Sequence Surveyor: Leveraging Overview for Scalable Genomic Alignment Visualization

Danielle Albers, *Student Member, IEEE*, Colin Dewey, and Michael Gleicher, *Member, IEEE*

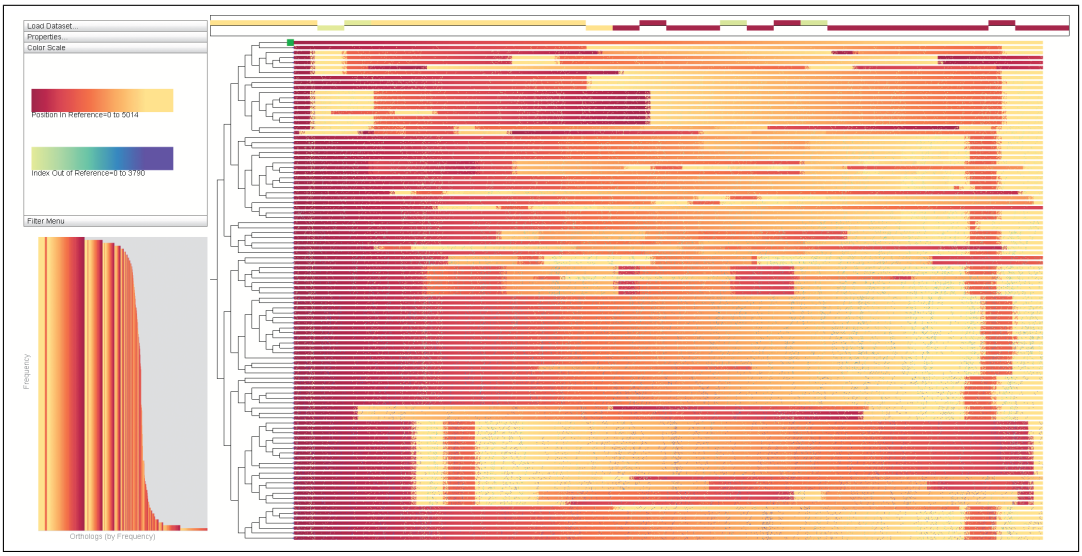


Fig. 1. Sequence Surveyor visualizing 100 synthetic genomes generated by an evolution simulation. Each genome is mapped to a row and genes are ordered by position. Color encodes the position of the gene within the chosen reference sequence (top row, indicated by the green box). Genes are aggregated, with each block's texture reflecting the overall distribution of colors in that block. The dendogram shows the phylogeny of the data set while the histogram shows the frequency distribution of orthology group sizes.

Abstract—In this paper, we introduce overview visualization tools for large-scale multiple genome alignment data. Genome alignment visualization and, more generally, sequence alignment visualization are an important tool for understanding genomic sequence data. As sequencing techniques improve and more data become available, greater demand is being placed on visualization tools to scale to the size of these new datasets. When viewing such large data, we necessarily cannot convey details, rather we specifically design overview tools to help elucidate large-scale patterns. Perceptual science, signal processing theory, and generality 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 whole-genome alignment overview visualization. Sequence Surveyor visualizes sequences in parallel, displaying data using variable color, position, and aggregation encodings. We demonstrate how perceptual science can inform the design of visualization techniques that remain visually manageable at scale and how signal processing concepts can inform aggregation schemes that highlight global trends, outliers, and overall data distributions as the problem scales. These techniques allow us to visualize alignments with over 100 whole bacterial-sized genomes.

Index Terms—Bioinformatics Visualization, Perception Theory, Scalability Issues, Visual Design.

1 Introduction

Sequence comparison is a fundamental task in the biological sciences. Scientists often need to compare genomic sequences, for example, to understand evolution, to infer common function, or to identify differences. Because sequences are often too long for manual examination, scientists rely on alignment tools that automatically identify matching subsequences. Tools for visualizing these alignments are commonly used when performing sequence comparison. A variety of approaches for displaying and exploring alignments exist, and have been incorporated into a wide variety of tools. Procter et al. [28] presents a recent survey of many popular tools.

The amount of sequence information available is growing rapidly. Scientists are exploring larger numbers of genomes and longer genomes. However, most tools by design focus on providing in-depth exploration of a small set of sequences for predefined tasks. Focusing on low-level details obscures the task of tracing high-level trends in large datasets (cf. Figure 10a). Looking at larger datasets at this fine level of detail is overwhelming, and does not scale to growing datasets.

In this paper, we introduce a different type of tool for exploring large multiple genome alignment datasets: overview visualization. Sequence Surveyor, our prototype system shown in Figure 1, provides flexible views of large datasets. It allows scientists to examine patterns and trends in multiple genome alignment datasets of unprecedented scale, such as a set of 100 bacteria genomes (Figures 11 and 13). Such large-scale overview comes at the expense of showing finer details: our design considers how visual principles suggest layouts that allow large-scale patterns to emerge, and includes abstraction mechanisms designed to retain salient features. Because we cannot know *a priori*

- *Danielle Albers is with University of Wisconsin - Madison, E-mail: dalbers@cs.wisc.edu.*
- *Colin Dewey is with University of Wisconsin - Madison, E-mail: cdewey@biostat.wisc.edu.*
- *Michael Gleicher is with University of Wisconsin - Madison, E-mail: gleicher@cs.wisc.edu.*

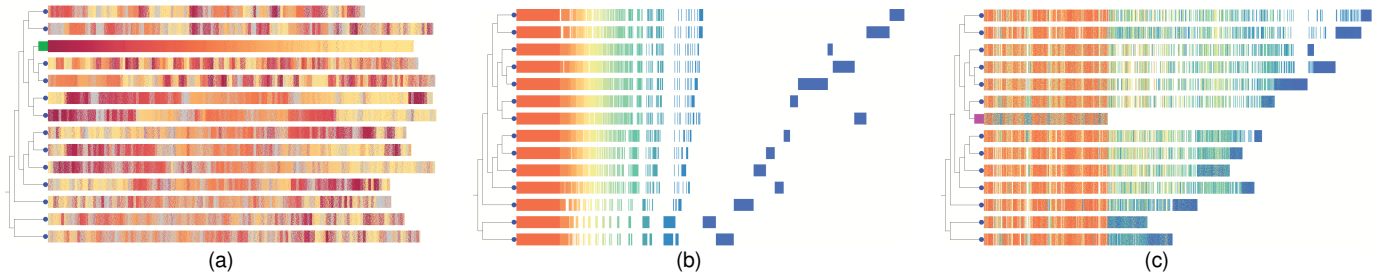


Fig. 3. Sequence Surveyor provides multiple color and position mappings that address different questions about data. (a) Coloring by the position of genes in a reference genome (green rectangle) shows that genomes most similar to the reference are not those most closely related, indicated by the preservation of the color ramp. (b) Membership frequency coloring and sorting by grouped frequency (the sets of genomes ortholog groups are conserved in) highlights patterns of presence and absence across species. Resulting bands of genes create conservation "fingerprints" for each genome that tend to line up best between the most closely related genomes. (c) Membership frequency coloring (most (red) to least frequent (blue)) and sorting according to position in a reference genome (magenta box) highlights uncommon regions of the reference: green columns in the reference show other species that also conserve these relatively unique regions.

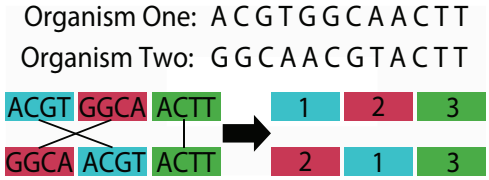


Fig. 2. Genome alignments are computed from genome sequence data by identifying matching subsequences (left), known as *orthologs*. Ortholog groups are identified by integer tags (right). Sequence Surveyor uses orthology data to explore genome alignments. In real data, orthologs are far longer than four nucleotides.

the kinds of questions the data will be used for, our approach provides flexible mappings that allow different kinds of patterns and trends to be made salient as the user explores the data. Mechanisms for filtering, zooming, and re-ordering the data help scientists find different kinds of large-scale features in the data and connect these to smaller sets of details for further exploration. While the focus of our work has been on genome alignment data, our tool can be used for other forms of sequence data, including traditional multiple sequence alignment tasks and applications in the Digital Humanities.

1.1 Biological Background

The primary task of alignment visualization involves viewing matching regions between a set of sequences. Alignment visualization is useful for many types of sequence data. While we focus on whole genomes (DNA sequences), the problems are similar for proteins and RNA. However, one important and complicating aspect of visualizing whole genome alignments is that there are potentially thousands of related elements which may occur in different orders and copy numbers in each genome. When trying to understand an alignment, a scientist often needs to consider other information such as the details of the sequences, annotation data, and expression information. As such, alignment visualization can include the issues involved in examining a single genome, many of which are surveyed in Peeters et al. [27].

The visualization of alignments is generally independent from the tools used to compute them. While pairwise alignments are most common, alignments between multiple sequences are becoming increasingly important as sequence information becomes more abundant and better understood. In this paper, we focus on visualizing whole-genome multiple alignments at the gene level. In such data, the DNA sequences are segmented into functional regions (i.e. genes), and each sequence is represented as an ordered list of genes (cf. Figure 2). Alignments identify groups of matching (evolutionarily-related) genes, known as *ortholog groups*, present in one or more genomes in the dataset. These groups, as computed by the alignment, serve as identifiers for related classes of genes. This type of data is technically called gene-level alignment, but, for the purposes of this paper, the more general term *alignment data* will be used. Details about the generation of the sample datasets is discussed in the supplement.

One of the primary tasks in analyzing whole genome alignment data is understanding patterns of conservation in the dataset. For the purposes of this discussion, conservation can be thought of as the preservation of orthologous genes between species. Understanding patterns in conservation can allow scientists to make conjectures about evolution and common function of different species. Orthology conservation can help answer questions about the conservation of genes at different loci in the genome, origins of replication (i.e. where rearrangements of genes between different species begin), and proportions of the genome shared between different organisms. These types of general questions make whole genome alignment data useful: by understanding conservation between genomes, we can begin to understand how different gene sequences function within an organism.

1.2 Solution Overview

Our approach to multiple genome alignment overview presents the data as horizontal stripes, with each row corresponding to a sequence. Color conveys attributes of the sequence data, yielding a dense field display. By selecting different mappings from the data onto the color field, different kinds of patterns are made salient and can be identified.

Mappings encode gene properties as horizontal position and color. A number of mappings are shown in Figure 4 and detailed in §3.2. Position encodings show sequences in order, left to right. Other views involve sorting orthologous genes by frequency (the number of genes matching it) or position in a selected reference genome. Similarly, genes may be colored according to their sequence position or attributes of their ortholog group. A user may experiment with different mappings and reference choices to find views that reveal interesting patterns. For example, positional ordering coupled with position in reference coloring identifies common genes and their rearrangements across the dataset, while ordering by frequency and coloring by position gives a sense of the conservation between sequences. Figure 3 illustrates some mapping combinations on real data.

Each sequence is displayed as a series of screen-space blocks, each containing a number of genes (see §3.3). Choices in how the contents of the blocks are aggregated help control the kinds of patterns that are visible: choices range from averaging the values in the blocks to emphasize overall trends to event striping displays that highlight outliers.

Various interaction techniques help to explore the sequence alignment views. Hovering the pointer over a block highlights blocks that share orthologs and provides a list of contained genes as a tool tip. A histogram of ortholog frequencies and a phylogenetic tree provide linked views that highlight subsets of the data. Zooming and detail displays help connect large patterns to specific details.

Contributions: Our overall contribution is the introduction of an approach to visualizing overviews of alignment data that allows for

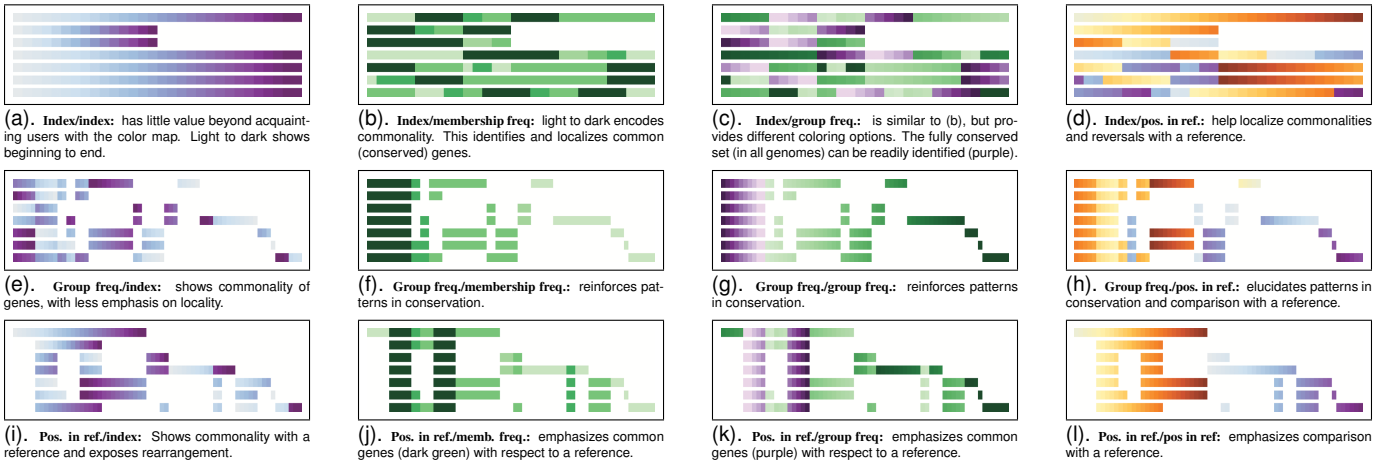


Fig. 4. Sequence Surveyor views shown on a toy dataset, each combining a position mapping and a color mapping. Different mappings make different patterns emerge in the color field. Subfigure rows show different position mappings, columns show color mappings (see subfigure captions). The top genome is the reference for both coloring and position. Nucleotide-level start position mappings do not apply in this example.

exploration of whole genome alignments at unprecedented scale.

In developing this approach, we make three specific contributions:

1. We show how perceptual principles confirm limitations in current alignment visualization approaches and also suggest a color-field design for alignment overview. While these principles are well known, they have not been applied to alignment tools.

2. We show how a set of flexible mappings can create a generalized display of alignment information as a color field. While some of the mappings have appeared before, we are the first to provide a general, flexible view of alignment data that can make different kinds of patterns and trends salient based on user control.

3. We provide a set of aggregation mechanisms for scaling large sequence data into limited space. User control over aggregation mechanisms can emphasize or suppress outliers.

These ideas are implemented in the Sequence Surveyor prototype, contributing a system useful not only for the genome alignment data that motivated its design, but also for other sequence visualization tasks such as those in the Digital Humanities (§5.3).

2 Related Work

2.1 Overview Visualization

Scalability presents an interesting challenge for designers: how can a visualization show large amounts of data in a way people can readily interpret. Fekete and Plaisant [10] discuss this challenge, recognizing the limitations of the perceptual system in processing information. They conclude that large-scale visualizations should leverage simplicity and only provide details and excess dimensionality on demand. This approach suggests the use of overview to handle large amounts of data. Shneiderman [30] explores scalable visualization over one billion data points. This work acknowledges that advancements in gigapixel displays offer scalability in terms of screen-space, but massive displays may fail to scale in terms of what the viewer can interpret. Aggregate visualizations instead approach this problem by leveraging available space through abstraction instead of showing exact values per point. We draw on both works, using aggregation for overview.

Many works use dense fields of color to encode information. Pixel-oriented visualization techniques [20] represent large datasets by mapping individual data values to pixels. These techniques explore the value of flexible encodings for such displays, an idea we adopt. Slingsby et al. [32] also discuss the value of flexible encodings. Chromogram [34] and Lasagna Plots [33] use color fields with a specialized encoding for sets of series, akin to our problems. None of these works have applied the color field design to sequence alignments or consider aggregation to scale to data sizes larger than the display.

2.2 Tools for Alignment Visualization

The importance of sequence comparison has led to numerous tools for alignment visualization. Meyer et al. [24] survey tasks seen in alignment visualization (albeit focusing on synteny browsing and pairwise alignments). Peeters et al. [27] survey general issues in sequence visualization, which has similar scaling concerns. They observe that different kinds of tasks occur at different scales. Neither work explicitly considers creating tools for exploring many large genomes.

Most alignment visualizations are variants of three basic designs: dot plots (scatter plots with sequence position on the axes), synteny views (which indicate matches relative to a reference), or parallel-coordinate views (which show alignment by drawing connections between sequences). Tools are suited to particular tasks and scales. For example, Viral Viewer [19] provides for comparison of dozens of viral genomes (viral genomes are small, and viral-scale viewers focus more toward nucleotide-level analysis problems). Mauve [7] is useful for a half-dozen or so medium (bacteria-sized) genomes, while Mizbee [24] supports pairwise comparisons of larger genomes.

Current approaches to alignment visualizations focus on preserving details in sequences, not providing insight into the overall trends in the data. Some scalability limits are explicit in system designs, for example comparing no more than two genomes, as in [9, 14, 18, 24]. Scalability limits may come from memory or performance issues as tools get bogged down with too much data. But often, scalability is hindered because the visual design breaks down: the displays simply become ineffective when there is too much data to display in detail. Tools such as [5, 7, 8, 12, 13, 15, 22, 23] do not explicitly discuss the scalability limitations of their approaches. In Sequence Surveyor, we focus on breaking scalability restrictions by limiting the level of detail explored in the visualization at any given time. This trade-off allows us to exchange detail for global patterns in the data.

Aggregation has been applied to create compact displays for sequence comparison. For instance, SequenceJuxtaposer [31] displays the majority base for groups of context bases and JalView [6] uses multiple coloring schemes to show the aggregate consensus. However, these tools operate on nucleotide-level alignments, changing the scale of the data retargetting problem: only four possible distinct entities must be displayed at the nucleotide level whereas Sequence Surveyor must visualize thousands of distinct ortholog groups. SequenceJuxtaposer and JalView also focus on local mutations and small rearrangements, which do not address the non-locality issues arising from whole genome alignment.

Genomicus [26] provides a visualization of a large number of possible genomic sequences in their phylogenetic context; however, this tool only visualizes regions orthologous to a particular region of some reference sequence, hiding any portions of the genome outside of this target region and losing information about the context and global position of orthologous gene groups in the dataset.

3 DESIGN

Our goal in the design of Sequence Surveyor is to create an alignment visualization tool able to scale to large numbers of genomes (dozens or more) and large genomes (thousands or more genes per sequence). At the same time, we must handle the full complexity of these alignments, including rearrangements, reference dependent and independent tasks, and gene repetition. Furthermore, the study of such massive datasets is new: the questions to be considered are wide-ranging and this display may offer the opportunity to discover new questions.

We focus on building a scalable overview tool. Exploring broad patterns in big datasets may come at the expense of providing the details usually shown in traditional tools. However, traditional tools can be used to examine details in subsets of the data identified in the overview, providing a multi-scale visualization approach.

A primary consideration was to use a visual design where patterns and trends can be viewed even in large-scale displays. An examination of perceptual principles (§3.1) suggests a color field design, rather than the designs more common in alignment visualizations. By encoding orthology using color instead of explicit connections, to some degree, we exchange accurate identification of individual connections for scalability. While color fields allow patterns and trends to "pop out" efficiently in large displays, this requires determining what should be made to pop out. As this is not known *a priori*, we instead define flexible mappings (§3.2) that allow user control and exploration. Similarly, we provide different schemes for aggregation (§3.3), not only allowing the system to scale to data sizes much larger than the number of pixels, but also controlling visual clutter and if trends or outliers are more significant to a particular exploration. Other aspects of the design include mechanisms for arranging the data for effective comparison (§3.4) and interaction techniques to aid exploration and connect to details (§3.5).

3.1 Leveraging Perception

A visual presentation of a large alignment dataset is necessarily complex. While the visual system can function in very complex environments, such as recognizing a scene or driving a car, this efficiency comes at the expense of flexibility needed to discover unknown patterns in novel environments. As the visual system is easily overwhelmed with many pieces of information [11], visualizations must rely on the kinds of patterns that emerge readily and support systematic search for details. While these mechanisms are known to the visualization community and often inform designs, they have not been considered in the design of large-scale alignment visualizations. In this section, we consider several perceptual principles relevant to large-scale genomic alignment visualization, and show how they explain the scale limits of an existing design and suggest a different design that better affords emergent patterns to be readily visible and systematic visual search for details. Specifically, we consider Mauve [7], as it is relatively successful at moderate scales and representative of a class of designs, and the color-field design used in Sequence Surveyor. Our survey suggests that while using color to encode orthology is not as exact as using connective encodings to identify orthologous matches, it scales to far larger datasets than traditional connective techniques.

3.1.1 Pre-Attentive Phenomena

Pre-attentive phenomena allow a viewer to 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 viewer is trying to find, it can use pre-attentive cues to highlight targets. Since pre-attentive cues can be processed in parallel, effective use of pre-attention can greatly ease the cognitive load of visual search tasks [17]. However, care must be taken to use such cues effectively and to avoid unwanted pop-out that may distract the viewer.

The impact of pre-attentive phenomena is well-known in the visualization community. 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 [7] employs a color scheme that causes pre-attentive association of unrelated regions: color can reinforce the orthology shown by connectivity, but does not encode it completely. Our design avoids false pre-attentive association through semantically informed 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 (see Figure 5a). For example, reversal of color ramps show inversions. Color mappings can be selected to make certain groups of genes pre-attentively stand out and help to reduce color aliasing by providing a wider range of color points for interpolation.

3.1.2 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 demanding and time-consuming [1]. By designing tools that cooperate with perceptual search mechanisms, users can more easily process the display for more rapid and efficient visual search. Pre-search processing pre-attentively collects structural and feature information to guide visual search.

Visual search is key in constructing scalable alignment visualizations. Establishing a pre-search contextual map can be achieved by encoding significant details in the low-resolution properties of visual encodings. This ensures that pre-processing mechanisms can readily identify objects of importance during early visual exploration: 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.

Users are generally comfortable searching displays using traditional reading orders [2]. However, orthology lines impose a non-linear reading order, forcing the viewer to follow these lines to search the data. Our design preserves a conventional reading order sorted according to viewer-defined preferences. This allows the viewer to methodically scan over the data in a logical ordering, thereby reducing the cognitive cost of visual search (see Figure 5b). To help guide systematic scanning in search, we provide horizontal guides (white lines between stripes) rather than a dense matrix of colors as often seen in heat maps.

3.1.3 Visual Clutter

Visual clutter occurs when item quantity, encoding, or layout hinders performance in search tasks. Clutter impairs the perceptual system by bogging down cognitive processes and slowing visual search. In data-processing tasks like sequence comparison, clutter reduction by adjusting semantic data granularity often proves far more effective than simply removing data and still preserves the overall data set [29].

Synteny and parallel views frequently become heavily cluttered at medium to large scales as ribboning techniques often create tangled webs of orthology lines (see Figure 5c). While techniques like edge-bundling can help reduce clutter [24], the color field design and aggregation techniques (§3.3) in Sequence Surveyor attempt to avoid clutter by limiting the amount of information shown. Our prototype becomes cluttered only when there is a high density of sequence events: the clutter effectively becomes a texture signifying many features, suggesting a region for closer exploration without obscuring other regions.

3.1.4 Summarization

Summarization is the ability to construct statistical summaries of non-attended regions [3]. Summarization phenomena can rapidly provide overview information without requiring the viewer's specific attention. This is 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

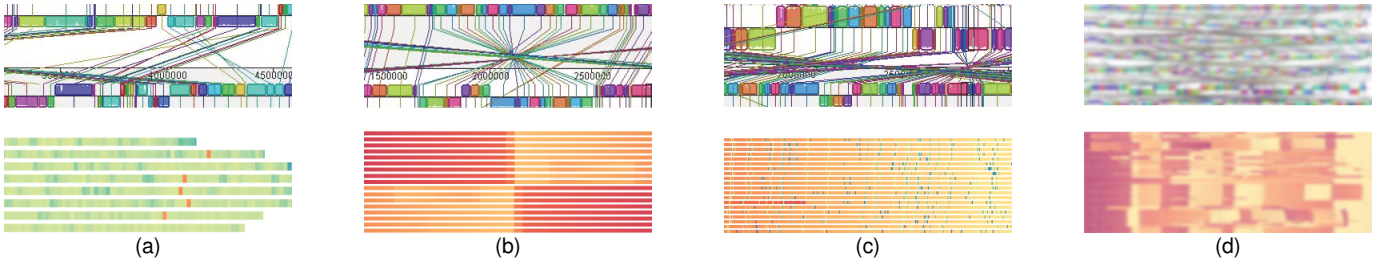


Fig. 5. Perceptual principles explain many scalability limitations in current tools, and inform our design. The top row shows views from Mauve [7], while the lower row shows similar data in our tool. (a) Pre-attentive processing states users will more readily distinguish highly conserved regions when they are mapped to bright colors than a series of orthology lines. (b) Orthology lines also impose a non-linear search order to the data, whereas a conventional reading order supports a more natural search pattern and large component color fields can be associated pre-attentively. (c) Visual clutter can form a dense texture in regions with high numbers of sequence events. (d) Unattended regions can guide visual search, but are processed by summarization mechanisms. Visualizations that support the low-resolution processing of these mechanisms can orient the viewer as to the overall data trends without requiring their explicit attention.

data. The lack of a uniform color scheme in Mauve coupled with the structural irregularity of the orthology lines and inversions inhibit summarization from conveying relevant data. In contrast, our design exploits summarization mechanisms to convey information about the data in non-attended regions. As a result, when blurred (which is similar to what summarization effectively does), a Mauve view becomes a gray mass, while our prototype's views retain useful features, such as large color fields and gradients (see Figure 5d).

3.2 Mapping

The color field design described in the previous section allows us to present a large field of information, yet have certain patterns and details emerge. However, the specific designs of the display will determine what kinds of information will form noticeable patterns. Unfortunately, we do not know the specific question that the display is meant to answer. On the contrary, a scientist may have many different kinds of questions, and our collaborators seem to have new kinds of questions emerge as they begin to explore new datasets. Therefore, we have chosen to provide a flexible set of mappings from the data to the display, giving the user control over what is encoded by horizontal position and color (cf. Figure 3). While many of these encodings have appeared in previous tools, our approach provides a generalized view of alignment data through user selectable mappings (Figure 4).

Each gene has many properties that may be encoded. Its *gene index* is its rank in the ordering of genes in the sequence, while its *start position* is its location in terms of the actual DNA (the lengths of different genes and gaps between them are included). The *position in reference* represents the gene index of a matching gene in a selected reference sequence. Frequency properties measure how many other genes match a given gene (the size of its ortholog group). *Membership frequency* counts how many different genomes contain at least one instance of an ortholog, while *gene frequency* counts how many times an ortholog occurs (this is typically greater than membership as a genes may be duplicated within genomes as paralogs). *Grouped frequency* further orders orthologs by the sets of genomes that contain them.

Any of the six gene properties may be mapped to color. Four properties may be mapped to position (of the frequency properties, only grouped frequency provides a total ordering required for a horizontal mapping). Different configurations make different kinds of patterns apparent in the display. Several of these mappings reflect the data mappings provided by common genomic visualization tools, while others present more unusual views of the data. See §5.1 for a discussion of how these mappings can be used in biological exploration.

Sequence Surveyor provides a series of eleven different color schemes: ten Color Brewer [4] ramps and one flat gray coloring. For several mappings, two color schemes are chosen. For example, the position in reference mapping uses one ramp for orthologs that match the reference, and a second ramp for those that do not (the solid gray ramp is particularly useful for this). In addition to providing aesthetic control, the color schemes provide the user with a certain degree of control over pre-attentive pop-out phenomena by allowing them the choice of color assignment for different attributes of interest.

Color mappings provide visual patterning over the data: blocks with similar properties map to similar colors. This creates visual "ramps" in the display and can highlight variations in these global trends. It also supports the pre-attentive association of various data points by creating large fields of color at regions of high similarity. Sorting mappings take advantage of the visual system's predisposition to clustering. Sorting according to particular parameters clusters visually on these parameters, imposing an orthology-based structure to the visualization: orthologous sets become spatially aggregated. This allows the viewer to quickly identify regions of interest to scan for patterns.

The large number of mapping options may intimidate users. We expect that with experience, a subset of the options will prove most useful, and we will be able to present this reduced palette with explanations of what the resulting views are useful for.

3.3 Aggregation

Most datasets of interest have sequences that are longer than we have horizontal spans of pixels. However, even if we could fit all of the data on the display, maximally dense displays may be cluttered and difficult to interpret. To manage this complexity, Sequence Surveyor performs aggregation in the horizontal direction. Genes are grouped into screen-space blocks (based on the chosen position mapping), and each block is depicted as a glyph. The block-based encoding was chosen over the more-straightforward approach of downsampling as it provides more control over how the information in the block is conveyed. The block-based approach does introduce aliasing: binning will cause blockiness and positional inaccuracy. However, we view these fine positional details as information that we are discarding in overview visualization: upon seeing the overview, other mechanisms can be used to see particular details. In return, blocking provides a mechanism for controlling clutter, and for using different visual encodings to convey information about the variety of content within the blocks.

Sequence Surveyor groups contiguous sets of genes into blocks. These groupings are determined by two primary mechanisms: block grouping width and gaps in the sequence sorting. Block grouping width is a perceptual parameter defined by the user. It specifies the minimum width gene grouping within a sequence in pixels, similar to a "bin size" parameter in a histogram. Gaps in the position mappings that are at least one pixel wide prematurely break a block grouping, creating a visible gap in the sequence encoding. This induced irregularity allows the viewer to see significant gaps in the sequence at an overview level while not overemphasizing small gaps which would otherwise be perceptually indistinguishable at scale. This gapping technique also supports the visual clustering of genes by treating physically separate gene clusters as independent blocks and mapping each cluster to the appropriate color values. The significance of sequence gaps varies with task and data priorities as well as position ordering.

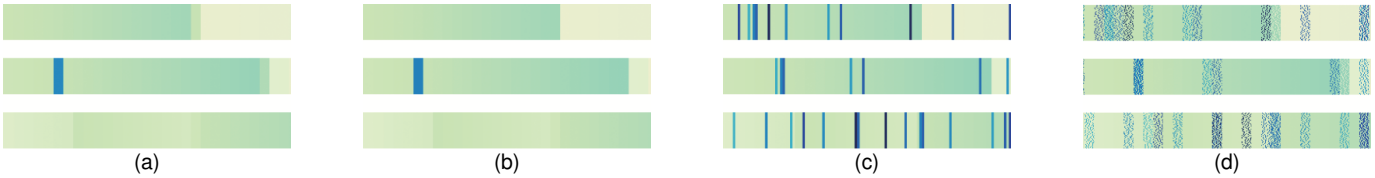


Fig. 6. The different aggregation schemes available in Sequence Surveyor. (a) Averaging reveals high-level trends in the blocks. (b) Robust averaging removes the influence of outliers from the average, resulting in smoother color fields conveying the dominant trends in the data. (c) Event Striping highlights outliers in the data. (d) Color weaving depicts the distribution of genes in the blocks.

Each block contains a number of genes, each with a parameter value mapped to a color. Different aggregation types lead to different ways of showing the contents of each block: blocks are large enough that they can encode information about the distribution of values within them. We provide four aggregation schemes, shown in in Figure 6: averaging, robust averaging, event striping, and color weaving.

- *Averaging* (Figure 6a) colors blocks by the mean of the component gene color values. Multiple color ramp mappings color according to the average gene values from the dominant color range within the block. This aggregation scheme summarizes overall trends in each block, as seen in Figure 13b.
- *Robust averaging* (Figure 6b) averages more intelligently by averaging the gene color values within one mean absolute deviation of the inner quartile range. This reveals dominant trends in the data by removing outliers from the average (cf. Figure 12a).
- *Event striping* (Figure 6c) flags outliers and changes to trends in a block as "events", drawing them as pixel-wide vertical stripes at the relative location of each event within the block. This prioritizes the drawing of outliers within the blocks, physically enlarging these regions of the glyph to highlight their existence, which may otherwise be lost to more dominant trends (cf. Figure 12b).
- *Color weaving* (Figure 6d) breaks block glyphs into individual pixels based on the technique defined in [16]. The gene positions are randomized and mapped over these pixels, approximating the distribution of gene values within the block. Randomization helps avoid misleading striping artifacts that may be introduced through repeating ordered data, as seen in Figure 13c.

By providing different aggregation filters, different properties of the data can be explored at the overview level without having to recompute the display properties of the entire set. Drilling deeper into these aggregated blocks can be accomplished with zooming (§3.5).

3.4 Data Display

The phylogenetic tree shows the relationship between genomes within the data set if this information is available. As a result, using the phylogenetic data as the backbone of the initial visualization structure clusters genomes according to an approximation of their pair-wise similarity. Organizing genomes by their similarity can help facilitate comparison by vertically clustering genomes with high conservation. The user may change this ordering during exploration. Sequence ordering can be specified through manual interaction or through tree files, which can be written and read by the tool during exploration.

Overall gene-level information is summarized in the histogram, displayed below the properties panel. The height of histogram bar represents gene frequency and the ortholog groups are sorted and aggregated according to the same frequency metric. The resulting shape conveys the overall frequency distributions of genes within the data set. A lasso-selection filter highlights interesting frequency clusters within the genome viewer. Brushing in the histogram coordinates with the phylogenetic tree by highlighting branches up to the most recent common ancestor shared by the genomes conserving the brushed orthologs and highlights these orthologs in the genome viewer.

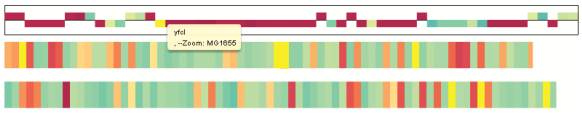


Fig. 7. Overview+detail zooming manages the non-locality issues arising in multiple genome alignments. As the user mouses over blocks in the genome view, component genes of those blocks are visualized in the zoom window (top), positioned vertically according to the strand where they are found and horizontally according to the position mapping. Zoom can be locked onto a block for interactive functionality.

3.5 Interaction: Exploration and Zooming

Overview visualizations rely heavily on data abstraction to present data in a meaningful way. However, the details of the underlying data are still significant to exploring large datasets. Sequence Surveyor uses interaction techniques to reveal detailed information hidden by overview abstraction. Information about the genes, chromosomes and sequences represented in a block are provided by a tooltip window. Brushing across blocks reveals genes conserved across different genomes: mousing over a particular set of genes highlights blocks containing orthologous genes. This brushing mechanism highlights the path between the target genome and its immediate sibling sequences in the phylogenetic tree, guiding organism-level comparison.

Inspired by synteny viewers, orthology lines can be drawn on demand. The user can click on blocks to visually link all blocks containing genes orthologous to the component genes of the block. Similarly, filters reduce the opacity of blocks not containing genes described by the filter parameters (gene name, ortholog group, reference chromosomes or genomes, and frequencies). Reducing the opacity of blocks outside of the filter preserves the overall context of the data while making the desired gene set more salient, visually emphasizing genes of interest with regard to the user's queries. Filters can be loaded and saved during exploration.

Gene-level sequence alignments do not guarantee that orthologous genes will be located in similar positions in all genomes. As a result, traditional zooming techniques, such as semantic and goal-directed zoom, can hide orthologous matches as the user drills down. In Sequence Surveyor, we attempt to circumvent this visual data loss through an adapted detail+overview zooming technique (Figure 7). Mousing over a block sets it as the zoom focus block. The component genes of the block are broken down in the zooming window at the top of the screen and visualized on either side of a reference line based on the strand of the DNA the gene is located on.

Zoom may also be focused on a block. This allows the user to brush over individual genes in the zoomed display to trace their locality. By locking to a block within the zoomed sequence, the user can also drill deeper into the data at finer semantic levels.

4 Implementation

Sequence Surveyor was implemented using ActionScript 3.0 with a SQLite database back end. The data model used in this database contains a reduced set of data properties for visualization: orthology data, positional data, and descriptive data (gene, chromosome, and genome names, descriptions, strand information). Format converters allow Se-

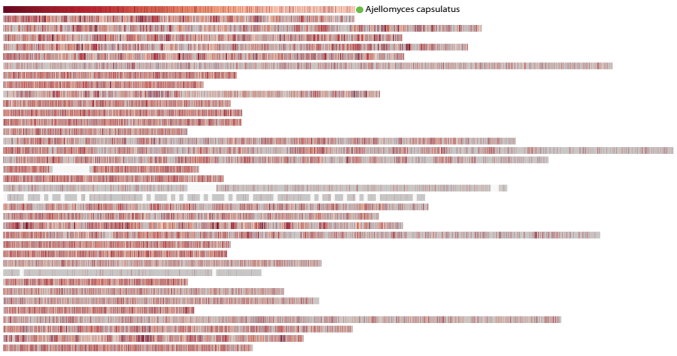


Fig. 8. Filtering reveals orthology with a single chromosome: blocks not containing any genes orthologous to a gene in the chromosome are reduced in opacity. Coloring according to a reference and aggregating by striping helps show detailed conservation patterns across thirty four fungal genomes. Interaction can be used to identify specific conservation patterns.

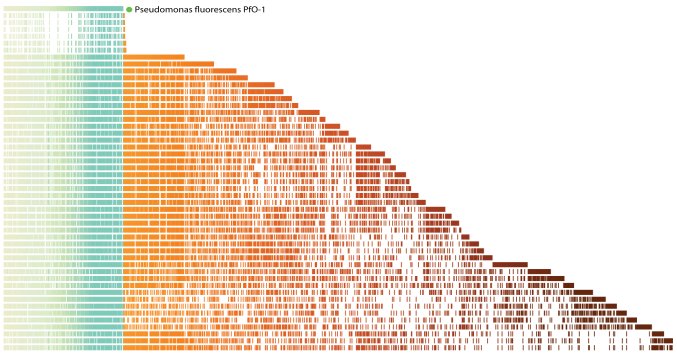


Fig. 9. Genes from 50 bacterial genomes are sorted according to their position in *Pseudomonas fluorescens PfO-1* (green circle). Genes not conserved by the reference are sorted according to their order in the remaining genomes (computed from the topmost genome downwards). Redundantly coloring by this reference reinforces this conservation: cool blocks are conserved whereas warm blocks are not.

quence Surveyor to support data from several different data formats.

5 Applications

5.1 Parallels to Existing Tools

By allowing the user to customize visualization parameters on demand, Sequence Surveyor is able to present views familiar to users of common tools. This provides familiar paradigms for exploring the datasets at scale. Here, we explore the views provided by Mauve, Mizbee [24], the Broad Institute Medea package [19], and the UCSC Genome Browser [21] and show how the information from these views can be displayed at scale in Sequence Surveyor.

Mauve: The Mauve viewer displays alignments using a parallel ribbon design: genomes map to rows and orthology is encoded by links. Despite the scalability issues discussed in §3.1, Mauve is effective for observing matchings between genes to see the patterns of conservation and rearrangement. By mapping gene position to start position and encoding matching genes with similar colors (for example grouped frequency or position in reference), Sequence Surveyor can convey similar information at much larger scale. For instance, inversions creating crossing formations in Mauve are reflected in Sequence Surveyor as inverted color ramps. While, in practice, crossing patterns are often much more salient than color for small inversions, detail-on-demand links can be used to supplement the color-based encoding. Additionally, Sequence Surveyor's flexibility in coloring makes it easier to see observations of interest (see Figure 10).

Mizbee: Mizbee's genome view shows conservation between two genomes by examining the conservation between particular chromosomes in a source genome and orthologous genes in a destination genome. Color maps to the destination chromosome that the conserved region is found in and conservation is further indicated by orthology ribbonning. Per-chromosome conservation information can be seen in Sequence Surveyor by filtering by orthology to the chromosome of interest and using interaction to explore more detailed conservation relationships (Figure 8). Mapping color to position in the destination genome reinforces the synteny coloring employed by Mizbee.

Medea: The Broad Institute's Medea suite provides five different visualization perspectives for viewing sequence alignment data for closely-related viruses: the Circular Genome Viewer, Stack Map, ChromoMap, Dot Plot, and Viral Viewer. Because these viral genomes are small and tend to have only point mutations, the Broad tools focus on reference-based displays: there are no issues of non-locality as matching regions are co-located in the data set. Sequence Surveyor can support similar explorations to the Medea suite by encoding data using position in reference.

UCSC Genome Browser: While the focus of the UCSC genome browser has traditionally been on exploring individual genomes, there is also limited support for visualizing multiple sequences simultaneously. This approach selects a reference genome and places all other genomes in parallel tracks beneath the reference. A box in a track represents a subsequence that is conserved in the reference. Conserved regions are ordered according to their position in the reference genome. This conservation data can be explored in Sequence Surveyor by sorting the genes according to their position in the desired reference. Any elements conserved from the reference will line up beneath their corresponding positions in the reference genome (see Figure 9).

5.1.1 Unconventional Mappings

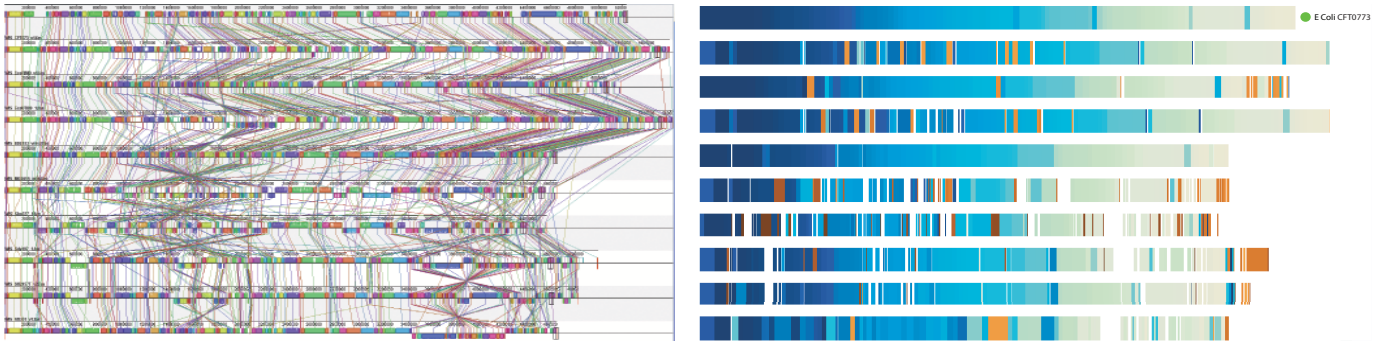
Sequence Surveyor supports exploration using less conventional mappings to provide insight into different properties of the data. Novel position mappings leverage pre-search processing to cluster genes more effectively than either color or orthology ribonning. Most existing tools do not explore gene position orderings besides sorting by start position. While this mechanism is useful for viewing data when gaps are relevant, it increases the number of objects on the screen, thereby increasing cognitive search loads. Alternatively, gene index sorting orders genes according to their local position in the genome, removing extraneous gaps in the data and dedicating more space to genes.

While many tools color by a reference genome, small regions not conserved in the reference genome can easily be obscured. Mapping gene position to a reference segregates ortholog groups according to their conservation in the reference, preserving small unconserved regions. This mapping also reveals the degree of homology between the source and other genomes: the smaller the reference genome becomes, the fewer ortholog groups it shares with the remaining genomes in the set. Similarly, sorting by grouped frequency visually clusters data according to the sets of genomes in which each ortholog group is contained. This provides insight into co-occurring ortholog groups. If paired with a start position or gene index coloring, these position mappings can display information about the organization of conserved regions in the data such as large-scale inversions and rearrangements.

Raw gene frequency is not commonly visualized in existing tools despite its intuitive meaning. However, coloring by gene frequency can reveal significant duplication patterns in the data set, potentially signalling significant genes or bugs in the underlying data. This coloring also reveals patterns where instances of an ortholog group occurs and visualizes many-to-many correspondences between the instances of a group in different genomes.

5.2 Use Cases

The data used in this paper comes from four groups of domain scientists: evolutionary biologists, a systems biologist, a yeast biologist, and a bioinformatician. All four groups have large genome alignment data that they want to explore, but no analysis tools to support that exploration. We use three bacterial datasets (100 genomes with up to



(a) (b)

Fig. 10. Ten *E. coli* and *Shigella* genomes visualized by (a) Mauve (reference is top genome) and (b) Sequence Surveyor by coloring by position in *E. coli CFT073* (green circle) and ordering by start position. The vertical genome order is the same in both cases. The conservation trends represented by orthology lines in Mauve become large color fields in Sequence Surveyor. Inversions appear as reversals in the color ramp. Regions not conserved appear as warm-colored blocks pre-attentively popping out of the visualization.

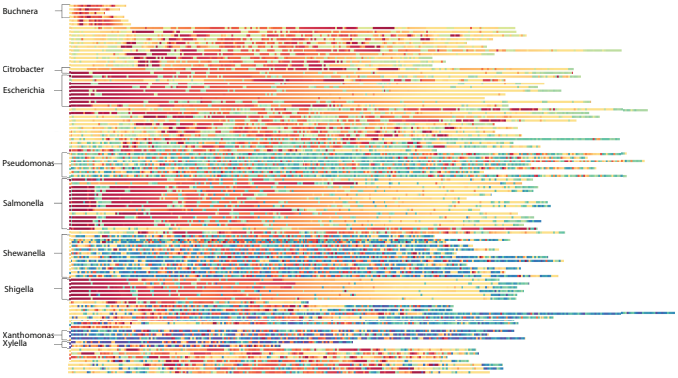
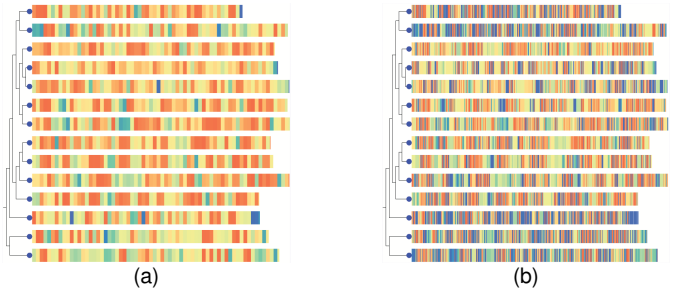


Fig. 11. Genome order can help reveal patterns between families of genomes. Sorting one hundred bacterial genomes by index and coloring by position in an *E. coli* organism highlights the high conservation between *Escherichia*, *Shigella*, *Salmonella*, and *Buchnera* genomes through warm colored bands and lack of conservation between *E. coli* and the *Pseudomonas* and *Shewanella* genomes.



(a) (b)

Fig. 12. Fourteen bacteria colored by membership frequency shows the conservation of genes and their spatial organization. (a) Robust averaging shows that genes are well conserved overall (warm-colored blocks). (b) Event striping highlights the outliers, exposing the distribution of more unique genes (blue).

6,037 genes per sequence (cf. Figures 11, 13), a subset of the 100 bacteria dataset with 50 bacteria and up to 5,765 genes per sequence (cf. Figures 9), and 14 plant pathogens with up to 4,507 genes per sequence (cf. Figures 3, 12)), one Mauve alignment of ten *E. coli* genomes (cf. Figure 10), and one draft multi-chromosomal fungal data set with up to 17,349 genes per genome (cf. Figure 8). We prepared the datasets from the domain scientists, including computing the alignments (the large alignments took 10 days of CPU time). Once the data was prepared, we worked with the domain scientists to introduce them to the tool and explore their data.

Users appreciated Sequence Surveyor as an examination tool useful for discovering and describing trends in data. They were immediately struck by the scale of the visualizations, not just in terms of size, but also in terms of diversity. Most were pleasantly surprised as they made observations comparing organisms thought to be unrelated. For example, filtering allowed them to quickly identify interesting genes and view the conservation of those genes even in unrelated sequences. The 100 bacteria data set aligns genomes from a variety of organisms, like *Yersinia pestis* (Black Plague), *E. coli*, *Salmonella*, and *Xylella fastidiosa* (a plant-bourne pathogen). The organization of the data according to evolutionary families proved to play an important role in comparing this diverse data set. Coloring by position in a reference organism from a given family highlighted high levels of conservation between related families (Figure 11). Closely related families generally conserve the reference color ramp, whereas less related families introduce new, more divergent color patterns. Furthermore, it allowed the biologists to identify *Citrobacter* genomes by eye from their conservation patterns and place these genomes near the related *Escherichia* genomes to better facilitate comparison. More global conservation patterns can be seen using grouped frequency sorting (cf. Figure 13a).

Sequence Surveyor allowed the scientists to quickly identify the set of genes that were conserved across the entire data set, also known as the "ancestral core", formed by the leftmost columns of genomes when mapping position to grouped frequency ordering. With respect to systems biology, the ancestral core is often composed primarily of essential metabolic genes. Being able to quickly identify these metabolic genes through this ancestral core can help highlight locality patterns between metabolic genes of interest from specific metabolic pathways. From an evolutionary standpoint, these core genes can reveal interesting functional properties of different genomic regions. The *Buchnera* genomes are drawn from insect parasites whose genomes have been pared down to an essential set of genes necessary for survival. By adjusting the parameters in Sequence Surveyor, this observation becomes readily apparent as nearly all component genes of these genomes appear as part of this ancestral core. The biologist can even gain insight into the loci at which these genes are conserved within other families of bacteria. The ability to manipulate the representation of the comparison of *Buchnera* genomes and the rest of the dataset is communicated very visually in Sequence Surveyor (Figure 13).

Our collaborators found Sequence Surveyor's ability to address different questions valuable. While exploring the data, position mappings like grouped frequency allowed them to quickly address questions that we had not previously considered, such as what set of genes is conserved only in a specific subset of the genomes. Also, they commented on how the tool's ability to blend location and conservation data in a

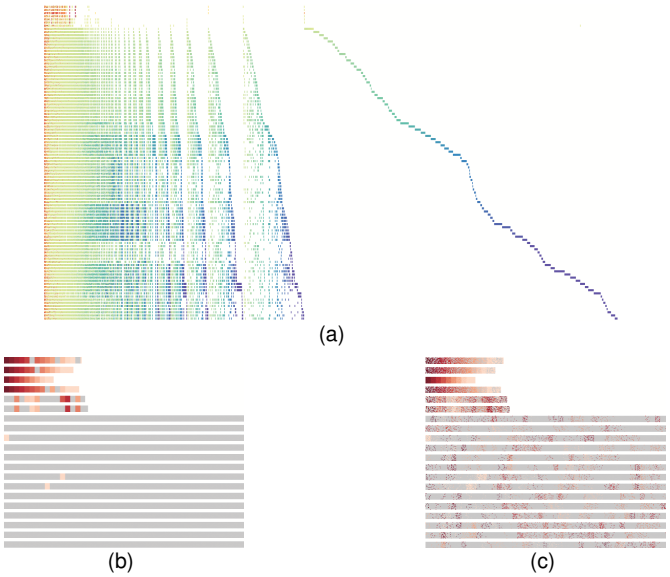


Fig. 13. One hundred bacterial genomes. (a) Grouped frequency position clusters the genes conserved in all organisms (the "ancestral core") in the leftmost columns. The top six genomes, *Buchnera* insect parasite genomes, are concentrated in this cluster, reinforced by position in reference coloring (red). Sorting by index shows the position of these *Buchnera* genes in the data: (b) averaging shows that few regions are dominated by this essential set, but (c) color weaving reveals their prolific distribution.

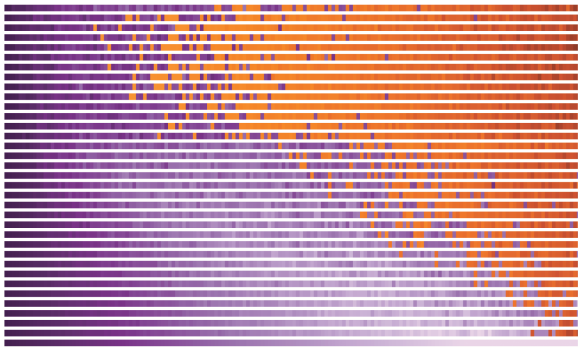


Fig. 14. The 5,000 most common words per decade since 1660. Sequences reflect the decade and are sorted chronologically. Words are sorted according to rank, and colored by position in the reference decade (2000-2010, bottom). Words in purple are the top 5,000 words in the reference decade, whereas words in orange are not. The upper-triangular pattern reflects the evolution of the English vernacular over time: while the most common words are constant, less common words evolve more quickly.

flexible setting would allow them to quickly identify the location of interesting clusters of genes and how tuning aggregation settings could support the exploration of unique features in their data (Figure 12).

Visualizations of large-scale patterns in data are also valuable for discovering bugs in datasets and alignment algorithms. As an example of Sequence Surveyor's value as a debugging tool, with it we were able to identify a number of problems with a draft alignment of 37 fungal genomes used during our testing and evaluation. First, the visualization quickly revealed that this data set contained a putative ortholog group of over 60,000 genes. This group popped out easily due to an extreme skew in the histogram and again by using gene frequency coloring. Upon a more detailed exploration of the genes in this group, the extreme duplication revealed itself to be a bug in the orthology assignments and was removed from the alignment. A second major issue revealed by the tool was that a number of genomes did not contain many genes that had orthologs in other species. This discovery prompted a manual inspection of parts of the alignment, which ultimately led to the identification of some inconsistencies in the labeling of genes by the alignment code. Without Sequence Surveyor, it is likely that it would have taken a lot more time for these problems to be discovered.

5.3 Other Applications

Sequence Surveyor can be applied to any analysis problem comparing datasets that can be mapped to a total ordering and similarity mapping ("orthology"). Obvious such extensions include the visualization of amino acid and nucleotide-level alignments. However, an unconventional application of these techniques is to the visualization of the Google N-Grams word count data [25]. We extracted the 5,000 most popular words per decade since the year 1600 from this data. Words are treated as genes, ranks as positions, and decades as genomes. Various Sequence Surveyor mappings reveal interesting patterns. For instance, Figure 14 shows the N-Grams data sorted by rank and colored according to the rank of the word in the decade from 2000 to 2010.

6 Discussion

The Sequence Surveyor prototype combines a perceptually-motivated color field display, flexible mappings, selectable aggregation strategies, and interaction techniques to provide overview visualization of multiple whole-genome alignments and similar data types. The initial feedback from domain collaborators suggests that they are excited to have a tool capable of allowing exploration at this scale.

The most immediate limitations in the prototype are implementation issues. The first prototype is built on a platform that has poor performance characteristics, making the application difficult to deploy and usability issues difficult to address. Practical features, such as easy connection to web reference databases need to be added. However, the promise of large-scale alignment exploration makes our domain collaborators patient with the prototype.

Our design does not yet address some aspects of alignment visualizations. For example, we provide no mechanism for displaying other data such as probabilistic alignments, match strengths, or annotations of the genes. Multiple selection, grouping, and conjunctive filtering are all mechanisms that could enhance the interaction techniques to widen the kinds of questions that can be explored easily.

Scaling to larger datasets poses new challenges. Handling longer genomes will require better zoom mechanisms than we currently provide, most likely including multi-scale views. Handling more genomes will require development of "vertical" aggregation strategies to group genomes and summarize the subsets, as well as interaction techniques for looking at detailed comparisons across sets. Experience working with domain experts will suggest a wider variety of questions that may require new view organizations to present data. While we are concerned that having too many choices for view control may challenge potential users, our collaborators seem to embrace the control it gives them in how their data is presented. A related point is that our collaborators were very excited with the use of overview tools for the presentation of their data when they publish their findings, which may have different needs than tools for exploration.

We have introduced Sequence Surveyor, a scalable genomic alignment overview visualization. Sequence Surveyor has been designed based on a series of perceptual observations and aggregation techniques to compose a generalized sequence comparison tool capable of interactively visualizing over 100 genomic sequences simultaneously. We have shown analyses from prior tools at scale and applications of Sequence Surveyor techniques beyond genome alignments.

Acknowledgments

This project was supported in part by DoE Genomics:GTL and SciDAC Programs (DE-FG02-04ER25627), NSF awards IIS-0946598, CMMI-0941013 and DEB-0936214.

REFERENCES

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

[2] R. Arnheim. The Perception of Maps. *Cartography and Geographic Information Science*, 3(1):5–10, Apr. 1976.

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

[4] C. A. Brewer, G. W. Hatchard, and M. A. Harrower. Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30:5–32(28), 2003.

[5] T. J. Carver, K. M. Rutherford, M. Berriman, M.-A. Rajandream, B. G. Barrell, and J. Parkhill. ACT: the Artemis comparison tool. *Bioinformatics*, 21(16):3422–3423, 2005.

[6] M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton. The Jalview Java alignment editor. *Bioinformatics (Oxford, England)*, 20(3):426–7, 2004.

[7] 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, 2004.

[8] C. Duran, Z. Boskovic, M. Imelfort, J. Batley, N. A. Hamilton, and D. Edwards. CMap3D: a 3D visualization tool for comparative genetic maps. *Bioinformatics (Oxford, England)*, 26(2):273–4, 2010.

[9] R. Engels, T. Yu, C. Burge, J. P. Mesirov, D. DeCaprio, and J. E. Galagan. Combo: a whole genome comparative browser. *Bioinformatics*, 22(14):1782–1783, 2006.

[10] J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. *IEEE Symposium on Information Visualization*, pages 117–124, 2002.

[11] S. L. Franconeri. The nature and status of visual resources. In D. Resiberg, editor, *Oxford Handbook of Cognitive Psychology*. Oxford University Press, 2011.

[12] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak. VISTA: computational tools for comparative genomics. *Nucleic Acids Research*, 32:W273–9, 2004.

[13] M. G. Grabherr, P. Russell, M. Meyer, E. Mauceli, J. Alföldi, F. Di Palma, and K. Lindblad-Toh. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics (Oxford, England)*, 26(9):1145–51, 2010.

[14] J. R. Grant and P. Stothard. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Research*, 36(suppl 2):W181–W184, 2008.

[15] L. Guy, J. Roat Kultima, and S. G. E. Andersson. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics (Oxford, England)*, 26(18):2334–2335, 2010.

[16] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1270–7, 2007.

[17] C. Healy. Perception in visualization. Web Resource, http://www.csc.ncsu.edu/faculty/healey/PP/index.html.

[18] P. Husemann and J. Stoye. R2Cat: Synteny Plots and Comparative Assembly. *Bioinformatics (Oxford, England)*, 26(4):570–1, 2010.

[19] 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.

[20] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6:59–78, 2000.

[21] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[22] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–45, Sept. 2009.

[23] D. Lee, J.-H. Choi, M. M. Dalkilic, and S. Kim. COMPAM: visualization of combining pairwise alignments for multiple genomes. *Bioinformatics*, 22(2):242–244, 2006.

[24] M. Meyer, T. Munzner, and H. Pfister. Mizbee: A multiscale synteny browser. *IEEE Transactions on Visualization and Computer Graphics*, 15:897–904, 2009.

[25] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[26] M. Muffato, A. Louis, C.-E. Poisnel, and H. R. Crollius. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics (Oxford, England)*, 26(8):1119–21, Apr. 2010.

[27] T. Peeters, M. Fiers, H. van de Wetering, J.-P. Nap, and J. J. van Wijk. Case Study: Visualization of annotated DNA sequences. *Eurographics*, 2004.

[28] 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, 2010.

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

[30] B. Shneiderman. Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 3–12, New York, NY, USA, 2008. ACM.

[31] J. Slack, K. Hildebrand, T. Munzner, and K. John. SequenceJuxtaposer: Fluid navigation for large-scale sequence comparison in context. In *German Conference on Bioinformatics*, pages 37–42, 2004.

[32] A. Slingsby, J. Dykes, and J. Wood. Configuring hierarchical layouts to address research questions. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):977 –984, 2009.

[33] B. Swihart, B. Caffo, B. James, M. Strand, B. Schwartz, and N. Punjabi. Lasagna plots: A saucy alternative to spaghetti plots. *Epidemiology*, 21(5):621–625, 2010.

[34] M. Wattenberg and F. Viegas. Beautiful history: Visualizing wikipedia. In J. Steele and N. Illinsky, editors, *Beautiful Visualization*, page 416. O'Reilly Media, Inc., 2010.