Applying Information Visualization Principles to Biological Network Displays

Tamara Munzner
University of British Columbia
Human Vision and Electronic Imaging 2011
25 Jan 2011

Outline
• visualization principles
• Cerebral system
  – combining interaction networks with microarray data
• Pathline system
  – combining multiple genes, time points, species, and pathways

Designing visual encoding and interaction tech
• visual encoding
  – marks: points, lines, areas
  – attributes: position, color, shape, size, orientation...
• interaction
  – selecting, regimeing, ordering...
• threat to validity: the way you show it doesn’t work

Creating algorithms to execute techniques
• classic computer science problem
  – create algorithm given clear specification
• threat to validity: your code is too slow

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

Separate visualization concerns into four levels

Characterizing problems of real-world users
• understanding domain concepts and current workflow
• finding gaps, breakdowns, slowdowns
  – where conjecture that vis would help
• threat to validity: users don’t do that

Abstracting into operations on data types
• operations
  – sorting, filtering, browsing, comparison, characterizing trends and distributions, finding anomalies and outliers, finding correlation...
• data types
  – number tables, relational networks, spatial
  – transform into useful configuration: derived data
• threat to validity: you’re showing them the wrong thing

What does visualization allow?
• discovering new things
  – hypothesis generation, discovery, eureka moment
• confirming conjectured things
  – hypothesis confirmation
• contradicting conjectured things
  – especially (inevitably!) data cleansing
• novel capabilities
  – tool supports fundamentally new operations
• speedup
  – tool accelerates workflow (most common!)

Principles in action: walk through examples
• vis work in many domains
  – topology
  – computer networking
  – computational linguistics
  – web logs
  – large-scale system administration
  – biology

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

Cerebral
Comparing Multiple Experimental Conditions Within Biologically Meaningful Network Context
joint work with:
Aimee Barrick, Jennifer Gardy, Robert Kiezak
http://www.pathogenomics.ca/cerebral/

MizBee
A Browser for Comparative Genomics Data
joint work with:
Mihir Mayer, Hanspeter Pfister
http://www.mizbee.org

MulteeSum
A Tool for Exploring Space-Time Expression Data
joint work with:
Mihir Mayer, Angela DePace, Hanspeter Pfister
http://www.multeesum.org

TreeJuxtaposer
Scalable Phylogenetic Tree Comparison
joint work with:
François Guimbretière, Serdar Tasiran, Li Zhang, Yunhong Zhou
http://olduvai.sf.net/tj

A Nested Model for Visualization Design and Validation
Munzner IEEE InfoVis 2009


http://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg
Systems biology model
- graph $G = (V, E)$
  - $V$: proteins, genes, DNA, RNA, tRNA, etc.
  - metadata: labels, biological attributes
  - $E$: interacting molecules
  - known from previous research

Cerebral video

Biological cells divided by membranes
- interactions generally occur within a compartment
- interaction location often known as part of model

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

Why not animation?
- global comparison difficult

Why not animation?
- limits of human visual memory
  - compared to side by side visual comparison


Cycle: model - experiment
- conduct experiments on cells
  - microarrays
  - measurements for each vertex in graph
- interpret results in current graph model
- propose modifications to refine model

- visual tool to accelerate workflow
  - integrated tool to see graph and measurements together
  - choose scope for problem complexity

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

Choice: Create custom graph layout
- graph layout heavily studied
  - given graph $G = (V, E)$,
  - create layout in 2D/3D plane
  - hundreds of papers
  - annual Graph Drawing conf.

Choice 2: Use small multiple views
- one graph instance per experimental condition
  - same spatial layout
  - color differently, by condition

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

Choice: Show measures and graph
- why not measurements alone?
  - data driven hypothesis: gene expression clusters indicate similar function in cell?

- clusters are often untrustworthy artifacts!
  - noisy data; different clustering alg. $\rightarrow$ different results
  - measured data alone potentially misleading
  - show in context of graph model

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

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

- 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

- simulated annealing
  - in $O(E\cdot \sqrt{V})$ vs. $O(V^3)$ time

Why not glyphs?
- why not measurements alone?
  - data driven hypothesis: gene expression clusters indicate similar function in cell?

- clusters are often untrustworthy artifacts!
  - noisy data; different clustering alg. $\rightarrow$ different results
  - measured data alone potentially misleading
  - show in context of graph model

Image credit: Dr. G. Weaver, Colorado University at Denver
Biological cells divided by membranes

Lay out using biological metadata
- similar to hand-drawn spatial position reveals location in cell

Choice: Lay out using biological metadata
- simulated annealing
  - $O(E \cdot \sqrt{V})$ vs. $O(V^3)$ time
Contributions

- Cerebral
  - supports interactive exploration of multiple experimental conditions in graph context
  - provides familiar representation by using biological metadata to guide graph layout
- tool deployment
  - open source, Cytoscape plugin
  - used by target group of collaborators
  - 5 citations, showcased in http://tinyurl.ca
  - many more independent adopters
  - 12+ bio lcs citations with Cerebral diagrams so far

Pathline
A Tool for Comparative Functional Genomics Data
joint work with
Miriah Meyer, Bang Wong, Mark Styczynski, Hanspeter Pfister
http://www.pathline.org

biologists measure it...
... how do the gene interactions vary across different species?

Pathline A Tool for Comparative Functional Genomics

consistent data... how do the gene interactions vary across different species?

functional genomics
how do genes work together to perform different functions in a cell?

functional genomics data
gene expression
molecular pathways

metabolic pathways
genes

gene expression

pathways

metabolic pathways
• 10 to 50 pathways of interest
• inputs/outputs called metabolites
• directed graph

pathways

gene expression
• 6000 genes and 140 metabolites
• 14 time points
• 14 species of yeast
• 3D table

gene expression

pathways

gene expression
• 6000 genes and 140 metabolites
• 14 time points
• 14 species of yeast
• 3D table

Tasks
study expression data as a time series
compare a limited number of time series
compare similarity scores along a pathway(s)
comparison of multiple similarity scores

collaborators: Regav Lab at the Broad Institute
biology: metabolism in yeast
data: multiple genes
multiple time points
multiple related species
multiple pathways
problem: existing tools can only look at a subset of this data

similarity scores

phylogeny
• evolutionary relationship
• binary tree

similarities
aggregate time series
over species
similarity of expression
across species
aggregate: Pearson, Spearman, others
quantitative value

phylogeny
• evolutionary relationship
• binary tree
Encode quantitative values with spatial position
- heatmap
- curvemap

Linearized pathway
- common axes to compare similarity scores
- bars and circles
- visual layer for selective attention
- color-code gene direction
- multiple similarity scores

Curvemap
- alternative to heatmaps
- base visual unit is a curve
- filled, framed line charts to enhance shape perception

Contributions
- Pathline
  - multiple genes, time points, species, and pathways
- new visual encoding techniques based on infovis principles and biology needs
- linearized pathway representation
- curvemap
- tool deployment
  - open source
  - used daily by several collaborators

Principle: use validation methods tuned to level
- is target problem really solved?
- what have we learned about tradeoffs in design space?

More information
- principles in more depth: vis intro book chapter
  http://www.cs.ubc.ca/~tmm/papers.html#akpchapter
- papers, talks, videos, courses
  http://www.cs.ubc.ca/~tmm
- this talk
  http://www.cs.ubc.ca/~tmm/talks.html#hvei11

A Nested Model for Visualization Design and Validation.
Muñoz, IEEE InfVis 2009.