Sequence Surveyor

Leveraging Overview for Scalable Genomic Alignment Visualization

Danielle Albers, Colin Dewey, and Michael Gleicher
University of Wisconsin-Madison
Department of Computer Sciences
IEEE VisWeek 2011
Viewing Genome Alignments
Viewing Genome Alignments
Perception

Scalable Design

Aggregation

Mapping
Scalable Design
Outline

The Data Domain
Sequence Surveyor
Design in Theory
  - Perception
  - Mapping
  - Aggregation
Design in Practice
Whole Genome Alignment

Identify related groups of genes appearing in a set of organisms

Organism One: A C G T G G C A A C T T

Organism Two: G G C A A C G T A C T T
Defining Scale

**Number of Genomes**
- 100
- 50
- 8

**Length of Genomes**
- Fungi (17,000+ genes)
- Bacteria (6,000+ genes)
- Viral (300+ genes)

**Types of Inquiry**
- Locality
- Co-occurrence
- Reference-Based
Outline

The Data Domain

Sequence Surveyor

Design in Theory
  - Perception
  - Mapping
  - Aggregation

Design in Practice
Our Solution
Our Solution

Phylogenetic Tree
Mapping Pane
Block Detail
Genomes
Histogram
Our Solution

Block Detail

Aggregation
Our Solution
Our Solution

Histogram

Phylogenetic Tree
Outline

The Data Domain
Sequence Surveyor

Design in Theory
- Perception
- Mapping
- Aggregation

Design in Practice
Perception

How the user processes dense data

Inform scalable design
- Limitations of current designs
- Insight into future designs

Four principles
Perceptual Principles

Pre-Attentive Phenomena

Visual Search

Visual Clutter

Summarization
Perceptual Principles

- Pre-Attentive Phenomena
- Visual Search
- Visual Clutter
- Summarization
Perceptual Principles

Pre-Attentive Phenomena

Visual Clutter

Visual Search

Summarization
Perceptual Principles

Pre-Attentive Phenomena

Visual Clutter

Visual Search

Summarization
Perceptual Principles

Pre-Attentive Phenomena

Visual Clutter

Visual Search

Summarization
Perception

Overview - Sacrifice detail for high-level comparison

Colorfield - Emphasize visual structure

Mappings – Emphasize key details

Aggregation – Do not overwhelm viewers
Mapping

Color Mapping
- Locality
- Frequency
- Reference

Color Schemes
- Sequential
- Diverging
- Split Schemes

Position Mapping
- Locality
- Frequency
- Reference
Combinations of different color and position mappings reveal interesting trends in the data

<table>
<thead>
<tr>
<th></th>
<th>Index</th>
<th>Membership Freq</th>
<th>Grouped Freq</th>
<th>Pos in Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grouped Freq</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pos in Reference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aggregation

Cannot show all the data at once
  - Limited screen real estate
  - Clutter

Blocking preserves local control
  - Display gene neighborhoods as glyphs

Four block encodings
Blocking

Group (relatively) continuous sets of neighboring genes into a single unit
Aggregate Encodings

- Average
- Robust Average
- Color Weaving
- Event Striping
## Interaction

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Block Brushing</strong></td>
<td>Highlight locations of block contents in overview, phylogeny, and histogram on mouse-over</td>
</tr>
<tr>
<td><strong>Block Linking</strong></td>
<td>Link locations of block contents in overview on click</td>
</tr>
<tr>
<td><strong>Detail Notes</strong></td>
<td>Details of genes in a block and matching genes of the set are presented in a separate window</td>
</tr>
<tr>
<td><strong>Non-locality Zoom</strong></td>
<td>Explore the contents of an aggregate block in the Block Detail Window on mouse-over</td>
</tr>
<tr>
<td><strong>Zoom Lock</strong></td>
<td>Fix the contents of a block in the zoom window to explore the distributions of specific genes</td>
</tr>
<tr>
<td><strong>Zoomed Gene Brushing</strong></td>
<td>Highlight locations of genes in overview, phylogeny, and histogram</td>
</tr>
<tr>
<td><strong>Zoomed Gene Linking</strong></td>
<td>Link locations of a set of matching genes in the overview</td>
</tr>
<tr>
<td><strong>Manual Rearrangement</strong></td>
<td>Drag-and-drop rearrangement of sequences and indicate branch crossings by opacity</td>
</tr>
<tr>
<td><strong>Filtering</strong></td>
<td>Highlight genes matching a set of names, id numbers, frequencies, genomes, or chromosomes</td>
</tr>
<tr>
<td><strong>Load Filter</strong></td>
<td>Load a filter set from a CSV</td>
</tr>
<tr>
<td><strong>Save Filter</strong></td>
<td>Save the current filter set to a CSV</td>
</tr>
<tr>
<td><strong>Histogram Brushing</strong></td>
<td>Highlight the locations of genes in a region of the frequency distribution in the overview and phylogenetic tree by mouse-over</td>
</tr>
<tr>
<td><strong>Load Tree</strong></td>
<td>Load different trees and arrangements from a tree file</td>
</tr>
<tr>
<td><strong>Save Tree</strong></td>
<td>Save the current tree structure and sequence arrangement to a tree file</td>
</tr>
</tbody>
</table>
Outline

The Data Domain
Sequence Surveyor
Design in Theory
  - Perception
  - Mapping
  - Aggregation
Design in Practice
Use Cases

100 Bacteria
   6,000 genes

50 Bacteria
   5,000 genes

35 Fungi
   17,000 genes

14 Pathogens
   4,000 genes

8 partial E. coli sequences
   300 genes
Parallels

Can use Sequence Surveyor to obtain information presented in existing tools at scale.

Mauve: Color by position in reference (arrow), order by start position
Anecdotes: Buchnera

Buchnera family of genomes and the ancestral core

Color by position in reference (arrow), order by set of genomes containing each gene
Anecdotes: *Buchnera*

**Averaging:**
No significant trend

**Color Weaving:**
Overall distribution
Anecdotes: *E. Coli*

Conservation relationships between different families of genomes

Color by position in reference (arrow), order by relative ordering
Bioinformatics applications allow users to test algorithms using visual checks

Color by overall frequency, order by relative ordering
Anecdotes: Fungi

Bioinformatics applications allow users to test algorithms using visual checks. Color by position in a reference, order by relative ordering.
Extensions

Proteins and nucleotide MSA

Any data with an orthology and ordered sets

Google N-Grams

Top 5,000 most popular words since 1660

Distribution of a word set in 2000 across time
Summary

Scalable whole genome alignment overview

Perception informs design

User-controlled mapping scales across queries

Aggregation filters data

Extends beyond the immediate biology
Acknowledgements

University of Wisconsin – Madison
Department of Computer Sciences Graphics & Vision Lab

University of Wisconsin – Madison
BACTER Institute for Computational Biology

University of Wisconsin – Madison
Genome Center Genome Evolution Laboratory
  Dr. David Baumler
  Dr. Eric Neeno-Eckwall
  Dr. Jeremy Glasner
  Dr. Nicole Perna

Funding by NSF awards IIS-0946598, CMMI-0941013 and DEB-0936214 and DoE Genomics: GTL and SciDAC Programs (DE-FG02-04ER25627)
Availability

Prototype and sample data package (coming soon):
http://graphics.cs.wisc.edu/Vis/SequenceSurveyor/

dalbers@cs.wisc.edu