First publ. in: Proceedings of the 11th Symposium on Graph Drawing (GD ´03) (LNCS 2912), 2004, pp. 111-122

Visualizing Related Metabolic Pathways in Two and a Half Dimensions
(Long Paper)

Ulrik Brandes[1,⋆], Tim Dwyer[2], and Falk Schreiber[3,⋆⋆]

[1] Department of Mathematics & Computer Science, University of Passau, Germany. brandes@algo.fmi.uni-passau.de
[2] School of Information Technologies, University of Sydney, Australia. dwyer@it.usyd.edu.au
[3] Bioinformatics Center (BIC-GH), Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany. schreibe@ipk-gatersleben.de

Abstract. We propose a method for visualizing a set of related metabolic pathways using $2\frac{1}{2}$D graph drawing. Interdependent, two-dimensional layouts of each pathway are stacked on top of each other so that biologists get a full picture of subtle and significant differences among the pathways. Layouts are determined by a global layout of the union of all pathway-representing graphs using a variant of the proven Sugiyama approach for layered graph drawing that allows edges to cross if they appear in different graphs.

1 Introduction

Metabolic pathways are subnetworks of the complete network of metabolic reactions. They differ across organisms because different species may for example have developed different ways to synthesize a specific substance. We are interested in visualizing several related metabolic pathways in such a way that the inherent differences can be explored by trained biologists in order to understand the evolutionary relationships among species. The structural characteristics of metabolic pathways make them particularly amenable to layered graph drawing methods [2,24,27]. We therefore propose to visualize sets of related pathways in $2\frac{1}{2}$ dimensions, i.e. to produce interdependent, two-dimensional, layered layouts for all pathways and stack them on top of each other so that the most similar pathways are adjacent.

To realize such a design, two graph drawing issues need to be addressed. We have to determine a suitable ordering to reduce the variation between consecutive pathways and we have to deal with dependencies introduced by the many substances and reactions present in more than one pathway. An interesting consequence is a new type of crossing number in which the weight of a crossing may be different for each pair of edges.

⋆ Partially supported by DFG under grant BR 2158/1-1.
⋆⋆ Supported by BMBF under grant 0312706A.

This paper is organized as follows. In Sect. 2, we give some background on the type of networks considered and define the graph model on which we operate. After briefly reviewing related approaches to visualizing similar graphs, we specify our visualization design in Sect. 3. In Sect. 4 we address the ordering problem while in Sect. 5 we discuss a method to determine a global layout of the stacked pathways considering that edge crossings are less severe if the edges involved do not coexist in the same pathway. To demonstrate the utility of our method it is applied to typical real-world data in Sect. 6.

2 Metabolic Pathways

Metabolic reactions are transformations of chemical substances which occur in living beings and are usually catalyzed by enzymes. A reaction changes certain substances (reactants) into different ones (products). Metabolic reactions form large and complex networks as for example shown on the well-known *Biochemical Pathways Poster* part 1 [21]. A *metabolic pathway* is a subnetwork of the complete network of metabolic reactions (see Fig. 1). Such a pathway can be given by biochemical textbooks and databases such as KEGG [17] or be defined by functional boundaries such as the network between an initial and a final substance.

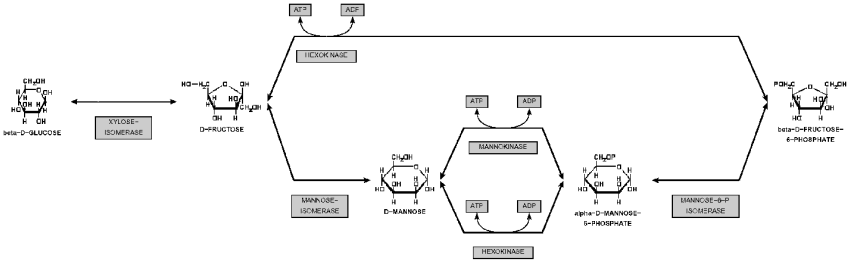


Fig. 1. A metabolic pathway

Metabolic pathways differ across organisms. Studies suggest significant variations even in most central pathways such as glycolysis [6]. Comparative analysis of pathways across species has several applications:

- understanding the evolutional relationships between species.
- development of species-specific drug targets (e.g. antibiotics).
- identification of previously unknown parts of pathways in a species.

From a formal point of view a metabolic pathway is a directed hyper-graph. The vertices represent the substances within a pathway, the hyper-edges represent reactions. A hyper-edge representing a reaction connects substances and

is labeled with the enzyme(s) that catalyze the reaction. As hyper-graphs are not commonly used for simulation and visualization purposes, metabolic pathways are often modeled by bipartite graphs, e.g. Petri-net representations of pathways [23]. Here the reactions themselves are vertices, and edges are binary relations connecting substances with reaction vertices.

Three structural properties of metabolic pathways are particularly relevant for visualization. They typically are:

1. small in size (often less than two dozen reactions)
2. sparse (because most substances are involved in few reactions)
3. directed and acyclic (due to a dominant direction of most reactions and a small number of cyclic pathways such as the citrate cycle)

Several similarity measures have been introduced to compare pathways. They are characterized by the combination of structural information about metabolic networks with additional data such as sequence information [11], the enzyme classification hierarchy [29], or information about the hierarchical clustering of reactions into pathways [20]. All these similarity measures require additional information to the metabolic network, for example the genome sequence of the organisms.

We suggest a more general measure for reaction or pathway similarity which only depends on the structure of the graph, i.e. on the presence or absence of vertices and edges. This facilitates comparison of metabolic pathways from different sources (databases, experiments) even if no additional data is available or the pathway boundaries are user-defined. Note that in general our visualization method is independent of the similarity measure which is only used to compute the order of the stacking.

2.1 Formalization

For the purpose of this paper the distinction between vertices representing reactions and substances is not important. A metabolic pathway is therefore modeled by a directed graph $G = (V, E)$, where a vertex $v \in V$ represents either a substance or a reaction, and an edge $(v, w) \in E$ indicates that substance v enters reaction w or that reaction v produces substance w.

Our goal is to visualize a set $\{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_r(V_r, E_r)\}$ of graphs that represent related pathways. Though the dissimilarity of two pathways $G_i = (V_i, E_i)$, $G_j = (V_j, E_j)$ can be measured in various ways, the arguments outlined above suggest that the following measure is appropriate.

Consider the *union graph* $G = (V, E)$, where $V = \bigcup_{i=1}^{r} V_i$ and $E = \bigcup_{i=1}^{r} E_i$, and a set of relevant elements $P \subseteq (V \cup E)$. The *Hamming distance* of two graphs G_i, G_j, $1 \leq i, j \leq r$, is defined as the cardinality of the symmetric difference of relevant elements present in either graph, i.e.

$$\delta_P(G_i, G_j) = |((V_i \cup E_i) \triangle (V_j \cup E_j)) \cap P)| \ .$$

Thus, dissimilarity can be defined in terms of missing edges, vertices, or both, by choosing P accordingly.

3 Visualizing Similar Networks

There are two common approaches to compare pathways in different species visually. Either the combination of all pathways into one diagram or the production of a drawing for each species. The first method is used in many textbooks, on a famous poster [21] and in systems such as BioMiner [27] and BioPath [12]. In general, the drawings contain either multiple (parallel) reaction-edges or single ones which are color-coded depending on the occurrence of a reaction in a set of species. An example for the second approach is the visual interface of the KEGG database [17] where all enzymes found in the gene catalog of a specific species are marked in the reference pathway map in order to identify the species-specific pathways. To compare pathways in r species, r diagrams are needed. A dynamic visual comparison method producing a diagram for each species is presented in [25].

These solutions are restricted to the comparison of only a few pathways, because either (in the first approach) the readability of the diagram decreases or (in the second approach) the size of the picture increases dramatically with each new pathway. Furthermore, none of the above mentioned methods deal with the problem of computing an appropriate order of the pathways.

3.1 Drawing Graphs in $2\frac{1}{2}$D

To allow users to compare several metabolic pathways, we want to draw them in different parallel planes, but with interdependent layouts. We call this type of representation a $2\frac{1}{2}$D drawing because the third dimension is used in a way fundamentally different from the other two. Note that, traditionally, the axes are interchangeable in 3D graph drawing.

The idea of treating the third dimension as an independent channel conveying a different kind of information [30] has been applied in numerous settings dealing with different types of network data. For instances, the data can be a single graph with vertex attributes that determine the third dimension [19,5], a graph together with a hierarchical clustering [9] or graphs that evolve over time [4,7]. The latter example relates $2\frac{1}{2}$D graph drawing to dynamic graph drawing and graph animation, since the difference in layout between consecutive states of the graph should be small. Another interesting example is the dependence graph of spreadsheet cells [26], which has a natural 2D layout that can be spread in three dimensions to indicate the data flow.

The case considered in this paper is somewhat different from those above. If we consider our pathways as subgraphs of the union graph, the data are characterized by the fact that

- there is a set of subgraphs with no given ordering,
- vertices and edges are likely to appear in more than one subgraph, and
- there are no edges to be drawn between different subgraphs.

Moreover, we aim at a representation that consists of a stacked set of two-dimensional (layered) drawings for each subgraph where a vertex has the same

2D-coordinate in each drawing in which it appears (so that it can be represented by a straight column).

Our approach can therefore be viewed as a generalization of both the Sugiyama method [28] and parallel coordinates [15].

4 Stacking Order

Let $\{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_r(V_r, E_r)\}$ be a set of graphs, and $G = (V, E)$ their union graph. We want to order these graphs so that those which are similar with respect to Hamming distance are close to each other. Variations of this problem arise in many applications, and two of them are especially relevant in our context. Let $P \subseteq (V \cup E)$ be a set of relevant elements.

Problem 1 (MIN SUM ORDERING). Find a permutation $\sigma = (\sigma_1, \ldots, \sigma_r)$ such that

$$\sum_{i=1}^{r-1} \delta_P(G_{\sigma_i}, G_{\sigma_{i+1}})$$

is minimum.

In the context of data transmission, this problem is also known as DOP (data ordering problem) [22]. A less restricted version (with an arbitrary distance matrix) is considered in [18], where the goal is to order parallel coordinates.

Theorem 1. *MIN SUM ORDERING is $\mathcal{NP}$-hard.*

Proof. Straightforward reduction from HAMMING DISTANCE TSP [1].

For a permutation $\sigma = (\sigma_1, \ldots, \sigma_r)$ we define the *lifetime* of an element $p \in P$ to be $\Lambda_\sigma(p) = \{1 \leq i \leq r \,:\, p \in G_{\sigma_i}\}$. An element $p \in P$ is called *persistent*, if its lifetime spans the entire interval $\{1, \ldots, r\}$, and *transient* otherwise. The number of *appearances* and *disappearances* of $p \in P$ are defined by $a_\sigma(p) = |\{1 < i \leq r \,:\, p \in G_{\sigma_i} \setminus G_{\sigma_{i-1}}\}|$ and $d_\sigma(p) = |\{1 \leq i < r \,:\, p \in G_{\sigma_i} \setminus G_{\sigma_{i+1}}\}|$. Note that persistent elements make no appearances or disappearances.

Corollary 1. *MIN SUM ORDERING is equivalent to minimizing*

$$\sum_{p \in P} \big(a_\sigma(p) + d_\sigma(p)\big) \ .$$

A related alternative objective is therefore to minimize the maximum number of times an element appears or disappears in an ordering.

Problem 2 (MIN INTERVAL ORDERING). Find a permutation $\sigma = (\sigma_1, \ldots, \sigma_r)$ such that

$$\max_{p \in P} \{a_\sigma(p), d_\sigma(p)\}$$

is minimum.

This problem is a generalization of the consecutive ones property, since, if $p \in P$ is transient, $\max\{a_\sigma(p), d_\sigma(p)\}$ is the number of lifetime intervals (and zero otherwise). We have the following complexity status.

Theorem 2. *MIN INTERVAL ORDERING is $\mathcal{NP}$-hard, but it can be determined in linear time whether there is an ordering such that each relevant element appears or disappears at most once.*

Proof. See [14]. Since the restricted problem corresponds exactly to the consecutive ones property, it is linear-time solvable using PQ-trees [3].

Since both problems arise in many contexts, a variety of algorithms is available to determine an ordering. For MIN SUM ORDERING, for instance, heuristics for the TSP are easily adapted to yield good orderings. Other alternatives include a simple greedy heuristic that sucessively inserts a new element where it causes the smallest increase of the objective (this method is claimed to perform well for instances in data transmission [22]) and one-dimensional projections of the distance matrix obtained by principal component analysis.

5 Global Layout

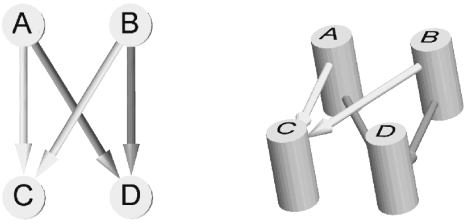
To maximize the similarity between visualizations of related graphs, we compute a layout only for the union graph. The drawing of vertices and edges thus remains unchanged throughout their lifetime. Since graphs representing metabolic pathways tend to be sparse and acyclic, the Sugiyama framework for layered graph layout [10,28] is widely used to visualize individual pathways [2,24,27].

While straightforward application of any of these methods to the union graph is possible, better results can be obtained by taking into account that the individual graphs are to be viewed in separate, parallel planes. To avoid confusion with layers in the Sugiyama approach, we refer to these as *strata*.

5.1 A New Crossing Minimization Problem

Our main modification with respect to standard variants of the Sugiyama framework is motivated by the observation that crossings of edges not present in the same stratum can be resolved by 3D rotation or stereo depth perception as shown in Fig. 2. Hence, when computing the global layout, we need not account for such crossings the same way as crossings between edges with overlapping lifetime intervals.

A common approach for crossing reduction in layered layouts is based on a layer-by-layer sweep, in which the ordering of vertices in a layer L_0 is fixed and vertices in an adjacent layer L_1 are permuted to reduce the number of crossings. Exact and heuristic methods for this one-sided crossing minimization problem are based on the *crossing matrix* $(c_{uv})_{u,v \in L_1}$, in which an entry c_{uv} corresponds to the number of crossings between the two layers caused by edges incident to u and v, if u is placed to the left of v.



(a) top view (b) rotated view

Fig. 2. Crossings between edges in different strata can be resolved by 3D rotation.

We define a strata-aware crossing matrix by multiplying the contribution of a pair of edges to an entry c_{uv} with the number of times that these edges are present in a common stratum. Note that this results in a weighted crossing matrix in which crossings are counted individually for each pair of edges, and that this weighting scheme is different from, say, assigning weights to edges and multiplying these if edges cross.

Obviously, our weighted crossing minimization variant is at least as difficult as standard crossing minimization.

5.2 Implementation

Since most crossing reduction methods based on the crossing matrix are oblivious to the definition of its entries, they can be applied in our case as well.

We adapted the open-source program dot[1] which uses a median heuristic coupled with an adjacent-exchange post-processing step [13]. New permutations generated by these heuristics are rejected if they lead to an increase in edge crossings according to our modified crossing matrix.

Since this method does not consider crossings until after a permutation is generated it was felt that it might not be readily compatible with the new definition of c_{uv}. As an alternative we also implemented the Integer Linear Programming (ILP) approach suggested in [16]. The ILP method directly uses $c_{uv} - c_{vu}$ as the coefficients of the variables of the cost function to find an exact solution.

In our experiments we found that the median heuristic outperformed the ILP approach, because it tends to achieve a reasonable global solution which the subsequent adjacent-exchange is able to improve according to our modified definition of c_{uv}. Moreover, it is significantly faster than the *branch and cut* algorithm for solving the ILP.

For $2\frac{1}{2}$D graph drawing, the horizontal coordinate assignment phase should be adapted as well. When routing the edges in $2\frac{1}{2}$D we can allow dummy vertices on different strata to overlap. In dot's implementation, an auxiliary graph is created in which edges of an arbitrary minimum length are inserted between

[1] Available at http://www.graphviz.org/.

adjacent vertices (and dummy vertices) in each layer to keep them separated. We therefore set the length of auxiliary edges between adjacent dummy vertices that do not coexist on the same stratum to zero, thus allowing such vertices, and hence edges, to overlap. This modification led to a significant improvement in aspect ratio of the final layout (approx. 40% in our densest test cases).

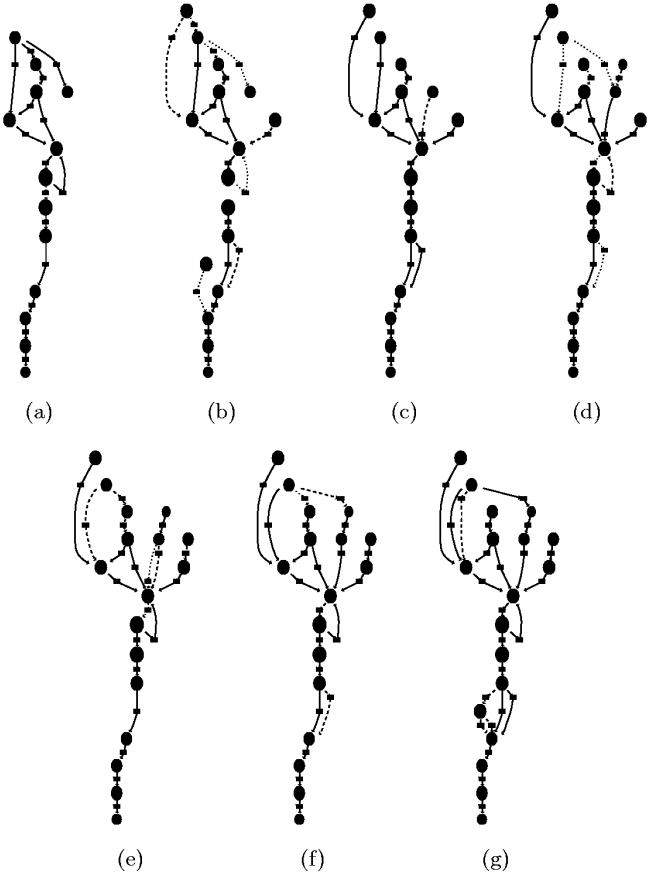


Fig. 3. Layouts of individual pathways obtained from a union graph layout. Appearing edges are shown dashed, disappearing edges dotted.

6 Application Example

The utility of our approach is demonstrated on pathways extracted from the KEGG database [17]. The data consists of parts of the glycolysis and fructose/mannose metabolism pathways in seven organisms that show significant

Table 1. Hamming-distance matrix for seven pathways

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	
(a)	0	21	23	21	43	40	56	*Haemophilus influenzae*
(b)	21	0	22	20	48	39	53	*Escherichia coli CFT073*
(c)	23	22	0	10	38	35	45	*Streptococcus pyogenes*
(d)	21	20	10	0	34	31	41	*Bacillus subtilis*
(e)	43	48	38	34	0	15	31	*Arabidopsis thaliana*
(f)	40	39	35	31	15	0	16	*Drosophila melanogaster*
(g)	56	53	45	41	31	16	0	*Homo sapiens*

differences. Table 1 gives Hamming distances between these pathways with all elements $P = (V \cup E)$ considered relevant. The order in which the organisms are listed is optimal with respect to MIN SUM ORDERING and was computed by enumeration.

Using our adapted version of the dot program described in the previous section, a layout of the union graph of these seven pathways was computed. The resulting individual layouts are shown in Fig. 3.

The $2\frac{1}{2}$D representations shown in Figs. 4 and 5 have been created with the WilmaScope 3D graph visualization system [8]. Edge appearances and disappearances are color-highlighted using green and red. By moving a semi-transparent plane through the image, users can navigate forward and backward in the similarity-ordered sequence of pathways.

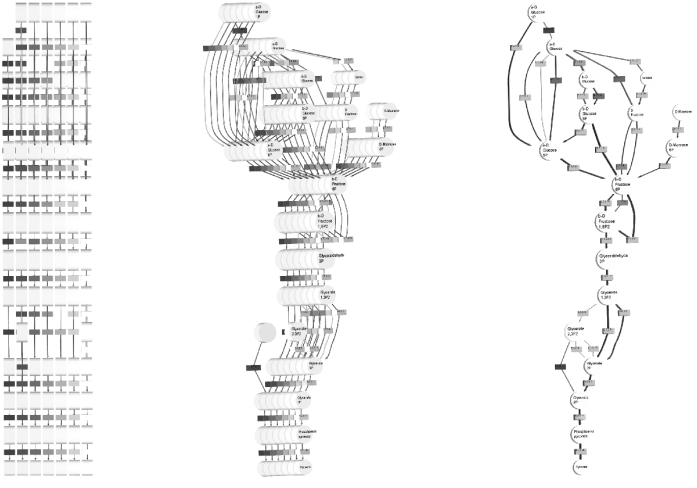
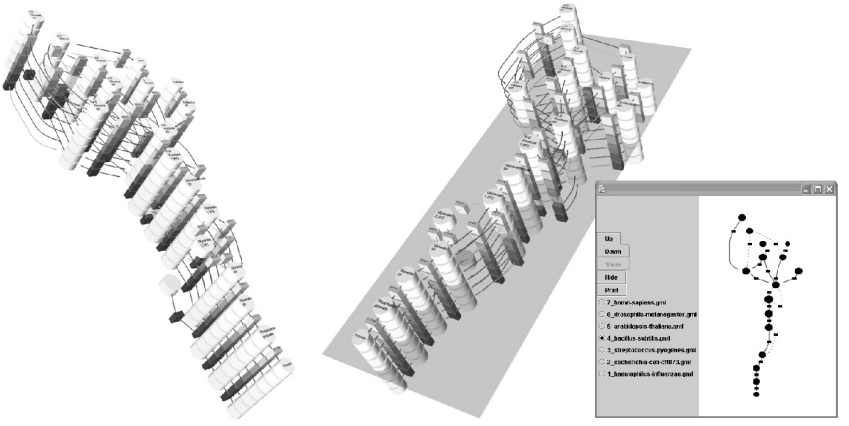


Fig. 4. $2\frac{1}{2}$D drawing of seven related pathways (parallel projection)



(a) A perspective projection (b) A screenshot showing the cross-section viewer for interactive exploration

Fig. 5. Perspective projection with cross-section viewer for interactive exploration

7 Discussion

We have presented an approach to visualize sets of related metabolic pathways. It should be noted that the usefulness of the stacking order is dependent on the quality of the data, which is frequently poor. However, the biologically meaningless order thus produced may in turn help discover data inconsistencies.

Straightforward extensions of our approach include varying distances between strata to indicate the actual dissimilarity and varying vertex thickness to represent numerical attributes like volume. A more challenging task is the computation of interdependent layouts for navigation along a phylogenetic tree rather than our one-dimensional ordering of pathways.

References

1. F. Alizadeh, R. M. Karp, D. K. Weisser, and G. Zweig. Physical mapping of chromosomes using unique probes. *Proceedings of the 5th ACM-SIAM Symposium on Discrete Algorithms (SODA '94)*, pages 489–500, 1994.
2. M. Y. Becker and I. Rojas. A graph layout algorithm for drawing metabolic pathways. *Bioinformatics*, 17(5):461–467, 2001.
3. K. S. Booth and G. S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *Journal of Computer and System Sciences*, 13(3):335–379, 1976.
4. U. Brandes and S. R. Corman. Visual unrolling of network evolution and the analysis of dynamic discourse. *Proceedings of the IEEE Symposium on Information Visualization 2002 (InfoVis '02)*, pages 145–151, 2002.
5. U. Brandes and T. Willhalm. Visualization of bibliographic networks with a reshaped landscape metaphor. *Proceedings of the 4th Joint Eurographics and IEEE TCVG Symposium on Visualization (VisSym '02)*, pages 159–164. ACM, 2002.

6. T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochemical Journal*, 343:115–124, 1999.
7. T. Dwyer and P. Eades. Visualising a fund manager flow graph with columns and worms. *Proceedings of the 6th International Conference on Information Visualisation (IV '02)*, pages 147–152. IEEE Computer Society Press, 2002.
8. T. Dwyer and P. Eckersley. The WilmaScope 3D graph drawing system. In P. Mutzel and M. Jünger, editors, *Graph Drawing Software*, Mathematics and Visualization. Springer, 2003.
9. P. Eades and Q. Feng. Multilevel visualization of clustered graphs. *Proceedings of the 4th International Symposium on Graph Drawing (GD '96)*, volume 1190 of *Lecture Notes in Computer Science*, pages 113–128. Springer, 1996.
10. P. Eades and K. Sugiyama. How to draw a directed graph. *Journal of Information Processing*, 13:424–437, 1990.
11. C. V. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *Journal Molecular Evolution*, 52:471–489, 2001.
12. M. Forster, A. Pick, M. Raitner, F. Schreiber, and F. J. Brandenburg. The system architecture of the BioPath system. *In Silico Biology*, 2(3):415–426, 2002.
13. E. R. Gansner, E. Koutsofios, S. C. North, and K.-P. Vo. A technique for drawing directed graphs. *Software Engineering*, 19(3):214–230, 1993.
14. P. W. Goldberg, M. C. Golumbic, H. Kaplan and R. Shamir. Four strikes against physical mapping of DNA. *Journal of Computational Biology*, 2(1):139–152, 1995.
15. A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. *Proceedings of the 1st IEEE Conference on Visualization (Vis '90)*, pages 361–378, 1990.
16. M. Jünger, E. K. Lee, P. Mutzel, and T. Odenthal. A polyhedral approach to the multi-layer crossing minimization problem. *Proceedings of the International Symposium on Graph Drawing*, Lecture Notes in Computer Science 1353, pages 13–24. Springer, 1997.
17. M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acid Research*, 28(1):27–30, 2000.
18. D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.
19. H. Koike. The role of another spatial dimension in software visualization. *ACM Transactions on Information Systems*, 11(3):266–286, 1993.
20. L. Liao, S. Kim, and J.-F. Tomb. Genome comparisons based on profiles of metabolic pathways. *Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES '02)*, pages 469–476, 2002.
21. G. Michal. *Biochemical Pathways (Poster)*. Boehringer Mannheim, Penzberg, 1993.
22. R. Murgai, M. Fujita, and S. C. Krishnan. Data sequencing for minimum-transition transmission. *Proceedings of the 9th IFIP International Conference on Very Large Scale Integration (VLSI '97)*, 1997.
23. V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman. Petri net representations of metabolic pathways. *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology (ISMB '93)*, pages 328–336, 1993.
24. F. Schreiber. High quality visualization of biochemical pathways in BioPath. *In Silico Biology*, 2(2):59–73, 2002.

25. F. Schreiber. Visual Comparison of Metabolic Pathways. *Journal of Visual Languages and Computing*, 14(4):327-340, 2003.
26. H. Shiozawa, K. Okada, and Y. Matsushita. 3D interactive visualization for inter-cell dependencies of spreadsheets. *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '99)*, pages 79–83, 1999.
27. M. Sirava, T. Schäfer, M. Eiglsperger, M. Kaufmann, O. Kohlbacher, E. Bornberg-Bauer, and H.-P. Lenhof. BioMiner – modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*, 18(Suppl. 2):219–230, 2002.
28. K. Sugiyama, S. Tagawa, and M. Toda. Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man and Cybernetics*, 11(2):109–125, 1981.
29. Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00)*, pages 376–383, 2000.
30. J. Wen. Exploiting orthogonality in three dimensional graphics for visualizing abstract data. Technical Report CS-95-20, Department of Computer Science, Brown University, 1995. http://www.cs.brown.edu/publications/techreports/reports/CS-95-20.html.