# Visualizing biological data—now and in the future

Seán I O'Donoghue[1], Anne-Claude Gavin[1], Nils Gehlenborg[2,3], David S Goodsell[4], Jean-Karim Hériché[1], Cydney B Nielsen[5], Chris North[6], Arthur J Olson[4], James B Procter[7], David W Shattuck[8], Thomas Walter[1] & Bang Wong[9]

Methods and tools for visualizing biological data have improved considerably over the last decades, but they are still inadequate for some high-throughput data sets. For most users, a key challenge is to benefit from the deluge of data without being overwhelmed by it. This challenge is still largely unfulfilled and will require the development of truly integrated and highly useable tools.



Computer-based visualization is widely used in biology to help understand and communicate data, to generate ideas and to gain insight into biological processes. This collection of reviews examines the key methods now being used to visualize genomes[1], alignments and phylogenies[2], macromolecular structures[3], systems biology data[4] and image-based data[5]. Here, we outline several common trends, challenges and recent advances that suggest the nature of future visualization in biology.

## Visualization goes mainstream

Twenty years ago, only experts could create computer images of a protein structure at atomic detail, a large phylogenetic tree, or a complex biochemical pathway. Today, software tools for creating these images are widely available and widely used. Of the different visualization areas in biology, molecular graphics[3] is perhaps the most mature, and as a result, molecular graphic images are widely used in textbooks, presentations and popular media. Other fields, such as genome visualization[1], are much younger; however, even here, molecular biologists have a rich toolbox of visualization software at their disposal, many of these tools amenable to use by non-experts[1].

A main reason for the increased accessibility and use of visualization software has been the advances in computer hardware and network access. Many visualization tasks that previously required expensive and specialized hardware can now be easily managed with a standard personal computer. However, an equally important factor has been the development of a wide range of methods and tools specialized in visualizing specific kinds of biological data. In this Supplement, we discuss over 200 tools selected from the much greater number now available. This diversity of tools can be confusing, but it is probably unavoidable, given the diverse nature of the biosciences. In fact, in many cases, biologists still find that their exact requirements are not met by current tools and often have to create custom solutions. This has helped spur a growing trend to allow reuse of visualization software, either by means of open source software libraries (for example, http://www.vtk.org/) or by means of architectures specifically designed to allow extensions (for example, Cytoscape[6]).

## Integration is improving

In the past, visualization tools were typically stand-alone programs designed to view data from a single experiment. In contrast, many of today's tools are integrated with remote databases and provide visualizations that integrate data from multiple sources. For instance, Jalview[7]—a popular tool for editing multiple sequence alignments—can connect to multiple data sources and displays not only alignments but also a wide variety of sequence feature information.

In addition, tools are increasingly being designed to interoperate directly with other visualization and analysis tools. Such interoperation can enable, for example, simultaneous interactive visualization of a multiple sequence alignment with corresponding three-dimensional structures (Procter *et al.*[2] and O'Donoghue *et al.*[3])—or of a network with corresponding heat maps, profile plots or phylogenetic trees and dendrograms (Gehlenborg *et al.*[4]).

Finally, many of today's visualization tools can be either directly embedded into, or launched from, web pages; and such tools are being used to construct integrated web applications for data mining and browsing, often using multiple visualization tools. For example, the UCSC Genome Browser[8] shows genomic sequences assembled from many laboratories and provides access to a diverse range of related data, including multiple sequence alignments among sequences from similar organisms, three-dimensional structures and *in situ* hybridization images.

The improved integration in visualization tools has been helped greatly by a trend toward increased consolidation of experimental data. An exemplary case of this trend is macromolecular three-dimensional structure: almost all experimentally determined structures are consolidated in a single resource (wwPDB[9]). Unfortunately, such consolidation is still the exception: it is more typical in biology to have equivalent data distributed over many resources. In the case of image data from high-throughput experiments, most of these data are never made publicly available, even though this would clearly be of value. Some preliminary steps

[1]European Molecular Biology Laboratory, Heidelberg, Germany. [2]European Bioinformatics Institute, Cambridge, UK. [3]Graduate School of Life Sciences, University of Cambridge, Cambridge, UK. [4]The Scripps Research Institute, La Jolla, California, USA. [5]British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, British Columbia, Canada. [6]Virginia Tech, Blacksburg, Virginia, USA. [7]School of Life Sciences Research, College of Life Sciences, University of Dundee, Dundee, UK. [8]Laboratory of Neuro Imaging, University of California, Los Angeles, California, USA. [9]Broad Institute of MIT & Harvard, Cambridge, Massachusetts, USA.
e-mail: sean.odonoghue@embl.de

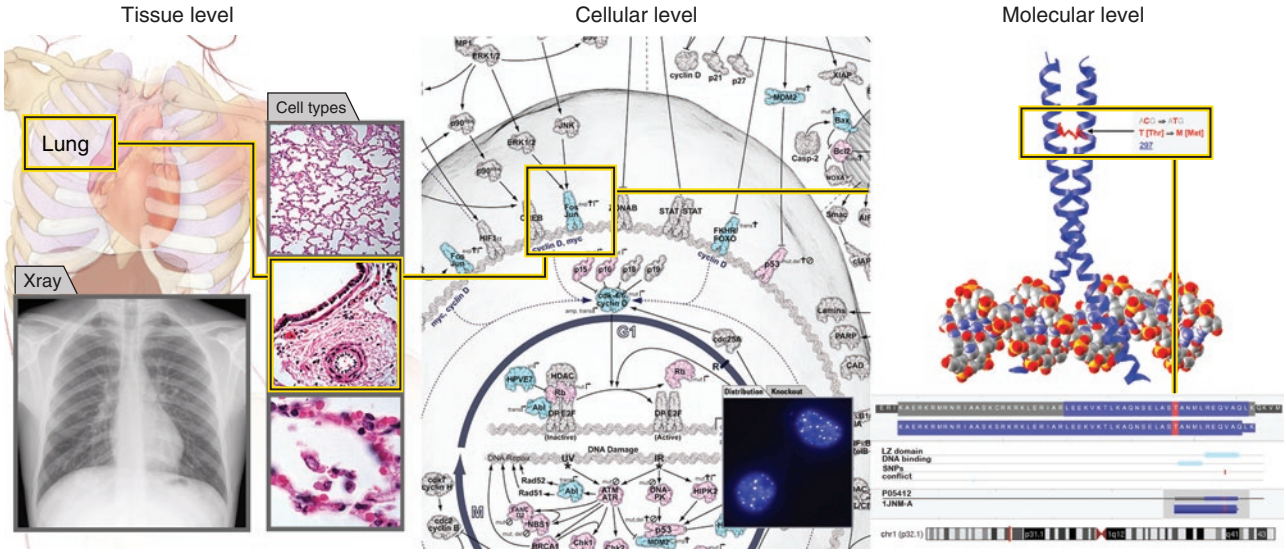


**Figure 1** | Possible integrated visualization environment. Soon, biologists may be able to seamlessly move between data from tissue, cellular and molecular scales, as well as data from genomes, networks and pathways. Many of these data will be organized around a cellular coordinate framework, and visualization of biochemical pathways will allow increasingly detailed representations of cellular topology and of proteins. For instance, selecting a tissue (left image) could automatically show micrographs of cell types; selecting a cell type could show relevant pathways (center image); selecting a protein from the pathway could access micrographs showing the cellular distribution and effects of knockdown experiments (bottom, center); in addition, selecting the protein could show atomic-detail three-dimensional structural information, sequence features, alignments and genomic location. Many of these interoperations are already being used today. Images courtesy of ClearScience (drawing), iStockPhoto (lung X-ray), Univ. of Kansas Medical Center (lung histology), Digizyme and Cell Signaling Technology (pathway). Protein structure and sequence alignment made using SRS 3D; chromosome image from UCSC Genome Browser[8].



are being made (for example, CCDB[10], http://ccdb.ucsd.edu/), but a truly consolidated resource for image data is likely to remain a distant goal owing to difficulties with defining standards for organizing and categorizing these data and to data set sizes that are prohibitively large for network-based transfer.

## User-interface challenges

Although visualization methods and tools have greatly improved, there has also been an exponential increase in the size and complexity of data sets studied in biology. A common challenge faced by many biologists is how to benefit from this data deluge without being overwhelmed by it. Visualization is clearly part of the solution; however, the sheer number and diversity of tools available can make the problem worse. Below we discuss several recent advances toward addressing these issues.

**Usability.** Very often, biologists fail to fully benefit from visualization methods because software tools are too difficult to learn. Making software that is easy to use often requires considerable work. Fortunately, there have been many advances in understanding principles of software usability[11]. These principles are increasingly being adopted by developers of visualization tools for biologists. Judging from progress over the past decade, we expect the usability standard to continue to improve. Unfortunately, improvements may be slow, because work on usability is usually less rewarded in science than is research on new methods.

**Visual analytics.** In the process of understanding and interpreting biological data, tools ideally would provide visualization for tasks that require human judgment, and other tasks would be automated where possible. But, finding a productive balance between automation and visualization is a challenge and is one of the goals of visual analytics methods[12], which involve studying the role of visualization in the whole process of analyzing and understanding data. Recently, these methods have begun to be applied to biological visualization tools, and, if successful, these developments will improve the ability of tools to provide meaningful biological insights[13] and to meet user requirements[14].

**Multiscale representation and navigation.** Biological data visualization often deals with a broad range of scales—for example, images may range from the atomic scale to the cellular level[3,5], and genomic browsers provide information from whole chromosomes down to an individual nucleotide position. To be useful, the graphical representations used need to adjust, ideally displaying the level of detail appropriate to a particular scale. For example, in showing the three-dimensional structure of a protein, ribbon representation is often used to hide all atoms except those involved in ligand interactions; as a user zooms out to see higher-order protein complexes, ribbon representation is too detailed and is replaced by an overall surface. Although the basic ideas are not new, the details of how to realize multiscale navigation vary greatly with the data type. This is the subject of ongoing research, particularly in visualizing genomic data, pathways and networks and joint visualization of image data sets acquired at different resolution, requiring multimodal image registration.

**Innovative representations.** In all areas of biology, new visual metaphors and graphical representations are being developed to convey information and to facilitate navigation. Innovation of representations is often inspired by the need to visualize new types of data or to support new analysis tasks. Examples include the need to display expression profile data together with pathway data (Gehlenborg *et al.*[4]) or the need to make genome assembly structures easier to see (for example, ABySS-Explorer[15]). In some cases, the innovations are brought in from outside of biology; for instance, partial order graphs are representations taken from discrete

mathematics that are now being used to create concise summaries of multiple alignment information (for example, POAVIZ[16]) and to visualize alternative gene splicing.

**Standardized representations.** Because visualization methods are still rapidly evolving, part of the difficulty faced by end users today arises from a lack of standards in representations. Although there is an obvious strength in diversity, and indeed a need for continued innovation in graphical representation, in many cases usability would be enhanced by the adoption of some standards in representation. In systems biology, there has recently been a significant community-driven proposal[17] toward developing a more unified standard for graphical notation of biochemical networks, and we anticipate similar proposals in other areas.

**Display hardware.** To help display and use complex biological data sets, large display devices and tiled arrays with improved resolution are likely to be of significant benefit[18]. As these devices become more affordable, they are likely to see more use. For instance, touch tables are promising for navigation and collaborative work on complex phylogenetic hierarchies (http://involvweb.org/).

**Adding a third dimension.** The use of three-dimensional visualization is being explored for networks[19], phylogenetic trees (Procter *et al.*[2]) and genomics data (for example, http://genodive.org/). Although the third dimension adds complexity to the user interface, three-dimensional visualization may be necessary for some very complex data sets. Visualization in three dimensions is helped greatly by hardware stereo, which is now becoming easily affordable.

**Augmented computer interaction.** For challenging data sets, we anticipate the increased use of methods that augment or improve the ability to interact with visual data. For example, tangible devices that give touch feedback are becoming more affordable and are promising for three-dimensional structure visualization[20]. Preliminary studies on augmenting visualization with auditory techniques ('sonification') have also been done, using molecular three-dimensional structure and sequence information, and preliminary results are encouraging[21].

## Computational challenges

Today, many visualization tasks are easily accomplished using a standard personal computer. However, in almost all areas of biology, visualization of cutting-edge data sets remains a challenge. For instance, a modern high-throughput image data set may consist of thousands of videos or hundreds of channels (each channel typically corresponding to one gene product)—and may be up to tens of terabytes in size. To interactively visualize these data, personal computers are often inadequate.

This situation is inspiring further innovation in software—especially in methods for dimension reduction and classification, which underlie visualization tools in many areas of biology. For example, the recently developed MCL clustering algorithm[22]—which enables fast network clustering—has been implemented in a range of visualization tools, particularly in systems biology. Although these advances will undoubtedly improve on today's limitations, our ability to collect data will also continue to improve, and it is certain to continually challenge our visualization capabilities.

## Future visualization

Ultimately, the goal of visualizing biological data is to provide biologists with an integrated framework they can use to gain insight into the processes in organelles, cells, organs and even whole organisms. Fulfilling this ambitious goal requires substantial further development in visualization methods, especially better integration of different tool types.

Several efforts to build such integrated visualization frameworks have begun—for example, using the framework of genomic coordinates to integrate increasingly diverse data[23]. Other frameworks based on commonly used systems biology data types are being developed[24]. And projects from the structural biology and microscopy communities aim to integrate biological data on the basis of a cellular coordinate framework (for example, Visible Cell[25] and others[26]) by synthesizing multiscale data, including data from cellular tomograms, cryo-electron microscopy, and atomic-detail three-dimensional structures, as well as inventories of expressed proteins, estimations of organelle shapes and distributions, and protein localizations and gradients. Probably no single framework will suit all biologists; however, the goals of these different efforts may eventually produce a standardized visualization environment that allows seamless integration of biological data (**Fig. 1**).

Creating such integrated visualization frameworks will require a collective effort, and several initiatives toward collaborative, community-based editing of biological image data have already begun (for example, CATMAID[27]). But all these efforts are still very much at the pioneering stage, and, to paraphrase Alan Kay, we could say that the revolution in biological data visualization hasn't started yet.

**COMPETING INTERESTS STATEMENT**
The authors declare no competing financial interests.

1. Nielsen, C.B., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. *Nat. Methods* **7**, S5–S15 (2010).
2. Procter, J.B. *et al. Nat. Methods* **7**, S16–S25 (2010).
3. O'Donoghue, S.I. *et al. Nat. Methods* **7**, S42–S55 (2010).
4. Gehlenborg, N. *et al. Nat. Methods* **7**, S56–S68 (2010).
5. Walter, T. *et al. Nat. Methods* **7**, S26–S41 (2010).
6. Shannon, P. *et al. Genome Res.* **13**, 2498–2504 (2003).
7. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. *Bioinformatics* **25**, 1189–1191 (2009).
8. Rhead, B. *et al. Nucleic Acids Res.* **38** (database issue), D613–D619 (2010).
9. Berman, H., Henrick, K. & Nakamura, H. *Nat. Struct. Biol.* **10**, 980 (2003).
10. Martone, M.E. *et al. J. Struct. Biol.* **161**, 220–231 (2008).
11. Shneiderman, B. & Plaisant, C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* 5th edn. (Addison Wesley, Reading, Massachusetts, USA, 2009).
12. Thomas, J.J. & Cook, K.A. *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (National Visual Analytics Center & IEEE, Richland, Washington, USA, 2005).
13. Saraiya, P., North, C. & Duca, K. *IEEE Trans. Vis. Comput. Graph.* **11**, 443–456 (2005).
14. Mirel, B. *J. Biomed. Discov. Collab.* **4**, 2 (2009).
15. Nielsen, C.B., Jackman, S.D., Birol, I. & Jones, S.J.M. *IEEE Trans. Vis. Comput. Graph.* **15**, 881–888 (2009).
16. Grasso, C., Quist, M., Ke, K. & Lee, C. *Bioinformatics* **19**, 1446–1448 (2003).
17. Le Novère, N. *et al. Nat. Biotechnol.* **27**, 735–741 (2009).
18. Ball, R. & North, C. *Comput. Graph.* **31**, 380–400 (2007).
19. Freeman, T.C. *et al. PLOS Comput. Biol.* **3**, e206 (2007).
20. Gillet, A., Sanner, M., Stoffler, D. & Olson, A. *Structure* **13**, 483–491 (2005).
21. Garcia-Ruiz, M.A. & Gutierrez-Pulido, J.R. *Interact. Comput.* **18**, 853–868 (2006).
22. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
23. Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R. & Stein, L. *BMC Bioinformatics* **2**, 7 (2001).
24. Shannon, P.T., Reiss, D.J., Bonneau, R. & Baliga, N.S. *BMC Bioinformatics* **7**, 176 (2006).
25. Burrage, K., Hood, L. & Ragan, M.A. *Brief. Bioinform.* **7**, 390–398 (2006).
26. Suderman, M. & Hallett, M. *Bioinformatics* **23**, 2651–2659 (2007).
27. Saalfeld, S., Cardona, A., Hartenstein, V. & Tomancák, P. *Bioinformatics* **25**, 1984–1986 (2009).