POINTS OF VIEW

Integrating data

Different analytical tasks require different visual representations.

Different data types have their own inherent structure that makes specific visualization techniques most fitting. For example, a matrix of gene expression values for given cell measurements can be highly informative when displayed as a heat map or parallel coordinate plot. The challenge is finding visualizations that will effectively combine data types. Many research studies depend on integrating data to comprehend underlying processes. Here we explore ways to merge data that are best represented as heat maps and node-link diagrams: two common but disparate graphing techniques.

Visualization approaches that are aimed at merging two or more graphical forms need to strike a balance between the optimal representation of one data type versus the other. As we discussed in previous columns, networks are naturally displayed as node-link diagrams or adjacency matrices[1], and the most effective visualizations for expression matrices are heat maps or parallel coordinate plots[2]. The goal of blending multiple data types into a single visualization is to discover correlations, common trends or potential causal relationships that would otherwise be difficult to deduce from the constituent data sets.

The design of a combined visualization depends on what the analysis task calls for. Take, for instance, a matrix containing expression data of genes over time or under different conditions (**Fig. 1a**) and a network defined by the interactions of the corresponding gene products (**Fig. 1b**). If the intention is to understand how changes in gene expression might be explained by how the genes are regulated, replacing the nodes in the interaction network with the expression profiles is a practical approach (**Fig. 2a**). The 'heat strips' make it possible to quickly find nodes in the network with uncommon or specific expression profiles. This strategy also allows one to study the behavior of individual expression profiles in the context of a network, but its utility is limited to a handful of time points.

With data sets containing more time points, examining each time point in succession is more manageable. To do this, use color to indicate the expression levels of the nodes in the network and allow users to interactively step through the sequence of frames (**Fig. 2b**). In this way, one can repeatedly toggle between states to understand the differences in expression between two time points in one or a small group of genes in the network. Although our perceptual system is exquisite at detecting changes between two consecutive images, using such an interactive 'sequence of stills' approach requires the viewer to keep in memory what he or she sees between frames and essentially limits the analysis to pair-wise comparisons. Alternatively, by plotting the networks as 'small multiples' arranged in a line or a grid, where each instance of the node-link diagram represents a time point, we can minimize the viewer's need to remember complex patterns (**Fig. 2c**). The ability to simultaneously see multiple time points also enables one to look for correspondences between a dozen or more conditions.

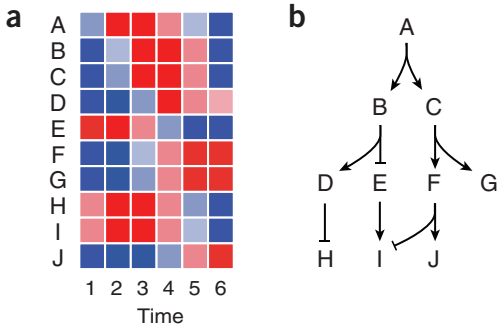

Figure 1 | Different representations for different data types. (**a**) Heat map showing gene expression levels across time. (**b**) A network relationship of the gene products from **a**.

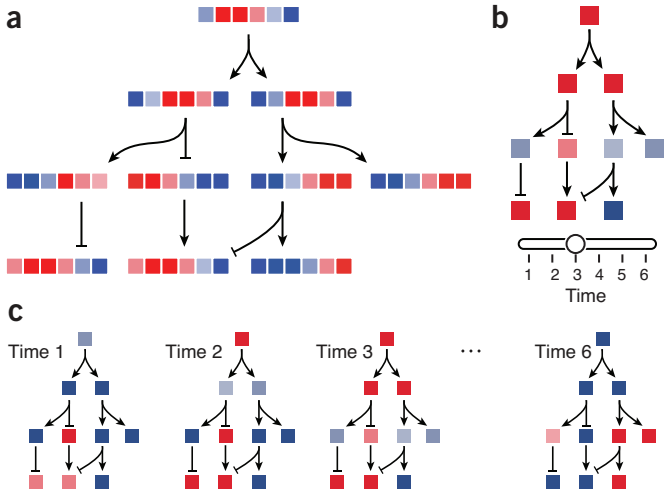

Figure 2 | Integrated views of data. (**a**) The complete expression profile for each node is displayed in the context of a network. (**b**) A network contains the expression values at one time point; users can interactively view time points in sequence. (**c**) Same as **b**, except all time points are presented simultaneously.

The suitability of the approaches discussed above strongly depends on the question one is trying to answer. Distinct graphing techniques emphasize different aspects of data and the ability to see data in discrete forms enables deeper understanding of the subject under study. It is useful to have tools that implement all or at least several of them in a single interface. A compelling example is the Cytoscape plugin Cerebral[3], which offers linked visualizations for a detailed node-link diagram, small multiple views of the interaction network as well as a parallel coordinate plot of expression profile. Such tools are well-suited for data exploration as they facilitate the process of switching between different data views and analysis tasks.

In future columns we will explore the design of data representations in genome browsers.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 115 (2012).
2. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 213 (2012).
3. Barsky, A. *et al. IEEE Trans. Vis. Comput. Graph.* **14**, 1253–1260 (2008).

Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute. Bang Wong is the creative director of the Broad Institute and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.