CellPress

Forum

Visualization: A Mind–Machine Interface for Discovery

Cydney B. Nielsen[1,*]

Computation is critical for enabling us to process data volumes and model data complexities that are unthinkable by manual means. However, we are far from automating the sense-making process. Human knowledge and reasoning are critical for discovery. Visualization offers a powerful interface between mind and machine that should be further exploited in future genome analysis tools.

Visualization for Discovery

As computer scientist Frederick Brooks put it 'a machine and a mind can beat a mind-imitating machine working by itself' [1]. He advocated for 'intelligence amplifying' systems that acknowledge the role of the human expert in the data interpretation process and enhance an individual's ability to reason about information. Visualization is a powerful modality for reasoning about data. First, it takes advantage of the pattern recognition capabilities of our visual system. By processing visual inputs in parallel, we can rapidly perceive complex patterns in plotted data that would be much slower if not impossible to perceive in tabular form (Figure 1A,B). Second, visualizations can reveal details that are hidden by computed statistics. This is perhaps most famously demonstrated by Anscombe's quartet [2]. These four sets of *x* and *y* values share identical summary statistics (i.e., mean of *x* values, mean of *y* values, variances, correlations, and regression lines); however they differ considerably when graphed (Figure 1B). Third, visualizations reduce the computational barrier to data analysis in particular as interactive interfaces, and empower the non-computational biologist to manipulate and investigate their data directly.

Visualization is perhaps best known for its role in scientific communication, which occurs once the discovery process is complete and the insights to be reported are well defined. However, the scientific process is an iterative cycle of generating hypotheses, data, and insights and visualization also plays an integral role in data interpretation and hypothesis generation (Figure 1C). Genome browsers are popular data exploration tools in genomics [3]. As an example from my own experience, a collaborator came across a striking localization of histone modifications within exon boundaries when inspecting human chromatin immunoprecipitation and sequencing (ChIP-seq) data in a genome browser. This observation sparked new hypotheses about the functional connections between chromatin structure and RNA processing. The initial exploratory visualization was itself not a final result, but rather the starting point for a systematic investigation of the enrichment of histone modifications across exons that was later published [4]. Such exploratory visualization is a powerful mechanism to deepen a researcher's understanding of the data and inspire new scientific ideas.

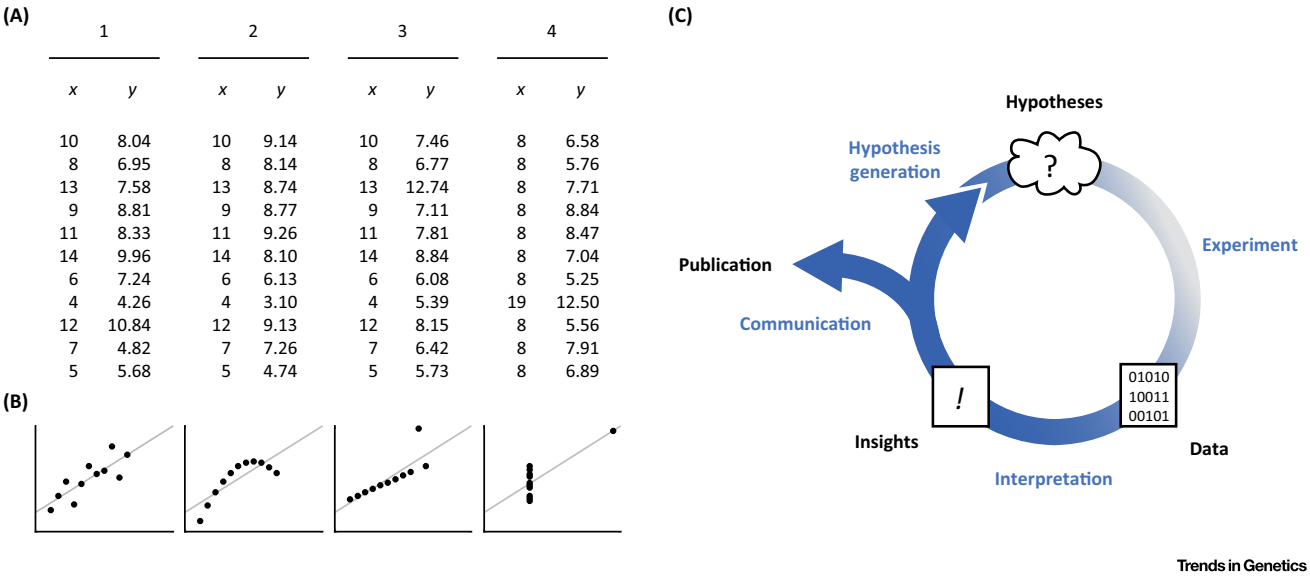

Figure 1. Visualization for Discovery. (A) Anscombe's quartet consists of four datasets of *x* and *y* values with the same mean of *x* values, mean of *y* values, variances, correlations, and regression lines [2]. (B) Visualization of the datasets in (A) with regression lines reveals very different patterns (adapted from [2]). (C) Visualizations play important roles throughout the scientific process (blue gradient), but are particularly powerful in enabling researchers to translate data into insights, to generate new hypotheses, and to communicate these findings to a broader community.

Improving Human–Computer Iteration to Accelerate Discovery

Both human and machine should tackle the tasks to which they are best suited. If a solution can be computed then it should be computed; visualization is a powerful approach whenever human input is required. Non-trivial amounts of analysis time are lost in the lag between computational and human processing. An interface that creates a seamless hand-off between the two results in greater efficiency in data interpretation, which today is very much a bottleneck. One approach is to equip the visualization tool with interactive computing capabilities. For example, EpiViz [5] creates a direct link between an interactive R computing session and a web-based genome visualization environment, thus reducing the iteration time between computational adjustment and visualization of results. Alternatively, the computation can be triggered from within the visual interface itself, for example, in Trackster [6], which allows researchers to interactively run computational tools using a range of parameters and to visually compare the outputs. Another approach is to use computation to guide researchers towards interesting data patterns. For example, StratomeX [7] enables visual comparison of patient stratifications based on different datasets (e.g., copy number profile, mRNA expression) and includes a query wizard to assist researchers in discovering interesting stratifications. Tools that reduce the number and time of analysis iterations between human and computer will go far in accelerating discovery.

With Great Power Comes Great Responsibility

Like any tool, visualizations can be misused. They can mislead instead of enlighten if not applied correctly. A common concern with making interactive visualization tools readily accessible to a broad community is that they open the door to ill-informed analyses conducted by novice users. For example, visualizations can facilitate cherry picking data patterns, such as correlations, that are consistent with a researcher's hypothesis. Some visualization tools report a *P* value for the significance of an observed effect; however, these do not consider the multiple testing done by the user in visually inspecting perhaps tens or hundreds of cases before finding the desired pattern and thus these handpicked effects are prone to be false positives. The problem in this example is that a tool intended for hypothesis testing is being used for data exploration. The onus falls on both tool users and designers. First, researchers who use statistical tests need to understand the conditions under which those tests are valid. Greater education in data analysis and statistics within the biological sciences is critical as large data become routine outputs of biological experiments. Second, tool designers must carefully consider the goals of the visualization and apply appropriate techniques. For example, they should use caution in reporting *P* values in an interface that promotes repeated querying, particularly if there is no tractable way to correct for multiple testing. They also need to consider their choice of statistic. Halsey *et al*. [8] recently discussed the fact that the *P* value exhibits wide sample-to-sample variability unless statistical power is very high. Instead, they advocate for more robust effect size estimates with confidence intervals. Third, it is our role as colleagues, supervisors, and reviewers to apply the same scientific rigor to data visualizations as to any other result. An attractive graphic with a *P* value is not the end of the story. The effects we observe need to be subjected to thorough independent validation such that spurious cherry picked observations are never treated as final results.

Figure 2. Genomics Pattern Discovery with Spark. In step 1, the user's genomic regions of interest and input data, such as those produced by ChIP-seq experiments, are preprocessed to enable rapid data retrieval in later steps. (Gray) Data enrichment peaks for two data samples; (vertical black boxes) user's regions of interest (r1–r5) centered on transcriptional start sites (TSSs). A data matrix is extracted for each input region and oriented according to strand. Rows in these matrices correspond to data samples, while the columns represent data bins along the genomic *x* axis; two bins per region are used in this diagram. The values are then normalized to be between 0 and 1, represented here by white and dark blue, respectively. In step 2, the matrices are clustered. $k = 2$ in this diagram, resulting in two clusters (c1 and c2). In step 3, the clusters and their region members are viewed in the Spark interactive visualization interface (figure adapted from [9]).

Future Opportunities: Beyond the Genome Browser

Great opportunities exist for innovation in genomic visual representation, interface flexibility, and scalability. First, there are vast possibilities for visualizations to move beyond the linear genomic axis. The urge to plot data directly is strong yet suboptimal for many analysis tasks. For example, genome browsers visually represent genome-aligned data verbatim, which is powerful for detailed inspection of an individual locus, but challenging for spotting patterns across loci. Spark [9] offers an alternative view that uses clustering across regions of interest to reorganize the genome into sets of related data patterns (Figure 2). This is one example of an alternative genomic view inspired by a researcher's analytic needs, but the area remains critically underexplored. Second, most genomic visualization tools address one or two analysis tasks and the abundance of specialized tools leads to interoperability challenges. Although it is unlikely that we will ever build a single monolithic tool that serves all our visualization needs, there is room for more customizable interfaces. For example, Tableau is a commercial product (http://www.tableau.com) that allows users to flexibly build and share combinations of simple interactive plots as web dashboards that are coupled to diverse data resources (e.g., spreadsheets, relational databases, cloud services). Similar principles could be applied to genomic visualization. Third, as genomic data resources continue to grow, visualization tools will need to tackle the challenges of scale. This involves designing new ways of visually integrating diverse genome-wide data types measured from potentially hundreds or thousands of samples. Graph structures that better capture the range of observed human genetic variation [10] provide an example in this direction. Moving forward, we should not be satisfied with the visualization tools on hand; rather we need to push our imagination and exploit the power of computation to build new visual interfaces that inspire hypotheses and deepen our understanding of genomic data.

Acknowledgments

I would like to thank the Canadian Cancer Society Research Institute and the Genome BC/Genome Canada Bioinformatics and Computational Biology program for research funding. I am grateful to the anonymous reviewers and the following colleagues for their valuable feedback on versions of this manuscript: Paul Boutros, Nils Gehlenborg, Martin Krzywinski, Sohrab Shah, and Bang Wong.

[1]Department of Pathology and Laboratory Medicine, University of British Columbia, British Columbia Cancer Agency, 675 West 10th Avenue, 4th floor, Vancouver, BC, V5Z 1L3, Canada

*Correspondence: cnielsen@bccrc.ca (C.B. Nielsen).

http://dx.doi.org/10.1016/j.tig.2015.12.002

References

1. Brooks, F.P. (1996) The computer scientist as Toolsmith II. *Commun. ACM* 39, 61–68
2. Anscombe, F.J. (1973) Graphs in statistical analysis. *Am. Statist.* 27, 17–21
3. Nielsen, C.B. *et al.* (2010) Visualizing genomes: techniques and challenges. *Nat. Methods* 7, S5–S15
4. Spies, N. *et al.* (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* 36, 245–254
5. Chelaru, F. *et al.* (2014) EpiViz: interactive visual analytics for functional genomics data. *Nat. Methods* 11, 938–940
6. Goecks, J. *et al.* (2012) NGS analyses by visualization with Trackster. *Nat. Biotechnol.* 30, 1036–1039
7. Streit, M. *et al.* (2014) Guided visual exploration of genomic stratifications in cancer. *Nat. Methods* 11, 884–885
8. Halsey, L.G. *et al.* (2015) The fickle P value generates irreproducible results. *Nat. Methods* 12, 179–185
9. Nielsen, C.B. *et al.* (2012) Spark: a navigational paradigm for genomic data exploration. *Genome Res.* 22, 2262–2269
10. Nguyen, N. *et al.* (2015) Building a pan-genome reference for a population. *J. Comput. Biol.* 22, 387–401