Visualization Analysis & Design for Biology

Tamara Munzner
Department of Computer Science
University of British Columbia

guest lecture, UBC CPSC 314
24 October 2014, Vancouver BC

http://www.cs.ubc.ca/~tmm/talks.html#biovis-cs314

Outline

- introduction
- Cerebral
- TreeJuxtaposer
- MizBee
- wrapup

Defining visualization (vis)

- vis: application area of computer graphics
  - biology: application area of vis

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

Identical statistics	
x mean	9
x variance	10
y mean	8
y variance	4
x/y correlation	1

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

Expression color scale
-2.5   0   2.5

[Cerebral:Visualizing Multiple Experimental Conditions on a Graph with Biological Context. Barsky, Munzner, Gardy, and Kincaid IEEE TVCG (Proc. InfoVis) 14(6):1253-1260, 2008.]

Why analyze?

Vis usage can be analyzed in terms of what data is shown, why the user needs it, and how the idiom is designed.

- abstractions
  - translate from specifics of domain to vocabulary of vis
    - *data abstraction*: **what** to show
      - might not draw what you're given: transform data into form useful for task
    - *task abstraction*: **why** they're looking at it
- idioms
  - *visual encoding idiom*: **how** to draw
  - *interaction idiom*: **how** to manipulate
- analysis framework: scaffold to think systematically about design space
  - huge, and most possibilities ineffective for particular task/data combination

What?
Why?
How?

Outline

- introduction
- Cerebral
- TreeJuxtaposer
- MizBee
- 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/

Cerebral:Visualizing Multiple Experimental Conditions on a Graph with Biological Context. Barsky, Munzner, Gardy, Kincaid IEEE Trans. Visualization and Computer Graphics 14(6):1253-1260 2008. (Proc. InfoVis 2008).

Cerebral video

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

→ Networks
  Link
  Node (item)

→ Tables
  Attributes (columns)
  Items (rows)
  Cell containing value

→ Attribute Types
  → Categorical
  → Ordered
    → Ordinal
    → Quantitative

What?
Why?
How?

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

→ Discover

♂ Actions
→ Compare

◎ Targets
→ Attributes
  → One
    → Distribution
→ Network Data
  → Topology

What?
Why?
How?

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

Arrange Networks And Trees
→ Node-link Diagrams
  Connections and Marks
  ✔ NETWORKS ✔ TREES

What?
Why?
How?

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

What?
Why?
How?

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

Expression color scale
-2.5   0   2.5

Facet
→ Juxtapose
→ Partition

	Data		
	All	Subset	None
Same	Redundant	Overview/ Detail	*Small Multiples*
Different	Multiform	Multiform, Overview/ Detail	No Linkage

Encoding

What?
Why?
How?

How: Juxtapose vs. animate

Manipulate
→ Change

LPSLL37_1

- comparison difficult across many frames with 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/

Slide 17 — Outline

Outline

- introduction
- Cerebral
- **TreeJuxtaposer**
- MizBee
- wrapup

## Slide 18 — TreeJuxtaposer

TreeJuxtaposer

Scalable Tree Comparison using Focus+Context with Guaranteed Visibility

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

http://www.cs.ubc.ca/labs/imager/tr/2003/tj/

TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility.
Munzner, Guimbretière, Tasiran, Zhang, Zhou. ACM SIGGRAPH 2003.

Slide 19 — TreeJuxtaposer video

TreeJuxtaposer video

Slide 20 — What and why: Data and task abstraction
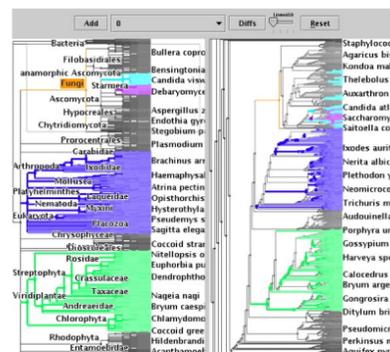
What and why: Data and task abstraction

- 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

Slide 21 — How: Idiom design decisions
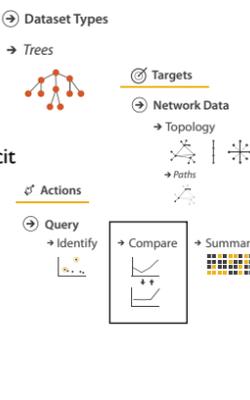
How: Idiom design decisions

- 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)

Slide 22 — How: Idiom design decisions
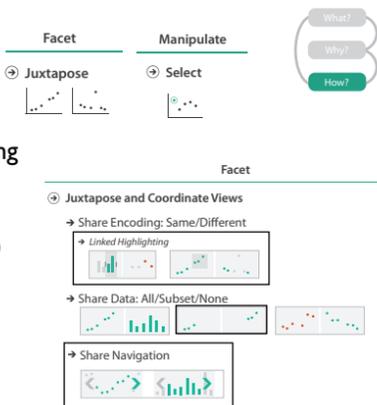
How: Idiom design decisions

- 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

Reduce
- Filter
- Aggregate
- Embed

Slide 23 — TreeJuxtaposer contributions
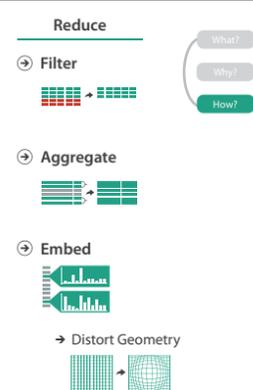
TreeJuxtaposer contributions

- 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

Slide 24 — Outline

Outline

- introduction
- Cerebral
- TreeJuxtaposer
- **MizBee**
- wrapup

Slide 25 — MizBee
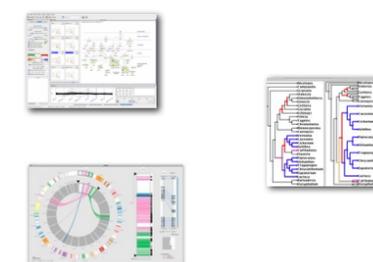
MizBee

A Multiscale Synteny Browser

joint work with:
Miriah Meyer, Hanspeter Pfister

http://www.cs.utah.edu/~miriah/mizbee

MizBee: A Multiscale Synteny Browser.
Meyer, Munzner, Pfister. IEEE Trans. Visualization and Computer Graphics 15(6):897-904, 2009 (Proc. InfoVis 2009).

Slide 26 — MizBee video

MizBee video

Slide 27 — What: Data abstraction
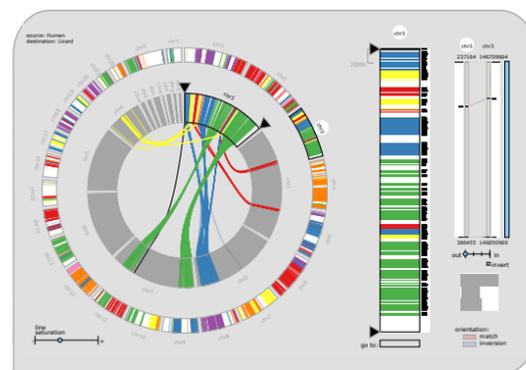
What: Data abstraction

- 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

Slide 28 — Why: Tasks in domain language

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?
      - ...

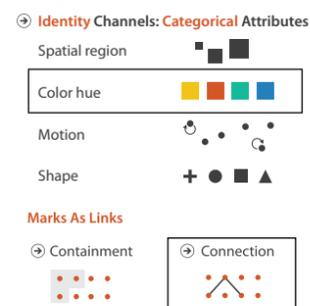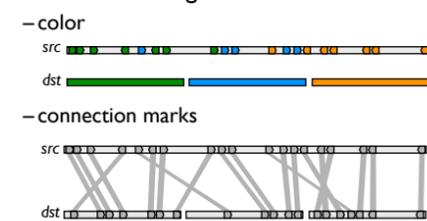## Slide 29 — Why: Tasks abstraction

Why: Tasks abstraction

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

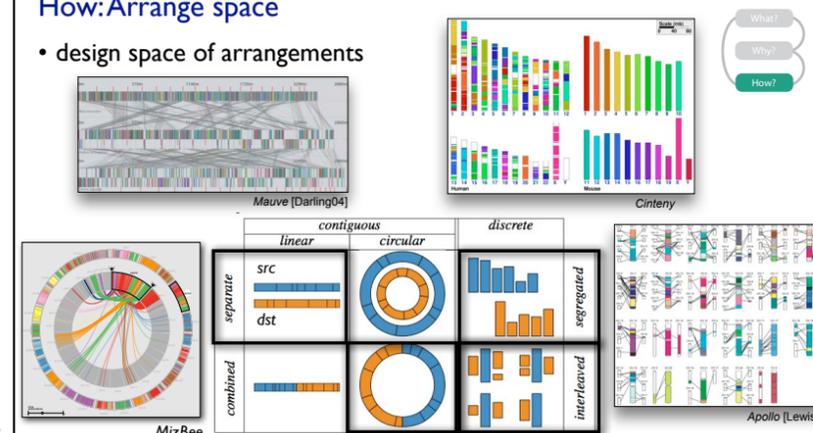|	genome	chromosome	block	feature	proximity / location	size	orientation	similarity
Which chromosomes share conserved blocks?	x	x					x	
For one chromosome, how many other chromosomes does it share blocks with?	x	x				x		
What is the density of coverage and where are the gaps on chromosomes' blocks?	x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

Slide 30 — How: Idiom design choices

How: Idiom design choices

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

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

Marks As Links
- Containment
- Connection

Slide 31 — How: Arrange space

How: Arrange space

- design space of arrangements

Mauve [Darling04]
Cinteny
MizBee
Apollo [Lewis02]

	linear	circular
separate	src / dst	
combined		

contiguous / discrete / segregated / interleaved

Slide 32 — How: Idiom design choices

How: Idiom design choices

- juxtapose linked views
  - *multiform overview-detail*
    - three views: genome, chromosome, block
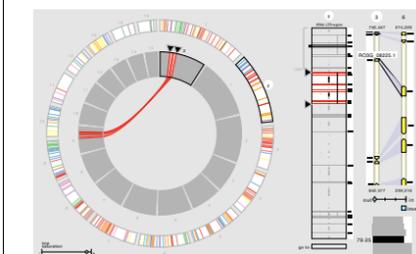   - different visual encoding in each
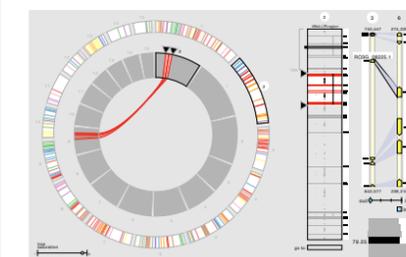
How: Idiom design choices

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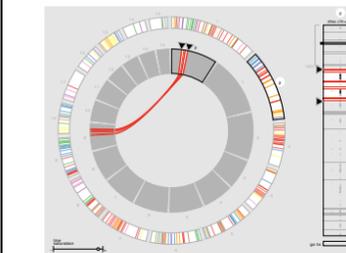   - aligned position more accurate than angle

Arrange

⊕ Axis Orientation
→ Rectilinear   ↑ Parallel   ✳ Radial

⊕ **Magnitude** Channels: **Ordered** Attributes

Position on common scale	
Position on unaligned scale	
Length (1D size)	
Tilt/angle	

33

How: Idiom design choices

- filter

Reduce

⊕ Filter

34

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
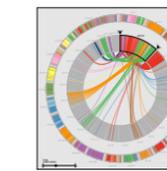
Manipulate

⊕ Select

♂ Actions

⊕ Query
→ Identify   → Compare   → Summarize

35

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

36

Outline

- introduction

- Cerebral

- TreeJuxtaposer

- MizBee

- wrapup

37
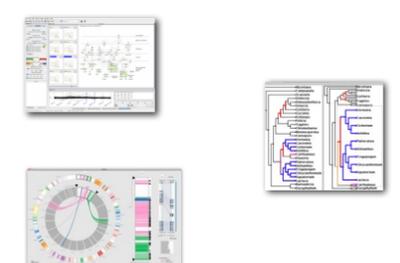
Connections Beyond Graphics

- many other application domains
  - biology only one of many

- machine learning
- computational geometry
- HCI
- cognitive psychology

38

Design Studies: Other Domains

- RelEx: automative networks

- Vismon: fisheries simulation/mgmt

- LiveRAC: large-scale system monitoring

- SessionViewer: web logs

39
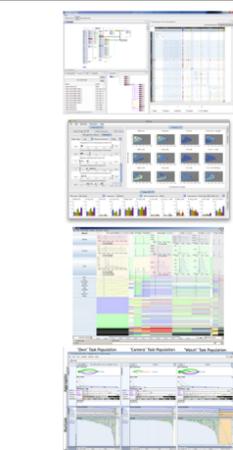
Techniques: Dimensionality Reduction

- Glimmer: GPU accelerated MDS

- DimStiller: visual dimensional analysis and reduction toolkit

- Glint: costly distance functions

40
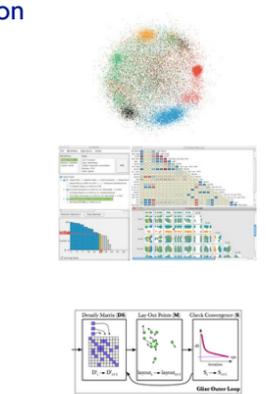
Techniques: Networks

- TopoLayout
  - multi-level network layout

- Grouse
  - multi-level network browsing

- GrouseFlocks
  - browsing space of all possible compound network hierarchies

- TugGraph
 - untangling complex multi-level networks interactively

41
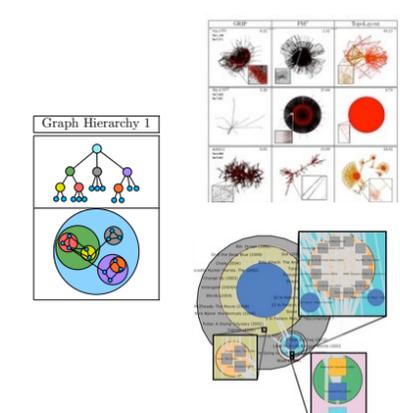
Evaluation: Dimensionality Reduction

- guidance on scatterplot/DR choices

- taxonomy of cluster separation factors
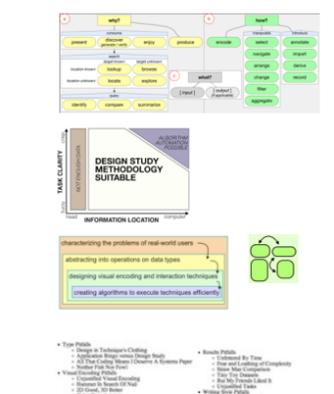
- 2D points vs 3D landscapes

42
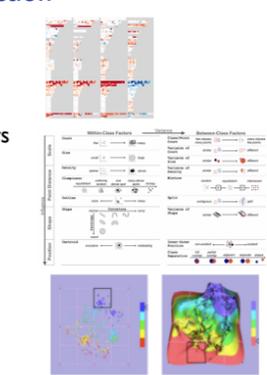
Evaluation: Focus+Context

- overviews: separate vs. integrated views

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

- impact of distortion on visual search, visual memory

43
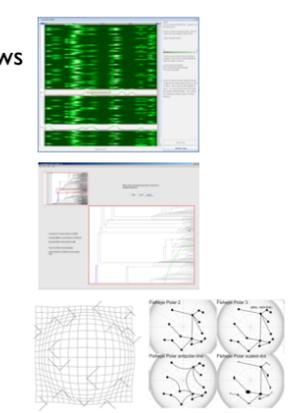
Theory/Models

- multi-level typology of abstract visualization tasks

- design study methodology

- nested model for vis design and validation
  - revisited: blocks and guidelines

- papers process and pitfalls

44

More Information

- this talk
  http://www.cs.ubc.ca/~tmm/talks.html#biovis-cs314
- papers, videos, software, talks, courses
  http://www.cs.ubc.ca/~tmm
- book *(to appear Nov 2014)*
  http://www.cs.ubc.ca/~tmm/vadbook

- acknowledgements
  - funding: Agilent, NSERC, NSF
  - talk feedback: Matt Brehmer

Visualization Analysis & Design
Tamara Munzner

45