Poster: Perceptual Principles for Scalable Sequence Alignment Visualization

Danielle Albers*
University of Wisconsin - Madison

Michael Gleicher†
University of Wisconsin - Madison

ABSTRACT

Sequence alignment visualization is an important tool for understanding genomics data. Current approaches have difficulty scaling to the larger data sets becoming available. In this work, we survey recent results from perceptual science and show how they provide ideas for creating more scalable alignment visualization tools. We identify several principles, discuss how they inform alignment visualization design, and show their relevance to the limitations of current approaches. We describe how these principles are used to inform the design of an alignment visualization prototype.

Index Terms: J.3.1 [Computer Applications]: Life and Medical Sciences—Biology and Genetics;

1 INTRODUCTION

Sequence comparison is a fundamental task in the biological sciences. Scientists often need to understand the similarities and differences between genetic sequences to understand evolution, to infer common function, or to identify differences. Because the sequences are too long for manual examination, scientists rely on alignment tools that automatically identify subsequences that match between the sequences being compared. Numerous approaches for displaying and exploring alignments exist and have been incorporated into a wide variety of tools, see [6] for a survey.

The quantity and complexity of genomics data is growing rapidly, increasing the challenges for data interpretation tools. Most obviously, tools must handle longer genomes and comparison between more genomes. Alignments might contain many-to-many correspondences, probabilistic information, and other types of complex data. Developing methods for presenting this data is difficult: systems will need to present vast amounts of information to the user. Different implementations currently target the problem at different scales. The Broad Viral Viewer [4] provides for rapid comparison of dozens of smaller (virus-sized) genomes; Mauve (Figure 2) [3] is useful for a half-dozen or so medium (bacteria-sized) genomes, while Mizbee [5] supports pairwise comparisons of larger genomes.

Our conversations with users suggest that existing tools break down as the interpretation problems scale up: the displays become too complex to be interpreted effectively. Some scalability limits are explicit in system designs or come from engineering issues. But often scalability is hindered because the visual design breaks down. This is not surprising as the human visual system is limited in the amount of information it can take in. By understanding these limits, we hope to design systems that can work within them.

Our goal is to build new visualization tools that will better scale to the challenges of modern genomics. We aim to provide overviews of large data sets that preserve complex relationships in the data, even if those overviews are simply used to guide exploration in existing tools. As a first step in this direction, we have been studying how results from perceptual science can inform the design of alignment visualization tools. This poster surveys perceptual principles and how they can inform the design of alignment visualization approaches, either in understanding scalability limitations or in suggesting ways to improve the current tools. We introduce initial efforts at using these ideas to design large-scale overview visualizations.

*e-mail: dalbers@cs.wisc.edu
†e-mail: gleicher@cs.wisc.edu

2 LARGE-SCALE OVERVIEW VISUALIZATION

We have developed a prototype alignment tool for large-scale overview visualizations. The prototype design is based on traditional parallel viewers and the perceptual principles discussed in this work. Horizontal position can be mapped to various properties of aligned regions, such as start position or ordering within the source genome. Vertical position is mapped to the source genome. Color can encode local and global properties of the aligned regions. Highlighting with brushing and selective ribboning are both used to associate matching aligned blocks. Horizontal aggregation groups blocks that are too small to effectively interpret individually.

3 PERCEPTUAL PRINCIPLES

Perceptual science has explored the capabilities and limitations of the human visual system. Here, we survey some relevant results and discuss how they lead to scalability issues in Mauve (Figure 2), and how we have used them to inform the design of our prototype.

3.1 Pre-Attentive Phenomena

Pre-attentive phenomena allow a viewer to very rapidly identify targets in cluttered environments. Manipulating pre-attentive features within an image simplifies visual search by making certain groups of objects "pop out." When the system knows what the user is trying to find, it can use pre-attentive cues to highlight the targets. However, care must be taken to use such cues effectively and to avoid unwanted pop-out.

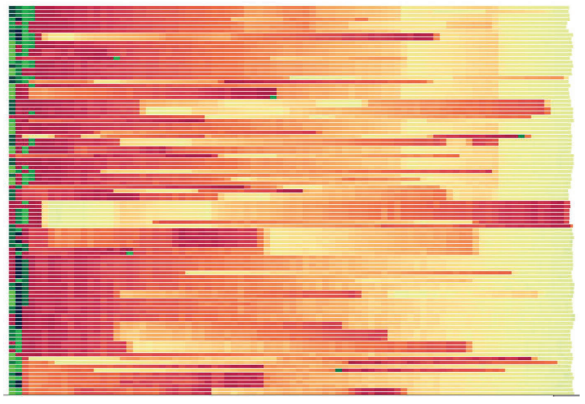

Figure 1: The current prototype visualization tool displaying 100 genomes, each composed of approximately 5,000 genes aggregated into perceptually cooperative units.

The impact of pre-attentive phenomena is well-known in the visualization community and key to informing the design of visualization systems. Current systems often take advantage of pre-attentive phenomena via highlighting and color schemes. However, the use of such schemes must be done with caution. Mauve, employs a color scheme that causes pre-attentive association of unrelated regions. Our prototype avoids false pre-attentive association through color choice. Our encodings allow for pre-attentive pattern finding and summarization: large fields of colors can be matched and texture patterns suggest sequence events.

3.2 Pre-Search Processing

Pre-search processing initiates the visual search process by collecting information to guide search. During the initial processing of a scene, pre-search processes gather structural and feature-based pre-attentive information to develop a contextual map of a scene. Blending structural scaffolding with feature data results in more rapid identification of regions of interest and more efficient search.

Alignment visualization can better facilitate pre-search processing by encoding significant details in the low-resolution properties of the visual encodings. This ensures that pre-processing mechanisms can readily identify objects of importance during early visual exploration. In Mauve, the tangle of lines cannot be effectively interpreted pre-attentively, thereby obscuring the structural scaffolding of the visualization. However, in our prototype, features of large regions, such as the quantity of sequence features or matches of interest, can be determined pre-attentively, helping the viewer determine where to direct their attention.

3.3 Visual Search

Visual search occurs when a viewer cannot find targets pre-attentively and must scan their attention over the scene to search for targets. Without perceptual aid, search tasks can be cognitively heavy and time-consuming [1]. By designing tools that cooperate with perceptual search mechanisms, users are better able to process the data for more rapid and efficient visual search.

Visual search is key in constructing scalable alignment visualizations. Users are generally comfortable searching displays using traditional reading orders. In tools like Mauve, the synteny lines impose a non-linear reading order, forcing the user to follow synteny lines to search the data. In our prototype, we preserve a conventional reading order sorted according to user-defined preferences. This allows the user to methodically scan over the data in a logical ordering, thereby reducing the cognitive cost of visual search.

3.4 Visual Clutter

Visual clutter occurs when item quantity, encoding, or layout causes performance degradation for search tasks. Visual clutter impairs the perceptual system by bogging down cognitive processes and slowing visual search. In data-processing tasks like sequence alignment comparison, clutter reduction by adjusting semantic data granularity often proves far more effective than simply removing data, while still preserving the overall data set [7].

Synteny and parallel views like Mauve frequently fall victim to visual clutter at medium to large scales as ribboning techniques often create tangled webs of synteny lines (see Figure 2). Edge-bundling techniques can be helpful in reducing synteny clutter [5]. Our prototype becomes cluttered only when there is a high density of sequence events: the clutter effectively becomes a texture that signifies many rearrangements; it does not obscure other regions. We also use interaction to help the user see through clutter.

3.5 Summarization

Summarization is the ability to construct statistical summaries of non-attended regions [2]. Summarization phenomena offer an opportunity to rapidly provide overview information without requiring

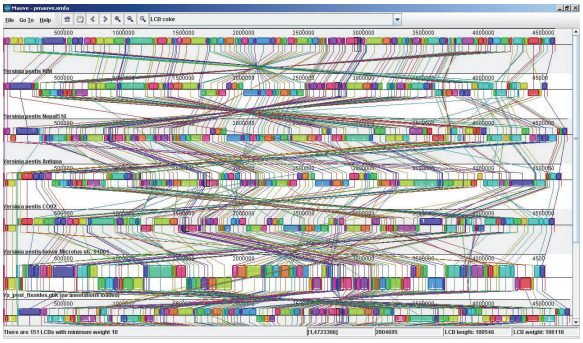

Figure 2: Mauve [3] is a successful tool for visualizing alignments among a small number of sequences, but breaks down as the number of sequences or the complexity of the relationships grow. Even the tool's authors feel it breaks down around 7 genomes. Much of the scalability concerns can be explained by perceptual principles.

the viewer's specific attention, which should be valuable in alignment visualization as often a scientist needs the context, not the details, of objects outside their immediate focus.

Given the large amounts of data in biological sequences, alignment visualization can take advantage of perceptual summarization by using visual encodings that provide a statistically accurate view of the data. The lack of a uniform color scheme in Mauve coupled with the structural irregularity of the synteny lines and inversions prevents summarization from conveying relevant data. By employing a perceptually uniform color scheme and regular structure in the prototype, we take advantage of summarization mechanisms to convey information about the data in non-attended regions. As a result, when blurred (which is what summarization effectively does), a Mauve view becomes a gray mass, while our prototype's views retain useful features, such as large areas and gradients.

4 Future Work

Our goal is to use these principles to develop an alignment visualization tool that provides meaningful overviews of large-scale genomic data sets. We intend to generalize the encoding mechanisms of the tool in order to accommodate a variety of user workflows. Furthermore, we would like to further incorporate the perceptual principles discussed in this survey in order to increase the overall scalability of the prototype and construct a fully scalable alignment visualization tool.

Acknowledgements: This project was supported in part by DoE Genomics:GTL and SciDAC Programs (DE-FG02-04ER25627) and NSF award IIS-0946598.

References

[1] G. A. Alvarez, T. Konkle, and A. Oliva. Searching in dynamic displays: Effects of configural predictability and spatiotemporal continuity. *J. Vis.*, 7(14):1–12, 12 2007.

[2] B. Balas, L. Nakano, and R. Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.*, 9(12):1–18, 11 2009.

[3] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*, 14(7):1394–1403, Jul 2004.

[4] D. Jen, L. Larson, C. Stolte, D. DeCaprio, T. Allen, B. Birren, M. Koehrsen, and M. Henn. Comparative viral genome visualization. IEEE InfoVis Poster Proceedings, 2009.

[5] M. Meyer, T. Munzner, and H. Pfister. Mizbee: A multiscale synteny browser. 15(6):897–904, Nov. 2009.

[6] J. B. Procter, J. Thompson, I. Letunic, C. Creevey, F. Jossinet, and G. Barton. Visualization of multiple alignments, phylogenies and gene family evolution. *Nature Methods*, 7(3):S16–S25, Mar. 2010.

[7] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *J Vis*, 7(2):17.1–1722, 2007.