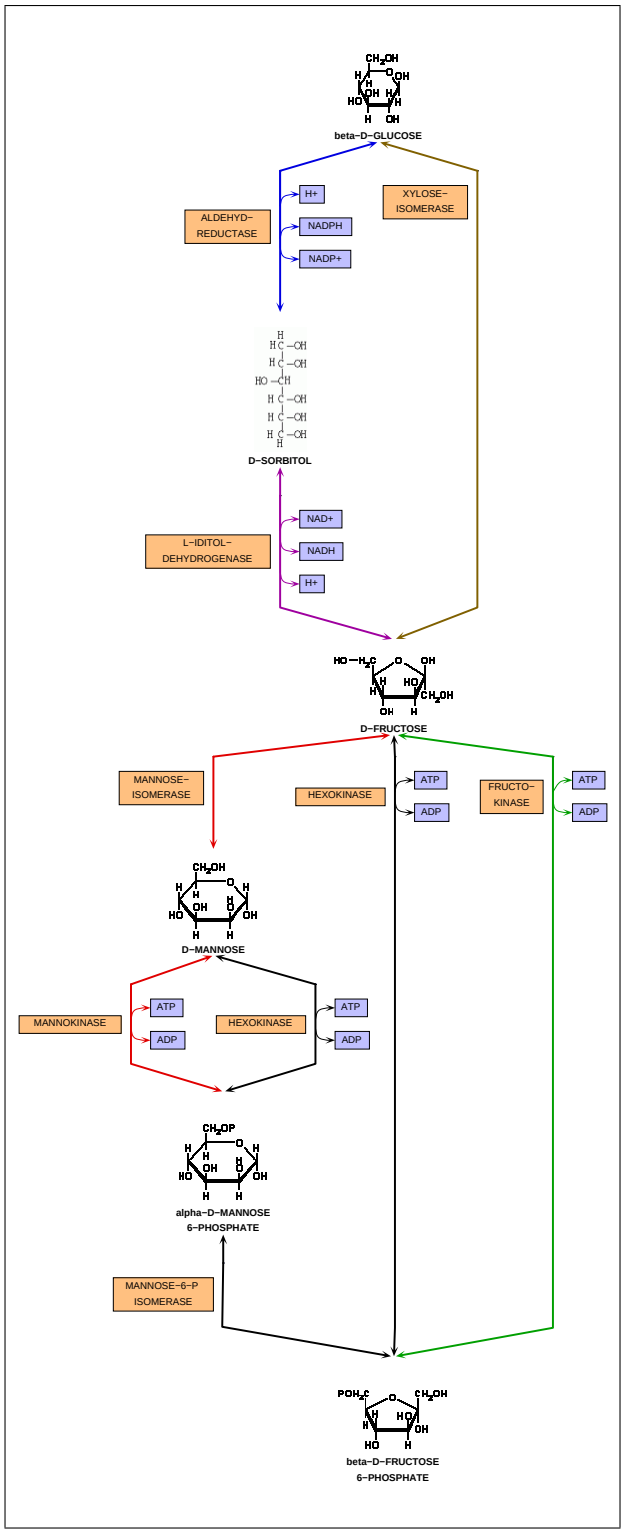

Figure 3: Similar metabolic pathways in different species can be compared by using different colours in a general pathway diagram.

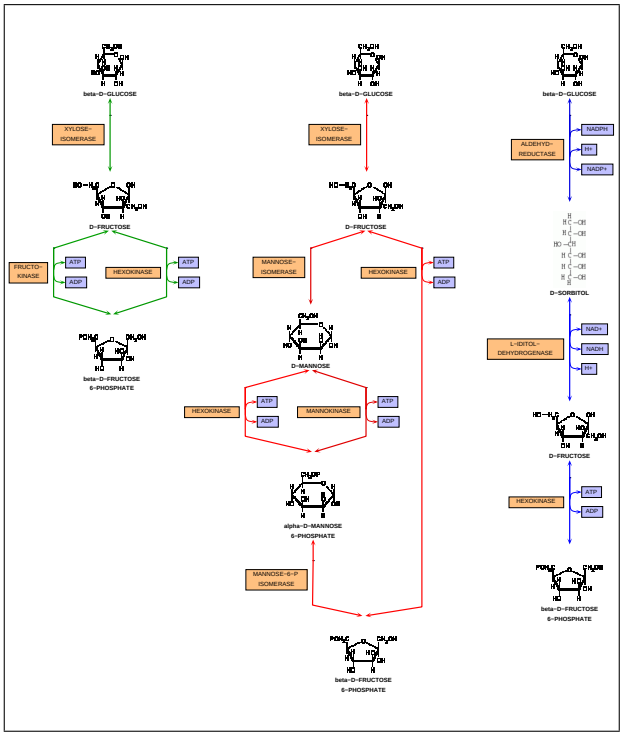

Figure 4: Separate visualization of the pathways shown in Figure 3 (left: in higher plants, middle: in unicellular organisms, right: in animals). This diagram shows similar pathways in these different organism classes, however, differences and common parts of these pathways are not obvious.

We represent a metabolic pathway as a directed bipartite graph $G = (V_1 \cup V_2, E)$, a modelling also used in Petri-net representations of pathways (Hofestädt & Thelen 1998). The nodes $v \in V_1$ represent the substances, co-substances, and enzymes within a pathway, nodes $v \in V_2$ represent the reactions. A *substance* can be a reactant or product; a *co-substance* can be a co-reactant or a co-product. A *compound* means all elements of a reaction (substances, co-substances, and enzymes). Edges are binary relations connecting compounds of reactions (nodes of V_1) with reactions (nodes of V_2). Figure 5 shows an example of this modelling.

First, we recall the main steps of the graph drawing algorithm (Schreiber 2002):

1. Compute local drawings for co-substances and enzymes of each reaction. For each reaction cluster the nodes for co-substances, enzymes and the reaction into a large node.
2. Insert layout constraints and temporary nodes for open and closed cycles. This step forces the subsequent steps to highlight the cyclic structure of specific pathways.
3. Reverse some edges to remove all cycles from the graph (restricted by *top-bottom* constraints).
4. Assign nodes to horizontal layers such that all edges are directed from top to bottom (the general drawing direction). This step computes the y-coordinates of nodes.
5. Compute a proper layering by inserting temporary nodes for edges and nodes involved in more than one layer.
6. Permute the order of nodes within each layer to reduce the number of edge crossings in the layered graph (under the restriction of *left-right* constraints).
7. Compute a balanced drawing without changing the pre-computed order in the layers. This step computes the x-coordinates of nodes.
8. Un-cluster large nodes and compute an edge routing by using the temporary nodes as additional base points for edges.

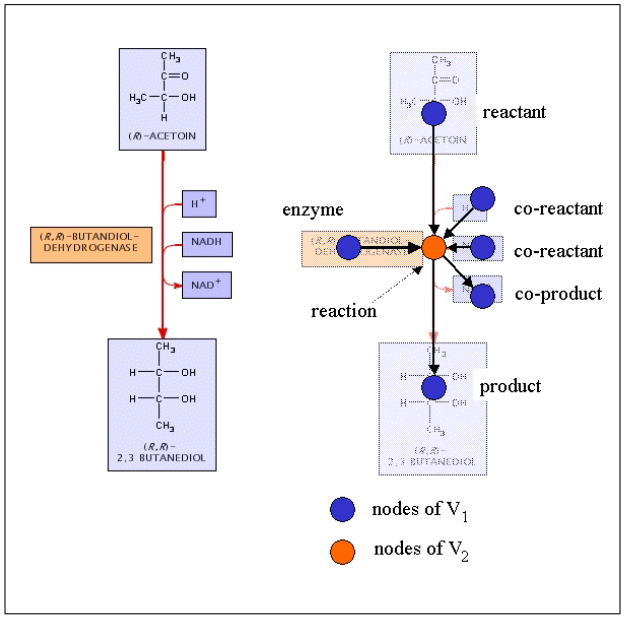

Figure 5: This image shows (left) a metabolic reaction and (right) its representation as a directed bipartite graph.

3 Visual comparison of pathways

In this section, we begin by describing the structure of the graph used for the modelling of metabolic pathways. Secondly, we consider a layout constraint used to compute better visual comparison of pathways. We describe the extension of the graph drawing algorithm to implement this constraint. This extended algorithm can be used to produce drawings for pathway data obtained from the BioPath system. Finally, we show how the algorithm can be adapted to compute visual comparisons of metabolic pathways with the drawing style used in KEGG.

3.1 Graph structure

Given data about similar pathways in different species we can represent these pathways as a directed bipartite graph $G = (V_1 \cup V_2, E)$, whereas nodes $v \in V_1$ represent substances, co-substances, and enzymes, and nodes $v \in V_2$ represent the reactions. Figure 5 shows a metabolic reaction and its representation as directed bipartite graph. Each pathway $pw_1, ..., pw_n$ of the species $1, ..., n$ results in a sub-graph $G_1, ..., G_n$ of graph G. Note that these sub-graphs are not connected, therefore graph G consists of at least n components. Furthermore, some sub-graphs or some parts of different sub-graphs may be identical. That is, the labels (names) of nodes representing substances are equal or nodes representing reactions in different sub-graphs are connected to equally typed and labelled nodes in each of these sub-graphs. These identical parts of sub-graphs represent identical substances and reactions in the pathways of different species.

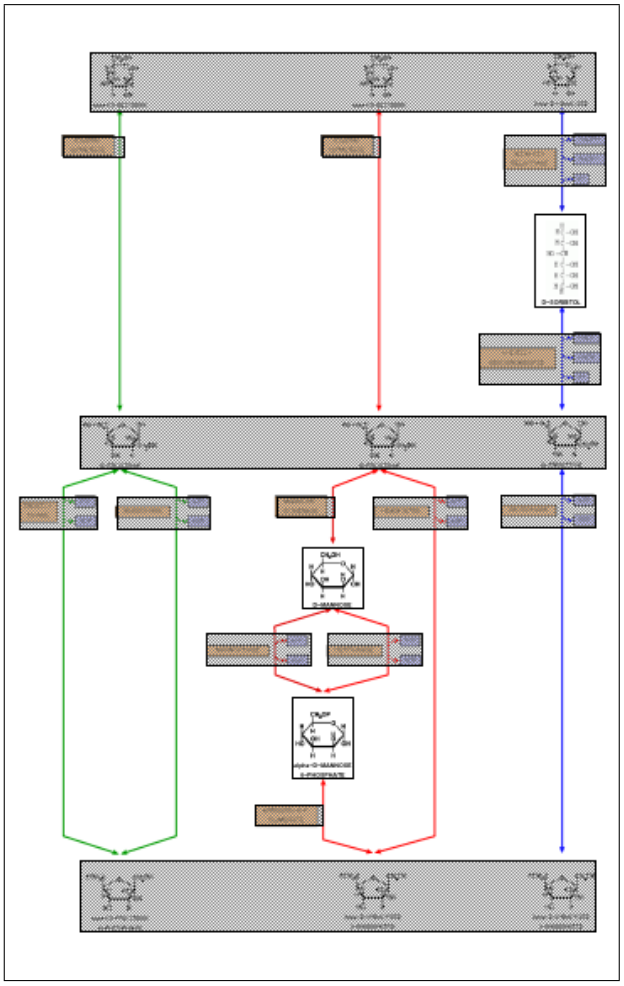

Figure 6: This diagram shows the temporary clustering of nodes. Step 1 of the graph drawing algorithm draws co-substances and enzymes of each reaction with a local drawing style and clusters these nodes into large nodes. (Note that these large nodes may consist only of one enzyme-node.) In step 2(a) nodes from different sub-graphs, which represent the same substance, are temporarily clustered into one node. For example, the three nodes on top of the image are clustered into one large node. All clustered nodes (after step 1 and after step 2(a)) are shown in grey.

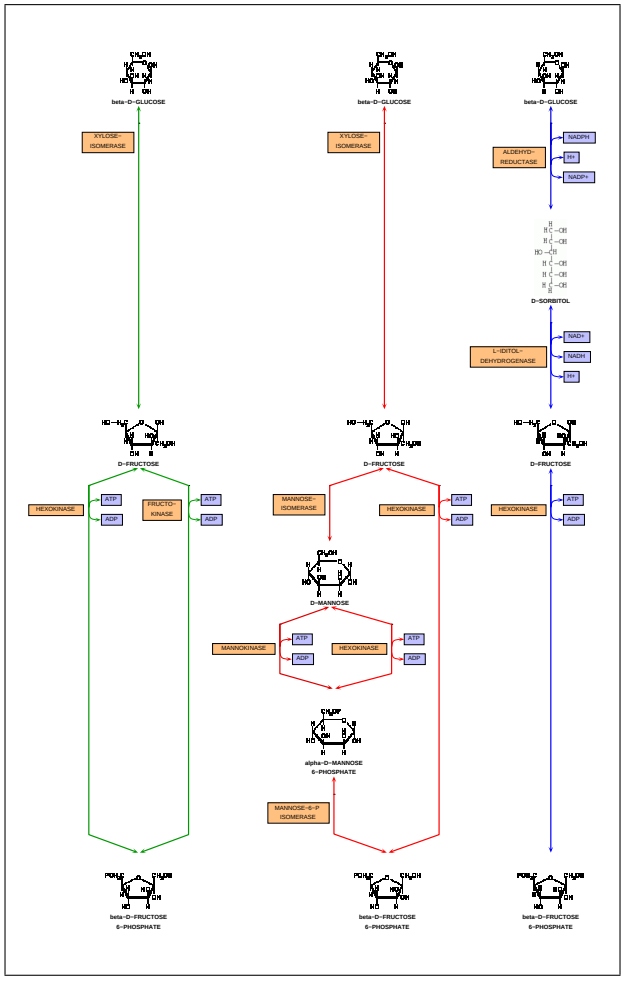

Figure 7: A visual comparison of the pathways shown in Figure 3 using the constraint graph drawing algorithm. The placement of equal substances on the same horizontal layer highlights similar and different parts in these pathways.

If we apply the graph drawing algorithm to the graph G we obtain a pathway diagram as in Figure 4. In this example the sub-graphs G_1, G_2, and G_3 represent similar pathways in higher plants (left), unicellular organisms (middle), and animals (right). These sub-graphs are not connected and parts of the different sub-graphs are identical. For example, there are nodes *β-D-glucose* (the top node of each pathway) and reactions *D-fructose* $\leftrightarrow$ *β-D-fructose-6-phosphate* catalyzed by the enzyme *hexokinase* in each of these sub-graphs. The pathways in Figure 4 are placed side-by-side, however, there is no direct visual comparison. The differences and common parts of these pathways are not highlighted.

3.2 Algorithm

How can we compute a better visual comparison of metabolic pathways? The idea is to place nodes of different sub-graphs on the same horizontal layer if these nodes represent the same substance. By placing equal substances on the same horizontal layer we also force identical reactions to appear in the drawing on a common layer. These reactions share equally labelled substance-nodes (which are placed on one layer) at the start of the reaction arrow, and the reaction itself is placed on the next possible layer.

Let us consider the pathway diagram in Figure 4 as an example. If we place all nodes representing the substance *β-D-glucose* on one horizontal layer, all nodes representing *D-fructose* on another layer, and all nodes for *β-D-fructose 6-phosphate* on yet another layer we obtain a drawing as shown in Figure 7. This pathway diagram highlights common parts such as the reaction *D-fructose* $\leftrightarrow$ *β-D-fructose-6-phosphate* catalyzed by the enzyme *hexokinase*, as well, as different parts such as two reaction steps between *β-D-glucose* and *D-fructose* in animals.

For a better visual comparison of metabolic pathways is it therefore sufficient to introduce *horizontal* constraints between nodes which represent the same substance in different sub-graphs.

To realize horizontal constraints we extend our algorithm by two additional steps: step 2(a) (between step 2 and step 3) and step 4(a) (between step 4 and step 5):

2.(a) Nodes from different sub-graphs, which represent the same substance, are temporarily clustered into one large node. See Figure 6 for an example.

4.(a) The large nodes obtained by step 2(a) are unclustered and all previously clustered nodes are

placed with the same y-coordinate as the corresponding large node.

In a hierarchical (top-bottom) drawing no edges are allowed between nodes in the same horizontal layer. Therefore the nodes within a cluster can be placed on the same layer only if they are not connected by an edge. To show that the extension of our algorithm is correct we have to prove that nodes in a cluster are not connected by an edge.

There are no edges between nodes within a cluster because:

1. Each sub-graph contains at most one node of a specific substance. This is guaranteed by the way BioPath computes a graph from the data in the database. Therefore step 2(a) builds clusters which contain at most one node from a specific sub-graph.

2. Nodes in different sub-graphs $G_1, ..., G_n$ are not connected by an edge.

Note that for left-right and top-bottom constraints the same conflict-solving strategies as in the original algorithm can be used because there are no such constraints between the nodes of a cluster.

3.3 KEGG-style diagrams

Pathway diagrams in KEGG (Kanehisa & Goto 2000) are different from the typical pictures in textbooks. These diagrams do not show co-substances and they place the enzymes on the reaction arrow instead of beside the reaction arrow as in typical drawings (see Figure 2). Furthermore, in KEGG only the names of substances are shown.

Our graph drawing algorithm can be easily adapted to compute this style of pathway diagrams. The general drawing remains the same, only the local drawings of enzymes and the information shown for substances are different. To achieve this drawing style it is necessary to change step 1 of the algorithm. This step computes local drawings for co-substances and enzymes of each reaction and clusters the nodes for co-substances, enzymes and the reaction into one large node. In the adapted version this step has to compute the local drawing only for the enzymes of a reaction. The large node than consists of a cluster of nodes for enzymes and the reaction.

Figure 8 shows the comparison of a similar pathway in *Homo sapiens* and *Saccharomyces cerevisiae* in the KEGG drawing style. The same substances are placed on a horizontal layer to allow the user an easy visual comparison of these pathways. Note that the modelling step has to guarantee that each sub-graph contains at most one node of a particular substance.

4 Discussion

Biologists frequently use visual representations of metabolic pathways. The visual comparison of pathways across different species can offer new insight into the different metabolic processes. However, visual interfaces to databases do not support such tasks very well.

We have presented a new approach for the visual comparison of pathways using a constraint graph drawing algorithm. Using horizontal constraints, the same parts of similar pathways in different species are placed on the same horizontal layer, therefore highlighting similarities and differences between these pathways. This visualization method has been tested with data obtained from the BioPath system and the KEGG database. The presented approach supports

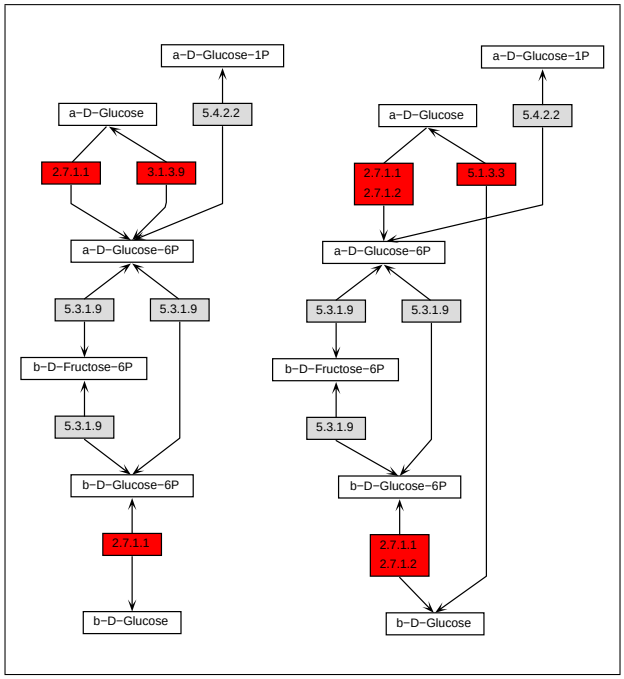

Figure 8: This pathway diagram shows a comparison of pathways in (left) *Homo sapiens* and (right) *Saccharomyces cerevisiae* in the KEGG drawing style.

different drawing styles and offers a user friendly way to compare metabolic pathways in different species visually.

Several extensions to our approach are possible. Colours can be used to further highlight differences between pathways as shown in Figure 8. In this pathway diagram a different colour is used to emphasize different substances and enzymes. In the current implementation the left-right order of nodes within a horizontal layer may be different in the sub-graphs. This can decrease their visual comparability. Left-right constraints can be useful to force the same left-right order of nodes in different sub-graphs.

This approach may also be applicable to other types of pathway diagrams such as visualizations of signal transduction pathways.

References

Becker, M. Y. & Rojas, I. (2001), 'A graph layout algorithm for drawing metabolic pathways', *Bioinformatics* **17**(5), 461–467.

Brandenburg, F. J., Gruber, B., Himsolt, M. & Schreiber, F. (1998), Automatische Visualisierung biochemischer Information, *in* R. Hofestädt, ed., 'Proceedings of the Workshop Molekulare Bioinformatik, GI Jahrestagung', Shaker Verlag, pp. 24–38.

Forster, M., Pick, A., Raitner, M., Schreiber, F. & Brandenburg, F. J. (2002), 'The System Architecture of the BioPath system', *In Silico Biology, to appear* .

Hofestädt, R. & Thelen, S. (1998), 'Qualitative Modeling of Biochemical Networks', *In Silico Biology* **1**, 39–53.

Kanehisa, M. & Goto, S. (2000), 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acid Research* **28**(1), 27–30.

Karp, P. D. & Paley, S. M. (1994), Automated Drawing of Metabolic Pathways, *in* H. Lim, C. Cantor & R. Bobbins, eds, 'Proceedings of the 3rd International Conference on Bioinformatics and Genome Research', pp. 225–238.

Mendes, P. (2000), Advanced Visualization of Metabolic Pathways in PathDB, *in* 'Proceedings of the 8th Conference on Plant and Animal Genome'.

Michal, G. (1993), *Biochemical Pathways (Poster)*, Boehringer Mannheim, Penzberg.

Schreiber, F. (2002), 'High Quality Visualization of Biochemical Pathways in BioPath', *In Silico Biology* **2**(2), 59–73.