Visualization in Bioinformatics

Kiara J. Esteves, Ushar Jaikarran & Shelley Luong

Abstract

In Biomedical Informatics, there is a constant need to represent large amounts of data, and visualization is used for this purpose. Formatting visualizations requires a series of strategies: the composition and layout of the presentation, as well as the choice of colors, must be superb. Additionally, the elements of a figure play a vital role in broadcasting the message. Without each of these components, the overall message will be misinterpreted by the audience and ultimately be useless. A collection of the "Points of View" articles in *Nature Methods* provides advice and tips for visualizing scientific data. Here we present a breakdown of these techniques for creating visualizations that will intrigue your audience, helping them to understand the message you are trying to convey.

Presentation

When designing anything, one must have a well-thought-out plan. When there is a well-designed plan, there will be a well-designed object. This object should not require instructions, but should be understood clearly by the audience based on the cues from the presentation. The author of "The Design Process" (one of the "Points of View" articles) believes that the "effectiveness [of the design] determines the user's ability to decode visual cues logically and finds the best solutions for objectives within given constraints" (Wong, 2011d). Producing a good structure requires the utilization of mathematical concepts since the placement of items relies on the ratio of the space to the rest of the page. Typically, our gaze lingers in densely clustered areas of an image. When lines are present, our eyes will follow these lines to connect objects. Additionally, our vision can detect the patterns in our work. For example, in a PowerPoint presentation, if image are placed in the same location on every slide, our eyes will look to that area first when the slide changes. *Therefore, putting the relevant information in that same spot manipulates the audience into reading it first.* To get them to look at another critical section, one can make that part larger to draw their attention. In this essay, each page has a margin with the same amount of white space. The margins are an example of utilizing layout to maintain uniformity.

Shapes

A series of principles can further define the placement of the objects. The "Gestalt Principles offer useful guidance to describe relationships between objects based on certain cues" (Wong, 2011b). These principles help explain how people group information based on how the information looks. We tend to group objects that look alike, are connected by lines, or enclosed in a common space as belonging together. When things are similar, people tend to formulate patterns out of the objects. Additionally, our mind needs conformity. When we see an object with an incomplete portion, our mind automatically fills in the space to make a shape that is familiar to us. While our minds automatically complete an object, our eyes can ignore certain elements. When reading this essay, a person will not notice every single period, comma, or quotation mark. We will be busy reading the surrounding words and glance over the periods because our minds acknowledge the pause.

Frame

The portions of the page that are unmarked by content are the negative space. Negative space increases visual appearance, and the effectiveness of posters, slides, and figures. The space between paragraphs tells the mind to take a break before continuing to the next section. The gaps and margins also provide a frame around the section, so the reader knows that everything within that frame describes the same thing. In presentations, the creator's first instinct is to fill the negative space because of their perception of it being expendable, or an indication of insufficient content. However, if we were to fill every blank spot with information, it will be difficult for the eye to focus on one particular aspect, and the reader is likely to miss information and miscomprehend the data. Overall, the contents of the page should take on a familiar shape such as a rectangle. This space helps our eyes focus on reading in a hierarchy. If words are sticking out at the end of a sentence, our eye will go to those words. By creating flat sides, we will read from top to bottom because that's the way we learned to read, and there will be more emphasis on the content within the frame.

Inconsistency

Inconsistency can make people pay attention to remote regions of an image or irrelevant information. “Salience is the physical property that sets an object apart from its surroundings” (Wong, 2011c). When looking at an excel spreadsheet, for example, not all of the information is important all of the time. In such instances, we can highlight the rows or columns that contain the most relevant information. Without highlighting, our eyes are overwhelmed and may focus on the data that isn’t pertinent to our study. With posters, information that is enlarged will be more easily seen and therefore read first. To get the audience to notice a particular image first, we can display this picture in color while the rest are in black and white.

Colors & Fonts

When we want to differentiate categories of information, color is utilized. Applying the rainbow of colors to represent broad ranges of values is usually the conventional approach. However, this isn’t always accurate because one color cannot represent a numerical value. Additionally, a person’s eye can have difficulty recognizing color changes then the data is subsequently misinterpreted. According to the article in “Points of View*,”* “every color is described by three properties: hue, saturation, and lightness" (Wong, 2010a). In a color picker software, a color wheel arranges hues with saturation decreasing the outside inward. Changes to all properties provides an endless selection of colors. With small datasets, we can afford to create a large variety of colors to represent the data. Large datasets cannot employ considerable variety because then bias can be built. Preference controls whether the viewers ignore valuable information. Arranging the font artistically and technically requires skills of typography. The quality of how we arrange the letters on a page can impact how people respond to our messages. There are two original letterforms known as serif and sans serif. Serif letterforms tend to be thinner and easier to read in long lengths of text, whereas sans serif is less readable in long stretches of text, so they are more appropriate for headings and labels.

Annotations & Symbols

Complex figures rely on labels to identify components and define terms. Labels are annotations, so they should be subordinate to their data points and not to other labels. Keep labels simple and easy to read. It is important to choose symbols that communicate relationships in the data. Symbols that have similar appearances can be easily missed, especially in regions where symbols overlap. It is best to use the first letter in the category name as a plotting symbol. Decoding figures are thus easier because the reader doesn't have to refer to the legend constantly.

Arrows

In the July 2011 issue of *Nature Methods*, there are nearly 300 arrows and more than half of the figures contain arrows. Arrows are guides for complex information. For example, an arrow with a right angle in molecular biology is a transcription start site or promoter. Arrows can have multiple meanings even if used in the same figure. In a study, college students were asked to evaluate diagrams with or without arrows. The results showed that participants who saw a chart with arrows included twice as much information in their description than those who had diagrams without arrows. Therefore, arrows focus the attention on the functional relationships between the elements. Arrows shouldn't be too big to distract us from the content they intend to emphasize.

Plots

Creating a successful plot requires you to understand the data. Bar charts and box plots are typically used to visualize quantities associated with something. But you have to choose the appropriate plot to represent the data accurately. Bar charts are used for counts, whereas box plots are used to visualize distributions. It is preferable to summarize with a box plot because when the numbers are too large for us to see them correctly, we can set a box plot with a range.

Diagrams & Pathway Diagrams

Sets are a general concept in scientific data analysis. An example is a set of discovered bacterial species located in a soil sample or variants found in a genome. A single task is the examination of the commonalities and differences of multiple sets by intersecting them. Junctions of assemblages are commonly illustrated using Euler or Venn diagrams. An alternative is to use ellipses, which produces an area-proportional solution. Adequate visualization of intersections for more than three sets requires a more scaled approach than Euler diagrams. One solution is to encode all intersections in the columns of a matrix using a binary pattern and to render bars above the matrix columns to represent the number of elements in each intersection. The bars can be logged to accommodate significant variations in junction size. Pathway diagrams describe the connectivity and flow in biological systems. The diagrams must depict patterns in connectivity, and they must show both direct and indirect relationships with the viewer to understand the pathway. Visual grouping creates a hierarchy for the flow of information in a channel and alignment emphasizes node relationships. Edges should connect to a fixed number of points on node shapes. Neural circuit diagrams show connections between neurons and brain regions. Simplification leads to greater clarity. The best way to simplify is to reduce the number of elements on a figure. When you show less on the screen, it demonstrates a greater emphasis on what is shown.

Heat Maps

Heat maps represent two-dimensional tables of numbers as shades of colors. It’s a favorite plotting technique used in biology to depict gene expression and other multivariate data. Heatmaps are well suited for the presentation of high-throughput data and rely on color encoding and reordering of the rows and columns. When either of these is compromised, then the visualization will suffer. In heat maps, clustered rows and columns create blocks of similarly colored cells that are easy to spot. One should avoid using red-green as a color combination because it limits accessibility to information for colorblind individuals.

Patterns

Time is unidirectional. It provides a natural order for events and has a semantic structure. Temporal data is cyclic and exhibits repeating patterns. The only challenge is that humans can not directly perceive time. There are approaches to visualizing temporal data; time is encoded using position, brightness, or animation. Consider the position first: time is mapped on the horizontal axis. When dealing with recurring patterns, compare the individual models in the data by breaking the time dimension into equal intervals and then aligning the intervals to emphasize the recurring pattern. If the cycle length changes over time, break the data into intervals of variable lengths and normalize them to a uniform cycle length to emphasize the recurring pattern, or you could leave the intervals unchanged to illustrate the difference in cycle lengths. A single plot scale showing all the data will most likely look like a jumble of lines with no patterns visible. Variable combinations can be distinguished using colors, dashed lines, and symbols. In a tight space, it can be challenging to find encodings that are readily distinguished. Figures of this type are confusing because many features are battling for emphasis, which can inhibit our perception of any pattern.

Dimensions

Three-dimensional data (3D) data is more complicated than two-dimensional (2D) data due to allowing another data dimension for space. It is challenging to understand the data since quantitative, relational and categorical data are difficult to represent in spatial relationships. When electing to use it, there will be partial occlusions, indications of depth and "perspective created by converging parallel lines, which enable us to estimate distances of objects from a certain vantage point" (Gehlenborg, N. & Wong, B.,2012). Unfortunately, gene expression and biological networks, such as a biological pathway, does not benefit from 3D spatial visualization; therefore, we would revert to 2D. Using 2D visualizations combined with multivariable data is efficient and reliable. This could be achieved by using plots such as parallel coordinate and scatter plots while applying color, sizes and shapes to it. As the complexity of the data increases, the difficulty in designing the appropriate figure also increases; this especially rings true for multidimensional data. The ideal figure would contain the structure and the value of the data but since there are too many variables, choosing to focus on the meaning of it would be a better alternative. In other words, the user should decide to concentrate on the relevant biology instead of the methodological aspects of it as it would provide a more efficient presentation of the data.

Sketching

Often, we do not pay much mind to one of the most basic forms of data collection—the utilization of pencil and paper—but we use it almost daily, whether it is to sketch or take notes. The strength of using pencil and paper to jot data down is the immediacy. By using a graphical representation, we take advantage of the ability for the human eye to track patterns, infer connections in the data and refine the hypothesis, whereas with tabulated data this would be almost impossible. When dealing with high dimensional data sets, it is imperative to avoid displaying too much data; the solution being to either leave some of the data out or provide a subset of the sample. The goal is to find the behavior amongst multiple components and the strategy is to restrict one plot per component. But what happens when it's a network, such as interactions between DNA, RNA, and small molecules, that needs graphical representation? Applying restrictions may work, but the audience won't get the full picture. Therefore, hubs and clusters may be used instead. When dealing with node-link diagrams, there are edges, which are the lines that connect the nodes; nodes could be directed and undirected meaning that the edges are asymmetric and symmetric respectively. This is very useful for complex interactions.

Networks

For large undirected networks, adjacency matrices could be used: these would show every node in a row and column; better yet, there are no data occlusions. Unfortunately using these also increases the difficulty in understanding relationships. When applying diagrams to represent the human genome, it would be challenging since the length of it is approximately 3 billion bases. Therefore, the genome would be divided into chunks to make the data more manageable. Then the genomic data would be displayed in either the accordion view, Hilbert-Curve display or as stacks of the region with a center combined with a statistical plot. Genomic data could also be represented structurally even when the DNA sequence has a deviation of one kilobase from the reference sequence. To depict the structural difference, the variant and reference sequences are needed as well as the utilization of breakpoints by using arcs in a linear layout. Although the components are necessary to display the data structurally in a straightforward manner, the design could just as easily fall victim to over-plotting. With all of these designs available to view genomic data, it is important to remember to make the design both excellent and effective in concisely telling the story. It is also up to the researcher's discretion to choose which design to use to display their data, keeping in mind that the design should emphasize the biological features of the data—making them the most obvious.

References

Gehlenborg, N. &. (2012, September). Into the third dimension. *Nature Methods,* 9(9), 213.

Gehlenborg,N. & Wong, B. (2012, April). Integrating data. *Nature Methods*, 9(4),315

Gehlenborg, N. & Wong, B. (2012, January). Networks. *Nature Methods*, 9, 115

Gehlenborg, N. & Wong, B. (2012, October). The power of the plane. *Nature Methods*, 9(10), 935

Wong, B. (2010a). Points of view: Color coding. *Nature Methods, 7*(8), 573-573. doi:10.1038/nmeth0810-573

Wong, B. (2010b). Points of View: Gestalt principles (Part 1). *Nature Methods, 7*(11), 863-863. doi:10.1038/nmeth1110-863

Wong, B. (2011b). Points of view: Layout. *Nature Methods, 8*(10), 783-783. doi:10.1038/nmeth.1711

Wong, B. (2011c). Points of view: Salience to relevance. *Nature Methods, 8*(11), 889-889. doi:10.1038/nmeth.1762

Wong, B. (2011d). Points of view: The design process. *Nature Methods, 8*(12), 987-987. doi:10.1038/nmeth.1783

Wong, B., & Gehlenborg, N. (2012, February 28). Points of view: Heat maps. Retrieved from http://www.nature.com/nmeth/journal/v9/n3/full/nmeth.1902.html

Nominating Faculty: Professor Evgenia Giannopoulou, Bioinformatics 3352, Department of Biological Sciences, School of Arts and Sciences, New York City College of Technology, CUNY.

Cite as: Esteves, K.J., Jaikarran, U., & Luong, S. (2018). Visualization in bioinformatics. *City Tech Writer, 13,* 41-46. Online at https://openlab.citytech.cuny.edu/city-tech-writer-sampler/