Visualization and Biology: Fertile Ground for Collaboration

Tamara Munzner
Department of Computer Science
University of British Columbia
June 2009

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

Why do visualization?
• pictures help us think
  – substitute perception for cognition
  – external memory: free up limited cognitive/memory resources for higher-level problems

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

http://www.cs.ubc.ca/~tmm/talks.html#eindhoven09

Collaboration: Complementary expertise
• vis researchers
  – vis design alternatives
  – human perceptual capabilities
  – scalable graphics algorithms
  – validation methodology
• domain scientists
  – deep knowledge of driving problems, data
  – both benefit from new tools
    – scientist: you get something helpful
    – vis researcher: we get to watch you use it
  – see if problem actually solved
  – need new knowledge back into our design principles

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

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

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

Validation: Is problem solved?
• humans in the loop for outer three levels

Good driving problems for vis research
• big data
• reasonably clear questions
• need for humans in the loop
• many areas of science are a great match
  – biology particularly appealing

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

http://www.innatedb.ca
http://www.pathogenomics.ca/cerebral/
open-source software download (Cytoscape plugin)
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

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

Immune system: $E=1263$, $V=760$

- bigger picture, target size for Cerebral

Human interactome: $E \approx 50,000$, $V \approx 10,000$

- too complex, beyond scope of tool

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

Model summarizes extensive lab work

- graphs come from hand-curated databases
  - dynamic, change with each new publication
- each edge has provenance from experimental evidence

Why not animation?

- global comparison difficult
- interactions generally occur within a compartment
- crossing membranes is interesting

Biological cells divided by membranes

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 $O(E + V)$ vs. $O(V^2)$ time

Use small multiple views

- one graph instance per experimental condition
  - same spatial layout
  - color differently, by condition

Why not animation?

- choose scope to manage complexity

TLR4 biomolecule: $E=74$, $V=54$

- very local view

Cerebral video

- similar to hand-drawn
  - spatial position reveals location in cell
  - simulated annealing in $O(E + V)$ vs. $O(V^2)$ time
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

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

InnateDB links to Cerebral
- InnateDB: facilitating systems-level analyses of the mammalian innate immune response

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

Phylogenetic (evolutionary) tree

Common dataset size today

Future goal: Full Tree of Life, ~10M nodes

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

Encoding and interaction design decisions
- guaranteed visibility of small marks
  - scaling up to millions of nodes

Operation: Comparing multiple trees video
- stretch and squish navigation
- linked side by side 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 Hillis Lab
- Scalable Tree Comparison using Focus+Context with Guaranteed Visibility
  - http://innatedb.ca
  - http://olduvai.sourceforge.net/tj

Adoption by biologists

InnateDB summary
- InnateDB: facilitating systems-level analyses of the mammalian innate immune response

Future goal: Full Tree of Life, ~10M nodes
- InnateDB: facilitating systems-level analyses of the mammalian innate immune response
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
  - 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

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

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

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

More information
- this talk
  [http://www.cs.ubc.ca/~tmm/talks.html#eindhoven09](http://www.cs.ubc.ca/~tmm/talks.html#eindhoven09)
- papers, videos
  [http://www.cs.ubc.ca/~tmm](http://www.cs.ubc.ca/~tmm)
- software
  [http://olduvai.sourceforge.net/tj](http://olduvai.sourceforge.net/tj)
  [http://www.pathogenomics.ca/cerebral](http://www.pathogenomics.ca/cerebral)
  [http://www.innatedb.ca](http://www.innatedb.ca)