Eurographics/ IEEE-VGTC Symposium on Visualization 2010
G. Melançon, T. Munzner, and D. Weiskopf
(Guest Editors)

Volume 29 (*2010*), *Number 3*

Pathline: A Tool For Comparative Functional Genomics

M. Meyer[1,2], B. Wong[2], M. Styczynski[3], T. Munzner[4], and H. Pfister[1]

[1]Harvard University, USA
[2]Broad Institute, USA
[3]Georgia Institute of Technology, USA
[4]University of British Columbia, Canada

Abstract

Biologists pioneering the new field of comparative functional genomics attempt to infer the mechanisms of gene regulation by looking for similarities and differences of gene activity over time across multiple species. They use three kinds of data: functional data such as gene activity measurements, pathway data that represent a series of reactions within a cellular process, and phylogenetic relationship data that describe the relatedness of species. No existing visualization tool can visually encode the biologically interesting relationships between multiple pathways, multiple genes, and multiple species. We tackle the challenge of visualizing all aspects of this comparative functional genomics dataset with a new interactive tool called Pathline. In addition to the overall characterization of the problem and design of Pathline, our contributions include two new visual encoding techniques. One is a new method for linearizing metabolic pathways that provides appropriate topological information and supports the comparison of quantitative data along the pathway. The second is the curvemap view, a depiction of time series data for comparison of gene activity and metabolite levels across multiple species. Pathline was developed in close collaboration with a team of genomic scientists. We validate our approach with case studies of the biologists' use of Pathline and report on how they use the tool to confirm existing findings and to discover new scientific insights.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

Biologists conduct comparative functional genomics studies in order to infer the evolution of gene regulation, and to understand how this evolution is linked to changes in gene activity. These studies focus on the *functional* outputs of genomes, looking at similarities and differences in gene activity over time and across species. Subsets of the genes work together in *pathways*, which represent specific chain reactions that occur in the cell. For example, many functionally related genes are involved in a cell signaling pathway that carries stimulus from the outside of the cell to the inside. Researchers use these pathways to focus the scope of their scientific inquiry and to gain a greater understanding of biology.

We work with a team of researchers who study a collection of pathways that make up metabolism in yeast. These pathways consist of specialized gene products that process chemical compounds called *metabolites*. The researchers need to analyze the levels of gene activity and metabolites belonging to multiple pathways over time and across multiple species. Their visualization needs were not met by currently available tools. Many tools focus on showing the topological structure of pathways, and can show only one experimental data point for each gene and metabolite [SMO*03,AHP*05,LYKB08]. Recently developed tools can show an entire set of values, for example a full time series, for each gene and metabolite [BMGK08, KPR02, JKS06, MSH*05, BW09]. Our collaborators, however, need to look at multiple sets of values simultaneously — across time, species, genes and metabolites — in order to compare trends between species.

In contrast, many visualization tools allow the comparison of multiple sets of functional experimental data points but do not attempt to show pathways. The heatmap visual encoding is popular with biologists, showing a matrix where a single gene activity value is encoded by color in

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and
350 Main Street, Malden, MA 02148, USA.

each cell, usually with rows representing different genes and columns representing species, time points, or treatment conditions [ESBB98, Sal04, SS02]. The ordering of rows and columns is often determined by, and shown with, a hierarchical structure such as a phylogenetic tree or a hierarchical clustering of the data. These tools, however, show only a single value for each cell rather than an entire time series. Moreover, the lack of a pathway representation limits the ability of researchers to extract meaningful insight about systemic biological questions from gene activity data [SND05].

To fill this gap, we present Pathline, an interactive visualization tool that shows time series data for both gene activity and metabolite levels over multiple pathways and multiple species. We present a characterization of the data our biology collaborators are studying, as well as a translation of their biology questions into data-centric tasks. In supporting these tasks, we make two novel visual encoding contributions. The first is a method for linearizing metabolic pathways for a visually concise overview that provides appropriate topological information and supports the comparison of quantitative data along pathways. The second is the curvemap detail view, depicting time series data with small multiples of filled line charts and overlaid curves that support comparison of gene activity and metabolite levels across multiple species. We validate our approach with case studies from our biology collaborators who used Pathline to both confirm existing knowledge and discover new scientific insights.

2. Biological Background and Data

Researchers at the Broad Institute study 14 species of yeast that span over 300 million years of evolution. They study genes involved with metabolism. *Metabolism* is a complex *network* of chemical reactions essential in all living organisms that allows the organism to grow and reproduce, maintain cellular structures, and respond to environmental conditions. The metabolic network is remarkably similar across species in terms of the reactions that it comprises. However each organism uses, and controls the use of, the reactions slightly differently. For example, the same gene may be turned on earlier in one species, versus later in another. Or, some species may have developed a different control mechanism for a particular gene, or have evolved a novel gene altogether to accomplish a certain metabolic task. These changes are hallmarks of the evolutionary process.

In metabolism, chemical compounds called *metabolites* are catalyzed from one form to another by the actions of *enzymes*, which are a type of gene product. Enzymes involved in metabolism can work in one of three directions: forward, moving a metabolite ahead a step; reverse, moving a metabolite back a step; or bidirectional, capable of catalyzing both forward and reverse reactions. The metabolic network is subdivided into *pathways*, which are a small set of related reactions that may contain cycles and branches. The products of one pathway may be the starting material of another.

The specific comparative functional genomics study of our biology collaborators depends on the following four main categories of data.

Gene activity and metabolite levels are measured for approximately 6,000 genes and 140 metabolites for each of the 14 species of yeast. Each gene and metabolite is measured at six physiologically relevant time points. While the number of species to study may grow in the future, the number of genes, metabolites, and time points is fixed due to the nature of the biological processes being studied. Each measurement of gene activity or metabolite level has three associated attributes: the time point, the name of the species, and the name of the gene or metabolite.

Metabolic pathway information in the form of a directed graph is taken from the publicly available BioCyc database [CFF*08]. Graph nodes are metabolites, and the edges represent small sets of genes, the products of which catalyze the reactions. The number of reactions in any given pathway studied by our collaborators is small, usually around a dozen. Researchers typically choose only a handful of pathways to look at simultaneously, filtering the full gene and metabolite dataset down from several thousand to several dozen.

Similarity scores, which are high-level aggregate scores of time series similarity, are computed for each gene or metabolite across a set of multiple time series. The set can either be all species, or any subset of species that interests the researcher. Currently our collaborators use the standard Pearson and Spearman correlation functions as direct similarity metrics and compute an aggregate metric of average pairwise similarity within a set.

Phylogenetic relationships showing the ancestral relationships between yeast species are also important data used in the analysis process. These relationships are represented by a tree where the leaf nodes are the living species and the internal nodes indicate speciation events, meaning a common ancestor that eventually gave rise to two distinct species.

3. Tasks

The level of gene activity, analogous to enzyme levels, and the level of the metabolites, change over time. Finding differences between species in the patterns of the these changes is an important part of comparative functional genomics. At the high level, understanding these differences will allow biologists to extrapolate the functioning of extant species (*i.e.*, those species that exist today) to that of their ancestors. They can then infer the evolution of specific cellular processes and of regulatory mechanisms in the genome.

More specifically, our collaborators would like to identify when different yeast species regulate metabolic processes the same way, either across all 14 species or by finding subsets of species that behave similarly. They look for trends and try to determine similarity between time series across a set of species at specific genes or metabolites along one or several pathways. The biologists need to carry out tasks at three levels:

- Detailed comparison of a limited number of time series.
- Aggregate comparison of the similarity score of genes and metabolites across a pathway.
- Multiple similarity score comparison.

At the detailed task level, the analysis involves inspecting the full time series of a small subset of carefully chosen genes and metabolites. The specific tasks at this level are:

- Look for trends in a set of time series for a gene or metabolite across species.
- Look for trends in a set of time series within a species for several genes and metabolites.
- Compare time series to find:
  - valleys that exist in some but not others.
  - time series that are a time shift of another, such as early peaks versus late peaks.
  - the most similarly shaped time series to one of interest.
  - which time series in a group are the same.
  - how many classes of time series exist in a set.

For the aggregate task level, researchers look at a single number for each gene and metabolite that represents their similarity across all of the species, or a subset of the species. This similarity score is an aggregation of the underlying set of time series for each gene and metabolite.

Researchers compare multiple possible aggregations at the third task level, looking for the differences between biologically interesting subsets of species and the combination of all of them. For example, they often look at three numbers for each gene and metabolite: an aggregation across all species, across one subset, and across another subset. They are interested in discovering when the members of the two subsets are similar themselves, but the entire set is not. They are also interested in discovering when only members of one subset are similar. At this level they also want to compare the results of different similarity metrics. They suspect that purely statistical methods such as the Spearman or Pearson correlations currently used to compute similarity scores, which take a shape-to-shape matching approach, may not expose meaningful biological relationships between time series. They would like to compare them with other more biologically inspired metrics in the near future.

Validating and understanding the results at the aggregate and similarity score levels requires frequent cross-checking with a detailed visual comparison at selected points.

4. Pathline

Pathline, shown in Figure 1, is an interactive prototype tool that supports the visual analysis of all four kinds of data and all three kinds of task levels discussed above. The tool was designed in a user-centered process with iterative refinement. Our design decisions were motivated by the specific needs of our genomics collaborators.

Using Pathline begins with loading a set of pathways and their associated gene and metabolite data. This choice implicitly filters the full gene and metabolite data set of 84,000 possibilities (from 6000 genes across 14 species) down to a much smaller set of a few hundred, as is the case with all viewers where a set of pathways is chosen by the user. The interface has two linked components:

Curvemap: This detail view shows a small-multiple matrix of filled line charts of time series data. Overlay columns on the right and the bottom show an overlay of all the curves for each row and column, respectively. The rows show the 14 species of yeast, ordered according to the phylogenetic tree shown to the left of the curvemap. The columns show the genes and metabolites chosen by the user in the order that they were selected.

Linearized pathways: This overview is a vertical strip showing the chosen pathways as grey segments, placed end to end. Each pathway segment contains the aggregate similarity scores for the genes and metabolites in the pathway, encoded with horizontal spatial position. The metabolites are encoded as lines, and the genes are encoded as points that are colored according to directionality. Up to three similarity scores can be viewed simultaneously. The pathways have been linearized to create an ordered list of genes and metabolites. Selecting a gene or metabolite adds a column showing all of its underlying data in the curvemap detail view.

We now describe each of these novel visual encodings in more detail, including justifications for our design decisions.

5. Curvemap Detail View

Each column in the curvemap detail view correspond to a single aggregate similarity score shown in the pathways overview. More specifically, each column corresponds to a selected gene or metabolite and shows the measured values as time series curves for all six time points. Each row shows the data for each of the species. The number of genes and metabolites that can be shown is limited by the screen resolution, and is typically between 4 and 15 columns.

The name *curvemap* alludes to its semblance to *heatmap*. Both have a matrix representing species and genes using rows and columns. A heatmap encodes a single value with color in each matrix cell. A curvemap shows a full time series curve in each matrix cell, encoding multiple values with spatial position. The phylogenetic tree to the left of the curvemap is analogous to the widely-used combination of trees with heatmaps: it shows the hierarchical structure of ancestral relationships that leads to the ordering of the species rows.

There are two obvious ways to extend a matrix view from showing one value per cell to showing multiple data points in each cell. In the language of Keim and Kriegel [KK94], who cast the problem as high-dimensional data visualization, the choice is to use multiple simple matrices side by side, with each matrix showing all the information about a single dimension; or use a single matrix with a more complex glyph in each cell, showing the values for the additional dimensions contiguously at each point. They call this latter approach *grouping*. In the cartographically-inspired language

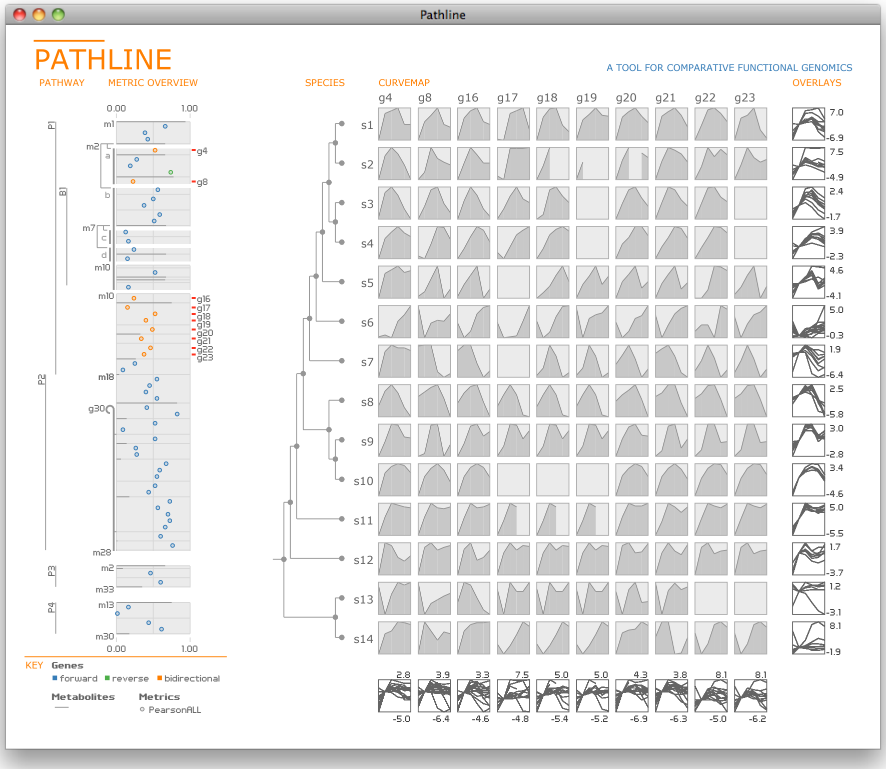


Figure 1: *Pathline, an interactive tool for the visualization of comparative functional genomics data. The left side shows the linearized pathways, whereas the right side shows the curvemap. Here, four different pathways are shown. This image, as well as all other images in this paper, can be found online at* `http://www.pathline.org/`.

of Slingsby *et al.* [SDW09], the question is which attribute to *condition* on at the deepest level of the multivariate attribute hierarchy: time point, species, or gene/metabolite.

The curvemap design is an example of grouping: time point conditions the deepest level of the hierarchy because the basic biological unit for comparison used by our collaborators is the time series of gene activity and metabolite levels. Before using Pathline, the biologists tried a heatmap-oriented version of grouping using TreeView [ESBB98, Sal04], which supports a nested mini-heatmap within each cell of the larger enclosing heatmap. Many of their tasks, however, were difficult to carry out using this representation.

The nature of the tasks carried out in the detailed analysis by the comparative functional genomics researchers drove the design decision to use curves rather than colored blocks. As discussed above, people can make more accurate absolute perceptual judgements for spatial position than color [CM84]. Moreover, people can make judgements about curve shapes that are far more subtle than those about color changes [LMK07]. The very language used by our collaborators to describe their detailed analysis tasks in terms of *peaks* and *valleys* reflects this sort of spatial thinking.

The curvemap uses filled line charts, also known as area plots, for the time series curves in the main matrix. We shade the area under the curve in dark grey against a background of lighter grey in order to make the shape of the curve more perceptually salient than with a line alone. (We note that the area under the curve is not directly meaningful with respect to the data.) We also surround the plots with a bounding box to create the clear perception of positive and negative space. Each time series curve is individually normalized such that the minimum value meets the bottom of the plot and the maximum value reaches the top. This normalization supports the comparison of the shape of the curves and the trends in the time series.

In addition to the main matrix, the curvemap view includes overlay multiples where all the curves for each column and row are superimposed in a single shared frame. These plots support the detection of trends in each column and row of curves. We normalize them on a per-column and per-row basis to provide an absolute scale that is important to understanding the differences across genes, metabolites, or species. The gene and metabolite measurements are unitless *fold-change* values that show whether the levels at one time point went up or down as compared to a baseline measurement. The biologists chose the second time point for their baseline because the yeast gene activity data are most robust to experimental error at this point. All of the curves thus cross at the second time point in these overlay multiples.

6. Linearized Pathways Overview

The linearized pathways overview is a high information density display showing multiple quantitative similarity scores for each gene and metabolite over multiple pathways. The design of this view focuses on supporting comparison of quantitative values along the pathways, a notable difference from previous systems that focus on the task of understanding pathway topology.

6.1. Linearization

Through discussions with our biology collaborators, we came to understand that their analysis calls for a schematic view that emphasizes a linear ordering of pathway elements, with topological information available at a secondary, rather than a primary, level. As discussed in Section 3, at the aggregate analysis levels the researchers want to understand when gene activity along specific pathways are similar between species. This type of inquiry requires only a coarse understanding of the pathway topology.

Many previous systems, however, include a detailed visual representation of a pathway's topological structure as a node-link graph, requiring a significant amount of screen real estate. Thus many systems, such as GeneShelf [KLK*09], show only a single pathway at a time, making comparison between multiple pathways and multiple metrics difficult. This restriction requires the user to navigate between multiple screens, each showing a single pathway, and to rely on memory to perform comparative tasks. In contrast, we designed the pathway overview to be a visually concise display supporting inline comparisons, which has been shown to be more perceptually effective than relying on the user's memory of what has been seen before [PW06].

As such, a fundamental design decision for the pathways overview is the transformation of each pathway from a directed graph to an ordered list of genes and metabolites. This linearization allows for a shared axis along which quantitative information is visually encoded using spatial position, a more perceptually accurate encoding than color [Mac86, CM84, LMK07].

Figure 2 shows the the logic behind the linearization process for a pathway that contains both a branch and a cycle. As we discuss in Section 2, the nodes are metabolites and the edges represent small sets of genes. Figure 2(a) shows a node-link representation of the pathway. In Figure 2(b) the cycle is unrolled and the branch is disconnected. The branch is then reinserted just above its reconnection point in Figure 2(c). This process is carried out recursively as branches can be nested.

Pathline requires a manually created ordered list of genes and metabolites as depicted in 2(c) as input, and produces the visual representation shown in 2(d). Our collaborators created linear pathways based on graphs from the BioCyc database using the set of rules described above. It is up to the users how to handle situations such as multiple metabolites at the starting point, or overlapping cycles, so that the ordered list best reflects their mental models of the data.

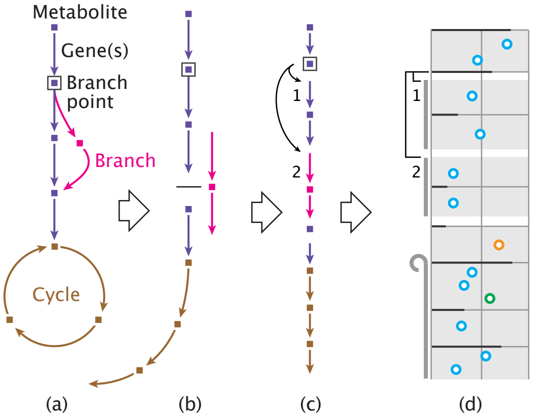


Figure 2: *Linearizing a pathway.* ***(a)*** *The node-link representation of the directed graph includes both a branch and cycle.* ***(b)*** *Loops are unrolled and branches are disconnected.* ***(c)*** *Branches are reinserted just above their reconnection points.* ***(d)*** *The pathway is represented as a grey segment, with genes encoded spatially with points and metabolites as lines. Short breaks in the pathway segment indicate branch points, along with stylized marks to the left of the blocks. Cycle start points are also shown to the left with another mark.*

In Figure 2(d), the entire pathway is rendered as a grey segment, with the individual genes and metabolites shown using different types of marks: points for genes and lines for metabolites. The directionality of the enzyme encoded by each gene is shown using color: blue for reverse, green for forward, and orange for bidirectional. The short breaks in the grey segment indicate branch points with additional marks to the left of the grey blocks indicating the topological structure of the branch. At a branch point, the two possible paths are shown with stylized branch marks and letter labels, along with vertical bars marking the extent of each branch alternative. A circular icon is shown to the left of the gene or metabolite at a cycle start point and a similar vertical bar shows the extent of the cycle.

In the linearized pathway overview in Figure 1, the extent of an entire pathway is indicated with a vertical line on the far left labelled with the pathway name. Some pathway segments, such as major branches, also have names, which are shown in the same way. The names of the genes and metabolites at the start of segments, at branch points, and at the start and stop points of a cycle are shown on the left. These names create visual landmarks that reflect the way the biologists typically refer to segments that do not have canonical names. Pathways that are not connected by a direct reaction are visually separated with longer white breaks between them.

The example in Figure 1 has four pathways labelled vertically on the left: `P1`, `P2`, `P3`, and `P4`. The `P1` pathway branches at metabolite `m2` into the segment marked `a`, and an alternative branch marked `b` that is also labelled with its name, B1. This branch contains another nested branch starting from metabolite m7, where the alternatives are labelled c and d. The P1 pathway runs directly into the P2 pathway, which includes a cycle. The cycle in P2 starts at g30 and ends at m28. The P3 and P4 pathways have no direct linkages to their preceding pathways.

6.2. Similarity Score Display

The horizontal extents of the grey pathway segments encode the quantitative data of the aggregate similarity score for each gene and metabolite, as shown in Figures 1 and 2(d). Up to three metrics can be displayed at once, visually encoded as an ordered bar-chart for metabolites, and as different shapes for genes. The three shapes used for the genes are a hollow circle, a plus sign, and a smaller filled-in rectangle. We hand-tuned the colors and shapes to have roughly equal visual salience.

Figure 5(a) shows the encoding of three different metrics. The PearsonALL aggregates over the entire set of species, whereas PearsonSubgroup1 and Pearson-Subgroup2 are metrics computed for two biologically interesting subsets of species. When more than one metric is shown, they are linked with a low-saturation bar that stretches between the two points. This was scientifically motivated as the researchers want to directly compare two subsets of species and use a third metric as a reference. For this reason we explicitly show the difference between two metrics.

We designed the visual encoding with mark type, mark shape, and color to create visual layers, allowing for selective attention when just one attribute type is interesting, or perceived together as a whole when the full context is required. Our collaborators need to see both the genes and metabolites along the pathways, but may need to focus on one or the other when looking for certain trends. A similar situation holds for all genes versus those of a particular directionality, or when comparing different metrics for aggregating similarity scores.

7. Interaction and Implementation

Pathline is built using multiple views linked through explicit clicks and lightweight mouseover interaction, following the general tradition of many previous information visualization systems [MMP09, SS02].

In the linearized pathways overview, mousing over the pathway elements shows the metadata of their name and the numerical value for the metric. Clicking on the point representing the aggregate similarity score for an element selects the underlying data for further investigation as a full column of time series curves in the curvemap detail view. In the overview where the click occurred, that element is highlighted with a small red bar drawn to the right of the segment, along with the name of the element. In Figure 1, 10 elements are selected across multiple pathways.

In the curvemap detail view, mousing over the name of a species in the phylogenetic tree highlights the curve associated with it in the overlay plots at the bottom of the main matrix, and mousing over a subtree highlights all of its associated curves. Similarly, mousing over a label at the top of a curvemap column highlights the associated gene or metabolite curve in the overlay plots to the right of the main matrix. This latter interaction also highlights the selected gene or metabolite name in the pathways overview.

The number of viewable genes and metabolites that can be added to the curvemap view and the size of the plots is determined by the window size. A vertical scrollbar appears in the pathway overview if the window is too short to render all of the pathways at once.

Pathline was implemented in the Processing language [RFM07]. Executables, source code, and example data are available at http://www.pathline.org.

8. Previous Work

We divide the most related previous work into general time series, networks and pathways, heatmaps, and genomics-specific time series visualization.

8.1. General Time Series Visualization

Several previous information visualization systems tackle the problem of visualizing time series data. TimeSearcher [HS01] offers good support for exploring a few long series, and for finding patterns within them. In contrast, the problem we address is to inspect a large collection of very short series. LiveRAC [MMKN08] is designed for a similar dataset, but the semantic zooming and guaranteed visibility techniques that support large-scale system management tasks are not appropriate for the comparative functional genomics analysis tasks presented in Section 3.

8.2. Network and Pathway Visualization

Cytoscape [SMO*03] is a leading example of a system designed to show the detailed topological structure of large biological networks made up of many pathways interconnected in complex ways. In contrast, Pathline addresses analysis tasks where only a small number of pathways are under consideration at one time, and their topological structure is a secondary, rather than primary, concern.

Many pathway visualization tools use color encoding to show a single experimental data value at each node, including Cytoscape [SMO*03], MicrobesOnline [AHP*05], and iPath [LYKB08]. Others show an entire set of experimental data values at each node. Cerebral [BMGK08] uses linked small multiples, with a separate node-link graph shown for each time point whose nodes are colored by the value at that time. Pathway Tools [KPR02] encodes a single data value with color at each pathway node, and uses animation to show the entire set of values. Several tools use a glyph to encode a set of values at each node, including VANTED [JKS06], PathwayExplorer [MSH*05], and GENeVis II [BW09].

All of these tools focus on graph layout, overlaying experimental data on a node-link graph representation where nodes are distributed in space to emphasize the connectivity relationships of nodes and edges. This representation is less useful for the tasks of our collaborators, where the goal is to accurately compare values of, and look for trends in, the experimental data. Thus, in Pathline we instead treat the experimental data as primary structure that drives spatial positioning, and relegate the topological structure to secondary status. Moreover, these tools are also limited to showing only a single set of data points at each pathway node, while our collaborators want to analyze multiple sets of data points.

8.3. Heatmaps

A different set of visualization tools focus on the task of comparing multiple sets of experimental data points; for biological data the most common visual encoding is a heatmap [WF09]. Heatmap visualizations are often coupled with clustering algorithms, as in TreeView [ESBB98,Sal04], the Hierarchical Clustering Explorer [SS02], and Genomica [LS10]. As discussed in Section 5, heatmaps show only a single value at each cell, and extending them to encode multiple values is nontrivial for perceptual reasons. Also, they do not explicitly show pathway information. Past work [SND05] has shown that the ability of scientists to extract biologically meaningful insights from gene activity data is severely hampered by the lack of this kind of contextual information.

8.4. Genomics Pathway and Time Series Visualization

GeneShelf [KLK*09], designed to be a lightweight web tool for exploring large public gene expression databases, handles data most similar to that supported by Pathline. The user selects a single pathway, which the system shows side by side with the time series data for the genes contained within the pathway. The time series data are shown using a small multiples matrix where each axis is an experimental condition, and each grid cell has a parallel coordinates view of the gene set over the time points. Any time point in a parallel coordinates view can be expanded to show a bar chart of the expression level of all the genes. Extending this approach to the time series curves that we show would disrupt the shape perception required for many of the detailed analysis tasks of our collaborators. As we discuss in Section 6.1, seeing only a single pathway at once is a limitation of GeneShelf that we explicitly address in Pathline, as is the problem of encoding aggregate quantitative values directly along the pathway.

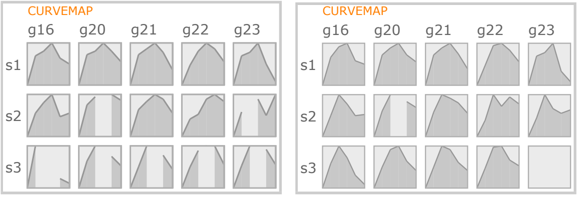
9. Case Studies

We collaborated with a team of seven biologists who have been collecting and analyzing comparative functional genomics data for several years — one of the authors on this paper is a member of the team. We conducted weekly meetings over the course of three months with members of the team to learn about their scientific questions, analysis needs, and available visualization tools. The team was previously using conventional heatmaps generated using Java TreeView [Sal04] to analyze gene activity and metabolite levels. They began using Pathline through a series of interactive prototypes that were gradually refined to the current design over the course of two months. The current version of Pathline is now their main visualization tool for analyzing this data. The biologists verified that Pathline can show known information more clearly than could be seen with their previous tools, and they directly attribute new insights into their data to the use of Pathline. We present their experiences with the tool as preliminary evidence towards the validity of our core design choices.

We have anonymized the species, gene, metabolite, and pathway names in these case studies by request from our collaborators, because they represent as-yet unpublished data and findings. Their insights derived from using Pathline have spurred them to undertake further analyses, and they anticipate publishable results.

9.1. Missing Data

One member of the research team noticed that the occurrence of missing data in many of the time series coincided with high values in the curve, as shown in Figure 3(a). Subsequent analysis of their data processing pipeline revealed a previously unknown bias for discarding high values. They then modified their pipeline; the reprocessed data shown in Figure 3(b) contain fewer missing data points for the same set of genes. Although the team was aware of the existence of missing values from the heatmap visualizations, they had not noticed that they occurred most often near high values.

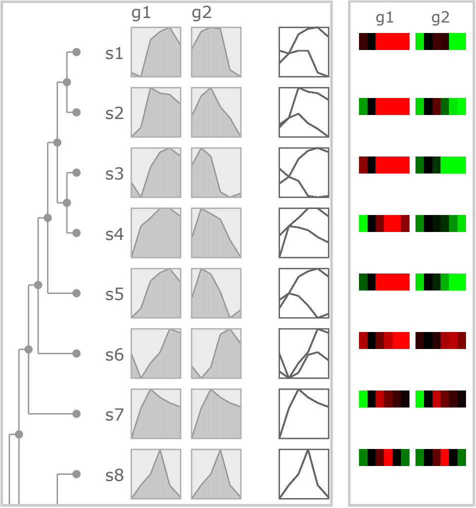


(a) (b)

Figure 3: ***(a)*** *Our collaborators noticed a correlation between high values and missing data in the curvemap detail view that had not been obvious when inspecting the heatmaps, and after investigation found a problem in the data processing pipeline.* ***(b)*** *After fixing the problem many missing data points were recovered.*

9.2. Whole Genome Duplication

A whole genome duplication event occurred in yeast some 150 million years ago. An ancestral yeast species gained an extra copy of all its genes, and thus living descendants of that common ancestor often have multiple copies of genes as a result of that duplication. Two such genes are `g1` and `g2`, which are shown in Figure 4. Scientists know that the genetic controls responsible for the activity levels of these genes have evolved to behave differently. A telltale sign of this phenomenon is the shift in the peaks of the time series curves for the `g2` gene activity patterns compared to those for g1 in the post-duplication species s1 to s5, shown in the first five rows of Figure 4(a). In the eighth row is one of the pre-duplication species, s8, where only one of the genes exists and the data have been duplicated for consistency, resulting in identical curves in both columns.



(a) (b)

Figure 4: *Whole genome duplication event.* ***(a)*** *The known post-duplication shift in activity patterns in the first five rows between the g1 and g2 genes is immediately obvious in Pathline, where the curves clearly have mirror symmetry.* ***(b)*** *The mirror symmetry is much less apparent in a conventional heatmap view showing the same data.*

The evolution of g2 in the post-duplication species was known by our collaborators, and was one of the first things they looked for using an early version of Pathline. According to the biologists, the expected shift is much more apparent in the curvemap encoding than in the conventional heatmap view, which is shown in Figure 4(b). The group remarked that these types of inquires generally take about 30 minutes to uncover in a heatmap, and require on the order of 5 minutes or less to see in Pathline — in this example, the trend was immediately obvious to them.

The biologists attribute this efficiency gain to several aspects of Pathline. One central aspect is the ability to identify similarities and differences in shapes in the curvemap views while still being able to understand the absolute magnitudes of the changes in the curve overlay plots. A heatmap typically relies on a fixed *saturation* value of the data corresponding to the brightest colors in the image; changing the saturation value significantly affects the viewer's perception of the trends in the data. Detecting the equivalent of similar curve shapes requires a detailed analysis of the heatmap at multiple saturation values — the identification of similarities across these multiple versions can be difficult. Combining the curvemap views with the curve overlays in the same window streamlines this process. Another central aspect of Pathline is the pathway-centered approach for customizing the curvemap view, which allows for the direct comparison of two genes or metabolites that would likely be displayed far from each other in a standard clustered heatmap, obscuring their key similarities and differences.

After using Pathline to see this previously known finding, the team continued their analysis and found something new. They noticed that although the s6 species in the sixth row is a post-duplication species, it exhibits behavior closer to the pre-duplication species in the bottom rows than its post-duplication relatives in the first five rows. Rather than diverging to display early versus late activation, these two different genes still display the same behavior, providing a possible hint as to where in the phylogeny the regulation of g2 changed to the dominant post-duplication behavior.

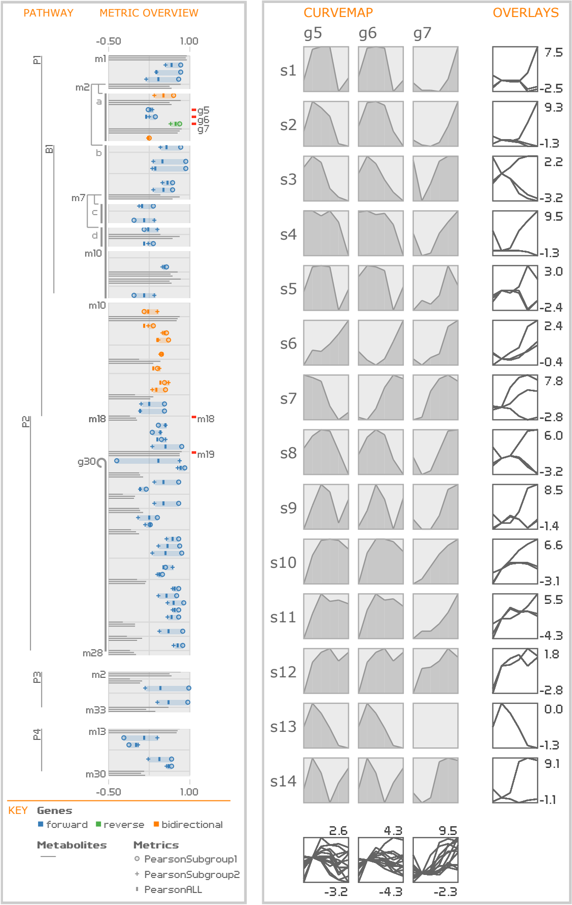
9.3. Pathway-Level Analysis

Our collaborators were also able to quickly identify some known pathway trends in their dataset. Shown in Figure 5(a), the metabolites in the P1 pathway and the P2 cycle show a general decline in similarity for elements later in the pathway with one notable outlier, the m19 metabolite. Identifying this trend and its outlier previously required a significant amount of effort; our collaborators stated that finding this same information using Pathline is straightforward and obvious using the linearized pathway view.

Our collaborators then probed deeper, and again analysis with Pathline led them to new insights. The metabolite m18 is acted upon by several enzymes to form m19. Comparing the m18 similarity scores for two important subgroups of species it was clear that m18 is not only poorly conserved across all of the species, but also within the two subgroups. This result is in contrast to the highly conserved nature of m19 across virtually the entire set of the species. By creating a curvemap only with m18 and m19, our collaborators immediately recognized some previously unknown behaviors. These behaviors have provided our collaborators with an interesting set of problems and hypotheses made possible by the link between the pathway view and the curvemap view in Pathline.

9.4. Gene-Level Relationships

Another interesting set of new insights involves the g5, g6, and g7 genes. g5 and g6 are both forward enzymes that work together to catalyze a single reaction in the P1 pathway, while g7 is a reverse enzyme that catalyzes the same reaction in the opposite direction — these three genes are selected in Figure 5(a). In the curvemap view, shown in Figure 5(b), it is evident that in almost every species g5 and g6 display nearly the same behavior. This fact would also be easy to detect in a heatmap because the genes would cluster together, making the rows adjacent — the proximity of the two genes to each other in the heatmap would be an indication that in each species the time series are nearly identical. The new insight, however, is that the temporal behavior of these two genes changes over the course of evolution, *i.e.*, across



(a) (b)

Figure 5: ***(a)*** *The metabolites along the* `P1` *and* `P2` *pathways show a general decline except for the outlier* `m19`*.* ***(b)*** *Using this curvemap view of the genes* `g5`*,* `g6`*, and* `g7` *the biologists confirmed known trends and discovered unknown gene duplication in the* `s7` *species.*

the species. Some species display an early peak, some a late peak, and others no strong peak at all. In contrast, the reverse gene `g7` is fairly similar in shape across all species, although the magnitude of changes varies, which is evident in the curve overlay plot at the bottom of the column. Our collaborators state that these latter two detailed analyses would have been difficult to perform using standard heatmaps: for instance, the different magnitudes of `g7` data would require inspection at many saturation levels to identify similarities across species. Additionally, they would have been unlikely to even probe for these specific behaviors due to the effort needed to produce targeted, customized heatmaps for small sets of genes.

In this same analysis, they quickly identified the extremely different behaviors in the `s7` species for `g5` and `g6` as the shape-based curvemap view made the trend immediately salient. From this observation they have since identified the cause of the behavior as a gene duplication event, and they now plan to investigate the potential function of the duplicate gene.

9.5. Bidirectional Enzymes

Another new insight sparked by the use of Pathline is the observation that the bidirectional enzymes all seem to have similar patterns, with just a few exceptions — these enzymes are selected in Figure 1. This trend had gone unnoticed using the team's previous visualization tools; visually encoding directionality of enzymes on the pathway view and supporting the direct comparison of this subset of enzymes in a curvemap view helped the biologists to notice this trend.

10. Conclusions and Future Work

We have presented the design of Pathline, an interactive prototype tool for visualizing comparative functional genomics data across multiple pathways, multiple genes and metabolites, and multiple species. Its curvemap detail view is an alternative to the color-based visual encoding of traditional heatmaps that supports detailed analysis of the shapes of time series curves across species and genes. The linearized pathways view provides an overview of multiple aggregate similarity scores for each gene across multiple pathways. We took a user-centered design approach in developing Pathline, working closely with our biology collaborators to refine the tool's design. These biologists used Pathline in their analysis process to confirm known findings and to generate new insights, and are using the tool to communicate these findings.

We believe that Pathline will be an effective visualization tool for many other biological problems. We have already identified two other research groups as potential users of Pathline — one group studies how stem cells give rise to the various types of blood cells in the body, while the other group is interested in abnormalities along cellular pathways that are involved with cancer. Furthermore, the curvemap and linearized pathway visual encodings are applicable to the much larger bioinformatics community that currently use heatmap and pathway visualization tools to explore a wide range of scientific topics. We hope the open-source release of Pathline will encourage a broader user base.

We would like to extend Pathline by allowing it to import metabolic and cellular pathway information directly from databases such as KEGG [KAG*08] and linearize it automatically. A very exciting direction for future work would be to add support for showing DNA sequence information, which could help researchers answer the deep question of how related genes have different gene activity patterns.

Acknowledgments

We thank our biology collaborators in the Regev lab for all of their time, enthusiasm, and support: Michelle Chan, Courtney French, Jay Konieczka, Jenna Pfiffner, Aviv Regev, and Dawn Thompson. This work was supported by the Data Visualization Initiative at the Broad Institute, and by the National Science Foundation under Grant 0937060 to the Computing Research Association for the CIFellows Project.

References

[AHP*05] ALM E., HUANG K., PRICE M., KOCHE R., KELLER K., DUBCHAK I., ARKIN A.: The MicrobesOnline web site for comparative genomics. *Genome Res 15*, 7 (July 2005), 1015–1022. 1, 6

[BMGK08] BARSKY A., MUNZNER T., GARDY J., KINCAID R.: Cerebral: Visualizing multiple experimental conditions on a graph with biological context. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2008) 14*, 6 (2008), 1253–1260. 1, 6

[BW09] BOURQUI R., WESTENBERG M. A.: Visualizing temporal dynamics at the genomic and metabolic level. In *IV '09: Proceedings of the 2009 13th International Conference Information Visualisation* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 317–322. 1, 6

[CFF*08] CASPI R., FOERSTER H., FULCHER C., KAIPA P., KRUMMENACKER M., LATENDRESSE M., PALEY S., RHEE S., SHEARER A., TISSIER C., WALK T., ZHANG P., KARP P.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research 36*, Database-Issue (2008), 623–631. 2

[CM84] CLEVELAND W. S., MCGILL R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association 79*, 387 (1984), 531–554. 4, 5

[ESBB98] EISEN M. B., SPELLMAN P. T., BROWN P. O., BOTSTEIN D.: Cluster analysis and display of genome-wide expression patterns. *Proc. National Academy of Sciences 95*, 25 (1998), 14863–14868. 2, 4, 7

[HS01] HOCHHEISER H., SHNEIDERMAN B.: Interactive exploration of time series data. In *Proc. Intl. Conf. on Discovery Science (DS)* (2001), Springer-Verlag, pp. 441–446. 6

[JKS06] JUNKER B. H., KLUKAS C., SCHREIBER F.: VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics 7* (2006), 109. 1, 6

[KAG*08] KANEHISA M., ARAKI M., GOTO S., HATTORI M., HIRAKAWA M., ITOH M., KATAYAMA T., KAWASHIMA S., OKUDA S., TOKIMATSU T., YAMANISHI Y.: KEGG for linking genomes to life and the environment. *Nucleic Acids Research 36*, Database-Issue (2008), 480–484. 9

[KK94] KEIM D. A., KRIEGEL H. P.: VisDB: Database exploration using multidimensional visualization. *Computer Graphics and Applications, IEEE 14*, 5 (1994), 40–49. 3

[KLK*09] KIM B., LEE B., KNOBLACH S., HOFFMAN E., SEO J.: GeneShelf: A web-based visual interface for large gene expression time-series data repositories. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2009) 15*, 6 (2009), 905–912. 5, 7

[KPR02] KARP P., PALEY S., ROMERO P.: The pathway tools software. *Bioinformatics 18* (2002), S225–S232. 1, 6

[LMK07] LAM H., MUNZNER T., KINCAID R.: Overview use in multiple visual information resolution interfaces. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2007) 13*, 6 (2007), 1278–1285. 4, 5

[LS10] LUBLING Y., SEGAL E.: Genomica, `http://genomica.weizmann.ac.il/`, last accessed 7 Apr 2010. 7

[LYKB08] LETUNIC I., YAMADA T., KANEHISA M., BORK P.: iPath: interactive exploration of biochemical pathways and networks. *Trends in biochemical sciences 33*, 3 (March 2008), 101–103. 1, 6

[Mac86] MACKINLAY J. D.: Automating the Design of Graphical Presentations of Relational Information. *ACM Trans. on Graphics (TOG) 5*, 2 (1986), 111–141. 5

[MMKN08] MCLACHLAN P., MUNZNER T., KOUTSOFIOS E., NORTH S.: LiveRAC: interactive visual exploration of system management time-series data. In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (2008), pp. 1483–1492. 6

[MMP09] MEYER M., MUNZNER T., PFISTER H.: MizBee: A multiscale synteny browser. *IEEE Trans. Visualization and Computer Graphics 15*, 6 (2009), 897–904. 6

[MSH*05] MLECNIK B., SCHEIDELER M., HACKL H., HARTLER J., SANCHEZ-CABO F., TRAJANOSKI Z.: PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Research 33*, Web-Server-Issue (2005), 633–637. 1, 6

[PW06] PLUMLEE M., WARE C.: Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. *ACM Trans. on Computer-Human Interaction (ToCHI) 13*, 2 (2006), 179–209. 5

[RFM07] REAS C., FRY B., MAEDA J.: *Processing: A Programming Handbook for Visual Designers and Artists*. MIT Press, 2007. 6

[Sal04] SALDANHA A. J.: Java Treeview – extensible visualization of microarray data. *Bioinformatics 20*, 17 (2004), 3246–3248. 2, 4, 7

[SDW09] SLINGSBY A., DYKES J., WOOD J.: Configuring hierarchical layouts to address research questions. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2009) 15*, 6 (2009), 977–984. 4

[SMO*03] SHANNON P., MARKIEL A., OZIER O., BALIGA N., WANG J., RAMAGE D., AMIN N., SCHWIKOWSKI B., IDEKER T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research 13*, 11 (November 2003), 2498–2504. 1, 6

[SND05] SARAIYA P., NORTH C., DUCA K.: An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans. Visualization and Computer Graphics 11*, 4 (2005), 443–456. 2, 7

[SS02] SEO J., SHNEIDERMAN B.: Interactively exploring hierarchical clustering results. *Computer 35*, 7 (2002), 80–86. 2, 6, 7

[WF09] WILKINSON L., FRIENDLY M.: The history of the cluster heat map. *The American Statistician 63*, 2 (2009), 179–184. 7