TechNews

THE ART OF BIG DATA

Sarah Webb explores the processes, challenges, and skills involved in visualizing large biological data sets.

Credit: Martin Krzywinski

The sheer size and complexity of today's biological data sets can be overwhelming. In fact, the human brain often can't process the information without assistance. However, those terabytes of data are teeming with information and interactions that could lead to novel hypotheses, new drug targets, and a more thorough understanding of biological systems. This is where visualization can play a vital role, according to Jan Aerts, a data visualization researcher at KU Leuven in Belgium, by changing a cognitive task into a perceptual task.

Such transformations are far easier said than done, though. Visualization researchers face a constant challenge as biological data sets continue to grow and change. The complexity and inherent uncertainty of many data sets are often at odds with the core principles of good visual design, such as simplicity and symmetry. The problem lies in both the size and heterogeneity of the data, says Nils Gehlenborg of Harvard Medical School. The human genome contains billions of bases that researchers might want to view at the chromosome, gene, or even base-pair level. And there can be additional types of data associated with specific parts of the genome. For example, cancer genomics data primarily consist of DNA sequences but can also include metadata on patient demographics and health—sex, age, tumor type, and a whole host of potentially interesting information from electronic medical records.

Visualization researchers bring together skills in bioinformatics, software development, cognitive science, and design to streamline the ways in which researchers interact with and interpret their data sets (see “A non-linear career path” below). The best visualizations methods and tools, although they can seem simple at first glance, are often the product of many iterations of brainstorming, design, and development.

Browsing genomes and beyond

With the advent of genomics, online genome browsers have emerged to help researchers examine data, look for patterns, and build hypotheses. Many such tools are available, each with slightly different features or focusing on different genomes. The University of California Santa Cruz (UCSC) Genome Browser has been available since 2000 and is a tool for exploring the human genome, various vertebrate genomes, and genomes of other model organisms (<http://genome.ucsc.edu/>). The core goal is to give researchers easy access to as much data as possible, says Associate Director Robert Kuhn, just one of nearly two dozen scientists who work on the UCSC Genome Browser. A biochemist and molecular biologist by training, Kuhn joined the browser staff in 2003 to help make the interface more user-friendly.

At the chromosome level, individual genes show up as scattered tick marks in the Genome Browser. As users zoom in on a gene, they can see various isoforms to learn how it is spliced. Darker colors indicate which isoforms are supported by more experimental evidence, boxes indicate exons, and arrows show the direction of transcription. Zooming further, dark and light stripes show the locations of specific codons.

Quality assurance is essential with such a widely used tool, Kuhn says. Any problems with data or display are sent immediately to a quality engineer, sometimes before users even report any issues. “We hate it when things are broken,” he says.

One technical challenge for the UCSC Genome Browser team has been keeping up with the exponential growth of genomics data. To manage this, they have automated data updates as much as possible, Kuhn says. For ex-

ample, every week an automated process downloads RNA data from the National Center for Biotechnology Information (NCBI) that hasn't been annotated in the browser yet. And the browser can point to and integrate data hosted at other locations as well, such as information on mouse knockouts from the Jackson Laboratory in Bar Harbor, Maine.

The ability to use the browser at other websites allows new projects to build on the UCSC Genome Browser's tools. While a postdoc at UCSC, Ting Wang worked on the Genome Browser and led a spin-off project that eventually became the UCSC Cancer Genomics Browser (<http://genome-cancer.ucsc.edu/proj/site/hgHeatmap/>). As a faculty member at Washington University in St. Louis, Wang has led an effort to build an epigenome browser, VizHub (<http://epigenomegateway.wustl.edu/>), as part of the larger Roadmap Epigenomics Project (www.roadmapepigenomics.org/), which currently hosts roughly 25,000 epigenetics data sets. The VizHub browser incorporates a meta-data color map so that users can aggregate samples based on additional factors such as tissue type. Researchers can also arrange genomics data nonlinearly to look at 3-D interactions between bases that might be megabases apart.

Jan Aerts of KU Leuven in Belgium says that good data visualization changes a cognitive task into a perceptual one. Credit: Courtesy of Jan Aerts.

A non-linear career path

Biological data visualization brings together a set of eclectic and interdisciplinary skills: software development, bioinformatics, and programming, along with an understanding of cognitive science, art and design. Few researchers start out with expertise in all of these areas, says Nils Gehlenborg. Overall, he thinks the computational skills are an important foundation and that trainees can pick up other pieces of the visualization puzzle, particularly art and design, as they work.

In practice, visualization researchers acquire these skills in a variety of ways. Gehlenborg and Lennart Martens started out in bioinformatics. Originally a plant scientist, Jan Aerts then moved into genomics. He first delved

Credit: Martin Krzywinski.

into visualization trying to gain insight in complex population-genomics data sets at the Sanger Institute.

Martin Krzywinski studied biochemistry as an undergraduate, followed by graduate studies in physics and math. He moved into visualization research at the Genome Sciences Centre in Vancouver, Canada after starting out as a systems administrator. Being a physicist rather than a biologist, he says, has allowed him to look for the simplicity within the data.--SW

appliedbiosystems

Uniquely designed for when precision is paramount

The new Applied Biosystems™ MicroAmp™ 8-Tube Strip with attached optical or domed caps offers a combination of features designed to lead to a successful PCR/qPCR result every time.

- Individually attached caps
- Etched labeling on individual tubes
- Dual side tabs for strip labeling
- Graduated 20 µL measuring markers

Available exclusively for Applied Biosystems™ thermal cyclers to provide optimal PCR results.

Take control of your PCR reactions at thermofisher.com/microamptubestripwithcaps

ThermoFisher
SCIENTIFIC

For Research Use Only. Not for use in diagnostic procedures. © 2010 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified. D90411461 0410

Martin Krzywinski and his colleagues introduced hive plots like these in 2011 as a more reproducible way for visualizing networks. These plots compare the sizes of regions within the human, mouse, and dog genome assemblies. Credit: Martin Krzywinski.

Even with the variety of browsers that are available, as their data and features grow, researchers can't always make the most of them without additional training. Initially, users find a way to use a browser for a specific purpose, but there might be a range of other features that they didn't know are available, Kuhn says. In recent years, his job has increasingly included leading workshops to teach people how to locate data and use the expanded features within the browser. "A flip side of that is not really knowing what data are available," he says. "They don't necessarily know where to look or what to look for." The browser staff is constantly looking for new ways to make data more accessible.

Probing the proteome

Proteomics also presents visualization challenges, according to Lennart Martens of Ghent University in Belgium. Similar to genomics databases, mass spectrometry databases are potential treasure troves for discovering novel interactions and generating new hypotheses he says. He estimates that the European Bioinformatics Institute's PRIDE database (www.ebi.ac.uk/pride/archive/) contains approximately 1 billion mass spectra at the moment, 70 percent of which haven't been identified. Clearly, not all of those spectra represent novel biochemistry, he says, but even if just 1 percent do, that's 7 million.

Martens, a bioinformatician, has focused much of his work on proteomics and, recently, metabolomics, trying to find better ways to represent mass spectrometry data both from newly generated data sets as well as from those that are publicly avail-

Tips for better data visualizations

Martin Krzywinski of the Genome Sciences Centre in Vancouver, Canada, who focuses on visualizing research results, says he gets frustrated with representations of data in research papers. Far too frequently the data visualizations are busy, difficult to follow, and do a poor job of communicating the concepts that the authors are trying to convey. It takes both experience and time to construct good data visualizations for large data sets, says Krzywinski, and it's a skill that scientists have to hone.

For researchers looking for a few practical tips that can immediately improve their data visualizations, Krzywinski has these suggestions.

1. Pay attention to typography and spacing.

The type style, type face and font size can communicate hierarchy and emphasis in your data set, he notes. If you use alignment and negative space to help group information and ideas, the rest can fall into place relatively quickly.

2. When using color, think blue.

Scientists tend to gravitate toward colors they like, but a rainbow palette can bias a reader's view of your data. Reds and greens can be poor choices for readers who are color-blind. In addition, the human eye doesn't perceive all color changes in the same way—a small change at the red end of spectrum can appear large, like a change from red to yellow, while that same change within the green palette would still look green. Bluer palettes, long used in cartography, are better choices, he says.

3. Find a go-to piece of imaging software and become adept at using it.

Too much biology visualization work is done in Microsoft Excel, a program with limited options for displaying data, Krzywinski says. There's a dearth of good options, he says, but he likes Origin by OriginLab, a PC-based tool. That program has good default settings that give researchers extra features such as more plotting options, ways to overlay data, axis breaks, and ways to manage labels. If one person in your group can become a competent user and help others, he says, "that can bootstrap your whole lab into more effective visualizations."

Just as language provides rules that facilitate written and verbal communication, visualization has its own "grammar," notes Krzywinski. There's little formal coursework to help biology researchers gain basic visualization fluency, but experts say that type of understanding is essential. "Everyone is responsible for doing a better job," he says. And simple improvements in data management and presentation can become standard laboratory practices, he stresses, much like using lab-coats, washing hands, and maintaining sterility.--SW

able. He concedes that it's often a struggle to condense many combinations of molecular fragmentation patterns in intuitive ways. "We don't always succeed," he says, "You can't infinitely condense this information."

For example, he described a visualization method that his team developed for examining protein fragmentation consistency. They wondered: If a peptide appears in multiple spectra taken at different times, will the fragmentation patterns always look the same? Such a comparison is relatively easy with two or even a handful of spectra. "But imagine if you had a thousand." Instead of a traditional spectrum that compares intensity on the y-axis versus mass/charge ratio (m/z) on the x-axis, they changed the y-axis to represent mass error. The size of circles indicates the mass intensity, and the position on the y-axis indicates the deviation between the recorded mass and the calculated mass. What's clear in the two plots—one including just two spectra and the other including 990—is that the fragments observed remain consistent, but the intensity and mass error vary widely.

Visualization design and development

Visualization researchers don't design tools such as genome browsers just for problems that affect large groups of users; visualization can also be an incredibly useful tool for exploring new relationships in "the long tail" of large data sets for specific questions. The idea is to help researchers identify patterns in their data that they weren't necessarily looking for, which can lead to new avenues of research

One challenge is that such tools can take a long time to develop. Visualization researchers must first understand the users and their needs, and then they have to understand the data sets and potential relationships in the data. Sometimes, gaining that mutual understanding can require a lot of back-and-forth discussion. "The problem is that biology is so messy," says Aerts, which can be challenging for someone who approaches visualization from a background in engineering or computer science. And sometimes it's not easy to get biologists—the end users—thinking beyond today's technological constraints.

Researchers begin to design typically by starting with ideas on pen and paper before moving to the computer, Aerts says. The true difficulty is harnessing new design possibilities outside of the conventional options such as bar charts and scatter plots. And then it's a highly iterative process of going back to the researcher with ideas, getting feedback, and refining the visualization.

Though that overall process is important, the field will need to develop new toolkits or other ways to navigate the process

This entropy plot (upper left) and the three bubble plots depict the variability in protein fragmentation patterns for a set of proteins. The entropy plot shows the overall degree of variability in the mass intensity of peptides for a group of proteins. Peptides represented on the lower left have very consistent mass intensities, while those in the upper right are the most chaotic. The three bubble plots visualize the fragmentation distribution for the three labeled proteins. HRC—the most consistent (lower left bubble plot)—shows fewer bubbles that are larger in size. LRC—the most chaotic (upper right bubble plot)—shows a large distribution in fragmentation intensity. Credit: Lennart Martens, Barsnes et al., *Proteomics*, 2011.

more quickly, says Gehlenborg. When researchers need to solve a problem, they can't wait a year to develop a visualization tool that works, he says.

Even though the importance of visualization is growing along with the size of biological data sets, basic instruction on building and interpreting data visualizations is not a part of most current biology training programs (see "Quick tips for better data visualizations" above). Much of the way we interpret images is based on culture and education, Martens says. Scatter plots and box plots are only familiar to individuals who have studied a lot of math. Other types of plots, such as bubble plots and radar plots, offer new ways to express ideas but almost always require explanation or take longer for a user to understand.

As a result, Martens says, researchers quickly move into different worlds in which they express their data using a particular visual metaphor that might not be familiar to others. For example, images of western blots can be confusing to outsiders who aren't accustomed to interpreting them quickly. "I think there we have a big problem that we can only solve by education," says Martens.

Written by Sarah Webb, Ph.D.

BioTechniques 61:107-112 (September 2016)
doi: 10.2144/000114448