

Biological Data Visualization

Edited by

Carsten Görg¹, Lawrence Hunter², Jessie Kennedy³,
Seán O'Donoghue⁴, and Jarke J. van Wijk⁵

- 1 University of Colorado, US, carsten.goerg@ucdenver.edu
- 2 University of Colorado, US, larry.hunter@ucdenver.edu
- 3 Napier University – Edinburgh, GB, j.kennedy@napier.ac.uk
- 4 CSIRO and the Garvan Institute of Medical Research, AU,
sean@odonoghuelab.org
- 5 Eindhoven University of Technology, NL, vanwijk@win.tue.nl

Abstract

The topic of visualizing biological data has recently seen growing interest. Visualization approaches can help researchers understand and analyze today's large and complex biological datasets. The aim of this seminar was to bring together biologists, bioinformaticians, and computer scientists to survey the current state of tools for visualizing biological data and to define a research agenda for developing the next generation of tools. During the seminar, the participants formed working groups on nine different topics, reflected on the ongoing research in those areas, and discussed how to address key challenges; six talks complemented the work in the break-out groups. This report documents the program and the outcome of Dagstuhl Seminar 12372 "Biological Data Visualization".

Seminar 09.–14. September, 2012 – www.dagstuhl.de/12372


1998 ACM Subject Classification H.5 Information Interfaces and Presentation, I.3 Computer Graphics, J.3 Biology and genetics

Keywords and phrases Information visualization, data visualization, biology, bioinformatics, user interfaces, visual analytics

Digital Object Identifier 10.4230/DagRep.2.9.131

1 Executive Summary

Carsten Görg
Lawrence Hunter
Jessie Kennedy
Seán O'Donoghue
Jarke J. van Wijk

License  Creative Commons BY-NC-ND 3.0 Unported license
© Carsten Görg, Lawrence Hunter, Jessie Kennedy, Seán O'Donoghue, and Jarke J. van Wijk

Introduction and Motivation

Biology is rapidly evolving into a 'big data' science, and as a consequence there is an urgent and growing need to improve the methods and tools used for gaining insight and understanding from biological data. Over the last two decades, the emerging fields of computational biology and bioinformatics have led to significant advances primarily in automated data analysis. Today, however, biologists increasingly deal with large, complex datasets (e.g., 'omics' data) where it is not known in advance what they are looking for and thus, automated analyses alone



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Biological Data Visualization, *Dagstuhl Reports*, Vol. 2, Issue 9, pp. 131–164

Editors: Carsten Görg, Lawrence Hunter, Jessie Kennedy, Seán O'Donoghue, and Jarke J. van Wijk



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

cannot solve their problems. Interactive visualizations that can facilitate exploratory data analysis and support biologists in creating new hypotheses lend themselves to complement automated analyses. Bioinformaticians already have built a variety of tools for visualizing different types of biological data and those tools are widely used in the community. So far, most bio-related visualization research has been conducted by people outside of the visualization community, people who have learned about visualization but are often not aware of research in the visualization community. Consequently, the current tools do not embody the latest advancements in design, usability, visualization principles, and evaluation.

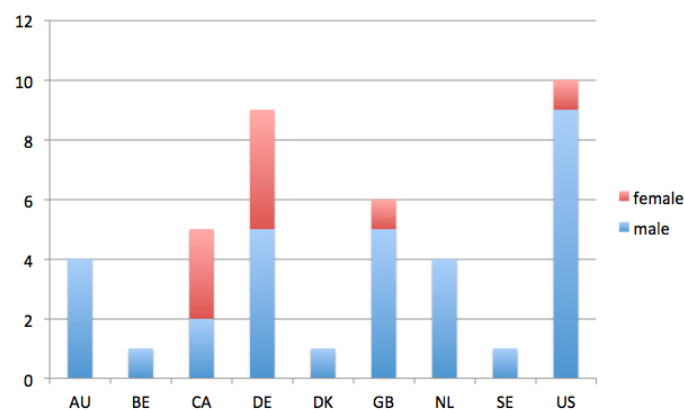
One main goal of this first Dagstuhl Seminar on Biological Data Visualization was to bring together the users (biologists), current visualization tool builders (bioinformaticians), and visualization researchers to survey the state-of-the-art of the current tools and define a research agenda for systematically developing the next generation of tools for visualizing biological data. Only a close collaboration of the researchers from all three communities can create the synergies necessary to address the challenges in analyzing and visualizing large and complex biological datasets.

Topics discussed during the seminar included:

- Challenges in visualizing biological data. Biological data is very heterogeneous. It contains spatial data, graphs, tabular data, and textual data. Challenges are wide spread: open-ended data quantity, open-ended exploratory tasks, long-term analyses, rich analytics, heterogeneous data, usability and evaluation of tools.
- Design and visualization principles, research in human-centered design, usability, and evaluation of interactive data-analysis tools.
- Creating a common research agenda and a common understanding of the problem field of biological data visualization.
- Integration of multiple visualizations for different data types and tasks into one tool to support more complex analysis scenarios.
- Designing an infrastructure for next generation visualization tools.
- Establishing collaborations between computer scientists and biologists.

Participants and Program

41 researchers from 9 countries participated in this seminar. Many participants came from the US and from Germany, others came from Canada, Australia, and a number of other European countries (see Figure 1). There was a good mix of researchers from the visualization, bioinformatics, and biology communities. About a third of the participants attended their first seminar at Dagstuhl.



■ **Figure 1** Participant statistics of the seminar.

Monday	Tuesday	Wednesday	Thursday	Friday
Introduction Personal Ads	Talk Reporting Session	Talk Reporting Session	Talk Breakout Groups	Reporting Session
Discussion: Topics for Breakout Groups	Breakout Groups	Discussion: Topics for new Breakout Groups	Breakout Groups	Discussion: BioVis Community
Talk Breakout Groups	Talk Breakout Groups	Excursion to Trier	Talk Breakout Groups	
Breakout Groups	Breakout Groups		Breakout Groups	

■ **Table 1** Final schedule of the seminar. The breakout groups on Monday/Tuesday discussed topics on Ontologies in Biological Data Visualization, Comparative Analysis of Heterogeneous Networks, Sequence Data Visualization, and Bridging Structural & Systems Biology; the breakout groups on Wednesday/Thursday discussed topics on Uncertainty Visualization, Infrastructure, Multiscale Visualization, Effective Visualization Design, and Evaluation.

Table 1 provides an overview of the final seminar schedule. The program was designed to facilitate in-depth discussions in small working groups. To get to know each other—the seminar brought together researchers from different communities—participants introduced themselves and their research interests with a ‘personal ad’ in the Monday morning session. This was a great way to set the tone for informal and engaging discussions during the seminar.

Previous to the seminar, the organizers collected interesting ideas and suggestions from the participants for possible topics for working groups. To allow participants to work on different topics and with different people, the topics and groups changed halfway through the seminar. On Monday morning and Wednesday morning all participants discussed and refined the suggested topics and formed groups according to their interests. The groups (four on Monday/Tuesday and five on Wednesday/Thursday) worked in parallel on their topics and reported regularly on their progress. The work in the breakout groups was complemented by a discussion on the BioVis Community on Friday and a number of talks given throughout the seminar:

- *Seán I. O'Donoghue*: BioVis Introduction: A Practitioner's Viewpoint
- *Daniel Evanko*: Visualization on nature.com
- *Matt Ward*: Biovisualization Education: What Should Students Know?
- *Arthur J. Olson*: The Promise and Challenge of Tangible Molecular Interfaces
- *Martin Krzywinski*: visualization – communicating, clearly
- *Bang Wong*: Concepts gleaned from disparate communities

These talks, presented to all participants in the morning sessions and after the lunch breaks, intentionally touched on broad and high-level topics to make them more interesting to the diverse audience in the seminar. The abstracts of the talks are presented in Section 3.

Discussion and Outcome

Some of the working groups followed a classical design process [6, 8] to structure their collaborative work. They split their discussions into a *problem phase* and a *solution phase*. Both phases featured divergent and convergent stages: *discover* and *define* for the problem phase and *develop* and *deliver* for the solution phase. Francis Rowland, a seminar participant with expertise in user experience design, facilitated these discussions.

Figure 2 shows some artifacts produced by the Ontologies in Biological Data Visualization working group that followed this design process. The *Four C's* approach (left) is an example for the discover and design stages. The group broke down their topic into four aspects: *Components* (parts), *Characters* (people involved), *Challenges*, and *Characteristics* (features and behavior). The *Four C's* approach helped the group to provide a holistic view on the design problem and to better define the topic. The *Draw the Box* approach (right) is an example for the develop and deliver stages. Members of the group collaboratively imagined an end product of their work that would be sold in a box on a shelf and designed its package. This approach helped the group members to gather ideas, visualize the outcome, and focus on the most important features of the product.



■ **Figure 2** Examples of design processes: the *Four C's* approach (left) and the *Draw the Box* approach (right).

The diverse outcomes from the nine working groups are summarized below. The detailed reports are presented in Section 4.

Comparative Analysis of Heterogeneous Networks: The analysis of the transcriptome produces a large number of putatively disrupted transcripts, and prioritizing which disruptions are most likely to be meaningful (causal or diagnostic) is a time-consuming process. To guide their interpretation researchers create heterogeneous networks by integrating information from a wide variety of annotation databases. The working group investigated how the analysis of the transcriptome can be facilitated by interactive visualizations of transcriptome assemblies and proposed a method to infer the functional consequence of a transcript's disruption based on the local structure of the annotation networks. A tight coupling of network analysis algorithms and interactive visualizations, specifically designed to support these analysis tasks, could accelerate identification of important transcript alterations.

Sequence Data Visualization: Genome-associated data is growing at a fast rate and genome browsers are still the tool of choice for integrating and analyzing different types

of data in one single representation. The working group analyzed the different challenges of visualizing genome-associated data and separated them into two different dimensions: problems associated with rearrangements of the genomic coordinates and problems with the abundance of data at each genomic position. To address these problems, the group discussed and developed a number of possible solutions, including the development of a reference-free gene-centric approach, compressing tracks by aggregation or summarization, and using meta-data or data itself as a novel way for selecting tracks. These approaches can lay the foundation for the development of new visualization tools.

Bridging Structural & Systems Biology via DataVis: There exist several gaps between the field of structural biology, which has yielded detailed insight into the molecular machines of life, and the field of systems biology, which has evolved more recently in the wake of the genomics revolution, but separately from the advances of the more structural view of biology. The integration of both fields and their visualization tools could create new tool sets to enhance the exploration and understanding of biological systems. The working group analyzed and described the existing gaps and proposed seven strategies to facilitate collaboration and professional advancement in structural biology, systems biology, and data visualization.

Ontologies in Biological Data Visualization: Ontologies are graph-based knowledge representations in which nodes represent concepts and edges represent relationships between concepts. They are widely used in biology and biomedical research, for the most part as computational models, in computational analyses, and for text mining approaches. The working group examined the potential impact of ontologies on biological data visualization. The group identified challenges and opportunities from the perspectives of three different stakeholders: ontologists (who create and maintain ontologies), data curators (who use ontologies for annotation purposes), and data analysts (who use ontologies through applications to analyze experimental data). Identified challenges include the dynamic nature of ontologies, scalability, how to utilize the complex set of relationships expressed in ontologies, and how to make ontologies more useful for data analysis. Identified research opportunities include the visualization of ontologies themselves, automated generation of visualization using ontologies, and the visualization of ontological context to support search. The group submitted a Viewpoints article on Ontologies in Biological Data Visualization to the IEEE Computer Graphics & Applications journal.

A Framework for Effective Visualization Design: Visualizations are not only an important aspect of how scientists make sense of their data, but also how they communicate their findings. The techniques and guidelines that govern how to design effective visualizations, however, can be quite different whether the goal is to explore or to explain. Unfortunately, scientists are often not aware of the spectrum of considerations when creating visualizations. To help clarify this problem, the working group has developed a framework to reason about the spectrum and considerations to help scientists better match their visualization goals with appropriate design considerations.

Uncertainty Visualization: Uncertainty is common in all areas of science, and it poses a difficult problem for visualization research. Visualization of uncertainty has received much attention in the areas of scientific visualization and geographic visualization; however, it appears much less common in information visualization and in biological data visualization. The working group analyzed and described the sources of uncertainty and types of uncertainty specific to biology. Uncertainty visualization in networks was identified as an open issue, including uncertainty in the network topology and uncertainty in attributes on nodes, edges, and their interdependencies. The group started a survey of the literature on uncertainty

visualization for biological data and proposed to construct a taxonomy of uncertainty visualization approaches, and investigate how they could be employed in the context of a collection of biological problems.

Evaluation: The working group identified two central problems with respect to the evaluation of tools for visualizing biological data: (1) How to motivate biologists to participate in evaluations? and (2) How to evaluate the tools? The answer to the first question was (simply) that biologists have to benefit from the evaluation to be motivated to participate, e.g. they might get a tool they can use to solve their problems. The second question was more complex and the working group discerned a number of dimensions, centered around what, why, when, where, and how. The discussion of these dimensions lead to the insight that there is a strong difference between approaches taken by designers working at a bio-institute and approaches taken by infovis researchers. Both approaches have merit, the challenge is to close the gap and combine them.

Multiscale Visualization: Biology involves data and models at a wide range of scales and researchers routinely examine phenomena and explore data at multiple scales. Visual representations of multi-scale datasets are powerful tools that can support data analysis and exploration, however, visualizing multi-scale datasets is challenging and not many approaches exist. The working group identified four common dimensions of biological multi-scale datasets: 3D space, time, data complexity (modality), and data volume (size). The group produced a short video to introduce each dimension independently in order to provide a quick and understandable view on the nature of the different scales and how they apply to biological data and exploration. Additionally, the group discussed in more detail a number of biological multi-scale data and models that can be visualized across multiple dimensions and introduced case studies to highlight issues like navigation, interaction, and human-computer interfaces. Carsten Görg presented a talk on the results from this working group at the 2012 Rocky Mountain Bioinformatics Conference.

Infrastructure: The working group discussed needs from both a technical and community standpoint regarding the challenges involved in the analysis of biomedical data and mechanisms to facilitate interactions between visualization communities in computer science and biology. Eight key criteria were identified: interoperability, reusability, compatibility, references & benchmarks, middleware, vertical integration, scalability, and sustainability. The group developed a model for a community-maintained, biological visualization resource that would enable biological questions, task descriptions, sample datasets and existing tools for the problems to be disseminated to the computational visualization and biological research communities. Additionally, the group developed a detailed use-case based on the data and analysis pipelines of the cancer genome atlas that will allow technical aspects of the eight key criteria to be explored and practical solutions proposed.

Finally, based on feedback from the participants (from the seminar questionnaire as well as from personal communication with the organizers) another important outcome of the seminar was to establish collaborations between computer scientists and biologists. The academic cultures in biology and computer science, including publication models, are quite different. In addition, biologists have a different mindset than computer scientists: biologists often work in a detail-oriented manner whereas computer scientists often seek to generalize. Understanding each other's culture is important for successful collaborations and the Dagstuhl seminar provided a unique setting to meet enthusiastic people from different communities, have long group discussions with a focus on problem solving, and form synergies with researchers that have a different outlook and expertise.

2 Table of Contents

Executive Summary

<i>Carsten Görg, Lawrence Hunter, Jessie Kennedy, Seán O'Donoghue, and Jarke J. van Wijk</i>	131
--	-----

Overview of Talks

BioVis Introduction: A Practitioner's Viewpoint <i>Seán I. O'Donoghue</i>	139
Visualization on nature.com <i>Daniel Evanko</i>	139
Biovisualization Education: What Should Students Know? <i>Matt Ward</i>	139
The Promise and Challenge of Tangible Molecular Interfaces <i>Arthur J. Olson</i>	140
visualization – communicating, clearly <i>Martin Krzywinski</i>	141
Concepts gleaned from disparate communities <i>Bang Wong</i>	141

Working Groups


Comparative Analysis of Heterogeneous Networks <i>Andreas Kerren, Corinna Vehlou, Jessie Kennedy, Karsten Klein, Kasper Dinkla, Michel Westenberg, Miriah Meyer, Mark Ragan, Martin Graham, Martin Krzywinski, and Tom Freeman</i>	142
Sequence Data Visualization <i>Jan Aerts, Jean-Fred Fontaine, Michael Lappe, Raghu Machiraju, Cydney Nielsen, Andrea Schafferhans, Svenja Simon, Matt Ward, and Jarke J. van Wijk</i>	143
Bridging Structural & Systems Biology via DataVis <i>Graham Johnson, Julian Heinrich, Torsten Möller, Seán O'Donoghue, Art Olson, James Procter, and Christian Stolte</i>	148
Ontologies in Biological Data Visualization <i>Sheelagh Carpendale, Min Chen, Daniel Evanko, Nils Gehlenborg, Carsten Görg, Lawrence Hunter, Francis Rowland, Margaret-Anne Storey, Hendrik Strobelt</i>	151
A Framework for Effective Visualization Design <i>Miriah Meyer, Jan Aerts, Dan Evanko, Jean-Fred Fontaine, Martin Krzywinski, Raghu Machiraju, Kay Nieselt, Jos Roerdink, and Bang Wong</i>	152
Uncertainty Visualization <i>Min Chen, Julian Heinrich, Jessie Kennedy, Andreas Kerren, Falk Schreiber, Svenia Simon, Christian Stolte, Corinna Vehlou, Michel Westenberg, and Bang Wong</i>	154
Evaluation <i>Jarke J. van Wijk, Kasper Dinkla, Martin Graham, Graham Johnson, Francis Rowland, and Andrea Schafferhaus</i>	155

Multiscale Visualization	
<i>Carsten Görg, Graham Johnson, Karsten Klein, Oliver Kohlbacher, Thorsten Möller, Arthur Olson, Francis Rowland, and Matt Ward</i>	156
Infrastructure	
<i>Seán O'Donoghue, Tom Freeman, Mark Ragan, Margaret Storey, Larry Hunter, Cydney Nielsen, Nils Gehlenborg, Jim Procter, and Hendrik Strobelt</i>	159
Acknowledgements	162
Participants	164

3 Overview of Talks

3.1 BioVis Introduction: A Practitioner's Viewpoint


Seán I. O'Donoghue (CSIRO and the Garvan Institute of Medical Research, AU)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Seán I. O'Donoghue

Experimental methods in biological research are delivering data of rapidly increasing volume and complexity. However, many current methods and tools used to visualize and analyse these data are inadequate, and urgent improvements are needed if life scientists are to gain insight from this data deluge, rather than being overwhelmed. I will discuss a recent switch in focus away from algorithmic bioinformatics towards data visualization and usability principles, illustrating how such a focus can have significant impact, illustrating these points with examples from work on macromolecular structures, systems biology, and literature mining. I will also discuss a recent, international community initiative that brings visualization experts together with computational biologists, bioinformatics, graphic designers, animators, and medical illustrators, and aims to raise the global standard of bioinformatics software (<http://vizbi.org/>).

3.2 Visualization on nature.com


Daniel Evanko (Nature Publishing Group, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Daniel Evanko

Nature very recently published results of the Encyclopedia of DNA Elements (ENCODE) project. To aid the discoverability of information in these manuscripts Nature Publishing Group developed the ENCODE Explorer and threaded presentations of the results to allow targeted reading of single selected topics through all 36 manuscripts of the project. We also created Javascript-based interactive figures with the intention of further developing and reusing these visualizations elsewhere. As a further aid to information discoverability, technical editors are beginning to annotate all gene, protein and chemical entities in original research papers published in a limited number of Nature research journals. We hope to make this information accessible through APIs.

3.3 Biovisualization Education: What Should Students Know?

Matt Ward (Worcester Polytechnic Institute, US)


License  Creative Commons BY-NC-ND 3.0 Unported license
© Matt Ward

In recent years, the level of activity in the area of visualization of biological data has greatly increased, both in terms of users and developers. An important question is what sort of training should be provided for students in this area? Can we use existing courses in biology and data/information visualization, or do we need one or more courses that fuse these two distinct fields? In this talk I describe my experiences in designing and delivering a course

on biovisualization to upper level undergraduate and graduate students majoring in either computer science or bioinformatics and computational biology. The project oriented course covers both basic principles of information visualization as well as the data models typically found in biology – sequences, networks, tabular data, and spatial structures. For each data type I describe the common analysis tasks performed on the data as well as a variety of visual mappings that can be applied. I also describe standard rules for effective visualization design and common methods for evaluating the resulting visualizations. I summarize my observations on the best and weakest aspects of the course and welcome feedback from the seminar attendees on ways to improve the course.

3.4 The Promise and Challenge of Tangible Molecular Interfaces

Arthur J. Olson (The Scripps Research Institute – La Jolla, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Arthur J. Olson

Structural molecular biology is a key science in connecting the worlds of physics and chemistry to biology. It is a discipline that focuses on three and four-dimensional relationships of complex shapes and functions. As such, it has been a fertile proving ground for novel technologies that can enhance interaction and visualization of such systems for the purposes of exploration, understanding and communication.


Physical models have been used for centuries to aid in the process of modeling and visualization in many areas of science. In the latter part of the last century computer graphics largely superseded physical models for these purposes. This advance in technology was accompanied by a loss of the perceptual richness inherent in the human interaction with real physical objects. The tactile and proprioceptive senses provide key cues to our ability to understand 3 dimensional form and to perform physical manipulations, but are now currently under-utilized in fields such as molecular biology.

We have been developing new ways to represent, visualize and interact with the molecular structures that make up the machinery of life. We are adapting two emerging computer technologies, *solid printing* and augmented reality, to create a natural and intuitive way to manipulate, explore and learn from molecular models. We create tangible models utilizing computer autofabrication. Each model can be custom made, with an ease similar to that of printing an image on a piece of paper. Specific model assembly kits can be made with this technology to create *molecular Legos* that go well beyond the chemical models of the nineteenth and twentieth centuries. Augmented reality is used to combine computer-generated information with the physical models in the same perceptual space. By real-time video tracking of the models as they are manipulated we can superimpose text and graphics onto the models to enhance the information content and drive interactive computation.

These models and tangible interfaces have been used in both research and educational settings. The talk will include a live demonstration of the models and interactive use in an augmented reality setting.

3.5 visualization – communicating, clearly


Martin Krzywinski (BC Cancer Research Centre, CA))

License  Creative Commons BY-NC-ND 3.0 Unported license
© Martin Krzywinski

We should think about visualization not only in terms of effective data encodings, but also in terms of design. We use visualizations to communicate patterns and concepts and are more effective if we incorporate design principles in our figures. Clutter and redundancy can muddle a figure – two pitfalls into which many figures fall. Using examples of redesigned figures, I will motivate how mitigating these two issues can improve visual communication. I will also work through a Nature figure redesign in detail to demonstrate the process and how you can apply it in your workflow.

3.6 Concepts gleaned from disparate communities

Bang Wong (Broad Institute of MIT & Harvard – Cambridge, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Bang Wong



4 Working Groups

4.1 Comparative Analysis of Heterogeneous Networks

Andreas Kerren, Corinna Vehlow, Jessie Kennedy, Karsten Klein, Kasper Dinkla, Michel Westenberg, Miriah Meyer, Mark Ragan, Martin Graham, Martin Krzywinski, and Tom Freeman

License © © © Creative Commons BY-NC-ND 3.0 Unported license
 © Andreas Kerren, Corinna Vehlow, Jessie Kennedy, Karsten Klein, Kasper Dinkla, Michel Westenberg, Miriah Meyer, Mark Ragan, Martin Graham, Martin Krzywinski, and Tom Freeman

Transcriptome sequencing of a large cohort is now a routine method of interrogating the profile of expressed gene products of many individuals. Analysis of the transcriptome produces a large number of putatively disrupted transcripts, and prioritizing which disruptions are most likely to be meaningful (causal or diagnostic) is a time-consuming process. Researchers integrate information from a wide variety of annotation databases, many of which are interaction or pathway networks, to guide their interpretation of how a disrupted transcript might affect the functioning of a cell.

We investigated how this analysis can be facilitated by interactive visualizations of transcriptome assemblies in the context of these networks. The goal of our proposed method is to infer the functional consequence of a transcript's disruption based on local structure of the annotation networks (Figure 3). For example, the investigator may have identified functional motifs in the network that are relevant to their hypotheses, and conclude that any disrupted transcripts found in these motifs are likely to be important.

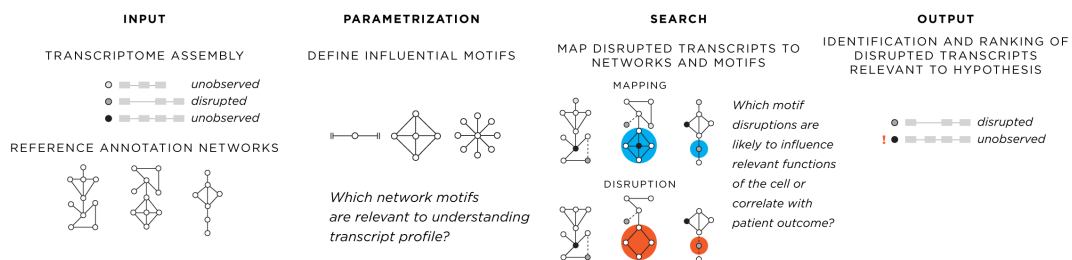


Figure 3 Transcript assemblies from the sequencing of a cancer genome produce indication of disrupted or unobserved transcripts. The existence of reference annotation networks onto which these transcripts can be mapped provides a method of assessing the functional implications of the disruption.

We claim that a tight coupling of network analysis algorithms and interactive visualizations, specifically designed to support these analysis tasks, would accelerate identification of important transcript alterations. A software system that realizes such a coupling could streamline the process of identifying influential network motifs, determining whether disrupted transcripts fall within these motifs, and support the process of deriving a priority rank based on the results of this search. By using a system of linked views, each showing one of the reference networks, the system could provide the researcher a means of mapping transcript disruptions onto the networks. The views would show constrained locales of each motif to decrease the information burden — the reference networks are typically very large (10,000+ nodes) — and preserve the users' mental map. This concept supports the analysis of disruptions in the context of different reference networks at a time and therefore helps users to assess the impact of the disruption.

Our plan is to formalize this analysis method and implement a software system to realize it in practice.

4.2 Sequence Data Visualization

Jan Aerts, Jean-Fred Fontaine, Michael Lappe, Raghu Machiraju, Cydney Nielsen, Andrea Schafferhans, Svenja Simon, Matt Ward, and Jarke J. van Wijk

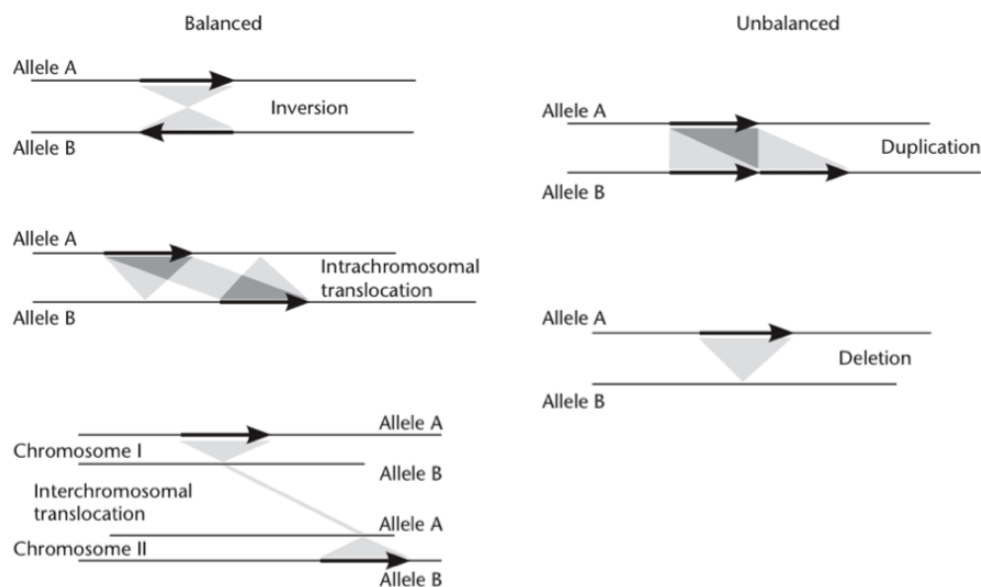
License © © © Creative Commons BY-NC-ND 3.0 Unported license
 © Jan Aerts, Jean-Fred Fontaine, Michael Lappe, Raghu Machiraju, Cydney Nielsen, Andrea Schafferhans, Svenja Simon, Matt Ward, and Jarke J. van Wijk

Introduction

Genome-associated data is growing at a fast rate. Since the advent of large genome sequencing efforts such as the Human Genome Project, the genome browser (e.g. UCSC genome browser, Ensembl) has been the tool bringing all types of data together into a single representation. This report serves to analyze the shortcomings of current genome browsers and to suggest options for solving the problems. We identified two main use cases in using genome browsers: hypothesis verification and hypothesis generation. In the first use case (*hypothesis verification*) users want to verify a hypothesis (e.g. the involvement of a certain gene in the development of cancer) by checking whether experimental data can support this hypothesis. Current genome browsers allow uploading custom experimental data in order to analyze it in context of other data. In the second use case (*hypothesis generation*), users often have access to much experimental data that needs to be interpreted and are looking for significant signals that fall into regions where there is evidence of functional relevance. Although generic genome browsers have clearly proven their worth, they are now starting to show clear shortcomings. These can be separated into two dimensions. First, using a *fixed genome coordinate system* works as long as one is only interested in the reference sequence and features that have fixed and clear positions on that reference sequence. When, however, considering structural genomic variations (i.e. duplications, deletions, inversions, and translocations) the paradigm of annotation vis-a-vis a fixed reference starts to break down. Second, the current concept of displaying features in different *tracks* becomes cumbersome with the immense growth of annotation data. The increasing number of annotation tracks to be selected/deselected for display makes keeping an overview of the available data nearly impossible. In addition to the issues when considering structural variation or large track lists, the integration of uncertainty in feature visualization is still lacking. This uncertainty exists at two different forms: statistical uncertainty and positional uncertainty. *Statistical uncertainty* reflects the confidence that one has towards the existence or correctness of that feature or not. Therefore, it is important to not only view summarizing annotations, but also to be able to investigate the underlying evidence. The representation of statistical uncertainty was the topic of a separate Dagstuhl breakout group, and therefore not further considered within the current group. *Positional uncertainty* considers the resolution of feature annotation. The boundaries (breakpoints) of deletions, for example, can often not be identified exactly, but can be known to lie within a certain range. At present this type of information is not displayed in the generic genome browsers. Below we examine the two dimensions to the problem of displaying genomic information, namely the reliance on a single genomic coordinate system and the large number of feature tracks.

Visualizing genomic structural variation

A structural variant consists of a DNA sequence, typically >1 kilobase, that deviates from a reference sequence in content, order and/or orientation. A distinction can be made between balanced variations (i.e. inversions and translocation) that do not change the total genome content, and unbalanced variations (i.e. deletions and duplications) that do result in a change of the total genome content (see Figure 4). The latter therefore are also known as copy number variations or CNVs. Although detection of structural variations has long been possible using e.g. fluorescent *in situ* hybridization (FISH) or array comparative genome hybridization (aCGH), they are now routinely detected using next-generation sequencing (NGS) technology. After paired-end sequencing of one or more samples and mapping these reads to the reference genome, patterns in read depth as well as aberrant distance between and/or orientation of paired-end sequences can indicate structural variations.



■ **Figure 4** Types of structural genomic variation (taken from [1]).

Problem Statement

The issue with representing structural variation using generic genome browsers is two-fold. First, the data to be represented (both underlying read mapping data and resulting variations) often involves features that are linked at two *different loci in the genome*. Read pairs, for example, consist of two reads that do belong together but can be mapped to very distant regions in the genome. Duplications and translocations also inherently constituted of two elements: the locus that acts as the source for the duplication/translocation, and the locus that acts as the target. Some efforts have been made to resolve this issue, e.g. by providing a split-pane view on the data (Integrative Genome Viewer; Robinson et al, 2011). There is however a second and more profound issue of reliance on a *single reference genome*. Any variation between a sample and the reference can only be displayed as a feature in a track of the genome browser (see Figure 5). As a result, two different samples cannot be compared directly to each other, but can only be compared by how each differs from the reference. In

addition, the reference-genome based representation does not reflect the *in vivo* configuration of the sample chromosome. For example, a region of the reference genome that is deleted in a sample can be highlighted with colored bars in a genome browser (see red bars in Figure 5). But this requires that the user build a mental-model of which genomic regions are now adjacent as a result of the deletion in the sample rather than being able to observe the new junction directly. This is fairly straightforward for simple variants, but quickly becomes challenging for more complex structural changes.

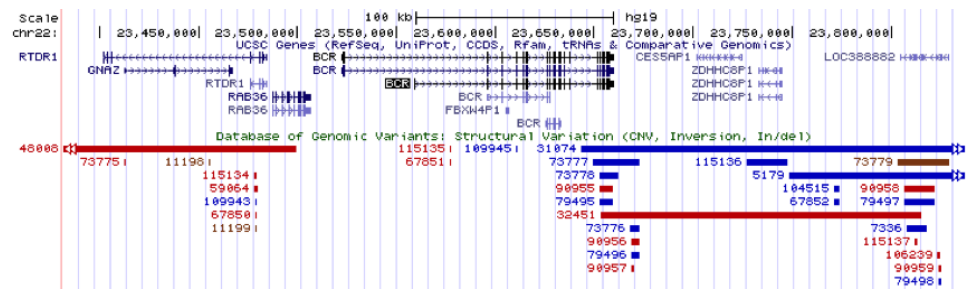
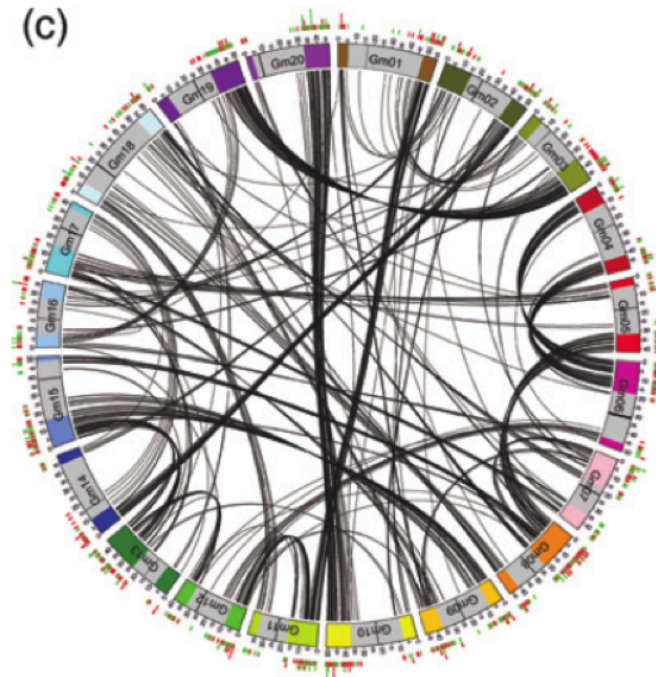


Figure 5 Structural variations annotated around the BCR gene on chromosome 22 (picture taken from the UCSC genome browser). Duplications are in blue, deletions in red, inversions (not shown here) in purple.

Our Approach

Some inroads have been made into representing features with more than one position (e.g. read pairs, or duplications and translocations). Visualization approaches and software tools have been reviewed previously ([13, 12]). Most notable is the use of the Circos viewer ([11]). This circular visualization maps the different chromosomes onto a circle and links related loci through the use of bezier curves (see Figure 6). Another approach is the dot plot in which the x and y axes correspond to the two sequences being compared, and points indicate sequence identity. Diagonal lines indicate corresponding sequence segments and the horizontal offset highlights reordering.

Both circular and dot plot representations emphasize the positions of structural variants on the genomic coordinate. Since the start of the genome browsers, the main nugget of information to be displayed has always been positional. Researchers have become accustomed to this habit, and novel visualization concepts are necessary. One option is to consider the genome as a collection of functional elements rather than a linear scaffold and emphasize the biological consequences of the variants rather than their genomic arrangement. Indeed, the location of a functional element (i.e. a gene together with any *cis*-acting regulators) on a chromosome is irrelevant for most purposes. A reference-free gene-centric approach will therefore be developed where the emphasis is on the contents of the genome rather than on its linear structure. In addition, a genome can be represented as a collection of segments that can be rearranged between different individuals in a graph-like structure. We can draw inspiration from previous work in this area ([14, 7]). By combining these representations with a circularized linear layout (such as Circos), we believe that a researcher can build a comprehensive overview of the effects of a structural variation.



■ **Figure 6** Example of Circos (taken from [17]).

Track compression

Many different annotations and measurements are associated with genomic positions. Since not all available data will fit on a screen, let alone be interpretable to a user, some compression has to take place. This can be achieved by selecting relevant data or by aggregating or summarizing different tracks. In this section, we analyze the chances and challenges of these approaches.

Data Description

The data associated with sequences can be separated into two main types: qualitative or enriched description of a region or quantitative data associated with positions. The region definition usually refers to the reference genome. In cases where one aims at comparing data stemming from different genomic sequences, mapping the data to a common reference scheme can already be a challenge (refer to previous section). The label usually refers to a function of the genomic region, e.g. “protein coding region” referring to the encoded protein, or a summary of quantitative data. In order to be informative to a user not familiar with the different data types, the label often needs to refer to several disparate pieces of information. Quantitative data usually detail the value of measurements per residue position. Different tracks can refer to very different types of measurements, (e.g. expression data) or type of sequenced sample (e.g. tissue or disease group). This can be a challenge to aggregating quantitative data, because accumulating or averaging may not be appropriate.

Track Selection

One approach to managing the multitude of data is to select that subset of data that is relevant to the subject under investigation. An ideal track selection tool would help the user select tracks relevant to their research, which can then be displayed in a genome browser. However, this selection is not trivial:

- The data sets are usually characterized by a concise name that is helpful for the respective domain experts. But for exploration it is hard to find out the content that might be relevant to a specific question. Even if one has found a signal in a region of interest, it is not obvious what that signal means. Reading through all linked data descriptions is very tedious.
- There might be significant overlap and redundancy in the data. Therefore, adding more tracks does not always add more information. In these cases aggregation might be possible. On the other hand information might be discovered in measurements the user is not aware of.

We identified two distinct approaches of selecting relevant tracks: using meta-data or using the data itself:

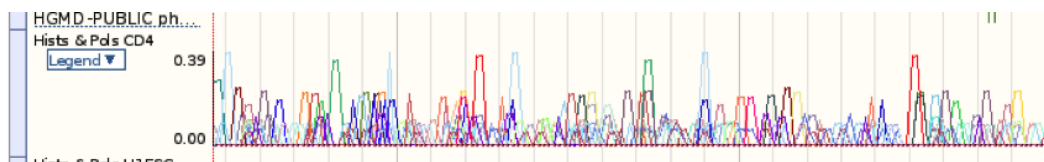
- The different tracks are currently characterized by names and descriptions. Various genome browsers use hierarchical organization and categorization to enable searching relevant data. However, we believe a more sophisticated use of controlled vocabulary or ontologies together with search systems could make the data more accessible. This categorization should allow to group data by different types of approaches, e.g. sample characterization (population, disease association, tissue, cell cycle), types of experiment (expression level, epigenetic modification), or type of summary annotation. An important requirement for this meta-data annotation is to allow grouping together tracks that can be aggregated without mixing apples and oranges.
- The data itself contains information that can help in selecting relevant selections or aggregation. Especially if users have experimental results they want to analyze in the context of the genomic information, looking for other signals that are statistically associated (correlated) with that data can help to find relationships and lead to new interpretations. This can also help to sort the tracks by relevance. Another related approach is to focus on region(s) of interest. Here, a simple filter can help to highlight those tracks that contain more than noise in the specified region. The remaining regions could be analyzed statistically to identify those tracks that show similar signals. On the one hand, this can help to aggregate related data to provide a more concise view; on the other hand the statistical analysis might point to common mechanisms governing the data and lead to new biological insight.

A very different approach to the selection of relevant tracks is the information contained in the community of users of the data. Very similar to the community recommendations used on sales web sites (e.g. Amazon) or social networks (e.g. LinkedIn), the data browser could suggest relevant tracks based on the usage statistics. Users with a high overlap of interest could make each other aware of interesting new directions to explore.

Data Aggregation

Another solution to making the data more manageable is aggregating different tracks into one. For example, if the user is only interested in finding out whether some variations fall into coding regions, all annotations of mRNA-mappings could be aggregated. However, here are a number of problems to solve:

- The semantical problem is to decide which tracks contain comparable information that may be aggregated, e.g. experimental results of a similar type or annotations of a similar type, but from different sources. Here, a good meta-data description, as described in the track selection section, is needed. This meta-data description should enable aggregation by different criteria, e.g. organized by cell or disease type.
- The visualization problem is how to show as much information as possible without overloading the user and without obscuring data. For example, if different quantitative measurements are shown simultaneously, using color or symbols to indicate the different measurements, it becomes hard to discern the different lines. Here a representation of e.g. the mean, the standard deviation, and specific outliers might be more helpful to show the most interesting aspects of the data.



Summary and Outlook

In the Dagstuhl workshop, we have analyzed the different challenges of visualizing genome-associated data and separated them into different dimensions: problems associated with rearrangements of the genomic coordinates and problems with the abundance of data at each genomic position. The problems and approaches for finding solutions outlined in this report will now be taken up in the development of new visualization tools. Although the discussion focussed on issues related to genomic sequence data, similar problems also exist in the realm of protein sequences. Therefore, the new concepts for visualizing data associated with genome sequences will hopefully also help to provide better overviews of protein sequences.

4.3 Bridging Structural & Systems Biology via DataVis

Graham Johnson, Julian Heinrich, Torsten Möller, Seán O'Donoghue, Art Olson, James Procter, and Christian Stolte

License Creative Commons BY-NC-ND 3.0 Unported license

© Graham Johnson, Julian Heinrich, Torsten Möller, Seán O'Donoghue, Art Olson, James Procter, and Christian Stolte

Introduction

Over 50 years of structural biology has yielded detailed insight into the molecular machines of life – from the scale of atoms to organs; the significance of this work with has been recognized by many Nobel Prizes. In contrast, systems biology has evolved over the past 20 years in the wake of the genomics revolution, but separately from the advances of the more structural view of biology. Visualization techniques for both structural and systems biology have both evolved in response to the need to analyze experimental data; in contrast, more general data visualization (datavis) approaches have evolved from a variety of application areas, where biology did not play a major role.

Describing the Gaps between Scientific Disciplines

Systems biology has traditionally utilized data ranging from genomics, proteomics, metabolomics, etc. which attempt to characterize in a systematic way the flow of molecular interaction and information/control. Structural biology, on the other hand, seeks to characterize the physical nature of such interactions and information flow by characterizing spatial and temporal structures. Connecting these two views is a challenge that requires techniques that have been developed on both sides of the gap. Likewise a gap exists between builders of computational tools from the biological community, and those from the computer science community. Within the visualization community, these gaps exist along several dimensions. Firstly, integrating the different data modalities and algorithmic approaches that arise from the structural and systems biology has been a significant challenge. Secondly, the viewpoint of the biologist focuses mostly on the biological question, while that of the visualization specialist focuses on the complete user experience. In addition, there also exist significant cultural gaps between these communities. The biology community and the visualization community publish in different ways, meet at different conferences, and evaluate their work using different criteria. Biology meetings are overloaded with data and urgent, important and unsolved problems, and unmet requirements – as a result, they are segmented into tiny subfields. By contrast, computer science meetings are data- and problem-hungry. These gaps are significant but surmountable, and bridging them holds the promise of pushing biology to the next level. The purpose of this white paper is to propose some strategies and tactics that may help to build these bridges.

Bridging the Gaps

The contrasting metrics for performance in each discipline mean that models for research dissemination do not allow sufficiently rapid transfer of new problems and new visualization solutions between the two domains. This is critical, however, and as a key outcome of this discussion, we define the following recommend strategies to facilitate collaboration and professional advancement in structural biology, systems biology, and datavis.

Mentoring and exchange programs amongst biological visualization, structural biology, and systems biology research groups. The most direct route to enable ideas and approaches to cross fertilize our fields is to facilitate interdisciplinary training. Orthogonal integration, where data visualization students and researchers are temporarily embedded in biology groups, will enhance the exchange of state of the art principles and approaches, and familiarize all parties with the tools, technology and data architectures involved.

Interdisciplinary conferences and symposia. The VIZBI and BioVis meetings already incorporate a range of mechanisms to encourage productive engagement between these communities; this could be strengthened by modeling other interdisciplinary meetings, such as ISMB. In addition, these mechanisms could be encouraged in the larger, more mainstream meetings in each field.

Co-localized hackathons and tutorial workshops. Much of the fundamental software tools used in structural biology today were created by bringing together specialists in numerical computation, physicists, physical chemists and biochemists. Focused workshops would enable engineers, theoreticians and applied researchers to identify new problems and design and prototype solutions informed by the latest visualization research. Whilst virtual participation is eminently feasible for these events, a one or two week period where specialists are physically co-located would maximize productivity. Tutorial sessions and facilitation

would be essential in these events to allow specialists to quickly understand and begin to apply their own knowledge to the problems at hand.

Exploitation of prepublication data repositories. Biological data repositories now play a key role for international collaboration, and could provide a means for data visualization researchers to access data and analysis problem solutions which would allow new solutions to be developed and evaluated in parallel with ongoing biological research programs.

Critical assessment of methodology exercises. Both communities have well established challenges that enable researchers to devise new solutions for data analysis problems. The model originally devised by Moulton et al. in 1991 (<http://www.predictioncenter.org>) applied for the assessment of biomolecular structure prediction approaches has led to a number of initiatives assessing approaches for biological text mining (<http://www.biocreative.org>) and biological systems reverse engineering (<http://www.the-dream-project.org>). These enterprises are distinct from the challenges in the data visualization field such as VAST and the BioVis challenge, since they employ real biological data which is accepted for publication but not yet released.

Clear definition of datavis challenges. If structural and system biologists were to clearly define specific visualization challenges, this would help the datavis community, enabling it to focus on more relevant problems that are likely to be adopted and to help advance the life science. The 2010 Nature Methods special issue (Vol. 7 No 3) and the ongoing VIZBI conference series are useful steps in this direction. The discussion group highlighted the following as key visualization challenges: analysis & comparison of macromolecular ensembles; uncertainty / confidence visualization; mapping of abstract data onto 3D structures, including text, URLs, community curations, as well as data from networks, pathways, populations, geographic distributions, and phylogenies. In accordance with the previous goal of critical assessment, it would help to define several concrete showcases: an example could be the 3D models of HIV being developed by Johnson et al. [2], as these combine many of the data types mentioned above.

Education. Science educators regularly employ data visualization techniques to communicate structural biology, but many biological systems have well established visual representations that conflict or entirely disregard best practices identified by data visualization practitioners. Communication of best practice is essential in order to ensure that new approaches for mesoscale structural visualization maximize the potential of these visual analysis tools, and correct terminology employed to allow biologists to discuss data visualization approaches with computer scientists.

Conclusions

Structural biology can now provide a detailed view of the information environment of systems biology. As the available data on structures and omic-scale systems grow and become more complex, the visualization tools developed in both communities need to be integrated with datavis methods into new toolsets for enhancing both exploration and understanding of these biological systems.

4.4 Ontologies in Biological Data Visualization

Sheelagh Carpendale, Min Chen, Daniel Evanko, Nils Gehlenborg, Carsten Görg, Lawrence Hunter, Francis Rowland, Margaret-Anne Storey, Hendrik Strobelt

License © ⓘ ⊕ Creative Commons BY-NC-ND 3.0 Unported license
 © Sheelagh Carpendale, Min Chen, Daniel Evanko, Nils Gehlenborg, Carsten Görg, Lawrence Hunter, Francis Rowland, Margaret-Anne Storey, Hendrik Strobelt

Introduction

Ontologies are graph-based knowledge representations in which nodes represent concepts and edges represent relationships between concepts. Ontologies have been used extensively as computational models in natural language processing, artificial intelligence, and the web sciences. A number of disciplines in which visualization plays an important role, including biology, are using ontologies to support the analysis of large and complex datasets. We examined how ontologies can be used to support biological data visualization and identified challenges and opportunities from the perspectives of three different stakeholders: ontologists (who create and maintain ontologies), data curators (who use ontologies for annotation purposes), and data analysts (who use ontologies through applications to analyze experimental data). A summary of the challenges and opportunities is presented below; we also submitted a more detailed discussion as a Viewpoints article on Ontologies in Biological Data Visualization to the IEEE Computer Graphics & Applications journal.

Challenges

A first challenge is centered around the dynamic nature of ontologies. Many ontologies constantly change and evolve due to discoveries and newly acquired knowledge in the domain they represent. The creation of multiple versions of ontologies is prone to inconsistencies and also adds downstream complexity for their users (humans as well as computer programs). Keeping track of evolution becomes even more daunting for ontologies that integrate multiple data sources. The evolution of ontologies affects all three types of stakeholders.

A second challenge is scale: many ontologies represent an overwhelming amount of data. These large ontologies are usually developed and maintained by a team of ontologists which requires a framework that supports the collaborative work on ontologies. Data curators often face the problem of finding the most appropriate concepts in these large ontologies when they annotate terms in documents or samples in experimental data. For data analysts, the amount of ontology annotations can be easily overwhelming, especially if documents or data are annotated with multiple ontologies.

A third challenge is related to the relationships and types that are represented in ontologies. So far, most applications do not take advantage of the complex set of relationships in ontologies but rather reduce them to a simple hierarchy. While this approach is certainly useful for data analysts it does not exploit the full potential of ontologies. The underlying problem here is that the representation and visualization of complex relationships is hard. The complex set of relationships within ontologies, combined with the large size of ontologies, makes their manual maintenance a considerable effort for ontologists.

Finally, to make ontologies more useful for data analysis, it is crucial to understand what analysts want to investigate and how they can use ontologies for their specific tasks. To this end, user studies are required to understand the workflows and aims of the analysts.

Research Opportunities

Visualization of Ontologies: Despite much research on the topic of ontology visualization, the majority of the tools developed thus far are focused on visualizing or navigating the ontologies themselves, rather than on visualizing content that has been annotated with ontological concepts. We propose that there is a need to develop tools that are both powerful and easy to use for curators of content as well as for users browsing ontologically annotated content (such as journal articles). Such tools can furthermore support the consumers of this content to do richer analyses of associated content.

Automated Generation of Visualization using Ontology: The availability of domain-specific ontologies provides an exciting opportunity for developing automated visualization methods and services. Although interaction remains as an important apparatus for facilitating data exploration, it may incur costly time and learning effort for using a visualization system. In many application scenarios, automatically-generated visualizations may serve users more efficiently and effectively, and can facilitate knowledge sharing among users.


Visualization of Ontological Context in Supporting Search: The application of ontologies and even multiple ontologies to annotate text corpora offers new potential and new challenges. Search is currently being explored within ontologies. This can be expanded to consider search across multiple related ontologies and inverted to include search via multiple ontologies. Ontologically annotated text provides semantically rich meta-data. Since visualizing even simple meta-data has been shown to enhance serendipity in information exploration, visualizing ontologically annotated text is very promising. However, the complexity of text visualization coupled with the complexity of ontology visualization makes this a big challenge.

Conclusion

By capitalizing on ontologies as knowledge representations, we will be able to make a significant step towards the realization of knowledge-assisted visualization [3]. It may take the form of automated visual annotation of texts, documents and corpora, automated construction of visualization for novice users, or automated visualization of ontological context in information retrieval.

4.5 A Framework for Effective Visualization Design

Miriah Meyer, Jan Aerts, Dan Evanko, Jean-Fred Fontaine, Martin Krzywinski, Raghu Machiraju, Kay Nieselt, Jos Roerdink, and Bang Wong

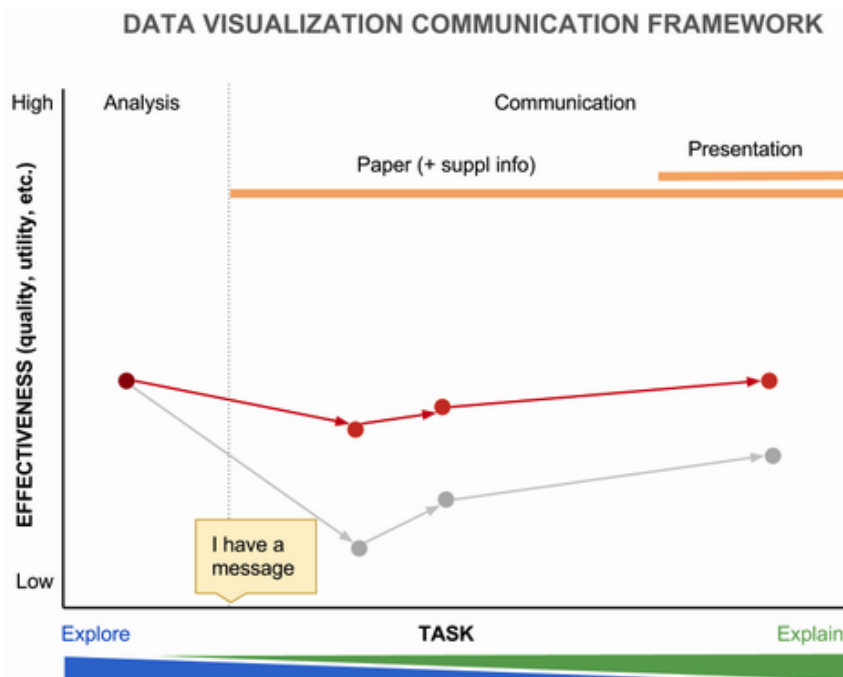
License  Creative Commons BY-NC-ND 3.0 Unported license

© Miriah Meyer, Jan Aerts, Dan Evanko, Jean-Fred Fontaine, Martin Krzywinski, Raghu Machiraju, Kay Nieselt, Jos Roerdink, and Bang Wong

Visualizations are not only an important aspect of how scientists make sense of their data, but also how they communicate their findings. The techniques and guidelines that govern how to design effective visualizations, however, can be quite different whether the goal is to *explore* or to *explain*. For example, if the goal is to support hypothesis generation from a large, genomics data set then techniques like multiple-linked views and data-rich visual representations are good considerations, as opposed to a visualization used in a conference presentation where significant abstraction of the data and simple visuals are necessary.

Unfortunately, scientists are often not aware of the spectrum of considerations when creating visualizations, resulting in ineffective figures, diagrams, and tools when there is a mismatch between the goal of the visualization and the design decisions. These considerations encompass audience, presentation modality, and the amount explanation versus exploration that is needed. To help clarify this problem, we have developed a framework to reason about the spectrum and considerations to help scientists better match their visualization goals with appropriate design considerations. We believe that awareness about this spectrum can improve visualization, particularly those targeting explanatory goals, as well as enable more fruitful discussions between scientists and visualization designers.

The framework (Figure 7) has a major axis that describes how exploratory or explanatory a visualization task is. On one end, pure exploratory visualizations are mostly likely to be interactive, and are often meant to support hypothesis generation, data and model validation, and scientific insight. On the other end, pure explanatory visualization are meant to communicate an idea, story, or scientific finding, usually in a highly abstract, simple way. It is important to map goals to a position along this axis because different design considerations exist from left to right. Things to consider are: who am I communicating to? Someone in my lab? In my department? In my scientific community? In the general public?



■ **Figure 7** Data Visualization Communication Framework.


These different locations on the task axis have different design consideration. For exploratory visualizations, these considerations are largely drawn from computer science, and are things such as interactivity, how to display or summarize large data sets, and how to support complex relationships. For explanatory visualizations, these considerations come largely from the design community, such as how to abstractly represent data and relationships, how to filter out unnecessary data and details, and how to tell a story visually.

The task axis also coincides with numerous other secondary consideration axes. Some of these are: considering the richness of the data, how much complexity can be shown; considering the amount of data, how much of the data must be filtered out; considering

hypothesis generation, can the viewer form new and different hypotheses; considering the time commitment of the viewer, can they get the message in 5 seconds, 5 minutes, or 5 hours; and considering the domain expertise, how much specific knowledge does it require. The second axis describes the effectiveness of the visualization. Considering this axis there is an important implication: moving back and forth has gravity. What this means is that an effective visualization at one point within the space will almost always be less effective when used directly for an application further along the major axis. For example, using an interactive genome browser with a full data set is effective for exploration, but will perform terribly in a conference presentation on a scientific finding. And conversely, a diagram for the general public explaining a scientific concept will almost certainly fail to produce new hypotheses for a research scientist. In summary, knowing where you are within the framework can help in picking appropriate design guidelines and visualizations for a specific communication goal.

4.6 Uncertainty Visualization

Min Chen, Julian Heinrich, Jessie Kennedy, Andreas Kerren, Falk Schreiber, Svenia Simon, Christian Stolte, Corinna Vehlow, Michel Westenberg, and Bang Wong

License  Creative Commons BY-NC-ND 3.0 Unported license

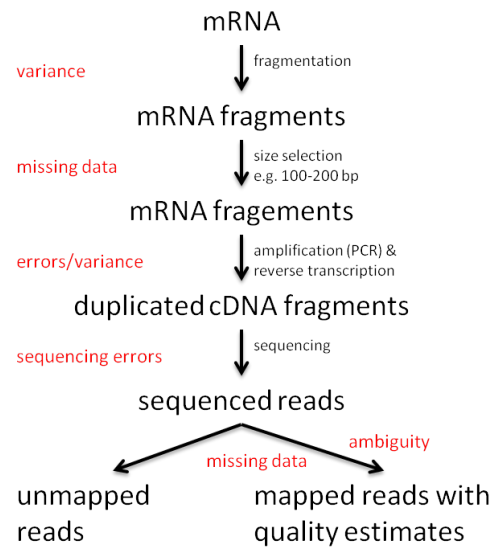
© Min Chen, Julian Heinrich, Jessie Kennedy, Andreas Kerren, Falk Schreiber, Svenia Simon, Christian Stolte, Corinna Vehlow, Michel Westenberg, and Bang Wong

Uncertainty is common in all areas of science, and it poses a hard problem to deal with in terms of visualization. There is no well-established definition of uncertainty, but several types of uncertainty are generally distinguished: measurement precision, completeness, inference, credibility, and disagreement [18]. Visualization of uncertainty has received much attention in the areas of scientific visualization and geographic visualization. Several techniques have been proposed employing special visual encodings (transparency, blur, error bars), addition of glyphs, modification of geometry, and animation, to name a few. Application of uncertainty visualization appears much less common in information visualization and in biological data visualization.

The working group looked at sources of uncertainty and types of uncertainty specific to biology. We studied a (RNA) sequencing (RNAseq [21]) pipeline, and identified the types of ambiguity that can be introduced in each step, see Fig. 8. Many steps are prone to introduce errors, some of which create a certain bias in what RNA fragments are ultimately amplified and sequenced. The output that comes from the pipeline is a set of mapped reads with an associated quality value that could be used (but is rarely in practice) in visualizing the sequences.

We also looked at computational models derived from the literature. Here, uncertainty is apparent in the model itself (granularity and structure), in the simulation (initial parameters, numerical inaccuracy), and verification of the simulated model by lab experiments (measurement errors). An open issue that we identified here concerns uncertainty visualization in networks (which represent the model): uncertainty in network topology, and the problem of dealing with dynamic (uncertain) attributes on nodes, edges, and their interdependencies.

The working group performed a quick scan of recent papers that employ some form of uncertainty visualization for biological data. We found several examples, including applications in population variability [4], expression data analysis [9, 22], data cleansing [15], and network modeling [16, 20].



■ **Figure 8** Uncertainty (red text) in the sequencing pipeline.

Our plan is to further extend the result of this quick scan into a literature survey paper that specifically addresses uncertainty visualization in biology. We propose to construct a taxonomy of uncertainty visualization approaches, and investigate how they could be employed in the context of a collection of biological problems.

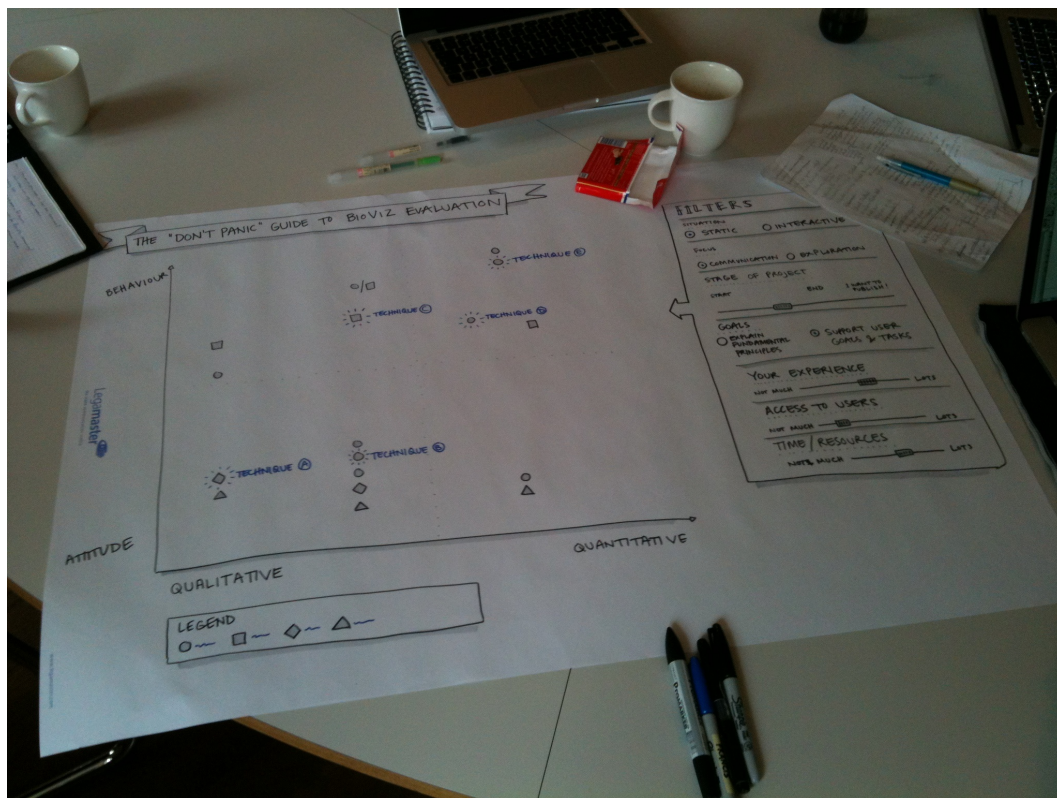
4.7 Evaluation

Jarke J. van Wijk, Kasper Dinkla, Martin Graham, Graham Johnson, Francis Rowland, and Andrea Schafferhaus

License Creative Commons BY-NC-ND 3.0 Unported license

© Jarke J. van Wijk, Kasper Dinkla, Martin Graham, Graham Johnson, Francis Rowland, and Andrea Schafferhaus

In our breakout group we discussed issues concerning evaluation in biovis. We agreed that two problems were central. First, how to get biologists motivated to participate in evaluations? During the talk we had a questionnaire send out and asked for feedback. The main result was (simply) that biologists have to get something out of it to motivate them: a tool they can use to solve their problems, and also chocolate and beer were suggested. Second, we focussed on how to evaluate. We discerned a number of dimensions, centered around what, why, when, where, and how. We found a strong difference between designers working at a bio-institute and infovis researchers. The former is characterized by close cooperation during development, a qualitative approach, and a focus on creating a tool; whereas the latter group performs evaluation typically afterwards, aims at quantitative results and creation of new techniques. Both have merit, the challenge is to close the gap. We decided that it would be great to develop a tool that shows various options for evaluation methods, given a specification of the problem, and designed a first mock-up, consisting of a set of filters/choices and a scatterplot showing approaches (see Figure 9). We aim to develop this idea further after the seminar.



■ **Figure 9** First sketch of a tool that shows various options for evaluation methods.

4.8 Multiscale Visualization

Carsten Görg, Graham Johnson, Karsten Klein, Oliver Kohlbacher, Thorsten Möller, Arthur Olson, Francis Rowland, and Matt Ward

License © © © Creative Commons BY-NC-ND 3.0 Unported license

© Carsten Görg, Graham Johnson, Karsten Klein, Oliver Kohlbacher, Thorsten Möller, Arthur Olson, Francis Rowland, and Matt Ward

Overview

In recent years the progress in the understanding of biological processes, in combination with the corresponding collection of large amounts of heterogeneous experimental data, led to raised requirements for corresponding visualizations. These visualizations should allow biologists to analyze data with respect to complex and even whole-organism models [10], that combine data of differing type, multiple dimensions such as space and time, as well as multiple scales in those dimensions.

While there are still visualization problems to solve for single dimensional data (e.g. effective comparison or representation of dynamics for protein interactions changing over time), the additional challenges here include the linking of different scales and types of data for presentation or interactive exploration. Stephen Prevenas (thelazygeeks.com) uses the following analogy to describe the shortcomings of just combining current single scale solutions:

Think of trying to watch *The Empire Strikes Back*, but the actors are on your TV,

the scenery is on your iPad, and the soundtrack is on an 8-track. Ok, maybe not an 8-track, that makes this analogy absurd. Say the soundtrack is on a CD, which I admit is only marginally more believable at this point, but I also shouldn't assume you have an iPad and an iPod.

We started to study these challenges, identify the underlying problems, and summarize them to lay the foundation for further research.

Goals and Discussion

We first had an exchange on the personal background and expectation of the group members regarding the working group. Even though the backgrounds were quite heterogeneous, including people from visualization, design, bioinformatics, and biology, the expectations were similar. We agreed that multiscale visualizations of biological data have many aspects and cover such diverse application fields that we first had to agree on a basic characterization of what multiscale means in the context of our discussion before we could even start working on the corresponding challenges. In addition, we wanted to collect examples of practical multiscale problems that combine multiple dimensions of scale, which could help us to derive a characterization.

We thus worked towards the following two aims:

1. Discussing and defining a formal characterization of multiscale visualizations of biological data.
2. Collecting multi-scale visualization examples to foster our search for a practically useful characterization and to present to the seminar participants.

It turned out that both problems were not easy to solve. We did not find examples that cover more than one or two dimensions (usually spatio-temporal dimensions), and several levels of scale; in fact, we were not aware of well-defined levels and dimensions besides the standard spatial and temporal dimensions. We also agreed that there is no sufficient, and at the same time unambiguous and generally accepted definition of multiscale in the domain of biology, and that this lack of definition would hinder further discussions in our group. We therefore decided that the first task of our group would be to develop our own definition, or at least try to cover the most important aspects in a characterization, and that we then should discuss the challenges and open problems with respect to the most interesting tasks and data.

The main questions we discussed in the following tried to link these two tasks: How can multiscale examples be systematically classified or categorized and what are reasonable dimensions in which scaling takes place. We spent half of the time of our meeting to discuss the nature of model and data modalities, different corresponding potential dimensions, and how they could be clearly separated. It turned out that it was a difficult task to agree on a simple but precise characterization that is useful as a base for further investigation. As our first approach to come to some understanding what defines a multiscale visualization challenge, we decided to describe corresponding processes, data, and tasks in terms of a coordinate system that is made up of the dimensions that fully characterize their multiscale nature. However, we discovered that the proposed dimensions often overlapped, were impossible to grasp and define formally, or seemed not to be useful enough for further investigations. In particular, there was a long discussion about the term “data complexity”. The questions we asked were, among others:

- How can “complexity” of data be captured, is it represented e.g. by the entropy, and does it include the size of the data?

- Is instead data size one of the scales, and if yes, does this cover problems where different data sizes are part of the visualization output, e.g. for comparison, or also cover problems where different amounts of input data are processed?

Obviously data volume does not directly translate into content information, but content information alone does not fully cover the corresponding problems. In the end, we decided that due to problems like limited human perception and computational complexity the data volume, i.e. the data size, has to be taken into account as one of the data dimensions. Since we also wanted to capture at least some aspect of complexity besides data volume, we discussed to also add the number of sources of the data, i.e. the method to generate it, and the number of representations as additional dimensions. In the end we reduced our characterization to an easy to understand four dimensional coordinate system that covers the complexity at least to a significant extent. These four dimensions of scaling include the quite natural and well-accepted time and space dimensions. In addition, we chose the number of modalities, which is a way to cover complexity without having to define a complexity metric for all kinds of data, and data volume as the two other dimensions.

After we agreed on this basic characterization we started collecting examples of multiscale visualizations to (1) provide an understandable view of the nature of the different scales and how they apply to biological data and exploration, and (2) to investigate the shortcomings of existing approaches. As a goal, we would like to bring together good and bad examples, where a classification according to the extent with which they cover the dimensions of our multi-scale characterization should allow better access to our collection.

We decided to produce a short informational video that presents multiscale visualizations for the different dimensions in our coordinate system, both for presentation and getting feedback from the seminar participants, as well as a prototype for a result to be published later. To have a clear and easy to understand demonstration of the different dimensions, we selected a common topic that allows to cover each of them. We chose the human body for that purpose and picked examples for each dimension that represent different aspects.


Another (short) topic of discussion was how to actually represent different scales like cellular and molecular level. Several approaches exist: (1) fly through the scales consistently from one to another, (2) combined visualizations (focus), and (3) multiple linked views (e.g. a magnifying glass). Linked to that is the problem of transition between different modalities (like a graph and the physical representation), and if they can be combined or need to be visualized in parallel. In order to make use of these representations, suitable interaction techniques need to be developed, but this discussion was outside the scope of this working group.

Outcome

“Understanding Multi-Scale Visualization” is a design and prototype for a short video exposition on the nature of multi-scale data, models and visualization in biology. Its goal is to provide a quick and understandable view of the nature of the different scales and how they apply to biological data and exploration. Our characterization could be used as a foundation for a taxonomy of multiscale techniques, and our example collection could show which parts of biological data space have been explored already.

4.9 Infrastructure

Seán O'Donoghue, Tom Freeman, Mark Ragan, Margaret Storey, Larry Hunter, Cydney Nielsen, Nils Gehlenborg, Jim Procter, and Hendrik Strobelt

License  Creative Commons BY-NC-ND 3.0 Unported license

© Seán O'Donoghue, Tom Freeman, Mark Ragan, Margaret Storey, Larry Hunter, Cydney Nielsen, Nils Gehlenborg, Jim Procter, and Hendrik Strobelt

Overview

The infrastructure working group discussed needs from both a technical and community standpoint regarding the challenges involved in the analysis of biomedical data derived from the Cancer Genome Atlas project and mechanisms to facilitate interactions between visualization communities in computer science and biology. Eight key criteria were identified: *Interoperability, reusability, compatibility, references & benchmarks, middleware, vertical integration, scalability, and sustainability*, and two outcomes. The first outcome is a model for a community-maintained, biological visualization resource that would enable biological questions, task descriptions, sample datasets and existing tools for the problems to be disseminated to the computational visualization and biological research communities. The second is a detailed use-case based on the data and analysis pipelines of the cancer genome atlas that will allow technical aspects of the eight key criteria to be explored and practical solutions proposed.

Key Criteria for BioVis Infrastructure

Interoperability. The success of systems such as Galaxy [5] and Vistrails (<http://www.vistrails.org>) demonstrates that BioVis tools developed by different groups in the community must interoperate, at the very least through consistent data exchange standards, but also in the provision of well designed and documented control interfaces to allow pipelining and orchestration.

Reusability. Best practices are needed to encourage groups to develop tools with standard interfaces that allow them to be embedded or combined with other tools (e.g. as widgets) in a variety of situations.

Comparability. Two aspects were identified: It should be straightforward to compare different tools that perform a similar function in order to assess which one is most appropriate for a use case. Effective BioVis tools should also provide visualizations that support comparative analysis of biological data.

References & Benchmarks. A standard model and repository is needed to allow reference biological problems to be described, along with representative datasets and analysis outcomes. Benchmark problems should be significant and representative of key biological questions, and task descriptions should be presented in a predigested manner such that a non-specialist in the field can understand the analysis processes required without deep knowledge. These reference descriptions can be used as benchmarks to evaluate existing BioVis tools and inform the design of new tools in the future.

Middleware. Middleware can facilitate *interoperability, reusability, comparability*, and *vertical integration*, providing it is easy to use, it also lowers the cost of entry to the field. In fact, several middleware libraries exist for biological data, but they are typically associated with a particular modality or biological domain. A new breed of middleware is needed that can draw together standard technical solutions from all relevant biological information

domains, such as image processing, text mining, and biomolecular sequence and structure analysis.

Vertical integration. Computational analysis pipelines are commonly used in biology, but it is essential that these can be published in a way that allows an analysis to be reproduced by other researchers. A number of systems that support provenance and pipeline dissemination have been developed in computer science (e.g. myExperiment, VisTrails, etc.) but the Bio & Vis communities must work to make their use routine in data driven biology.

Scalability. It is safe to assume that any notion of ‘big data’ currently described will be relatively small compared to the volume of data that must be handled by our tools in the future. Technical solutions (e.g. data/processing clouds) already exist, but the users of the system are often the most serious bottleneck and data and derived analysis results need to be delivered in a usable manner.

Sustainability. Open Sustainability models are required for software as well as data created by grant-based biological research, to ensure the community at large can access the outcomes of research. Such practices are not commonplace in computer science laboratories, where prototypes serve only as proof-of-concepts to be abandoned rather than refined to make them usable by biologists. New standards for software tools must be declared in both communities, and public repositories (such as the one described below) should be created to enable rapid interchange of new tools and datasets, and maintenance of previously developed tools.

Outcomes

We decided to develop two outcomes following our initial discussions. A community resource for biological task, data and tool dissemination and a case study to explore the software infrastructure necessary to support a current biomedical research problem.

Community resource for benchmark problems, datasets and available tools

A web platform to support community maintained descriptions, datasets and instances of tools relevant to BioVis could be provided that would act as a bridge between the CS and biological BioVis communities by integrating with both the biovis.net and vizbi.org sites. It will provide:

- Biological problem/analysis task descriptions described in a way that is accessible to lay-biologists and computer scientists.
- Data sets relevant to problems, using standard file formats or links to archive quality web based databases.
- Descriptions of available tools. Each tool should be linked to, or archived on the site so that it can be launched, along with instructions describing how to perform the task with this tool.

It is essential that tasks, datasets, and tools provided by the resource are *significant*, *representative*, *selective*, and *predigested* in order to ensure that the resource is relevant to both the computer science and biological community. A number of starting points were proposed, including reaping problem descriptions and tools that solve problems from the burgeoning number of Stack Overflow style sites [19] and deriving descriptions from the reviews created by the VIZBI community. In order to be sustainable the resource will require a community of editors to be recruited and the engagement of tool authors to maintain the public descriptions of their work.

Use case based on the Cancer Genome Atlas (CGA)

“The CGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.” (from the CGA’s own materials). Although at a preliminary phase, the numbers involved in this experimental study are already staggering: 20 tumor types are being investigated by analyzing affected and unaffected tissue from 500 patients to identify variation in single nucleotide polymorphisms (SNPs), copy number (CNVs), DNA methylation, mRNA and microRNA transcripts, and gene mutations.

Data from each experiment requires one or more computational analysis pipelines, and the results must be validated, integrated and understood in order to elucidate the driving mechanisms in each tumor type, and evaluate the efficacy of available therapies for each individual. The CGA have developed a data and result staging system, *Firehose*, and a visualization tool – *StratomeX*, based on *Caleydo* (an Eclipse Rich Client Platform (RCP)), that provides genome-scale integrated genome/transcriptome views across these data. However, a number of data curation and deep biological analysis tasks are very difficult.

The following criteria were identified for a next generation CGA visualization system:

1. Data Provenance. System needs to display the origin and processing pathway for the data currently visualized, and ideally allow comparison of results of alternate processing pathways.
2. Standardized representation of data and analysis processes. Provenance models require well defined representations such as the Predictive Model Markup Language [Wikipedia] to describe the transformations that data undergoes prior to visualization. Formal representations also enable interoperability and reproducibility.
3. Remote access to data. Most sets of data are too large to fit into memory – the system needs aggregation and subsetting mechanisms to allow browsing of the complete dataset on any reasonable user platform.
4. Pluggable architecture. Alternate visualizations for the same data, or additional visualizations for new or derived data facilitate deep analysis, and encourages contribution from third-parties. Architecture will also allow new data format and analysis process support.
5. Communication between plugins/modules. Synergies are important: communication between distinct visualization modules with shared selections, colorings, etc. allow greater insight. The modules in the system should also be able to select appropriate modules for performing particular purposes – such as computing a distance matrix for particular biological entities and then clustering to yield an appropriate visualization.
6. Global communication/user experience. Several different types of users will enter and use the system in different ways – a core system event model will need to support arbitrary routes through the data and analysis process.

These criteria were explored further, to identify base layer software components and the key visualizations that would be needed and the kind of provenance associated with each one. A user story describing how a typical cancer biologist will employ the system was proposed to explore how the different analysis and visualization components will be required to interact.

Conclusion

The infrastructure working group identified a number of technical and social requirements that should be addressed by the community. We proposed the development of a community resource to collate problems and solutions for biological visualization tasks, and this will be explored further. We also developed a detailed use case based on a current biomedical research

problem to help identify technical and conceptual challenges in biological visualization that the community should prioritize in the future.

5 Acknowledgements

We would like to thank all participants of the seminar for their contributions and lively discussions; we also would like to thank the reviewers of our initial proposal for their constructive feedback and the scientific directorate of Dagstuhl Castle for providing us with the opportunity to organize this seminar. Finally, the seminar would not have been possible without the untiring help of the (scientific) staff of Dagstuhl Castle, including Ms. Susanne Bach-Bernhard, Ms. Jutka Gasirowski, and Dr. Marc Herbstritt.

References

- 1 Jan A Aerts and Chris Tyler-Smith. *Structural Variation in Great Ape Genomes*. John Wiley & Sons, Ltd, 2001.
- 2 autoPACK model of HIV 1.4 running in PMV with Screen Space Ambient Occlusion Narrated. <http://www.youtube.com/watch?v=W84yW9HIzCI>.
- 3 Min Chen, D. Ebert, H. Hagen, R.S. Laramée, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *Computer Graphics and Applications, IEEE*, 29(1):12–19, jan.-feb. 2009.
- 4 M. Corell, S. Ghosh, D. O'Connor, and M. Gleicher. Visualizing virus population variability from next generation sequencing data. In *Proc. IEEE Symp. Biological Data Visualization (BioVis)*, pages 135–142, 2011.
- 5 J. Goecks, A. Nekrutenko, J. Taylor, E. Afgan, G. Ananda, D. Baker, D. Blankenberg, R. Chakrabarty, N. Coraor, J. Goecks, G. Von Kuster, R. Lazarus, K. Li, A. Nekrutenko, J. Taylor, and K. Vincent. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11(8):R86, 2010.
- 6 D. Gray. *Gamestorming: A Playbook for Innovators, Rulebreakers, and Changemakers*. O'Reilly, 2010.
- 7 A. Herbig, G. Jager, F. Battke, and K. Nieselt. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, 28(12):7–15, Jun 2012.
- 8 L. Hohmann. *Innovation Games: Breakthrough Products Through Collaborative Play: Creating Breakthrough Products and Services*. Addison Wesley, 2006.
- 9 C. Holzhüter, H. Schumann, A. Lex, D. Schmalstieg, H.-J. Schulz, and M. Streit. Visualizing uncertainty in biological expression data. In *Proc. SPIE 8294, Visualization and Data Analysis (VDA 2012)*, 2012.
- 10 J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival, N. Assad-Garcia, J. I. Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, Jul 2012.
- 11 M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res.*, 19(9):1639–1645, Sep 2009.
- 12 C. Nielsen and B. Wong. Points of view: Representing genomic structural variation. *Nat. Methods*, 9(7):631, Jul 2012.
- 13 C. B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, and T. Wang. Visualizing genomes: techniques and challenges. *Nat. Methods*, 7(3 Suppl):S5–S15, Mar 2010.
- 14 B. Paten, M. Diekhans, D. Earl, J. S. John, J. Ma, B. Suh, and D. Haussler. Cactus graphs for genome comparisons. *J. Comput. Biol.*, 18(3):469–481, Mar 2011.

- 15 T. Paterson, M. Graham, J. Kennedy, and A. Law. VIPER: a visualisation tool for exploring inheritance inconsistencies in genotyped pedigrees. *BMC Bioinformatics*, 13(Suppl 8):S5, 2012.
- 16 H. Rohn, A. Hartmann, A. Junker, B. H. Junker, and F. Schreiber. FluxMap: a Vanted add-on for the visual exploration of flux distributions in biological networks. *BMC Systems Biology*, 6:33, 2012.
- 17 A. Roulin, P. L. Auer, M. Libault, J. Schlueter, A. Farmer, G. May, G. Stacey, R. W. Doerge, and S. A. Jackson. The fate of duplicated genes in a polyploid plant genome. *Plant J.*, Sep 2012.
- 18 M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualization. In *Proc. AVI'08*, pages 376–379, Napoli, Italy, 28–30 May 2008.
- 19 Margaret-Anne Storey, Christoph Treude, Arie van Deursen, and Li-Te Cheng. The Impact of Social Media on Software Engineering Practices and Tools (DCS -338-IR). In *FSE/SDP Workshop on the Future of Software Engineering Research (FOSER 2010)*, pages 359–364, 2010.
- 20 C. Vehlow, J. Hasenauer, A. Kramer, J. Heinrich, N. Radde, F. Allgöwer, and D. Weiskopf. Uncertainty-aware visual analysis of biochemical reaction networks. In *Proc. IEEE Symp. Biological Data Visualization (BioVis 2012)*, volume 2012, pages 91–98, 2012.
- 21 Z Wang, M Gerstein, and M Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.
- 22 M. A. Westenberg, J. B. T. M. Roerdink, O. P. Kuipers, and S. A. F. T. van Hijum. SpotX-plore: a Cytoscape plugin for visual exploration of hotspot expression in gene regulatory networks. *Bioinformatics*, 16(22):2922–2923, 2010.

Participants

- Jan Aerts
K.U. Leuven, BE
- Sheelagh Carpendale
University of Calgary, CA
- Min Chen
University of Oxford, GB
- Kasper Dinkla
TU Eindhoven, NL
- Daniel Evanko
Nature Publishing Group, US
- Jean-Fred Fontaine
Max-Delbrück-Centrum, DE
- Tom Freeman
University of Edinburgh, GB
- Nils Gehlenborg
Harvard University, US
- Carsten Görg
University of Colorado, US
- Martin Graham
Edinburgh Napier University, GB
- Julian Heinrich
Universität Stuttgart, DE
- Lawrence Hunter
University of Colorado, US
- Graham Johnson
UC – San Francisco, US
- Jessie Kennedy
Edinburgh Napier University, GB
- Andreas Kerren
Linnaeus University – Växjö, SE
- Karsten Klein
The University of Sydney, AU
- Oliver Kohlbacher
Universität Tübingen, DE
- Martin Krzywinski
BC Cancer Research Centre, CA
- Michael Lappe
CLC bio, DK
- Raghu Machiraju
Ohio State University, US
- Miriah Meyer
University of Utah – Salt Lake City, US
- Torsten Möller
Simon Fraser University – Burnaby, CA
- Cydney Nielsen
BC Cancer Agency's Genome Sciences Center, CA
- Kay Nieselt
Universität Tübingen, DE
- Sean O'Donoghue
CSIRO – North Ryde, AU
- Arthur J. Olson
The Scripps Research Institute – La Jolla, US
- James Procter
University of Dundee, GB
- Mark Ragan
The Univ. of Queensland, AU
- Jos B.T.M. Roerdink
University of Groningen, NL
- Francis Rowland
EBI – Cambridge, GB
- Andrea Schafferhans
TU München, DE
- Falk Schreiber
IPK Gatersleben, DE
- Svenja Simon
Universität Konstanz, DE
- Christian Stoltz
CSIRO – North Ryde, AU
- Margaret-Anne Storey
University of Victoria, CA
- Hendrik Strobelt
Universität Konstanz, DE
- Jarke J. Van Wijk
TU Eindhoven, NL
- Corinna Vehlou
Universität Stuttgart, DE
- Matthew O. Ward
Worcester Polytechnic Inst., US
- Michel A. Westenberg
TU Eindhoven, NL
- Bang Wong
Broad Institute of MIT & Harvard – Cambridge, US

