

DESIGNING VISUALIZATIONS FOR BIOLOGICAL DATA

Miriah Meyer, School of Computing,
University of Utah, Salt Lake City,
UT, 84112, U.S.A.
E-mail: miriah@cs.utah.edu

Submitted: <leave for Editor to date>

Abstract

Visualization is now a vital component of the biological discovery process. In this article I present visualization design studies as a promising process for creating effective, visualization tools for biological data.

The field of biology has been transformed over the last decade with the rapidly decreasing cost of data. Today, the focus is on combining huge genomics databases with large amounts of molecular data, and then augmenting this data with long-term clinical outcomes across large populations of people. Biologists believe that embedded in these complex, massive datasets are scientific goldmines like a cure for cancer. But how do you combine data that spans from the molecular level up to the population level in a meaningful way?

Visualization has emerged as an important way to make sense of this data. Lying at the intersection of computer science, design, and biology, visualization of biological data enables scientists to understand data by encoding meaning through images and supporting exploration through human-computer interactions. My research shows that producing agile visualizations by matching the rate of software development to the biologists' rate of experimentation produces tools that not only support scientists, but also influence them as they tackle complex questions.

A recent trend in the visualization community is to develop influential tools by conducting *design studies* in close collaboration with scientists and other expert end users. In this paper I describe a methodology for conducting design studies, and illustrate their effectiveness with an example from my own work.

Visualization Design Studies

A design study is a "project in which visualization researchers analyze a specific real-world problem faced by domain experts, design a visualization system that supports solving this problem, validate the design, and reflect about lessons learned in order to refine visualization design guidelines"[1].

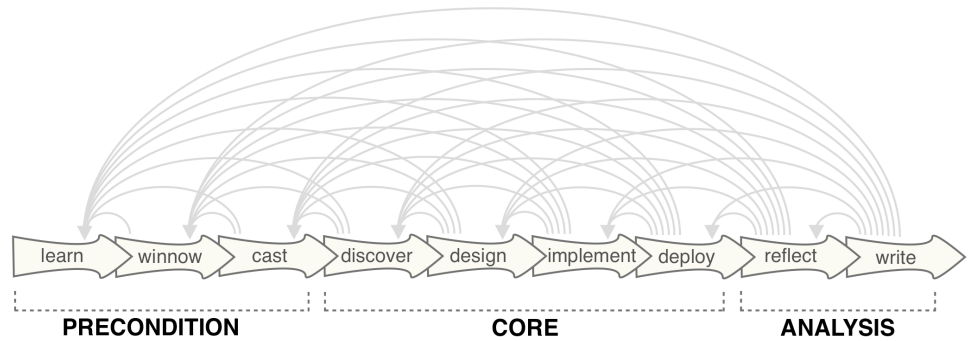


Fig. 1. This nine-stage framework [1] provides practical guidance for designing visualization tools that focus on solving specific, real-world problems. Copyright Miriah Meyer, 2012.

Design studies are often highly collaborative projects that, at their heart, rely on a visualization practitioner acquiring a deep understanding of a problem and exploring the broad space of design solutions. Through progressively refining prototypes based on feedback from the target end-users, effective visualization tools emerge for solving a real-world analysis problem.

The nine-stage framework [1] lays out a practical process for conducting design studies, shown in Figure 1. This framework contains three *precondition* stages that focus on learning the space of visualization techniques and establishing synergistic collaborations. The inner four *core* stages describe the process for characterizing a real-world problem followed by the design, implementation, and evaluation of a visualization solution. The final two *analysis* stages describe steps for analyzing lessons-learned during the project and communicating those findings to the visualization research community. The entire process is iterative and practitioners often jump backwards to earlier stages when refining ideas.

Pathline: A Design Study

To illustrate how the nine-stage framework applies to a real-world, visualization design problem, I'll discuss the process of creating the tool Pathline [2], a visualization system for exploring molecular biology data.

In this project we collaborated with a group of biologists at the Broad Institute of Harvard and MIT who are studying how the same sets of genes can facilitate different kinds of cellular functions in related species. Our collaborators specifically study metabolism in yeast, and they spent several years conducting experiments and collecting data for many different genes in fourteen species of yeast. Their goal is to understand patterns and trends in this data in the context of known information about

chemical reactions that occur in cells. The problem they faced at the onset of our collaboration was that existing tools could only look at a subset of this data at a time.

During the *discover* stage of the project we conducted weekly meetings with the group of 7 biologists over the course of 3 months to learn about their scientific questions, data analysis needs, and existing visualization tools.

From these meetings we came to understand that the group was working with four kinds of data: several dozen **graphs** that describe cellular reactions that are catalyzed by genes; a large **table of quantitative values** that describe how much each gene in each species is on or off over time during an experiment; a **tree** depicting the evolutionary relationship of the fourteen species of yeast; and **quantitative correlation values** that describe how similar the temporal values are for a specific gene across the species.

We performed a similar analysis on the questions our collaborators were asking and translated them into a list of data analysis tasks. We characterized these tasks as operating at four different levels. Briefly, those tasks, starting at the lowest level, are:

1. Study the gene data over time to observe dynamic trends in the data.
2. Compare a limited number of time series, such as those corresponding to a specific gene across all the species, to characterize meaningful differences.
3. Compare the correlation values across one or more of the reaction graphs to understand where the genes behave the same, and differently, in the species.
4. Compare multiple correlation values to validate which correlation metric is the most informative for the specific analysis needs of the group.

This data and task abstraction [3] served as the input for the *design* stage. In this stage we focused our designs

around the idea of encoding quantitative values with a spatial encoding channel instead of other, less effective channels such as color [4]. This means that instead of using a node-link diagram to visualize the reaction graphs which relies on spatial position to show topology [5], we instead linearize the graphs and use position to layer on the correlation values. We also use spatial encoding to show the temporal gene data using a matrix layout of line charts, called a curvemap, as opposed to the widely used color encoding of the heatmap display [6]. The linearized graph representation and the curvemap display are shown in the left and right of Figure 2 respectively. This figure is a screenshot of the Pathline visualization interface.

These two views are linked together by interaction. A user selects genes of interest on-the-fly in the linearized representation which populates the curvemap display by adding columns of temporal gene data. Thus, the linearized representation serves as an overview of the data, guiding the user to select interesting genes for more detailed analysis in the curvemap display [7].

During the *implement* stage of the design study process we used an iterative refinement scheme over the course of two months. We created numerous low-fidelity prototypes using illustration software, followed by rudimentary interactive prototypes using the Processing programming language [8]. Each prototype was refined based on feedback from our collaborators. Source code, executables, and example data for the final version of Pathline can be found at <http://www.pathline.org>.

In the *deploy* stage we released the final version of Pathline to our collaborators – it is now one of the primary analysis tools used by the group. The biologists verified that Pathline shows known information more clearly than could be seen with their previous visualization tools, and they directly attributed new insights into their data to the use of Pathline.

We compiled a series of case studies [9] that illustrate how the biologists use Pathline and what types of analysis they are able to perform. These case studies describe how, using Pathline, the biologists were able to: discover a bias in their data processing pipeline that produced missing data; confirm known findings orders of magnitude faster than using conventional visualization tools; develop numerous hypotheses, one of which led to a follow-up experiment and

the discovery of a previously unknown evolutionary event.

The design study project that underlies Pathline has several visualization research contributions, including a detailed characterization of this subdomain in biology, the invention and validation of new visualization representations, and the creation of the first visualization tool to combine all four kinds of biological data used by the group.

Conclusions

Today, biological data holds the promise of explaining the origins of life, curing human disease, and helping us to live healthy and happier lives. Reaching these goals, however, relies on making sense of vast amounts of biological data. This challenge has placed visualization directly within the scientific discovery process.

Based on close collaborations between visualization practitioners and domain experts, design studies are a visualization method for creating tools that support complex, real-world, data analysis problems. Design studies have proven to be a powerful method for designing effective visualization tools – this trend is expected to continue as the flood of biological data continues to grow.

References and Notes

* This paper was presented as a keynote talk at Arts, Humanities, and Complex Networks – 3rd Leonardo satellite symposium at NetSci2012. See <http://artshumanities.netsci2012.net>

1. M. Sedlmair, M. Meyer, T. Munzner, “Design Study Methodology: Reflections from the Trenches and the Stacks”, *IEEE TVCG* **18**, No. 12 (2012), pp. 2431-2440.
2. M. Meyer, B. Wong, M. Styczyski, T. Munzner, H. Pfister, “Pathline: A Tool for Comparative Functional Genomics”, *Computer Graphics Forum* **29**, No. 3 (2010), pp. 1043-1052.
3. T. Munzner, “A Nested Model for Visualization Design and Validation”, *IEEE TVCG* **15**, No. 6 (2009), pp. 921-928.
4. W. Cleveland and R. McGill, “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods”, *J. American Statistical Association* **79**, (1984), pp. 531-554.
5. G. Di Battista, P. Eades, R. Tamassia, I. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs* (Prentice Hall PTR, 1998).
6. L. Wilkinson and M. Friendly, “The History of the Cluster Heat Map”, *American Statistician* **63**, No. 2 (2009), pp. 179-184.
7. B. Shneiderman, “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”, *Proc. IEEE Symp. Visual Languages*, (1996), pp. 336-342.
8. B. Fry, *Visualizing Data* (O’Reilly Media, 2008).
9. P. Cairns and A. Cox, *Research Methods for Human-Computer Interaction* (Cambridge University Press, 2008).

Fig. 2. Pathline is a tool for exploring molecular biology data [2]. The view on the left is a linearized representation of graphs that describe cellular reactions. The view on the right is a curvemap display showing time-series data for several user-selected genes across related species. Copyright Miriah Meyer, 2012.

