Visualization and Biology: Fertile Ground for Collaboration

Tamara Munzner
Department of Computer Science
University of British Columbia

February 2009

http://www.cs.ubc.ca/~tmm/talks.html#harvard09

Outline

• visualization ideas and background
• combining interaction networks, microarray data
  – Cerebral system
• comparing phylogenetic trees
  – TreeJustposer system
• discussion

Why do visualization?

• pictures help us think
  – substitute perception for cognition
  – external memory: free up limited cognitive/memory resources for higher-level problems
• finding gaps, breakdowns, slow downs
  – where conjecture that vis would help

When should we bother doing vis?

• need a human in the loop
  – augment, not replace, human cognition
  – for problems that cannot be (completely) automated
• simple summary not adequate
  – statistics may not adequately characterize complexity of dataset distribution

Characterizing problems

• understanding domain concepts and current workflow
• finding gaps, breakdowns, slowdowns
  – where conjecture that vis would help

Abstracting into operations on data types

• data types
  – tables of numbers
  – relations: networks/graphs, hierarchies/trees
  – spatial data: geographic, positions in space
• operations
  – sorting, filtering, browsing, comparison, characterizing trends and distributions, finding anomalies and outliers, finding correlation...
  – relations: following path through network...

Designing encoding and interaction

• visual encoding
  – marks: points, lines, areas
  – attributes: position, color, shape, size, orientation,
• interaction
  – selecting, navigating, ordering...

Creating efficient algorithms

• classic computer science problem
  – create algorithm given clear specification

Design decisions

• huge space of design alternatives
  – conflicting tradeoffs
  – iterative refinement often necessary
• many/most choices are ineffective
  – wrong visual encoding can mislead, confuse
  – principled reasons to make choices usually not obvious to untrained people

Validation: Is problem solved?

• humans in the loop for outer three levels

Collaboration: Complementary expertise

• vis researchers
  – vis design alternatives
  – human perceptual capabilities
  – scalable graphics algorithms
  – validation methodology
• domain scientists
  – deep knowledge of driving problems, data

Good driving problems for vis research

• big data
• reasonably clear questions
• need for humans in the loop

Outline

• visualization ideas and background
• combining interaction networks, microarray data
  – Cerebral system
• comparing phylogenetic trees
  – TreeJustposer system
• discussion

Visualization and Biology: Fertile Ground for Collaboration

Tamara Munzner
Department of Computer Science
University of British Columbia

February 2009

http://www.cs.ubc.ca/~tmm/talks.html#harvard09

Outline

• visualization ideas and background
• combining interaction networks, microarray data
  – Cerebral system
• comparing phylogenetic trees
  – TreeJustposer system
• discussion

Why do visualization?

• pictures help us think
  – substitute perception for cognition
  – external memory: free up limited cognitive/memory resources for higher-level problems
• finding gaps, breakdowns, slow downs
  – where conjecture that vis would help

When should we bother doing vis?

• need a human in the loop
  – augment, not replace, human cognition
  – for problems that cannot be (completely) automated
• simple summary not adequate
  – statistics may not adequately characterize complexity of dataset distribution

Characterizing problems

• understanding domain concepts and current workflow
• finding gaps, breakdowns, slowdowns
  – where conjecture that vis would help

Abstracting into operations on data types

• data types
  – tables of numbers
  – relations: networks/graphs, hierarchies/trees
  – spatial data: geographic, positions in space
• operations
  – sorting, filtering, browsing, comparison, characterizing trends and distributions, finding anomalies and outliers, finding correlation...
  – relations: following path through network...

Designing encoding and interaction

• visual encoding
  – marks: points, lines, areas
  – attributes: position, color, shape, size, orientation,
• interaction
  – selecting, navigating, ordering...

Creating efficient algorithms

• classic computer science problem
  – create algorithm given clear specification

Design decisions

• huge space of design alternatives
  – conflicting tradeoffs
  – iterative refinement often necessary
• many/most choices are ineffective
  – wrong visual encoding can mislead, confuse
  – principled reasons to make choices usually not obvious to untrained people

Validation: Is problem solved?

• humans in the loop for outer three levels

Collaboration: Complementary expertise

• vis researchers
  – vis design alternatives
  – human perceptual capabilities
  – scalable graphics algorithms
  – validation methodology
• domain scientists
  – deep knowledge of driving problems, data

Good driving problems for vis research

• big data
• reasonably clear questions
• need for humans in the loop

Outline

• visualization ideas and background
• combining interaction networks, microarray data
  – Cerebral system
• comparing phylogenetic trees
  – TreeJustposer system
• discussion
Systems biology model
- graph $G = (V, E)$
  - $V$: proteins, genes, DNA, RNA, tRNA, etc.
  - $E$: interacting molecules

Model - Experiment cycle
- conduct experiments on cells
- interpret results in current graph model
- propose modifications to refine model
- vis tool to accelerate workflow?

Goal: Integrate model with measurements
- system model
  - interaction graph $G = (V, E)$
  - meta-data for each $v$ in $V$
    - labels, biological attributes
  - experimental measurements
    - multiple floats for each $v$ in $V$
    - microarray data

Model summarizes extensive lab work
- graphs come from hand-curated databases
- each edge has provenance from experimental evidence
- FRAM: an iterative model refinement technique
- Toll-like receptor pathway
- MAPK and TNF signaling pathways
- HAIR-PROTEIN interaction
- Francisella tularensis

Cerebral layout using biological metadata
- similar to hand-drawn
- spatial position reveals location in cell
- simulated annealing in $O(E^{1/2}V)$ vs. $O(V^3)$ time

Human interactome: $E=50,000$, $V=10,000$
- too complex, beyond scope of tool

Cerebral video
- interactions generally occur within a compartment
- crossing membranes is interesting
- biological cells divided by membranes

Encoding and interaction design decisions
- create custom graph layout
  - guided by biological metadata
- use small multiple views
  - one view per experimental condition
- show measured data in graph context
  - not in isolation

Traditional graph layout
- given graph $G=(V,E)$
- create layout in 2D plane
  - hundreds of papers
  - annual Graph Drawing conf.
  - Circular (Six and Tollis, 1999)
  - Force-directed (Fruchterman and Reingold, 1991)
  - Hierarchical (Sugiyama, 1989)

Existing layouts did not suit immunologists
- graph drawing goals
  - visualize graph structure
- biologist goals
  - visualize biological knowledge
  - some relationships happen to form a graph
  - cell location also relevant

Biological cells divided by membranes
- interactions generally occur within a compartment
- crossing membranes is interesting

TLR4 biomolecule: $E=74$, $V=54$
- very local view

Immune system: $E=1263$, $V=760$
- bigger picture, target size for Cerebral

Human interactome: $E=50,000$, $V=10,000$
- too complex, beyond scope of tool

Why not animation?
- global comparison difficult

Use small multiple views
- one graph instance per experimental condition
  - same spatial layout
  - color differently, by condition

Hand-drawn diagrams
- cellular location encoded spatially
- infeasible to create by hand in era of big data

Cerebral layout using biological metadata
- similar to hand-drawn
- spatial position reveals location in cell
- simulated annealing in $O(E^{1/2}V)$ vs. $O(V^3)$ time

Why not animation?
- global comparison difficult

Human interactome: $E=50,000$, $V=10,000$
- too complex, beyond scope of tool
**Why not animation?**
- limits of human visual memory
  - compared to side by side visual comparison

**Why not glyphs?**
- embed multiple conditions as a chart inside node
  - clearly visible when zoomed in
  - but cannot see from global view
  - only one value shown in overview

**Show measured data in graph context**
- data driven hypothesis
  - clusters indicate similar function?
  - same pattern of gene expression → same role in cell?
- clusters are often untrustworthy artifacts!
  - noisy data: different clustering alg. → different results
- measured data alone potentially misleading
  - show in context of graph model

**Adoption by biologists**

**Phylogenetic (evolutionary) tree**

**Common dataset size today**

**Future goal: Full Tree of Life, ~10M nodes**
- Plants
- Protists
- Fungi
- Animals

**Operation: Comparing multiple trees**
- presentation: single tree shown as final result
  - exploration: determine true tree from many possibilities
    - different biological conjectures or data
    - different phylogenetic reconstruction algorithms
    - multiple alternatives from same reconstruction algorithm
  - most previous work on browsing
    - necessary but not sufficient for comparison

**Limitations of paper: Scale and speed**
- literal: actual paper
  - figurative: interfaces with same semantics as paper
  - need to focus on details
  - yet maintain context

**TreeJuxtaposer**
- collaboration with biologists at UT-Austin Hills Lab
  - open-source software download
  - http://evolution.sourceforge.net/

**Encoding and interaction design decisions**
- linking tree views through node correspondences
  - showing structural differences
- guaranteed visibility of small marks
  - scaling up to millions of nodes

**InnateDB links to Cerebral**
- InnateDB: facilitating systems-level analyses of the mammalian innate immune response
  - http://innatedb.ca

**Data cleansing example**
- incorrect edge across many compartments
  - in well studied dataset
  - not obvious with other layouts

**Cerebral summary**
- supports interactive exploration of multiple experimental conditions in graph context
- provides familiar representation by using biological metadata to guide graph layout

**Outline**
- visualization ideas and background
- combining interaction networks, microarray data
  - Cerebral system
- comparing phylogenetic trees
  - TreeJuxtaposer system
- discussion
Showing differences via contiguous groups
- clade: ancestor + all descendants
- biological questions to support
  - is a clade in one tree also a clade in other?
  - is some group a clade?

Best corresponding node between trees

Guaranteed visibility
- marks are always visible
  - structural differences, search results, user selections
- easy with small datasets
  - regions of interest shown with color highlights

Guaranteed visibility challenges
- hard with larger datasets
- reasons a mark could be invisible

Guaranteed visibility challenges
- hard with larger datasets
- reasons a mark could be invisible

Guaranteed visibility benefits
- with GV
  - no mark is visible means no need to explore area further
- without GV
  - risk of false negative conclusions, or
  - user must do tedious exhaustive search to ensure nothing missed
- algorithm scalability challenge
  - rendering complexity based on number of onscreen pixels
    - not total number of items in dataset
    - \( O(n \log^2 n) \), vs. naive \( O(n^2) \)

Constrained navigation for visibility
- stretch and squish navigation
  - stretch out part of surface, the rest squishes
  - borders nailed down
  - integrated focus and context
- items never fall outside camera
  - but squished regions can have many items per pixel

Guaranteed visibility challenges
- hard with larger datasets
- reasons a mark could be invisible
  - mark outside the window
    - solution: constrained navigation
  - mark underneath other marks
    - solution: use 2D not 3D layout
- mark smaller than a pixel
  - solution: smart culling

Smart culling for small item visibility
- naive culling does not draw all marked items
  - graphics cards optimized for realism: small items far away and thus not important
  - rendering infrastructure for visualization semantics: small items might be critical

Guaranteed visibility benefits
- with GV
  - no mark is visible means no need to explore area further
- without GV
  - risk of false negative conclusions, or
  - user must do tedious exhaustive search to ensure nothing missed
- algorithm scalability challenge
  - rendering complexity based on number of onscreen pixels
    - not total number of items in dataset
    - \( O(n \log^2 n) \), vs. naive \( O(n^2) \)

Constrained navigation for visibility
- stretch and squish navigation
  - stretch out part of surface, the rest squishes
  - borders nailed down
  - integrated focus and context
- items never fall outside camera
  - but squished regions can have many items per pixel

Guaranteed visibility challenges
- hard with larger datasets
- reasons a mark could be invisible
  - mark outside the window
    - solution: constrained navigation
  - mark underneath other marks
    - solution: use 2D not 3D layout
- mark smaller than a pixel
  - solution: smart culling

Guaranteed visibility challenges
- hard with larger datasets
- reasons a mark could be invisible

TJ summary
- first interactive tree comparison system
  - automatic structural difference computation
  - guaranteed visibility of small marks
- scalable to large datasets
  - 250K to 500K total nodes: original
  - up to 4M nodes: later, with PRISAD
  - subquadratic preprocessing
  - sublinear realtime rendering
  - depends on number of pixels, not number of nodes

Crosscutting themes
- workflow speedups
  - inspecting microarray data with graph
    - minutes vs. hours/days
    - comparing clades between trees
    - seconds vs. hours/days
- contributions from biologist collaborators
  - driving problems and data
  - tool use during iterative refinement

Outline
- visualization ideas and background
    - combining interaction networks, microarray data
      - Cerebral system
    - comparing phylogenetic trees
      - TreeJuxtaposer system
- discussion

Many more, bio/health + others...
- NIH/NCF Visualization Research Challenges Report
  - Johnson, Moorhead, Munzner, Pfister, Rheingans, and Yoo (eds.), IEEE Press
  - http://vgtc.org/wpmu/techcom/?page_id=11

Many other bio/vis research areas
- multiple sequence alignment
  - SequenceJuxtaposer
    - open-source software download
      - http://olduvai.sourceforge.net/sj

Many other bio/vis research areas
- microarray data
  - Hierarchical Clustering Explorer
  - Seo and Shneiderman, U Maryland

Many other bio/vis research areas
- microarray data
  - Hierarchical Clustering Explorer
  - Seo and Shneiderman, U Maryland


Crosscutting themes
- workflow speedups
  - inspecting microarray data with graph
    - minutes vs. hours/days
    - comparing clades between trees
    - seconds vs. hours/days
- contributions from biologist collaborators
  - driving problems and data
  - tool use during iterative refinement
Vast opportunities

- young field, still much to be done
- think about your current workflow
  - what could you speed up by swapping in perception for cognition?
  - exploit the familiar, yet consider breadth of design alternatives
- finding some friendly neighborhood vis collaborators
  - IEEE VisWeek 2009 (Vis, InfoVis, VAST)
    Oct 11-16, Atlantic City
    http://vis.computer.org/VisWeek2009
  - EuroVis 2009: Jun 10-12, Berlin
    http://www.zib.de/eurovis09

More information

- this talk
  http://www.cs.ubc.ca/~tmm/talks.html#harvard09
- papers, videos
  http://www.cs.ubc.ca/~tmm
- software
  http://olduvai.sourceforge.net/
  http://www.pathogenomics.ca/cerebral
  http://www.innatedb.ca