Tools for visual analysis of biological networks

Anton Zoubarev
December 18th, 2009

Abstract: The trend in tools for visualization and analysis of cell pathways is to go beyond small-scale static depictions of pathways and towards large-scale networks capturing cellular dynamics. Such tools attempt to construct pathways automatically and integrate data from large datasets in genomics, proteomics and transcriptomics to give a more meaningful representation of cell activity. We look at techniques that existing tools use to visualize large networks, to analyze their topology, to find patterns in microarray data and to study cell dynamics in the context of the pathways.

1 Introduction

Tools for visualization and analysis of biological pathways are becoming more important as the bottleneck in the research of cell biology shifts from data generating experimental stage to the data analysis and visualization stage. The complexity of a living cell, amount of data involved in a typical experiment, and diversity of available databases present a challenge for visualization.

1.1 Cell

A cell is a large complex system containing tens of thousands of genes and an even larger amount of proteins and other biological entities. Unlike the relatively unchanging genome, the dynamic proteome (set of proteins expressed by a genome) changes from minute to minute in response to tens of thousands of intra- and extracellular signals. The amount and diversity of cell components, interactions between them and their functions makes study of such system a challenge. Understanding complex choreography of cell activities is important for advancing biology and has applications in medicine.

1.2 Cell pathways

One of the steps towards better understanding of activities inside the cell is to capture underlying biology in a diagram. Historically, biologists studied pathways on a small scale. Even at this scale it sometimes takes years to understand a simple pathway. Years of research left hundreds of hand-drawn pathways capturing a glimpse of cell functioning. Such drawings are used to share research results, to teach, and to provide context for interpreting new experimental results.

1.3 High-throughput cell biology

The availability of techniques such as microarray technology and other high-throughput biology

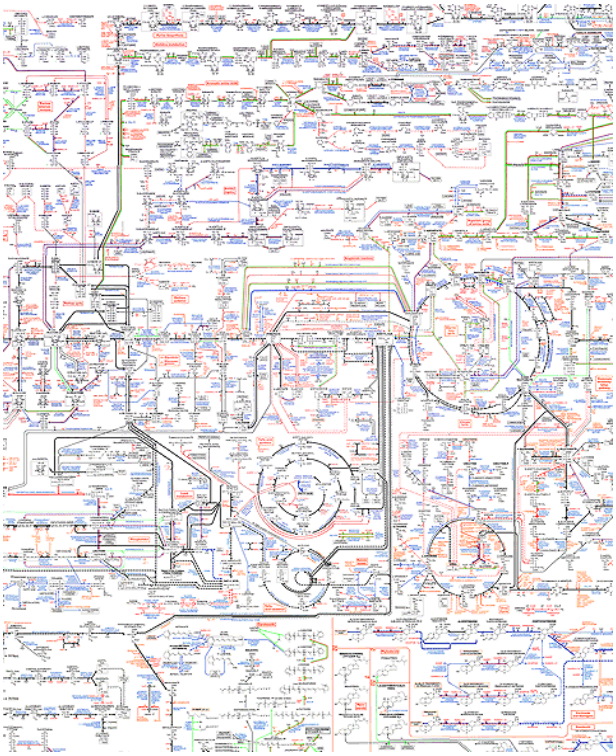

Figure 1: Fragment of manually constructed E. coli metabolic network. About one third of the network is shown. (http://www.g-language.org/g3/)

methods makes it possible to gather snapshots of cell activity across thousands of entities. These technologies greatly speed up generation of experimental data. Powerful tools are required to visualize and analyse large amounts of information produced by these methods.

1.4 Databases

Data generated by studying the transcriptome (set of all RNA molecules produced by the cell), the proteome and the metabolome (complete set of small-molecule metabolites) of the cell are stored in biological databases which are becoming widely available. They contain such information as protein-protein and protein-gene interactions, gene and protein functions and other biological data. These datasets are being linked together, cross-referenced and annotated.

1.5 Larger and richer pathways

With the availability of all these data, it is now possible to attempt to visualize entire cell networks or add more details to existing pathways by combining data from multiple sources and overlaying it; for example, gene regulation data can be added to existing metabolic pathway to highlight how each of the enzymes involved is expressed and controlled by the cell.

The motivation behind looking at larger pathways is this. Small, well-studied pathways are neither isolated nor independent from one another. In some cases, it is important to study all the pathways simultaneously; for example, some conditions span multiple pathways such as cell's response to stress. Even in cases where researchers study only a small portion of the cell network, the bigger picture view can help them see what happens upstream and downstream from the region of their interest.

1.6 Paper scope

In this paper we look at existing techniques that are employed by analysis and visualization tools used to better understand inner workings of the cell. These include tools to visualize large networks, to analyze their topology, to find patterns in microarray data, and to study cell dynamics in the context of the pathways.

2 Visualizing pathways

Representing pathways visually can allow researchers to explore and find interesting features in networks and pathways 'by eye'. For visualization to be effective for this purpose it has to capture essential biological information, provide enough detail and context, as well as respect limitations of human visual system.

Traditionally, pathway diagrams were hand drawn with biologists making sure the diagram conveys all necessary meaning while at the same time remaining easy to interpret. Construction, layout and visual encoding are in the hands of the researcher. Since such diagrams are usually contain around 30-40 elements it is not too difficult to deal with these issues. Sometimes a certain amount of creativity is required to make diagram concise and easy to understand.

Visualization process can be broken down into three steps. First, pathway has to be constructed. Second, layout and placement of pathway entities has to be chosen. Third step involves choice of visual encoding to show types of entities and information associated with them.

2.1 Manual and automated pathway construction

Before pathway can be visualized it has be assembled. Databases such as KEGG (Kanehisa 2000) contain comprehensive list of cellular pathways. The visualizations provided by such archives are static. Tools like Pathway Editor (Sorokin A. 2006) allow users to create pathway visualizations manually. Systems like Pathway Studio (Nikitin 2003), PathwayFinder (Daming 2004), and PubGene (Jenssen 2001) attempt to build pathways by processing research articles retrieved from publicly available databases. GenePath (Zupan 2003) infers pathways based on microarray data. Vector PathBlazer (Reshetnikov 2003) can create pathways by combining information from different reference databases such as KEGG (Kanehisa 2000).

2.2 Layout

Layout of constructed pathways can be done either manually or in automated fashion. Manual placement of elements allows more control and precision but it is time consuming. On the other hand, layouts produced by algorithms are fast but often lack essential biological details.

2.2.1 Manual layout

Usually, for manually constructed pathways layout is produced manually as well, either by drawing or using pathway editor tools. In order to move to larger scale networks one approach is to simply draw a network combining many known pathways of an organism. This method has its uses, but the approach does not scale well because constructing such a map for a new organism will take a long time. Figure 1 shows roughly a third of E. coli metabolic network to give an idea of the scale involved.

Another approach is to keep all manually constructed pathways in a database but provide query mechanisms to find pathways of interest such as pathways that would be affected by introduction of some drug. This approach ignores relationships between pathways and makes it difficult to see how pathways interact with each other.

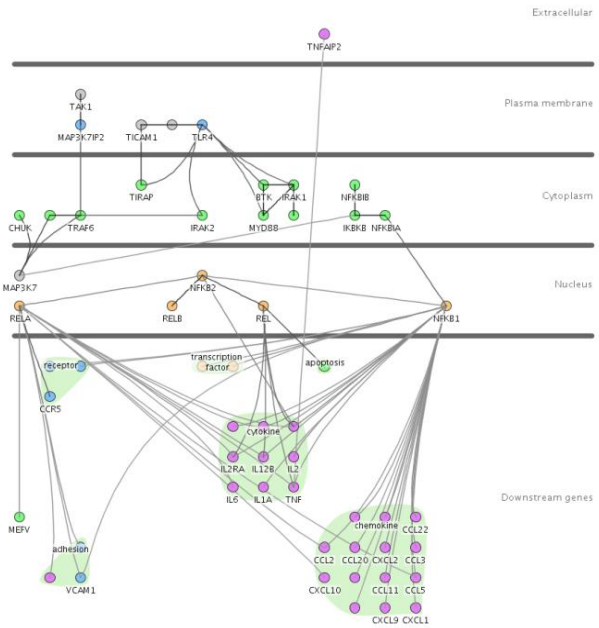

Figure 2: Sub-cellular localization is shown by Cerebral (Barsky 2007) by placing nodes in different regions of the screen.
(http://www.cytoscape.org/retreat2007/applications.php#barsky)

2.2.2 Automated Layout

Many available biological network tools make use of generic graph visualization algorithms like simulated annealing, force-directed layouts, and so on. Layouts produced by these algorithms often have little meaning to a biologist since the layout optimization criteria does not take into consideration domain-specific knowledge such as protein complexes, sub-cellular localization or order of the events in a pathway. Some tools do attempt to use domain-specific knowledge while rendering pathways.

Pathway Studio (Nikitin 2003), Patika (Demir 2002) and VisANT (Hu 2005) visualize composite nodes representing molecular complexes and pathways as single nodes that can be interactively expanded to show individual members (Figure 3).

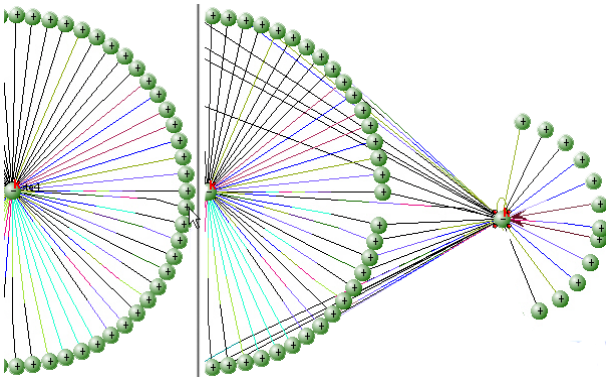

Figure 3: Nodes can be collapsed and expanded in VisANT (Hu 2005). **(http://visant.bu.edu/vmanual/)**

Cerebral (Barsky 2007) and some other tools place nodes according to their compartment inside the cell, as shown in Figure 2.

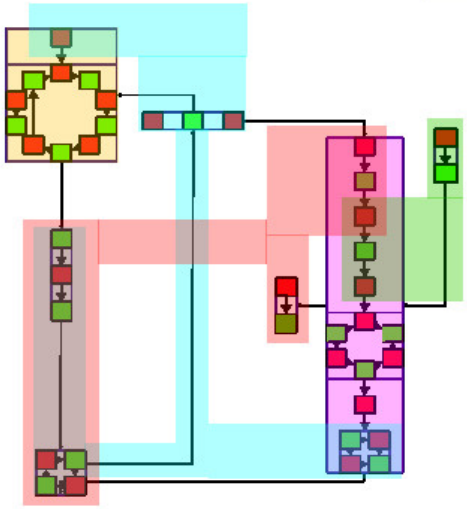

Figure 4: Graph rendered by MetaViz (Bourqui 2007). Cycle and cascade features are visible. Overlapping pathways are highlighted. (image taken from Bourqui 2007 paper)

MetaViz (Bourqui 2007) enables drawing of a genome-scale metabolic network that takes into account some pathway drawing conventions used in biological community such as the notion of cascade and cycle. Constructing large networks from multiple pathways that sometimes overlap requires a decision to be made on which of the overlapping pathways to draw in full and which ones to break down into sub-pathways. MetaViz allows the user to control such choices thus allowing researchers to emphasize different sub-pathways depending on the question they are trying to answer (Figure 4).

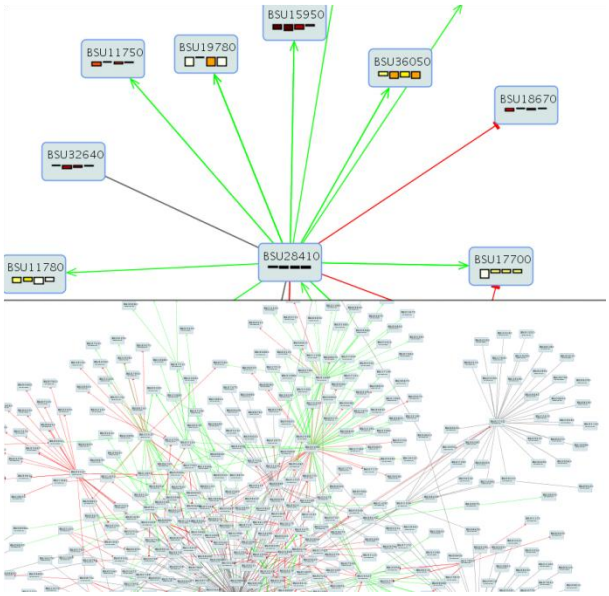

Figure 5: GeneVis (Baker 2003) draws labels in world space which makes them readable only with sufficient amount of zoom (top). When viewing large networks, labels are illegible (bottom).

2.3 Visual encoding

There is no official standard notation to represent pathway elements. Every tool uses some different variant of visual encoding. There are a number of attempts to standardize the notation, for example, Systems Biology Graphical Notation (Le Novère 2009).

All pathway visualization tools display pathways as graphs where nodes may represent a range of biological features of interest: genes, proteins, macromolecular complexes, or even cellular pathways. These nodes are connected by different types of edges that encode many types of interactions that are possible between components. Edges can denote movement, inhibition, activation, binding, and so on.

Typically, simple visual attributes for nodes (colour, size or shape) and edges (colour, type of arrow head or style of the line) are used to convey some biological meaning of interest: type of entity, amount or type of interaction.

For visualization to be meaningful to a biologist, more domain-specific information has to be shown. One common technique is category-based grouping and highlighting. Related components, such as those involved in performing the same biological function, are displayed together and highlighted. In Figure 2 nodes of related entities are grouped together and outlined with highlighting. VisANT (Hu 2005) allows connections to be defined between groups of nodes, which makes it possible to have a layer of abstraction by hiding details and viewing only relationships between sub-networks.

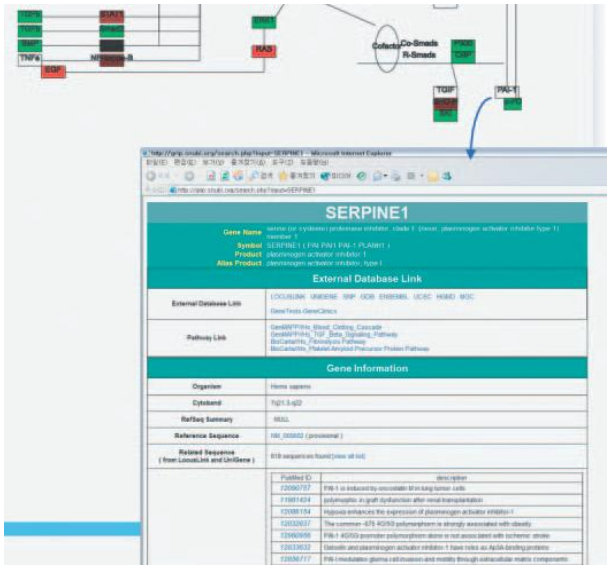

Figure 6: Clicking on a node in ArrayXPath (Chung 2004) opens a new window with additional information about biological entity. This includes biological processes involving this compound, links to relevant studies, organisms, and so on. (http://www.snubi.org/software/ArrayXPath)

Other biological information such as gene or enzyme names is displayed using labels. Tools differ widely in how they handle labelling. Some choose to show all labels which results in many of them overlapping and occluding each other. Others draw labels in world space resulting in labels that are too small and not legible except at high levels of zoom (Figure 5). A better approach that's implemented in Cerebral (Barsky 2007) seems to be to draw labels in screen space and use algorithms that prevent overlapping labels. This reduces the number of labels that are displayed, but all labels that are shown are legible. Users can manipulate label density to choose appropriate settings for their task.

Another common method to add more biological information is to bring up references about the entity from biological databases with links to relevant studies when that entity is selected by the user. This is shown in Figure 6.

2.4 Scalability

As network gets larger and overwhelm our visual abilities, we have two choices: to show less information on the screen or to show it better.

One major way to deal with too much information is simply not showing a portion of it. Querying, filtering and navigation are all examples of this approach.

2.4.1 Navigation

Navigation by allowing the user to zoom and pan around the network is very common; it is implemented by VisANT (Hu 2005), Patika (Demir 2002), Pathway Studio (Nikitin 2003) and most other tools. This approach works well when area of interest is localized to a small region of a network and other regions are irrelevant. One problem with this method is that it makes it impossible to see sufficient level of detail in the region of interest while keeping overall context in mind. One technique that addresses this shortcoming is fisheye view (Toyoda 2003). One survey suggests that such views could be confusing to biologists and could require some time to get used to (Saraiya 2005).

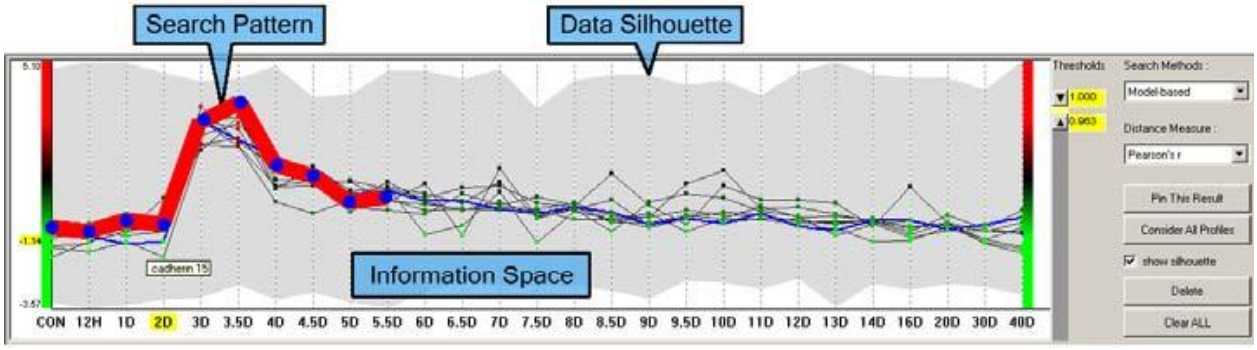

Figure 7: HCE (Seo 2006) allows searching gene profile in parallel coordinates by simply drawing a desired gene expression profile. (http://www.cs.umd.edu/hcil/hce/)

2.4.2 Querying and filtering

If we want to see patterns that span the entire network, in many cases some of the available

information is not relevant and can be hidden. Some of the existing tools allow querying and filtering to restrict focus only to nodes of interest. Pathway Studio (Nikitin 2003) contains an SQL-like language that allows users to query the network using some simple topological contractions with node and link attributes.

2.5 Abstraction

Another powerful approach to deal with visual complexity is abstraction. In this case some information is lost but relevant features are preserved. The use of it is limited in biological visualization as there does not seem to be common well-defined layers of abstraction. Some abstractions can be created using existing tools. Representing entities composed of multiple elements as a single node can be done in VisANT (Hu 2005). MetaViz (Bourqui 2007) can show thumbnail image of pathway as a 'node'.

2.6 Reducing clutter

Sometimes it is possible to reduce visual clutter while preserving all information. The edge bundling method makes the screen less cluttered by bundling edges together. Cerebral's version of edge bundling can be seen in Figure 2. Another technique that seems to make image easier to understand is placing nodes on invisible grid.

3 Pathway analysis

There is a limit on what can be discovered just by visually studying the pathway. A majority of pathway visualization tools come bundled with multiple analysis methods; these include determining basic topological characteristics of a network or overlaying data from microarray experiments and comparing pathway activity under different treatment conditions.

3.1 Topology analysis

The topology of the network can shed some light on underlying biology. Researchers can run all sorts of graph related algorithms to try to answer questions of interest. Some researchers are looking at conservation of pathways across multiple organisms. PathBLAST (Kelley 2004) finds similarities in networks belonging to different organisms. Osprey (Breitkreutz 2003) is able to superimpose one network on top of another to highlight similarities and differences.

Often topological features have a meaningful biological interpretation. In protein interaction and genetic interaction networks, for example, the degree of a node is often its importance to the cell. Slight changes to such a node will have a dramatic impact on cell functions.

It may also be useful in some contexts to find the shortest path between two genes, proteins, complexes or pathways. The overall lengths of such pathways may be related to the immediacy or breadth of signal response.

3.2 Cell dynamics analysis

While it is important to know what pathways the cell has, it is essential to see which pathways are active in the cell at a given time. High-throughput methods are used to take snapshot of quantities of biological entities of interest. Several pathway visualization tools incorporate gene expression data within the network. Such mapping of gene expression information provides a glimpse of systems-wide dynamics in biological systems and makes it possible to view changes of cell activity in the context of a pathway.

3.2.1 Overlaying microarray data

In a typical microarray experiment, gene expression activity is sampled across multiple treatment conditions and at multiple time points or stages of cell cycle. Although tens of thousands of genes are typically measured, the researchers do not usually work with an interaction graph model covering the entire dataset because of its overwhelming complexity.

Tools use different approaches to integrate the collected data into an interaction graph: mini heat-map at each node as in GenMAPP (Salomonis 2007), mini line chart as in Pathway Editor (Byrnes 2009), and animation of node colour. Some of them are shown in Figure 8.

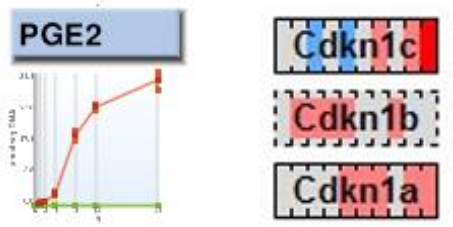

Figure 8: Pathway Editor (Byrnes 2009) shows mini line chart graph for each node with gene expression data (left). GenMAPP (Salomonis 2007) overlays microarray data as mini heat map on each node (right).

3.2.2 Querying and filtering

To aid in study of dynamics of cellular processes, a parallel coordinates view is part of a large number of pathway visualization tools. In parallel

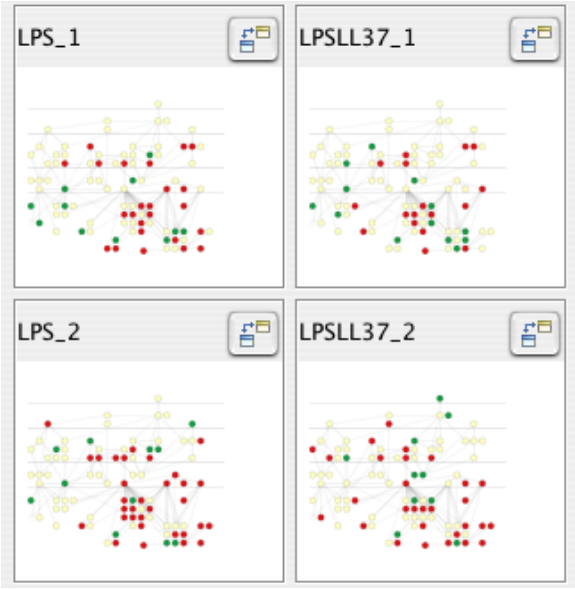

Figure 9: Small multiples view used in Cerebral (Barsky 2007). (http://www.pathogenomics.ca/cerebral/manual.html)

coordinates, points in n-dimensional space are represented by polylines with vertices on the parallel axes such that each vertex of the polyline corresponds to the coordinate of the point in one of the dimensions. A very convenient way to find genes of interest in parallel coordinates is to simply draw a desired expression profile and genes matching it will be selected. One implementation of this method by HCE (Seo 2006) is shown in Figure 7. To reduce clutter while preserving all the information visual clustering in parallel coordinates can be used that is somewhat similar to edge bundling (Zhou 2008). This approach is shown in Figure 10 and is not implemented by any of the tools we looked at.

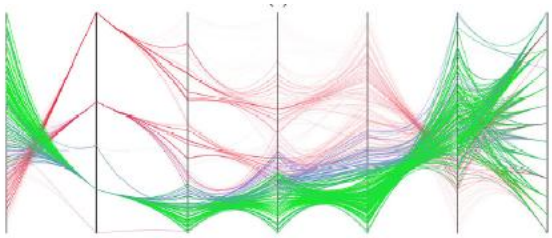

Figure 10: Visual clustering in parallel coordinates (Zhou 2008) reduces visual clutter by bundling polyline edges. (Image taken from Zhou 2008 paper)

Some of the elements in the visual field can be removed by filtering. Different filters can be applied to the data. Fold-change filter removes data points that do not have the specified minimum gene expression level. This minimum limit can be set by researcher depending on the experiment. Another example of a filter is a significance filter. When the microarray dataset is pre-processed, a statistical significance score is assigned to each measurement in each condition. A significance filter allows removal of the data points with specified range of scores.

3.2.3 Comparing multiple conditions

In another set of tasks, researchers could be interested in how a pathway is affected by different treatments. Small multiple views, as shown in Figure 9, can be used to allow quick comparison of different conditions. Such views present multiple smaller versions of the pathway under different treatment condition. It makes it easier to spot differences of gene expression across multiple conditions. The tool can also calculate the difference of gene expression data by drawing a pathway where nodes encode change in gene expression between two conditions which eliminates the need to compare it by eye.

3.3 Interactions

Two Interaction techniques that researchers seem to find very useful is grouping nodes according to

some criteria and linking some biological information with the group (Saraiya 2005).

Linked views are also helpful. They provide views of the same dataset but present it using different visualization techniques, such as heat map, parallel coordinates, scatter plots, and so on. Each visualization approach offers a unique perspective of the data. Selection in one view automatically selects entities in all others making exploration easier and more intuitive (Figure 11).

4 Microarray data analysis

In some cases it is useful to look at an entire microarray dataset from an experiment. Tools like HCE (Seo 2006) and GeneSpring (Agilent) allow the study of large amounts of microarray data. By using these tools, researchers try to find interesting patterns of gene expression. The tasks are mostly exploratory in nature that is a researcher may not have a particular goal in mind when starting analysis.

4.1 Visual encoding

Microarray data from an experiment is a large multidimensional dataset. Typically such data sets are visualized as heat maps with dendrograms produced by clustering. Parallel coordinates and scatter plots are also used to provide greater level of detail when needed. All these views are linked so that selecting element in one view also highlights it in others. Figure 11 shows linked views in Hierarchical Clustering Explorer (Seo 2006).

4.2 Clustering

One of the exploration tasks is to try to see patterns and correlations in the data. To aid researchers, various approaches to highlight clusters of data with some degree of similarity are used. Many clustering algorithms are available. Hierarchical clustering, k-means clustering, self-organizing maps are among popular ones. Other techniques from statistical data analysis are present in some tools; GeneSpring supports PCA and QT clustering.

One scenario where clustering can be used is to verify if microarray data looks reasonable. Usually researchers have some knowledge of each experimental condition and based on that they would expect a clustering where similar experimental conditions are clustered closely with one another. A quick look at dendrograms can confirm their expectations.

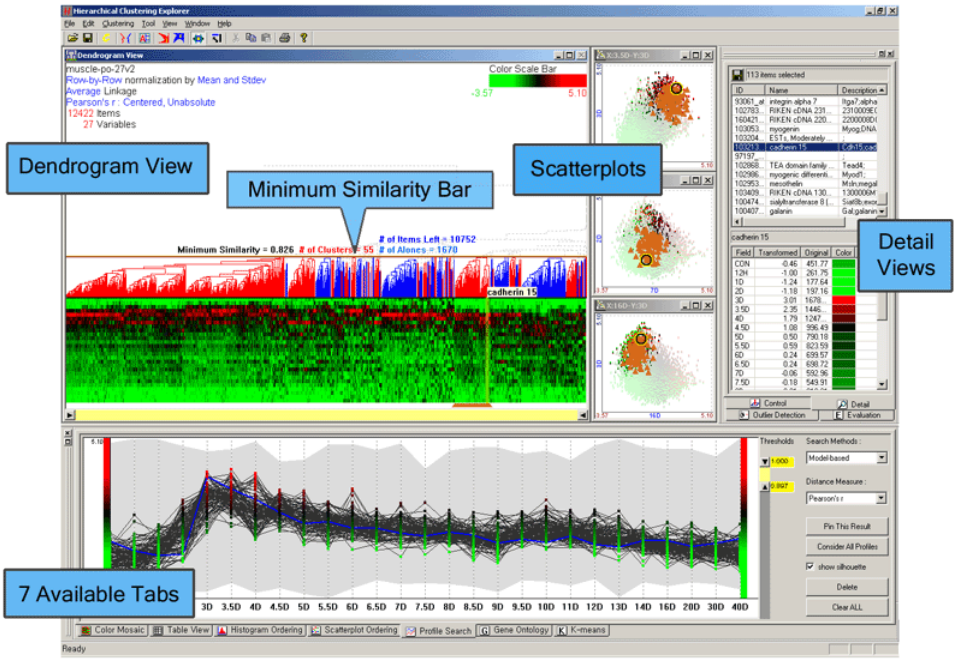

Figure 11: Linked views in HCE (Seo 2006). (http://www.cs.umd.edu/hcil/hce/)

4.3 Scalability

For large datasets, when there are more data points than pixels on the screen, an approach that renders heat map and dendrogram overview can be used (Chen 2009). This approach offers a way of simplifying the graphical display while maintaining essential information and providing support for easy navigation and display of contextual information. This strategy links an overview dendrogram and a detail-view dendrogram, each integrated with a re-orderable heat-map. The overview displays only a user-controlled, limited number of nodes that represent the skeleton of a hierarchy as shown on the left side of Figure 12. The detail view, shown on the right side of Figure 12, displays the sub-tree represented by a selected meta-node in the overview.

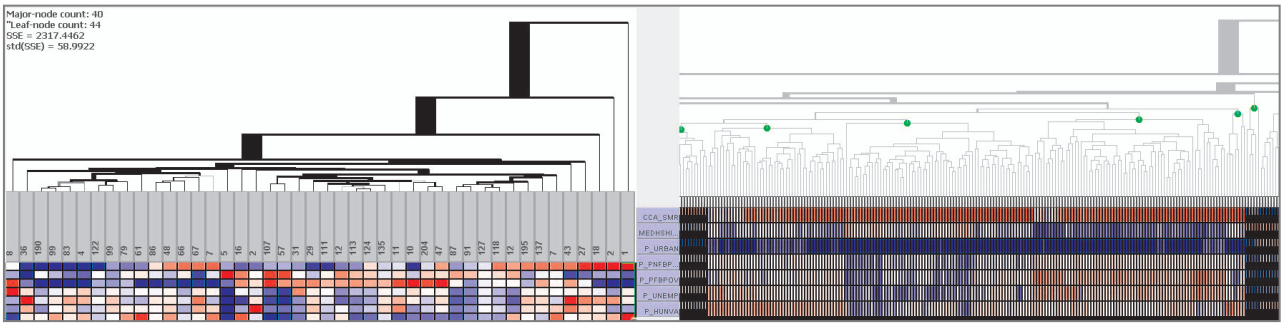

Figure 12: The overview displays only a user-controlled, limited number of nodes that represent the skeleton of a hierarchy as shown on the left side. The detail view, shown on the right side, displays the sub-tree represented by a selected meta-node in the overview. (Image taken from Chen 2009 paper)

4.4 Problems

Tools in this group provide the ability to discover relationships, clusters, gaps, outliers, and other features in the data. However, since many of the methods were adopted from non-biological areas, a major challenge is getting to biological meaning from discovered relationships.

Another drawback of clustering approaches is that they can potentially bias users into a particular line of thought too quickly (Saraiya 2004).

5 Conclusion

Visualization and analysis tools for biological pathways are moving towards richer, bigger pathway visualizations that integrate high-throughput experimental data and information from multiple databases. Tools in this area have to have simple querying, navigation, and offer powerful analysis methods to help users study large datasets.

A survey published in 2005 (Saraiya 2005) identifies lack of domain-specific biological context as the main shortcoming of existing tools. While newer tools attempt to improve on this: for example, area of sub-cellular localization has seen some improvement; in general, more biologically rich pathway visualizations are necessary. Niche tools with narrower focus and well-defined tasks are usually better with providing biological context than more general tools.

To deal with scale, most current graph rendering techniques used produce ball-and-stick graphs that are hard to read for most dataset sizes (Figure 5). Graph layout algorithms more tuned towards biological needs would be valuable. Also, some types of biological interactions are not necessarily well modeled by simple binary graphs; experimenting with hyper-graphs might prove to be beneficial. Defining common layers of abstraction could enable more scalable visualizations.

Current tools offer some querying and filtering capabilities. More powerful query languages can greatly enhance usefulness of the tools. The great help would be the ability to issue queries like: “Find pathways where selected entity affects cell-cycle progression” (Suderman 2007) .

References

Agilent. *GeneSpring.* www.agilent.com/chem/genespring.

Baker, C. et al. "GeneVis: simulation and visualization of genetic networks." *Information Visualization* 2, no. 4 (2003): 201-217.

Barsky, A. et al. "Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation." *Bioinformatics* 23, no. 8 (2007): 1040-1042.

Bourqui, R. et al. "Metabolic network visualization eliminating node redundance and preserving metabolic pathways." *BMC Systems Biology* 1, no. 28 (2007).

Breitkreutz, B. et al. "Osprey: a network visualization system." *Genome Biology* 4 (2003).

Byrnes, B. et al. "An editor for pathway drawing and data visualization in the Biopathways Workbench." *BMC Systems Biology* 3, no. 99 (2009).

Chen, J. et al. "Constructing Overview + Detail Dendrogram-Matrix Views." *IEEE Transactions on Visualization and Computer Graphics* 15, no. 6 (2009): 889-896.

Chung, H. et al. "ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics." *Nucleic Acids Research* 32 (2004).

Daming, Y. et al. "PathwayFinder: paving the way towards automatic pathway extraction." *ACM International Conference Proceeding Series.* 2004.

Demir, E. et al. "PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways." *Bioinformatics* 18, no. 7 (2002): 996-1003.

Hu, Z. et al. "VisANT: data-integrating visual framework for biological networks and modules." *Nucleic Acids Research* 33 (2005).

Jenssen, TK. et al. "A literature network of human genes for high-throughput analysis of gene expression." *Nature Genetics*, no. 28 (2001): (21-28.

Kanehisa, M. et al. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28, no. 1 (2000): 27-30.

Kelley, B. et al. "PathBLAST: a tool for alignment of protein interaction networks." *Nucleic Acids Research* 32 (2004).

Le Novère, N. et al. "The Systems Biology Graphical Notation." *Nature Biotechnology* 27 (2009): 735 - 741.

Nikitin, A. et al. "Pathway studio—the analysis and navigation of molecular networks." *Bioinformatics* 19, no. 16 (2003).

Reshetnikov, V. et al. "Vector PathBlazer: A New Pathway Discovery, Analysis, and Visualization Tool." *Proc. ISMB.* 2003.

Salomonis, N. et al. "GenMAPP 2: new features and resources for pathway analysis." *BMC Bioinformatics* 8, no. 217 (2007).

Saraiya, P. et al. "An Evaluation of Microarray Visualization Tools for Biological Insight." 2004. 1-8.

Saraiya, P. et al. "Visualizing biological pathways: requirements analysis, system evaluation and research agenda." *Information Visualization* 4, no. 3 (2005): 191-205.

Seo, J. et al. "An Interactive Power Analysis Tool for Microarray Hypothesis Testing and Generation." *Bioinformatics* 22, no. 7 (2006): 808-814.

Shannon, P. et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Research* 13, no. 11 (2003): 2498-2504.

Sorokin A., Paliy K., Selkov A.,Demin O. V., Dronov S., Ghazal P., Goryanin I. "The pathway editor: a tool for managing complex biological networks." *IBM Journal of Research and Development* 50, no. 6 (2006): 561 - 573.

Suderman, M. et al. "Tools for Visually Exploring Biological Networks." *Bioinformatics* 23, no. 20 (2007): 2651-2659.

Toyoda, T. et al. "GSCope: a clipped fisheye viewer effective for highly complicated biomolecular network graphs." *Bioinformatics* 19, no. 3 (2003): 437-438.

Zhou, H. et al. "Visual Clustering in Parallel Coordinates." *Computer Graphics Forum* 27, no. 3 (2008): 1047 - 1054.

Zupan, B. et al. "GenePath: a system for automated construction of genetic networks from mutant data." *Bioinformatics* 19, no. 3 (2003): 383-389.