Abstract of "Improved Scientific Analysis through Domain Driven Visualization and Support for Analytic Deliberation" by Radu Jianu, Ph.D., Brown University, May 2012.

This dissertation introduces and evaluates novel visualization methods that enable researchers to derive and test hypotheses from available scientific data faster and more accurately than before. Following the traditional visualization approach, we introduce novel ways of visualizing and interacting with scientific data that support and accelerate researchers' data analysis workflows. Following the visual analytics path, which advocates for supporting the reasoning process itself, we quantify the degree to which interface design elements can be used to unobtrusively guide researchers towards applying verified and established analysis techniques in their research.

We first present novel visualization methods that were developed in response to analytic needs indentified through collaborative efforts in three concrete application areas. In neuroscience we enable faster interaction with diffusion tensor imaging (DTI) datasets by creating planar representations of the inherently 3D data. In proteomics we facilitate the visual collation of experimental data and existing protein interaction information and accelerate the discovery process by uncovering and supporting elements of the proteomic analysis workflow. In genomics we increase the accessibility of analyzable visualizations of microarray data and eliminate the overhead of creating visualizations and learning new systems by implementing and evaluating a novel data distribution method.

Finally, we use the concepts of persuasive technology and "choice architecture" which state that a user of a system can be unobtrusively guided towards behavioral patterns that are more efficient, in terms of self-assumed goals, by slight alterations in the system interface. We provide quantitative experimental support for the hypothesis that we can use subtle changes in the interfaces of visual analysis systems to influence users' analytic behavior and thus unobtrusively guide them towards improved analytic strategies. We posit that this approach may facilitate the use of visual analytics expertise to correct biases and heuristics documented in the cognitive science community.

Improved Scientific Analysis through Domain Driven Visualization and Support for Analytic Deliberation

by
Radu Jianu
B. S., Polytechnic University of Timisoara, 2005
Sc. M. Brown University, 2007

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2012

This dissertation by Radu Jianu is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date ______________ ____
David H. Laidlaw, Director

Recommended to the Graduate Council

Date ______________ ____
Odest Chadwicke Jenkins, Reader

Date ______________ ____
Ben J. Raphael, Reader

Approved by the Graduate Council

Date ______________ ____

Dean of the Graduate School

Acknowledgements

I would like to thank the following people who have more or less indirectly contributed to the writing of this dissertation.

My advisor, David H. Laidlaw, for teaching me to think like a scientist, to find and solve problems worth solving, and for providing a stable and nurturing environment in which I could develop both as researcher and person. David will now continue to guide me in my academic career by serving as a role model for good teaching, research, and advising. I cannot imagine a better person to have worked with closely for so many years.

My undergraduate advisor, Adrian Rusu, for instilling in me the desire to pursue a graduate degree and an academic career, and for his initial guidance that made my years here at Brown possible.

The readers of this dissertation, Ben Raphael and Chad Jenkins, for their great feedback and helpful comments.

Cagatay Demiralp for being a great collaborator, co-author, friend and for not being too upset about the way I spelled his name. My collaborators outside the department, in particular Arthur Salomon and Christophe Benoist, for helping me find and solve the interdisciplinary problems featured in this dissertation.

My very close Providence friends for helping me recharge my batteries between coursework, paper submissions, and project deadlines. Thank you Misha, Wenjin, Aparna, Aggeliki, Babis and Olya for many good times. Without your company, my time at Brown would have been unbearable (even if perhaps shorter). At the same time I'd like to thank the Providence Tango community for helping me forget about work during unforgettable dance nights.

Lastly, but perhaps most important of all, I would like to thank my family: my wife, Doria, and my parents. Doria, for joining me in this adventure, for putting up with late work nights, and for being a true life companion. My parents, for being there whenever I needed them and for continuously supporting my decisions.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

The aim of this dissertation is to further visualization research by introducing novel techniques that let researchers hypothesize about their scientific data faster and more accurate than before. First, we use domain driven design to create data visualizations that enable new or accelerate existing data analysis workflows in three specific domains — neuroscience, proteomics, and genomics. This allows researchers to extract insights from their data more efficiently. Second, we want to support the process of distilling these insights into actionable hypotheses by quantifying the extent to which interface elements can be used to align scientists' analysis strategies to verified problem solving techniques.

Data visualization explores how to represent complex data graphically such as to maximize the ability of the human visual system and intuition to extract meaning from it. Visualization is therefore by definition tied to data, and a visualization's value often judged by its ability to improve the scientific workflows of researchers from different domains. This is why visualization perfectly embodies Fred Brooks' perspective on computer scientists as toolsmiths: "...hitching our research to someone else's driving problems, and solving those problems on the owners' terms, leads us to richer computer science research" [27]. In accordance with this vision, an important part of this dissertation is dedicated to introducing and evaluating novel visual representation and interaction techniques that help scientists understand their data better in a few concrete application areas: neuroscience, proteomics, and genomics.

However, visualization is a research area in its own right. Novel approaches that support the core of the scientific analysis process and are generalizable across multiple domains have broad, long-lasting impact and define visualization as a stand-alone research area. As such, this dissertation combines visualization contributions that are specific to the above mentioned domains, with a general approach that allows system designers to guide scientists, regardless of their domain and visualization they use, towards improved analytic strategies. The three domain-specific contributions and the final general one correspond to four goals that in concert implement the dissertation's aim of enabling researchers to more efficiently derive and test hypotheses from their scientific data.

The first three goals are approached following the traditional visualization method. We start by interviewing domain experts and observing their workflows to identify shortcomings in tools they commonly use to analyze their particular scientific data. We then introduce novel visualization methods to alleviate these shortcomings. Finally, we demonstrate, through formal or informal evaluations with our collaborators, that the developed methods let researchers understand the data faster and with less effort than before.

Using this methodology we first allow neuroscientists to more efficiently interact with and understand diffusion tensor imaging (DTI) datasets. We create abstract planar representations of DTI datasets and link them to traditional 3D models of the same data. We show, both quantitatively and qualitatively, that this accelerates interaction tasks that are at the heart of white matter analysis.

Second, we enable proteomic researchers to relate their experimental data to the existing body of knowledge with the hope of accelerating the process of discovery. To this end, experimental data is visually collated with existing protein interaction information in ways that speak to proteomic researchers' intuition. Annecdotal feedback indicated that our methods accelerated previous workflows for analyzing proteomic experimental data significantly.

Third, we increase scientists' access to visual representations of data and eliminate the overhead of installing and learning new applications, and creating visualizations. Specifically, we present a novel way of disseminating data as large pre-rendered visualizations distributed via the Google Maps API. We evaluate this approach on genomic data but show how it can be generalized to other domains. Annecdotal feedback showed that this simplified mode of data access prooves particularly useful for exploring new datasets, in casual settings, and by users who are not computer savvy.

Finally, we aim for an improvement of the scientific analysis processes, independent of particular scientific fields and visualizations, by quantifying the degree to which interface design choices can lead to improvements in analytic strategies used by researchers. This last goal extends the visual analytics reasearch agenda which, amongst others, advocates for the improvement of the reasoning process itself. Inspired by persuasive technology [64] and Sunstein and Thaler's "nudge" concept [159], which state that a system's design can unobtrusively guide a user towards behavioral patterns that are more efficient, in terms of self-assumed goals, we show how we can use subtle changes in the interfaces of visual analysis systems to unobtrusively guide users towards analytic strategies that cognitive science deems as more efficient.

The dissertation exemplifies how visualization and interface design support scientific discovery workflows from raw data interpretation to hypothesis elicitation and testing. The four goals that compose this dissertation illustrate the generality continuum spanned while supporting scientific data anlysis workflows in their entirety. Discovering efficient data representations and interactions is often data and domain specific. Our first two goals (Chapters 2,3) demonstrate how a tight collaboration with domain experts can lead to novel visual representations and interactions and reveal shortcomings and improvements to existing ones. Making these representations available to scientists (e.g. stand-alone systems, web applications) represents an opportunity for research that is less domain dependent (Chapter 4). Finally, hypothesis elicitation and testing, while sometimes influenced by domain specific particularties and constraints, can often be thought of in abstract terms using cognitive science and problem solving principles (Chapter 5). In concert, the four contribution areas presented in this dissertation support a significant part of the scientific discovery process.

1.2 Overview and Contributions

The broad contributions of this dissertation are novel visualization methods that improve analysis in proteomics, genomics and neuroscience and a better understanding of how we can leverage user interface design to improve scientific data analysis. This section provides a short introduction to these four application areas and a break-down of specific contributions.

1.2.1 Visualization Contributions in Neuroscience

White matter connects gray matter regions in the brain and is composed of bundles of myelinated axons. White matter research has important clinical applications as changes in white matter are associated with a wide range of neurological diseases. Diffusion Tensor Magnetic Resonance Imaging (DTI) enables the *in vivo* reconstruction of white matter as collections of 3D integral curves that reflect the paths of the white matter bundles. Due to the intricacy of the connectivity in the brain, such 3D models are visually dense, making it difficult for practitioners to identify and interact with anatomical and functional structures.

To address this, starting from the inherently three dimensional data, we create simplified planar representations that summarize important properties of the data but are more suited for interaction, given their two-dimensionality. We then link these representations to traditional 3D curve models such that all visualizations can be used in concert for a better understanding and manipulation of the data. We demonstrate in a small quantitative user study that this synergy accelerates interaction with white matter datasets. This work was published in [85, 88, 90, 70, 87, 89].

Contributions:

- Faster interaction with and better understanding of 3D models of white matter structures in the brain by linking them to two-dimensional abstractions defined from the same data
- A novel two-dimensional representation of white matter data that has the desirable properties of low-dimensional representations (e.g visual clarity or ease of selection) while preserving anatomically meaningful coordinates
- An evaluation that quantifies the benefits of augmenting a 3D white-matter model of DTI data with linked planar representations
- Improved accessibility to white-matter visualizations for browsing purposes by disseminating them using the Google Maps API.

- A concrete visualization system for analyzing white-matter datasets

1.2.2 Visualization Contributions in Proteomics

Proteins within a cell interact to regulate the cell's activity. Cascades of protein interactions peculiar to specific cells or cellular outcomes are called signaling pathways. An in-depth understanding of these pathways will let researchers discover efficient drugs that influence a cell's behavior without causing unwanted side-effects.

Experimental data is an important component in understanding how signaling pathways function. However, to efficiently interpret experimental results, they need to be collated with existing knowledge that can explain experimental observations and provide additional insight. One of the most common such data types are protein-protein interactions stored in public databases.

We present visualization techniques and design guidelines for combining interaction and experimental data in ways that harness proteomic researchers' intuitions. These include: integrating protein interaction networks into familiar signaling pathway images instead of drawing them as general graphs; enabling interaction level analysis of dense networks; and driving exploration by comparative analysis of multiple experimental datasets. An evaluation with domain experts motivates and demonstrates the utility of this work. The results were published in [86, 122, 95, 180].

Contributions:

- Design guidelines for visualizing protein interaction networks and experimental proteomic data together

- An anecdotal evaluation with domain experts revealing proteomic analysis workflows

- A method for augmenting static pathway diagrams with dynamic interaction information

.

1.2.3 Visualization Contributions in Genomics

Scientists today have access to many large datasets that describe biological processes. Advanced systems for visualizing such data exist but have associated costs that depend on a scientist's computer abilities and familiarity with the data type and content. Thus, when handed unfamiliar datasets, researchers often assess the time commitment these require and determine whether the analysis costs are justified. This may lead to wasted analysis time or ignoring of potentially useful data.

In this context we explore the benefits of using the Google Maps API, a pan-and-zoom interface that is well supported and highly familiar, to distribute raw data along with pre-rendered visualizations derived from it. We use five concrete visualization examples to show that integrating Google Maps with established visualization techniques offers a low-overhead way of exploring a dataset to assess its relevance and facilitates lightweight analyses of datasets outside a researcher's immediate focus. A collaborative design process and evaluation revealed that the pre-rendered browser approach works well in the genomic domain. We hypothesize that this distribution model may be extended to other application areas as well. This work was published in [91, 92].

Contributions:

- The concept of disseminating genomic data as precomputed visualizations using the Google Maps API
- An anecdotal evaluation showing the advantages and disadvantages of this approach
- Five examples of specific visualizations and their evaluation with domain specialists
- Design elements, challenges and opportunities when working with pre-computed visualizations and the Google Maps API

1.2.4 Improving Scientists' Analytic Strategies through User Interface Changes

Ample cognitive science research revealed that human thinking is subject to heuristics and biases that often lead to suboptimal problem solving [76]. A proposed solution in the fields of behavioral economics and human-computer interaction relies on designing choice-layouts (i.e., how choices are presented to consumers) and computer interfaces that "nudge" users towards decisions that increase their chances to make choices in their own interest.

We extend this "nudge" concept into the visualization domain by providing experimental support for the hypothesis that subtle changes in user interfaces of visual analysis systems can unobtrusively steer researchers away from cognitive biases and heuristics. Specifically, we report results from a controlled study in which subjects were asked to complete three analysis sessions using a system consisting of a visualization and an analysis support module. Two sets of non-functional changes were made to the analysis support interface before the second and third sessions. These changes were designed to improve three hypothesized or observed analytic deficiencies. Results of the study show that the changes succeeded in alleviating the targeted deficiencies. This work was published in [94, 93].

Contributions:

- An evaluation that quantifies the effect of design variation on analytic performance
- Qualitative observations about users' analytic strategies in a network analysis task

1.3 Background and Motivation

This section motivates the work presented in this dissertation by relating it to previous approaches and findings in the fields of visualization, visual analytics and cognitive science. Subsequent chapters will describe related work pertaining to each specific topic in more detail.

Visualization leverages the human visual system to enhance our ability to process large amounts of data and to facilitate cognition. Visual representations of information allow analysts to perceive patterns in the data, see data in context, and draw comparisons. In their anthology "Using Vision To Think", Card, Mackinlay, and Shneiderman [32] describe how visualization supports the process of sensemaking, in which information is collected, organized, and analyzed to form new knowledge and drive analysis.

Visualization draws from the interplay between graphically represented data and a human's perception and analytical abilities and is as such intrinsically tied to the users it aims to help.

In accordance with this view, an important part of this dissertation is dedicated to introducing novel visualization techniques and refining existing ones to help scientists in a few concrete application areas interact with their data better than before. We found that feedback on existing visualization methods and analysis workflows from domain experts in neuroscience, proteomics and genomics revealed important shortcomings that we could address. Specifically, we discovered that the traditional mode of visualizing and interacting with data representing white matter in the brain [19, 118] can be improved by providing alternative, abstract views of the same data alongside the original representation. We found that simply drawing protein interaction networks as general graphs, as it was done previously [82, 148], does not correspond to how proteomicists think of protein pathways. Finally, we found that when it comes to making visual representations accessible to end-users, feature-rich and highly adaptive environments [113, 102, 152] are not necessarily optimal.

As visualization is applied to increasingly complex problems and data, simply immersing the scientist in a visualization with the hope of discovering the unexpected becomes unfeasible. The primary product of visualization tools is insight [155, 150] defined by [141] as "an individual observation about the data by the user, a unit of discovery". Insights can be used as evidence or in best cases as hypotheses seeds, but they are rarely full fledged, testable hypotheses. Visualization can thus be regarded as a support mechanism for generating evidence. This restrictive view on the role of visualization implies that users are left to their own devices when aggregating scattered pieces of information into high-level hypotheses.

This should be regarded as a limitation because significant evidence from cognitive science suggests that human thinking is subject to heuristics and biases that deviate from normative rationality and lead to erroneous analysis. Such effects occur in all stages of analysis and many have not only been demonstrated, but also quantified by controlled psychological studies: Wason's 2-4-6 study [174] shows that hypothesis confirmation is used instead of the normatively correct hypothesis disconfirmation; Simon [151] reveals "satisficing", a heuristic that limits analysis to a hypothesis that is good enough. Dunbar [54] shows that such effects hold in scientific analysis settings as well.

Fortunately, evidence also suggests that reasoning can be improved by following rational recipes for analysis and external aids that amplify cognition. For example, Dunbar [54] shows how bias toward seeking confirming evidence can be overcome, and describes how analogy and unexpected findings often lead to consideration of multiple hypotheses in scientific domains [55]; Savikhin et al [144] uses a specific example from economic reasoning to prove that visualization can help overcome error-prone heuristics used in decision making.

In the field of visualization these aspects have been recognized and investigated by visual analytics, a sub-field fueled by growing intelligence needs after 2001. *Illuminating the Path* [165] introduced and defined this emerging field as "the science of analytical reasoning facilitated by interactive visual interfaces". Since its introduction visual analytics has advanced science with both theoretical and applicative results. Examples of the former include a five-stage sense-making model [140, 23] derived through Cognitive Task Analysis (CTA) or valuable insight into the workflows of collaborative sense making [138, 84]. Representative of the latter are a plethora of applications that probe the feature and design space of analysis-support software such as The Analyst's Notebook [124], The Scalable Reasoning System (SRS) [132] or Entity Workspace [22]. Most of these systems offer a large set of evidence and hypothesis management features which is likely to increase the cognitive span of the users and allow them to make associations that they couldn't make before.

The work presented in this dissertation complements such research by using a visual analytics methodology to create a link between observed analytic deficiencies and corrected behavior. Although cognitive biases and the need to leverage cognitive science expertise to alleviate them had been recognized within the field [73, 158], few visual analytics attempts have been made to bridge the gap between descriptive analysis (i.e., how humans actually analyze problems) and normative analysis (i.e., rational strategies of analysis). Here, we test the hypothesis that careful design of features included in a visualization system can unobtrusively guide users towards normatively correct analysis.

This approach was inspired by Thaler and Sunstein's work on libertarian paternalism [164] and the idea of "choice architecture". The authors rely on the assumption that anyone who designs how choices are presented is necessarily influencing decision-making behavior, and advocate for designing choice structures that "nudge" users to make decisions in their best interests. We used these concepts to demonstrate that interface design can be leveraged in a targeted way to guide scientists towards using an analysis system more, to pursue multiple hypotheses in parallel, and to gather more evidence per hypothesis.

1.4 Road Map

The document is broadly structured as follows. Chapters 2-4 share a similar structure and describe visualization contributions in three specific areas: neuroscience, proteomics and genomics. Chapter 5 describes the results of a user study which tests whether interface nudges can be leveraged to guide users towards rational analysis strategies. The dissertation ends with a concluding chapter. Below is a detailed description of this structure.

Chapters 2-4 describe visualization contributions made in neuroscience, proteomics and genomics and share a similar structure. Each chapter starts with an introduction to the problem, motivation of its significance and overview of contributions. A detailed related work section that contrasts the methods presented in this dissertation to previous approaches follows. Next are design choices and methods. Then we describe the evaluation procedure and the results obtained. Each chapter ends with a discussion and conclusion.

Chapter 2 demonstrates that linking 3D stream-tube models of DTI datasets to planar abstractions derived from the same data can accelerate interaction with and exploration of the datasets. It also describes novel and effective planar representations, and an application that can be used for DTI analysis. Chapter 3 introduces methods for collating publicly available protein interaction data with experimental results in ways that support workflows and intuitions of proteomic researchers. Chapter 4 shows how scientific data can be disseminated as a variety of pre-computed visualizations served through Google-Maps and presents anecdotal feedback from domain experts that demonstrates the usefulness of this approach.

Chapter 5 describes how software interface elements can be used to correct analytic behavior that is affected by cognitive biases and heuristics. We first motivate our approach by citing ample cognitive science research and relating it to existing efforts in visualization and visual analytics. We then describe a user study that tests our hypothesis and present its results. We end with a discussion and conclusion.

Chapter 6 concludes this dissertation with a reiteration of contributions, a statement on the impact of the presented work, a few discussion points pertaining to the dissertation as a whole, a description of potential future directions, and a short summary.

Chapter 2

Planar Exploration and Analysis of 3D White Matter Tractograms

Diffusion Tensor Magnetic Resonance Imaging (DTI) enables the exploration of fibrous tissues such as brain white matter and muscles non-invasively in-vivo [18]. It exploits the fact that water in these tissues diffuses at faster rates along fibers than orthogonal to them. Integral curves that estimate fiber tracts by showing paths of fastest diffusion are among the most common information derived from DTI volumes. Such curves are generated from DTI data by following the principal eigenvector of the underlying diffusion tensor field bi-directionally and are commonly referred to as fiber tracts. Sets of DTI tracts are known as tractograms and their study is called tractography.

In this dissertation we discuss DTI visualization in the context of white matter tractography. White matter in the brain ensures the connectivity between various regions of gray matter and is composed of bundles of myelinated axons. White matter tractography has important applications in both clinical and basic neuroscience research as lesions in white matter are associated to a wide range of neurological diseases.

DTI curves are often visualized as 3D models composed of streamlines or variations of streamlines (streamtubes and hyperstreamlines) in 3D [117, 182]. Reflecting the intricacy of the connectivity in the brain, these 3D models are generally visually dense and, with increasing DWI resolutions, this complexity is bound to become greater. It is thus often difficult for practitioners to see tract projections clearly or identify anatomical and functional structures easily in these dense curve collections. Typical interaction tasks over tracts, such as fine bundle selection, are also difficult to perform and have been a focus of recent research [8, 7].

In this context, we present a novel interaction paradigm and demonstrate both qualitatively and quantitatively, that it can accelerate the exploration of white matter tractograms. Starting from the inherently three dimensional data, we create abstract planar representations. These representations summarize important properties of the data and are suitable for interaction, given their two-dimensionality. We link such abstract representations to traditional 3D tractogram models through interaction, such that operations performed in one of the views are mirrored into the others. Users can thus create a mental mapping between the different modes of representation and use them in concert for a better understanding and manipulation of their data.

In a first developmental iteration, given a tractogram, we linked conventional 3D white matter tractograms to a planar embedding and a hierarchical clustering tree (see Figure 2.1). Both 2D visualizations are representations of a similarity matrix obtained by computing pairwise "distances" which reflect the geometrical similarities between fiber tracts. The planar embedding is obtained by considering each fiber tract to be an individual 2D point and placing it on a drawing canvas such that the distances between points approximately reflect the distance relations between their corresponding fiber tracts. The hierarchical clustering tree representation, or dendrogram, is obtained by applying the average linkage hierarchical clustering algorithm on the similarity matrix. As shown in Figure 2.1, these two abstract representations can be linked to a traditional 3D tractogram model implicitly through interaction and explicitly through a perceptually uniform coloring.

This first iteration was evaluated by interviewing experts and gathering feedback in an informal setting. Results suggested that this type of coordinated interaction has the potential to enable faster and more accurate manipulation of dense fiber tract collections. Work done concurrently on low dimensional brain representations and described in [42] confirms our findings. The authors there quantitatively compare linked views of low dimensional representations of DTI data and traditional 3D models to several state-of-the-art DTI visualization systems.

The main drawback of such abstract representations, as found in our first evaluation session, is that they lack an explicit anatomical interpretation. This means that little or no spatial correlation can be found between the abstract views and the anatomical views. It is therefore challenging for practitioners to create lasting mappings between abstract representations and their corresponding 3D tractograms, even in the presence of non-spatial links between the views (i.e., interaction and color).

Motivated by this problem, we introduce two-dimensional neural maps which package desirable properties of low-dimensional representations into views that preserve meaningful anatomical coordinates. Starting from a hierarchical clustering of a white matter tractogram and a given clustering cut, bundles of tracts and their corresponding centroids are computed. These centroids are then projected along the three principle projection planes: saggital, axial and coronal. The result is a set of projected neural paths in the plane, similar to illustrations in medical textbooks. We link these two-dimensional path projections to the original 3D white matter model as shown in the interactive system illustrated in Figure 2.2.

We assess the usefulness of neural path projections in two consecutive studies, one anecdotal and one quantitative. Anecdotal study results indicate that this new representation is intuitive and easy to use and learn. Results of the quantitative study show that users are faster and more confident with the neural path projections than with traditional 3D interaction or with linked abstract planar representations.

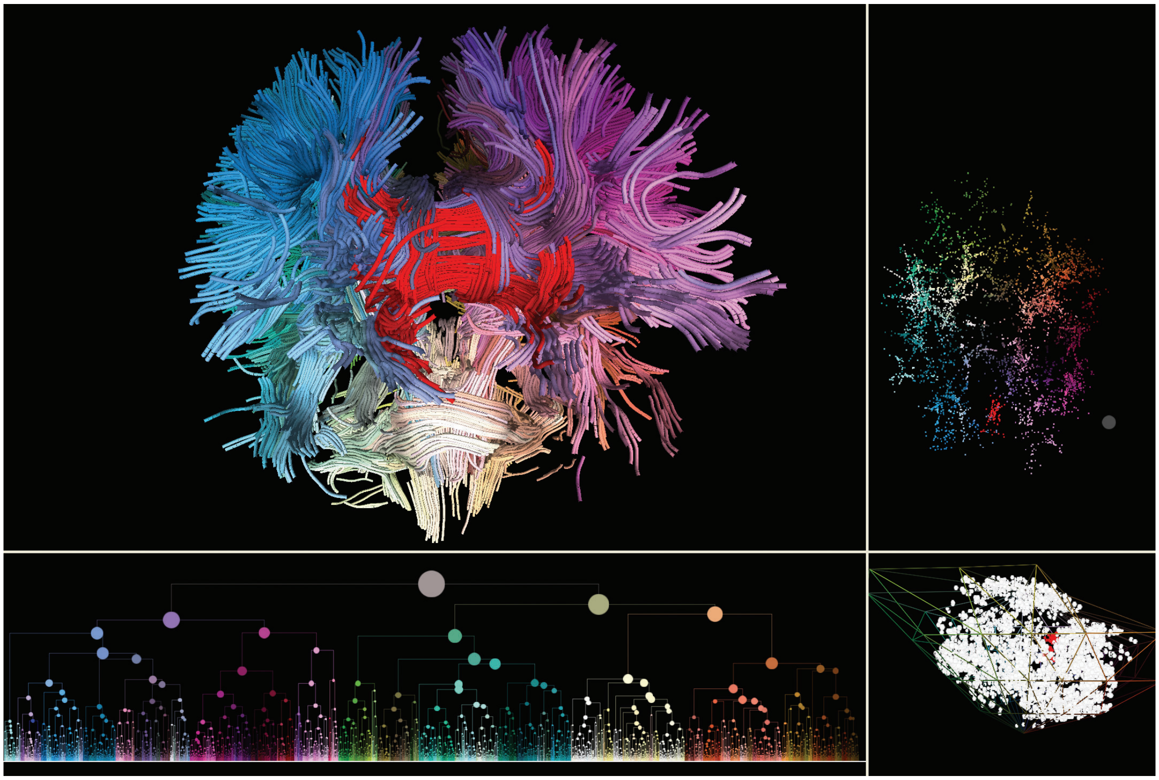


Figure 2.1: Coordinated DTI tractogram model exploration in lower dimensional visualizations: 2D embedding (upper-right), hierarchical clustering (lower-left), and L*a*b* color embedder (lower-right). A selection of a fiber-bundle (red) in the hierarchical clustering is mirrored in the other views.

2.1 Related Work

Here we discuss existing techniques that the present work builds on: techniques for visualizing and interacting with DTI datasets, methods for visualizing similarity relations and the use of multiple, coordinated views for visualization.

2.1.1 Visualizing and Interacting with DTI Datasets

The most commonly used technique to visualize DTI data is streamline tracing; in DTI-specific literature this is also called fiber tracking [117] or tractography [19]. This method is used in our 3D DTI visualization.

Interacting with streamline DTI models is not trivial. A common interaction task is the selection of fiber bundles. This is usually done directly on the model by placing 3D regions of interest (ROIs) along the presumed path of the desired bundle and then having the application select fibers that intersect those ROIs [33, 172, 112]. More recently, Akers et al. [7] introduced a sketching and gesture interface for pathway selection: the user paints a 2D freehand stroke and the selection algorithm selects tracts that cross the brush path. Finally, concurrent research by Chen et al. [37] also links

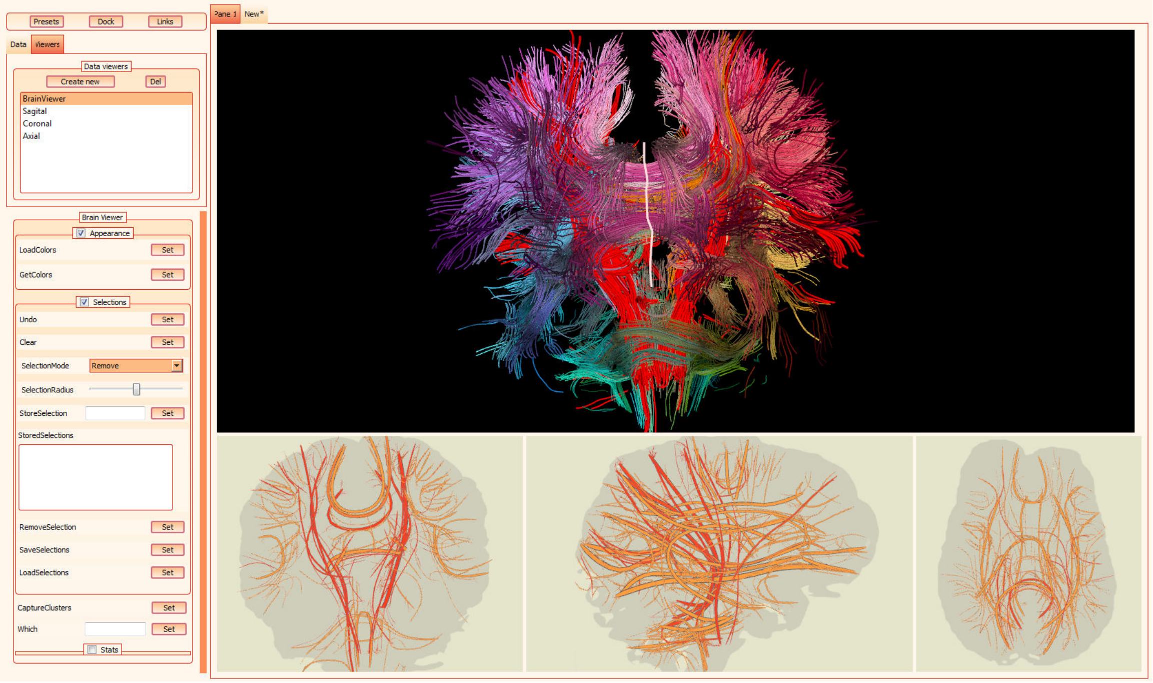


Figure 2.2: An interactive analysis system using linked views and planar tract-bundle projections. Three planar representations, along the coronal, transverse and sagittal planes (bottom panels), are linked to a 3D stream-tube model (upper left) and a 2D point embedding of tract similarities (upper right). Selections in the projection views can be performed by clicking or cutting across cluster curves and are mirrored in the 3D view. Points corresponding to the selected tracts are interactively embedded into the plane and used to refine selections at tract level.

2D embeddings to DTI datasets and finds that it accelerates interaction. The work presented in this dissertation differs by incorporating hierarchical clustering trees, using perceptually variable coloring to link views, and, most importantly introducing neural path projections as a novel type of low dimensional representations that maintain an anatomical framework.

Automatic DTI fiber clustering methods have been developed to support DTI model interaction and visualization. For a review of such methods consult [116]. Fiber clustering relies on a similarity metric that captures the geometric similarity between integral curves. For example, closest point measures like the Hausdorff distance [41], and the Fréchet distance [9] only measure the Euclidean distance between two selected points on a pair of curves. Conversely, the average point-by-point distance between corresponding segments defined in [51], the mean of closest distances defined in [41], and the mean of thresholded closest distances defined in [182] summarize all points along two curves as the mean Euclidean distance along their arc lengths.

Fiber similarity can be mapped to color as was first done in [28] by assigning distinct colors to clusters, and more recently in [48] by immersing a 3D embedding into the L*a*b* color space.

2.1.2 Visualizing Similarity

Visualization literature describes several methods for conveying similarity relationships between entities. Most of them have been researched in the context of multidimensional visualization, where the distance is derived from the position of a point along each dimension. However, a subset of these methods can be used for entities over which an arbitrary similarity function is specified. In the following, we will only review this category. For a more detailed discussion on multi-dimensional visualization techniques, Keim [100] provides a good overview.

An intuitive way of making distance apparent is by using a scatterplot. In its simplest form this method can only be used for data with at most three dimensions and explicit vector values. **Multi-dimensional scaling** (MDS) techniques can overcome this limitation. They attempt to map the multi-dimensional points to a visualizable lower dimension while preserving distance relations between points.

So called non-linear MDS methods are suited for computing representations when distances between points are given explicitly but coordinate values for the points are unknown, as is the case in the tract similarities computed as part of this research. These methods use the distance between data points to define an error measure that quantifies the amount of distance information lost during the embedding. Gradient descent or force simulation is then used to arrange the points in the low dimensional space so as to minimize the error measure. A good example of such an approach is Force Directed Placement (FDP) [68] originally proposed by Eades [57] as a graph drawing approach. It simulates a system of masses connected by springs of lengths equal to the distances that need to be embedded. The points are initially placed at random and are then iteratively moved by displacements derived from forces computed by Hook's spring laws. After a number of iterations the spring system will reach a local minimum energy state that represents the resulting embedding. We use this method as part of our work.

An iteration of the original FDP model is $O(n^2)$, and since at least n iterations are necessary to reach equilibrium, the final complexity is $O(n^3)$. This makes the computation for high-resolution, complete brain models expensive. One method that addresses this problem is called Force Scheme. Proposed by Tejada et al. [161], it reduces the overall complexity to $O(n^2)$ by requiring fewer iterations to reach the final state. A complexity of $O(n^{5/4})$ was achieved by Morrison et al. [119] by creating a hybrid model based on approximations using samples and interpolations. In this dissertation we use another algorithm, with linear iteration time, developed by Chalmers [34].

A MDS can be used in conjunction with a perceptually uniform color space to **display similarity as a color cue**. We use this technique to reflect the variation of tract similarity as a perceptual variation of colors: similar tracts receive perceptually similar colors while dissimilar tracts get perceptually distant colors. A color space is said to be perceptually uniform if the perceptual difference between any two colors in just noticeable difference (JND) units is equal to the Euclidean distance between the two colors in that color space. The L*a*b* color space is perceptually uniform and thus a 2D or 3D embedding can be immersed into L*a*b* to obtain a similarity color coding. It should be noted, however, that the perceptually uniformity in the L*a*b* is an empirical approximation and assumes a particular calibration setting for individual monitors.

A **dendrogram** is another method for visualizing similarity that does not require explicit vector values for points and as such is suited for displaying tract similarity. It is a tree-like visual representation of results produced by hierarchical agglomerative clustering algorithms ([15, 99]). Because they are used in a wide range of scientific domains they have become intuitive tools for many scientists.

2.1.3 Coordinated Views for Visualization

Visualization techniques are usually task and data specific. Different views are therefore frequently used to show data from multiple perspectives, combine the strengths of any individual technique, and distribute the cognitive and interpretative load of complicated data and tasks across multiple views [16].

However, the task of aggregating the different views into a unitary single mental image factors in the complexity of the visualization itself [16]. This effect can be reduced by coordinating the content, appearance and behavior of the views [123]. This is achieved either implicitly, through coordinated appearance or behavior, or explicitly through visual cues, such as color or lines linking the separate windows. In this dissertation we use both approaches. Shneiderman [149] offers a good review on multiple-view coordination techniques such as brushing and linking or details on demand.

Multiple views applications are often used to aid in the understanding and exploration of complicated datasets. In [30], the authors show several examples of how brushing and linking techniques can be used to map a complicated data space into multiple simple views that, when explored together, convey the overview data picture. Gresh et al. [74] present an approach that links 3D visualizations to statistical representations to facilitate the effective exploration of medical data. XmdvTool [173] and Visulab [145] attempt to maximize a user's understanding of multidimensional data by linking multiple representational techniques such as scatterplots, glyphs, or parallel coordinates. Finally, work such as [129] and [26] propose domain independent, extensible multiple-view architectures that satisfy general requirements of the visualization domain.

2.2 Design Elements

Here we describe methods underlying the DTI interactions described in this chapter. To reiterate, these interactions build on the concept of linking traditional 3D visualizations of DTI datasets to abstract planar representations derived from the same data.

We first introduce the 3D stream-tube visualization which is used throughout the work. This incorporates established techniques and interactions. We continue with several types of planar abstractions that, when linked to traditional 3D tractograms, improve interaction and data understanding. All such abstractions are based on a geometrical similarity measure between 3D tracts, which is discussed first. The different types of abstractions will be divided into two categories, discussed in separate sections that follow. First, explicit visualizations of tract similarity are based on well established techniques but lack anatomical meaning. Second, a novel type of visual abstraction combines the strengths of low dimensional representation with a meaningful anatomical relation to the original 3D data. Finally, we describe a multiple-views system that makes this novel interaction paradigm available to neuroscientists.

2.2.1 A 3D Stream-Tube Visualization of DTI Data

Datasets used as part of this dissertation were 3D white matter tractograms text-formated as 3D poly-curves. Details on how raw DTI datasets were acquired and how tractograms were derived from them can be found in [85]. We display the tracts as 3D stream-tubes (see Figure 2.1).

The following interaction modes are available on the 3D model. The 3D sphere selection tool will select any tubes passing trough a sphere which the user can position with the mouse in the XY plane and with the mouse-wheel along the depth dimension. The size of the sphere is adjustable. Alternatively, the 2D brush tool allows a user to draw a freehand brush stroke over the 3D model to select tracts whose screen projections intersect the brush stroke. Given a current selection, a new set of selected tracts, generated with either the 2D brush or the 3D sphere, can be added to it, removed from it or intersected with it. Finally, the following novel operation is implemented: once a set of tracts selected, users can grow the selection by gradually adding tracts which are close to the current selection.

Selections performed on a specific brain model can be saved for future analysis. Moreover, statistics such as average tube length, number of tubes or average fractional anisotropy can be generated interactively on sets of selected tracts.

2.2.2 Similarity Between Fiber Tracts

The similarity between two tracts is quantified using the weighted chamfer distance discussed in [48]. This measure tries to capture how much any given two tracts follow a similar path, while giving more weight to the points closer to tract ends. Distances between each pair of fiber tracts are computed using $\lambda = 0.5$ as described in [48] and assembled to create a distance matrix. While this distance measure is a good approximation of the notion of similarity in the domain, the methods described in the following sections are independent of the particular distance measure used.

2.2.3 Explicit Visualizations of Tract Similarity

In the following three sections we describe three traditional ways of representing distance visually and how they can be adapted to the particularities of DTI datasets.

2D Point Embeddings

Planar embeddings of tract similarity are computed using Eade's [57] force directed method in concert with Chalmer's [34] acceleration technique which reduces the complexity of the computation to O(n) by using a sampling strategy.

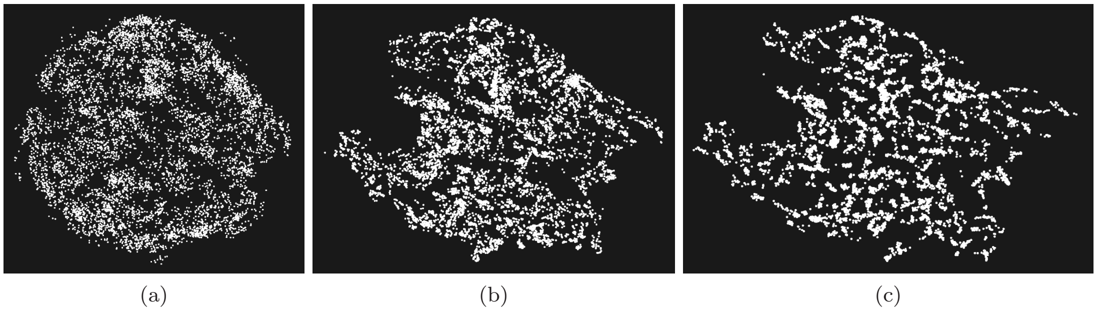


Figure 2.3: 2D tract embedding for different spring force settings. a) Spring force with absolute distance displacement. b) Spring force with absolute distance displacement, weighted by decay function and with repulsive force. c) Spring force with relative distance displacement, weighted by decay function and with repulsive force. In c) clusters are tighter making selection and understanding of manifold recognition easier.

For the force computation we use Hook's law $F = -k\Delta X$ where ΔX is the spring displacement and k is the spring constant. We experimented with variations of this force to obtain embeddings that are better suited for neurotract interaction and analysis: a sharper definition of clusters can improve bundle selection and manifold recognition while small distances should take embedding priority over large distances. We tried the following approaches: using squared distance to exaggerate large distances and make clusters more defined, using relative displacement instead of absolute difference in distance to give larger distances more arrangement flexibility, and using a combination of weighting forces with a factor inversely proportional to distance and adding a repulsive force between points. As Figure 2.3 shows, good visual results were obtained by combining relative distance displacement, forces weighted by a decay factor ($e^{-\sigma/d}$ with σ a decay factor and d the distance), and a repulsive force ($F_{rep} = k_{rep}/d_{embed}^2$, where k_{rep} is a constant and d_{embed} is the embedded distance) between all points.

2D point interactions include point selection and point coloring. Selection is performed by clicking and dragging; multiple selection can be performed to select points from non-adjacent regions. For coloring, the 2D coordinates of the embedding can be interpreted as the (a, b) coordinates in the L*a*b* color space, and, for a given luminance, colors can be attributed to points. The result is that close points will receive perceptually close colors. However, this color embedding is not ideal due to the particularities of the L*a*b* color gamut: it has an irregular shape and saturated colors close to the boundaries. The 2D coordinates need to be scaled to fit into the gamut and will thus occupy within the gamut a small, central region that corresponds to unsaturated colors.

A 3D Color Embedder

A better coloring can be obtained, as seen in Figure 2.1, by using a 3D color embedder. We compute an approximation of the L*a*b* color gamut, as visible on the lower-right panel of Figure 2.1, and use it as a container for force directed embedding. To avoid having to adjust a repulsive container force, which would likely need a hard-to-control, steep gradient, we perform a physically accurate simulation with container contact detection. The embedding begins in the center of the gamut and is gradually expanded until most of the space is filled. During implementation we observed that the largest distances are often embedded along the luminance axis (y-axis of color gamut). This is problematic because luminance offers little resolution and can be interpreted as a lighting effect. We therefore apply a ”flattening” force at the beginning of a simulation cycle to force large distances to lie in the horizontal plane. These force components, acting on the y-axis towards the center of the gamut, wear off as the embedding moves towards a steady state. The force computation used is the same as for the 2D embedding, with straightforward 3D modifications. In terms of interaction, the color embedder only supports collapsing and color grabbing.

Dendrograms

Dendrograms are visual representations of hierarchical trees obtained through agglomerative clustering. We use an average linkage clustering whereby the distance between two clusters is computed as the average of all inter-cluster distances. Minimum linkage does not give consistent results because of so called “broken tracts” introduced by the fiber tracking algorithms short tubes placed between major tract bundles will cause these bundles to be glued together by a minimum linkage algorithm.

To compute the tree layout we use the method described in [136]: for each subtree the layouts for the two child trees are computed recursively and placed next to each other aligned at the bottom; the root is then placed one unit above their bounding box and in the middle of its horizontal axis. For single node trees a unit bounding box is used.

The following interactions are implemented for dendrograms: multiple node selections, collapsing and expanding of individual nodes, or collapsing nodes automatically through cluster cuts.

2.2.4 Hierarchically Projected Neural Paths

2D point embeddings and dendrograms suffer from a major drawback as will be shown in the findings section: they lack an anatomical interpretation. This section describes hierarchically projected neural paths (see Figure 2.2) which is a type of representation that packages the strengths of abstract, low dimensional representations in an anatomical framework. The following two sections describe how hierarchically projected neural paths can be constructed.

Clustering and Projection

Hierarchically projected neural paths are schematic views of major tract bundles projected on a few selected planes. In this work the sagittal, coronal and transverse planes were chosen as the main modes of representation (see Figure 2.2).

We first compute a clustering tree using an average-linkage hierarchical clustering algorithm on the tract distance matrix (e.g., [53]). We choose the average-linkage criterion because it is less sensitive than the minimum-linkage to broken tracts due to tracking errors. We obtain a clustering of tracts by manually setting a cut threshold on the dendrogram. This threshold can be also

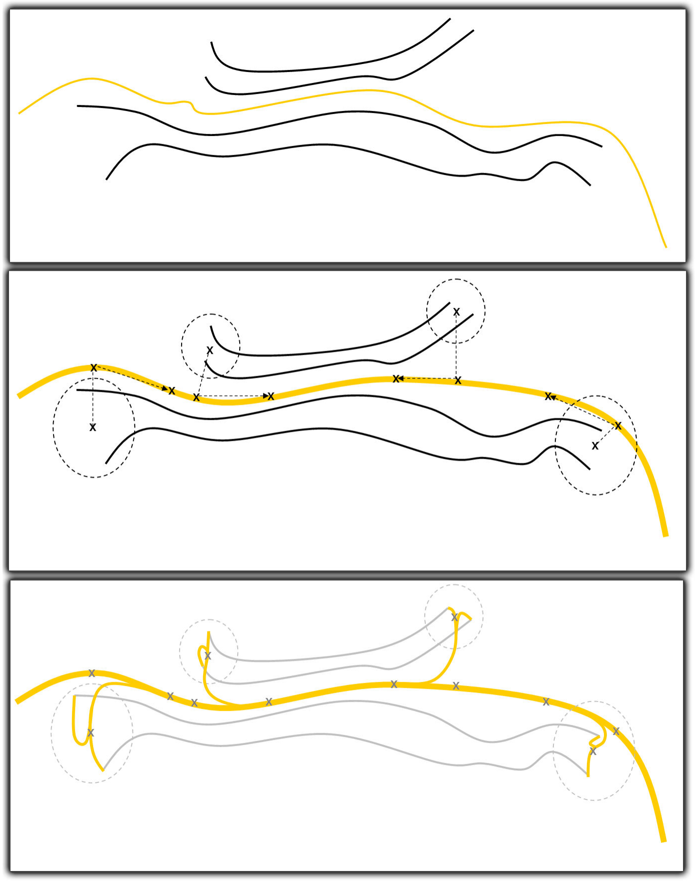


Figure 2.4: Schematic tract-cluster representation. (Top) 2D projections of a tract-bundle, with an associated centroid curve (orange), are determined from a hierarchical clustering of initial 3D tracts. (Middle) The centroid curve is smoothed by a spline and the endpoints of non-centroid curves are clustered using their initial 3D coordinates (four clusters); for each cluster, three control points linking the center of the cluster to the centroid spline are computed. (Bottom) Splines are run from each curve endpoint through the control points of its corresponding cluster.

interactively changed by users to control the coarseness of the clustering. A constant cut at 60% of the clustering tree's height gave consistent results across the six datasets we experimented with.

Next, we create simple orthogonal projections of tracts on each plane. We cull out tracts that do not contribute significantly to the projection. If the ratio of projected tract length to true tract length is under a threshold value, we remove the tract from the corresponding cluster. We set the culling threshold to 0.65 for the projections used in our experiments.

Finally, we compute a centroid for each cluster by choosing the tract with the smallest maximum distance to any other tract in the cluster. We found that for illustration purposes it is desirable to avoid broken tracts. We therefore weigh the centroid selection to favor longer tracts by dividing the maximum distance from each tract to any other tract by the tract's length.

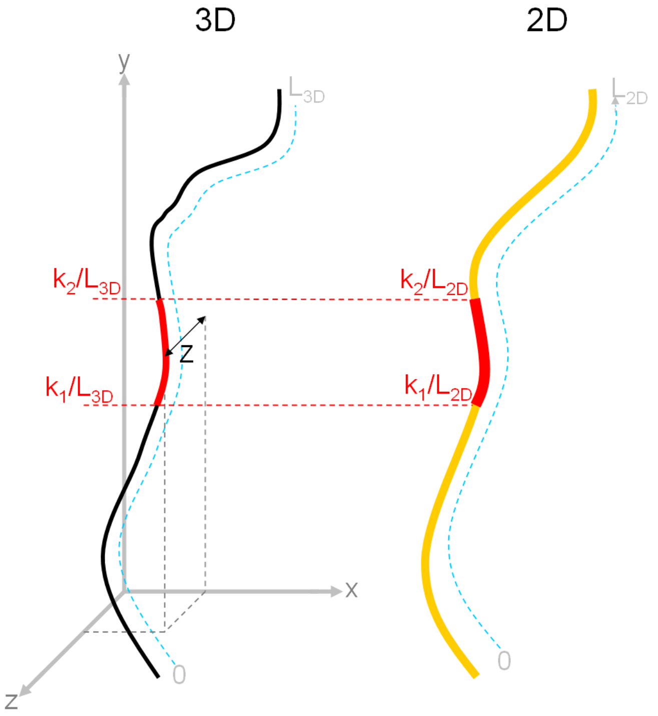


Figure 2.5: Depth ordering of 2D paths. For each segment of a 2D spline, we locate a corresponding segment on the 3D curve from which the spline was derived by traveling the same fractional distance along both curves. The depth of the 2D segment is the same as the depth of the middle of its corresponding 3D segment.

Visual Representation

We opted for an illustrative rendering of brain projections. Illustrative visualization uses abstraction to reveal structure in dense visualizations and to harness scientists' familiarity with textbook representations [170]. Both criteria apply to white matter tractograms: fiber bundles provide a natural abstraction of 3D anatomy that avoids the clutter of large streamtube collections, while textbook illustrations [72] shape the intuition of neuroscientists. These advantages have also been recognized and explored by Otten et al [127].

The rendering assumes a given clustering with assigned centroid tracts, which can be computed as described in the previous section. Our approach is inspired by Holten's hierarchical edge bundles [81] in attempting to group all fiber-tracts from a bundle into one, visually salient structure. However, hierarchical edge bundles operate on abstract connections that are unconstrained by concrete geometrical shapes. They can therefore be drawn according to visual aesthetics principles alone. Conversely, fiber tract paths play important anatomical functions and should be preserved in tractogram visualizations. To this end, we perform our bundling by routing tracts along the path of the most representative tract in their bundle. Thus, the centroid tracts will define a schematic neural skeleton on top of which the non-centroid tracts are scaffolded.

Projections of centroid curves are smoothed prior to rendering to achieve a schematic representation and to reduce clutter. This is done by sampling a number of evenly distributed control points (five in our implementation) along the tract projection and using them as control points for a spline. In our implementation the spline is piecewise cubic and consists of 30 segments. The thickness of a centroid curve is proportional to the square root of the number of tracts in the bundle.

Once centroid tracts are represented as 2D splines, endpoints of non-centroid curves are linked to their cluster's centroid spline following the procedure illustrated in Figure 2.4. First, the end-points of non-centroid curves in a bundle are clustered based on the end-points' initial 3D coordinates. Two endpoints are placed in the same cluster if the distance between them is less than 2 mm. Then, for each such endpoint cluster we compute three control points that link the geometrical center of the endpoint cluster to the centroid spline: the first point is the center itself, the second is a point on the centroid spline closest to the center point, and the third is determined by traveling from the second point down the centroid spline, towards each curve's other endpoint, for a predefined distance (e.g., half of the distance between the first two points). Ultimately, splines are run from each tract endpoint through its cluster's three control points, thus linking each endpoint to the centroid path. The thickness of these endpoint linkage splines gradually increases from unit thickness (i.e., single-tract thickness) at the tract endpoint to a thickness proportional to the square root of the endpoint cluster size, where it merges with the centroid spline.

We depth-order spline segments so that 2D centroid splines crossings can indicate the depth ordering of their corresponding 3D shapes. The depth ordering is done differently for centroid splines and non-centroid splines, since while centroid curves are close representations of actual 3D tracts, non-centroid curves are abstract representations obtained through the process described above. Furthermore, the depth ordering is approximate (as discussed in the following paragraph) and may produce artifacts.

For centroid splines, the depth of a spline segment is computed by finding a matching segment on the 3D tract from which the spline was derived, and taking the depth of that segment's center (Figure 2.5). The matching segment on the 3D tract has its endpoints at the same fractional distance from the start of the 3D tract as the 2D segment's distance from the start of the 2D spline. This per-segment ordering was chosen because of the intricacy of white matter tractograms. Tracts often wrap around each other such that a correct per-tract depth ordering cannot be determined. Treating each curve segment independently maximizes the probability that the 2D rendering remains truthful, at least within the resolution of the tract segmentation. Conversely, non-centroid splines are completely abstract 2D representations. The depth of any non-centroid spline is determined by averaging the depth of the corresponding 3D tract. Wrapping fiber tracts are therefore not captured by this latter process.

Finally, bundle-color, texture or thickness can be used as additional depth cues. While we have not fully integrated and evaluated such encodings in our current prototypes, we have experimented with color cues and found those to be useful.

In the following section, we give details on how we use 2D neural path representations as part of an interactive application and as standalone digital maps.

2.2.5 A Multiple-Views System for Exploring DTI Datasets

Using QT GUI and G3D graphics libraries we created a framework that allows for any of the previously described visualizations to be linked together. Figure 2.2 shows the application with a traditional 3D stream-tube view linked to three path projections. Operations performed in one view are mirrored in all other linked views. For example, selecting points in the 2D embedding will result in a selection in the brain model, while color grabbing in the 3D color embedder will cause tracts to receive the corresponding coloring information (see Figure 2.1).

In terms of the system's implementation, following an interaction or any change in its state, a visualization can broadcast a message that informs linked views that one of its properties has changed. Linked visualizations, depending on their implementation, can either act on such a message or ignore it.

We have recently also developed a digital map interface that coupled the projected path representations with the Google Maps API to enable web-accessible, ligtht-weight visualizations of DTI data. This mode of distribution is described in Chapter 4.

2.3 Evaluation and Findings

We evaluated the methods both anecdotally, by interviewing domain experts, and quantitatively by measuring subjects' bundle selection times as part of a formal user study. The results show that while planar abstractions in general are likely to accelerate the exploration of DTI tractograms, the hierarchical path projections introduced in section 2.2.4 offer the most significant improvement due to their anatomically grounded representation. Below we detail the evaluation procedures and results for both the anecdotal evaluation and the quantitative users study.

2.3.1 Anecdotal Evaluation

In a **first evaluation** we gathered feedback about the value of linking the explicit planar abstractions presented in section 2.2.3 to traditional 3D models. We showed our prototype to a group of experts, including one research neuropsychiatrist and three neuropsychologists. They were all interested in the relationship between fiber tracts and cognitive and behavioral function in the brain and have either seen or interacted with streamtube representations of fiber tracts before. A think-aloud protocol was used; we demonstrated the prototype using a projector while asking questions and collecting their feedback.

The experts agreed that the proposed paradigm can supplement the existing tools and would be particularly useful in accelerating the selection of tract bundles. They found the coloring method to be helpful and visually appealing, which was argued to be an important factor for adoption of a visualization tool. They found the hierarchical clustering tree to be more useful than the 2D scatterplot representation. One interaction scenario proposed was to select a rough region in the brain model using sphere selection and then gradually refine it in the hierarchical clustering tree.

This feedback was backed by concurrent research on linking planar representations to DTI datasets presented in [37]. The authors ran a quantitative study and find that a system which linked a scatterplot representation to a 3D tractogram lead to lower selection times of major tract bundles, as compared to several other leading DTI visualization systems.

On the downside there was concern that learning the correspondence between the 2D point-cloud representation and the actual fiber-tract collection can be non-trivial. Following this feedback the hierarchical path projections described in section 2.2.4 were developed. The goal was to package the benefits of dimensionality reduction techniques in an anatomically valid representation.

This new representation mode was evaluated in a **second anecdotal study**. Three neuroscientists took part in an informal evaluation: we demonstrated the prototype while asking questions and collecting participant's feedback. Two of the experts also tried both interfaces themselves by selecting a set of major bundles: the CC, cingulate bundle, uncinate anterior internal capsule, and the corticospinal tract. There was agreement that our new interface was significantly more intuitive and easier to use and learn than the abstract low dimensional representation.

2.3.2 Quantitative User Study

As noted in the previous section, authors in in [37] compare a system that links a 3D stream-tube model to a standard 2D point representation to several state-of-the art DTI visualization systems that don't employ linked planar views. It was therefore judged to be a valid baseline to compare the improved projected path representation against. Their interaction consists of a brush tool that works similarly to ours in 3D and as a lasso tool on the 2D point representation. Users are able to select tracts or points and then remove them or, conversely, remove everything else from a current selection.

The user study involved four subjects with general neuroanatomy knowledge and all had some experience with tractography visualization tools. The first subject was a neuroscience graduate student who had previously used Diffusion ToolKit (DTK) for six months. The second subject had five years experience in diffusion MRI clinical research and had used BrainApp, Slicer and TrackVis. The third subject was a biomedical engineering graduate student and had significant tract selection experience using BrainApp. Our last subject was a computer science graduate who was developing automatic algorithms for white matter analysis. Two of the users were male and two female. All users were right-handed.

The user study involved the timed selection of three major bundles in two distinct datasets, using the two systems. The three targeted bundles were the bilateral cingulate bundle, the bilateral corticospinal tract, and the right superior longitudinal fasciculus. The order in which the systems were used was alternated: two of the subjects started with the projected path representation while the other two were asked to first use the 2D point embedding system. For each system, users were first given a brief description of the underlying visualization concepts and were shown a brief demo.

	time (secs)				confidence			
	cb	cs	slf	mean	cb	cst	slf	mean
2D point	227	361	234	274	4.1	3.3	3.1	3.5
2D path	136	165	215	172	4.1	3.8	3.7	3.9

Table 2.1: User performances on bundle selection task.

They were then trained on the same three bundles as they would use in the real task, but on two different datasets. Following training, they were asked to select the three bundles while being timed on each selection. After each selection they also provided a five point subjective confidence estimate for their selection. This methodology aimed to capture users' performance once they have already developed strategies for bundle selections, and as such more closely model what would happen over extended use. After completing the task on both systems, subjects were asked to complete a post-questionnaire in which they provided qualitative feedback on their experience.

Results from the quantitative study are conclusive. The average times and confidence measures for each user, over all datasets and tract bundles are shown in Figure 2.6 using a paired t-test (projected path measures subtracted from 2D points measures). As seen, there is a significant drop in selection time using the novel projected path representation. Results are less conclusive for the subjective confidence measure, with two of the differences lying within standard error. Table 2.1 summarizes users' overall and per-bundle mean performances on each tool.

In addition to the quantitative measurement, by observing our users' actions, we distinguished several typical behaviors. Two distinct selection strategies were used in the projected path visualization. Two of the users consistently brushed over large areas of the projection to ensure that the targeted bundle was selected and then relied on the 3D view to clean up the selection. The other two users aimed for fine selections in the 2D projections and then inspected the 3D view to determine whether any fibers escaped the selection. They added the missing tracts using short, targeted brush strokes and then removed tubes that were erroneously added during this operation. These users seemed to have a better understanding of the mapping between the 3D view and the 2D projections which perhaps explains the difference in strategies.

All subjects used the 2D point representation relatively rarely. The most common operation was to remove points they were completely confident were not part of the selection (e.g half of the brain, peripheral U-shaped bundles). However, as one of the users noted, in absence of a clear mapping between the views subjects were hesitant to perform bold operations in 2D. This confirms concerns expressed in the first anecdotal evaluation.

In a few cases users took significantly longer on a single task than other users for the same dataset and tract. This outlier effect happened when users switched to a ”rigorous” refinement mode and lost track of time. Interestingly enough a ”rigorous” refinement mode did usually not result in a better subjective confidence rating.

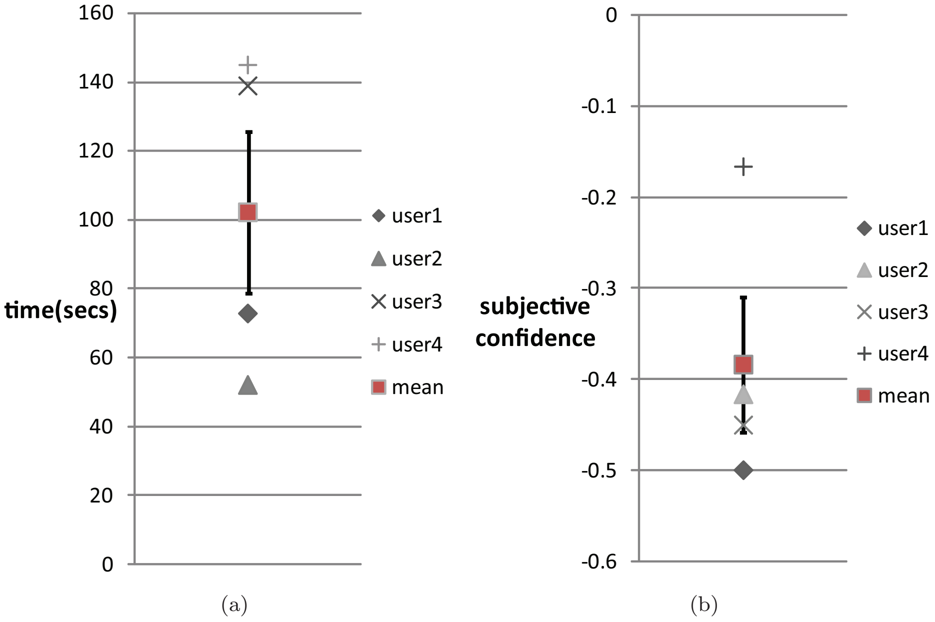


Figure 2.6: Per-user differences between (a) time and (b) confidence measurements on the two tools. Differences are obtained by subtracting 2D path tool performance values from 2D point tool performance values. Red squares show the mean performance difference between the tools. Errors bar around the red squares indicate the standard error of per-user differences.

2.4 Discussion

Each dimension in a visualization comes with extra cognitive and perceptual load. While there are clear advantages to three-dimensional visualizations in some contexts, previous work shows that humans are better at understanding two-dimensional representations [40, 143]. Beyond reducing cognitive and perceptual load, dimensionality reduction techniques have been popular in data mining because the “intrinsic dimensionality” of data is often much lower than the dimension of the space where data is immersed. In this context, it is not difficult to imagine fiber tracts as points on a low-dimensional manifold sitting in a high-dimensional space, particularly when we consider fiber tracts’ locally continuous and smooth variation in the brain. So, low-dimensional representations can go beyond being interaction gadgets and provide new “windows” into the intrinsic structure of data.

For instance, while clearly being inferior to hierarchically projected paths for some tasks, the 2D point embeddings can be used for a comparative analysis of multiple distance measures. In Figure 2.7 three different embeddings computed for the following distance measures are shown: the one proposed in [48], the one used in this dissertation, and Haussdorf distance. The embeddings were linked to the model they were derived from. By selecting points in either embedding, changes in relationships are highlighted in the others, while the corresponding tracts are displayed in a fourth window. Figure 2.7 illustrates how a fiber bundle is embedded depending on the particular distance measure: the first type of measure uses only tract curvature and will thus place the three tracts

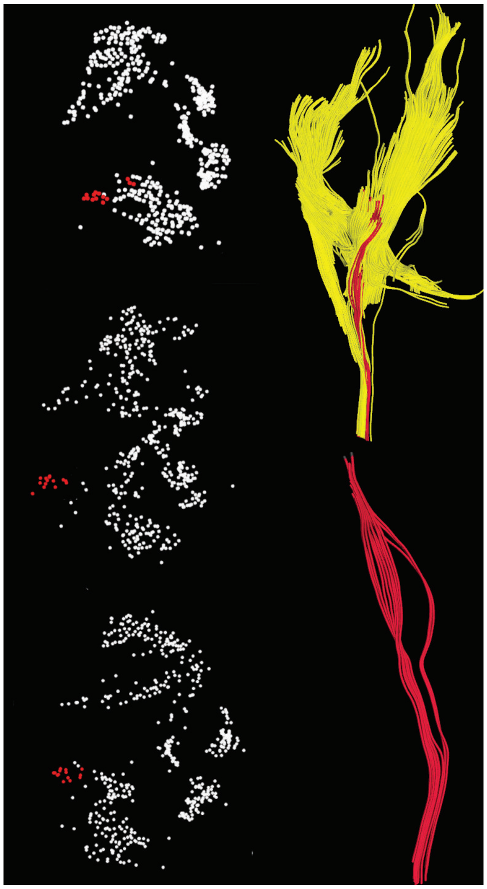


Figure 2.7: Comparing 2D embeddings for multiple tract distance measures. On the right, three types of distance measures were embedded: no end-point weight (top), weighted end-points (middle), Haussdorf (bottom). A few tract-points were selected. On the right, the corresponding 3D model is shown (top), together with the selected tracts in isolation from unselected ones (bottom).

that deviate from the bundle path further apart from the rest; the second measure adds weight to tract endpoints and as such ignores the bend in the three tracts and places all of them into the same cluster; finally, the Haussdorf distance considers only the minimum point-to-point distance and will thus place the tract-points together but also in the vicinity of other tracts that, while having close individual points, don't necessarily display any curvature similarity.

As shown, the explicit distance visualizations have the drawback of lacking anatomical interpretation. This might be alleviated by incorporating abstract representations of anatomical landmarks into the representations. For instance, points that correspond to tracts which come close to the brain ventricle can be highlighted in the planar representations. Alternatively, the ventricle could be approximated by a set of fictional tracts that could be projected along with the rest of the tracts but represented differently. While these methods could alleviate some of the mapping problems, the projected path representations are likely to maintain a significant advantage.

Finally, the 2D neural path representation uses simple heuristics but relies heavily on the quality of fiber tracking, distance measure and clustering. All of these factors can reduce the esthetic and functional success of the representation. While this can be seen as a limitation it also means that progress in any of these techniques will result in improved projected path representations.

2.5 Concluding Remarks

A new method for visualizing and navigating through tractography data, combining two-dimensional representations of fiber tracts with streamtube models was presented. Based on the geometrical similarity between tubes, planar abstractions were created from the tractographic data: a 2D point embedding, a dendrogram, and a novel visualization based on projecting major bundles onto three projection planes. These planar representations are linked to traditional 3D stream-tube visualizations of white matter tractograms through interaction and color. Two anecdotal and one quantitative evaluations show that such planar abstractions can improve data understanding and interaction in general, but are most effective if they preserve anatomical features such as in the case of the projected bundles technique.

Chapter 3

Exploring and Analyzing Protein Networks

Proteins within a cell interact with one another in order to regulate the cell's activity. The nature of these interactions is diverse. Among others, an external event can be transmitted to the inside of a cell through interactions of signaling molecules; a protein binds to another protein to alter its function; or a protein will carry another protein to a specific cell location.

A cascade of protein interactions peculiar to a specific cell, stimulation, or cellular outcome is called a signaling pathway. An in-depth understanding of these pathways will, among other outcomes, let researchers discover efficient drugs that can influence a cell's behavior without causing unwanted side-effects.

Experimental data is an important component that researchers use to understand how signaling pathways function. For instance, researchers can artificially stimulate a cell and measure how the proteins within it respond, possibly over a series of time-points. To efficiently interpret the results of such experiments, they need to be collated with existing knowledge that can explain some of the observations and provide valuable insights for hypothesis generation. One of the most common such data used in signaling pathway analysis are protein-protein interactions extracted from proteomic publications and stored in online databases.

Advances in proteomic experimental techniques and improved analytical methods have enabled researchers to produce vast quantities of experimental data. Combining it with the sheer complexity of protein interaction networks increases the information space even more. Thus, thinking about the data at its original low level has become impractical. New computational techniques are required that either extract relevant information automatically or let researchers process data faster by looking at condensed visual representations.

This necessity has been acknowledged by the research community and analysis frameworks that build on traditional graph drawing to visualize protein interaction networks have emerged. However, findings presented in this chapter, as well as results from more recent work, suggest that additional research is needed to ensure that the visualization methods employed are adequate for proteomic research.

This chapter presents a design study on several novel visual and interaction paradigms for the analysis of quantitative proteomic data, canonical signaling pathway models, and protein interaction networks along with the proteomic analysis requirements that motivated them. The methods are evaluated anecdotally with domain experts to determine their overall ability to accelerate the proteomic discovery process.

The methods described are general and discussed in terms of their benefits as components of established protein networks analysis applications such as Cytoscape. However, for concrete exemplification, they will occasionally be framed in the context of the testbed application used to develop and evaluate them.

Figure 3.1 illustrates the main visualization and interaction paradigms presented in the paper: harnessing the researcher's existing mental schema and intuition by integrating dynamic interaction data into static but familiar signaling pathway images provided by the user; enabling proteomic specific interaction level analysis of dense networks by integrating a novel Focus+Context technique; and driving exploration by comparative analysis of multiple experimental datasets.

First, we relate results to previous work in protein interaction network visualization and related techniques. We then introduce the methods by presenting an overview of the visualization workflow and then detailing each of its components. We then present our results as findings and evaluations of how the techniques improve the proteomic workflow. A discussion of design choices follows. A distillation of the findings concludes the chapter.

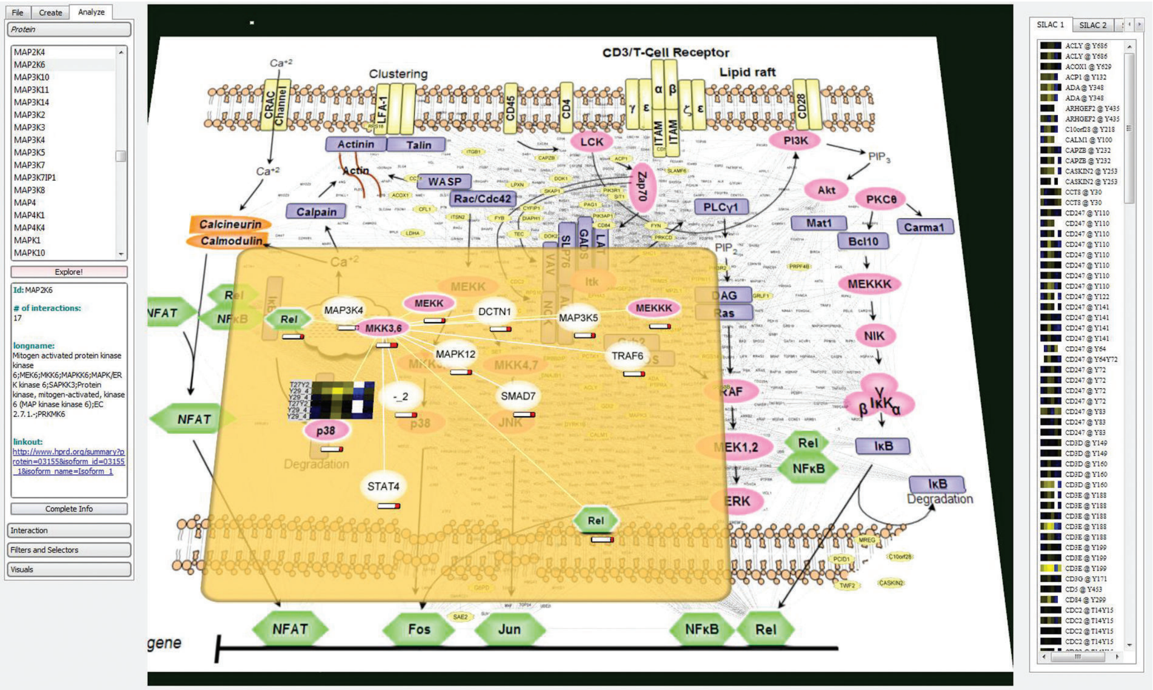
3.1 Related Work

Here we present a few related work topics pertaining to this research: visualizing protein interaction networks in particular and networks in general, and focus+contex visual exploration methods.

3.1.1 Visualizing Signaling Pathways and Protein Interaction Networks

The first representations of protein interaction networks had the form of static, schematic drawings of signaling pathways. Several papers such as [25, 114] discuss guidelines and approaches to drawing such representations. However, the static nature and manual assembly became serious limiting factors when protein-protein interaction databases were first created – researchers needed a way to generate visualizations on the fly based on database queries.

Many popular protein interaction databases – examples include the Human Protein Reference Database (HPRD) [131], Molecular INTeractions Database [181], STRING [171], and the Database of Interacting Proteins [176] – started to provide on their websites visual components that let users navigate the protein interaction space. Most of these visualizations represent protein-protein interactions via a node-link paradigm and produce visual layouts with spring models or other force-directed methods. Recently, more advanced standalone visualization systems have emerged; notable



Id: MAP2K6

# of interactions:
17

longname:
Mitogen activated protein kinase kinase 6;MEK6;MKK6;MAPKK6;MAPK/ERK kinase 6;SAPKK3;Protein kinase, mitogen-activated, kinase 6 (MAP kinase kinase 6);EC 2.7.1.-;PRKMK6

linkout:
http://www.hprd.org/summary?protein=03155&isoform_id=03155_1&isoform_name=Isoform_1

Figure 3.1: Analysis of a protein interaction network in the context of the T-cell pathway. Proteins and interactions dynamically extracted from the HPRD database (small fonts scattered between the protein icons in the pathway view
) are integrated directly into an imported image of a canonical signaling pathway. Heatmaps representing quantitative data from multiple experiments appear on the right and are used to drive
analysis. Focus+Context is implemented as a semitransparent plane hovering over the global image and allows researchers to navigate through complex networks in a one-level-at-a-time mode.

among them, Cytoscape [148] and VisANT [82] offer multiple representation methods, session-saving capabilities, and numerous features for pathway analysis. Moreover, users can add features and customize the software using plugin architectures.

Nevertheless, aspects of these visualization systems can still be improved. For instance, using generic techniques devised by the graph-drawing community sometimes yields visualizations that are far from intuitive to proteomic researchers, since their failure to incorporate protein cellular location and signaling pathway drawing conventions detracts from the visualization's familiarity. This problem is also recognized by [17] and [98].

Another topic not sufficiently investigated is the integration into protein interaction visualizations of quantitative data from large-scale proteomics experiments. Cytoscape uses a flexible plug-in architecture to address this and other functionality needs; other systems simply let one load textual annotations onto a protein network. The visual display, analysis, and comparison of results from multiple quantitative proteomic experiments are still an area of active research. The most recent work identifying and addressing the issues of both layout and experimental data is [17]. It extends Cytoscape with a new protein network layout algorithm that organizes proteins in cellular layers, based on an annotation file supplied by the user. Quantitative data can be loaded and viewed as color mappings on the proteins. Multiple experimental conditions are shown using small multiples (i.e., multiple iconic representations of the protein network for each experimental condition) and a parallel coordinate view. The work presented in this chapter differs by offering an alternate way of drawing the protein network, a different representation of the experimental data and the ability to load multiple experiments, each with several conditions, and in identifying and supporting the need for exploring biological networks at global and local levels simultaneously.

3.1.2 Visualizing and Exploring Networks

There are many techniques or systems for displaying general graphs such as [46, 67, 65, 166]. However they often fail to translate well to biological networks. Protein network layouts require a constraint-based approach in which general aesthetic graph-drawing criteria are met, while satisfying other biological or user-defined constraints. Dwyer and Marriott [56] is the state of the art in constraint-based graph layout but its complexity, while powerful in its adaptability, makes it hard to implement and control. Like [17], we chose to implement an algorithm that is easier to adapt to our specific problem. The layout algorithm itself is close in several aspects to the one described in [66] for drawing evolving graphs. They place new nodes at the barycenter of existing ones, with subsequent force-directed steps. we use a similar approach to place database-extracted proteins in relation to pathway proteins.

The idea of scaffolding graph drawings on another structure, as done in this work, is found in [120]. Here, domain knowledge is used to identify spanning trees within graphs, and the simpler tree layouts are used as scaffolds for the general graph structure. Similarly, [6] automatically computes spanning trees as graph scaffolds and demonstrate their methods in the context of biological networks.

3.1.3 Focus and Context

Revealing global aspects of data while also granting access to details is commonly known as Overview+Detail. A subcategory of Overview+Detail is formed by so-called Focus+Context techniques which show the global and detailed views simultaneously. They are often preferred over more traditional Overview+Detail, such as zooming and panning, which can leave the global picture out of view when zoomed in on details. Quantitative evidence that may explain this preference was published by Plumlee and Ware [134] — they show that the cognitive cost is higher when zooming and panning than when viewing local and global aspects of the data simulatanously on side-by-side displays.

Several Focus+Context techniques have been devised. For instance, [137] leverages trained human 3D perception by displaying trees in 3D and using the proximity of objects as a direct focusing mechanism. Another popular Focus+Context approach is to distort the representation space to give more screen real estate to focused regions as opposed to context regions. Other examples of such techniques are [142], [120] or [162].

The Focus+Context method presented in this chapter is closely related to [160], which interposes a separate viewing plane between the viewer and the actual scene. Although similar to a regular lens, this space can be used to display detailed information about the underlying scene.

3.2 Design Elements

This section introduces the design principles and implementation details employed by the visualization methods presented in this chapter. First, an overview of the proposed visualization workflow is given. Then, details about each of its components are provided.

Researchers analyze their data in the following workflow:

1. import a model of a canonical pathway representation either by loading a signaling pathway image and preprocessing it to help the system infer the structure (Figure 3.2, lower left) or by specifying the model explicitly by placing proteins and interactions on an empty canvas (Figure 3.2, lower right);

2. load one or more quantitative experimental datasets;

3. automatically extract proteins and interactions from protein interaction databases such as HPRD and build a network around the pathway model specified in step 1 and the quantitative data from step 2;

4. represent the network graphically using a novel canonical pathway-oriented layout (Figure 3.3);

5. explore and analyze the network guided by interesting features noted in the experimental data; investigate the network at interaction level using a Focus+Context technique; analyze how known information blends with the new experimental results using such features as clustering of quantitative proteomic data, filtering, highlighting, and information on demand (Figure 3.1);

6. derive insights or generate new hypotheses, design and run new experiments, and restart from step 2.

3.2.1 Pathway Model Specification

The solution described in this chapter requires the user to specify, using a simple interface, the canonical pathway representation of the signaling pathway under investigation. This can be done either by putting proteins and interactions on an empty canvas or by using a pathway image that is preprocessed to help the system extract the pathway structure; the preprocessing entails drawing single, continuous strokes over or around each pathway element – proteins, interactions and other entities. These strokes aid the software in identifying image features (Figure 3.3) as detailed below.

If the stroke endpoints are far apart compared to the stroke length, the image feature is probably an interaction and the endpoints are matched against protein positions to find which proteins are involved. The interaction strokes snap to image features in a manner similar to a lasso tool. This

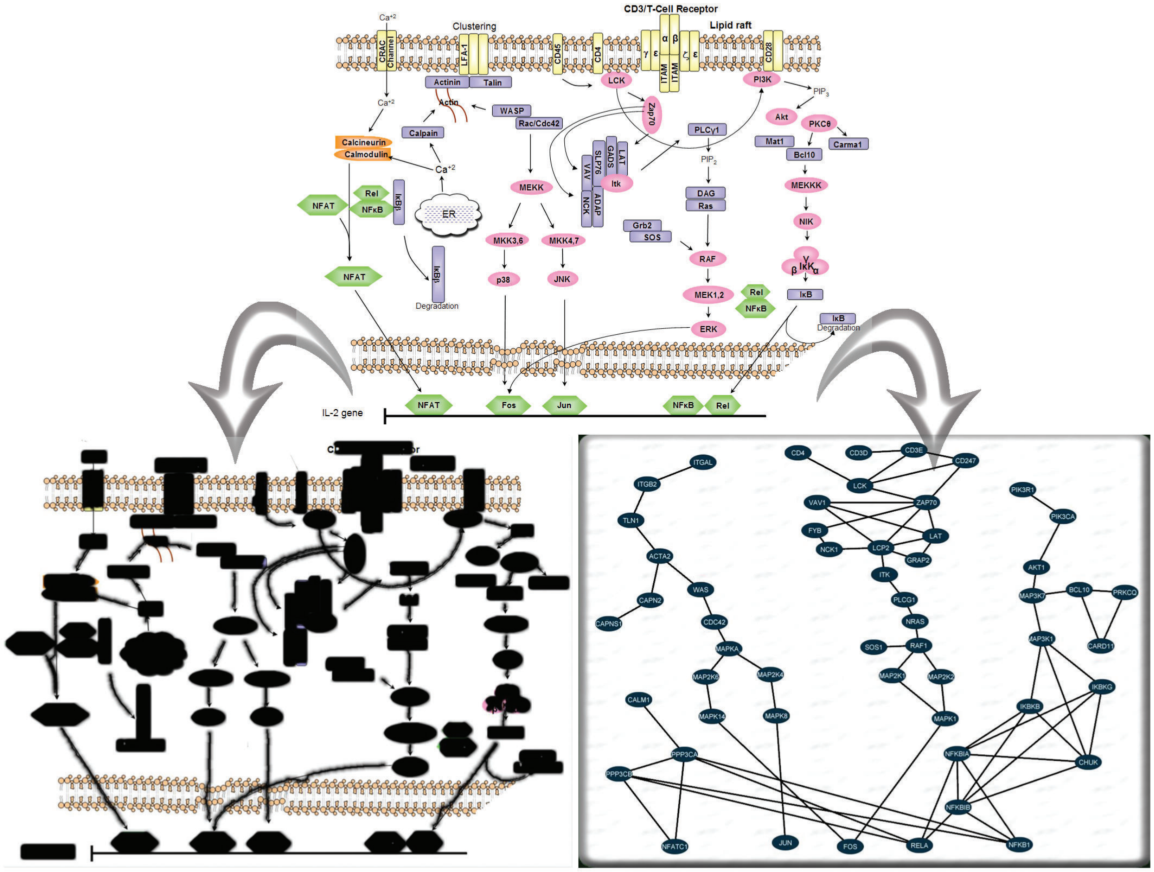


Figure 3.2: Structuring protein interactions around familiar canonical pathways provides intuitive visualizations. A canonical signaling pathway representation (top) can be imported into the system in two ways: on the lower left the pathway image itself is loaded into the system and preprocessed by circling proteins and drawing over interactions; the pathway features are then inferred from the user strokes and image features and shown here in black; or, on the lower right, protein and interaction icons are placed and dragged on an empty canvas to create a new pathway model. After positional assignment of each protein, the software aids in associating interaction database accession numbers to each of the newly defined canonical pathway proteins.

is done in order to obtain the correct image region that the interaction is covering, for reasons described in Section 3.7.

If the stroke endpoints are close relative to the total length of the stroke, the feature-detection algorithm decides to classify the feature as a protein. It then computes an average color for the area enclosed by the stroke and removes all points dissimilar to it. In most cases this leaves only the image shape selected. The protein position on the canvas can be inferred through this computation.

If a selection is unsatisfactory the user can cancel it and try again – depending on the previous selection, the algorithm will attempt to correct the image-processing parameters for the second try. For instance, if the area selected by the user is much larger than that returned by the algorithm, the color-similarity threshold is increased.

Once the graphical model is complete, either by pathway processing or by pathway drawing, the placed proteins need to be linked to protein identifiers in the protein interaction database. The user chooses the correct protein by searching the database for keywords using a dedicated dialog box. In our test cases this process took between 15 and 30 minutes for medium pathways such as those in the figures, but these times vary with image complexity and user training.

3.2.2 Interaction Data

The experimental prototype described here uses the HPRD protein interaction database. HPRD is a protein interaction and metadata source based on manual literature search. The database information is stored and loaded as flat files.

The network exploration paradigms defined here could be used with any protein interaction database. One of the main challenges in supporting a protein interaction database is providing access to useful metadata from other databases. This is due to the inherent difficulty of translating protein identifiers across independent protein databases.

3.2.3 Experimental Data

The quantitative proteomic data is loaded as XML or flat files upon pathway creation and can contain multiple quantitative data points as well as protein identifiers and other metadata. For graphical representation, the quantitative proteomic data are transformed into a colored heatmap representation (Figure 3.1) indicating fold changes of a given peptide across different experimental conditions (time course of receptor activation or comparison between wild type and mutant cells). The following color-coding is used: blue – decrease of proteomic quantity, yellow – increase of proteomic quantity, black – no change.

If multiple experimental files are loaded, as in a comparison between wild type and mutant cells, special types of heatmaps are computed for each pair of experiments to reflect changes between experiments: yellow then indicates a major change between the two experiments, while black corresponds to no change. A single protein can have multiple heatmaps, one for each assigned peptide. The heatmap icon appears in two places: displayed in the expanded network exploration upper plane, attached to proteins revealed in the experiment (Figure 3.1), and in a dedicated panel on the right (Figure 3.1) containing all peptides discovered in an experiment.

For multiple quantitative data sets, the heatmap experimental data panel on the right (Figure 3.1) is configured to contain tabs not only for each separate experimental data-set but also for changes observed between pairs of data-sets. For instance, in a phosphoproteomic receptor activation timecourse experiment involving wild type and cells lacking critical signaling proteins, the heatmap tab contains one tab dedicated to timecourse phosphopeptide heatmaps in the wild type cell, another tab for the mutated cell, and a third tab displaying the fold change of individual phosphopeptides observed between the two cell types through the receptor activation timecourse. This feature can be particularly useful in knockout-type experiments since the differences in behavior between a normal

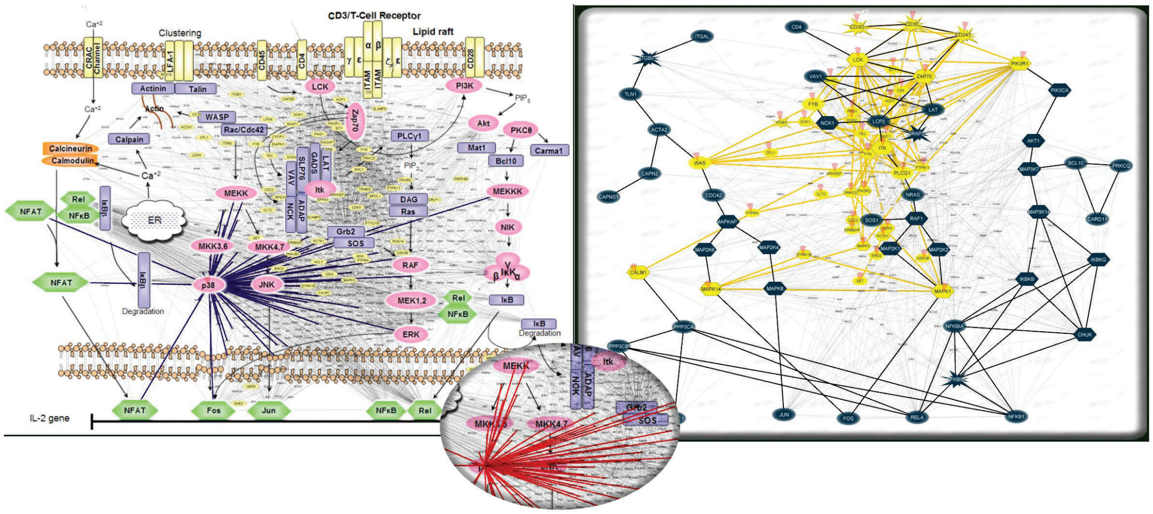


Figure 3.3: Proteins and interactions from HPRD (small fonts) that are connected to the canonical pathway model are: (left) integrated directly into the signaling pathway image with one protein selected and its interactions highlighted; (right) structured around a user-constructed model; different classes of proteins have different appearances: experimental proteins are colored yellow, kinases are drawn as hexagons and receptors as irregular stars; several experimental proteins are not known to be connected to the pathway and are therefore located in the lower right corner. HPRD proteins are placed in a structured manner between the pathway proteins based on their separation from the pathway proteins. (cutout) Disadvantage of simply drawing the network on top of the pathway image: HPRD interactions obscure elements of the canonical pathway; compare to the improved method (left) in which important pathway elements remain in the foreground.

and a mutated cell become evident immediately.

The experimental data panel is kept visible at all times so that researchers can use it to explore the new quantitative data systematically. The items in the experimental data panel can be used to start the exploration by linking directly to Focus+Context representation.

Using experimental data to guide exploration was also discussed in [17]. The work here differs both in the way the information is presented to the user and in the emphasis that is put on comparative analysis of multiple experiments. Such analysis can also be performed in the other system, but the small multiple approach is likely to overload the display if used with dense networks and large quantities of experimental data. Their parallel coordinates view was also not extended for both multiple time-points and multiple experiments.

3.2.4 Network Generation

From the user-provided pathway skeleton, the software constructs a protein-protein interaction network by loading proteins and interactions from the HPRD database. The network is grown iteratively in a breadth-first manner: first, proteins interacting directly with the canonical signaling pathway model are imported, and then in subsequent steps, proteins interacting with those added in the previous iteration are extracted from HPRD and included. Finally, interactions among all proteins are loaded.

The number of levels to grow the network and optional filters used to exclude proteins from the build process are specified by the user. However, growing the pathway from the user-specified proteins alone may leave experimental proteins outside the network. To ensure inclusion of all experimental proteins in the final visualization, the network is also grown from the experimental proteins themselves. This solution increases the chances of linking the experimental proteins to the pathway since two networks are grown simultaneously toward each other.

3.2.5 Computing Protein Positions

While the canonical pathway proteins have user-provided predefined positions, the system must compute where to put the proteins extracted from the interaction database. These proteins are placed depending on their distance, in terms of number of interactions, from each of the pathway proteins. If protein P is interacting directly with protein A and is three interactions away from protein B, it is placed on the line segment between A and B, closer to A. The distances are not necessarily directly proportional to the path lengths: they can be weighted so that direct connections are much shorter than longer interaction paths.

Essentially the nodes are placed at a path-length weighted barycenter of the pathway nodes. Barycenter positioning was also used in [66] to place new nodes in relation to already existing ones in the context of evolving graph drawings. This algorithm produces positions close to those computed by a traditional spring layout algorithm, since a node is dragged by the edge springs to a similar location.

This methodology leads to identical positions for some proteins, however, and a force-directed approach based on [67] is used to perturb the layout and remove overlaps; a simple linear grid approach is used to improve the performance of the layout algorithm by using vicinities to reduce the number of comparisons needed to compute forces on protein-nodes.

The sizes of nodes are taken into consideration when computing repulsive forces. The aspect ratio of nodes in relation to the force vectors can also be taken into account so that forces are applied anisotropically. This leads to slightly longer run times but minimizes overlap, especially in augmented pathway images where some nodes can be much larger in one direction.

As a special case, positions cannot be computed for proteins linked only to the experimental data and not to the known pathway. These are placed in the lower right side of the display, yielding a cluster of proteins that are not known to be connected to the pathway.

This algorithm is relatively fast, interactive, and achieves the desired results without the complexities of more powerful constraint-based techniques such as [56]. The layouts in Figure 3.3 took around 2 minutes to compute. We also experimented with simulated annealing methods. These, however, were much slower and did not improve the layouts significantly due to the high network density. Some parameters inherent to force-directed methods still require user adjustment.

3.2.6 Augmenting a Pathway Image with Dynamic Data

The case of specifying a pathway image and integrating dynamic information seamlessly into the already existing representation is more complicated than assembling a completely new visualization. Simply drawing the database extracted elements on top of the pathway image has several disadvantages, as shown in the cutout of Figure 3.3. In contrast, the method presented in this section creates the illusion that the proteins and interactions drawn dynamically are part of the pathway image (see Figure 3.3, left).

The following specialized operations are used to create the illusion that the HPRD proteins and interactions are part of the pathway image. The shapes and locations of proteins and interactions in the image are computed in the image preprocessing step. They are then used in the layout stage to minimize overlap (dynamically loaded proteins tend to move to "empty" image areas). Finally, they are copied from the image and redrawn as masks on top of the final network. This technique ensures that the pathway model stays on top of the dynamic network and gives the illusion that the canonical pathway representation and the dynamic network coexist and interact (see Figure 3.3, left).

3.2.7 Exploring the Network

In our design the interaction network can be explored at two levels simultaneously: at a global level, where the signaling pathway and other high-level structures are evident, and at a local level, where only one protein and its neighbors appear in detail as the researcher jumps from protein to protein in the network. The two types of visualization coexist as two parallel planes, the local one gliding above the global one (Figure 3.1). With these complementary views of the pathway space, the user explores the network in the detailed space that is rich in focused protein information while maintaining an overview of the explored area and orienting the expanded exploration to his or her location within the global view.

Exploration is done in a plane that hovers above the global view and shows in detail only one protein and its interactors. Initial access to the exploration plane can be obtained by double-clicking proteins in the global-view, in the experimental lists, or in a list of all proteins present in the visualization. While in exploration view, clicking one of the interactors shifts the center of the view to this selected protein, a change performed through smooth animation to maintain context understanding. Standard zooming and panning using mouse controls are also available, but testing has found them less favored by users. Proteins in the exploration plane are arranged so as to mimic their placement in the global layer while satisfying aesthetic criteria such as minimum distances between proteins or interaction overlap (Figure 3.4, left). The effect is achieved by applying a simulated annealing [46] algorithm that attempts to maximize layout similarities while ensuring a pleasing drawing. The area allocated to the exploration view is computed dynamically on the basis of the number of proteins to be displayed. A view that places the main protein in the center and its interactors circularly around it is also provided.

Clicking a protein in the exploration view highlights it and its neighbors in the lower plane, making it easier for the user to establish a correspondence between the two.

3.2.8 Visualization Prototype

A compact set of features were added to the testbed system, allowing our researchers to operate on the network data and pose visual queries. For instance, selectors and the ability to adjust appearance allow the researcher to highlight interesting aspects of the visualization. In the right panel of Figure 3.3, a user has selected various groups or classes of proteins and attached to them special visual attributes such as shape and color a technique often used in stylized signaling pathway representations. The method described in [121] is used to highlight interactions of one or more selected proteins. Easily extensible filters allow a researcher to remove proteins deemed uninteresting. One potentially useful filter with significant effects keeps only proteins that connect a set of user selected proteins.

3.2.9 Implementation Details

The prototype application was written in C++. The G3D 6.7 graphics library was used for 3D graphics and rendering and the Qt 4.3 library for user interface elements. The HPRD database can be downloaded as flat files together with the application.

3.3 Evaluation and Findings

The results of this work are findings about ways to improve analysis of protein interaction networks and quantitative proteomic data, and novel visualization and exploration paradigms motivated by these findings.

The research presented in this chapter was driven and validated by insights obtained during our collaborative development process and by an anecdotal evaluation with domain experts. Results indicate that applying these concepts in the context of systems for visualizing protein-protein interaction networks may accelerate the discovery of new connections among quantitative proteomic data, interacting proteins, and canonical signaling pathways. While a controlled study may still be needed to verify and quantify the benefits of individual aspects of our methods, the anecdotal evaluation with domain experts is a preferable approach in an iterative design setting, with no predefined requirements since it can provide fast, easy access to usability information on high-level analysis tasks.

Evaluation was performed on the analysis of phosphoproteomic experiments with the help of four proteomic researchers interested in research of the T-cell and Mast cell. These experts artificially stimulate cells and measure the amount of phosphorylation that occurs on proteins as a result. Phosphorylation is an important cellular process by which a phosphate is added to a protein or other molecule. A protein can be phosphorylated in multiple places, called phosphorylation sites.

In a single experiment setting, phosphorylation measurements over multiple time-points can provide causality hints. More importantly, however, researchers can run separate experiments before and after inhibiting an investigated protein. By comparing changes in measured phosphorylation values they can hypothesize about the role of the investigated protein in the cellular pathway.

Finding 1: Visually combining experimental data and known protein interactions enhances analysis

Results presented in this section augment previous results from [17] and [148] with similar findings in a different analysis setting. Specifically, results suggest that coupling new experimental data with protein interaction data extracted from public databases within a unified visual analysis can shorten the analysis process of a new experimental dataset from weeks to days. In addition to the straightforward time gain, shorter time intervals between individual data observations lets researchers integrate them more efficiently into a cohesive hypothesis.

By using the prototype, one of the collaborators who took part in the evaluation quickly discovered a meaningful biological fact that eluded her in previous analyses of a T-cell related phosphoproteomic dataset. She started by browsing through the list of experimentally measured proteins, displayed as seen in Figure 3.1 on the right hand side of our prototype. She then decided to take a closer look into the protein Slp76, because of the variation reflected by its heatmap. Double-clicking on the list item opened a detailed exploration view, as shown in Figure 3.1. The visualization revealed that this protein was known to interact with the protein VAV. Metadata available within the software then revealed that the particular measurement could indeed be related to that specific interaction. In addition, a novel phosphorylation site was detected on SHP1. An interaction with SLP76 and meta-data about this interaction were easily accessible in the software and led to the hypothesis that SHP1 negatively regulates SLP76.

These insights may have been eventually produced using the collaborators' previous strategy of manually querying each experimentally measured protein and gathering information about them. The integration of experimental data and the protein interaction network reduced the time needed to make this discovery.

Finding 2: Canonical pathway-driven layout is intuitive for proteomic researchers

Structuring dynamically extracted protein interactions around a familiar canonical pathway (see Figure 3.1) provides an intuitive visualization that helps proteomic researchers orient themselves and learn the interaction network quickly. A proteomic experiment revealing hundreds to thousands of protein modification sites overwhelms users with the many unfamiliar proteins. Becoming familiar with the proteins in such an experiment is greatly facilitated by placing those proteins within signaling pathway-structured protein interaction networks.

This pathway-structured method was motivated by negative feedback on an initial prototype that used a standard force-directed network layout. This feedback suggested that generic network-drawing algorithms fail to place proteins in positions that are meaningful either from a biological or a pathway-conventions standpoint (receptors can end up near the nucleus). Moreover, proteomic researchers were overwhelmed by the unstructured node-link diagrams such methods produce and tried to map the new visualization to the signaling pathway they were using before. This was also found by [17] and [98] to be an important issue in systems that employ traditional graph drawing algorithms to display protein interaction networks. The work presented here differs from theirs by introducing a novel visualization paradigm to address this problem.

In a broader visualization context, integrating dynamic connectivity information into static diagrams is a potentially useful concept because it facilitates the integration of new information into existing thinking schemas. We demonstrate its perceived usefulness in a proteomic context. More targeted research is needed to establish whether the perceived benefits translate to actual task improvement, and to identify other areas of application.

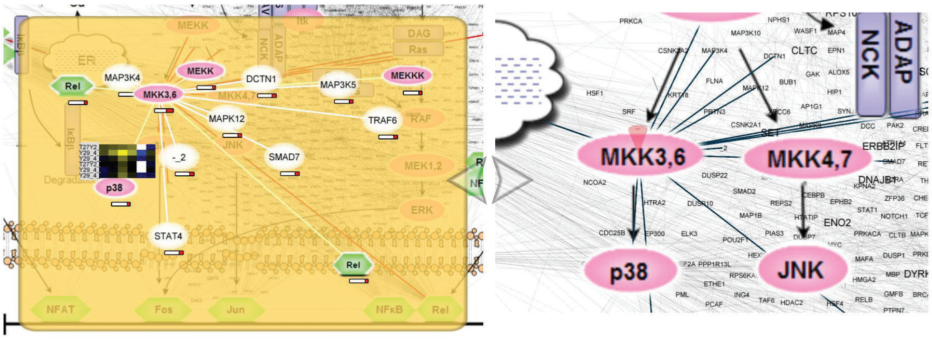


Figure 3.4: Exploration plane versus zoom-and-pan. (left) The network is explored in a separate plane showing only one protein and its interactors. Selecting an interactor changes the view of that particular protein via a smooth animation. This interaction network crawling method allows systematic discovery of connections among proteomic data and existing protein knowledge. Transparency keeps the global view visible and the same protein is highlighted within both planes. The protein layout in the exploration plane mimics the layout in the global plane, but is slightly distorted to achieve a more attractive representation. Changes in peptide abundance are represented as linear heatmaps. (right) Zooming and panning, while also available to explore the network, have several drawbacks: the view is cluttered, some interactors reach outside the viewing area, there is no space for additional details, and the global perspective is lost.

Evaluation

Overall, the experts preferred this layout method over two network visualizations they tried before: Cytoscape [148] and an earlier interaction network prototype. At the time of their use, both systems used traditional graph drawing algorithms and were criticized for their lack of structure.

Conclusions drawn from specific user comments were: the familiar pathway model that seeds the exploration is visually appealing and reduces the initial ball-of-strings shock associated with most network visualizations; it helps users orient by providing a familiar context; it gives protein placements more meaning and ensures that well known proteins are placed in familiar locations.

A problem identified in early testing was that growing the pathway from the user-specified proteins alone omitted many experimentally observed proteins from the network due to the lack of connections between these proteins and the known pathway within the protein interaction database. This problem was addressed by growing the network not only from the pathway proteins but also from the proteins indicated by the experiment, thus ensuring their inclusion within the network. However, some experimental proteins will still not be connected. These proteins are placed in the lower right corner of the representation, essentially forming an island of proteins revealed in the experiment but not known to be connected to the user-provided signaling pathway skeleton.

This approach has its benefits, as one test case revealed. After loading a large phosphoproteomic dataset onto the well established insulin pathway, a user immediately observed that many of the experimental phospho-proteins were connected to the signaling pathway, while the "island" of unconnected proteins was fairly small. This increased the user's confidence in both the experimental results and the visualization.

Finding 3: Global and local exploration modes (multilayer, multiscale views)

The evaluation shows that researchers prefer to explore an interaction network by using a local view of each protein, looking only at its direct interactors at a time (Figure 3.4, left). This initial hypothesis guided the design choices and was validated during evaluation.

During the testing stages much time was spent in local view instead of global view. This finding suggests that protein network analysis benefits from views that isolate one protein and its interaction from the rest of the network. Current interaction network visualization frameworks lack Focus+Context capabilities, and little research exists to address this issue.

Evaluation

The evaluation revealed that the exploration plane was indeed the most popular mode of protein-network exploration. A second demonstration and usage session with a separate proteomic group led to the same conclusion. The global view was used to apply filters, browse through the data and jump-start exploration. It also created an important first impression of the visualization as a whole and kept the users engaged. For reasoning about connectivity however, researchers rarely looked directly at interactions in the global view, even though zooming and panning were available. The navigation plane was used instead.

Observations of proteomic workflows during development and evaluations suggest that current proteomic analysis happens mostly at interaction level. This explains why our Focus+Context method was preferred over traditional global exploration: a single protein and its interactors can be viewed without clutter from any other network elements; all interactors are visible at once without panning; the space can be distorted to make room for additional glyphs and information associated with the proteins; and both views – global and local – are visible at the same time, with an emphasis on the local view.

Given the synergy between local and global viewing, with a stronger emphasis on local exploration for accurate, analysis tasks, the method presented in this chapter appears to be adequate. The local view is in the focus, while the entire global view is maintained in the background as a mental anchor. The user can switch between views immediately using an intuitive operation that requires minimal mental transformations.

Probably the main contribution of this result is that techniques for the exploration of networks concurrently at varying degrees of detail are suited for proteomic analysis tasks and should be included in specialized systems. While the presented novel technique works well in this domain, other Overview+Detail paradigms, such as the ones described in our related work section, may also produce good results.

We note that this research used unfiltered interactions directly from proteomic databases. This resulted in dense networks. Curating interactions that are placed in the pathway could allow all information to be visible at the same time as seen in some networks presented in [17]. In this case zooming and panning may be sufficient for interacting with the network.

Finding 4: Comparative displays of multiple experiments help identify important pathway players

Test cases showed that the ability to load and compare multiple experimental results, for example from cells containing deleted or mutated proteins, helped researchers link cell behavior to experimental results. Also, researchers found it useful to have the experimental data permanently visible to drive the exploration.

Evaluation

The first prototype that was developed did not present the user with an explicit list of experimental proteins. Instead they were marked on the network. Users argued that they prefer to be able to go through their experimentally derived proteins systematically, preferably in a list. This lead to the addition of an experimental proteins list in one of the system's submenus. Further testing showed that users referred to that list throughout their analysis. The conclusion was that having it permanently displayed and linked to all the views would speed up their analysis process.

In the final evaluation, the typical analysis workflow consisted of systematically going through the experimental protein list, selecting ones with interesting patterns as suggested by their heatmaps, and opening them in local exploration.

The following test scenario showed the usefulness of this approach: in an experiment, a known T-cell signaling protein ZAP70 was removed from the cell and quantitative phosphoproteomic perturbations were recorded before and after the removal.

The user started his analysis by examining the heat-maps indicating the fold changes between the two experiments. The heatmap profile signaled an interesting change on the Lck protein, an upstream component of the pathway: the phosphoryaltion of Lck was greatly delayed when the downstream protein ZAP70 was removed. By bringing up Lck in the exploration plane, a direct interaction was discovered that connected Lck to Zap70 and explained the change.

3.4 Discussion

3.4.1 General Considerations

Maintaining a tight collaboration between researchers from computer science and proteomics lead to a better understanding of the requirements and specifications of proteomic visualizations. The canonical pathway-driven network layout and experimental data-guided network exploration are tangible results such a collaboration.

Good proteomic visualizations should support and automate part of proteomic researchers' data analysis workflows. But identifying these workflows is nontrivial and often varies among individual labs and researchers. The novelty of experimental data and constantly evolving proteomic methodologies make it hard for the researchers themselves to describe their workflows clearly. However, the process of workflow discovery, while laborious for both proteomic and computer science researchers, is beneficial for both parties since it identifies where computers can help most.

3.4.2 Layout

One drawback of the canonical pathway-guided layout is the overhead associated with specifying the canonical signaling pathway within the software. The most laborious step is not so much inputting the structure but searching for correct protein identifiers in the interaction database; this can be time-consuming due to naming ambiguities, multiple matches, missing proteins, and inconsistencies across protein databases. Initial testing revealed that identifying correctly canonical signaling pathway proteins within the protein database is aided by additional cues and metadata such as number of interactions or interacting partners.

The average time required by users to input the pathway skeleton and attach database identifiers was around 20 minutes for medium-sized pathways like those shown in the figures here. This overhead is acceptable if one considers that researchers commonly spend months or years studying a few pathways. Moreover, a canonical signaling pathway skeleton, once constructed, can be used to build multiple networks for different proteomic experiments and parameters.

Proteins imported from databases are by default displayed smaller than pathway proteins and are sometimes not legible without zooming. This mode of display was motivated by the desire to keep the pathway structure in the foreground and by the need to save canvas space and minimize overlaps in dense protein networks. However, the default size settings are adjustable and users can customize them for individual classes of proteins.

Another issue related to protein glyphs is that proteomic researchers often place several icons corresponding to the same protein in various places on the canvas, usually depending on the specific function and context. Many-to-one correspondences between graphical icons and data entities are uncommon in network representations. The testbed application allows this type of representation by automatically adding numbered suffixes to identical proteins to differentiate them. However, extensive use of this feature tends to clutter the representation with redundant information since interactions are replicated for each copy of the same protein.

Augmenting a pathway image with dynamic data does not always work. While the application is designed to accept any type of image, low image quality or high complexity can render the system unable to extract the pathway structure. The feature detection algorithm is flexible and can automatically adjust its parameters based on user feedback. However, the image-processing techniques were not the focus of this research and are not state of the art.

3.4.3 Focus and Context Exploration

The local exploration plane received positive feedback and was used extensively during evaluation. Its simplicity is both an advantage and a limitation. Users can easily understand what the display is showing and how to crawl around the network, while the visualization avoids clutter and in most cases does not require zooming or panning. Showing a single network level, however, can make it difficult to determine the optimal direction for future exploration. Unfortunately, real-life uncurated protein networks have high graph degrees that limit the number of levels we can show without clutter. Possible solutions to this problem are: hyperbolic views, automatically adjusting the number of levels that can be displayed without clutter, or attaching glyphs to nodes that provide cues about interesting exploration directions.

The decision to place the exploration plane on top of the global view rather than using a separate window was primarily motivated by the desire to save screen real estate. This choice has the disadvantage of occlusion, but we believe this is outweighed by the ability to use the entire display area for exploration while preserving a view of the global layout in the background. This situation arises frequently in protein interaction networks since many proteins are highly connected and need large display areas. The area assigned to the exploration plane is computed dynamically depending on the number of proteins to be displayed, thus minimizing occlusion as much as possible. The transparency of the exploration plane is also adjustable. The current proliferation in screen real estate, even in common analysis settings, opens the way to placing the two views next to each other. This approach would remove the occlusion problem but the need for frequent changes in focus across views and to spatially relate elements across the two views might lead to an additional cognitive cost.

Observations of the users confirmed this design choice: the global view was used mainly as a visual reference, especially for large networks, and as support for posing visual queries using selectors and filters. These tasks are not significantly hindered by occlusion. Proteins and interactions were rarely looked at closely in the global view, a task that occlusion would affect more.

It is also possible that using some filtering criteria on protein interactions will lead to sparser, more relevant networks like those featured in [17]. These could then be fully legible and explorable at a global view, potentially minimizing the need for a separate exploration view. However, the domain experts who took part in the evaluation have not identified in the biological databases they currently use any such criteria that can be automatically applied.

The placement of interacting proteins within the upper expanded view plane is designed to mimic the placement within the global lower plane while preserving aesthetic criteria such as node overlap. In addition to highlighting in the global view the interacting proteins that are being explored, this allows the user to better relate the exploration views to the global view.

The view during exploration can be either tilted or parallel to the view plane. An in-depth analysis of the benefits of each type of projection was out of the scope of this work. However, several users a strong preference for the tilted view, but this preference can be attributed to the superior visual appeal for a 3D representation rather than a perceptual benefit. Negative comments about distortions caused by the perspective projection seem to support this hypothesis.

3.5 Concluding Remarks

This chapter introduced several novel visualization methods and paradigms for the analysis and quantitative comparison of multiple proteomic data sets in the context of published protein-protein interaction networks and known signaling pathways. The effectiveness of the methods was evaluated in terms of data insights, hypothesis generation, and improvements in analysis time. Specifically, we showed that tightly coupling known protein interaction information with new experimental data, scaffolding protein interaction information around familiar signaling pathway models, and exploration at varying degrees of detail will increase adoption rates among proteomic researchers and accelerate knowledge extraction from massive quantitative proteomic datasets.

Chapter 4

A Map Inspired Framework for Accessible Data Visualization and Analysis

Visualization of biological data can be thought of as lying on a continuum. At one end, database-driven websites provide sparse representations of small, manageable bits of data. At the other end, complex stand-alone visualization systems offer many different visualization options and analysis features. Both approaches have merit and are widely used, but both have task-specific limitations. In terms of usability, the former have low visual expressivity and usually do not incorporate large data sets or complex computations, while the latter have significant overhead associated with setting up and learning to control the environments. In our experience, most biology researchers use one or two established analysis environments but are generally unable to invest time in learning new, experimental visualization tools. From a dissemination standpoint, scientists producing data lack the expertise required to set up and maintain a database-driven website. Finally, turning a prototype into a usable system can be a daunting task for visualization researchers due to the high costs and low benefits of GUI refactoring, automatic parameter tuning and creation of user manuals.

In this context the chapter introduces scientific data maps, pre-rendered visualizations of most data tied to a subdomain or scientific problem. They are handled over the web, possibly through the Google Maps API, and have a simple and intuitive set of interactions that can be learned with minimal overhead. An evaluation with domain experts shows that this approach is a viable solution for specific users and tasks and provides advantages from both ends of the visualization continuum while limiting many of the drawbacks mentioned above.

The key differences between traditional approaches and scientific data maps are as follows. Instead of the data-query-specification/recompute paradigm, maps contain all of a user's data or the views derived from them. Data query is thus done through zooming and panning during visualization. Traditionally, it is the end user's job to construct a visualization (query specification and

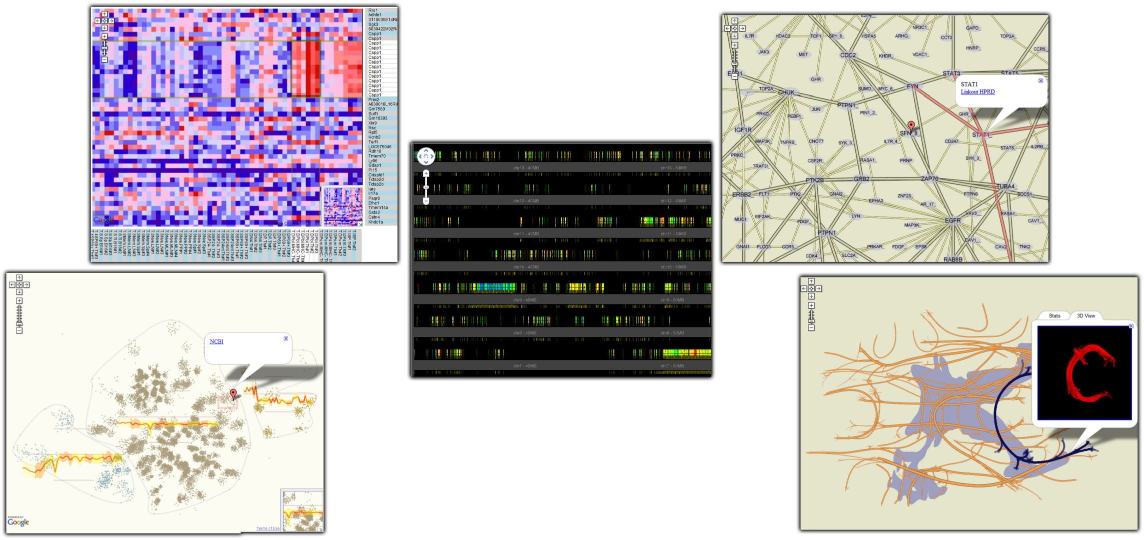


Figure 4.1: Five examples of digital map visualization (from left to right): gene co-expression and heatmap representations, a genome-viewer, a protein interaction network and a brain tractography projection.

parameter definition), while maps are built by visualization experts or bioinformatics staff in larger labs. Finally, the goal of visualization systems is to give users complex functionality that answers a large array of questions. Maps, on the other hand, aim primarily to provide fast, intuitive access to visual data; their functionality is therefore balanced with a sparse set of interactions, close to what is available in regular Google Maps.

Advantages of data maps occur on both the user and the visualization researcher sides. For users, including scientists browsing and analyzing data as well as those producing data, visualizations become easy to access, learning time is significantly reduced, users worry only about the data, and disseminating visual results is simplified. On the side of the visualization researcher, fast prototyping can be used for the rendering application with little effort invested in interfaces; there is no concern for computation and rendering time; visualizations are easy to distribute to both test and end users; and powerful synergies with web-based libraries, such as Protovis, can be produced by creating focus+context explorations in which the map provides context.

The motivation behind the work is to let labs publish data and results in visual form along with raw textual data so that users can access readily analyzable perspectives on the data without additional overhead. Specifically, the work is driven by the Immgen project [4], a collaborative effort aimed at generating a complete microarray dissection of gene expression in the immunological system of the mouse. The data-map concept enables the dissemination of the project's microarray data as precomputed visualizations that can be accessed on the project website. However, we show how the map concept is general enough to be also applied to the two application areas presented in Chapters 2 and 3.

The chapter is structured around five specific visualizations, which are implemented as maps: a genome viewer, a 2D embedding and heatmap of gene expression data, a protein-interaction network, and a white-matter tractography map. All have been evaluated informally with domain experts and two have been deployed and are in use on the Immgen website.

4.1 Related Work

Here we discuss existing approaches that are relevant to our work. The section starts with a discussion on web based visualization, it continues with an exposition on the Google Maps API and a presentation of visualization systems for biological data. We end with techniques relating to each of the five specific examples.

4.1.1 Web Based Visualization

Data visualization has been available on the web for many years but has usually displayed a limited amount of data using basic graphics and interaction. More recently visualization research started to target this environment and advanced applications have emerged. ManyEyes [169] paved the way for everyday data visualization, while other studies [168, 43] prove the need for accessible web visualization.

While web-development toolkits such as Protovis [24] greatly aid web development, large scale web visualization is hampered by inherent browser capabilities [96]. Alternatively, stand-alone systems have been made available as applets or to be run as client applications directly from websites [107, 148]. However, users still have to control the parameters involved in producing visualizations, specify their data queries and learn system features. This often constitutes an undesired overhead.

Yet another approach, more similar to our work from an implementation standpoint, is to use Ajax (asynchronous JavaScript and XML) technology to do the rendering on the server side and serve images asynchronously to the client browser. A specific call for Ajax-based applications in bioinformatics is made in [12], while [21] and [75] exemplify this approach. The essential difference between this work and traditional offline visualization systems is that control and display happens in a separate place from rendering and computation. The methods presented in this chapter differ by attempting to limit regular users' effort in creating visualizations and assigning this task to experienced personnel, by introducing large visualizations that contain most of the data associated with a problem, and by using the Google Maps API, a readily available Ajax implementation of pre-rendered images.

Closest to this methodology are X:MAP [179] and Genome Projector [11] which present implementations of genome browsers using the Google Maps API. This idea is expanded here to a broader visualization context, visualization solutions are introduced for four specific examples and an evaluation of both the preference of Google Maps powered visualizations in general and of the four specific visualization examples is presented.

4.1.2 Google Maps

We use the Google Maps API, an Ajax framework used to render large maps, to display our visualizations. It receives as input image data in the form of a set of small images, called tiles, that when assembled together form the different zoom levels of the map. Each zoom level consists of a rectangular grid of tiles of size $2^{zoom}X2^{zoom}$. The API decodes the zoom level and coordinates of the currently viewed map region to retrieve and display the visible tiles. The developer can load a custom set of tiles in the API by implementing a callback function that translates numerical tile coordinates and zoom level into unique paths to the custom tiles.

The API provides basic functionally such as zooming and panning and allows programmatic extension or customization with markers and polyline overlays, information pop-ups and event management. The API can be easily integrated into any Javascript-powered web-page.

Our visualization development environment was extended with the option to render any of our visualizations into a set of image tiles instead of the screen, for a specified number of zoom levels. In addition, each visualization must export auxiliary data to be used for interactive purposes (e.g. coordinates of genes for gene selection).

4.1.3 Biomedical Visualization

Many advanced systems for biological data analysis have been developed over the past decade. Examples targeting microarray expression data include free software packages such as Clusterview [60], TimeSearcher [80], and Hierarchical Clustering Explorer (HCE) [147] or commercial systems such as Spotfire [3] and GeneSpring [2]. GenePattern [107] is a broad effort aiming to facilitate the integration of heterogeneous modules and data into a unitary, web-managed framework for microarray data analysis.

Similarly, several tools exist for pathway and network analysis: Cytoscape [148], VisANT [82], Ingenuity [5] and Patika [47], all of which provide features aimed at complex analysis of microarray data or pathways.

The goal of the work presented in this chapter is to offer no-overhead visualizations that will be used primarily for casual data exploration by users unable to spend time learning advanced systems. In that regard, this work comes closer to applications providing primarily look-up functionality such as tools published on the NCBI website or the genome browser at USCS [102]. In contrast to these efforts, the data maps aim to provide visualizations that include more computation and visual cues and less complicated query specifications.

4.1.4 Multidimensional Scaling

A more complete discussion on multidimensional scaling is presented in Chapter 2. Similarly, the work here uses the same algorithm with linear iteration time proposed by Chalmers [34]. However, here, this algorithm is combined with elements from HiPP [130], an algorithm using a hierarchical clustering to drive a 2D embedding.

4.1.5 Genome Browsers

Genome viewers are used to explore genome structure from the chromosome down to the sequence level. They can be used to browse the structure of genes, to investigate whether function is linked to genomic location and to understand genomic conservation or alteration. Many implementations of genome browsers exist, ranging from basic web-based ones [156, 102, 153] to more advanced browsers integrated into complex analysis systems [113, 152, 45]. For visual display most genome viewers string chromosomes linearly, with some exceptions such as Circos [1] and Mizbee [113] that display them radially.

The approach presented here comes closest to X:Map [179] and Genome Projector [11], which use the GoogleMaps API to display precomputed images. The genome viewer presented here differs in relying on a visual mapping of expression data onto the genome and full genome views to drive exploration, and in presenting results obtained from evaluation.

4.1.6 Graphs and Protein Interaction Networks

A comprehensive discussion on networks was made in Chapter 3. However, the work presented here relies on two concepts that were not previously discussed. The following two paragraphs introduce these concepts.

Most existing protein-interaction visualizations handle only small networks, making the selection of relevant sub-networks essential. However, clear guidelines for this task do not exist, and simply selecting nodes within some separation leads to exponential growth because of the typically high degree of such biological networks. A solution to this problem was proposed in van Ham's work [167], which uses the degrees of interest (DOI) concept introduced by Furnas [69] to select meaningful graph regions. Similar computations are used in one of the map implementations presented in this chapter to achieve a zoom-based filter.

To reduce clutter in the network, Eades and de Mendonca's vertex-splitting operation [58] is used. This proposes that nodes which exhibit high tension due to layouting forces acting on them should be split into multiple copies. There is little work on using vertex splitting for drawing graphs. Henery et al [78] use vertex splitting in visualizing social networks; however, their method is not applicable to proteomic networks, which are in most cases unsuited for clustering.

4.2 Design Elements

This section starts with five specific visualizations that examplify the scientific data map approach. It ends with a distillation of design guidelines and methods for creating visualizations such as the ones presented in this Chapter.

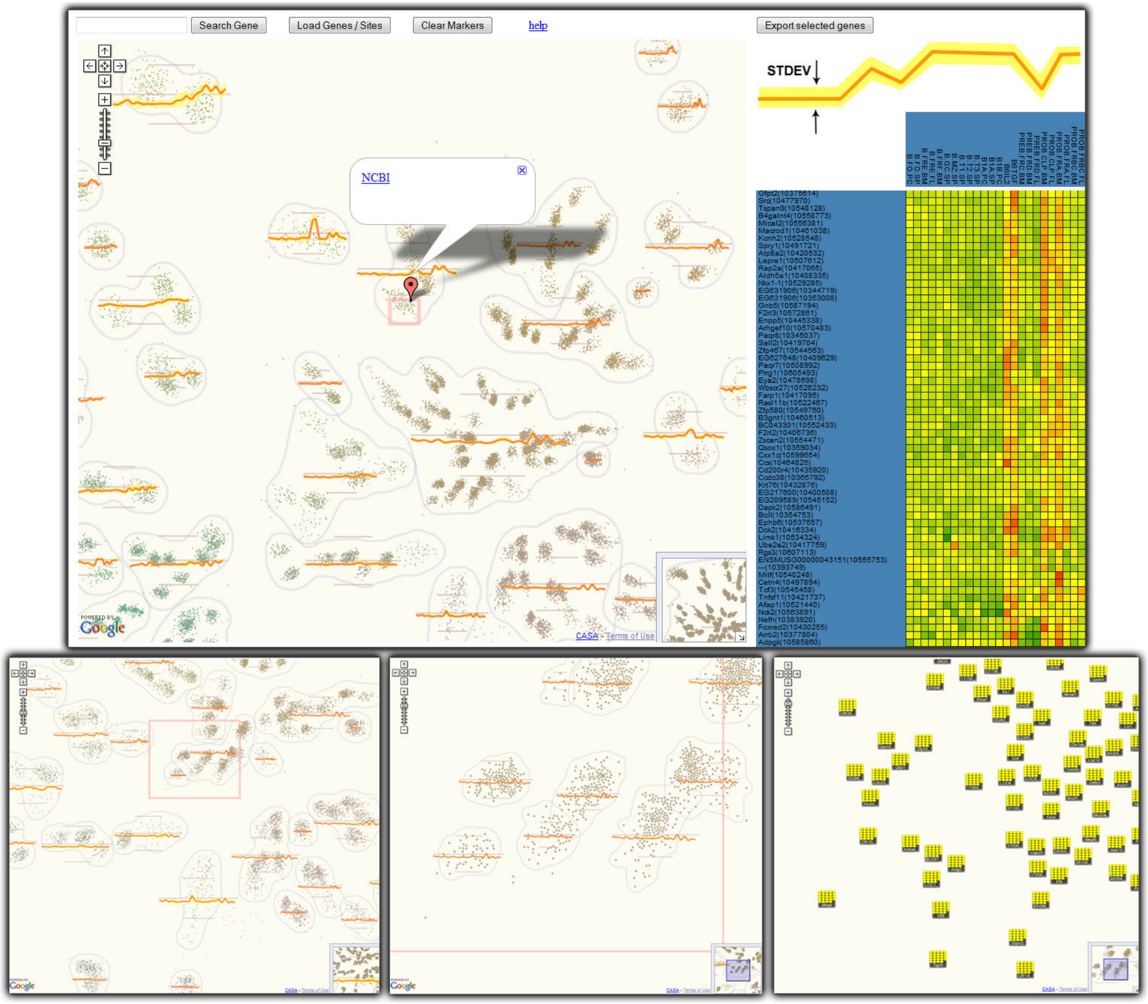


Figure 4.2: Co-expression map of 23k genes over 24 cell types of the B-cell family exemplifies map concept. The top view illustrates how maps are combined with client-side graphics: the map is at the center of the display while selecting genes by drawing an enclosing rectangle generates a heatmap on the right. Maps have multiple levels of zooming (bottom 2 rows), each with a potentially different representation. For example, genes are drawn as heatmap glyphs at the high zoom (lower right), and as dots at low zoom. Expression profiles of collocated genes are aggregated and displayed as yellow glyphs over the map. As zoom increases, expression profiles are computed for increasingly smaller regions. Interactions are not limited to zooming and panning; pop-up boxes link out to extra data sources, and selections of genes bring up a heat map (top panel).

4.2.1 Example 1: Gene Co-Expression Map

Description and usage scenario: Given genes with expression measurements over multiple biological conditions, we construct a 2D map where genes are placed so that their proximity is proportional to the similarity of their expression profiles. Scientists can use the T-cell co-regulation map in Figure 4.2 to find other genes that co-regulate with genes of interest and to understand how these genes co-regulate given the set of conditions described by the map. Immunologists can browse co-regulation maps to understand expression patterns in the featured conditions. Finally, scientists interested in downloading unfamiliar data can perform a preliminary investigation using maps hosted on the project website.

Our embedding algorithm was inspired by HiPP [130] but employs a different layout technique. As in HiPP, we use bisecting k-means to create a hierarchical clustering of the data. We then compute the clustering distance of two genes as the length of the path between their nodes in the clustering tree. We multiply this distance by the Euclidian distance between genes in the high-dimensional space described by the biological conditions. Finally, we use Chalmer's embedding [34] to project this combined distance in 2D. The discrete component introduced by the clustering tree is responsible for the clear demarcations between clusters observable in Figure 4.2. We initially used a standard projection but user feedback indicated that the lack of visible clusters detracted from analysis. Users considered the modified version preferable even when made aware that cluster boundaries were introduced artificially.

In rendering, glyphs are drawn over map regions, showing the aggregated expression profile of genes in that particular region along with the standard deviation. The size of aggregated regions is zoom-dependent; as zoom level increases averagings are performed over smaller sets for increased averaging accuracy. This is achieved by linking zoom to cluster-cutting of a hierarchical clustering of 2D projected distances. In low-level zooms, genes are represented by heatmap glyphs that color-code the expression value of that gene at each condition, giving users access to individual data values. The color scheme chosen was blue-green-yellow-red to maximize the perceived expression difference.

For the Google Map implementation, the visualization was rendered to tiles, gene positions were exported to a text file, and gene expressions were coded as one-byte values to limit size and were exported to a text file. These elements are used in the Javascript + Google Maps + Protovis map implementation in Figure 4.2. Users can search for a gene and highlight it via a marker. They can also select a group of genes by drawing a selection rectangle. If the selection is small enough (100 genes in our implementation), a heatmap representation is rendered using the Protovis library. The list of selected genes can be exported for further analysis.

4.2.2 Example 2: Gene Expression Heatmaps

Description and usage scenario: Given a list of genes, each with multiple expression measurements corresponding to a set of biological conditions, a rectangular heatmap representation is constructed in which each row corresponds to a gene, each column to a condition and each cell is a color-coded expression value. Rows and columns are arranged so that co-regulated genes and conditions are placed together. Scientists interested in T-cells can access a number of heatmaps corresponding to different types of genes to understand regulation patterns.

This map (see Figure 4.3) exemplifies a low-cost map implementation. The R library was used to generate a heatmap clustered on both genes and conditions. Text files with the genes and conditions in the ordering occurring on the heatmap were exported. The heatmap image was split into tiles and used to generate a Google map. Protovis was used to attach to the right and the bottom sides

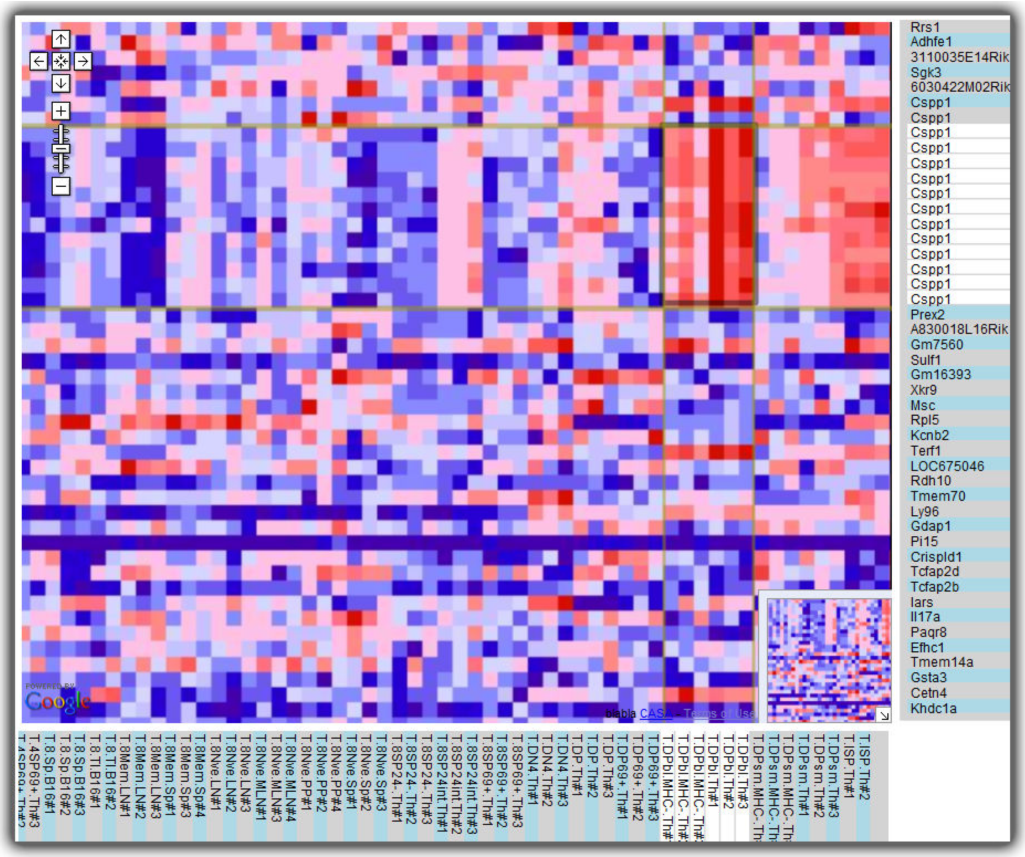


Figure 4.3: A heatmap representation is displayed as a map, with gene and cell type axes implemented in Protovis attached on the right and at the bottom. The axes are linked to the map's zooming and panning so that users can identify which genes and cells they are looking at. Selection of an area of interest prompts the highlighting of the corresponding cell types and genes.

of the map axes with gene and condition labels. These axes are synchronized to the map's zoom and pan operations so that labels for the currently viewed regions of the heatmap are always within view. Users can select a region on the map and prompt the highlighting of the corresponding genes and samples.

4.2.3 Example 3: Genome Map

Description and usage scenario: Given expression values over a set of conditions for any gene, we create color-coded expression glyphs at genes' genomic coordinates. Scientists can use this map to analyze connections between gene function and genomic location and can identify co-located genes that exhibit similar expression. Such maps can also be used to query and highlight on the map regions of the genome that are enriched in genes belonging to particular classes, either defined by an expression (e.g. genes that have higher expression in condition 1 relative to condition 2), or by functional category (e.g. all tyrosine phosphatases).

Gene expression is mapped to a blue-green-yellow-red color scheme. Glyphs color-coding expression values in every condition are created for each gene; a gene-name label is included. The 21 mouse chromosomes are arranged vertically, each extending horizontally. Following user feedback, no space warping or distortions, such as in [1, 113], have been used. The expression glyphs are mapped onto this space based on gene location. We use no aggregation of expression for different zoom levels because inspection of genomic maps and user feedback indicate that co-located genes most often do not have similar expression patterns.

Genes are not uniformly distributed on chromosomes; instead, regions with high and low gene density alternate. In high-density regions the space available to render a gene, assuming finite zooming, is limited and often insufficient to ensure visibility of the glyph elements. We therefore spread gene glyphs apart while keeping them anchored through a line to their true genomic positions, as seen in Figure 4.4. We use an iterative force method for this purpose.

The visualization is rendered to tiles and gene positions exported to a text file; these elements are used in a Google-Map implementation.

Gene search and highlighting of sets of genes are supported. For the latter, results of queries of type 'genes with expression in condition A higher than expression in condition B' are exported along with the map. Users can select these queries to highlight genes. The highlighting marker is an image with high alpha in the center and fading alpha towards the boundaries, so that the closer two highlighted genes are, the more their markers amplify each other, as seen in Figure 4.4. This ensures that regions with a high density of marked genes stand out even at overview zoom levels.

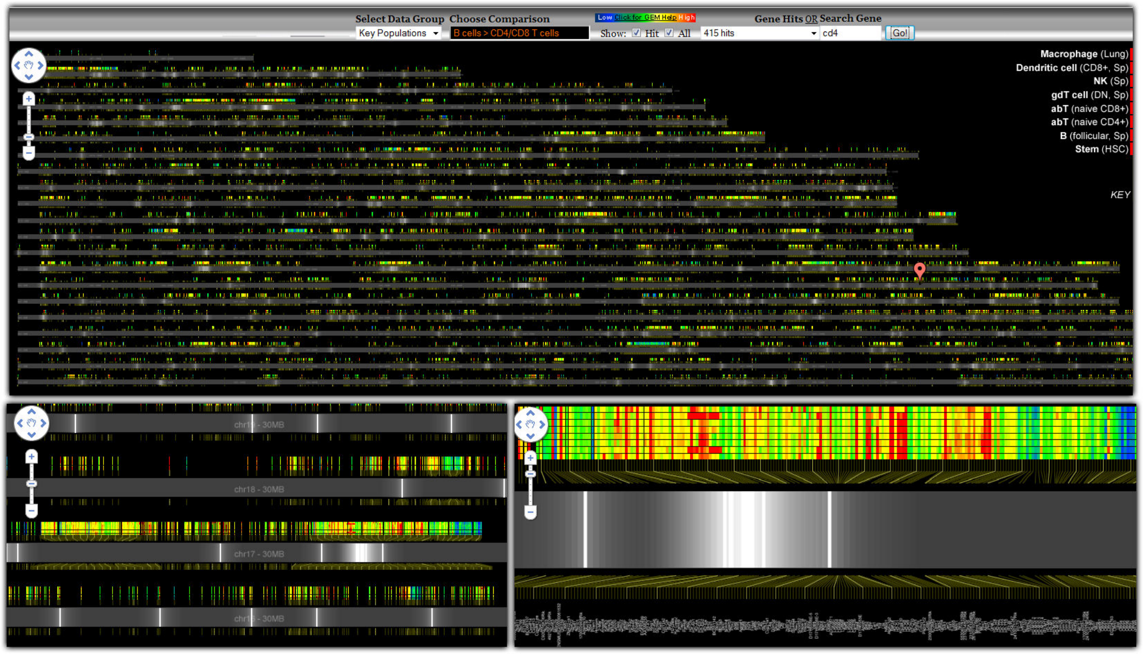
The map makes possible three levels of visual queries depending on the zoom: regions of highlighted genes stand out at the whole genome perspective; at a slightly zoomed-in level, regions with similar expression stand out by virtue of similar color patterns in gene glyphs; at full zoom, individual gene expression patterns become visible.

4.2.4 Example 4: Protein Interaction Networks

Description and usage scenario: Given proteins and interactions between them obtained from a public database, a node-link map is created (see Figure 4.5). Following a proteomic experiment, a set of active proteins is determined; a large percentage of them are unknown to the scientist in terms of function and interactions. To start the analysis, the scientist loads our map and superposes the experimental proteins. The scientist goes through experimental proteins, learns their interaction neighborhoods and determines if they are candidates for further analysis.

Drawing protein interaction networks as static maps is challenging due to clutter and because related data are not necessarily co-located. Also, because of long edges, zooming may not define a useful data query. To overcome these challenges, we use vertex splitting and zooming to filter out proteins based on a protein importance measure. These design decisions were subsequently validated during our evaluation.

To ensure co-location of linked proteins and to reduce clutter, we use Eades' [58] vertex splitting with a layout algorithm inspired by [68]. The layout space is interpreted as a rectangular grid and



Select Data Group: Key Populations

Choose Comparison: B cells > CD4/CD8 T cells

Show: ☑ Hit ☑ All

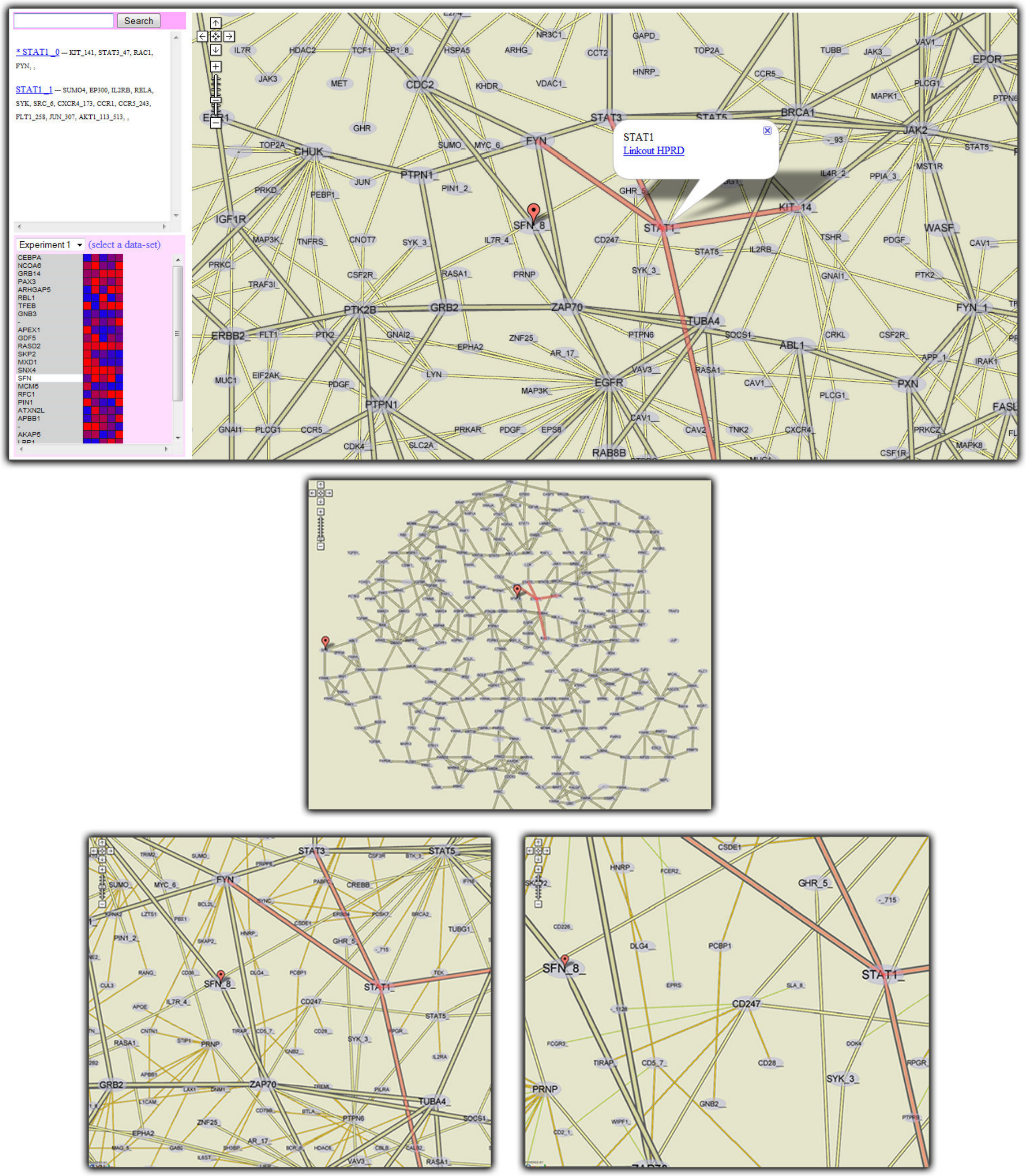
Gene Hits OR Search Gene: 415 hits cd4

Figure 4.4: Gene expression data measurements over eight cell types of the entire mouse genome are mapped onto genome coordinates. The top view shows the general analysis framework as presented on the Immgen website; zoomed-in views appear at the bottom. Three types of visual queries can be performed, depending on the zoom. At an overview, lists of relevant genes can be highlighted using Google markers with custom icons - white lines with alpha gradients on each side marking regions with interesting expression characteristics. At an intermediate zoom (lower left), regions with similar expression can be identified: a blue low-expression region is visible at center right. At a zoomed-in level individual expression values and gene names can be identified.

pair-wise force computations between nodes are restricted to nodes located in neighboring grid cells. Discontinuities at cell boundaries are reduced by fuzzy assignment of nodes to cells. Specifically, a node close to its geometric cell's boundaries, will have a fair chance of being considered as being part of a neighboring cell.

Once the spring system reaches stability, tensions on nodes determine the opportunity for a vertex split. Given a dividing line running through a node, force vectors acting on each side of the line are added together and projected on a direction perpendicular to the dividing line. Due to performance issues, our system never reaches perfect equilibrium; thus we consider the tension on a node as the minimum of the two opposing force magnitudes. Multiple division lines are probed to find the maximum tension on a node. The node with maximum tension is split if the tension exceeds a threshold (since our objective is not to planarize the graph). The splitting process involves creating two copies of the node and assigning edges corresponding to whether the force vectors they created were on one side of the split or the other.

To deal with clutter, we chose to prioritize what nodes are shown at overview zoom levels by computing a relevance measure for proteins. As in [167], this relevance measure is computed as a



Search

* STAT1_0 — KIT_141, STAT3_47, RAC1, FYN, ,

STAT1_1 — SUMO4, EP300, IL2RB, RELA, SYK, SRC_6, CXCR4_173, CCR1, CCR5_243, FLT1_258, JUN_307, AKT1_113_513, ,

Experiment 1 (select a data-set)

CEBPA
NCOA6
GRB14
PAX3
ARHGAP5
RBL1
TFEB
GNB3
-
APEX1
GDF5
RASD2
SKP2
MXD1
SNX4
SFN
MCM5
RFC1
PIN1
ATXN2L
APBB1
-
AKAP5
LRP1

STAT1
Linkout HPRD

Figure 4.5: Analysis of quantitative proteomic data in the context of a protein interaction network. The top panel shows an overview of the analysis setup. Time-course proteomic data is displayed on the lower left. The experimental protein selected in the list is highlighted on the map. A second protein was selected on the map and has its interactors and meta-information displayed. All instances of this protein are listed on the upper left, together with their interactors. Three additional zoom levels are shown on the lower row; as zoom level increases, less relevant proteins are added to the display.

function of a protein's intrinsic relevance and a relevance diffused from neighboring nodes. We alter the diffusion term to avoid elevating the relevance of proteins connected to a highly relevant protein but nothing else, a common situation in satellites of large protein hubs. Given a protein P, we first compute breadth-first-search subtrees rooted in P's neighbors, P itself not included. We name the subtree corresponding to neighbor N_i as N_{P,N_i} - the neighborhood of P through N_i - and compute $R(N_{P,N_i})$ - the relevance of N_{P,N_i}. As in [167], this is the maximum of the intrinsic relevance of each node in N_{P,N_i} weighted by a factor that decays exponentially with increasing distance from the node to P. We then consider P as connecting all possible pairs of N_{P,N_i}. We name this the connectivity relevance and compute it as $R_c(min(R(N_{P,N_i}), R(N_{P,N_j})))$. The final diffusion term is a maximum of all R_c's. A protein's intrinsic relevance is computed as a mix of the following: protein degree, occurrence in a specific pathway (e.g. T-cell), and occurrence in experimental data-sets obtained from our collaborators.

The relevance score is used to place proteins in bins, much like the city-versus-town distinction in a map analogy. Proteins are first sorted in descending order of relevance. A number of levels for the visualization is decided, five in our examples. Each bin i then receives a contiguous set of proteins from the ordered list. The layout is performed in stages, one for each bin. The most relevant proteins are laid out first, and their positions are then frozen for the second bin of proteins to be placed on the map. The discrete nature of the approach makes the layout suboptimal since higher-level layouts are not aware of lower-level graph topology. To alleviate this problem, we allow two or more levels to coexist while having a single one be current. Network elements in non-current levels exert less force than those in the current level and are less likely to be split. In a sense, then, non-current levels provide guidance for the current-level layout.

At rendering, nodes are displayed only if their bin-index is lower than a threshold based on the current zoom. Node sizes are adjusted by zoom level to reflect differences in relevance while preserving a sense of uniform scale throughout different zooms.

The visualization is then rendered to image tiles. To facilitate node selection, we export for each tile a corresponding text file containing the bounding boxes of proteins, or parts of proteins that appear in the tile. Upon a mouse-click on the map, the tile contents file is retrieved and the information it contains is used to check for intersections with proteins. For edges, each protein points to a file containing endpoint coordinates of its interactions. This implementation conforms to the Google Maps architecture and avoids the loading and client-side storage of large data files. As shown in Figure 4.5, we use polyline overlays to achieve node selection using the constellation technique [121], information pop-ups to display protein meta-data, and markers to highlight proteins of interest such as experimentally derived ones. To navigate among different copies of the same protein, a window on the side of the map lists all protein copies and their interacting proteins: clicking on the lines in the list causes the display to jump to the specific map location.

Time-course proteomic experimental data can be loaded and displayed as colored heatmaps on the left-hand side of the map, as in [86]. Multiple experimental datasets, for instance for normal and mutated cells, can be loaded and toggled between during analysis. Upon an experimental protein selection, Google markers will indicate the map location of the protein.

4.2.5 Example 5: Planar DTI Tractography Maps

Description and usage scenario: Given a DTI streamline dataset, three planar schematic representations show projections of important tract bundles onto the three principal projection planes: sagittal, coronal and transverse. These representations are distributed in Google Maps, with tract-bundle selection capabilities. Bundle statistics, in both textual and image forms, are computed and accessible in info boxes for tract bundles. Users can easily navigate through a large set of tractograms published as 2D maps and analyze differences in statistics for major structures such as the corpus callosum or cingulum bundle, or find datasets exhibiting desired statistical properties for more detailed analysis in an interactive system.

In Chapter 2 we describe a novel planar visualization of white matter tractograms. Here we introduce a web interface for this type of planar representation by integrating it into Google Maps and enhancing it with labels, statistics, and links. (see Figure 4.6).

The visualization system described in Chapter 2 was extended to render 2D projections into a set of image tiles instead of the screen. For each cluster, including both tract-bundled and endpoint clusters, information required for interaction and browsing is exported.

Selection information consisting of evenly spaced points along splines and thickness radii for splines contained in a cluster is exported. In line with the tile paradigm, instead of exporting this information to a single large file, it is divided geometrically across corresponding tiles and written as multiple tile-content text files. Upon user selection, the content file of a clicked tile is fetched from the server and its data analyzed for an intersection. This approach avoids loading and searching through large files.

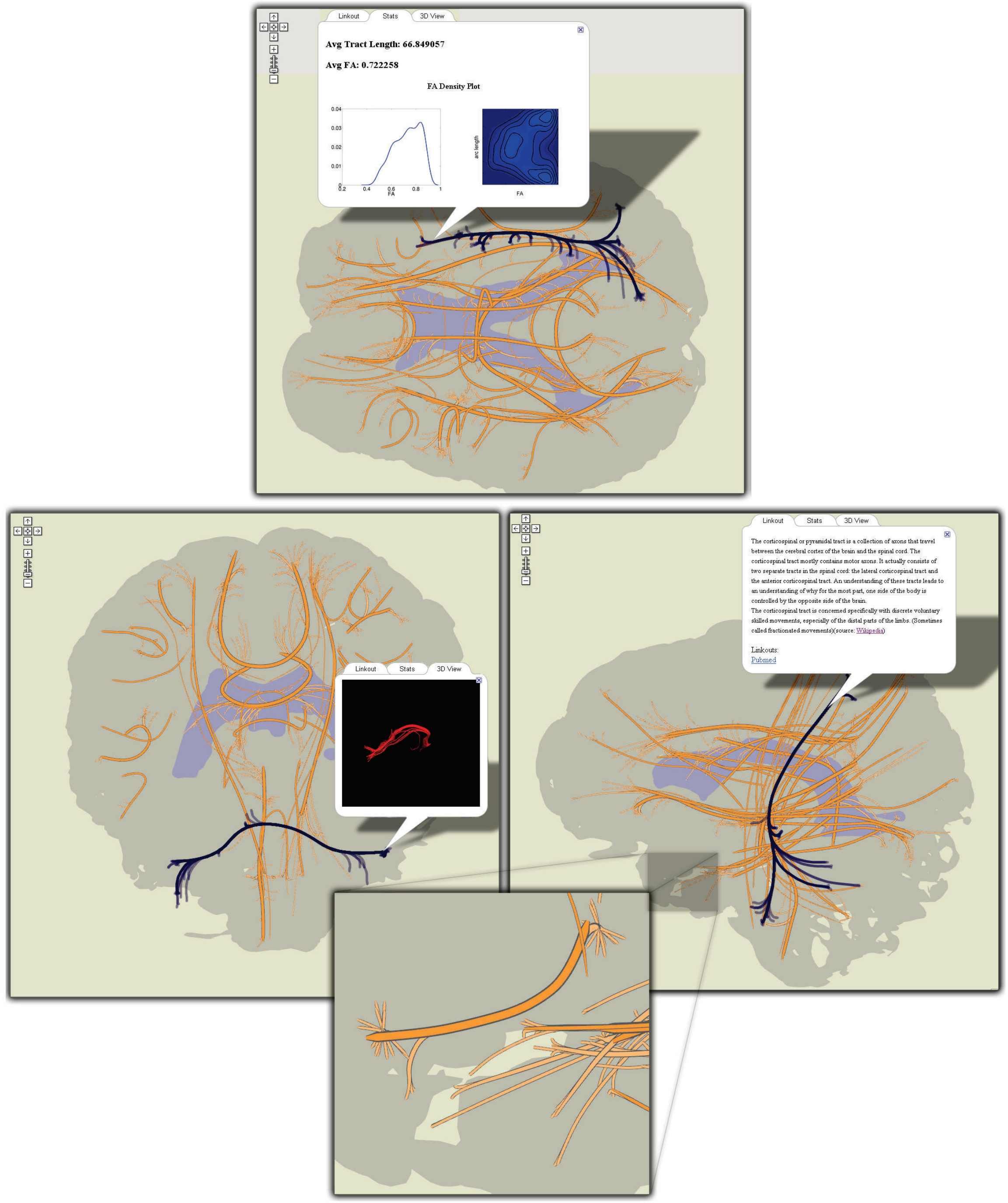
A valid cluster selection is marked on the map with polyline overlays running over tract splines contained in the selected cluster (see Figure 4.6). For this purpose, spline coordinates for each cluster are exported to files indexed by a unique cluster identifier.

Finally, for each tract cluster a variety of metadata accessible during map browsing in information-boxes, as shown in Figure 4.6, is also exported. A short description and links to the most relevant publications or research can be manually added for major tracts. A few 3D poses of each tract bundle are pre-rendered and exported as animated GIF images, indexed by the cluster identifier. Statistical data, in both textual and graphical form are computed for each cluster and written as HTML content to cluster indexed files. This information is loaded and displayed in tabbed information boxes at the user's request.

4.2.6 General Design Elements

Here are some general design guidelines that can be distilled from the previous examples.

Data size and specification: To compensate for their static nature, pre-rendered visualizations should encompass all data associated with a scientific problem. Thus, a visualization can be useful



Avg Tract Length: 66.849057

Avg FA: 0.722258

FA Density Plot

The corticospinal or pyramidal tract is a collection of axons that travel between the cerebral cortex of the brain and the spinal cord. The corticospinal tract mostly contains motor axons. It actually consists of two separate tracts in the spinal cord: the lateral corticospinal tract and the anterior corticospinal tract. An understanding of these tracts leads to an understanding of why for the most part, one side of the body is controlled by the opposite side of the brain.

The corticospinal tract is concerned specifically with discrete voluntary skilled movements, especially of the distal parts of the limbs. (Sometimes called fractionated movements)(source: Wikipedia)

Linkouts:
Pubmed

Figure 4.6: DTI tractography data projected onto the sagittal, coronal and transverse planes. Major tract bundles are represented schematically by their centroid tract; individual tracts in bundles are linked from the centroid bundle to their projected end points. Zooming in allows access to smaller clusters of tracts. Bundles can be selected and pre-computed statistical data along with 3D views of the tract bundle ("brain view") can be displayed.

for many queries, since data specification can be done during visualization through zooming, panning and highlighting. Our work exemplifies this approach. In the genome viewer three different visual queries can be performed based on zoom level: highlighting regions at an overview zoom, identifying regions with similar expression levels at intermediate zoom, and access to gene name and expression at a detailed zoom.

Individual visualizations sometimes need to be adapted to suit this approach. Our protein interaction networks use vertex splitting to enable queries by zoom-and-pan and a zoom-linked filter to address clutter. Our co-regulation map uses expression glyphs that guide users towards gene groups with specific expression patterns.

Use: Unlike advanced analysis systems, we have only targeted exploratory, preliminary and casual browsing of data or lightweight analysis tasks. As we will show in the following section, fast and intuitive access to visual perspectives of a dataset, even if less flexible than complex systems in terms of interaction and queries, can help in some cases accelerate analysis. It is hard, however, to determine how suited this approach is in the context of more complex functionality.

Users: Users can be divided into data consumers and data producers. In our experience, the former often perceive a dataset to have a low reward-effort ratio when they are unfamiliar with the type of data, are generally computer averse or lack access to a computational infrastructure. The browser visualizations targeting such users should be sparse and intuitive. This may seem self-evident, but state-of-the-art visualization systems commonly require scientists to understand visualization-specific jargon (e.g., select a specific graph-drawing algorithm).

Data producers want to distribute visualizations along with their raw data so that fellow researchers need not run their own analysis. Data producers will use an interactive system to create the browser visualizations. The assumption is that they are specialists in the data they are distributing, so that a system can use more complex visualization metaphors.

Development overhead: Development overhead can vary greatly among visualizations: our heatmaps are just static images augmented with basic interactivity, co-regulation information had to be first projected in 2D, and protein interaction networks required an entirely new drawing algorithm. A simple heuristic is that the overhead depends on the effort required to planarize the information displayed (e.g., relational data is harder than projected multidimensional data) and on the amount of data shown.

Deployment: Google Maps visualizations can be designed to work without dependencies on databases and server-side scripting. In such cases they can be deployed by simply copying a directory structure to a web server. This was an important factor for our collaborators in deciding to adopt this mode of representation.

4.2.7 Implementing Interaction

While reiterating that complex interactions are not the focus of this approach, we give below a few interaction patterns common in visualization that are possible in implementations based on Google Maps.

Selection/Brushing: For selection, positions of selectable elements have to be exported in data files, along with the pre-rendered visualization. This information is used to translate coordinates of mouse events into selections. In the co-regulation viewer and heatmap, users select genes by drawing enclosing rectangles. In the white-matter visualization we export curve trajectories for each tract-cluster, and use the proximity of a mouse click to a curve as a selection heuristic.

Highlighting: Elements selected through interaction or search can be highlighted using markers or polylines (traditionally used to highlight routes in digital geographic maps). Figure 4.2 illustrates a group of selected genes identified by markers. Polylines are used to implement Munzner's *Constellation* technique [121] of highlighting node and neighbor selections on the protein interaction network (see Figure 4.2) and to highlight tract-cluster trajectories on the white-matter visualizations (see Figure 4.6). Finally, images shown as markers can be customized to create more complex effects. In the genome browser for instance, multiple co-located markers with alpha gradients create an additive visual effect.

Semantic zooming: Our protein interaction network illustrates semantic zooming by displaying additional proteins with each increase in zoom level. The map framework allows developers to show different images at each zoom level. A scene can thus be pre-rendered at different zoom levels, each with its own visual abstractions. Two important factors to consider are that a visualization can have only as many abstractions as zoom levels and that exported images double in pixel size with each additional zoom level. This should be taken into consideration in designing the number of abstractions, as thirteen-level visualizations are infeasible to distribute (see following section).

Filtering: Semantic zooming can be used to implement filtering. As mentioned before, our protein interaction network (Figure 4.5) illustrates this concept. While not implemented in any of our visualizations filtering could also be achieved by rendering multiple complete tile-hierarchies for pre-determined filtering conditions. Completely dynamic filtering is infeasible using pre-rendered visualizations.

Data aggregation/abstraction: In our co-regulation viewer we average expression values over groups of genes at varying levels of specificity. In the genome viewer we contemplated displaying aggregated expression values over larger genome regions at overview zooms to deal with gene density, but chose a different approach following user feedback. Semantic zooming is, however, a good way to implement varying degrees of data abstraction. Another way is to use combinations of markers with custom icons to create glyphs that show aggregated data; this has the advantage that such effects can be created programmatically at run time. A simple example is seen in our genome browser where selection glyphs create an aggregated visual effect.

Details on demand: Figures 4.2,4.3,4.5,4.6 illustrate how information popups are used to retrieve information about visualization elements. Figure 4.6 shows how pre-computed statistical data and 3D-poses can even offer different perspectives of selected data subsets. A second detail-on-demand implementation is shown in Figure 4.3: mouse hovering generates a tooltip overlay. For more interactivity, browser-side graphics can be coupled with Google Maps. The co-regulation map (Figure 4.2) uses Protovis to show expression values of user-selected genes as heatmaps. We note

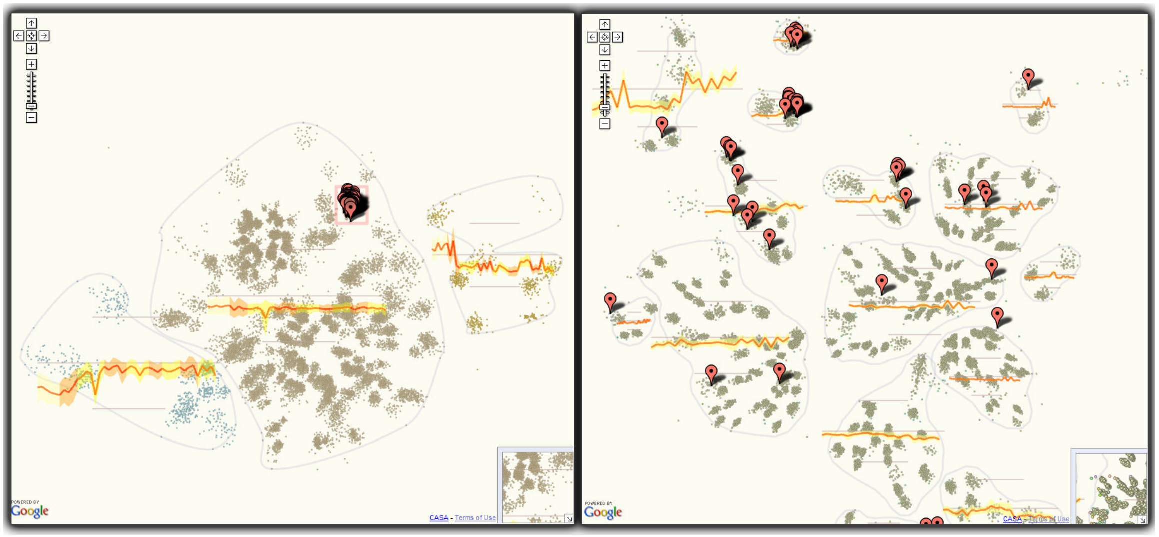


Figure 4.7: Linked co-regulation maps of the T-cell (left) and B-cell (right) families. A selection in the T-cell map is reflected onto the B-cell map. A few groups of genes that are co-regulated in both cell families are noticeable by inspecting the upper part of the B-cell map.

that information used in the detail views (e.g. expression values, 3D-poses etc) must be exported along with the rendered tiles.

Overview+Detail: The implicit Overview+Detail mechanism in Google Maps is the mini-map. However, more complex interactions can be achieved with browser-side graphics or multiple synchronized Google Maps on the same page. The closest feature to this in our implementations is the dynamically generated heatmaps in the co-regulation viewer. However, it would be easy to extend the protein interaction network by a linked Protovis viewer that displays local network information for selected proteins.

Brushing and Linking: Two of our evaluation subjects noted that linking several of our visualizations together can be beneficial. For example, linking co-expression views (e.g. for different cell families) can answer questions about conservation of gene function over multiple conditions. This functionality was implemented for the co-expression maps using browser cookie-polling, as shown in Figure 4.7.

4.2.8 Improving Performance

Below are a few considerations for improving the performance of tiled visualizations.

None of our visualizations required more than nine zoom levels. Assuming a tile size of 256 pixels, these translate into square images with $2^8 * 256 = 65536$ pixels on the side, at the largest zoom level. Furthermore, the number of tiles quadruples at each additional zoom level such that these visualizations consisted of $\sum_{i=0}^{8} 2^i * 2^i = 87381$ image files. Efficient image compression is desirable to reduce space requirements and speed up tile loading. Tile numbers can also be reduced by exploiting that visualizations often contain areas of empty background. Thus, many tiles can

	All tiles		Non-empty tiles	
	PNG	JPG	PNG	JPG
Co-reg.	(5461 37.6)	(5461 39.9)	(3505 35.1)	(3505 30.2)
Heatmap	(5461 23)	(5461 29.4)	(2811 12.7)	(2811 19)
Networks	(5461 32.2)	(5461 33.6)	(4620 29.8)	(4620 25.3)
Brain	(5461 37.6)	(5461 39.9)	(3505 34.1)	(3505 32.2)
Genome	(5461 35.1)	(5461 38)	(4051 27.1)	(4051 30.3)
Genome*	(87381 263.4)	(x x)	(17630 100)	(x x)

Table 4.1: Number of tiles and disk space(MB) for the five visualizations with different image compression (PNG vs. JPG) and all tiles vs. non-empty tiles. First five rows stand for visualizations with 7 zoom levels; the last row corresponds to a 9 level genome browser.

be represented by a single background-tile. Coordinates of background tiles are exported at the time of rendering and subsequently decoded by the Javascript implementation. Empty tiles are usually compressed into smaller files by default (due to uniform coloring) and their number is visualization dependent. Still, performance gains remain meaningful and typically grow considerably with increases in a visualization's zoom levels. Table 4.1 summarizes these improvements on several of our visualizations.

As mentioned in the previous section, interaction and data on demand rely on exporting additional information at rendering time that must be fetched and used by the browser visualization. Loading this data at once, during initialization, can freeze the visualization and result in large memory loads. Instead, in line with the tile approach, the information should be split in multiple files and retrieved only when an interaction demands it. For example, information about the shape of the curves in the white-matter visualization is split over a 10 × 10 grid spanning the visualization. Upon a mouse click, the corresponding cell content is fetched and tested for intersections. If an intersection with a tract cluster is found, a file containing information about this cluster (e.g., cluster trajectories for highlighting, metadata to be displayed in information pop-ups) is retrieved. This ensures that visualizations remain responsive during interactive tasks.

4.3 Evaluation and Findings

An anecdotal evaluation of map visualizations yielded the following general feedback: in a considerable number of tasks scientists liked lightweight, familiar visualizations; our visualizations were deemed intuitive and easy to use; users pointed out the possibility of collaborative work; one lab coordinator appreciated the ease of disseminating data and has decided to change his database-driven distribution to our map approach; each specific visualization offers improvements over existing methods.

Four proteomic researchers interested in T-cells and Mast cells, from two separate labs, evaluated the protein network map. Four geneticists working with T-cells and NK cells evaluated the coregulation map, heatmap and the genome browser at the end of the development process. Three neuroscientists evaluated the brain projection maps. The Immgen project coordinator was consulted throughout the development process and evaluated the genome browser and the heatmap during implementation. One neuroscientist provided his input on the tractography maps. Below is a summary of the feedback, followed by specific feedback on each of the five visualizations.

4.3.1 Evaluation Summary

All users rated ease of use as higher than other systems they have worked or experimented with. They were excited to be able to run the visualizations in a browser and several stated that this makes them more likely to use the visualizations. Most subjects said the available features are enough for quick data analysis. The main workflow for the biological maps was to project genes or proteins of interest onto existing maps. The neuroscience experts found the web interface with the digital map interaction useful for both quick data inspection and collaboration. Most users were content with the provided feature sets, interaction and visualization, while some asked for more hyperlinking and metadata features. A majority of our subjects identified the static nature of the maps as a non-issue. The Immgen project coordinator commented about the benefits of being able to accompany raw data with relevant visualizations and the minimal overhead in both maintaining the map systems and using them. He is actively considering switching the labs database-driven distribution system to our map approach.

4.3.2 Gene Co-Expression Map

All subjects agreed that the co-expression map is useful. Three subjects would use the maps by projecting their own genes of interested onto one or more cell spaces. Our fourth subject would also look for global patterns of co-regulation, possibly over multiple maps, and suggested we link multiple maps in separate browser tabs. One subject suggested using this application to create customized datasets by selecting subsets of co-regulated genes from explored datasets.

All subjects found the interactions intuitive and the maps easy to use. One of our subjects thought data maps could be useful for researchers new to a lab since they could start analyzing data right away. She then extended this idea to non-Immgen members and mentioned she would like such visualizations to be present in other data sources as well. She added that her particular lab has good technical support, but that since she is close to graduating and considering doing research on her own, this approach seems very appealing.

Two of our users did not consider the static nature of the maps a drawback. The other two expressed the desire to customize the cell types over which genes are projected. However, they agreed that there are relatively few cell subsets that they would choose from and that multiple maps covering these possibilities would probably work. We note that our users were highly familiar with the Immgen data and their analysis was in most cases past the exploratory stage, explaining the desire for increased flexibility.

In terms of features, two users explicitly complimented the superposed expression profiles, stating that they summarize data well and can guide exploration. All users were happy with the heatmap upon-selection mechanism and with the ability to export selected sets of genes.

4.3.3 Gene Expression Heatmap

Our collaborator asked for browser-based heatmaps in the context of distributing Immgen data in heatmap form to the immunologic community. His concern was that the currently deployed static images are limited in exploring such representations, especially since overview analyses involve up to 500x1000 matrices. Our fix of providing gene/cell axes coordinates with map zoom and pan was deemed a solution to this problem.

Of our four evaluation subjects, only one used large heatmaps as part of his analysis. He was excited about our distribution mode and said it was a significant improvement over his current analysis. He said the sticky axes made navigation much easier and complimented the ability to select a rectangular region on the map to highlight genes and cell types. Commenting on this type of visualization from the perspective of a user who is highly familiar with his data and would like to build his own maps for analysis and publication, he said that the ability to share visualizations via web-links makes collaboration easier.

4.3.4 Genome Map

The genome viewer benefited from iterative development and evaluations between implementation stages. The initial insight was a need for an overview analysis of gene expression in the genome space. Our collaborators' assessment was that bringing forth correlated regulation of neighboring genes would help gene-regulation variation with cellular differentiation be better understood. The specific question was to what extent do genes that are adjacent share co-regulated expression patterns.

A first design item validated during development was to not use data aggregation for different zoom levels and to rely on additive visual cues of individual items. For instance, while individual gene and cell expression values are not discernible at an intermediate zoom level, the average expression and variability of neighboring genes remain salient. This proved effective: our collaborator noted that this allowed him to identify inactive regions of the genome. For Immgen-specific cells, lymphocytes and other blood-borne cells, these seemed to be primarily the regions carrying long clusters of olfactory receptors.

Another design choice was to introduce regional highlights that let users visualize areas in which several genes meet a chosen pattern of expression. Contrary to his expectations, our collaborator found that such regions proved relatively rare once he could get a whole-genome perspective. He noted that genes with comparable patterns of activity tend to be dispersed, and that co-regulated clusters exist but are relatively rare. He also noted a striking example of neighboring genes with divergent patterns of expression in a few genes interspersed in the middle of the olfactory receptor clusters, which are known to be quite active in lymphoid cells. He concluded that there is likely a higher order of genomic organization that the genome map can help explore.

In our genomic evaluation at the end of development, results were less consistent. We believe this is because the subjects' interests were focused on specific genes of interest rather than on overview analyses of regulation patterns. However, the workflow noted in the heatmap and co-regulation map became immediately apparent in this visualization: users would load their own genes of interest and see how they fit on a specific genome map. Both co-location and expression similarity would be of interest to them.

Our subjects were positive about the map approach. We switched from a standalone system implementation to the web-map approach during development, responding to our collaborators' need to publish the projects data on the web. The ease of use, the visual appearance, and the mode of distribution were noted by several members of the Immgen project during our first release. The static nature of the maps seemed to be non-issues for our collaborator and most of the subjects in our final evaluation. One of our subjects expressed a desire to build his own maps, but agreed that researchers who are not very familiar with Immgen data would find this a good first contact with the data.

4.3.5 Protein Interaction Networks

Our subjects were excited about looking at large interaction networks in their browsers. The consensus was that the browser setup is highly effective and that they would choose it over other systems they are familiar with. They explained their choice by remarking that they don't like to spend time installing software and learning new features, and found the techniques we demonstrated intuitive and easy to use. At the time of the demonstration the prototype did not provide sufficient access to meta-data. That was the most common feature that was requested.

As for the design decisions underlying the map visualization, relevance-based filtering and vertex duplication, feedback was positive. The unanimous opinion was that relevance filtering was intuitive, finding that it corresponded to how they normally approach a new network: identify important or familiar proteins and then drill down to learn more about their neighbors. Another comment was that seeing familiar proteins and connections early helps reinforce their confidence in the visualization. None of our subjects thought that not seeing the whole network at once obstructed their exploration, while one explicitly stated that the simplified view is superior to cluttered network visualizations he has seen before. Three of our subjects were satisfied with how we currently compute protein relevance, while one thought protein connectivity was enough since highly connected proteins correspond to important proteins.

In our first demonstration, hyperlinks to protein copies were placed directly on the map next to protein glyphs. The first researcher we interviewed said he found the concept of split proteins disorienting and would probably not be able to work with it. We then put the hyperlinks in a list on the side of the display, as in the current prototype 4.5. The researcher thought this visualization was improved because he could go through the copies of a protein systematically to reconstruct its neighborhood. His final response was that splitting proteins is not desirable but is acceptable if it can simplify the visualization. Our other three subjects stated that multiple copies of proteins would not get in the way of their analysis at all; one even said he preferred looking at proteins this way because it made their interaction neighborhoods more apparent. Another subject made the point that pathway drawings often had multiple copies of proteins. When we argued that these copies are biologically motivated while ours aren't he agreed but said they are still familiar with the technique. Generally, it seemed that the primary task they need to perform on protein interaction networks is finding all interactors of a protein. Our subjects thought that the list of protein copies, hyperlinked to the map, allows them to do that without obstructing analysis.

4.3.6 DTI Brain Maps

This visualization was evaluated with a neuroscientist following an informal protocol. The static map implementation of the projection-based DTI visualization was evaluated alongside a stand-alone system implementing the same views interactively linked to a 3D stream-tube model. Our scientist commented that he would primarily use the system because complex tract selections were required that cannot be performed in the static map. However, he pointed out the unique opportunities offered by the map implementation: collaborating with other scientists by sending links, being able to look at datasets anywhere, any time, and browsing through datasets before importing a model into the stand-alone application. He described the projection-based visualizations as intuitive, especially compared to other dimensionality-reduction methods, and requiring no learning overhead. He also appreciated the 3D poses of tract bundles and statistics available in the map info-boxes. This evaluation led us to conclude that static maps are less suited for the 3D domain where complex interactions are needed, but can occupy a task-specific niche such as collaborative work and casual analysis.

4.4 Discussion

In this chapter we advocate for the dissemination of visualizations as large, static data maps with a small set of intuitive interactions. Here we present a few general consideration pertaining to this topic and a list of opportunities that map visualization opens.

4.4.1 General Considerations

As suggested by our evaluation, the low-overhead tile based approach we exemplify seems to be particularly attractive to researchers lacking access to a strong computational infrastructure, for unfamiliar datasets, and for casual data browsing. Our evaluation of the white matter visualization shows that in other domains this approach might be more narrowly useful. From our experience, the Google Maps API can also be a useful medium for gathering feedback on visual encodings, possibly developed as part of another system. Collaborators are more likely to provide feedback on visualizations that they can access and use with minimal overhead than on ones they must install and learn. Furthermore, concerns such as deployment and platform, rendering speed and interactivity, GUI and data formats become non-issues.

This work explores only the Google Maps API. However, we hypothesize that other Ajax tiled approaches would probably also be suitable for this approach. More generally, zoom-and-pan frameworks (e.g. Bing Maps API, Silverlight, OpenZoom) can be used in conjunction with a subset of the desing elements discussed in this chapter to develop similar visualization. Moreover, the development of a tiled frame-work designed to support data visualization rather than geographical maps could prove useful. Such a framework, if open source, would also alleviate concerns about licensing, support and stability associated with commercial products. Principles of sparsity and intuitiveness should remain the foundation of tile frameworks, since the proposed browser visualizations should not seek to rival complex systems.

Our examples also demonstrate the synergy between maps and interactive web elements implemented in Protovis. Focus+context visualizations can be created so that maps offer the context while focus views are implemented in Protovis. However, we note that an essential guideline we advocate is simplicity; merely replicating the complexity of stand-alone systems on the web is not our goal.

4.4.2 Opportunities

We end the discussion with a list of opportunities for map-like visualization:

Linking maps: A biological concept is rarely explained by a single perspective on the data, so that linking multiple maps together can be beneficial. For instance, linking the genome map to the 2D co-regulation map can be used to test the hypothesis that co-regulation has a genomic location component. As indicated by one of our subjects, linking multiple co-expression maps (e.g. for different cell families) can answer questions about conservation of gene function over multiple conditions and would be a desirable addition to our framework. An initial prototype of this functionality is shown in Figure 4.7

Viewing maps on large displays: Most large-display setups have a way to display static images. In this case, zooming would be performed by moving closer to the display. Limitations, however, specifically on semantic zoom, must be imposed on the visualization.

Collaborative work: During our evaluation the users were excited about the opportunities of collaboration offered by maps. Exchanging interactive images rather than static ones and sending links rather than datasets was positively received. This concept can easily be extended to support collaborative work; the static nature of the visualizations are in this case an advantage. We would like to add annotation capabilities to our maps to enable researchers to exchange ideas. The static nature of maps is an advantage here too, since it ensures that each user has the same view of the data and that shared comments target the same visualization elements.

Easy instrumentation: An important component of visualization is understanding how visualizations are used. Due to the minimal interaction advocated, maps should be easy to instrument. In fact, one of our deployed maps has recently been instrumented using the Google Analytics framework. User interaction capture will be implemented shortly.

4.5 Concluding Remarks

A series of cognitive studies led Hegarty et al [77, 154] to conclude that "cognitive science research indicates that the most effective visual representations are often sparse and simple. When given control over interactive visualizations, people do not always use these technologies effectively or choose the most effective external representations for the task at hand."

We presented a low-overhead approach that can facilitate browsing for a range of unfamiliar scientific datasets, that relies on pre-computed visualizations carefully prepared by data experts for distribution with sparse interactions, so that end users can access readily analyzed views of scientific data. We build on the familiarity of the Google Maps framework and leverage its functionality to distribute those views. In an anecdotal evaluation we showed that this data-distribution mode is particularly suited for exploring unfamiliar datasets, for casual data analysis such as at home or during commuting, and for lab biologists who lack access to strong computing infrastructures. Additionally, we lay out design guidelines benefiting those wanting to create such visualizations and we describe five concrete example visualizations.

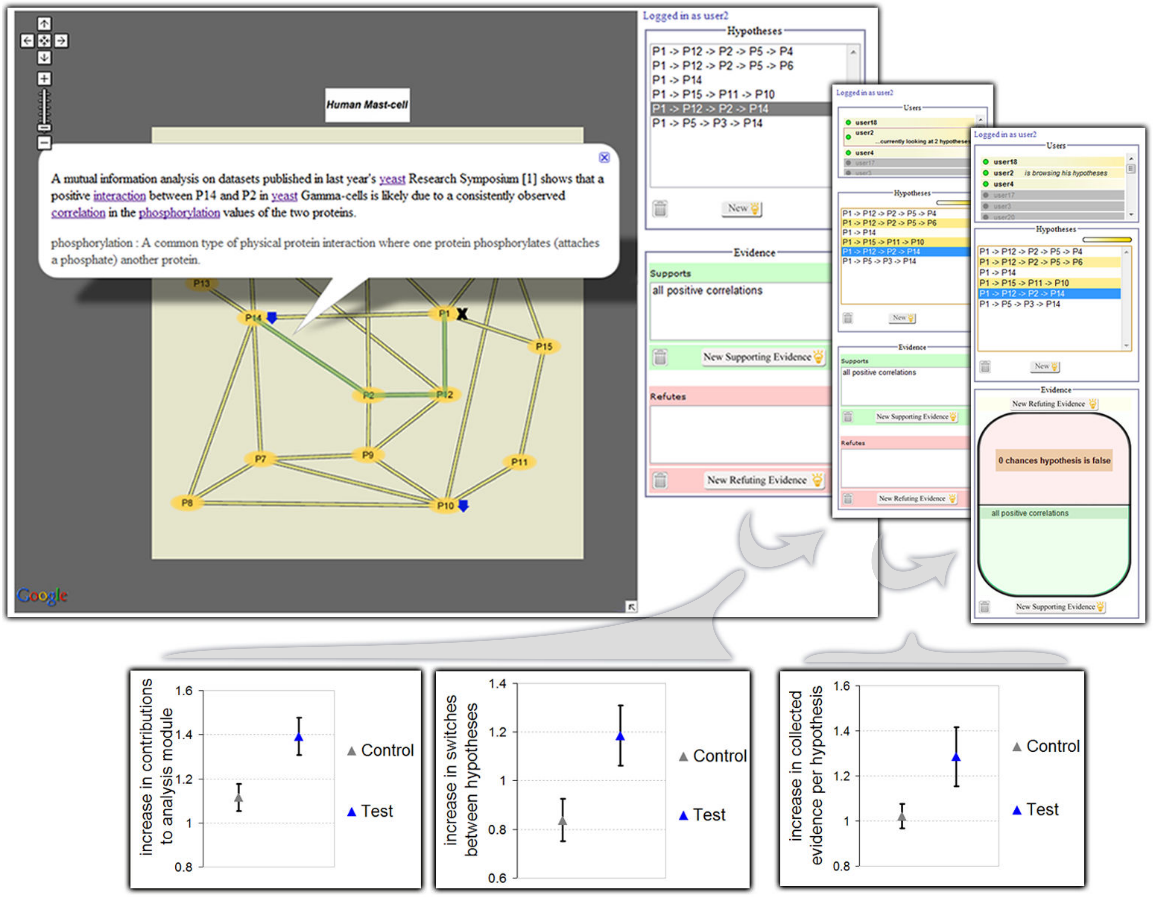
Chapter 5

Improving Scientists' Analytic Strategies through User Interface Changes

Sensemaking is a cyclical process in which humans collect information; examine, organize, and categorize that information; isolate dimensions of interest; and use the results to solve problems, make decisions and take action [32, 133, 135, 140]. Visualizations improve sensemaking by accelerating the search for information, facilitating the discovery of patterns, and providing means for evaluating various hypotheses [32]. Traditionally, analysis was limited to few datasets and, in the absence of high-performance visualization methods and systems, understanding and exploring one dataset would play a major part in the analysis process. However, data gathering and visualization have evolved to the point where, jointly, they can describe the behavior of systems with intricate structures (e.g., biological systems). Such systems, and even their functional sub-units, are rarely described by just a few pieces of evidence, whose discovery visualization may facilitate. In such scenarios, the aggregation of individual pieces of evidence into high-level, experimentally testable hypotheses represents the more significant proportion of the sense-making process.

A research opportunity thus arises: designing methods and interfaces that work in concert with visualization systems and allow researchers to aggregate their findings into cohesive scientific stories. Catalyzed by growing intelligence needs after 2001, the new field of *Visual Analytics* (VA) emerged out of traditional visualization efforts to address such problems. Illuminating the Path [165] introduced and defined visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces".

Here, we advance the VA agenda by providing experimental support for the following hypothesis: we can use subtle changes in the interfaces of visual analysis systems to influence users' analytic behavior and thus unobtrusively guide them towards improved analytic strategies. An overview of our results and methodology is shown in Figure 5.1.


Human Mast-cell
P13
P14
P1
P15
P2
P12
P7
P9
P11
P8
P10
Google
Logged in as user2
Hypotheses
New
Evidence
Supports
New Supporting Evidence
Refutes
New Refuting Evidence
Users
user18
user2
...currently looking at 2 hypotheses
user4
user17
user3
user2 is browsing his hypotheses
user20
0 chances hypothesis is false
increase in contributions to analysis module
increase in switches between hypotheses
increase in collected evidence per hypothesis
1.6
1.4
1.2
1
0.8
0.6
▲ Control
▲ Test

A mutual information analysis on datasets published in last year's yeast Research Symposium [1] shows that a positive interaction between P14 and P2 in yeast Gamma-cells is likely due to a consistently observed correlation in the phosphorylation values of the two proteins.

phosphorylation : A common type of physical protein interaction where one protein phosphorylates (attaches a phosphate) another protein.

P1 -> P12 -> P2 -> P5 -> P4
P1 -> P12 -> P2 -> P5 -> P6
P1 -> P14
P1 -> P15 -> P11 -> P10
P1 -> P12 -> P2 -> P14
P1 -> P5 -> P3 -> P14

all positive correlations

Figure 5.1: By making subtle, non-functional changes in the interface of an analysis support module (top) we generated statistically significant changes in users' analytic behavior in a visual problem-solving task. A first set of changes nudged subjects to increase their use of the analysis module by 39% (lower left, $p = 0.02$) in an attempt to support our subjects' working memory. It also caused them to switch among hypotheses 19% more often (lower center, $p = 0.03$), indicating more consideration of alternative hypotheses. A second set of changes then led subjects to gather 26% more evidence per hypothesis (lower right, $p = 0.01$). These three increases compare to smaller or negative variations in a control group (+15%, −17%, −2%).

This work was motivated by cognitive science research showing that human thinking is subject to heuristics and biases that often lead to suboptimal decision making [76]. Simon [151] for instance shows that humans are subject to "satisficing", a heuristic that limits the search for possible hypotheses to the first that is good enough. Wason introduces the famous 2-4-6 study [174] which shows that hypothesis confirmation is used instead of the rational strategy of hypothesis disconfirmation. Such effects are not limited to laboratory studies or naive subjects but manifest in scientific research as well [54]. Cognitive science research also shows that such biases can be partially overcome with external or contextual help: Dunbar shows that peoples' confirmation bias can be overcome, and describes how analogy and unexpected findings often lead to consideration of multiple hypotheses in scientific domains [54, 55].

Specifically, we report results from a controlled study in which subjects were asked to complete three analysis sessions using a system consisting of a visualization and an analysis support module. Two sets of non-functional changes were made to the analysis support interface before the second and third sessions. These changes were designed to improve three hypothesized or observed analytic deficiencies: analysts' excessive reliance on memory, an inability to consider hypotheses in parallel, and an insufficient search for evidence. Our quantitative results show that the interface changes succeeded in alleviating these deficiencies. Compared to a control group, our test subjects used the support module more, they switched between hypotheses more often, and they collected more evidence per hypothesis. Our data doesn't merely show that changes in interfaces translate into different user behavior, but demonstrates that we can leverage interface design and cognitive principles in controlled ways to overcome known analytic deficiencies.

Our approach was inspired by two similar paradigms: Thaler and Sunstein's work on libertarian paternalism popularizes the notion of "choice architecture design" [164]; Fogg [64] introduces and defines the concept of "Persuasive Technology". Both approaches advocate for designing a "choice architecture" or system's interface such that users are "nudged" to make decisions in their and the society's best interests. We posit that this approach may facilitate the use of visual analytics expertise to correct biases and heuristics documented in the cognitive science community.

To the best of our knowledge, few concrete attempts have used visual analytics techniques to align *descriptive analysis* (i.e., what people actually do to derive a solution), to *normative analysis* (i.e., rational strategies of deriving the best solution), and none have done it using the "nudge" paradigm. Instead, VA research is traditionally aimed at understanding and modeling the sensemaking process [23, 133] and operations that need to be supported [84, 138] or at creating systems that offer ways of storing, annotating, exploring, and querying evidence sets. Such features support analysis in the same way visualization does, by facilitating access to information, but they do not necessarily structure analysis, a task still left entirely at the analyst's discretion.

We note that the aim of this work is not to introduce novel analysis support features or interface design guidelines, but to quantitatively measure the ability of a small set of such elements to nudge users towards normative analysis practices.

5.1 Related Work

In this section we show how our work is motivated by and relates to existing research. We start with an overview of analytic biases and heuristics. We continue with a description of previous work that inspired our approach: libertarian paternalism and persuasive technology. We end by illustrating how our results relate to and advance current visual analytics research. We note that the results presented in this chapter have been published in [93].

5.1.1 Limitations of Human Analysis

Humans are prone to a range of decision making biases and heuristics which can occasionally result in sub-optimal results [76]. A specific manifestation of such effects occurs in the context of hypothesis driven analysis. Simon [151] coins the term *satisficing*, a heuristic that limits analysis to a hypothesis that is good enough. Bruner and Potter [29] show that subjects cling to initial hypotheses and are unable to consider alternative explanations in an experiment involving image-slides with varying degrees of focus. More recently, Danner et al [44] show that three or more retrievals from memory of a specific means towards a goal will succeed in inhibiting competing means for the same goal. Finally, multiple studies have shown that the use of a single hypothesis leads to a bias in the way subjects evaluate evidence [111, 35].

Biases are also present in gathering and evaluating evidence pertaining to a hypothesis. According to the scientific method, the best way to test a hypothesis is to attempt to disconfirm it. However, researchers have found that subjects usually try to confirm their hypotheses rather than disconfirm them. That is, subjects will choose experiments that generate results predicted by their hypotheses. This is known as *confirmation bias* and is eloquently demonstrated in Wason's card test [174].

Many of the studies mentioned above have been conducted with naive subjects. As such, a question about the degree to which these observations hold in scientific or clinical reasoning remained open. Several studies show that such biases and heuristics manifest in the scientific and clinical domain as well, albeit perhaps to a lesser degree. Dunbar [54] used a scientific setting, replicating the discovery of a real biological finding, to demonstrate that single-hypothesis and confirmation strategies were predominately used, and that such strategies inhibited the subjects to replicate the scientific discovery. Similarly, Ben-Shakhar et al. [20] showed that clinicians at a Jerusalem hospital showed strong agreement with a priming suggestion when deciding on a diagnosis. Finally, Rodgers and Hunter [139] found that researchers investigating a favored hypothesis selectively deleted studies from a meta-analysis. On the bright side, Klayman argues that even though a confirmation bias exists, under certain circumstances, this is a good strategy to use [104]. Furthermore, in an "in vivo" study involving observations of scientists performing everyday analysis, Dunbar [55] found that while subjects do try to confirm hypotheses, their hypotheses will often change in the face of inconsistent findings.

There is significant evidence that analysis can be improved by using prescriptive analysis techniques, training in normative thinking, and external aids that amplify cognition. For instance, it is thought that bounded memory and attention inhibit the analyst's ability to consider multiple hypotheses. The distributed cognition theory suggests that individuals use their environment as an external aid to amplify their cognition. Clark and Chalmers [39] point to research in support of this theory. In visualization, authors in [32] state that explicit visual thinking increases an analyst's cognitive span. Dunbar's [54] study demonstrates that indeed if subjects were conditioned to pursue alternative hypotheses and disconfirming evidence, solutions to a scientific puzzle were reached more often. Dunbar [55] also describes how analogy and a string of unexpected findings can often lead to consideration of multiple hypotheses and novel findings. Elstein et al [61] describe a study that reveals that medical students using hypothesis-driven analysis outperformed those that used a data immersion approach. Scientific studies mentioned in [76] show how *multiple attribute utility theory* (MAUT) can reduce biases and heuristics such as the prominence effect that causes subjects to base a decision on a single attribute which they consider most important. Finally, literature rooted in the field of intelligence analysis provides anecdotal evidence as to the benefits of applying prescriptive techniques and algorithms to complex analysis tasks [97, 79, 76].

Visualization literature also points out that analysis can be improved by using visual aids. In [50] it is shown that subjects given a problem statement in visual form perform better than subjects given the same information in textual form. Savikhin et al [144] uses a specific example from economic reasoning to prove that visualization can help overcome heuristics used in decision making. In [23, 133] the authors call for augmenting the analysts working memory to increase the attention span for evidence and hypotheses and improving divergent thinking by encouraging users to consider alternative hypotheses.

5.1.2 Guiding User's Choices: Nudges and Persuasive Technology

Thaler and Sunstein's work [164, 159] in the field of behavioral economics popularized the term *choice architecture* — how a set of choices is presented to a consumer — and the concept of *libertarian-paternalism* — designing choice architectures that "nudge" consumers towards making decisions in their own interest (paternalistic) while unrestricting choice (libertarian). A similar concept was proposed in the HCI domain by Fogg [64] who defines *persuasive technology* as "interactive information technology designed for changing users' attitudes or behavior". More recently, Lockton [109] generalized Fogg's persuasive technology and linked it to Thaler and Sunstein's choice architecture model by introducing the "design with intent" concept. Broadly, this refers to design intended to guide user behavior across a range of disciplines from architecture to software. We build on these previous approaches and demonstrate empirically how the nudge paradigm can further the visual analytics agenda.

Sunstein and Thaler as well as Fogg motivate their approaches with two arguments, which they support with experimental evidence. First, any choice architecture or computer interface necessarily influences decision-making behavior, whether intentionally or not. This statement is demonstrated by studies showing how potentially unintentional choice designs, such as state-by-state opt-in versus opt-out organ donation programs, significantly impact people's choices. Second, as shown previously, ample research indicates that people's choices and behaviors are not necessarily aligned with their goals. Ill-formed preferences, default rules, framing effects, and starting points, all dominate important decisions and thinking processes. From a visual analytics perspective, if an analyst's objective is to select the optimal course of action based on available data, cognitive biases and heuristics can steer him towards erroneous results.

Both approaches have inspired scientific results that validated their feasibility. Thaler and Bernartzi [163] use an array of cognitive effects such as *mental discounting* (i.e., weighing current events more than future events) or *default options* to persuade employees of a company to increase their contributions to their retirement plan. In the technological realm, the enhanced speedometer [108] changes visual appearance based on the current speed limit (when known) encouraging users to stay within speed limits, while the smart sink [13] augments a normal sink with visual cues that make energy consumption apparent. These works have provided inspiring design models for the analysis nudges presented in this dissertation.

Finally, Thaler, Sunstein and Fogg, as well as subsequent research articles, address ethical questions raised by influencing choice or behavior. Thaler's view is that paternalism is unavoidable and that libertarian paternalism should ensure, as a general rule, that people can easily avoid the paternalist's suggested option. Fogg proposes a thorough investigation of the gains and losses of all parties involved in the development, distribution and use of a particular system to determine its ethicality. Subsequent papers augment Fogg's persuasive technology approach with additional ethical constraints or guidelines. According to Oinas-Kukkonen [126], persuasive systems may be defined as "computerized software or information systems designed to reinforce, change or shape attitudes or behaviors or both without using coercion or deception" and proposes that persuasion should always be open and unobtrusive. It also disputes some of Fogg's initial design suggestions, such as surveillance and conditioning as ethically unacceptable. The nudges used in our work abide byf these ethical principles.

5.1.3 Supporting Analysis through Visual Analytics

The work presented in this chapter extends the VA research agenda which focuses on designing interfaces and visualizations that support the aggregation of data insights into cohesive scientific theories.

Work in the field of VA falls broadly into two categories: theoretical research based on existent psychological studies or user evaluations, and applicative work. In the theoretical domain, work in [23, 133] presents a five stage sense-making model derived through Cognitive Task Analysis (CTA) and verbal protocol experiments with analysts to identify leverage points for visualization. Authors in [84, 138] analyze how users synthesize multiple collections of evidence in a collaborative setting, using a physical, visual medium. Their results, a break-down of analysis tasks with observed frequency/duration and insights into the workflows of collaborative sense making, are useful for deciding which analysis tasks to support. Finally, multiple position papers advocate for leveraging the expertise of cognitive science and intelligence communities in the context of visualization supported workflows [73, 158]. Our work is tangential to and motivated by such results.

At the opposite spectrum, new applications probe the feature and design space of analysis-support software. Several applications for thought mapping and evidence management use the paradigm of laying out reasoning artifacts on a canvas, either freely or as a tree/graph structure. Examples of such systems are: The Concept Maps [31], MindManager [115], The Analyst's Notebook [124], Visual Links [83], The Scalable Reasoning System (SRS) [132] and The nSpace Sandbox component [175]. Several systems depart from the canvas paradigm. Entity Workspace [22] operates only on textual evidence and uses grouping and linking as an organizing paradigm in a highly structured medium. In

HARVEST [71] users can not only visualize existing information, but also construct new analytical knowledge from existing information and use visualization on it. Authors in [177, 178] apply similar principles to multi-dimensional visualizations and use specific visualization characteristics to drive the organization of evidence. Finally, work in [59] departs from conventional methods by structuring analysis as short stories hyperlinked to evidence, a paradigm based on a narrative theory [63] suggesting that people are storytellers and excel at evaluating a story for consistency, detail and structure. In our evaluation we use design elements which we distilled from these existing analysis systems.

To the best of our knowledge, few concrete attempts have used visual analytics techniques to bridge the gap between descriptive analysis and normative analysis. As such, our work complements current research by using a visual analytics methodology to create a link between observed analytic deficiencies and corrected behavior. Perhaps closest to the work presented here are results by Savikhin and Maciejewsk [144] who demonstrate experimentally that a targeted visual representation can induce normatively correct decisions in an otherwise biased economic choice task. We extend this result by linking it to the more general nudging approach proposed by Sunstein, Thaler and Fogg, by using interface design in general, and by providing an experimental validation on a high-level analytic task.

5.2 User Study Design

We conducted a controlled user study to test our hypothesis that small changes in a visualization system's interface can be used to produce targeted modifications in users' analytic workflows. This section presents the design of this study. We start with an overview description of the methodology used and continue with an in-depth presentation of each aspect of the study.

5.2.1 Study Overview

Subjects completed an analysis task inspired by a real scientific problem using a visualization and an analysis support interface (Figure 5.1, top). Each subject performed three such analysis sessions at one week intervals. Each session lasted roughly one hour.

Thirty-six subjects, mostly undergraduate and graduate students, were divided into two groups: 21 test and 15 control subjects. The control group solved all three tasks using the same analysis support interface. Conversely, test-group subjects were given slightly different versions of the analysis support interface in each session. Specifically, two sets of interface nudges were added to the analysis system before the second and third sessions. We hypothesized that, while changes between sessions would be observed in both groups due to task-learning effects, the test group would exhibit additional effects due to the interface nudges.

The analysis task was inspired by the proteomic domain: finding causal paths in protein interaction networks to explain the interdependency of pairs of proteins that are not directly connected. None of the subjects was familiar with the task or background material beforehand. They were given a 20 minute tutorial at the beginning of the study. Our approach of using a relevant scientific setting and naive subjects was inspired by a study by Dunbar et al [54].

Our test system was instrumented to automatically log users' interactions. Subjects were also asked to distill their analysis in a written questionnaire at the end of each of the three analysis sessions. We analyzed the datasets both quantitatively, to find support for our nudging hypothesis, and qualitatively, to gain insight into how subjects approached their task.

5.2.2 Task Description

Subjects were asked to solve three artificially constructed analysis tasks inspired by workflows used by proteomic researchers studying protein signaling pathways.

Proteins are functional molecules within cells. They interact with one another forming complex causal pathways that determine the response of cells to events. Such protein interactions are the object of intense scientific research because understanding cellular pathways would allow researchers to devise efficient drugs that can influence a cell's behavior without causing unwanted side-effects. Proteomicists often use visualizations of interaction networks to understand changes in protein activation patterns measured in proteomic experiments. A distinct class of experiments is *knockout-experiments*: here researchers deactivate particular proteins, and compare protein activation levels before and after the removal. A more detailed description of protein interaction mechansisms, experimental techniques and visualization methods can be found in Chapter 3.

Our subjects were given network visualizations that were said to depict protein interactions documented in recent publications. Figure 5.1 shows one of three distinct networks that subjects were asked to analyze. The networks were manually created and laid out. The familiar Google Maps interface was used to display the network images and offer basic interaction. Clicking on nodes or edges opened information bubbles referring to these particular elements. Interactions were described by short, fictional paper abstracts detailing the particularities of each interaction and the context in which it was discovered.

Subjects were told that a knockout experiment had been performed on a specific type of cell. They were informed that a protein was removed from the cell and that researchers subsequently observed changes (positive or negative) in the levels of several proteins. These changes were marked on the network with arrows. Finally, subjects were asked to use the available information to determine network paths likely to have produced those changes, and to rank them based on their plausibility. This network task represents a visual, complex, and open-ended implementation of causal reasoning tasks which have been typical choices of cognitive studies [174].

Our networks used proteomic terminology but introduced fictional proteins, interactions and interaction mechanisms. Thus, the probability of a regulation chain was determined by the logical consistency of the presented evidence. The key rules that subjects were expected to extract from the evidence and use in their analysis were: the probability of a depicted interaction is lower if it was documented in species and cells other than those investigated in the knockout experiment; a correlation between two proteins should be treated as an edge with uncertain directionality; interactions could describe direct or inverse regulation mechanisms; and the edges sequence in a solution path should correctly explain the sign of the observed change. These assumptions, along with a general description of protein signaling, were illustrated in a 20 minute tutorial (text and video) and were clarified further on request. Moreover, essential terms were highlighted in all evidence text and in-situ explanations were displayed upon mouse clicks (Figure 5.1).

The order in which the three networks were presented to users was alternated to minimize the chance of network differences influencing the global result. Thus, in the test group, six subjects solved the networks in order $1, 2, 3$, seven subjects solved them as $2, 3, 1$, and the remaining six solved them as $3, 2, 1$. A similar division was used for the control group.

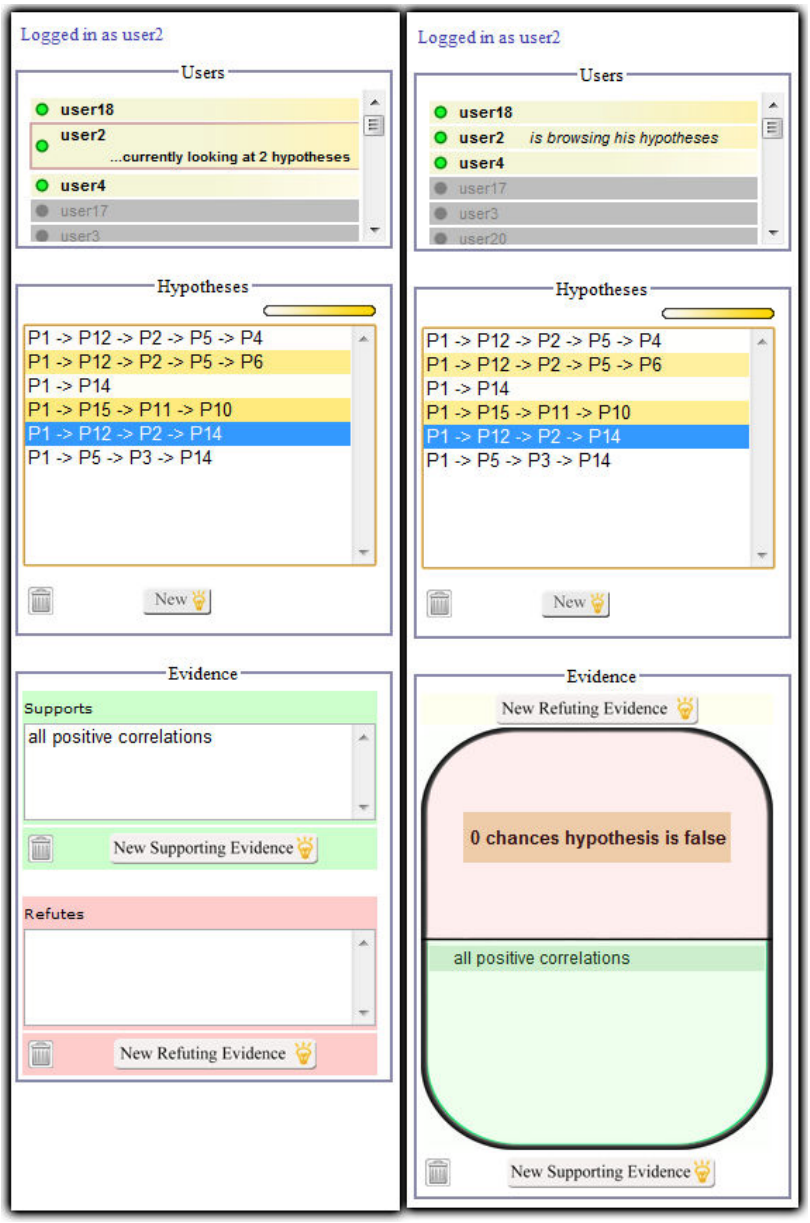
5.2.3 Analysis Interface and Evaluated Nudges

In addition to the protein network viewer, an analysis support interface augmented the experimental environment (Figure 5.1). As noted in Section 5.2.1, control subjects used the same base analysis interface in all three sessions. Test subjects started with an identical interface but were then exposed to upgraded versions in the second and third session. These versions, obtained by incrementally including two sets of evaluated nudges in the base interface, are shown in Figure 5.2.

The base analysis module contained three lists in which users could store their hypotheses, confirming, and disconfirming evidence. Hypotheses were entered into the system as noncyclical network paths by clicking on sequences of connected nodes. Evidence was inserted into either the confirming or disconfirming category by typing free text in a pop-up box. Selecting existing hypotheses would highlight their corresponding paths on the visualization and display their associated evidence, thus allowing subjects to revisit and compare hypotheses. Subjects were familiarized with these features in the tutorial video at the beginning of the study.

Three nudges were designed to alleviate three analytic deficiencies. First, we assumed that subjects would rely on their working memory rather than use the analysis system. Second, based on cognitive science studies, we assumed that subjects would have trouble considering multiple hypotheses in parallel. Third, we hypothesized that subjects would gather mostly confirming evidence for their hypotheses, and ignore the aspect of disconfirming evidence. Results of our initial session one runs caused us to adjust our last assumption: subjects were gathering approximately equal amounts of confirming and disconfirming evidence but in overall small amounts. We refined the design of our last nudge to better target this issue.

As already noted the **first evaluated nudge** (Figure 5.2, left) aimed to increase users' reliance on the analysis module. Our design rested on the assumption that if subjects knew other users were actively interacting with the module they would do so as well. To test this assumption, a section listing online users was added to the base analysis module. As users interacted with the module, this was reflected in a publicly visible status message (e.g., "user is browsing his hypotheses", "user has entered new evidence"). Fake user-bots were added to ensure that a nudge factor was at all times present. This design was inspired by research on conformity effects and motivational factors for



Logged in as user2

Users

user18

user2 ...currently looking at 2 hypotheses

user4

user17

user3

Hypotheses

P1 -> P12 -> P2 -> P5 -> P4

P1 -> P12 -> P2 -> P5 -> P6

P1 -> P14

P1 -> P15 -> P11 -> P10

P1 -> P12 -> P2 -> P14

P1 -> P5 -> P3 -> P14

Evidence

Supports

all positive correlations

Refutes

Logged in as user2

Users

user18

user2 *is browsing his hypotheses*

user4

user17

user3

user20

Hypotheses

P1 -> P12 -> P2 -> P5 -> P4

P1 -> P12 -> P2 -> P5 -> P6

P1 -> P14

P1 -> P15 -> P11 -> P10

P1 -> P12 -> P2 -> P14

P1 -> P5 -> P3 -> P14

Evidence

0 chances hypothesis is false

all positive correlations

Figure 5.2: The two modified analysis interfaces include three evaluated nudges: a box listing online users actively interacting with the analysis module (left), a color gradient (white to gold) shows recently analyzed hypotheses (left), and a redesigned, larger, evidence box asks users to commit to the implications of a hypotheses not having associated evidence (right).

online contributions. Specifically, humans change their behavior to match that of others [38, 14, 36] to gain social approval [49], or because they derive utility information from observing what others do [157, 101]. In addition, visibility and status recognition encourages users of social networks to increase their online contributions [10, 125, 125].

The **second nudge** was designed to encourage users to compare and contrast hypotheses in parallel rather than perform a sequential search in hypothesis space. Initially we planned to evaluate this nudge by itself but ultimately merged it with the first one to make the length of the study manageable. The design involved assigning each hypothesis a recency score that decayed over time but increased with any interactions targeting the hypothesis (e.g. selection, adding evidence). Recently active hypotheses were highlighted in the hypotheses list by using a color-gradient based on the recency score. Finally, thresholding the recency score allowed us to determine the number of a user's active hypotheses, display this information in the user status (Figure 5.2, left), and sort users based on how many hypotheses they were investigating. This offered a visual and status reward. While a user could trick the system by quickly switching between hypotheses, this was accounted for during the data analysis stage (see Results section) and we have observed just two intentional instances of it. These first two nudges were integrated into the analysis interface before the second session.

Finally, the **third nudge** was deployed before the last session and aimed to encourage test subjects to gather more evidence for the same number of hypotheses. To that end we modified the evidence collection part of the interface (Figure 5.2, right). First, the evidence collection area was more visually interesting and distinct from the rest of the interface. Second, if no confirming or disconfirming evidence had been entered for a hypothesis, the evidence boxes would read "0 chances that hypothesis is false" or "hypothesis is unlikely". This essentially required subjects to commit to extreme cases something that humans are known to avoid [52]. Third, an unintentional modification that we introduced while implementing the design was that the evidence boxes in this nudge were larger than in the base interface.

We hypothesize that this nudge could be restricted to disconfirming evidence only, in which case it could potentially alleviate confirmation biases [174]. As noted, our subjects did not exhibit a confirmation bias in the early stages of the study, so that we resorted to testing the more general case of increasing the amount of total evidence.

5.2.4 User Pool

Our study included a total of 36 subjects. Of these, 16 were women and 20 men. Six of them were young professionals, 18 were undergraduates, and 12 were graduate students. In terms of field or major, 26 of the subjects were active or majoring in sciences and engineering, while 10 were humanities students. None of the subjects had previous experience with proteomic analysis. As such, all subjects relied solely on the tutorial provided at the beginning of the study.

Subjects were randomly distributed in control (15) and test (21) groups such that the two groups had similar distributions of gender and age (undergraduate, graduate or postgraduate). Subjects were compensated for their participation.

5.2.5 User Study Limitations

Ease of hypothesis elicitation: A pilot run showed us that free-text specification of hypotheses would have lead to considerable variability in what users entered as hypotheses. To be able to compare results across subjects we limited hypotheses to paths of connected proteins. This interaction mode, reinforced by the tutorial video, gave subjects an easy "recipe" for generating hypotheses: any network path represented a valid hypothesis.

Lack of motivation: Our study did not involve monetary incentives to encourage subjects to provide valid solutions. As a result, several subjects appeared not to devote significant effort in searching for clues beyond those immediately noticeable.

Unforeseen problem solving strategies: A few of our early subjects copied the network on paper and annotated each interaction and protein. This strategy is not scalable to real protein interaction networks and it does not capture the exploratory nature of analysis. To avoid this we instructed the rest of the subjects to not use such exhaustive analysis strategies.

Varying degree of task difficulty: One of the three networks was perceived by several users as less difficult than the other two. We anticipated this problem and alternated the order in which tasks were presented to subjects.

Task misunderstanding: Instead of constructing short paths that linked the knockout protein to each changing protein, two subjects looked for long paths that linked the knocked-out protein and all arrow-proteins (i.e., proteins with changed levels) together. We used these results because the subjects used this interpretation consistently in all three sessions.

Visualization too limited: Several subjects expressed the need for a more feature-loaded visualization and two of them changed their analysis strategies between sessions to accommodate their requirements. Specifically, the two subjects realized that one interaction can be in multiple hypotheses. As a result, one used pen and paper to do her analysis at interaction level while one added single interactions as hypotheses and then assembled those into higher level paths. For one of the subjects we discarded the last dataset, while for the other we interpreted the data such as to reconcile the two strategies.

Analysis times varied: We urged users to spend approximately 60 minutes on each session. Several subjects however insisted on finishing earlier. Moreover, some datasets showed prolonged intervals of inactivity and several users were observed to take web-browsing or texting breaks. In our analysis we eliminated intervals with no activity and normalized all measurements by the time spent on the task.

Low number of subjects: Our sample size was relatively low for the open ended tasks our study involved. However, we note that the trends in the data became apparent with as few as six users in each group and changed very little throughout the experiment.

Effect of change is not captured: Our study does not capture the amount by which interface changes amplify the saliency of our nudges. It may well be that nudges are less observable and effective if they are introduced into the first system release.

5.3 Results

Here we describe quantitative and qualitative results from our user study. All data and analyses are available online [183] and have also been published in [93].

5.3.1 Data Preparation and Analysis

Thirty-two subjects completed all three sessions while four completed only the first two for a total of $28*3+4*2 = 104$ datasets. Four of the subjects, two from each group, solved the tasks on paper using exhaustive annotation of the networks. Three additional users also switched to this approach in the final session. All these data were discarded from the analyses leaving $104-4*3-3*1 = 89$ datasets from 13 control subjects and 19 test subjects.

We measured and analyzed three quantitative indicators to support our nudging hypothesis. First, we recorded the number of hypotheses and evidence entered into the system as a proxy for the degree to which subjects' relied on the interface to trace their analysis. This number was normalized by the time, in minutes, subjects spent on each session. Second, we measured the number of times a subject switched between hypotheses and normalized it by the number of hypotheses, as an indicator of the degree to which hypotheses were analyzed in parallel during analysis. Third, we recorded the number of evidence-items collected and divided it by number of hypotheses.

In the case of hypotheses switches we ignored selections lasting less than 5 seconds because we observed that users sometimes cycled rapidly through hypotheses as a method of gaging progress. We also ignored switches occurring in the last part of the analysis while subjects were filling in the answer questionnaire. We found that by default most users did a comparative analysis of hypotheses at the very end. Our nudge however was designed to encourage users constantly to consider alternative explanations.

In a second phase we also performed a qualitative analysis of our subjects' workflows. Our goals were to understand the dominant analytic strategies and behavioral patterns, and to verify the degree to which biases and heuristics applied.

5.3.2 Quantitative Support for Nudging Hypothesis

The premise of our experiment was that interface nudges would cause test subjects to change their behavior between sessions differently from how control subjects' behavior would evolve naturally as a consequence of learning or boredom. Figures 5.3-5.5 demonstrate the validity of our premise by contrasting the relative changes in performance measures between consecutive sessions in both experimental groups. As expected, change was negligible in control subjects (means of all triangles are close to one), but was significant for test subjects when a nudge was present (means of black squares higher than one). However, test group behavior remained constant whenever performance measures were not specifically targeted (e.g. change in contributions between the last two sessions). This suggests that subjects were not simply responding to interface changes but to nudges targeting particular performance measures.

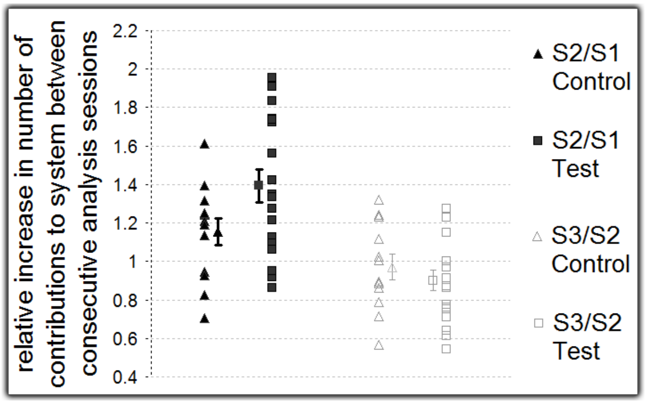


Figure 5.3: Changes between the first two sessions (black) caused test subjects (square) to increase the number of hypotheses and evidence items entered into the analysis system by an additional 24% over the control subjects'(triangle) relative increase. The interface changes made before the third session did not have a significant impact on this performance measure (grey).

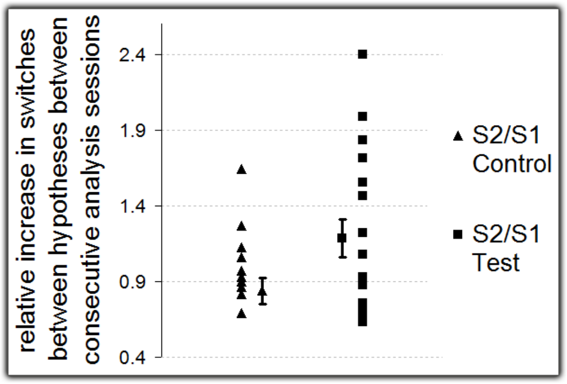


Figure 5.4: Changes between the first two sessions caused test subjects (square) to increase their switching between hypotheses by an additional 35% over the control subjects'(triangle) relative increase.

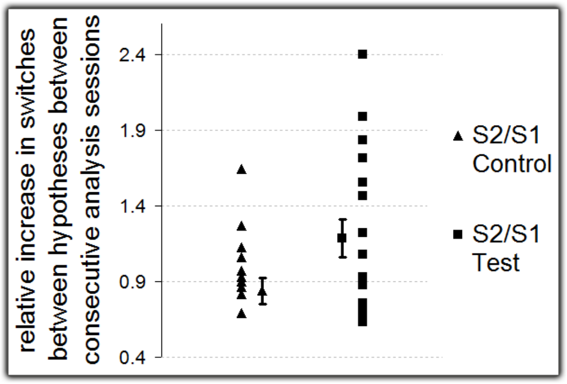


Figure 5.5: Changes between the last two sessions (black) caused test subjects (square) to gather 24% more evidence for their hypotheses as opposed to a constant evidence/hypotheses ratio (-2%) between all consecutive control sessions (triangle). Changes in the test group before the second session (gray) produced nonsignficant changes in the evidence collection as compared to the control group.

Test subjects contributed 39% more hypotheses and evidence items to the analysis module in the second session than in the first. This compares to an increase of only 15% in the control group (Figure 5.3). A t-test found this difference to be statistically significant ($t(29) = -2.07, p = 0.02$). Contributions remained close to constant between the second and third sessions in both the control and the test group (Figure 5.3). This conforms to the expected behavior since no nudge targeting contributions was added between these sessions.

The difference in switches between hypotheses was an increase of 18% in test subjects versus a decline of 17% in control subjects (Figure 5.4). The difference was significant, as indicated by the t-test ($t(25) = -1.89, p = 0.03$). The first two nudges were both added before the second session. Thus, we cannot assign either of the observed changes to any single nudge but to all interface changes made between the first two sessions.

The amount of evidence collected per hypothesis remained fairly constant between sessions in the control group with a decrease of 2% (Figure 5.5). Test subjects however, gathered on average 24% more evidence per hypothesis in the third condition than the second. This difference was also found to be statistically significant ($t(38) = -2.28, p = 0.01$).

5.3.3 Qualitative Analysis of Subjects' Workflows

Our subject's logs allowed us to qualitatively assess their workflows to extract common strategies and to determine the extent to which subjects rely on analytic biases and heuristics. The following paragraphs summarize our conclusions.

Observed workflows: More than half our subjects started with an initial exploration of the network. This exploration was not hypothesis driven and typically lasted between three and six minutes. Subjects then moved on to a hypothesis driven analysis, trying to connect "arrow" proteins to the knockout protein (Figure 5.1). We could discern two strategies for entering hypotheses. Most subjects would pick a candidate path, do a pre-evaluation of its likelihood, enter it into the system provided it was plausible, and then follow with a second pass to summarize and document evidence. These users would often revisit hypotheses and compare them. A few subjects added hypotheses without prior exploration and then summarized evidence in a following pass. Generally, they did not reevaluate those hypotheses again until a final pass when they decided on a global likelihood ordering.

Observed biases and heuristics: We also analyzed the data in terms of biases and heuristics documented in cognitive science literature. In particular we looked for confirmation bias, *single-attribute analysis* (i.e., focusing on a single most prominent attribute and using that to rank options), *conjunction fallacies* (i.e a specific condition is deemed more likely than an ecompassing general one), and inabillity to operate with varying degree of probabilities. Our findings are interesting because they show that some of these effects are not as dominant in a close to real analysis setting as cognitive science suggests and because they describe potential scenarios that trigger such behavior.

A first interesting finding was that confirmation bias is not dominant. In fact, subjects gathered slightly more disconfirming evidence than confirming evidence. Moreover, several users gathered almost exclusively disconfirming evidence, while others pruned paths that had strong negative evidence. Also, one subject would copy entire sections from the information bubbles and enter it directly as confirming evidence, but would always carefully summarize negative evidence. This suggests that she recognized the higher diagnostic value that the disconfirming evidence would have in her final ranking.

A known heuristic that we found in several datasets was *single-attribute analysis* (i.e., focusing on a single most prominent attribute and using that as a means of ranking options). We noticed several cases in which subjects added complicated paths before shorter, more intuitive ones. On a closer inspection we found that they had selected a single attribute (e.g. cell type) and were using it to include or discard paths from their analysis.

We also noticed an inability to operate with varying degrees of probability. Several subjects seemed to postpone the consideration of paths involving a complex probability judgment (e.g. multiple interactions with associated uncertainty) and instead concentrated on paths that allowed a binary decision.

Our network setup was well suited to discovering *conjunction fallacies* which occur when a specific condition is deemed more likely than a general one. In our network task short paths should be more likely candidates for analysis than longer paths. In general, our subjects seemed to be aware of this principle. In fact most new hypotheses abided by this rule. Additionally, several subjects added the short length of a path as positive evidence. However, we noticed that subjects' analytic strategies tricked them into the conjunction fallacy in a significant number of cases. We observed three main scenarios leading to this.

First, the favored method of expanding ones set of hypotheses was to modify an existing one by rerouting part of its path. At the very least subjects would use interactions that they were already familiar with. Most subjects avoided picking completely new routes, especially in network areas where they had already done some analysis. Such small changes to initially short paths lead subjects to analyze increasingly longer paths. Ultimately, subjects spent considerable time on long paths that were less likely than unexplored shorter options.

Second, subjects occasionally considered longer paths that linked multiple arrow-proteins together more likely than short paths from the knockout protein to each of those arrow- proteins. We hypothesize that users were looking for good unifying stories, a known cognitive tendency. Interestingly, one of the subjects confessed that he was aware of the conjunction fallacy but that the "story was too good" to be irrelevant.

Network layout: The third reason for multiple instances of conjunction fallacy is tied to the network layouts. The way paths were displayed visually had a significant impact on which ones were chosen for analysis. The majority of subjects preferred paths that described fairly continuous visual arches, or that were symmetric with ones they had already looked at. Sharp-angled paths were usually selected last even if they were shorter than already analyzed hypotheses. Another interesting effect observable in several datasets was that symmetrical paths were more often compared to each other than to other hypotheses.

Interestingly, the detrimental effect that the network layout had on our subjects' elicitation of new hypotheses is amplified if we consider that those hypotheses will be further expanded as noted above.

5.4 Discussion

In this section we discuss the broader impact of our contributions, alternative methodologies, and open questions.

5.4.1 Significance

Some of the findings reported in this chapter may seem unsurprising. That interface design can alter analytic workflows is evident, as is the fact that online visibility is correlated with increased online activity [125]. However, our study data doesn't only show that changes in interfaces translate into different user behavior. Our contribution lies in demonstrating that interface elements can be leveraged in controlled ways to unobtrusively correct users' strategies: our subjects' deficiency in supporting their hypotheses with evidence was observed in the first session and alleviated by a redesigned analysis support module in the third session. We believe this approach is valuable because it has the potential to correct and improve users' strategies without having to rely on coercive or obtrusive elements such as pop-up messages or help-agents.

5.4.2 Applicability

The analytic biases and heuristics targeted in our study were chosen because they are amply documented in the cognitive science literature. It is likely that one or more of these effects do not manifest or are beneficial in some areas or settings. In fact *naturalistic decision making* [105], a distinct research area, models situations (e.g., crisis control, time-sensitive operations) in which heuristics are an efficient analytic strategy.

The aim of this study was not to eliminate a specific set of biases and heuristics but to demonstrate that once such effects are identified we can use interface elements to reduce their occurrence. For example, we posit that in settings typically modeled by naturalistic decision making, heuristics are part of the rational analytic model. As such, excessively deliberative and time-consuming analysis could be considered erroneus or suboptimal and discouraged through the use of nudges.

5.4.3 Design Guidelines

Our work was primarily aimed at providing experimental support for applying the nudge paradigm in the visual analytics domain rather than providing a set of design guidelines. The nudge design space warrants a more exhaustive exploration because it can either provide a tool for guiding users towards better analytic strategies or help us understand how our interfaces unintentionally shape users' exploratory and analytic patterns. Our work exposed us to interesting questions about the ability and degree to which tutorials, ways of entering and storing hypotheses, and even simple design choices such as text-area size and color can influence users' behavior.

A few loose design guidelines can be distilled from our work however. First, placing collaborative elements and conformity triggers in analysis systems can nudge users to change their behavior.

We hypothesize that artificial "model-analysts", like we used in our experiment, could nudge users towards conforming to a desired behavior. Second, visual rewards, such as our recency score, will encourage users to consider options in parallel if this is desirable. Third, messages in text-areas, perhaps in conjunction with box-size, may be employed for boxes that should not be left empty. Finally, based on the qualitative analysis of our subjects' workflows we hypothesize that ways of automatically suggesting hypotheses may alleviate some of the observed conjunction fallacies and that subjects would benefit from support for multiple attribute analysis. Both such mechanisms would need to be domain specific and are beyond the scope of our work.

5.4.4 General Considerations

The data distributions may suggest that nudges, rather than uniformly targeting all subjects, tend to be particularly effective for a subgroup and less so for the rest. As seen in Figures 5.3-5.5, measurements obtained from test subjects appear to form two clusters: one with values similar to those measured in the control group, and one with distinctively higher values. These clusters do not correlate with the order in which networks were presented to users. However, the data gathered as part of this study is insufficient to test this hypothesis.

Our study did not replicate several biases and heuristics documented in the cognitive science literature. Most notably, humans are thought to be unable to elicit many hypotheses and to be biased towards gathering predominantly confirming evidence. Conversely, our subjects generated many hypotheses and showed no confirmation bias. We see two possible explanations for this. First, two of the study limitations may be responsible: the ease of generating hypotheses and the lack of subjects' motivation lead them to pursue multiple hypotheses and not develop attachments to favored ones. An alternative explanation is that humans are able to switch from a normal working mode to an analysis mode in which normative principles are more carefully observed. Research by Dunbar [55] hints at this hypothesis.

This latter possibility supports our choice of analysis task. Shorter and more focused tasks like the ones used in many cognitive experiments can be applied to large numbers of users and provide clean data. It is not clear, however, to what extent they translate to the exploratory analysis typical of scientific discoveries. As noted in the related work section several science studies indicate that there are observable differences between laboratory settings and real scientific or clinical situations.

Similarly, our study might have been more informative had we tested domain experts in their field of research rather than naive users on unfamiliar tasks. It remains uncertain whether domain experts, who generally follow well established workflows, can be nudged as easily as our subjects. Moreover, a high familiarity with an analysis system may also cause expert subjects to overlook new interface nudges.

Unfortunately, domain experts are scarce and the variability in the scientific problems they solve is high. Thus, quantitative studies that faithfully replicate real life scientific settings are improbable. Our choice of task and users implements a realistic approximation that provides insight into how to minimize the impact of biases and heuristics in scientific workflows. This endeavor is important because, as described in the beginning of the chapter, domain experts are not immune from cognitive biases and heuristics and often benefit from normative analysis strategies.

5.5 Concluding Remarks

We presented results from a quantitative user study demonstrating that controlled changes in the interface of an analysis system can be employed to correct potential deficiencies in users' analytic behavior. Specifically, we manipulated the design of a basic visual analysis tool over a set of three analysis sessions to produce three changes in our subjects' analysis. First, subjects were nudged to increase their reliance on the analysis support module which accompanied the visualization. Second, subjects were nudged to analyze hypotheses in parallel rather than sequentially. Third, subjects were nudged to gather more evidence for their hypotheses. The results of our user study led us to conclude that once deficient analytic behavior is identified in a scientific workflow supported by visual interfaces, it is likely that those interfaces can be redesigned to correct that behaviour.

The significance of our work is three-fold. First, we give an account of how even the simplest design decisions shape users' analytic behavior. Second, we advance visual analytics efforts by introducing and validating an approach that leverages visualization environments to correct analytic biases and heuristics reported by cognitive science literature. Third, we provided a short overview of analysis workflows, and biases and heuristics that our subjects used on a scientifically inspired analysis task.

Chapter 6

Discussion and Conclusion

This dissertation exemplifies how visualization supports data driven scientific discovery from data representation, through exploration and understanding to hypothesis elicitation and testing. It captures the interplay between domain specific contributions, designed and evaluated through close collaborations with domain experts, and wide ranging contributions that extend to multiple fields, are inspired by general theories and are evaluated through rigorous user studies.

This concluding chapter starts by reiterating the concrete contributions of this dissertation and by emphasizing its impact. It continues with a few discussion points pertaining to the dissertation as a whole and several future research directions it directly inspires. It ends with a brief summary of the work.

6.1 Contributions

The contributions of this dissertation are improvements through novel visualization techniques in neuroscience, proteomic and genomic data analysis, a novel visualization distribution mode, as well as a quantitative study of how interface design can be used to "nudge" scientists towards more efficient and correct analytic practices.

In neuroscience a novel interaction paradigm is introduced: 3D stream-tube models of white matter in the brain are linked to two-dimensional abstractions defined from the same data. In particular, a novel two-dimensional representation of white matter tractography that has the desirable properties of low-dimensional representations while preserving anatomically meaningful coordinates was developed. A concrete visualization system for analyzing white matter tractograms has been built. Accessibility to DTI visualizations for browsing purposes is enhanced by using a novel dissemination mode based on the Google Maps API. Data gathered from a formal and an anecdotal evaluation demonstrates the benefits of these approaches.

In proteomics design guidelines for visualizing protein interaction networks and experimental proteomic data are presented and evaluated with domain experts. Drawing protein networks by scaffolding publicly available protein interactions onto stylized pathway drawings enhances proteomicists' interpretation of the network data. Exploring protein networks at multiple levels of detail using focus+context techniques supports proteomic workflows. Combining publicly available protein interaction networks with new experimental data accelerates the discovery process.

In genomics we show that disseminating data as precomputed visualizations using the Google Maps API can efficiently support many analysis tasks and reduce or eliminate several overheads. Examples of specific visualizations of micro-array data and their evaluation with domain specialists are presented. We also show that this method is potentially domain independent. Further contributions include design elements, challenges and opportunities when working with pre-computed visualizations and the Google Maps API.

Finally, following the visual analytics path, a quantitative user study shows how scientific analysis can be improved in terms of hypothesis correctness and closeness to normative analysis guidelines by variations in a system's interface design. We posit that this approach may facilitate the use of visual analytics expertise to correct biases and heuristics documented in the cognitive science community.

6.2 Impact and Generality of this Dissertation

Our three domain specific contribution areas provide immediate design guidelines for building visualization systems that support the work of scientists in those fields. Moreover, these contributions are visualization and computer science contributions in their own right because they fuel further innovation within computer science and because they are, at least to some extent, generalizable to other application areas. Finally, our nudging contribution is independent of a particular domain and even of visualization itself. It can therefore impact any computer driven analysis and thus contribute to better research in a wide range of domains.

We are confident that the contributions presented in Chapters 2-4 can directly impact neuroscientists, genomic and proteomic researchers, help them understand the inner workings of the human body and mind, and devise better drugs and treatments that will improve our quality of life. Our confidence is rooted in the fact that these contributions are a product of collaborative, interdisciplinary design. They are motivated by analysis shortcomings in neuroscience, proteomics and genomics identified by domain experts in those fields. Their design and evaluation were performed with the assistance of those ultimately benefitting those methods.

Additionally, these contributions can fuel the inspiration of other visualization researchers to build upon our results. Such work already exists [128, 146, 106] but we believe our results have potential to further impact the design of visualization methods for domain driven network exploration, neural circuitry analysis, or genomic data browsing.

Also, such domain driven visualization contributions directly benefit the domains that motivated them but often generalize to other application areas as well. Existing work that builds on our approaches to create visualization solutions in domains such as archeology [110] or seismology [62] vouches for that. Our application of the Google maps paradigm to neuroscience, though originally motivated by genomics, is another example. Interacting with white matter tractograms via simplified, two-dimensional proxies can be applied to any stream-line visualization such as those created for fluid-flow datasets. Finally, many of the benefits of collating new experimental data with known network information that is projected in meaningful spaces, as described in Chapter 3, are likely to be transferable to neural circuitry analysis (see section 6.4.2).

Moreover, it is important to note that the generality of contributions is often tied to their specificity. In the previous paragraph we gave a few examples of specific paradigms that are almost directly transferable to other domains. However, high level contributions are likely to be applicable to a wider range of application areas, albeit with additional refactoring work. Examples of such contributions from this dissertation are: interacting with complex data types through alternate, simpler proxies that abstract certain aspects of the data; using linked views as a path towards understanding complex data sets; tightly coupling visualizations of many relevant data-sources together into a unitary system (e.g. new experimental data, known information, publications etc.); or distributing raw data along with readily analyzable views of it. While such contributions do not provide a step-by-step recipe for building a system in a new domain, they do provide overarching design principles that can guide visualization developers.

Finally, the "nudging" paradigm, applied to the visual analytics domain, represents a novel and viable solution for bridging the gap between descriptive and normative analysis using visual analytics methodologies. The need of supporting human analysis against analytic biases and heuristics had been recognized as one of the major objectives of visual analytics [73, 158]. The results presented here support the hypotheses that interface changes can be leveraged to guide users towards normative analytic support. Beyond its purely scientific significance, this finding can help catalyze research into supporting analytic deliberation through the use of interfaces and visualization. Specifically, it can provide a foundation for future visual analytics research to determine biases and heuristics that would benefit from "nudging", perhaps across multiple domain areas, efficient interfaces nudges, and evaluations of their effects. At a high level, this contribution is independent of a particular domain and even of visualization itself and can therefore impact any computer driven analysis and contribute to better research in a wide range of domains.

6.3 Discussion Items

This section presents a few discussion points pertaining to the dissertation as a whole. More detailed items for each of the four specific topics presented in this dissertation can be found at the end of Chapters 2-4.

Visualization can be thought of as a box of tools and building materials. Visualization researchers use these tools to build analysis systems that help researchers understand and hypothesize about their data faster, with less effort, or more correctly. Many visualization systems, although assembled using the same tools and materials, are unique, novel and useful through their design and the problems they solve. The work we presented exemplifies how general building blocks are shared across multiple contributions areas: brushing and linking in multiple views helps both neuroscientists and proteomicists, low dimensional representations are applicable for both genomic data and 3D neurological datasets, while the use of a digital map framework can be useful in the biology realm as well as in neuroscience.

While each chapter can be thought of as an independent unit, covering its own research agenda, methodology and contributions, the dissertation is unified by its attempt to use visualization to help scientists perform their analysis more accurately and more efficiently. It seizes on the opportunity to demonstrate the benefits of visualization innovation in three concrete areas but at the same time proposes a quantitative approach to improve scientific analysis workflows that is independent of any particular visualization and domain.

The contributions of the presented work cover a continuum of specificity to generality. Chapters 2 and 3 target specific analysis workflows and tasks: white matter tractography and tract-bundle selection, and analysis of proteomic experimental data in the context of available protein interaction information. Chapter 3 shows that representing data as pre-computed digital maps is desirable in a range of analysis tasks, and, while inspired by genomics, is not limited to a particular scientific field. Finally, the quantitative study on analysis "nudging" is domain independent and demonstrates how interface elements can be used to guide users towards better analysis regardless of the visualizations they use.

The dissertation combines qualitative and quantitative evaluation to measure the effectiveness of novel techniques or design principles that are introduced. Quantitative evaluations, as part of controlled user studies, are generally desirable as means of evaluation because they show numerical proof of a new method's efficiency, and quantify the performance gain at the same time. However, many of the methods presented in this dissertation address domain experts and complex, targeted scientific tasks. In such cases it is often hard to find enough users to achieve statistical relevance and to design tasks that are simple enough to be measured and reproduced yet complex enough to represent a meaningful scientific analysis unit. Anecdotal evaluations with domain experts are convenient for iterative development and tight collaborative settings and will generally offer a good approximation of user preferences and performances. For example, the finding that proteomic researchers are deterred from analysis by unstructured network representations has been determined through anecdotal evaluation, and has been independently verified by two concurrent studies. Similarly, our finding that planar abstractions attached to 3D white matter tractograms accelerated tract selection was also confirmed and quantified by concurrent research.

A unifying principle throughout this document is Brooks' "computer scientists as toolsmiths" paradigm. Each of the contributions presented in this dissertation is a direct result of collaborative work with domain experts from neuroscience, proteomics, genomics and cognitive science. Using researchers' real problems to drive visualization innovation, while often laborious for both domain experts and computer science researchers, will ultimately identify where computers can help most and benefit all parties.

6.4 Open research opportunities

In line with the collaborative nature of visualization, future research opportunities in the few specific domains presented here should be identified by continuing the collaborative discovery processes with researchers in neuroscience, proteomics and genomics. A few immediately foreseeable opportunities, however, are inspired by the interplay between the different components of this dissertation and by feedback and observations gathered as part of our work.

6.4.1 Data Infrastructure for Distributing, Analyzing and Cross-Referencing Neurological Data

A tighter integration of neurological datasets, along with appropriate visualization and querying capabilities, into both clinical and research neuroscientific communities is desirable yet missing. Genomic and proteomic data, results, and publications are distributed across a wide range of websites, databases and systems that provide querying, visualization, processing and analysis algorithms for biological data. In general, such data-sources are tightly cross-referenced to each other. This distributed knowledge system allows small research communities to build on top of data infrastructures created and maintained by bigger research laboratories and institutions. It also allows researchers to easily relate their new data and hypotheses to existing and evolving knowledge.

Such a data infrastructure is lacking in neuroscience. Datasets are sparsely available, can usually be only accessed as raw data files, and cross-referencing is nearly inexistent. We believe part of this landscape is a consequence of the inherently 3D data and visualizations which are not easily translatable to the text-based web and querying paradigms. Our mapping approach is a first step in the direction of web-accessible neurological datasets. Additional work is necessary however, to create web-deployed tools and visualizations that are more robust, dynamic and suited for querying and analyzing neurological data. We believe visualization technologies fueled by this application area could then be applied to a plethora of domains relying on the analysis of three dimensional data.

6.4.2 Creating Tools for Analyzing Neurological Networks

Our results on protein interaction pathways inspire work on network analysis tools for neurological data. The ensemble of neurons in the human brain essentially forms networks that connect different regions of the brain at different scales and complexities. Just as protein pathways, such neural networks are associated with different cognitive functions, all worth studying and understanding.

Advances in data acquisition techniques (DTI, fluorescent microscopy) provide new and efficient ways of mapping neural connectivity in the brain, at scales that span both inter- and intra-brain regions. Efforts to document such connections in open databases have started to emerge. This creates opportunities similar to the ones already tackled in proteomics: developing novel neural network visualizations that collate data from multiple sources: connectivity databases, experimental data, and meta data (e.g., publications, anatomical atlases).

The particularities of the neuroscience domain render this problem interesting and challenging at the same time. First, unlike protein interaction networks, neurological circuitry resides in an anatomical space that is highly relevant for neuroscientists. Moreover, this space is three-dimensional and thus poses representational and interaction challenges. Second, neurological networks can be analyzed at different scales: connections that link major brain regions, connections within those regions, or individual neurons. Third, anatomical particularities of each individual brain, especially in diseased cases, influence the integrity and strength of brain connectivity.

6.4.3 Automatically Suggesting Viable Hypotheses in Protein Pathway Analysis

The qualitative analysis in our "nudging" study (Chapter 5) revealed that our subjects exhibited cognitive biases while exploring our protein networks and identifying potential hypotheses. Moreover, protein interaction networks are dense, complex and becoming increasingly so. Thus, simply immersing scientists into network visualizations, even under the "guidance" of an experimental dataset, may not be the most effective way towards new discoveries.

An alternative is using bioninformatic algorithms to automatically suggest sets of likely hypotheses to researchers. We posit that even the simplest of methods, shortest-paths computation for example, would perhaps accelerate the analysis process. More advanced analytic methods, however, such as Petri-Net or Baysian modeling, are likely to complement human analytic abilities and to constrain the analysis space to only the most viable hypotheses. In general we believe that a tight integration between human abilities, be it cognitive or perceptual, and computer specific strengths is the likely path to optimal analysis.

6.4.4 Cognitive and Domain Driven Analysis Tools and Visualizations

An immediate extension required before the "nudging" paradigm presented in Chapter 5 can gain practical applicability, is in quantifying the nudging potential of a set of common interface design elements and constructions. We imagine one or more studies that would use short decision or deliberation tasks to quantify the nudging abilities of interface characteristics, and, in general, of elements that are part of the distribution and presentation of an analysis tool. Examples of tested items could include widget type, placement, color, size and styling, alert and help messages, or tutorials. A careful design would allow us to crowd-source such user studies [103] thus making them manageable, extensible and sufficiently representative. Such studies could yield readily applicable design guidelines for practitioners to use in the development of new user interfaces.

A second extension required for "nudges" to be used in the ways envisioned in this dissertation, is in determining desired and undesired behavior. While in our current work we considered general biases and heuristics, as described by the cognitive science community, we believe that domain specific particularities can determine what constitute optimal or deficient decision strategies. This view is subscribed to the more general goal of understanding how to adjust analysis tools and visualizations to match the cognitive particularities of specific domains, domain categories, and analysis situations.

To exemplify, *naturalistic decision making* (NDM) [105] represents a branch of cognitive science that models cognitive processes that occur in time-sensitive situations and suggests effective decision making strategies in such environments (e.g., disaster response, emergency medics, military operations). While a few visual analytics efforts have already prototyped disaster response applications, the main supported scenario involved collaborative and distributed decision making in which field agents equipped with small mobile displays communicate with an operational headquarter. As such, contributions were generally limited to extreme resolution visualization (e.g., small for field operatives, large for HQ), and supporting distributed and collaborative analysis (e.g., co-located around a table-top display, distributed between operatives).

We find a different, yet interesting approach is to tailor analysis methods based on the cognitive particularities, constraints and limitations of individuals placed in NDM situations. For example, the nudging paradigm as described in this dissertation would have to be adapted for a NDM setting: nudging may need to encourage fast, heuristic decision making rather than highly deliberative analysis. Moreover, certain interface design patters, interaction techniques or analytic workflows may overload NDM users. A wide range of cognitive principles applicable for time-sensitive decision making should thus be matched to appropriate analytic tools.

Such thinking could be applied to visualization as well. Visual attributes and cues, aesthetic and design principles, and even entire visualization methods that are effective in ordinary analysis settings might be unsuited for decision-making that is quick, heuristical and driven by limited cognitive bandwidth. A potential hypothesis is that sparsity, aggregation, emphasis on strong visual cues, and a limitation of subtle cues, data attributes and dimensions might encourage fast decision making and benefit NDM settings even if representing the data less faithfully. Another hypothesis is that some visualization methods require more cognitive bandwidth than others, thus making them less suited for quick analysis. A few concrete questions that could make the object of such research spring to mind. Are heatmap representations more suited for quick decision making than parallel coordinates? Is plotting a small number of the most representative datapoints in a dataset plot less visually inhibiting than showing the entire data? Is a heatmap using a discrete and limited color range easier to interpret than one using a wide and continuous color palette? While binary answers to some of these questions may seem trivial, quantifying differences in perceived complexity, and whether they translate into an inhibition to make a decision, would allow visual designers to create systems that fit the constraints of specific application areas.

In line with the visualization researcher as toolsmith paradigm, we envision developing and evaluating such visualization and analysis principles by coupling rigorous experimentation with strong collaborations in other domains, and development of concrete analysis applications scaffolded on our findings. Ultimately, this work can lead to analysis systems that are better at supporting and complementing human perceptual and cognitive abilities given specific analytic contexts.

6.5 Summary

The work presented in this dissertation enables researchers in three specific scientific areas, neuroscience, proteomics and genomics, to do better analysis in less time. It also lays the foundation for the development and quantitative evaluation of user interface elements that can unobtrusively guide scientists towards more efficient and correct analysis workflows, regardless of their fields of research.

The methods introduced have been developed following the traditional visualization methodology, which aims to improve the way data is visualized and interacted with, as well as the visual analytics methodology which aims to improve the analysis process itself.

Chapter 2 shows how traditional 3D white matter brain models can be linked to planar representations, some of which novel, to accelerate typical interaction tasks and improve data understanding in neuroscience. Chapter 3 shows how publicly available protein interaction information and proteomic experimental data can be combined visually to answer scientific question in ways that harness the researchers' intuition and support their workflows. Chapter 4 introduces a novel way of disseminating genomic data and shows how it can facilitate or accelerate data browsing, lightweight analysis and data dissemination. Chapter 5 embodies the visual analytics approach by combining the "nudge" paradigm and elements from persuasive technology with a quantitative user study to show how individual user interface design elements can guide users towards more correct analysis behavior.

Bibliography

[1] Circos. `http://mkweb.bcgsc.ca/circos/`.

[2] Cutting-edge tools for expression analysis. `www.silicongenetics.com`.

[3] Decision site for functional genomics. `http://www.Spotfire.com`.

[4] Immgen project. Website. `http://www.immgen.org/`.

[5] Ingenuity. `http://www.ingenuity.com/`.

[6] A.T. Adai, S.V. Date, S. Wieland, and E.M. Marcotte. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *Journal of Molecular Biology*, 340(1):179–190, 2004.

[7] D. Akers, A. Sherbondy, R. Mackenzie, R. Dougherty, and B. Wandell. Exploration of the brain's white matter pathways with dynamic queries. In *Proc. of Visualization*, pages 377–384, 2004.

[8] David Akers. Wizard of Oz for participatory design: Inventing an interface for 3d selection of neural pathway estimates. In *Proceedings of CHI 2006 Extended Abstracts*, pages 454–459, 2006.

[9] H. Alt and M. Godau. Computing the Frechet distance between two polygonal curves. *International Journal of Computational Geometry and Applications*, 5(1):75–91, 1995.

[10] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, page 980. ACM, 2007.

[11] K. Arakawa, S. Tamaki, N. Kono, N. Kido, K. Ikegami, R. Ogawa, and M. Tomita. Genome Projector: zoomable genome map with multiple views. *BMC bioinformatics*, 10(1):31, 2009.

[12] G. Aravindhan, G.R. Kumar, R.S. Kumar, and K. Subha. AJAX Interface: A Breakthrough in Bioinformatics Web Applications.

[13] E. Arroyo, L. Bonanni, and T. Selker. Waterbot: exploring feedback and persuasive techniques at the sink. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, page 639. ACM, 2005.

[14] S.E. Asch. Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological monographs*, 70(9):1–70, 1956.

[15] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber. *Multivariate Analysemethoden: Eine anwendungsorientierte Einfuhrung.* Springer, 2005.

[16] M.Q.W. Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119. ACM New York, NY, USA, 2000.

[17] A. Barsky, T. Munzner, J. Gardy, and R. Kincaid. Cerebral: visualizing multiple experimental conditions on a graph with biological context. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1253–1260, 2008.

[18] Peter J. Basser, James Mattiello, and Denis LeBihan. Estimation of the effective self-diffusion tensor from the nmr spin echo. *J Magn Reson B*, 103(3):247–254, March 1994.

[19] P.J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, and A. Aldroubi. In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, 44(4):625–632, 2000.

[20] G. Ben-Shakhar, M. Bar-Hillel, Y. Bilu, and G. Shefler. Seek and ye shall find: Test results are what you hypothesize they are. *Journal of Behavioral Decision Making*, 11(4):235–249, 1998.

[21] S.I. Berger, R. Iyengar, and A. Ma'ayan. AVIS: AJAX viewer of interactive signaling networks. *Bioinformatics*, 23(20):2803, 2007.

[22] E. Bier, E. Ishak, and E. Chi. Entity Workspace: an evidence file that aids memory, inference, and reading. *Intelligence and Security Informatics*, pages 466–472, 2006.

[23] J.W. Bodnar. Making sense of massive data by hypothesis testing. In *International Conference on Intelligence Analysis*, pages 2–4.

[24] M. Bostock and J. Heer. Protovis: A Graphical Toolkit for Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1121–1128, 2009.

[25] D. Botstein and K.W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10(8):2703–2734, 1999.

[26] N. Boukhelifa, JC Roberts, and PJ Rodgers. A coordination model for exploratory multiview visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2003. Proceedings. International Conference on*, pages 76–85, 2003.

[27] F.P. Brooks Jr. The computer scientist as toolsmith II. *Communications of the ACM*, 39(3):61–68, 1996.

[28] A. Brun, H.J. Park, H. Knutsson, and C.F. Westin. Coloring of DT-MRI fiber traces using Laplacian eigenmaps. *Lecture Notes in Computer Science*, pages 518–529, 2003.

[29] J.S. Bruner and M.C. Potter. Interference in visual recognition. *Science*, 144(3617):424–425, 1964.

[30] A. Buja, JA McDonald, J. Michalak, W. Stuetzle, and M. Bellcore. Interactive data visualization using focusing and linking. In *IEEE Conference on Visualization, 1991. Visualization'91, Proceedings.*, pages 156–163, 1991.

[31] A.J. Cañas, R. Carff, G. Hill, M. Carvalho, M. Arguedas, T.C. Eskridge, J. Lott, and R. Carvajal. Concept maps: Integrating knowledge and information visualization. *Lecture Notes in Computer Science*, 3426:205, 2005.

[32] S.K. Card, J.D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[33] M. Catani, R.J. Howard, S. Pajevic, and D.K. Jones. Virtual in vivo interactive dissection of white matter fasciculi in the human brain. *Neuroimage*, 17(1):77–94, 2002.

[34] M. Chalmers. A linear iteration time layout algorithm for visualising high-dimensional data. In *Proceedings of the 7th conference on Visualization'96*. IEEE Computer Society Press Los Alamitos, CA, USA, 1996.

[35] L.J. Chapman and J.P. Chapman. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74(3):271–280, 1969.

[36] T.L. Chartrand and J.A. Bargh. The chameleon effect: The perception-behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893–910, 1999.

[37] Wei Chen, Ziang Ding, Song Zhang, Anna MacKay-Brandt, Stephen Correia, Huamin Qu, John Allen Crow, David F. Tate, Zhicheng Yan, and Qunsheng Peng. A novel interface for interactive exploration of dti fibers. *IEEE TVCG (Proc. of Visualization)*, 2009.

[38] R.B. Cialdini and N.J. Goldstein. Social influence: Compliance and conformity. 2004.

[39] A. Clark and D. Chalmers. The extended mind. *Analysis*, 58(1):7, 1998.

[40] Andy Cockburn and Bruce McKenzie. Evaluating the effectiveness of spatial memory in 2d and 3d physical and virtual environments. In *CHI'02*, pages 203–210, 2002.

[41] I. Corouge, S. Gouttard, and G. Gerig. Towards a shape model of white matter fiber bundles using diffusion tensor MRI. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004*, pages 344–347, 2004.

[42] S. Correia, J.A. Crow, D.F. Tate, and Q. Peng. Wei Chen Member, IEEE Zi'ang Ding Song Zhang Member, IEEE Anna MacKay-Brandt. *IEEE Transactions on Visualization and Computer Graphics*, 15(06).

[43] C.M. Danis, F.B. Viegas, and M. Wattenberg. Your place or mine?: visualization as a community component. In *Proceedings of CHI*, 2008.

[44] U.N. Danner, H. Aarts, and N.K. de Vries. Habit formation and multiple means to goal attainment: Repeated retrieval of target means causes inhibited access to competitors. *Personality and Social Psychology Bulletin*, 33(10):1367, 2007.

[45] A.C.E. Darling, B. Mau, F.R. Blattner, and N.T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394, 2004.

[46] R. Davidson and D. Harel. Drawing graphs nicely using simulated annealing. *ACM Transactions on Graphics (TOG)*, 15(4):301–331, 1996.

[47] E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, G. Nisanci, R. Cetin-Atalay, and M. Ozturk. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 18(7):996, 2002.

[48] Cagatay Demiralp and David H. Laidlaw. Similarity coloring of dti fiber tracts. In *Proceedings of DMFC Workshop at MICCAI*, 2009.

[49] M. Deutsch and H.B. Gerard. A study of normative and informational social influences upon individual judgment. *Journal of abnormal and social psychology*, 51(3):629–636, 1955.

[50] G.W. Dickson, G. DeSanctis, and D.J. McBride. Understanding the effectiveness of computer graphics for decision support: a cumulative experimental approach. *Communications of the ACM*, 29(1):47, 1986.

[51] Z. Ding, J.C. Gore, and A.W. Anderson. Classification and quantification of neuronal fiber pathways using diffusion tensor MRI. *Magnetic Resonance in Medicine*, 49(4):716–721, 2003.

[52] W.M. DuCharme. Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, 85(1):66–74, 1970.

[53] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley-Interscience Publication, 2nd edition, 2000.

[54] K. Dunbar. Concept discovery in a scientific domain*. *Cognitive Science*, 17(3):397–434, 1993.

[55] K. Dunbar. What scientific thinking reveals about the nature of cognition. *Designing for science: Implications from everyday, classroom, and professional settings*, pages 115–140, 2001.

[56] T. Dwyer, Y. Koren, and K. Marriott. IPSep-CoLa: An incremental procedure for separation constraint layout of graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):821–828, 2006.

[57] P. Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42(149160):194–202, 1984.

[58] P. Eades and CFX De Mendonca. Vertex splitting and tension-free layout. *Lecture Notes in Computer Science*, pages 202–211, 1995.

[59] R. Eccles, T. Kapler, R. Harper, and W. Wright. Stories in geotime. *Information Visualization*, 7(1):3–17, 2008.

[60] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

[61] A.S. Elstein and A. Schwarz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *Stroke*, 33:493–6, 2002.

[62] C. Engelsma and D. Hale. Visualization of 3d tensor fields derived from seismic images.

[63] W.R. Fisher. Narration as a human communication paradigm. *Contemporary rhetorical theory: A reader*, page 265, 1999.

[64] BJ Fogg. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 225–232. ACM Press/Addison-Wesley Publishing Co., 1998.

[65] A. Frick, A. Ludwig, and H. Mehldau. A fast adaptive layout algorithm for undirected graphs. In *Proceedings of the DIMACS International Workshop on Graph Drawing*, pages 388–403. Springer-Verlag London, UK, 1994.

[66] Y. Frishman and A. Tal. Online dynamic graph drawing. *IEEE Transactions on Visualization and Computer Graphics*, pages 727–740, 2008.

[67] T.M.J. Fruchterman and E.M. Reingold. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21(11):1129–1164, 1991.

[68] T.M.J. Fruchterman, E.M. Reingold, Dept. of Computer Science, and University of Illinois at Urbana-Champaign. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[69] GW Furnas. Generalized fisheye views. *ACM SIGCHI Bulletin*, 17(4):23, 1986.

[70] Steven Gomez, Radu Jianu, and David H. Laidlaw. A fiducial-based tangible user interface for white matter tractography. In *Proceedings of ISVC 2010*, 2010.

[71] D. Gotz, M.X. Zhou, and V. Aggarwal. Interactive visual synthesis of analytic knowledge. In *Proceedings of the IEEE Symposium on Visual Analytics Science & Technology*, pages 51–58, 2006.

[72] Henry Gray. *Anatomy of the Human Body.* Lea & Febiger, 1918.

[73] T.M. Green, W. Ribarsky, and B. Fisher. Building and applying a human cognition model for visual analytics. *Information Visualization*, 8(1):1–13, 2009.

[74] DL Gresh, BE Rogowitz, RL Winslow, DF Scollan, and CK Yung. WEAVE: A system for visually linking 3-D and statistical visualizations, applied to cardiac simulation and measurement data. In *Proceedings of the conference on Visualization'00*, pages 489–492. IEEE Computer Society Press Los Alamitos, CA, USA, 2000.

[75] B. Gretarsson, S. Bostandjiev, J. ODonovan, and T. Hollerer. WiGis: A Framework for Scalable Web-based Interactive Graph Visualizations.

[76] R. Hastie and R.M. Dawes. Rational choice in an uncertain world. *Journal of the Indian Academy of Applied Psychology*, page 107, 2003.

[77] M. Hegarty. Dynamic visualizations and learning: Getting to the difficult questions. *Learning and Instruction*, 14(3):343–352, 2004.

[78] N. Henry, A. Bezerianos, and J.D. Fekete. Improving the Readability of Clustered Social Networks using Node Duplication. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1317–1324, 2008.

[79] R.J. Heuer. *Psychology of intelligence analysis.* United States Govt Printing Office, 1999.

[80] H. Hochheiser, E.H. Baehrecke, S.M. Mount, and B. Shneiderman. Dynamic querying for pattern identification in microarray and genomic data. In *Proceedings of IEEE International conference on Multimedia and Expo*, volume 3, pages 453–456. Citeseer, 2003.

[81] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):741–748, 2006.

[82] Z. Hu, J. Mellor, J. Wu, T. Yamada, D. Holloway, and C. DeLisi. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic acids research*, 33(Web Server Issue):W352, 2005.

[83] Visual Analytics Inc. Website. `http://www.visualanalytics.com`.

[84] P. Isenberg, A. Tang, and S. Carpendale. An exploratory study of visual information analysis. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1217–1226. ACM, 2008.

[85] R. Jianu, C. Demiralp, and D. Laidlaw. Exploring 3d dti fiber tracts with linked 2d representations. *IEEE TVCG (Proc. of Visualization)*, 15(6):1449–1456, 2009.

[86] R. Jianu, K. Yu, L. Cao, V. Nguyen, A.R. Salomon, and D.H. Laidlaw. Visual integration of quantitative proteomic data, pathways and protein interactions. *IEEE Transactions on Visualization and Computer Graphics*.

[87] Radu Jianu, Cagatay Demiralp, and David H. Laidlaw. Exploring brain connectivity with two-dimensional neural maps. In *IEEE Visualization 2010 Poster Compendium*, 2010.

[88] Radu Jianu, Cagatay Demiralp, and David H. Laidlaw. Exploring the brain connectivitywith two-dimensional neuralmaps. *IEEE Transactions on Visualization and Computer Graphics*, 2010.

[89] Radu Jianu, Cagatay Demiralp, and David H. Laidlaw. Exploring the brain connectivitywith two-dimensional neuralmaps. IEEE Visualization Poster Compendium, 2010.

[90] Radu Jianu, Cagatay Demiralp, and David H. Laidlaw. Visualizing and exploring tractograms via two-dimensional connectivity maps. In *Proceedings of ISMRM'10*, 2010.

[91] Radu Jianu and David H. Laidlaw. Visualizing gene co-expression as google maps. In *ISVC Proceedings 2010*, 2010.

[92] Radu Jianu and David H. Laidlaw. Visualizing protein interaction networks as google maps. In *IEEE Visualization 2010 Poster Compendium*, 2010.

[93] Radu Jianu and David H. Laidlaw. An evaluation of how small user interface changes can improve scientists analytic strategies. In *Proceedings of SIGCHI (CHI) 2012*, 2011.

[94] Radu Jianu and David H. Laidlaw. Guiding visualization users towards improved analytic strategies using small interface changes. In *IEEE Visualization 2011 Poster Compendium*, 2011.

[95] Radu Jianu, David H. Laidlaw, and Arthur Salomon. Visualizing phosphorylation experiments data in the context of known protein interactions. IEEE Visualization Poster Compendium, 2006.

[96] D.W. Johnson and TJ Jankun-Kelly. A scalability study of web-native information visualization. In *Proceedings of graphics interface 2008*, pages 163–168. Canadian Information Processing Society Toronto, Ont., Canada, Canada, 2008.

[97] M. Jones. Thinker's Toolkit: 14 Powerful Techniques for Problem Solving, 1998.

[98] F. Jourdan and G. Melançon. Tool for metabolic and regulatory pathways visual analysis. In *Proceedings of SPIE*, volume 5009, page 46, 2003.

[99] L. Kaufman and P.J. Rousseeuw. Finding groups in data: an introduction to cluster analysis. *New York*, 1990.

[100] DA Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[101] D.T. Kenrick, J.K. Maner, J. Butner, N.P. Li, D.V. Becker, and M. Schaller. Dynamical evolutionary psychology: Mapping the domains of the new interactionist paradigm. *Personality and Social Psychology Review*, 6(4):347, 2002.

[102] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, et al. The human genome browser at UCSC. *Genome research*, 12(6):996, 2002.

[103] A. Kittur, E.H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456. ACM, 2008.

[104] J. Klayman and Y.W. Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211–228, 1987.

[105] G.A. Klein. A recognition-primed decision (rpd) model of rapid decision making. *Decision making in action: Models and methods*, pages 138–147, 1993.

[106] J. Klein, M. Scholl, A. Kohn, and H.K. Hahn. Real-time fiber selection using the wii remote. In *Proceedings of the SPIE*, volume 7625, 2010.

[107] H. Kuehn, A. Liberzon, M. Reich, and JP Mesirov. Using GenePattern for gene expression analysis. *Current protocols in bioinformatics/editoral board, Andreas D. Baxevanis...[et al.]*, 2008.

[108] M. Kumar and T. Kim. Dynamic speedometer: dashboard redesign to discourage drivers from speeding. In *CHI'05 extended abstracts on Human factors in computing systems*, page 1576. ACM, 2005.

[109] D. Lockton, D. Harrison, and N. Stanton. Design with intent: Persuasive technology in a wider context. *Persuasive Technology*, pages 274–278, 2008.

[110] A. Loomis and M. Watters. Multimodal volume visualization of geophysical data for archaeological analysis.

[111] C.G. Lord, L. Ross, and M.R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

[112] M. Maddah, A.U.J. Mewes, S. Haker, W.E.L. Grimson, and S.K. Warfield. Automated atlas-based clustering of white matter fiber tracts from DT-MRI. *Lecture Notes in Computer Science*, 3749:188, 2005.

[113] M. Meyer, T. Munzner, and H. Pfister. MizBee: A Multiscale Synteny Browser. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):897–904, 2009.

[114] G. Michal. On representation of metabolic pathways. *BioSystems*, 47(1-2):1–7, 1998.

[115] MindManager. Website. `http://www.mindjet.com`.

[116] Bart Moberts, Anna Vilanova, and Jarke J. van Wijk. Evaluation of fiber clustering methods for diffusion tensor imaging. In *Procs. of Vis'05*, pages 65–72, 2005.

[117] S. Mori and P.C.M. Van Zijl. Fiber tracking: principles and strategies-a technical review. *NMR in Biomedicine*, 15(7-8):468–480, 2002.

[118] S. Mori and P.C.M. van Zijl. Fiber tracking: principles and strategies ˇ2013 a technical review. *NMR in Biomedicine*, 15(7-8):468–480, 2002.

[119] A. Morrison and M. Chalmers. A pivot-based routine for improved parent-finding in hybrid MDS. *Information Visualization*, 3(2):109–122, 2004.

[120] T. Munzner. H3: Laying out large directed graphs in 3D hyperbolic space. In *IEEE Symposium on Information Visualization, 1997. Proceedings.*, pages 2–10, 1997.

[121] T. Munzner, F. Guimbretière, and G. Robertson. Constellation: a visualization tool for linguistic queries fromMindNet. In *1999 IEEE Symposium on Information Visualization, 1999.(Info Vis' 99) Proceedings*, pages 132–135, 1999.

[122] Vinh Nguyen, Lulu Cao, Jonathan Lin, Anna Ritz, Norris Hung, Radu Jianu, Benjamin Raphael, David H. Laidlaw, Laurent Brossay, and Arthur Salomon. A new approach for quantitative phosphoproteomic dissection of signaling pathways applied to T cell receptor activation. *Molecular and Cellular Proteomics*, 8(11):2418–2431, 2009.

[123] C.L. North, B. Shneiderman, and Human/Computer Interaction Laboratory. *A taxonomy of multiple window coordinations.* Human-Computer Interaction Laboratory, Institute for Advanced Computer Studies, 1997.

[124] A. Notebook. i2 Analyst Notebook. *i2 Ltd,¡ http://www. i2. co. uk/¿ Viewed at*, 31, 2007.

[125] O. Nov. What motivates wikipedians? *Communications of the ACM*, 50(11):64, 2007.

[126] H. Oinas-Kukkonen and M. Harjumaa. Towards deeper understanding of persuasion in software and information systems. In *First International Conference on Advances in Computer-Human Interaction*, pages 200–205. IEEE, 2008.

[127] R. Otten, A. Vilanova, and H. Van De Wetering. Illustrative White Matter Fiber Bundles. *Computer Graphics Forum*, 29(3):1013–1022, 2010.

[128] R. Otten, A. Vilanova, and H. Van De Wetering. Illustrative white matter fiber bundles. In *Computer Graphics Forum*, volume 29, pages 1013–1022. Wiley Online Library, 2010.

[129] T. Pattison and M. Phillips. View coordination architecture for information visualisation. In *Proceedings of the 2001 Asia-Pacific symposium on Information visualisation-Volume 9*, pages 165–169. Australian Computer Society, Inc. Darlinghurst, Australia, Australia, 2001.

[130] F.V. Paulovich and R. Minghim. HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE transactions on visualization and computer graphics*, 14(6):1229–1236, 2008.

[131] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, TKB Gandhi, M. Gronborg, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13(10):2363–2371, 2003.

[132] W.A. Pike, R. May, B. Baddeley, R. Riensche, J. Bruce, and K. Younkin. Scalable visual reasoning: supporting collaboration through distributed analysis. In *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*, pages 24–32, 2007.

[133] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 2005, pages 2–4, 2005.

[134] M.D. Plumlee and C. Ware. Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(2):179–209, 2006.

[135] Y. Qu and G.W. Furnas. Sources of structure in sensemaking. In *CHI'05 extended abstracts on Human factors in computing systems*, page 1992. ACM, 2005.

[136] EM Reingold and JS Tilford. Tidier drawings of trees. *IEEE Transactions on Software Engineering*, pages 223–228, 1981.

[137] G.G. Robertson, J.D. Mackinlay, and S.K. Card. Cone trees: animated 3D visualizations of hierarchical information. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 189–194. ACM New York, NY, USA, 1991.

[138] A.C. Robinson and G. Center. Collaborative synthesis of visual analytic results. In *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST'08*, pages 67–74, 2008.

[139] R. Rodgers and J.E. Hunter. The discard of study evidence by literature reviewers. *The Journal of Applied Behavioral Science*, 30(3):329, 1994.

[140] D.M. Russell, M.J. Stefik, P. Pirolli, and S.K. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276. ACM, 1993.

[141] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, 2005.

[142] M. Sarkar and M.H. Brown. Graphical fisheye views of graphs. *Communications of ACM*, 37(12):73–84, 1994.

[143] Debra MacIvor Savage, Eric N. Wiebe, and Hugh A. Devine. Performance of 2d versus 3d topographic representations for different task types. In *HFES Annual Meeting*, 2004.

[144] A. Savikhin, R. Maciejewski, and D.S. Ebert. Applied visual analytics for economic decision-making. In *IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST'08*, pages 107–114, 2008.

[145] C. Schmid and H. Hinterberger. Comparative multivariate visualization across conceptuallydifferent graphic displays. In *Scientific and Statistical Database Management, 1994. Proceedings., Seventh International Working Conference on*, pages 42–51, 1994.

[146] T. Schultz. Feature extraction for dw-mri visualization: The state of the art and beyond. In *Proc. Schloss Dagstuhl Scientific Visualization Workshop 2009*, 2010.

[147] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, pages 80–86, 2002.

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

[149] B. Shneiderman. Book Preview-Designing the User Interface: Strategies for Effective Human-Computer Interaction. *Interactions-New York*, 4(5):61, 1997.

[150] B. Shneiderman, S.K. Card, J.D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[151] H.A. Simon. Rationality as Process and as Product of Thought. *The American Economic Review*, 68(2):1–16, 1978.

[152] A.U. Sinha and J. Meller. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC bioinformatics*, 8(1):82, 2007.

[153] M.E. Skinner, A.V. Uzilov, L.D. Stein, C.J. Mungall, and I.H. Holmes. JBrowse: A next-generation genome browser. *Genome Research*, 19(9):1630, 2009.

[154] H.S. Smallman and M. Hegarty. Expertise, spatial ability and intuition in the use of complex visual displays. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, volume 51, pages 200–204. Human Factors and Ergonomics Society, 2007.

[155] R. Spence. *Information visualization.* Addison-Wesley Harlow, 2001.

[156] J. Stalker, B. Gibbins, P. Meidl, J. Smith, W. Spooner, H.R. Hotz, and A.V. Cox. The Ensembl Web site: mechanics of a genome browser. *Genome research*, 14(5):951, 2004.

[157] C. Stangor, G.B. Sechrist, and J.T. Jost. Social influence and intergroup beliefs: The role of perceived social consensus. *Social influence: Direct and indirect processes*, pages 235–252, 2001.

[158] J. Stasko, C. Gorg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

[159] C.R. Sunstein and R.H. Thaler. Libertarian paternalism is not an oxymoron. *U. Chi. L. Rev.*, 70:1159, 2003.

[160] DeRose. T., EA Bier, M. Stone, K. Pier, and W. Buxton. Toolglass and magic lenses: the see-through interface. In *Proceedings of SIGGRAPH*, volume 93, pages 73–80.

[161] E. Tejada, R. Minghim, and L.G. Nonato. On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization*, 2(4):218–231, 2003.

[162] S.T. Teoh and K.L. Ma. RINGS: A technique for visualizing large hierarchies. *Lecture Notes in Computer Science*, 2528:268–275, 2002.

[163] R.H. Thaler and S. Benartzi. Save More Tomorrow: using behavioral economics to increase employee saving. *Journal of political Economy*, pages 164–187, 2004.

[164] R.H. Thaler and C.R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness.* Yale Univ Pr, 2008.

[165] J.J. Thomas and K.A. Cook. Illuminating the path: The research and development agenda for visual analytics. *IEEE Computer Society*, 2005.

[166] D. Tunkelang. A practical approach to drawing undirected graphs, 1994.

[167] F. van Ham and A. Perer. Search, Show Context, Expand on Demand: Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):953–960, 2009.

[168] F.B. Viégas, M. Wattenberg, M. McKeon, F. Van Ham, and J. Kriss. Harry potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools. In *Proc. HICSS*, 2008.

[169] F.B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121, 2007.

[170] I. Viola, M.E. Groller, M. Hadwiger, K. Buhler, B. Preim, M.C. Sousa, D.S. Ebert, and D. Stredney. Illustrative visualization. In *IEEE Visualization*, page 124, 2005.

[171] C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(Database Issue):D433, 2005.

[172] S. Wakana, H. Jiang, L.M. Nagae-Poetscher, P.C.M. van Zijl, and S. Mori. Fiber Tract-based Atlas of Human White Matter Anatomy 1, 2004.

[173] MO Ward. XmdvTool: integrating multiple methods for visualizing multivariatedata. In *IEEE Conference on Visualization, 1994., Visualization'94, Proceedings.*, pages 326–333, 1994.

[174] P.C. Wason. Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20(3):273–281, 1968.

[175] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The Sandbox for analysis: concepts and methods. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 801–810. ACM New York, NY, USA, 2006.

[176] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303, 2002.

[177] D. Yang, E.A. Rundensteiner, and M.O. Ward. Nugget discovery in visual exploration environments by query consolidation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 603–612. ACM New York, NY, USA, 2007.

[178] D. Yang, Z. Xie, E.A. Rundensteiner, and M.O. Ward. Managing Discoveries in The Visual Analytics Process.

[179] T. Yates, M.J. Okoniewski, and C.J. Miller. X: Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Research*, 36(Database issue):D780, 2008.

[180] K. Yu, L. Cao, R. Jianu, R. Park, C. Gatsonis, D. Laidlaw, and A. Salomon. A software suite to expedite the study of cell signaling pathway: automated acquisition, organization and annotation. In *55 th ASMS Conference Proceedings*. American Society for Mass Spectrometry, 2007.

[181] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS letters*, 513(1):135–140, 2002.

[182] S. Zhang, C. Demiralp, and DH Laidlaw. Visualizing diffusion tensor MR images using streamtubes and streamsurfaces. *IEEE TVCG*, 9(4):454–462, 2003.

[183] Experimental data. `http://graphics.cs.brown.edu/research/sciviz/nudges/`.