Sequence Surveyor: Scalable Multiple Sequence Alignment Overview Visualization

Danielle Albers
Department of Computer Sciences
University of Wisconsin-Madison
dalbers@cs.wisc.edu

Colin Dewey
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
cdewey@biostat.wisc.edu

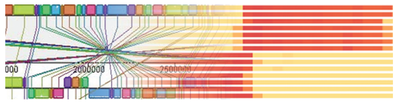
Michael Gleicher
Department of Computer Sciences
University of Wisconsin-Madison
gleicher@cs.wisc.edu

Sequence alignment visualization is an important tool for understanding genomics data. As sequencing techniques improve and more data becomes available, greater demand is being placed on existing tools to scale to the size of these new data sets. However, current tools do not scale to the challenges of growing data sets, as they focus on visualizing details instead of global trends. When viewing such large data, we necessarily cannot convey small details, rather we specifically design overview tools to help elucidate large scale patterns.

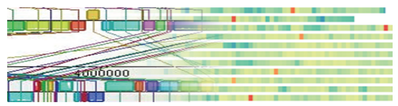
Perceptual science and signal processing theory provide a framework for the design of such visualizations that can scale well beyond current approaches. We present Sequence Surveyor, a prototype that embodies these ideas for scalable multiple sequence alignment overview visualization. We demonstrate how perceptual science and signal processing concepts can be used to support scalability in visualization and use these techniques to simultaneously visualize over 100 aligned genomic sequences.

Leveraging Perception

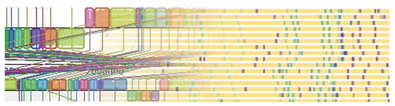
Perceptual science explores the capabilities and limitations of the human visual system.
To understand these limitations, we surveyed recent results from the perceptual literature and derived a set of five principles outlining how perception can guide the design of scalable visualization techniques.



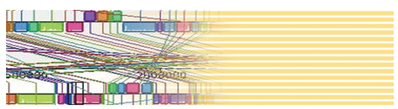
Visual search occurs when the user consciously scans their attention over the scene to identify regions of interest. Orthology lines can suggest an irregular exploration structure of the data. Sequence Surveyor uses color and position to encode orthology, preserving a standard reading order for exploration.



Pre-attentive phenomena suggest that certain elements of the visual field can "pop out" before a scene is consciously processed. In tools like Mauve, color correspondence does not always imply semantic correspondence. Sequence Surveyor's color mappings give color a semantic meaning, allowing pre-attentive mechanisms to orient the user as to interesting features in the view.



Visual clutter arises when the visual information in a given space bogs down cognitive processes. In cases of significant conservation, dense orthology lines can clutter the visual field. In Sequence Surveyor, clutter acts as a texture: cluttered regions identify portions of the data set with a high concentration of sequence events.



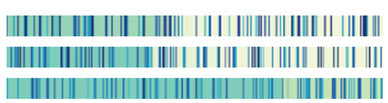
Pre-search processing analyzes the spatial and semantic structure of the scene to guide visual search. Orthology lines in regions of high conservation may obscure overall spatial structure; however, Sequence Surveyor uses matching color fields to encode similarity: large regions of color can be pre-attentively associated.

Summarization describes how the cognitive system processes unattended information. Summarization delivers a low-resolution version of the visual field: effectively a blurred version of the unattended scene. When Mauve is blurred, it becomes a gray mass while Sequence Surveyor preserves the overall patterns of the visual scene.

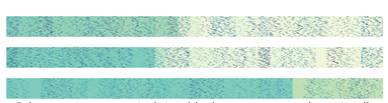
Aggregation Techniques

In genomic sequence data, there will generally be significantly more genes than pixels for a given row. Aggregation filters intelligently limit the amount of visual information while still preserving key properties of the data. Sequence Surveyor supports aggregation by grouping genes together into spatially-defined units called 'blocks'. The information displayed per block can be controlled using one of three aggregation filters.

Averaging maps the color of a block to the mean of the color values in that block, capturing general trends.

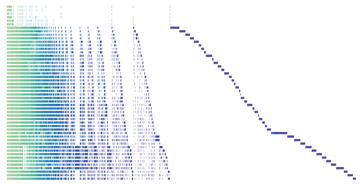


Event striping ensures that any values disrupting the data trend are mapped to a stripe, highlighting outliers in the data.



Color weaving maps pixels in a block to component data, visually approximating the distribution of data values within a block.

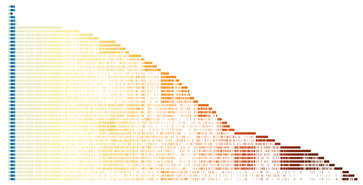
Flexible Exploration



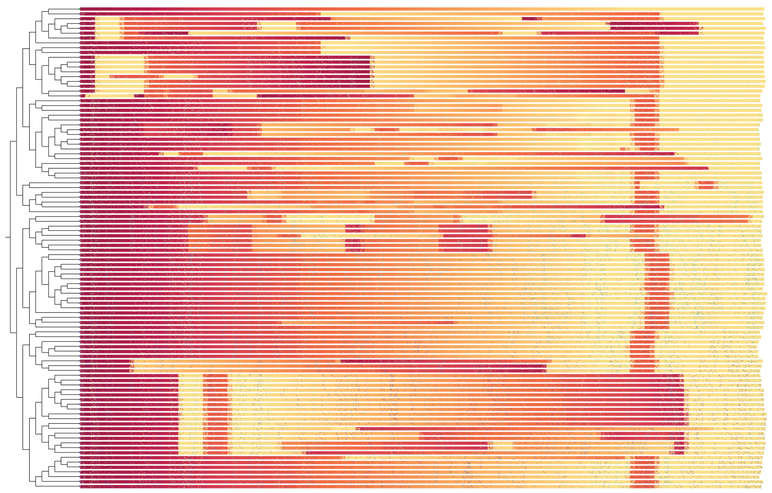
Sortings provide immediate visual clustering of genes with similar properties. The above visualization clusters commonly co-occuring genes by sorting by the genomes they occur in.

Sequence Surveyor assumes the user is the expert: they know what is"interesting" in their data set. This qualitative scalability is provided via user-controlled color and sorting mappings. Several such mapping combinations reflect existing genomic visualization tools at scale, while others present more unique views of the data.

Color reveals immediate visual trends and patterns for different properties among all elements of a data set. Coloring by position in reference shows whole genome orthology patterns.

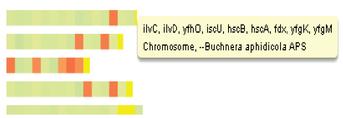


Both sorting and coloring by different reference genomes provides insight into divergence events such as those between organisms from the same set of parents.



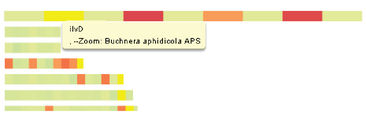
Sequence Surveyor applied to 100 genomes derived from an evolutionary simulation. The genomes are sorted by the local ordering of the genes within the genomes and colored according to their position in the first genome. Aggregation is composed via color weaving. The dendogram on the left displays the evolutionary relationships presented in the data.

Detail on Demand

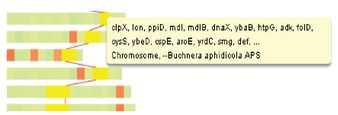


Global homology relationships are revealed via a brushing mechanism: mousing over a block highlights blocks containing genes orthologous to its component genes. Tooltips reveal the genes represented by the block.

Once the user has located a region of interest, they often want to explore more intricate details about that region. Sequence Surveyor supports high-level exploration of details within a data set through a set of interaction techniques.

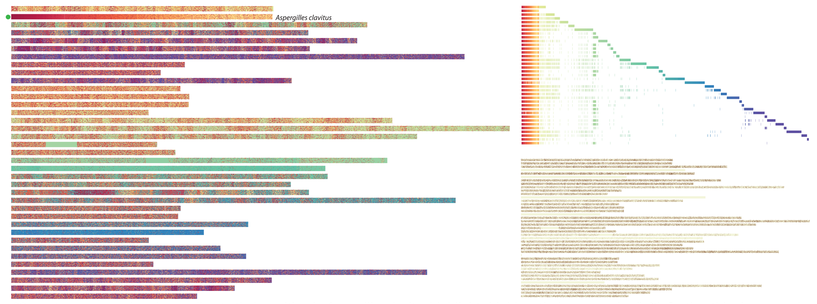


Due to non-locality of genes in sequences, traditional zooming hides important homology details in the data set. Sequence Surveyor uses overview+detail to zoom: zooming shows the genes of a block as a new sequence at the top sequence position.

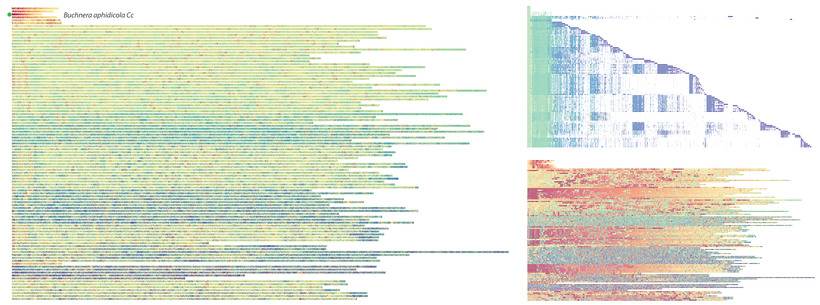


Homology ribboning can be applied to the visualization on demand. Selecting a block of interest visually links orthologous blocks in the visualization.

Visualizing Real-World Data



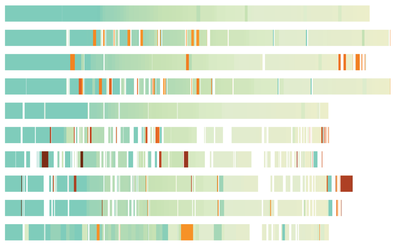
33 yeast genomes colored according to their position in *Aspergilles clavitus* highlights highly orthologous species (left). Sorting this view according to the position in this reference shows a significant set of differences between the orthologous gene groups in these related species (upper right). Color weaving by frequency reveals the extreme duplication (over 63,000 times) of a single orthologous gene group (lower right).



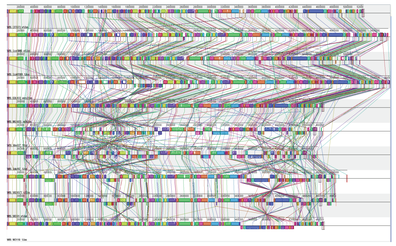
100 bacterial genomes colored according to their position in *Buchnera aphidicola Cc* , a bacterial genome that been reduced to an essential set of genes by evolutionary mechanisms. The proliferation of these genes in the data set is seen by the abundance of red and orange blocks throughout the scene (left). Sorting by this reference and coloring by grouped frequency reveals that, in general, this reduced set of genes is found in most genomes in the data set (upper right). Recoloring with respect to an *E. coli* genome highlights the similarities between the *E. coli, Shigella, Salmonella, Buchnera* and *Yersinia* genomes in the data set (lower right).

Comparison with Existing Tools

Mauve 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. Sequence Surveyor can mimic traditional Mauve views by coloring according to position in the first genome and sorting by start position. The below images show both the Mauve and Sequence Surveyor visualizations of ten *E. coli* genomes. The conservation trends represented by orthology lines in Mauve become large color fields in Sequence Surveyor. Unlinked regions in Mauve appear as warm-colored blocks, pre-attentively popping out of the visualization.



Sequence Surveyor on ten E. coli genomes.



Mauve on the same ten E. coli genomes.

Acknowledgments

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

References

Genome Evolution Laboratory. *Mauve: Multiple Genome Alignment*. http://asap.ahabs.wisc.edu/mauve/, 2010.