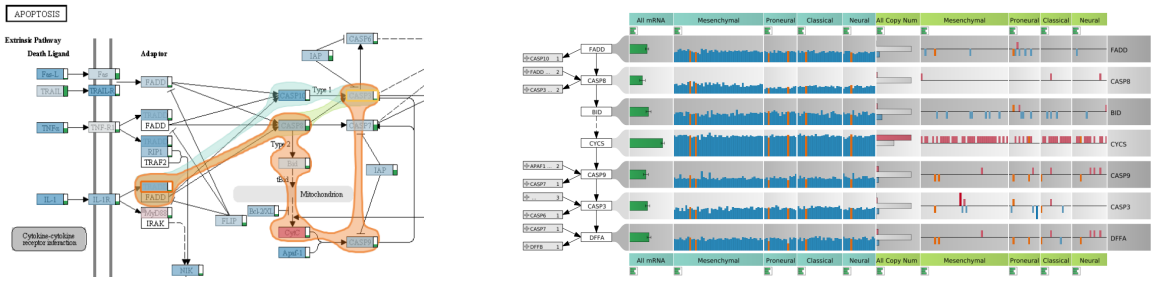
enRoute: Dynamic Path Extraction from Biological Pathway Maps for In-Depth Experimental Data Analysis

Christian Partl*
Graz University of Technology
Alexander Lex*
Graz University of Technology
Marc Streit†
Johannes Kepler University Linz
Denis Kalkofen*
Graz University of Technology
Karl Kashofer‡
Medical University of Graz
Dieter Schmalstieg*
Graz University of Technology



(a) (b)

Figure 1: **enRoute** A path highlighted in orange in the pathway map in (a) is extracted and shown next to associated experimental data in (b).

Abstract

Pathway maps are an important source of information when analyzing functional implications of experimental data on biological processes. Associating large quantities of data with nodes on a pathway map and allowing in depth-analysis at the same time, however, is a challenging task. While a wide variety of approaches for doing so exist, they either do not scale beyond a few experiments or fail to represent the pathway appropriately. To remedy this, we introduce enRoute, a new approach for interactively exploring experimental data along paths that are dynamically extracted from pathways. By showing an extracted path side-by-side with experimental data, enRoute can present large amounts of data for every pathway node. It can visualize hundreds of samples, dozens of experimental conditions, and even multiple datasets capturing different aspects of a node at the same time. Another important property of this approach is its conceptual compatibility with arbitrary forms of pathways. Most notably, enRoute works well with pathways that are manually created, as they are available in large, public pathway databases. We demonstrate enRoute with pathways from the well-established KEGG database and expression as well as copy number datasets from humans and mice with more than 1,000 experiments at the same time. We validate enRoute in case studies with domain experts, who used enRoute to explore data for glioblastoma multiforme in humans and a model of steatohepatitis in mice.

Keywords: Pathway analysis, path extraction, cellular networks, biological processes, expression data, copy number variation.

Index Terms: H.5.m [Information Systems]: Information Interfaces and Presentation—Miscellaneous; I.3.8 [Computing Methodologies]: Computer Graphics—Applications

*e-mail:{lastname}@icg.tugraz.at
†e-mail:marc.streit@jku.at
‡e-mail:karl.kashofer@medunigraz.at

1 Introduction

Pathway maps are an important tool for studying biomolecular processes. There is a wide variety of pathway maps available that model interactions of proteins, chemical reactions and their catalyzing enzymes, as well as cellular signaling processes. Pathways are small to medium-sized graphs representing consensus knowledge that is backed up by literature and typically either describe a process in healthy organisms or a specific disease. Consequently, pathways are often activated or inactivated in particular conditions, as diseases or other influences change the processes within the cells.

A common approach to study specific influences on cellular processes is to concurrently analyze experimental data for one or multiple conditions. A condition in this sense describes a group of measurements that are semantically homogeneous. Examples for this homogeneity criterion are samples taken from one species in a multi-species analysis, or samples of patients belonging to a subtype of a form of cancer. Looking at pathways in the context of experimental data can tell analysts that, for example, branches of a graph are inactive for a given condition. Examples of such observations are omnipresent in the literature. For instance, the gene *PTEN* is known to regulate a signaling pathway relevant to the regulation of cell-growth (phosphoinositide 3-kinase signaling pathway) [4]. If *PTEN* is mutated, however, the pathway becomes deactivated, leading to unchecked cell division and tumor growth. In other well-studied examples, pathways such as the *p53 signaling pathway* or the *tyrosine kinase* pathway are affected by mutations and de-regulations of *PIK3R1*, *NF1*, and *ERBB2*, which play a role in subtypes of glioblastoma multiforme (GBM), a brain cancer [29]. In previous work, we described a case where different subtypes of *GBM* have highly differential behavior in mRNA expression patterns in the *Glioma* pathway [16]. It is also well documented that these molecular subtypes and their effect on biochemical processes captured in pathways have serious implications for prognosis, treatment, and patient well-being (e.g., [29, 21]). Our goal with the development of enRoute is to visually support the analysis processes leading to such insights.

Due to the relevance of the problem, it is not surprising that a

multitude of approaches, techniques, and implementations for augmenting pathways with experimental data exist. However, because of the complex nature of the biological processes involved, this task is not trivial. In this paper, we present enRoute, a novel approach to explore experimental data in the context of pathways. We augment the original pathway map to hint at parts with interesting underlying experimental data, and let analysts select paths for detailed analysis. A selected path is highlighted in the pathway map and extracted. The extracted path is displayed in a separate view in linear form side-by-side with visualizations of experimental data. Using our method, we can solve many critical problems in visualizing experimental data for pathways, such as showing datasets of large scale, displaying multiple different datasets, and resolving multi-mapping issues common in pathway representations.

enRoute is integrated in a suite of visualization techniques available in the Caleydo visualization framework[1]. It enables detailed analysis of pathways that were identified using other methods (e.g., [16]). However, enRoute can also be used as a stand-alone tool, provided the analyst is aware of pathways of relevance for the studied conditions. While we focus on how enRoute can be used to explore experimental conditions for nodes in pathways, the underlying concept of extracting paths can be equally applied to the general case of graphs with many node attributes.

2 Biological Background

Biological pathways describe molecular processes in living cells, explaining the complex relationships between biomolecules in the intracellular compartments and membranes in a time and treatment dependent way. The functional biomolecules are mostly proteins and metabolites, however, other biomolecules, like nucleic acids, lipids and ions also have regulatory effects. Last but not least, pathways also consider the spatial separation of compartments within cells with specialized pathways for membrane bound signaling cascades or in mitochondria.

Biological pathways can be categorized into groups. Metabolic pathways describe buildup and degradation of chemical substances; an example is the *TCA cycle* pathway. Another important group of pathways describes intracellular signaling cascades, like the *mTOR signaling pathway*, which regulates apoptosis. Disease related pathways focus on the molecular aspects of a disease, like the *diabetes mellitus type 1*, or the well known *Pathways in Cancer* pathway. One also has to keep in mind that each pathway is individually designed to highlight the functional aspects in the center of the researchers' interest.

In common biological pathway maps, nodes represent functional entities, like proteins, metabolites or chemical compounds. As proteins are generated from mRNA, which is itself transcribed from the genes encoded in DNA, a protein node is often identical with a gene node. Functional nodes that catalyze reactions often have extensive multi-mapping of many genes onto the same node, due to the ability of many proteins to catalyze this particular reaction. As in chemistry, one enzyme can often be substituted for another one, albeit with different efficiency.

Links between nodes describe biological reactions, like protein modifications, biochemical reactions or changes in gene expression. Pathways are usually drawn as maps, where the structure of the graph is depicted in node-link diagrams following specific drawing conventions. Manually curated pathways try to simplify the complex network of interactions to make the presentation visually appealing, while maintaining the scientific content. Pathway maps are available from a wide variety of databases, where KEGG [11] is a prominent example. *Protein-protein interaction networks* are a more general form of a biological network that captures the reactions and interactions between proteins without the necessity of a functional context.

[1] http://www.caleydo.org/

In recent years, profound advances in molecular biology have made a multitude of large datasets available for pathway modeling. High throughput mass-spec analysis of proteins, next-generation sequencing of nucleic acids, and NMR for metabolites generate vast amounts of data that have to be put into the right biological context to be useful. Data generated by these new techniques fall into two categories, one is a direct measurement of protein amounts, gene expression levels, or metabolite concentrations. The second category of data, like copy number variation of genes, mutations in genes, or methylation patterns, is not directly represented in pathway maps, but has a profound influence on the biological system represented in the pathway. Data of the second category is often used to reason about the causes of differences in the expression levels of genes or the concentrations of metabolites in different experimental conditions. It is this scientific comparing, reasoning, and explaining that we want to support with the tools presented in this paper.

3 Requirement Analysis

In collaboration with our partners from the Medical University of Graz, we have elicited the challenges experts face when analyzing pathways and associated experimental data. In the following, we discuss these challenges and, later on, evaluate the related work as well as our own solution in light of this analysis.

R I: The Scale Requirement – A common challenge in any type of visualization is scalability. In the context of pathway analysis the scale of the graph is hardly a problem, as the graphs are grouped into semantical units. The problem of scale in the context of pathway analysis is primarily concerned with the large number of experiments and experimental conditions. Scaling to several hundreds of experiments is a requirement for integrated pathway analysis.

R II: The Heterogeneity Requirement – While mRNA expression data is still the most prevalent data type analyzed in the context of pathways, next-generation sequencing has made other types of data readily available. Copy number variation and mutation status data are relevant examples, as mutation and copy number variation are often the cause of a change in a path, while mRNA only measures an effect. Additional datasets increase the scale problem, but it is also important to make clear distinctions in the representations used to avoid confusing analysts. Also, as different datasets are of different data type (copy number data is often ordinal, mutation status nominal), different visualization techniques are required.

R III: The Multi-Mapping Requirement – The most important nodes in pathways are gene products. They summarize various entities such as RNA, enzymes, proteins, etc., which have complex relationships. One gene can be the template for multiple proteins with slightly distinct domain composition called isoforms. Additionally, multiple genes sometimes encode proteins with similar functions, which are then consolidated into a gene family. As a consequence, a node in a pathway can be associated with multiple measurements of a single experiment. In fact, multi-mappings are quite common in KEGG. This significantly increases the complexity of visualizing experimental data.

R IV: The Layout Constraint Requirement – The layout of pathways is either produced manually by experts, or automatically. Manual pathway layout follows biological drawing conventions, by, for example, drawing cycles in circles or using pseudo-orthogonal edges. Also, these carefully hand-crafted layouts contain rich metadata and annotations. Automatic layouts either aim to respect those conventions to some degree (e.g., [13]) or use a force-directed layout (e.g., [24]). Our experience has shown that biologists prefer manually created pathways, or at least representations following these biological conventions, over arbitrary layouts. A reason for this might be that biologists are often intimately familiar with the layout of particular pathways and are reluctant to see it changed, as

this requires additional effort on their side. Integrating large quantities of experimental data in a constrained layout is more complicated than doing so in a free layout, since the free layout can be adapted to fit the data.

R V: The Topology-Attribute Coexistence Requirement – Analyzing experimental data in the context of cellular processes can be described as tasks on a graph. The process contains topology-based tasks, as well as attribute-based tasks. *Topology-based tasks* are, for example, those that look for node accessibility (which nodes are reachable from a source node) or connectivity (which nodes are connected, where are articulation points) [14]. An example for a topology-based task in pathway analysis is to find all processes that are influenced by a receptor at the cell surface. *Attribute based tasks* either focus on edge- or on node-attributes [14]. Common edge attributes in pathways characterize the type of a relationship between two nodes, for example transcriptional activation or inhibition, protein modification by cleavage, ubiquitination or phosphorylation, or biochemical conversion. The topology and the edge-attribute information is typically contained in the pathway maps themselves. The pathway maps also contain node-attributes (e.g., specifying the type of node; whether it is a protein, a compound, etc.), but mostly node-attributes are available in the form of mapped experimental data. Typically, graph visualization techniques are optimized for one or two of these tasks. Path-related topology-based tasks are, for example, well supported in node-link layouts, while edge-attribute-based-tasks are, for example, better supported by matrix layouts. The biggest challenge in pathway visualization is that all three types of tasks are equally relevant and the states and properties of all three – topology, edge attributes and node-attributes – influence the others. The node attributes, for example, influence the topology as experimental evidence can show that the topology is not valid for a particular condition. Consequently, a suitable visualization technique for pathways including experimental data has to enable all three types of tasks.

4 Related Work

Of the aforementioned challenges, the *scale requirement (R I)* sets the relevant body of related work apart from the wider sub-field of graph visualization. In standard node-link diagrams up to three or four node attributes can be encoded by assigning different visual attributes, such as color or size, to the nodes [3]. Consequently, we discuss techniques addressing node and/or edge attributes in excess of these numbers in addition to pathway visualization approaches (see the article by Gehlenborg et al. [6] for a review of the latter).

An example for a technique **adapting the layout to accommodate large amounts of node attributes** is the table-based graph visualization technique by Schulz et al. [23], where each node corresponds to a row in a table that can have multiple columns for multiple attributes. An approach by Pretorious and van Wijk [22] uses recursive partitioning for multiple node attributes. Another technique in this class is *GraphDice* by Bezerianos et al. [3], which positions the nodes in a scatter-plot according to the values of a pair of selected node attributes. For the examples mentioned, we observed that the accommodation of node attributes significantly impair the ability to understand the topology of the graph, violating *R V* as well as *R IV*. A number of systems strive for a compromise between node-size, embedded experimental data visualization, and topology information. Examples for automatically routed pathway graphs (violating *R IV*) are bar charts [9, 30] or line plots [8] used inside of nodes.

The approach of using a **separate linked view** is a widely used alternative. Shannon et al., for example, use a linked parallel coordinates view to visualize attributes in metabolic and protein-protein interaction networks [25]. Streit et al. have previously used linked parallel coordinates as well as a heat map to show associations between experimental data and pathways with the *Bucket* technique [28]. The recent *GraphPrism* [10] shows graph measures in stacked histograms and highlights nodes in a node-link-layout based on selections in the histograms. This approach could be easily extended to node attributes. *Cerebral* by Barsky et al. [1], a *Cytoscape* plug-in, also contains a parallel coordinates view linked to a node-link-layout depicting protein-protein interaction networks. The main drawback of separate linked views is that it requires interaction to see the association to the experimental data and the number of simultaneously associated entities is severely limited thereby violating *R V*.

Cerebral also employs **small multiples**, where each of the multiples contains a topologically identical node-link layout, but has different experimental data mapped to the node color. Lex et al. have used this approach to show differences of a small number of cancer subtypes on pathways [16]. While this approach is a good choice for a limited set of experiments or conditions, it can not handle more cases (*R I*) or heterogeneous attributes (*R II*).

Approaches that fulfill the *layout constraint requirement* (*R IV*) typically employ **on-node mapping** using color-coding with multiple glyphs (e.g., [17, 20, 27]) or color coding in combination with animation and selection [12]. As the available screen space for encoding the information is limited to the node size, this approach does not scale to more than a handful of experimental attributes (*R I*). Interactively switching experimental conditions requires significant cognitive effort when comparing conditions. However, it can be a suitable technique for topology-based tasks that only consider one condition.

The work most closely related to our own is *Pathline*, by Meyer et al. [19]. Pathline uses a set of visual encoding techniques to represent cycles, branches, and directionality of a linearized pathway. The linear layout allows for a simple comparison of functional data for genes and metabolites. Pathline also introduces the *Curvemap* view, which is used to compare temporal expression data between multiple species. While the linearized pathway is very space efficient, its biggest drawback is its unconventional layout that can hinder topology-based tasks (*R V*). Also, the linearized pathways currently need to be manually produced and therefore cannot make use of the wide body of pathways available in public databases.

5 The enRoute Visualization Technique

Creating a solution that meets all five requirements formulated in Section 3 is challenging. The only two options discussed so far that fulfill *R V*, i.e., that support topology-based as well as attribute-based tasks concurrently, are direct on-node mapping and small multiples. However, both of these techniques fail to address *R I - R III*. In order to create a technique that fulfills *R V* and allows experts to investigate a large number of experiments (*R I*) that potentially belong to different datasets (*R II*), using the original pathway map layout (*R IV*), we can make use of an observation: high-level topology-based tasks (e.g., identify the sub-part of the pathway relevant for a situation) are not conducted at the same time as low-level attribute based tasks (e.g., explore whether a de-regulation in a receptor in one experiment influences the rest of the path for this experiment).

This observation can be exploited by following Schneiderman's visual information seeking mantra: "*Overview first, zoom and filter, then details-on-demand*" [26]. The analyst starts by investigating the pathway in its standard layout as taken from one of the major databases (meeting *R IV*). She then selects a concrete path for which she wants to investigate experimental data, executing a *zoom and filter* operation. The chosen path is then shown in the enRoute view in a linear form. Next to the nodes we now have space available to concurrently show all mapped experimental data in a tabular format. In contrast to the classical multiple coordinated view approaches discussed before, this technique makes it possible to inspect the complete set of experimental data (meet-

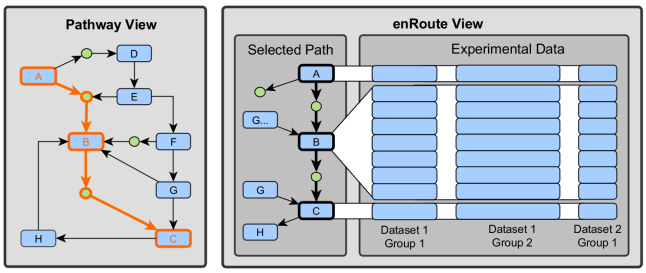


Figure 2: The components of the enRoute visualization technique. An analyst can choose a path in the pathway view, which is then shown side-by-side with the associated experimental data.

ing *R I, R II*), including resolved multi-mappings (meeting *R III*), along the path of interest, while the original layout is preserved (fulfilling *R IV*). Therefore, the overall process can be divided into three independent steps: *pathway brushing* in the source pathway, *path extraction*, and *sample data encoding*. These steps are covered by two tightly-coupled views that act together, as illustrated in Figure 2. The *pathway view* provides the complete topological information as well as an overview of the experimental data, while the *enRoute view* contains the linear path and the experimental data visualization. The elaborate interplay of all these systems can solve the critical *Topology-Attribute Coexistence Requirement (R V)*: The topological information for the whole pathway is preserved in the pathway view, while the topological information for a path and the experimental data is shown in the enRoute view. However, to fully support all of the requirements using such a setup, several important design decisions have to be made, which are explained in the next sections.

5.1 Pathway View

Pathway maps, as they are provided by, for example, KEGG, are available as annotated image files. We use these images and augment them to present information and enable interaction, as described in previous work [27]. The interactive version of the pathway makes it possible to select nodes, allowing synchronized highlighting with other views, but also to define paths (i.e., a series of nodes). Pathways can be chosen through a drop down list, via a search interface, or by clicking embedded pathway nodes.

5.1.1 Experimental Data Mapping on Pathways

While direct on-node mapping of experimental data suffers from the drawbacks discussed previously, there are two applications where it is beneficial. The first is the *overview* task, when deciding which path to choose. Initially, we use color-coding of average mRNA expression values of all experiments and multi-mappings of a node to indicate the "general trend" of the experimental data for this node. Since this feature itself hides all variation between experiments and experimental conditions, we additionally encode the standard-deviation from the mean value in a small bar at the side of the node (see the green vertical bars at the right of each node in Figure 3). Nodes with a high standard-deviation hint at underlying inhomogeneous and thus interesting experimental data. This enables experts to get a rough overview of the experimental data, which is particularly valuable for selecting a path in the first place. Color coding is also valuable if high-level topological information for a condition or an experiment is required. In these cases, we map the precise value of an experiment, the average and standard-deviation across multi-mapping nodes, or the average and standard-deviations of the condition. We use a blue-white-red color map by default, avoiding the more traditional red-black-green color map, which is problematic for color blind people. In cases where no experimental data is available for nodes, we indicate the possibility of interaction by a small black rectangle in the upper left corner of the node, see for instance the *CAM* gene in Figure 3.

5.1.2 Path Selection

As the enRoute visualization technique builds upon the idea of providing experimental data along a path in the pathway, the user-driven determination of the path is a critical step in the overall process. Selecting paths in graphs can be either done by letting the user interactively brush a series of edges or nodes that form a path (*iterative approach*), or by specifying a start and stop node (*start-stop approach*). While the former results in a unique path, the latter can produce multiple alternative paths. Being able to quickly investigate alternative paths interactively is an additional benefit of the start-stop approach. Also, specifying longer paths is faster using the start-stop approach. Consequently, we have chosen the start-stop approach as the default behavior for path selection. We calculate the set of alternative paths between two nodes using a variant of the Bellman-Ford algorithm [2]. By default, the shortest path is selected and loaded into the enRoute view. However, the user can browse through the alternative paths by using the mouse wheel and therefore can easily choose the most relevant one for the current task considering both topological and attribute information. However, in some cases it is desirable to extend a path (in either direction). For these cases, nodes can be added by selecting them while holding the control key, de-facto enabling an iterative approach as well. In addition to the path selection in the pathway view, paths can also be modified by selecting branching nodes in the enRoute view, which is discussed in Section 5.2.

5.1.3 Path Representation

To visualize the chosen path and its alternatives, we use a slightly modified version of the *Bubble Sets* technique [5]. Using Bubble Sets has several advantages compared to highlighting edges. First, they are more salient, due to their size, but especially due to their curved features, standing out compared to the otherwise largely orthogonal layout found in many pathway map databases [7]. Second, the precise routing of edges is often not available and one would have to resort to drawing not-exactly matching edges. Third, higlighting with Bubble Sets can also resolve ambiguities found in the original pathway textures. Figure 3 demonstrates how we use Bubble Sets with the interactive features previously discussed. In (a) the system shows two possible paths between two genes that were selected using the start-stop approach. The expert extends the currently selected path in (b) and chooses an alternative one in (c). An example of a resolved ambiguity can be seen at the cell membrane (the vertical double-lines) in Figure 3. Only from the image it is not clear which of the receptor nodes are connected to those farther right. Using the path overlay however, it is now obvious that *IGFR* is indeed linked to *PLCy* (see edge from second to third node highlighted by the orange path in Figure 3 (a)), although the original pathway graph does not explicitly show this.

While the original Bubble Sets technique is meant for visualizing a set of items, it is not developed to highlight a certain path between the members of a set exclusively. Thus, in order to force each Bubble Set to strictly follow a certain path, we have modified the Bubble Sets technique. Instead of allowing arbitrary branching between the nodes within a Bubble Set, we connect only those, which are connected by an edge in the pathway.

5.2 enRoute View

After a path was chosen in the pathway view, the enRoute view enables a detailed analysis of this path in context of the experimental data. The top-down, linear layout of the path is optimized for the node-attribute based task; if a node has a lot of mapping *data rows*, the spacing is adjusted to allow for a uniform row height in

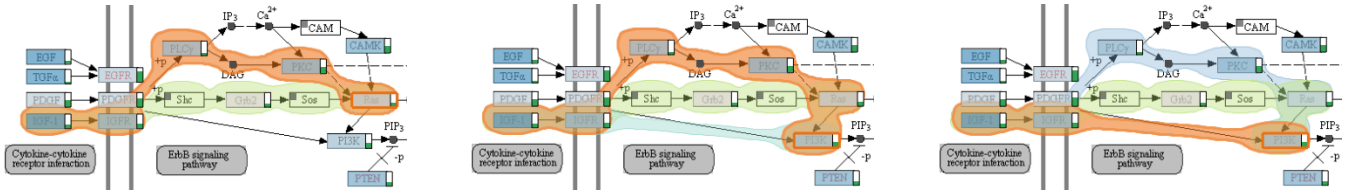


Figure 3: Pathway overlay showing all possible paths between a user-chosen start and end node using the Bubble Sets technique. In (a) the expert has selected *ICG-1* as start and *Ras* as end node, which results in two possible paths that can be chosen for an in-depth investigation in the context of experimental data. The system selects the shorter path by default (orange). In (b) the user extends the path by the *PI3K gene*, which adds one additional alternative path that is finally chosen by the user in (c).

the experimental data display. The nodes are connected to the rows in the experimental data display using ribbons, as is illustrated in Figure 2. While nodes that map to a single data row are unambiguously associated with its row through the position, multi-mappings and complex nodes can not be associated with rows using position alone. A complex node contains multiple nested nodes, which in turn can again contain multiple mappings. An example is shown in Figure 7, where the complex node contains five embedded nodes, which map to multiple rows each. The ribbons make these subtle associations obvious.

The enRoute experimental data view follows the divide-and-conquer visualization strategy [15]. Experiments are grouped based on a homogeneity criterion, which can be based on semantics (e.g., as all experiments in a group are from the same species, while other groups are from different species), or based on statistics (for example, obtained through a clustering algorithm). As shown in Figure 2, the groups are spatially separated, resulting in a matrix layout. While we require the data within a group to be from a single dataset, the groups themselves can be from arbitrary combinations of datasets, addressing the *heterogeneity requirement (R II)*.

5.2.1 Visualizing the Path

To preserve more of the topological information, enRoute also shows where branches join or leave the path. A branch is represented by its first node relative to the existing path and connects to the left side of the node where the branching occurs. In case of multiple branches coming into or leaving from a node, all incoming, respectively outgoing nodes are abstracted into one expandable node, keeping the visualization compact (see Figure 4(a)). The abstract nodes show how many branches they contain and display labels, if enough space is available. Each of these nodes can be expanded on demand to reveal the individual branch nodes. When a node is expanded, all other branches are grayed out, and the expanded nodes are rendered on top of them. The expanded nodes show a preview for their associated experimental data, as demonstrated in Figure 4(b). This facilitates the identification of potentially interesting branches.

A user can interactively switch to a branch, as shown in Figure 4(c). Depending on whether the branch is incoming or outgoing, the branch either replaces the nodes above or below the node where the branching occurs in the original path. The new branch contains all nodes that are in an unambiguous path, up to the next branch. Changes in the path triggered in the enRoute view are propagated to the pathway view, where the Bubble Sets surface is updated. All components of the enRoute visualization technique use linking and brushing. This helps to associate one of multiple branches in the enRoute view with the corresponding branches in the pathway view. The visual appearance of the path is modeled to resemble the KEGG pathway maps. Different designs for other data sources are possible.

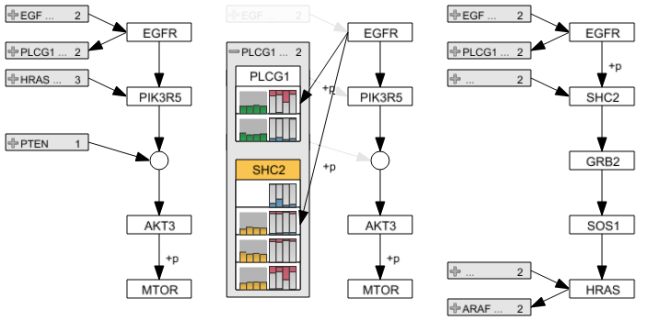


Figure 4: Illustration of the properties of the path representation in the enRoute view. (a) shows the selected path from *EGFR* to *MTOR*. By expanding the abstract node leaving *EGFR* in (b), the branch nodes *PLCG1* and *SHC2* are revealed, showing abstract previews of their associated data. Selecting the node *SHC2* causes its branch to replace all nodes in the path succeeding *EGFR* in (c). As *SHC2* is followed by an unambiguous path of nodes, all of them are added.

5.2.2 Visualizing Experimental Data

A scalable visualization of heterogeneous experimental data is one of the core challenges for the enRoute visualization technique. As previously mentioned, enRoute supports multi-dataset analysis for one of two data types: quantitative and ordinal. Heat maps are a common choice for visualizing quantitative as well as ordinal data in biomolecular data visualization. However, we have decided not to use heat map views, since changes in hue or value are known to be inferior to changes in position and length for quantitative and inferior to position for ordinal data [18]. Meyer et al. have recently given an example for expression data, where a mirroring effect was apparent in a line-plot but much less so in a heat map. Heat maps or more generally pixel-oriented displays, are, however, superior, as far as scalability is concerned. Clustered heat maps can convey trends even if more data values are visualized then pixels are available. However, while enRoute requires significant scalability in terms of the number of experiments, the number of genes is limited by the number of nodes in the path.

Consequently, we chose to use bar charts instead of heat maps for the visualization of both quantitative and ordinal data. Figure 5(a) illustrates a case for quantitative data, where each bar represents one mRNA expression value for one experiment. We use a slight cushioning to make the borders between bars apparent. As previously mentioned, the experiments are grouped based on homogeneity. Groupings can contain overlapping sets of experiments. The groups have captions at the top and bottom, the background of which color-codes the dataset type. In Figure 1, for ex-

(a) (b) (c) (d)

Figure 5: The four types of visual encoding for experimental data (quantitative data in (a-b) and ordinal data in (c-d)). (a) Each experiment has its own vertical bar. (b) Abstraction of several experiments into a bar chart with error bars. (c) Redundantly encoded ordinal values where reduced copy numbers are shown in blue bars pointing downwards, increased copy numbers are shown in red bars pointing upwards. (d) Histogram abstracting a group of copy number values.

ample, groups of mRNA data have a turquoise background, while copy number data captions have a green background. Depending on the task, it might be sensible to use an abstraction for multiple experiments instead of a separate bar for each one. Our solution for abstractions of quantitative data is shown in Figure 5(b). Instead of vertical bars we use bar charts with error bars, where the bars encode the mean value of the underlying experiments for the row and the error bars encode the standard-deviation. Abstracted groups have a constant width, in contrast to individual bars, where the width adapts to the available space. This guarantees the ability to compare the lengths across columns in the matrix. Not only is this representation less cluttered, it also facilitates a better comparison of values of the same group along the path. If, in contrast, a comparison between groups is more relevant in a particular situation, the individual (vertical) bars are preferable, as they facilitate comparison across groups. Aside from the orientation, the abstract bars also are of another color to make the difference evident.

For ordinal data, we employ a redundant encoding using both length and color. Figure 5(c) shows an encoding optimized for ordinal copy number datasets. Copy number is often categorized into five categories – deleted on both alleles, deleted on one allele, regular copy number, low amplification, and high amplification. Our encoding shows nothing for a regular copy number, a bar pointing downward from a baseline in light blue for a deletion on one allele, and a longer bar in dark blue for a deletion in both alleles. Increased copy numbers are encoded using bars pointing upwards in either light or dark red. This visual encoding is also suitable for cases where copy number data is available in hybrid form: quantitative values in case of increased copies (e.g., 10 vs. 100 copies) and ordinal values for the deletion states. In line with the abstract display of quantitative data for multiple experiments, the abstraction of ordinal data is rendered as horizontal bars, more specifically as a horizontal histogram. Thereby, the same color coding is used.

The aforementioned previews of experimental data for branches in the path use a similar visual encoding. The branches contain one bar for each group of experiments. Here, the abstract bars are rendered vertically, due to the space constraints in the preview. Instead of a histogram, a stacked bar is shown for ordinal data.

The experimental data display uses extensive linking and brushing. Not only is it synchronized with all other views as far as genes are concerned, it also utilizes brushing within the experimental data display. This is particularly valuable, when the same sample is contained in multiple columns, possibly even in different data types. Figure 6(b), for example, shows a brush (in gold) for those samples that have a high-level amplification of the gene *PDGFRA*, which allows to look for influences of copy number variation on mRNA expression.

6 Implementation and Scalability

The enRoute visualization technique is part of Caleydo, an open source biomolecular data visualization framework [28]. Caleydo is implemented in Java and uses JOGL for rending. The path overlay on top of the KEGG pathway maps is created using a modified version of a free implementation of the Bubble Sets technique[2].

enRoute scales to hundreds of experiments covering even the most extensive datasets currently available. Figures 1 and 6 show public mRNA and copy number datasets from *The Cancer Genome Atlas* (TCGA)[3] containing 550 samples and ∼20,000 genes each. Figure 6 contains the whole set of experiments twice (in different groupings, once abstracted, once showing all values). If the length of the path exceeds the available screen space, we use scroll-bars to navigate to the off-screen parts. We found this to be reasonable due to the linear nature of the exploration process along the path.

7 Case Studies

enRoute was designed in collaboration with the fifth author, a biologist from the Medical University of Graz. We evaluate the enRoute visualization technique using case studies conducted with this biologist and two different datasets. The first is the aforementioned TCGA dataset containing mRNA and copy number data for patients suffering from *glioblastoma multiforme* (GBM), a type of brain cancer, the second is a dataset collected at the Medical University of Graz for a mouse model of steatohepatitis containing mRNA expression data.

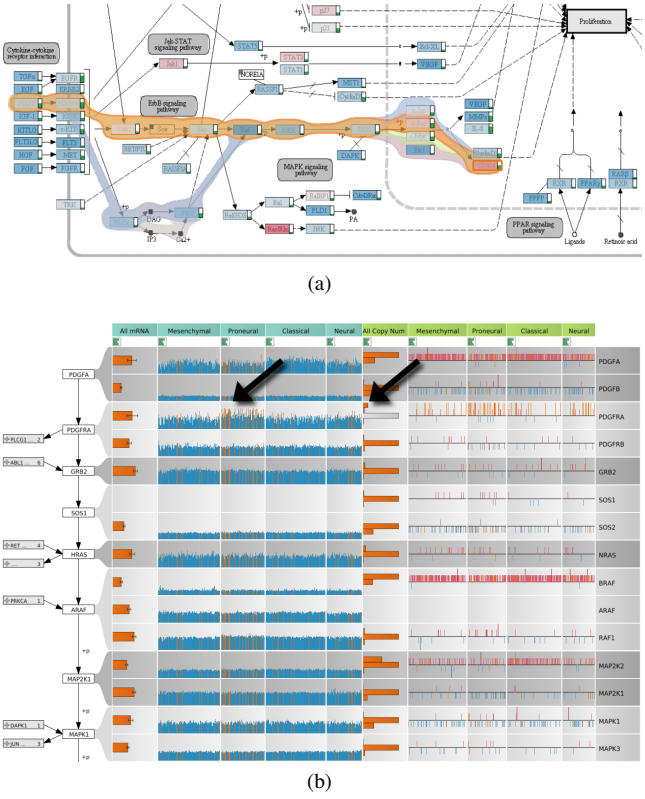
7.1 Glioblastoma Multiforme

Our first case study demonstrates the path extraction feature by visualizing a part of the large *Pathways in Cancer* map from KEGG. We use gene expression and copy number variation data generated by the TCGA project to ask the question if the signaling cascade from platelet derived growth factor A via map kinases to the cell cycle regulators *CDK4* and *CyclinD1* plays a role in the different subtypes of GBM. The *Pathways in Cancer* map encompasses many important regulatory mechanisms involved in tumor proliferation like angiogenesis, metastasis, apoptosis evasion, resistance to chemotherapy and cell cycle activation. The cell cycle is activated by the transcription factor *c-myc*, which is a fundamental event that leads to unrestricted growth of tumor cells. The signaling cascade leading to *c-myc* activation can either follow the canonical path via g-protein coupled receptors, *Grb2*, *Ras*, *Raf*, *Mek*, *Erk* to the *AP-1* complex or alternatively via protein kinase c mediated phosphorylation of *Ras* and *Raf*. enRoute allows the selection of all sub-paths leading to the activation of the cell cycle in the *Pathways of Cancer* map and automatically selects one of the possible connections, as shown in Figure 6(a).

Using the mouse-wheel, the researcher can now highlight and select the different possible paths in the large map. Selection of the path leads to the extraction of a linear representation of the path into the enRoute view. Here the available experimental data is dynamically displayed in bar charts, which are grouped, as can be seen in Figure 6(b). The grouping in this case is based on a classification by Verhaak et al. [29], where each group corresponds to a clinically relevant subtype. The researcher stated that visualization of the data in this manner allows him to compare the gene expression in the sub-types of GBM and correlate it to the copy number variation data. Using this visualization, it is easily detectable that in GBM, the *PDGFA* gene has become replicated multiple times in the genome of the tumor, whereas the *PDGFB* gene is commonly lost (see Figure 6(b)). The replication or loss of the *PDGF* genes does not, however, impact the expression levels of these growth factors indicating tight control over protein levels downstream of gene count. This is different for the copy number variation of *PDGF-receptor A (PDGFRA)*. The amplification of this gene is associated with overexpression of the protein, which can be highlighted by selecting the experiments with high amplification in the data plots (orange in Figure 6(b)). Interestingly, this copy number variation and

[2]http://github.com/JosuaKrause/Bubble-Sets

[3]http://cancergenome.nih.gov



(a)

(b)

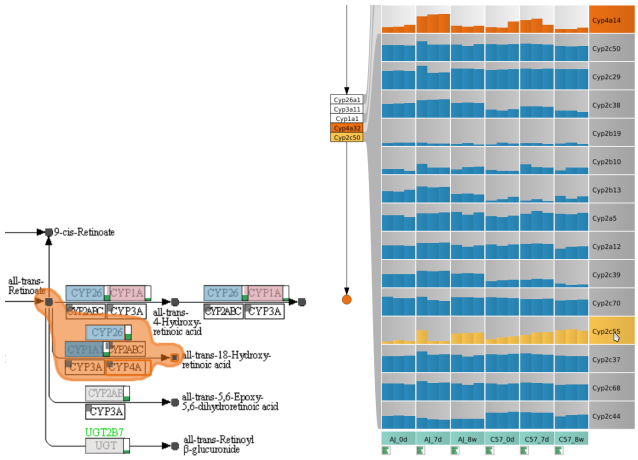
Figure 6: The GBM case study. (a) Alternative branches from the receptors to *cell proliferation* are highlighted, the orange alternative is selected. (b) enRoute view for the selected path and the GBM dataset showing about 550 samples for copy number and gene expression data each. Notice the correlation between the increased copy number status of *PDGFRA* and increased expression levels in the *proneural* subtype.

upregulation of *PDGFRA* seems to be specific to the *proneural* subtype of GBM. Further downstream in the signaling cascade we find that nearly all tumor samples show amplification of the *BRAF* locus and the *MAP2K1* locus also known as *MEK1*. This demonstrates that the activation of the cell cycle by the *map-kinase pathway* is an important feature of GBM indicated by multiple gene amplifications and changes in the expression of the proteins involved in this signaling cascade.

In summary the pathway extraction via the enRoute tool allowed the expert to study gene expression data and copy number variation data in an easily selectable sub-path of the complex pathway maps contained in the KEGG database. The expert stated that the visualization of the flow of signal transduction along the vertical axis and the horizontal organization of the different sample types (i.e., tumor subtypes) is very intuitive and helped him in efficiently extracting biologically relevant information from the dataset.

7.2 Steatohepatitis Mouse Model

The second use case demonstrates another advantage of the path-extraction feature of enRoute for the analysis of biochemical pathways. In biochemical pathways, the links between nodes represent chemical conversions catalyzed by proteins, which are then called enzymes. Enzymes are usually capable of converting multiple substrates to products, albeit with different efficiencies. This partially unspecific reaction, and the fact that enzymes often have many protein isoforms generated by alternative splicing, lead to heavy multi-mapping of gene names to nodes in biochemical pathways. It is



(a) (b)

Figure 7: Complex multi-mapping in the KEGG *retinol metabolism* pathway map. (a) A path over a complex node is selected. (b) The gene expression data for two mouse strains (C57 and AJ) at different time points of intoxication with DDC (0d, 7d, 8w). *Cyp4a14* (orange) is induced in short term toxicity in the 7d treatment. *Cyp2c55* (gold) is differentially regulated in the AJ and C57 mouse strains.

extremely difficult for a researcher to understand the biology of a metabolic conversion using pathway maps, if the node of interest contains a multitude of involved gene names. Using enRoute, the researcher can select the nodes upstream and downstream of the enzymatic reaction and thus obtain a linearized representation of all genes involved. This is demonstrated in Figure 7(a), where the conversion of *all-trans-Retinoate* to *all-trans-18-Hydroxy retinoic acid*, a part of the *Retinol* (Vitamin A) metabolism, was selected. The single node in between these two metabolites encompasses 18 different proteins that can contribute to this conversion. The enRoute tool can display all these genes in a convenient map (see Figure 7(b)) allowing to study the gene expression of each single gene in all experimental conditions. In this example we have loaded data generated by gene expression profiling of the livers of experimental animals (two mouse strains; AJ and C57) during the course of DDC (*3,5-diethoxy-carbonyl-1,4-dihydrocollidine*) treatment. Expression was measured at three time points, after 0 days, 7 days and 8 weeks of intoxication. This treatment induces histological changes in the liver *parenchyme*, which resemble closely the changes seen in human *steatohepatitis*, making it a model for *non-alcoholic steatohepatitis*. These morphological changes are the result of oxidative stress, which is usually connected to the activity of *cytochrome p450* enzymes. These enzymes catalyze the oxidation of organic substances and are major enzymes involved in drug metabolism and bioactivation. Studying the aforementioned conversion of *retinoate* to *retinoic* acid by *cytochrome p450* enzymes, it can now be seen that *Cyp2c55* (highlighted in gold) is differentially regulated in the AJ and C57 mouse strains and that *Cyp2b13* is predominantly expressed in mouse strain A. Additionally it can be detected that *Cyp4a14* (highlighted in orange) is induced in short term toxicity in the 7d treatment timepoint. All this information was not visible to the researcher using conventional on-node mapping approaches and was successfully visualized using enRoute.

8 Conclusion and Future Work

Developing a solution that enables experts to analyze functional implications of a large number of experimental data on cellular processes is a challenging and yet unsolved task. We have introduced five requirements that are crucial for creating such a system. In

short, an optimal solution needs to allow experts to concurrently investigate hundreds of samples in dozens of experimental conditions and even considering multiple, heterogeneous datasets in the context of pathway maps. We propose the enRoute visualization technique that addresses all five requirements in a tightly-coupled dual-view approach. Experts can select a path from a pathway map, which is then highlighted in the map. The selected path is extracted and shown in linear form side-by-side with the associated experimental data. Our case studies showed that enRoute enables analyses that are not possible by other means. Feedback from our collaborators as to the utility was enthusiastic throughout. The enRoute visualization technique will be publicly available with the next release of the Caleydo software.

The enRoute system integrates about 500 pathways from the KEGG database for the two most researched organisms, human and mouse. enRoute can be easily extended to other organisms covered by KEGG. However, although KEGG has a very broad focus and is nowadays widely spread in the community, we plan to integrate further pathway resources of different kinds, which often have a special focus on certain types of cellular processes. A valuable extension would be, for instance, to include the EBI IntAct database, which centers on protein interaction networks. Furthermore, we plan to allow experts to not only investigate the association of experimental data in the context of a single pathway at a time, but to concurrently see their interdependencies within the cellular network to other related pathways.

Acknowledgements

We would like to thank Hans-Jörg Schulz for his input. This work is supported by the following grants: Caleydoplex (P22902, FWF), Tumorheterogeneity (GZ:A3-22.M-5/2012-21, state of Styria), IMGuS (Austria Wirtschaftsservice), and inGeneious (385567, FFG).

References

[1] A. Barsky, T. Munzner, J. Gardy, and R. Kincaid. Cerebral: Visualizing multiple experimental conditions on a graph with biological context. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '08)*, 14(6):1253–1260, 2008.

[2] R. Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16:8790, 1958.

[3] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J. D. Fekete. GraphDice: a system for exploring multivariate social networks. *Computer Graphics Forum (EuroVis '10)*, 29(3):863–872, 2010.

[4] L. C. Cantley and B. G. Neel. New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-Kinase/AKT pathway. *Proceedings of the National Academy of Sciences*, 96(8):4240–4245, 1999.

[5] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '09)*, 15(6):1009–1016, 2009.

[6] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A. Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7(3):56–68, 2010.

[7] R. Hoffmann, P. Baudisch, and D. S. Weld. Evaluating visual cues for window switching on large screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*, pages 929–938, 2008.

[8] Z. Hu, J. Hung, Y. Wang, Y. Chang, C. Huang, M. Huyck, and C. DeLisi. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Research*, 37(Web Server):W115–W121, 2009.

[9] B. H. Junker, C. Klukas, and F. Schreiber. VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7(1):109, 2006.

[10] S. Kairam, D. MacLean, M. Savva, and J. Heer. GraphPrism: compact visualization of network structure. In *Proceedings of the ACM Conference on Advanced Visual Interfaces (AVI '12)*, 2012.

[11] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, and et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database-Issue):480–484, 2008.

[12] P. D. Karp, S. Paley, and P. Romero. The pathway tools software. *Bioinformatics*, 18(Suppl 1):S225–S232, 2002.

[13] A. Lambert, J. Dubois, and R. Bourqui. Pathway preserving representation of metabolic networks. *Computer Graphics Forum (EuroVis '11)*, 30(3):1021–1030, 2011.

[14] B. Lee, C. Plaisant, C. S. Parr, J. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proceedings of the AVI Workshop on BEyond time and errors: novel evaluation methods for information visualization (BELIV '06)*, page 15, 2006.

[15] A. Lex, H. Schulz, M. Streit, C. Partl, and D. Schmalstieg. VisBricks: multiform visualization of large, inhomogeneous data. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '11)*, 17(12):2291–2300, 2011.

[16] A. Lex, M. Streit, H. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg. StratomeX: visual analysis of Large-Scale heterogeneous genomics data for cancer subtype characterization. *To appear in: Computer Graphics Forum (EuroVis '12)*, 31(3):fff–lll, 2012.

[17] H. Lindroos and S. G. E. Andersson. Visualizing metabolic pathways: comparative genomics and expression analysis. *Proceedings of the IEEE*, 90(11):1793–1802, 2002.

[18] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

[19] M. Meyer, B. Wong, M. Styczynski, T. Munzner, and H. Pfister. Pathline: A tool for comparative functional genomics. *Computer Graphics Forum (EuroVis '10)*, 29(3):1043–1052, 2010.

[20] B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo, and Z. Trajanoski. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Research*, 33(Web Server issue):633–637, 2005.

[21] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, and et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510–522, 2010.

[22] A. J. Pretorius and J. J. Van Wijk. Visual inspection of multivariate graphs. *Computer Graphics Forum (EuroVis '08)*, 27(3):967–974, 2008.

[23] H. Schulz, M. John, A. Unger, and H. Schumann. Visual analysis of bipartite biological networks. In *Proceedings of the Eurographics Workshop on Visual Computing for Biomedicine (VCBM '08)*, pages 135–142, 2008.

[24] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[25] R. Shannon, T. Holland, and A. Quigley. Multivariate graph drawing using parallel coordinate visualisations. Technical report, 2008.

[26] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages (VL '96)*, pages 336–343, 1996.

[27] M. Streit, M. Kalkusch, K. Kashofer, and D. Schmalstieg. Navigation and exploration of interconnected pathways. *Computer Graphics Forum (EuroVis '08)*, 27(3):951–958, 2008.

[28] M. Streit, A. Lex, M. Kalkusch, K. Zatloukal, and D. Schmalstieg. Caleydo: Connecting pathways and gene expression. *Bioinformatics*, 25(20):2760–2761, 2009.

[29] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, and et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, 2010.

[30] M. A. Westenberg, S. A. F. T. Van Hijum, O. P. Kuipers, and J. B. T. M. Roerdink. Visualizing genome expression and regulatory network dynamics in genomic and metabolic context. *Computer Graphics Forum (EuroVis '08)*, 27(3):887–894, 2008.