Visualization Analysis & Design for Biology

Tamara Munzner
Department of Computer Science
University of British Columbia

VanBUG: Vancouver Bioinformatics Users Group
8 October 2015, Vancouver BC

http://www.cs.ubc.ca/~tmm/talks.html#vanbug15 @tamaramunzner
Outline

• introduction
• Cerebral
• MizBee
• TreeJuxtaposer
• wrapup
Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.
Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

- many analysis problems ill-specified, not clear what questions to ask in advance – don’t need vis when fully automatic solution exists and is trusted

Anscombe’s Quartet

<table>
<thead>
<tr>
<th>Identical statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x mean</td>
<td>9</td>
</tr>
<tr>
<td>x variance</td>
<td>10</td>
</tr>
<tr>
<td>y mean</td>
<td>8</td>
</tr>
<tr>
<td>y variance</td>
<td>4</td>
</tr>
<tr>
<td>x/y correlation</td>
<td>1</td>
</tr>
</tbody>
</table>
Why use an external representation?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- external representation: replace cognition with perception

Analysis framework: Four levels, three questions

• **domain** situation
  – who are the target users?

• **abstraction**
  – translate from specifics of domain to vocabulary of visualization
  • **what** is shown? **data abstraction**
    • often don’t just draw what you’re given: transform to new form
  • **why** is the user looking at it? **task abstraction**

• **idiom**
  • **how** is it shown?
    • **visual encoding idiom:** how to draw
    • **interaction idiom:** how to manipulate

• **algorithm**
  – efficient computation


Validation methods from different fields for each level

- anthropological/ethnography
  - **Domain situation**
    - Observe target users using existing tools
  - **Data/task abstraction**
    - Visual encoding/interaction idiom
      - Justify design with respect to alternatives
    - Algorithm
      - Measure system time/memory
      - Analyze computational complexity
  - Analyze results qualitatively
  - Measure human time with lab experiment (*lab study*)
  - Observe target users after deployment (*field study*)
  - Measure adoption

- design
  - Problem-driven work

- computer science
  - Technique-driven work

- cognitive psychology

- anthropology/ethnography

- mismatch: cannot show idiom good with system timings
- mismatch: cannot show abstraction good with lab study
Angles of attack

- **technique-driven work**
- **problem-driven work**
- **theoretical foundations**
- **evaluation**
Technique-driven work: Networks

• scaling up networks
  – multilevel networks, 10K-100K nodes
    • topologically aware decomposition, layout, browsing
  – trees, millions of nodes
    • guaranteed visibility of semantically meaningful marks

*Figure 7.25: GrouseFlocks uses containment to show graph hierarchy structure. (a) Original graph. (b) Several alternative hierarchies built from the same graph. The hierarchy alone is shown in the top row. The bottom row combines the graph encoded with connection with a visual representation of the hierarchy using containment. From [Archambault et al. 08], Figure 3.*

**TopoLayout**
**Smashing Peacocks Further**
**Grouse**
**GrouseFlocks**
**TugGraph**

http://youtu.be/t1Xbt6XOWp8
http://youtu.be/AWXae8zykt8

**TreeJuxtaposer**
**PRISAD**

http://youtu.be/fq8EIAOutvs
http://youtu.be/GdaPj8a9QEO
Technique-driven work: Dimensionality reduction

• close overlap with machine learning
  – Glimmer: MDS on the GPU
  – Glint: DR for costly distances
  – QSNE: sparse documents
    • high quality for millions of items

Glimmer
http://youtu.be/PLaBAPM6qLI
Evaluation: Dimensionality Reduction

• guidance on scatterplot/DR choices

• taxonomy of cluster separation factors

• 2D points vs 3D landscapes
Evaluation: Focus+Context

• overviews: separate vs. integrated views

• navigation: stretch and squish vs. pan/zoom navigation

• impact of distortion on visual search, visual memory
Theory/Models

- multi-level typology of abstract visualization tasks
- design study methodology
- nested model for vis design and validation
- papers process and pitfalls
- book: Visualization Analysis and Design
Problem-driven work: Many domains

LiveRAC: system management time-series

http://youtu.be/id0c3H0VSkw

RelEx: in-car overlay networks

http://youtu.be/89lsQXc6Ao4

Vismon: fisheries management

http://youtu.be/h0kHoS4VYmk

Overview: investigative journalism

http://vimeo.com/71483614
Problem-driven work: Genomics

MizBee
http://youtu.be/86p7brwuz2g

Variant View
http://youtu.be/AHDnv_qMXxQ

MulteeSum
Pathline

Cerebral
http://youtu.be/76HhG1FQngI
Design Study Methodology

Reflections from the Trenches and from the Stacks

joint work with:
Michael Sedlmair, Miriah Meyer

http://www.cs.ubc.ca/labs/imager/tr/2012/dsm/
Design Studies: Lessons learned after 21 of them

Cerebral genomics
MizBee genomics
Pathline genomics
Mulfetum genomics
Visman fisheries management
QuestVis sustainability
WiKeVis in-car networks

MostVis in-car networks
Car-X-Ray in-car networks
ProgSpy2010 in-car networks
RelEx in-car networks
Cardiogram in-car networks
AutobahnVis in-car networks
VisTra in-car networks

Constellation linguistics
LibVis cultural heritage
Caidants multicast
SessionViewer web log analysis
LiveRAC server hosting
PowerSetViewer data mining
LastHistory music listening
DESIGN STUDY METHODOLOGY SUITABLE
Problem-driven work and you

• enormous opportunity for bioinformatics
  – apply human-centered design methods beyond visualization!
  – task analysis: what do the target users really need?
Outline

• introduction

• Cerebral

• MizBee

• TreeJuxtaposer

• wrapup
Cerebral
Visualizing Multiple Experimental Conditions on a Graph with Biological Context

joint work with:
Aaron Barsky, Jennifer Gardy, Robert Kincaid

http://www.pathogenomics.ca/cerebral/
What: Data abstraction

- dataset types
  - network
    - nodes: genes
    - links: known interactions between genes
  - table
    - quantitative attributes
      - gene expression levels for nodes across different experimental conditions
    - categorical attributes
      - subcellular location of interaction
      - functional groups

© Attribute Types
- Categorical
- Ordered
- Quantitative
Why: Task abstraction

- task: interpret microarray experiment results with respect to gene network
  - goal: accelerate existing discovery workflow
  - compare distributions between attributes
    - experimental conditions
  - interpret attributes in context of current network topological structure

Actions

- Compare

Targets

- Attributes
  - One
    - Distribution
- Network Data
  - Topology
How: Idiom design decisions

• arrange space for networks
  – custom node-link diagram layout
    • points for nodes
    • connection marks for links
  – vertical compartment according to subcellular location attribute
  – cluster according to functional grouping
How: Arrange space

- automatic layout similar to hand-drawn diagrams
  – vertical compartment according to subcellular location attribute

http://www.nature.com/nri/focus/tlr/nri1397.html
How: Idiom design decisions

• facet: partition data into multiple views
  – juxtapose views side by side
    • same encoding, different data: *small multiples*
    • nodes in each view colored by expression levels for experimental condition
How: Juxtapose vs. animate

Manipulate

Change

• comparison difficult across many frames with many changes everywhere

• rule of thumb: eyes beat memory
  – principle: external cognition vs. internal memory
    • easy to compare by moving eyes between side-by-side views
    • harder to compare memory of what you saw to visible view
Cerebral contributions

• multiple juxtaposed views support interactive comparison between gene expression level experimental data and network context

• automatic network layout algorithm in spirit of hand drawn diagrams
  – localization and functional group attributes affect spatial position

• open source
  – Cytoscape plugin
  – InnateDB database integration

http://www.pathogenomics.ca/cerebral/
Outline

• introduction

• Cerebral

• MizBee

• TreeJuxtaposer

• wrapup
MizBee

A Multiscale Synteny Browser

joint work with:
Miriah Meyer, Hanspeter Pfister

http://www.cs.utah.edu/~miriah/mizbee
MizBee video
What: Data abstraction

• data: multiscale lists
  – features: hundreds of thousands
    • ordered attribute: position in chromosome sequence coordinates
    • categorical attributes: orientation, chromosome of matching feature
    • quantitative attributes: length, similarity score
  – syntenic blocks: thousands
    • contiguous sets of features on same chromosome
    • combine thresholded features if
      – destination chromosome and orientation match
      – close together
  – chromosomes: dozens
  – genomes: two
Why: Tasks in domain language

• analyze conservation (similarity) relationships between genomic features
  – high-level biology questions
    • evolution
      – how long ago did two species share common ancestor?
    • function
      – which segment of the genome is responsible for specific function in the cell?
    • ...
  – low-level data-centric questions
    • algorithm refinement
      – are paired features within a block contiguous?
      – which chromosomes share conserved blocks?
      – are similarity scores alike within block?
      – ...
## Why: Tasks abstraction

- **relationship types:** proximity, size, orientation, similarity
- **data scales:** genome, chromosome, block, feature
- **topics:** algorithm in/out, block reliability, high-level science

<table>
<thead>
<tr>
<th>Question</th>
<th>genome</th>
<th>chromosome</th>
<th>block</th>
<th>feature</th>
<th>proximity / location</th>
<th>size</th>
<th>orientation</th>
<th>similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which chromosomes share conserved blocks?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For one chromosome, how many other chromosomes does it share blocks with?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What is the density of coverage and where are the gaps on: chromosomes?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Where are the blocks: on chromosomes? around a specific location on a chromosome?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are the sizes and locations of other genomic features near a block?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How large are the blocks?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do neighboring blocks go to the same: chromosomes? relative location on a chromosome?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the orientations matched or inverted for: block pairs? feature pairs?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do the orientations match for pairs of: neighboring blocks? features within a block?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are similarity scores alike: with respect to neighboring blocks? within a block?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the paired features within a block contiguous?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How large is a feature relative to other genes within a block?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are the sizes, locations, and names of features within a block?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What are the differences between individual nucleotides of feature pairs?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How: Idiom design choices

- encode match relationships between chromosome segments with both
  - color
    - $src$
    - $dst$
  - connection marks
    - $src$
    - $dst$

Identity Channels: Categorical Attributes
- Spatial region
- Color hue
- Motion
- Shape

Marks As Links
- Containment
- Connection
How: Arrange space

• design space of arrangements

Mauve [Darling04]

Cinteny

MizBee

Apollo [Lewis02]
How: Idiom design choices

• juxtapose linked views
  – *multiform overview-detail*
  • three views: genome, chromosome, block
  • different visual encoding in each

Facet

Juxtapose

<table>
<thead>
<tr>
<th>Data</th>
<th>All</th>
<th>Subset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same</td>
<td>Redundant</td>
<td>Overview/Detail</td>
<td>Small Multiples</td>
</tr>
<tr>
<td>Different</td>
<td><em>Multiform</em></td>
<td>Overview/Detail</td>
<td>No Linkage</td>
</tr>
</tbody>
</table>
How: Idiom design choices

• axis orientation
  – radial: genome
  – rectilinear: chromosome, block
  • aligned position more accurate than angle

Why?
How?
What?
How: Idiom design choices

• filter
How: Idiom design choices

• outer ring: summarize relationships with color
  – select one chromosome from set of source chromosomes

• inner ring:
  – destination chromosomes around copy of selected source chromosome
  – show relationship details with connection marks as well as color
MizBee contributions

• first synteny browser with side-by-side linked views
  – across the range of scales
  – encoding all four conservation relationship types
    • proximity, size, orientation, similarity

• open source
  http://www.cs.utah.edu/~miriah/mizbee
Outline

• introduction

• Cerebral

• MizBee

• TreeJuxtaposer

• wrapup
TreeJuxtaposer

Scalable Tree Comparison using Focus+Context with Guaranteed Visibility

joint work with:
François Guimbretière, Serdar Tasiran, Li Zhang, Yunhong Zhou

TreeJuxtaposer video
**What and why: Data and task abstraction**

- **data: trees**
  - phylogenetic tree reconstruction
    - siblings unordered, interior nodes inferred

- **task: compare topological structure**
  - larger query scopes require more explicit tool support
    - compare several is more difficult than identify/inspect one
      - even trickier: summarize all

- **derived data: structural differences**
  - best corresponding node in other tree

---

**Tables**

<table>
<thead>
<tr>
<th>Dataset Types</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>Query</td>
</tr>
<tr>
<td></td>
<td>Identify</td>
</tr>
<tr>
<td></td>
<td>Compare</td>
</tr>
<tr>
<td></td>
<td>Summarise</td>
</tr>
<tr>
<td>Network Data</td>
<td></td>
</tr>
<tr>
<td>Topology</td>
<td></td>
</tr>
<tr>
<td>Paths</td>
<td></td>
</tr>
</tbody>
</table>

---

**Why?**

- What?
- How?

---

**Actions**

- Identify
- Compare
- Summarise
- Query
- Derive
How: Idiom design decisions

• juxtapose linked views
  – show two tree layouts side by side
  – linked navigation

• encode with color: linked highlighting
  – structural differences
  – corresponding subtree (click select)
  – best corresponding node (hover select)
How: Idiom design decisions

• embed focus+context in single view
  – reduce with complex combination of filtering and aggregation

• distort geometry
  – metaphor: stretch and squish navigation
  – shape: rectilinear
  – foci: multiple
  – impact: global
TreeJuxtaposer contributions

• first interactive tree comparison system
  – derive structural difference data to support comparison task
    • subquadratic algorithm: best corresponding node
  – juxtapose views with cross-dataset linked highlighting

• embed focus+context information in single view with stretch and squish navigation
  – sublinear algorithm: guaranteed visibility of structure marks even when squished

• open source
  http://olduvai.sf.net/tj
Outline

• introduction

• Cerebral

• MizBee

• TreeJuxtaposer

• wrapup
Visualization Analysis and Design

http://www.cs.ubc.ca/~tmm/vadbook/
### What?

- **Data Types**
  - Items
  - Attributes
  - Links
  - Positions
  - Grids

- **Data and Dataset Types**
  - **Tables**
    - Items
    - Attributes
  - **Networks & Trees**
    - Items (nodes)
    - Grids
    - Positions
  - **Geometry**
    - Items
    - Positions
  - **Clusters, Sets, Lists**
    - Items

- **Dataset Types**
  - **Tables**
  - **Networks**
  - **Fields (Continuous)**

- **Geometry (Spatial)**

- **Dataset Availability**
  - Static
  - Dynamic

### Why? How?

- **Attribute Types**
  - Categorical
  - Ordered
    - Ordinal
    - Quantitative

- **Ordering Direction**
  - Sequential
  - Diverging
  - Cyclic
• \{action, target\} pairs
  – discover distribution
  – compare trends
  – locate outliers
  – browse topology
### Encode

<table>
<thead>
<tr>
<th>Arrange</th>
<th>Order</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Express</td>
<td>Align</td>
<td></td>
</tr>
<tr>
<td>Separate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Manipulate

<table>
<thead>
<tr>
<th>Change</th>
<th>Juxtapose</th>
<th>Facet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Reduce

<table>
<thead>
<tr>
<th>Filter</th>
<th>Aggregate</th>
<th>Embed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More Information

• this talk
  http://www.cs.ubc.ca/~tmm/talks.html#vanbug15

• papers, videos, software, talks, courses
  http://www.cs.ubc.ca/group/infovis
  http://www.cs.ubc.ca/~tmm

• book
  http://www.cs.ubc.ca/~tmm/vadbook
    • 20% promo code for book+ebook combo: HVN17
    • http://www.crcpress.com/product/isbn/9781466508910

• acknowledgements
  – funding: Agilent, NSERC, NSF