Getting Into Visualization of Large Biological Data Sets

Martin Krzywinski*, Inanc Birol, Steven Jones, Marco Marra

Genome Sciences Center, BC Cancer Research Center, Vancouver, BC, Canada

Abstract

Creating impactful visualizations in biology requires focused and productive dialogue between computer scientists, the algorithm and implementation experts, and biologists, the data domain experts who formulate the hypotheses. The utility and longevity of a visualization method can be increased if the system is made independent of specific analysis methods, given flexibility to address a variety of data types and hypotheses and useful for both data exploration and communication. From experience in creating and maintaining Circos [1], the community standard for visualizing genome comparisons, and Hive Plots [2], a method for quantitative network visualization, we present strategies to facilitate this inter-disciplinary communication and identify critical components of visualizations designed for print.

Keywords: genome visualization, network visualization, genomics.

Index Terms: J.3 [Computer Applications]: Life and Medical Sciences—Biology and Genetics; I.3.6 [Computer Graphics]: Methodology and Techniques—Graphics data structures and data types

1 Sharing and Understanding Biological Data

Effective collaboration between biologists and computer scientists requires that both have familiarity with each other's skills, data and, importantly, context and biases indigenous to their respective fields. Awareness of factors that confound visualization of biological data sets is critical: the size of the data set (e.g. 1000 mammalian genomes), ambiguity (plurality of transcripts and the extent of their confirmation) and sparseness (~2% of the genome codes for protein).

1.1 The Role of the Biologist

When sharing complex data, the biologist should be clear about what meaningful information it can contain and the kind of patterns that might arise as a result. It is likely that the computer scientist will not be familiar with the data modality, its statistics or artefacts (e.g. sequencing read depth, reasons for non-uniform or missing coverage) and will seek a fundamental representation of the data (e.g. a series of integer values keyed by position, with regions with missing or biased sampling). The biologist should emphasize the connection between hypotheses and data patterns in order for the computer scientist to operationalize questions into graphical answers.

1.2 The Role of the Computer Scientist

The computer scientist should prepare for the major challenge being one of managing scale and annotation [3], rather than the complexity of data representation or lack of capable visual mappings. Though there are exceptions, such as in the case of networks [2,4,5], new visual forms are not usually needed. From among many state-of-the-art visualization methods, the best solutions are often the simplest ones.

A thorough understanding of the biologist's hypothesis must inform the design of the visualization, which should saliently distinguish positive and negative outcomes. A good system should be flexible to address many hypotheses and map statistical significance onto visual weight.

1.3 The Role of the Visualization

Visualizations can aid in data exploration or communication, and their target audience may be expert or non-specialist – these choices inform their design.

Visualizations to explore data should be interactive, provide a variety of information channels, and offer easy access to both overview and detail. The user must be able to compare and group data sets, by similarity or difference, and because the range of data values often varies, assess differences in distributions.

Visualizations meant for print that communicate results should be designed around a message, rather than data, and present conclusions saliently by highlighting important patterns. They may show the data in the background for context, but should do so if the emphasis on the message is retained. The printed medium is less forgiving than an interactive space because the reader does not have the opportunity to manipulate the figure. Here, we will illustrate several concepts we deem essential in designing print visualization.

2 Properties of Effective Visualizations in Print

2.1 Generalizability

The system should be able to help visualize a variety of data types and remain agnostic to their meaning. It should support well-established formats (e.g. GFF, WIG), or have very simple input format requirements. The fewer assumptions and dependencies (e.g. on analysis) inherent in the system, the larger the number of users it can serve and problems that it can address.

The implementation should avoid excessive use of both jargon and abstractions. A balanced use of vocabulary specific to the field (e.g. read coverage) with terms that are general (e.g. data track) encourages users within the field to view their data more abstractly and lowers the barrier to entry for those outside it. For example, Circos refers to its chromosome representation as *ideograms* because these segments may represent other structures, such as sequence contigs, genes, or any scaled axis. One area in which Circos' design was not productive was the early semantic and syntactic distinction between cytogenetic bands and highlights, the latter being a more general data type with greater functionality. A better approach would have been to present a method of drawing bands as a specific highlight recipe – emphasizing that bands are a type of highlight.

2.2 Flexibility

The visualization should be flexible enough to address many hypotheses in the problem domain. The biologist must be able to tune the visualization to specifically address their hypothesis, or to selectively emphasize patterns to saliently present the results.

Our hive plots implementations ask that the user provide the node-to-axis and node position mapping. By giving the user

* martink@bcgsc.ca

control to represent their network in terms of meaningful properties, we avoid making any assumptions about what the user considers important. To get a user started, common recipes and sensible defaults are included. This is in contrast to conventional black-box layout algorithms which cannot be tuned to specific hypotheses.

Circos supports dynamic rules that affect how data is displayed based on data values and properties. Using rules, a user can adjust the display to show only the data they are interested in using meaningful filters (e.g. position, size, annotation) which are flexibly implemented (range checks, regular expressions, code snippets). Rules separate data from presentation and make it easy to focus the viewer's attention on a particular part of the figure (e.g. by hiding or fading other elements).

2.3 Transparency, Predictability and Robustness

The algorithm to generate the visualization should be easily explained to anyone familiar with the data. Positions, shapes and colors in the visualization should be based on properties of the data set, not heuristics driven by aesthetics.

Transparency and predictability are sorely missing in most network visualizations. For example, the result of a conventional layout algorithm is essentially impossible to predict or explain. Although in principle the layout should intuitively appeal to the viewer's understanding of the underlying connections, in practice this is not the case. It is common to see layouts that obfuscate patterns in structure (Fig 1), or create them where none exist. Comparing these layouts is impossible.

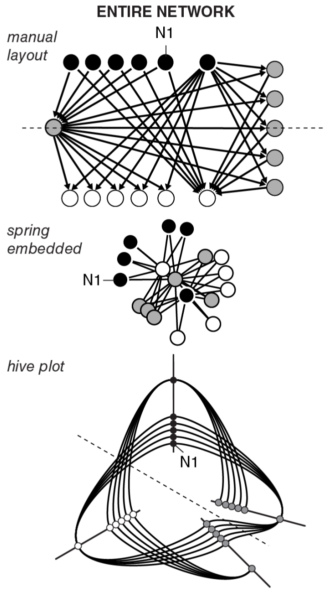

Figure 1: Hive Plots are a robust and quantitative visual form of networks based on meaningful properties chosen by the user. Unlike in a conventional layout, the removal of a node from a network can be easily spotted in a hive plot.

In addition to being transparent and predictable, visualizations must be robust with respect to the data. Changes in the data set should be reflected by proportionate changes in the visualization. Figure 1 demonstrates how the removal of a single node from a network drastically alters its spring embedded layout, but only mildly affects the hive plot.

We proposed the hive plot because conventional methods of network visualization lacked these fundamental properties, which we take for granted in other visualizations.

2.4 Scale, Sparseness and Adjacency

Biological data sets that annotate genomic sequence are typically high-resolution (changes at base pair level are meaningful), sparse (distances between changes are orders of magnitude greater than the affected areas) and connect distant regions by adjacency relationships (gene fusions and other rearrangements). It is difficult to take these properties into account on a fixed linear scale, the kind used by traditional genome browsers. In Circos we address this by a circular composition of axis segments (arguably the simplest modification of a linear layout) and by giving the user the ability to crop and order axis segments arbitrarily and apply a scale adjustment to a segment or portion thereof (Fig 2).

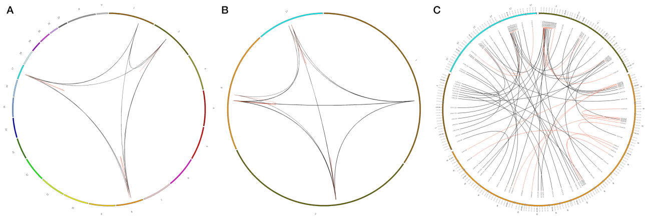

Figure 2: Patterns in densely distributed breakpoint positions between chromosomes 1, 2, 9 and 17 are emphasized when scale is zoomed. (A) full genome (B) chrs 1, 2, 9, 17 (C) 31 regions on chrs 1, 2, 9, 17 totaling 223 kb.

2.5 Perceptually Based Color Selection

When choosing colors, their perceptual characteristics must be taken into account. We encourage the use of Brewer palettes, which make color comparison quantitative and add positive aesthetic. The use of hue to encode magnitude must be avoided and palettes proofed against color blindness should be considered. Circos implements all the sequential and diverging Brewer palettes for its heat maps.

3 Conclusion

Visualization must be designed towards answering a hypothesis or delivering a core message and be sensitive to domain-specific nuances in the data, such as sparseness, variability and incompleteness. The visual mapping must show meaningful patterns saliently. New visual representations should replace those that fail to meet fundamental criteria, such as robustness.

References

[1] M. Krzywinski, *et al*., "Circos: an information aesthetic for comparative genomics," *Genome Res*, vol. 19, pp. 1639-45, Sep 2009.

[2] M. Krzywinski, *et al*., "Hive plots--rational approach to visualizing networks," *Brief Bioinform*, Dec 9 2011.

[3] M Meyer, *et al*., "Pathline: A Tool for Comparative Functional Genomics." *Computer Graphics Forum* (Proc. EuroVis 2010), 29(3):1043-1052.

[4] C. B. Nielsen, *et al*., "ABySS-Explorer: visualizing genome sequence assemblies," *IEEE Trans Vis Comput Graph*, vol. 15, pp. 881-8, Nov-Dec 2009.

[5] T. Munzner. "Applying information visualization principles to biological network displays," *Proc. SPIE* 7865, 78650D, 2011.