POINTS OF VIEW

Multidimensional data

Visually organize complex data by mapping them onto familiar representations of biological systems.

The biological researcher can access many methods to rapidly interrogate molecular structures and mechanisms. Such experiments typically involve numerous independent variables, such as substrates, measurement modalities and experimental conditions. Many of these variables may be causally correlated, and the data likely address multiple hypotheses. This multidimensional complexity can make it difficult to design a figure that clearly presents both the structure and value of data in a manner relevant to the inquiry.

When communicating complex data, focus on their meaning instead of structure—anchor the figure to relevant biology rather than to methodological details. What are the interesting findings, and what representation would communicate them clearly? Answering these questions may mean forgoing the conventional approach to displaying multidimensional data (Fig. 1). Instead, it may be better to project the data onto familiar visual paradigms, such as a protein network or pathway, to saliently show biological effects in a functional context.

Figure 1 | Dimensions can be encoded as spatial or visual elements, such as along x and y axes or by color, size or symbol. The number of dimensions and the selection and layering of encodings can have a profound effect on clarity.

An example of an effective presentation of multidimensional data is shown in Figure 2, from a study of drug effect on a network of signaling proteins across a variety of immune cell types¹. The figure uses the method of small multiples: each table cell is based on a schematic of the protein network, onto which quantitative data are projected as colored circles. Rows and columns represent experimental conditions. The figure is readily understood by experimentalists because it leverages biological context to relate the organizational details of the experiment.

The design decision that makes Figure 2 so effective is the use of spatial encoding to present the data domain (the protein network). It

Figure 2 | Overview of the impact of a drug class on a signaling network in different cell types. Colored circles encode EC_{50} and percent inhibition using the scheme in Figure 3c. Adapted from ref. 1.

Figure 3 | Design schematic for Figure 2, showing data structure, variable type and visual mapping. (a) Identification of nested data dimensions informs the levels of organization in the figure. (b) Data types and encodings used. (c) The protein dimension is spatially encoded into a diagram of the signaling network and tabulated by experimental condition. The adjacency of proteins signifies involvement in the same pathway, and vertical position relates to intracellular position. Protein nodes are combined into shapes. Perceptually accurate size and hue encodings² are used for EC_{50} and percent inhibition.

maintains the functional relationship between the proteins, making it possible to assess the drugs' impact on the network, which is the intention of the study. Had the spatial encoding been used for the quantitative variables, as exemplified by Figure 1, this relationship would be muddled and the pathway analysis confounded. Figure 2 scales well without being overwhelming—the original shows 392 different cell type–drug combinations¹.

In planning the design for a complex figure, it is helpful to list the relevant variables of the experiment (Fig. 3a). The next step is to classify the variables and select the encoding method (Fig. 3b). Effective encodings will maintain the nesting and multiplicity of the data structure in the final version (Fig. 3c).

Tabular small multiples are well suited for applications that offer interactive exploration. The scope of data can be focused (such as by transcription factors), the range of data narrowed (by high-potency effects) or the table rearranged. Remember that when presenting tabular data, the order of rows and columns can both reveal and hide patterns.

The final design (Fig. 3) is unencumbered and accommodates selective emphasis of pathways (via colored highlighting) or proteins (via thicker strokes). The ability to focus the reader's attention on specific elements in displays of complex data is desirable and is made possible by a light visual style. Row and column numbers are used to aid data lookup.

In the design of your figures, look to leverage existing biological conceptual models to organize the presentation of your high-dimensional data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Martin Krzywinski & Erica Savig

1. Bodenmiller, B. *et al. Nat. Biotechnol.* **30**, 858–867 (2012).
2. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 769 (2012).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Erica Savig is a PhD candidate in Cancer Biology in the laboratory of Garry P. Nolan at Stanford University.