Entourage: Visualizing Relationships between Biological Pathways using Contextual Subsets

Alexander Lex, Christian Partl, Denis Kalkofen, Marc Streit, Samuel Gratzl, Anne Mai Wassermann, Dieter Schmalstieg, and Hanspeter Pfister

Abstract—Biological pathway maps are highly relevant tools for many tasks in molecular biology. They reduce the complexity of the overall biological network by partitioning it into smaller manageable parts. While this reduction of complexity is their biggest strength, it is, at the same time, their biggest weakness. By removing what is deemed not important for the primary function of the pathway, biologists lose the ability to follow and understand cross-talks between pathways. Considering these cross-talks is, however, critical in many analysis scenarios, such as judging effects of drugs. In this paper we introduce Entourage, a novel visualization technique that provides contextual information lost due to the artificial partitioning of the biological network, but at the same time limits the presented information to what is relevant to the analyst’s task. We use one pathway map as the focus of an analysis and allow a larger set of contextual pathways. For these context pathways we only show the contextual subsets, i.e., the parts of the graph that are relevant to a selection. Entourage suggests related pathways based on similarities and highlights parts of a pathway that are interesting in terms of mapped experimental data. We visualize interdependencies between pathways using stubs of visual links, which we found effective yet not obtrusive. By combining this approach with visualization of experimental data, we can provide domain experts with a highly valuable tool. We demonstrate the utility of Entourage with case studies conducted with a biochemist who researches the effects of drugs on pathways. We show that the technique is well suited to investigate interdependencies between pathways and to analyze, understand, and predict the effect that drugs have on different cell types.

Index Terms—Pathway visualization, biological networks, subsets, graphs, biomolecular data

1 INTRODUCTION

All living organisms can be considered as highly complex networks of biomolecules (genes, gene products, and metabolites) and biochemical reactions. It is the sum of tightly controlled and regulated interactions between these components that determines an organism’s form and function. In the study of biological networks, the series of actions among biomolecules that lead to specific biological effects are commonly described as biological pathways. In other words, a pathway is a meaningful subset of biomolecules and reactions whose interplay fulfills a function in a cell or organism. Some pathways describe metabolic processes, e.g., the production of the amino acid tyrosine (Tyrosine metabolism pathway), whereas other pathways highlight the processes involved in a disease, such as the Glioma pathway, which describes the molecular mechanisms dysregulated in brain cancers.
2 Domain Goals and Biological Background

While the analysis of relationships between multiple pathways is an important task in many application scenarios, our development of the Entourage technique was driven by three domain goals in drug discovery. We have been in close collaboration with an early stage drug discovery research group from a large pharmaceutical company over a period of several months, which included several meetings with larger groups of researchers and weekly meetings with one of their domain experts. In the following we describe their analysis goals and the datasets required to achieve these goals.

Understand a drug’s mechanism of action and drug sensitivities of cell lines. The target of established drugs is typically known. In many cases, such drugs inhibit one or multiple gene products. However, there are fluctuations on how cell lines with distinct genomic profiles respond to the drugs [4]. Finding out why, for example, some cancer cells are killed by the drugs while others survive is one objective.

Judge side effects and safety of drugs. Although drugs are often designed to modulate only one particular biological pathway, their action on the cell and the organism as a whole must be considered in their development to better assess their safety. Being able to analyze cross-talk between pathways can help in judging the potential risks associated with a compound early on in the drug development process [32].

Identify potential for repositioning of drugs. Two alternative routes are usually taken in drug discovery: (1) developing new chemical entities and (2) finding new uses for already existing or previously failed drugs that have shown an adequate clinical safety profile [2]. The second route is usually more effective as such drugs can be approved quicker. As more and more knowledge about biological interactions and refined pathway maps become available, it is quite possible that existing drugs can be repurposed for a broader spectrum of therapeutic indications. Inter-pathway analysis can help to identify potential new therapeutic uses for approved drugs.

The data to be analyzed in these tasks can be classified into three categories: pathway data, and two forms of experimental data: genomic and pharmacologic profiling data. We have already introduced pathway data, and will now briefly explain the experimental data.

Genomic profiling data refers to datasets that measure the activity or structural variation of genes. An example of genomic activity is gene expression (or mRNA expression), which indicates how much of a functional gene product, such as a protein, is produced. Changes in gene activity can cause pathway dysregulation and a diseased state. One reason for a change in gene activity can be a structural variation. Structural variations occur on different scales. They can affect only a single base pair in the DNA or modify a whole chromosome. Two common forms of structural variation data are copy-number variation data, which records large scale duplications or deletions of genes, and mutation data, which captures smaller changes within an individual gene. Structural variation can result in changed activity or even loss of function. For example, a mutation in the gene PTEN results in uncontrolled signaling in a pathway promoting cell survival, which can lead to tumor growth [5]. The joint analysis of pathways and genomic experimental data makes it possible to identify such effects.

Pharmacologic profiling data essentially measures how cells react to compound exposure. A common measure is the half maximal inhibitory concentration (IC$_{50}$), which reports the concentration at which a drug achieves 50% inhibition, e.g., the drug concentration that is required to kill half of the treated tumor cells. The lower the concentration, the more effective the drug and the more sensitive is the cell line under study to the drug treatment. Pharmacologic and genomic profiling data are commonly jointly analyzed to, for example, identify reasons for the differential response that cell lines show to drug treatment. Integrating pathways into such an analysis can make it much more targeted, since it allows analysts to focus on the processes influenced by the compound and to explore related processes. Currently, several distinct tools are used for this analysis. Entourage is the first system to combine the different data types into one integrated interactive visualization.
3 ANALYSIS TASKS

All of the domain goals described in the previous section have pathway analysis as pivotal parts of them. In this section we break down these domain goals into generalizable analysis tasks. We have elicited these tasks through interviews and feedback sessions with our collaborators. We classify the detailed tasks into two categories, the Pathway Interconnectivity tasks and the Pathway-Experimental Data Linking tasks. Note that both of those tasks are among the most critical requirements in pathway analysis [26] and are considered open problems in biological network analysis [1].

The Pathway Interconnectivity tasks deal with finding pathways related to each other and analyzing the relationships between pathways. The analysis tasks are:

Find related pathways. While an initial pathway is typically known for the stated goals, it is important to easily find related pathways, as cross-talk and other interdependencies are more likely between highly related pathways. We consider two pathways as related when they either share one or multiple genes (nodes), have an edge crossing from one pathway to the other, and/or if one is contained or referenced in the other.

Identify high-level relationships of pathways. When related pathways are found, it is also important to see how they are related. For example, it is interesting to see whether the same sub-process is contained in both pathways, or whether one pathway is contained or referenced within the other pathway.

Identify the role of a gene in multiple pathways. Identifying the role of a gene in other pathways is important to determine the different cellular processes that a gene is involved in, which, for example, is valuable knowledge when assessing the suitability of a gene as a drug target.

Find path intersections. As a change in gene activity, e.g., caused through a mutation or modulation by a drug, can influence the activity of subsequent genes in a path, it is important to not only look for the role of the originally altered gene but also to explore the role of genes that might be influenced by it. This can be done by exploring the relationships of the nodes downstream of the original gene, i.e., by finding pathways that intersect the path of a changed gene.

One-Network Approaches. As an example of the first class, KEGG [13] provides a high-level overview of the metabolic network (the KEGG Atlas) and lets analysts drill down into individual pathways. Other tools, like iPath [33] or Pathway Projector [16] use the same underlying data but improve the interaction with the atlas. While the original KEGG Atlas and iPath both use hyperlinks to replace the network overview with pathway maps on demand, Pathway Projector embeds node information directly on the all-encompassing map, thereby relying wholly on zooming and panning for navigation. Pathway Projector differs from the other two also in respect to how it represents gene nodes. The KEGG Atlas and iPath both represent genes and embedded pathways only as edges between the metabolites (the intermediate products of the metabolism), while the Pathway Projector actually shows nodes for genes, enzymes, and metabolites, which allows for direct mapping of experimental data.

All of these techniques show an incredibly large amount of data on a single screen. For example, the KEGG Atlas for E. coli, a comparatively simple organism, summarizes 1,365 genes, 1,813 enzymes, and 1,572 metabolites. In order to handle all this data, these techniques rely on selecting a focus, either by zooming and panning, or by changing into a different view altogether. This makes it very hard to identify intersections to parts of the network outside of the currently visible area. All of these examples rely on a static layout, making features such as the layout lens [31] that pull connected nodes into focus, impossible.

Multiple Pathway Maps Approaches. The approach of showing multiple pathways and visualizing relationships among them is taken in earlier versions of the Caleydo system, for example, in the Jukebox [29] and the Bucket [19] techniques. Both of them arrange multiple pathways in a 2.5D layout, while one of the pathways serves as focus. Visual links are used to connect related items on demand. As both techniques use thumbnails for context pathways, labels or even individual nodes are hard to see. Also, none of the techniques can show more than one relationship at a time.

Similar to the Caleydo techniques, Justafi et al. extended the Vanted system to show multiple pathways as thumbnails [12]. They use navigation glyphs that show how individual nodes are connected to other pathways. A glyph has one petal for each possible link to other pathway groups, where the length of the petal encodes whether there is a link to a pathway group or not. These petals can be used to navigate to other pathways. The system, however, does not show any relationships on a node level.

Hybrid Approaches. VizANT [11] takes a hybrid approach by showing a larger network as a set of meta-nodes which can be uncollapsed to reveal the underlying nodes and their cross-pathway connections. In the visualization community such meta-nodes are typically referred to as super nodes and are supported by a wide range of general purpose graph visualization frameworks such as Tulip [3] or CGV [31]. The concept of meta- or super nodes is significantly different from that of our contextual subsets as the nodes are not smart with respect to the context. A super node is either collapsed or not, whereas contextual subsets can show the elements that are contextually relevant while hiding the others.

The approach by Klukas and Schreiber [15] also employs super nodes. It uses a force-directed graph layout of abstracted pathway nodes and their relationships. Each abstract pathway node can interactively be expanded to show all the nodes of the corresponding pathway in detail. These nodes are arranged using the KEGG layout. They show all cross-pathway connections of individual nodes for multiple pathways at the same time. While this approach is reasonable for a limited set of pathways, the adding of more pathways continuously introduces clutter and reduces node size.

Rohrschneider et al. [25] use a similar approach in terms of showing multiple pathways at the same time, but use a grid-based automatic layout for the overall metabolic network. Their navigation approach is of particular interest, because they use the table-lens metaphor to switch between pathway super nodes and detailed renderings of the pathway. However, as with all super node approaches, the nodes are either expanded or collapsed but do not allow a context-based preview.
Consequently, the technique also can only provide detail and context for a very small set of pathways simultaneously.

**General Subset Techniques**  Our approach is also related to subset visualization techniques such as VisBricks [18] or Portals [23, 10]. VisBricks partition numerical datasets into subsets and show each subset with the visualization technique most suitable for the contained data and task. Portals are local regions within a visualization that show a different view on the area they cover. Olston and Woodruff employ portals to show data overlayed on maps [23], while the Hadlak et al. in-situ visualization uses portals (also) for graphs [10]. None of these techniques however, use any form of semantic context. Notice that we use the term portal to refer to shared nodes between pathways and that this usage is unrelated to Olston and Woodruff’s term.

**Visualizing Experimental Data in Pathways** We use the enRoute technique [24] to visualize experimental data. enRoute uses path extraction and a separate linked view, which we integrated into Entourage. In contrast to Entourage, enRoute is strictly limited to a single pathway and has no notion of finding or presenting related or multiple pathways. Entourage and enRoute are complementary: the former addresses the problems of large and disjoint networks, while the latter makes it possible to visualize many node attributes.

**5 Visualizing Pathway Relationships**

Enabling the Pathway Interconnectivity task requires a joint analysis of multiple pathways. Current techniques, however, lack the flexibility required for exploring interdependencies across pathway boundaries. The main problem one must address is scale. Current approaches either cannot show individual nodes sufficiently large or cannot show relationships between multiple pathways. We developed contextual subsets to remedy this issue.

Figure 2 illustrates the difference between a traditional multiple pathway analysis and the contextual subsets method. The traditional approach depicted in Figure 2(a) shows all nodes for all pathways. The pathways in this example share several nodes. We refer to such shared nodes as portals as they allow us to jump from one pathway to another. Figure 2(b) illustrates the same set of pathways using the contextual subsets technique. Instead of showing all pathways in detail, we distinguish between focus pathways, shown at full scale (PW 1 in Figure 2(b)), and context pathways, which are smaller and show only a contextually relevant subset of their graph (PW 2 and 3 in Figure 2(b)). What is contextually relevant is driven by a user-selected focus node (A, purple in Figure 2(b)). The context pathways only show limited subsets of their network that also contain node A. In the example shown in Figure 2(b), the most important path is shown for each occurrence of the focus node, while other branches are only indicated, as is evident in Pathway 2.

This simple yet effective principle makes use of an observation: analysts do want to see all the details of one pathway map (their focus pathway), but do not need to see all the intricate details of other potentially involved processes (the context pathways) to judge interdependencies to their focus pathway. Entourage utilizes the observation that the focus of attention shifts serially to optimize the visible content to what is currently relevant to the analyst. The challenge we have to address is the continuous change of attention, the adoption of the analysis focus in the process of an exploration. Entourage employs a series of visual encodings and interaction techniques to make these changes as convenient and transparent as possible.

**5.1 Overview**

Figure 3 shows Entourage’s main components. The focus pathway takes up the majority of the space, while the context pathways are shown at the side. In this example E2F was selected as the focus node and the context pathways show their paths related to this node. Details on how context paths are selected are explained in Section 5.2, as are our methods to find relevant pathways.

Changes in focus are driven by user selections. However, choosing a meaningful focus is not always easy. Sometimes analysts will need to explore high-level relationships of pathways beforehand they can set a sensible focus. Visualization is ideally suited to convey such high level relationships. Relationships between pathways are largely driven by portals as they connect two pathways. Showing portals and where they link to is therefore the most important aspect of showing high-level relationships between pathways. Figure 3 shows our approach for visualizing portals. We use a combination of stubs, which are shown for all portals at the same time, and visual links, i.e., visible edges, which are shown on request. These visual encodings efficiently convey high-level relationships between pathways and enable an analyst to set good focus points. Our visual encodings for showing relationships are explained in Section 5.3.

Finally, we need to address how to efficiently manage display space, as multiple focus and context elements compete for the limited screen real estate. We use an intelligent arrangement of pathways as well as multiple levels of detail for context pathways to optimize the display space, which are described in Section 5.4.

**5.2 Determining Context Paths and Pathways**

As discussed before, the contextual subsets concept is based on showing contextual information for a user-chosen focus, i.e., a focus node of a pathway. Which context information is eventually displayed depends on two factors: which paths in a pathway contain a focus node and which pathways are considered in the first place.

**Determining Context Paths** Context paths are selected by searching the graph for occurrences of the focus node or for immediately related nodes. Related nodes are, for example, nodes belonging to the same gene family. As it is common in nature that several distinct genes can fulfill the same role, albeit often with varying efficiency, pathway maps use both, a single label for the whole family or individual label for each of the family members. We consider these multi-mappings in our choices of relevant paths. This is the reason why occasionally differently labeled nodes are connected in Entourage.

Paths can either be unambiguous, as is the case in Pathway 3 of Figure 2(b), or contain branches, as in Pathway 2. If a path contains branches we automatically determine the branch that is likely to be most interesting by calculating the most variable branch in terms of the underlying experimental data. We do so by calculating the standard deviation across all experiments for each of the mapped datasets for every possible branch and choosing the branch that exhibits the highest deviation. As discussed in previous work [24], we preserve as much of the topology in the vicinity of paths as possible. Incoming and outgoing branches are collapsed into abstract nodes to save space, but can be extended to full-size nodes and switched-in to replace the main branch on demand. We decided against more complex attempts of linearizing larger portions of the network and including branches and cycles [22], to make the paths easy to understand for the analyst. Furthermore, we limit the length of automatically determined paths to what fits
Focus Pathway

Portals with Stubs

CyclinD1
CDK4

Window Stub

Context
Pathways

Matching
Focus Nodes

Focus Node

Portal Links

Fig. 3. The major components of Entourage. The focus pathway shows all details while the context pathways only show what is relevant in the context of the focus node. The insets at the top show how we indicate connections between pathways.

conveniently in the available space constraints, but give analysts the ability to extend the paths manually.

**Determining Pathways**  Entourage shows paths only for manually selected pathways but suggests pathways that are relevant for a current focus node. Figure 1 shows a list of pathways on the left side. This list contains all pathways that contain the currently selected focus node, or a node of the same gene family. The pathways in the list are ranked by their similarity to the current focus pathway. We calculate a similarity score for each pathway by computing the number of nodes shared with the focus pathway and normalize it by its size. The score is shown as a bar next to the pathway name. To quickly determine which pathways have already been added to the workspace we mark loaded pathways using a dark gray background. In some situations analysts are interested in pathways that are generally similar to a selected pathway, without choosing a focus node. We use a similar algorithm to calculate scores of pathways in this scenario.

An alternative to the automatic, similarity based list is an alphabetic list which can be searched using keywords or regular expressions. This is especially helpful to find an entry point of an analysis. Finally, since pathway maps often embed related pathways, we enable the adding of such pathways to the workspace by clicking on embedded pathway nodes. The *Melanoma* focus pathway shown in Figure 3, for example, contains six embedded pathways indicating that these pathways play an important role in the context of the focus pathway. One of them (*Cell cycle*) is also a current context pathway, which is indicated by its purple border.

5.3 **Visualizing Connections**

To find path intersections and to enable the identify high-level relationships task we need to visually communicate which portal nodes connect two pathways. This requires visual encodings to (a) convey that a node (either in a focus or in a context pathway) is a portal and to (b) tell the analyst to which other portals it can be connected to.

Since related pathways often contain a substantial quantity of portal nodes, obvious approaches, such as color-coding or drawing visible edges, may easily fail. Even though objective (a) could be addressed by using a color-based highlighting of portal nodes, objective (b) would potentially require assigning many different colors to a single node. Visual links (i.e., visible edges), on the other hand, can connect a node to many others; but can result in significant clutter, given the many nodes and the dense layout of pathway maps, even if they were intelligently routed [28]. Therefore, we have chosen to primarily use stubs to encode relationships between nodes. Stubs were shown to be effective for indicating a connection without cluttering the display [7]. Figure 3 illustrates our stubs implementation. The two insets at the top show them in detail. For each pair of related portal nodes we render a pair of stubs pointing at each other. The direction of a stub thus indicates the location of its target. We attach the stubs to the side of the node closest to the target and we quickly let them fade while they are converging to a point. We also show portals only with respect to the “active” pathway, i.e., stubs only point to and from the pathway on which the mouse pointer rests. This reduces the set of portals, minimizing clutter and ambiguities while showing all relevant connections.

As context pathways only show a subset of nodes, potential portals might not be displayed. Nevertheless, we also want to communicate the presence of hidden portals. To achieve this, we show that a pathway has a relationship to the active pathway through one or multiple hidden portals by placing a stub on its window’s title bar, as shown in Figure 3 (labeled window stub).

As is evident from the figures in this paper, stubs are excellent at indicating connections between many portal nodes without introducing a high amount of visual clutter. However, they can be ambiguous at times, especially when the angle between two stubs attached to the same node is small. To resolve potential ambiguities, we show the exact connections out of a portal node by using visual links when the mouse hovers over the portal (labeled portal links in Figure 3).
we describe how we optimize the arrangement, size, and amount of layout to arrange pathways as opposed to a free layout, since matrix-like makes a high-level comparison of two focus-pathways possible. Promotion or demotion of pathways between these levels of detail and limited. In situations where there is not enough space to show any is typically large enough to convey a sense of the overall topology of the pathway. In order to aid orientation, we highlight the route of the context path(s) in the thumbnail, as shown in the inset of Figure 4. However, we consider this topological information less relevant than the actual context, which is why we omit the thumbnail if space is limited. In situations where there is not enough space to show any context paths, we resort to showing only the pathway titles. While this is not ideal, it is better than removing the pathway, since (a) it can be conveniently brought back into focus or any other level of detail and (b) it still indicates whether there is context information to be shown. Promotion or demotion of pathways between these levels of detail and the focus can be triggered manually but is also done automatically. Automatic actions can be disabled for individual pathways. This also makes a high-level comparison of two focus-pathways possible.

5.4 View Management
Using contextual subsets significantly reduces the number of elements that need to be displayed yet preserves the relevant context. Nevertheless it is prudent to make good use of the available screen space. Here we describe how we optimize the arrangement, size, and amount of data shown in the various pathways under analysis. When optimizing a layout for pathway analysis, one is confronted with a range of partially conflicting goals. The first and most obvious goal is to maximize the amount of relevant content shown. This often conflicts with the goal to ensure legibility of all elements. Following the contextual subsets concept, we always use at least one focus pathway, for which we comply with the legibility goal, thus limiting the remaining space for contextual information. To deal with the varying amounts of space, we promote and demote pathways to various levels of detail and optimize the pathway layout.

Levels of Detail Our approach to efficiently layout pathways requires us to change their size. We achieve this by introducing three levels of detail for context pathways: high, medium, and low, which are illustrated in Figure 4. The thumbnail used in the highest level is typically large enough to convey a sense of the overall topology of the pathway. In order to aid orientation, we highlight the route of the context path(s) in the thumbnail, as shown in the inset of Figure 4. However, we consider this topological information less relevant than the actual context, which is why we omit the thumbnail if space is limited. In situations where there is not enough space to show any context paths, we resort to showing only the pathway titles. While this is not ideal, it is better than removing the pathway, since (a) it can be conveniently brought back into focus or any other level of detail and (b) it still indicates whether there is context information to be shown. Promotion or demotion of pathways between these levels of detail and the focus can be triggered manually but is also done automatically. Automatic actions can be disabled for individual pathways. This also makes a high-level comparison of two focus-pathways possible.

Layout Optimization We decided to use a rigid column-based layout to arrange pathways as opposed to a free layout, since matrix-like layouts are more space efficient when it comes to layouting rectangular shaped objects like pathway maps. Also, a column-based layout is well suited to reflect the history of the analysis process by sorting the pathways by age. Entourage can accommodate as many columns as are reasonable for a given screen resolution, but always enforces at least one context column as well as a minimum width for a column.

Our initial implementation followed the goals outlined above, always aiming to maximize the visible context information while ensuring legibility. Early feedback, however, triggered the realization that another factor is essential: layout stability. We observed that our collaborators were irritated by layout changes, even though they were animated. As a consequence we added the goal of minimizing layout changes. We also found that changing the size of a particular pathway is much less irritating than changing its position, either within or between columns. Consequently, our layout algorithm now prohibits position changes unless the focus pathway is exchanged, but permits resizing and switching between levels of detail of context pathways.

Within a context column, we maximize the vertical space between individual pathway windows. While this might not be as aesthetically pleasing as stacking them on top of each other, this strategy serves a purpose: It helps to avoid ambiguities of stubs pointing to the pathways by increasing the angle between stubs.

To fulfill our goal of maximizing the amount of relevant content displayed we promote and demote pathways intelligently. Automatic demotion of pathways is triggered when the horizontal or vertical display space is insufficient for displaying all elements at a reasonable size, while automatic promotion is triggered as space becomes available. An important decision in this regard is which pathways to demote or promote. This primarily depends on the causes of the space change. For example, if the vertical space is exceeded by the pathways in a context column, only pathways within that column have to be considered for demotion. In contrast, if there is too little horizontal space, the demotion of any pathway can potentially free up space.

To ultimately decide which of the pathways to demote or promote, we use three attributes of various priorities. The highest priority is given to pathways that contain a user-selected path (see Section 6). The second-highest priority is given to pathways that currently contain context paths. Finally, pathway “age” is considered as the lowest priority, where “young” pathways, i.e., those that were recently in focus, are given priority. We calculate a ranking of the candidate pathways based on these attributes and eventually demote the pathway with the lowest priority.

Taken together, our layout considerations guarantee a stable and predictable management of many pathways.

The techniques and encodings discussed in this section allow analysts to take a detailed look at one pathway while always keeping an eye open for cross-connections to other pathways. By showing only the information relevant to the current analysis, the important parts of the data can be shown at full scale. We thus provide an analyst with the necessary tools to address the Pathway Interconnectivity tasks.

6 Experimental Data Analysis
So far we have focused on how to visualize relationships between pathways considering only the pathways and the underlying network. In this section we will introduce (a) how experimental data can be leveraged to select interesting pathways and cross-connections in the first place and (b) how to visualize experimental data in the context of pathways and pharmacologic data. We will thereby address the two Pathway-Experimental Data Linking tasks.

The first of these tasks is to identify subsets of pathways that warrant detailed investigation based on experimental data. To accomplish this task, we need to provide information on which nodes are interesting in terms of the mapped experimental data. The most common approach to supplement pathways with experimental data is to color-code the nodes [20]. Other approaches include small bars, line plots, etc. For a comprehensive analysis refer to the review by Gehlenborg et al. [8]. All of these approaches aim at encoding experimental data on top of the pathways. However, such attempts are futile when dealing with large and heterogeneous datasets, but are helpful for single, homogeneous datasets [24]. Consequently, we make it possible to map individual
datasets by color coding the nodes, if desired (see Figure 5), but by default take a different approach: We point analyst to parts of a pathway that are either interesting for exploring the underlying experimental data or that are relevant to consider in other pathways. We do so by calculating the standard deviation of experimental data associated with each gene. If this deviation is higher than a threshold, we show an exclamation mark, as shown in Figure 5. The color of the exclamation mark encodes the dataset where the deviation was observed, which is also used in the data mapping view (see Figure 1). If multiple datasets show a large deviation, we encode only the largest. We chose a glyph since it nicely supplements the color-coding of nodes we use for showing average values of a single, selected dataset. This feature addresses the aforementioned task well, as typically variability in the data is of most interest.

For visualization of the actual underlying experimental data we employ the enRoute technique [24], which is part of the Caleydo framework. enRoute requires analysts to select a path in the network for which detailed experimental data is shown. Selected paths are highlighted using Bubble Sets [6]. Figure 5(a) shows a simple example of a selected path, the resulting extracted path is shown in Figure 5(b). Entourage always keeps track of the selected path in the selected path view, shown in Figure 1 on the right. By default, only the path is shown, but this view can be expanded to show enRoute, as demonstrated in Figure 7. Notice, that the enRoute view can also be shown full-screen, thereby occluding the pathways but giving more space to the experimental data analysis. Which data and which stratifications (groupings) of experiments are shown is driven by analyst choices made in the data mapping view shown in Figure 1 at the bottom.

While the original enRoute technique can only be used for paths in a single pathway, Entourage is ideally suited to select paths across pathways, as is shown in Figure 1. Notice that pathway boundaries are included in the path representation. We chose not to extend the bubble sets across pathways but instead use the visual links we also use for portals, as the connecting portal nodes are in fact the same node.

Finally, to address the identify relationships between cell line responses to drug treatment and genomic data task, we extended enRoute to show contextual data that is not associated with genes. Such data is shown above the gene-associated data and uses the same ordering of samples. Figure 6 shows the compound sensitivity of ovary CCLE cancer cell lines to the drug AEW541 on top of the expression (on the left) and copy number values (on the right) associated with the RAF gene family (BRAF, ARAF, RAF1). Here, low bars indicate high sensitivity, i.e., low $IC_{50}$ values. Notice that the samples are sorted with respect to their sensitivity to the compound, which is a simple yet effective way to search for relationships between genomic and pharmacologic data. This visual encoding can successfully address the task of associating cell line responses and genomic data. Moreover, since, for the small set of compounds the targeted processes and genes are known, it is easy to identify paths where interesting relationships between genomic and pharmacological data occur.

### 7 Implementation and Scalability

Entourage is part of the open source Caleydo Biomolecular Data Visualization Framework. Caleydo is implemented in Java and uses JOGL [9] for rendering. Entourage will be a part of the next Caleydo release. We use an adapted version of the freely available implementation of Bubble Sets [17] for highlighting selected paths. Our Entourage implementation works with pathway maps from the KEGG [13] and WikiPathways [14] databases (see supplementary material for examples). Although we use the layouts provided by these databases in our current implementation, our technique is not limited to those and can equally be applied to automatically generated pathway layouts.

Depending on the size of the current focus pathway, Entourage can display up to ten pathways simultaneously, where one pathway is the focus pathway while the other pathways are at least in “medium” level of detail on a full HD notebook display (see supplementary material for examples). On larger, higher resolution screens, this number increases. The space for pathways can be increased by hiding currently unused support views (the data mapping, pathway list, and path views shown in Figure 1). Feedback from our collaborators indicates that this number of simultaneously explorable pathways is sufficient in all but the rarest cases and superior compared to other systems. We believe that our visual encodings are also suitable to point at interesting relationships outside an analyst’s primary field of view, making it suitable for the increasingly large displays that are becoming commonplace. On conventional displays we typically limit the number of focus pathways to one in order to guarantee readability. This number, however, can be temporarily increased if detail about the structural relationship of pathways should be shown.

### 8 Case Studies

Entourage was developed in a user-centered design process including weekly meetings between the visualization developers and multiple domain experts. As a result of these meetings we have established the 3 domain goals and the 6 analysis tasks. We deployed various iterations of Entourage and our primary contact, a biochemist, who is also an author of this paper, used Entourage over a period of four weeks. During this period we were in constant contact with her and refined various aspects of the system. The case studies presented here report on her observations. Prior to the deployment of Entourage the team was using conventional pathway tools and had to resolve any questions concerning pathway relationships manually. For visualization of experimental data they mainly relied on tools like TIBCO Spotfire [30], whereas interactions between biomolecules were analyzed with network visualization tools like Cytoscape [27] or the KEGG web interface. Hence, Entourage allowed them to integrate two analysis steps that were previously carried out separately into one single task.

In the following we describe case studies for the domain goals concerning drugs’ mechanisms of action and drug repositioning, which we found representative for demonstrating Entourage’s functionality. The case studies either describe a novel observation or clearly demonstrate how a known effect can be rationalized with Entourage.

---

1http://caleydo.org
While this observation was made for breast cancer tissue, exploring the related context pathways by setting the focus node to Ras to explain different compound sensitivities of cell lines, our collaborator largely overlapped, although Lapatinib showed a broader spectrum of activity than Erlotinib. She found a strong co-occurrence between ErbB2 mRNA over-expression and sensitivity to Lapatinib in lung and breast cancer cell lines, a trend that was less apparent or not observed at all for other responsive cell lines.

8.1 Relating Genomic Features to Compound Sensitivity

To explain different compound sensitivities of cell lines, our collaborator used the previously introduced CCLE dataset. This dataset contains data on the inhibitory effects of 24 drugs against roughly 500 cell lines from different cancer tissues and genomic data. Ideally, a drug completely inhibits the growth of these cell lines at minimal concentrations. First, she wanted to investigate factors that sensitize cell lines to the drugs Lapatinib and Erlotinib that inhibit members of the ErbB gene family and are used in cancer treatment. The ErbB family is a family of epidermal growth factor receptors that are known to play an important role in tumor growth. The drug Lapatinib is a dual inhibitor of EGFRL and ErbB2, while Erlotinib is a known inhibitor of only EGFR, all of which belong to the aforementioned family. Due to its immediate relevance, the expert started by loading the ErbB signaling pathway into Entourage. By searching for related pathways she found several cancer-specific pathway maps. The pathways Glioma and Non-small cell lung cancer ranked among the top on the list (see Figure 7). She commented that this indicates that the ErbB signaling pathway is a key player in these diseases. For the ErbB pathway map, our collaboration partner was interested in the experimental data for the genes in the path that leads from ErbB receptors to MYC, a gene known to regulate cell growth. She also noticed that ErbB2 was highlighted with a red exclamation mark indicating high variance in the copy number data. She thus selected the genes of this path for an in-depth analysis. She then looked at this path’s gene expression data in the embedded enRoute view and combined it with sensitivities to Lapatinib. For the analysis, cell lines were grouped by their tissue of origin (e.g., breast, ovary, liver, etc.) and sorted by sensitivity to Lapatinib. Her first observation, when looking at the experimental data, was that the two drugs displayed inhibitory activities across cell lines from many different tissues. The cell lines from lung, breast and three other tissues were in general most responsive. The set of cell lines that were responsive to Erlotinib and Lapatinib largely overlapped, although Lapatinib showed a broader spectrum of activity than Erlotinib. She found a strong co-occurrence between ErbB2 mRNA over-expression and sensitivity to Lapatinib in lung and breast cancer cell lines, a trend that was less apparent or not observed at all for other responsive cell lines.

She then chose to focus on cell lines from breast and also investigated copy number variation for these cell lines. For most breast cancer cell lines that over-expressed ErbB2, high copy numbers of this gene were found, i.e., the increased expression could generally be traced back to an increased copy number. Interestingly, only two breast cancer cell lines that showed strong over-expression of ErbB2 did not respond to Lapatinib treatment. The columns of these two cell lines are highlighted gold and orange in Figure 7. She then tried to find the cause for this effect and examined the expression of downstream genes in the pathway. She found that for these two cell lines, the gene Ras was strongly under-expressed (also shown in Figure 7).

It is straightforward to assume that this under-expression further down the path counteracts the over-activation of the pathway by increased ErbB2 expression, explaining the resistance to Lapatinib treatment that reduces the effects of ErbB2 expression. Our collaborator stated that this highlights the importance of being able to analyze genomic data in a pathway context because compound sensitivities can often only be explained by the interplay of multiple genomic features.

Based on this observation she started to investigate whether other cancer-related pathways contain the same signaling cascade, i.e., path. She selected Ras as the focus node of her analysis, which revealed several other pathways that contain the same ErbB signaling cascade. Figure 7 shows an example where it is obvious, due to the many parallel stubs that the cascade is indeed identical. Our collaborator commented that it would interesting to investigate in the future whether a similar gene expression pattern in these cancer types would also entail resistance to Lapatinib.

8.2 Rationalizing Successful Drug Repositioning

Graft-versus-host disease (GVHD) is frequently observed after tissue or organ transplantation and is caused by immune cells that originate from the donor and were transplanted with the tissue. These immune cells perceive tissue of the recipient as foreign and attack it, thereby causing damage. The Graft-versus-host disease pathway identifies TNF-alpha, a gene involved in inflammation, as an important player.
in the disease. Accordingly, molecules counter-acting (inhibiting) the effect of TNF-alpha have been evaluated for preventing GVHD in transplantation patients, with no success. However, the roles of TNF-alpha in the organism are manifold, as our collaborator was able to demonstrate using Entourage, when she chose TNF-alpha as her focus node, revealing all associated pathways. She found that one of the highest scoring and therefore most similar pathway to the Graft-versus-host disease map is the Rheumatoid arthritis pathway, shown as the focus pathway in Figure 8. Indeed, as clinical safety for TNF-alpha inhibitors had been proven in the initial trials for GVHD patients, the molecules were revisited and tested for their efficacy in patients suffering from rheumatoid arthritis. In this case, anti-TNF alpha therapy showed the desired clinical effect and today TNF inhibitors are part of the standard treatment of rheumatoid arthritis. The domain expert pointed out that Entourage also ranks the Apoptosis, MAPK signaling, and NF-kappa B signaling pathways, which explain the controversial role of TNF-alpha in cancer. While the Apoptosis pathway shows the process by which TNF-alpha leads to cell death, the other two pathways point out how TNF-alpha contributes to cell survival. Accordingly, the benefit of TNF-alpha inhibitors in anti-cancer treatment remains an open question and clinical trials are awaited to further explore the potential use of these molecules in malignancies.

They highly valued the ability to immediately see all relevant related processes for a pathway and being able to compare them easily and to see experimental data in the context of pathways.

We have demonstrated the utility of Entourage in two case studies highly relevant for pharmacological research. These case studies reflect current needs of pharmaceutical research, but we believe that our technique is equally applicable in domains such as systems biology or general molecular biology, as interconnections between pathways influence virtually all domains involved with biomolecular data.

Moreover, we argue that the contextual subsets approach can be applied to general graph analysis. It is conceivable that automatically created clusters of a graph can be used instead of manually partitioned pathways. We also believe that several other aspects presented in this paper can be generalized to other visualization applications. In particular, our methods for visualizing relationships could be used for supplementary relationships in graphs, while our view management approach is applicable to all techniques using flexible multi-view setups.

In the future we aim to investigate how to represent compound data and its influences on the biological network. For the CCLE data, visualization of protein-compound interaction, for example, is irrelevant, since the compounds covered by this dataset are few and well understood. There are, however, similar datasets being created that contain data for hundreds or even thousands of compounds, about which there is only limited knowledge available. Early research in this area [21] highlights the important role visualization can play in this domain.

Another potential future line of inquiry is that of comparative analysis of multiple paths. Consider an example where two branches converge into a single node. Current visualization techniques are either not able to deal with the quantity of node attributes necessary to conduct a sensible analysis or fail to represent the topology efficiently, opening opportunities for interesting visualization research.

9 Conclusions and Future Work
Analyzing relationships between pathways to accommodate analysis scenarios such as accounting for pathway cross-talk, repurposing drugs, and relating genomic features to drug sensitivity are challenging and to date unsolved problems. Previous approaches have aimed to show all considered pathways as a whole or even tried to represent the whole network at the same time. We have argued that doing either is not scalable and at the same time unnecessary. Our approach uses a strict focus and context approach, where the focus as well as the relevant context is presented at any given time. We use a combination of carefully chosen visual encodings and analytical support to help experts find the important parts of their data.

Overall, our collaboration partners were excited about the analytical capabilities of Entourage and mentioned that they perceived a significant improvement over their previous tool-chain. While, for example, the KEGG interface could be used to conduct an analysis similar to the one described in the second case study, doing so would be very tedious, as KEGG provides no support for analyzing relationships of pathways.


