Visualization and Biology: Fertile Ground for Collaboration

Tamara Munzner
Department of Computer Science
University of British Columbia

June 2009

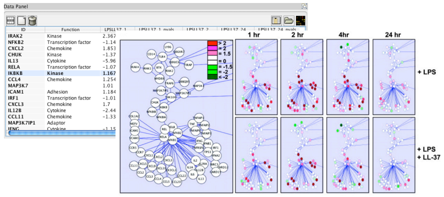
http://www.cs.ubc.ca/~tmm/talks.html#eindhoven09

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



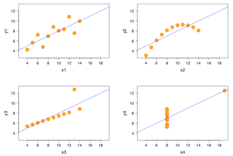
Data Panel

ID	Function	LPSLL37_1	LPSLL37_1_pvals	LPSLL37_2	LPSLL37_24	LPSLL37_24_pvals
IRAK2	Kinase	2.367				
NFKB2	Transcription factor	-1.14				
CXCL2	Chemokine	1.853				
CHUK	Kinase	-1.37				
IL13	Cytokine	-5.96				
RELA	Transcription factor	-1.07				
IKBKB	Kinase	1.167				
CCL4	Chemokine	1.254				
MAP3K7		1.01				
ICAM1	Adhesion	1.184				
IRF1	Transcription factor	-1.01				
CXCL3	Chemokine	1.7				
IL12B	Cytokine	-2.44				
CCL11	Chemokine	-1.33				
MAP3K7IP1	Adaptor					
IFNG	Cytokine	-1.15				

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

Anscombe's quartet: same
- mean
- variance
- correlation coefficient
- linear regression line



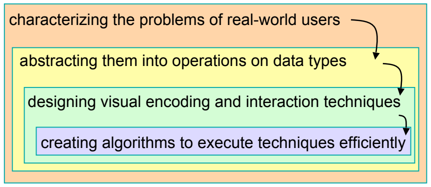
http://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg

What does visualization allow?

- discovering new things
  - hypothesis 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!)

Multiple levels of problem-driven vis

- cascading levels: output above is input below



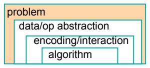
characterizing the problems of real-world users

abstracting them into operations on data types

designing visual encoding and interaction techniques

creating algorithms to execute techniques efficiently

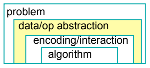
Characterizing problems



problem
data/op abstraction
encoding/interaction
algorithm

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

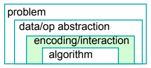
Abstracting into operations on data types



problem
data/op abstraction
encoding/interaction
algorithm

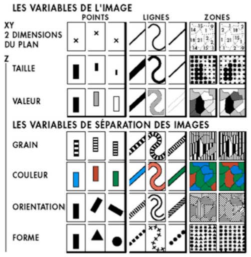
- data types
  - tables of numbers
  - relations: networks/graphs, hierarchies/trees
  - spatial data: geographic, positions in space
- operations
  - sorting, filtering, browsing, comparison, characterizing trends and distributions, finding anomalies and outliers, finding correlation...
  - relations: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

Designing encoding and interaction



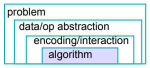
problem
data/op abstraction
encoding/interaction
algorithm

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



Semiology of Graphics. Jacques Bertin, Gauthier-Villars 1967, EHESS 1998

Creating efficient algorithms



problem
data/op abstraction
encoding/interaction
algorithm

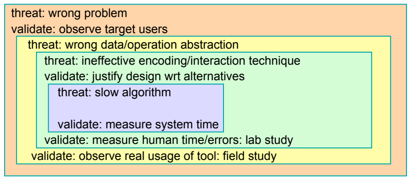
- classic computer science problem
  - create algorithm given clear specification

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



threat: wrong problem
validate: observe target users
threat: wrong data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify design wrt alternatives
threat: slow algorithm
validate: measure system time
validate: measure human time/errors: lab study
validate: observe real usage of tool: field study

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
    - feed new knowledge back into our design principles

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

Cerebral

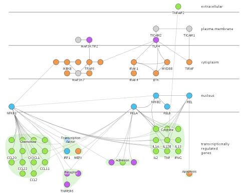
collaboration with researchers at UBC Hancock Lab studying innate immunity

Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context
Aaron Barsky, Computer Science, UBC
Tamara Munzner, Computer Science, UBC
Jennifer Gardy, Microbiology and Immunology, UBC
Robert Kincaid, Agilent Technologies
IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 2008) 14(6) (Nov-Dec) 2008, p 1253-1260.
http://www.cs.ubc.ca/labs/imager/tr/2008/cerebral/
http://www.cs.ubc.ca/labs/imager/th/2008/BarskyMscThesis/

open-source software download (Cytoscape plugin)
http://www.pathogenomics.ca/cerebral/
deployed in InnateDB (mammalian innate immunity database)
http://www.innatedb.ca

Systems biology model

- graph G = {V, E}
  - V: proteins, genes, DNA, RNA, tRNA, etc.
  - E: interacting molecules

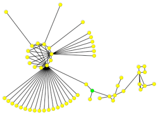


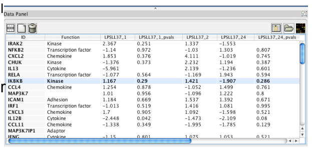
Model - Experiment cycle

- 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

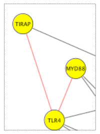




ID	Function	LPSLL37_1	LPSLL37_1_pvals	LPSLL37_2	LPSLL37_24	LPSLL37_24_pvals
IRAK2	Kinase	2.367	0.251	1.337	-1.553	
NFKB2	Transcription factor	-1.14	0.972	-1.03	1.303	0.807
CXCL2	Chemokine	1.853	0.376	4.111	-1.019	0.745
CHUK	Kinase	-1.376	0.373	2.232	1.194	0.387
IL13	Cytokine	-5.961		2.139	-1.236	0.601
RELA	Transcription factor	-1.077	0.564	-1.169	1.943	0.594
IKBKB	Kinase	1.167	0.29	1.421	-1.907	0.286
CCL4	Chemokine	1.254	0.878	-1.052	1.499	0.761
MAP3K7		1.01	0.956	-1.096	1.222	0.8
ICAM1	Adhesion	1.184	0.669	1.537	1.392	0.671
IRF1	Transcription factor	-1.013	0.519	1.416	1.081	0.995
CXCL3	Chemokine	1.7	0.905	1.092	-1.598	0.521
IL12B	Cytokine	-2.448	0.042	-1.473	-2.109	0.08
CCL11	Chemokine	-1.338	0.349	-1.995	-1.785	0.129
MAP3K7IP1	Adaptor					
IFNG	Cytokine	-1.15	0.801	1.075	1.053	0.521

Model summarizes extensive lab work

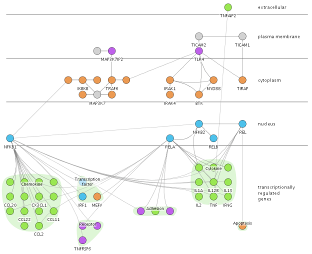
- graphs come from hand-curated databases
  - dynamic, change with each new publication
- each edge has provenance from experimental evidence
  - TIRAP: an adapter molecule in the Toll signaling pathway. *Horng T, Barton GM, Medzhitov R.*
  - Mal (MyD88-adapter-like) is required for Toll-like receptor-4 signal transduction. *Fitzgerald KA, Palsson-McDermott EM, Bowie AG, Jefferies CA, Mansell AS, Brady G, Brint E, Dunne A, Gray P, Harte MT, McMurray D, Smith DE, Sims JE, Bird TA, O'Neill LA.*



- choose scope to manage complexity

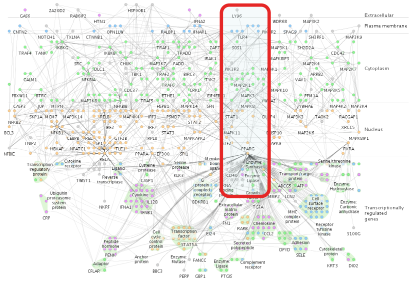
TLR4 biomolecule: E=74, V=54

- very local view



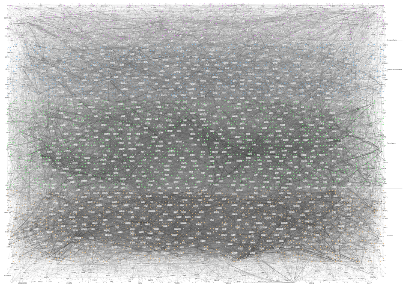
Immune system: E=1263, V=760

- bigger picture, target size for Cerebral

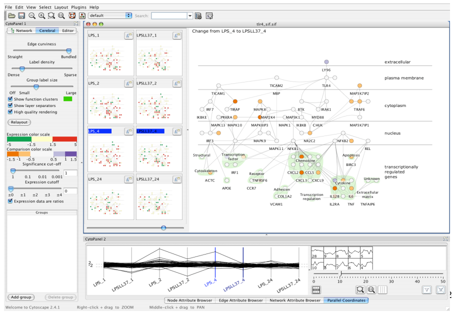


Human interactome: E~50,000, V~10,000

- too complex, beyond scope of tool



Cerebral video

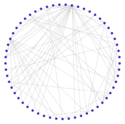


Encoding and interaction design decisions

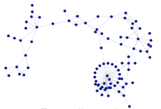
- create custom graph layout
  - guided by biological metadata
- use small multiple views
  - one view per experimental condition
- show measured data in graph context
  - not in isolation

Traditional graph layout

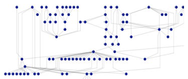
- given graph G={V,E}
- create layout in 2D plane
- heavily studied
  - hundreds of papers
  - annual Graph Drawing conf



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

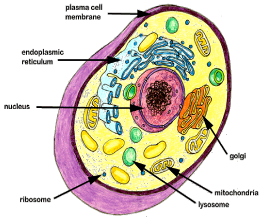
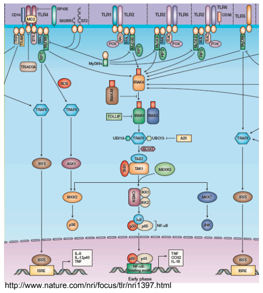


Image credit: Dr.G Weaver, Colorado University at Denver

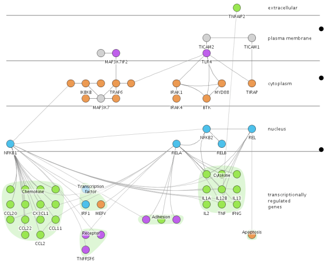
Hand-drawn diagrams



http://www.nature.com/nri/focus/tlr/nri1397.html

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

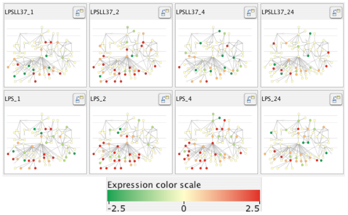
Cerebral layout using biological metadata



- similar to hand-drawn
- spatial position reveals location in cell
- simulated annealing in O(E$\sqrt{V}$) vs. O(V^3) time

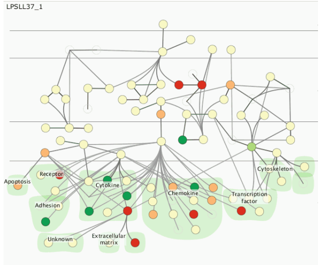
Use small multiple views

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



Why not animation?

- global comparison difficult



Why not animation?

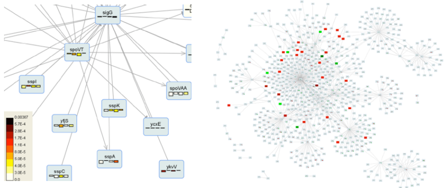
- limits of human visual memory
  - compared to side by side visual comparison

- Matthew Plumlee and Colin Ware. Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. *ACM Trans. Computer-Human Interaction (ToCHI)*,13(2):179-209, 2006.

- Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247-262, 2002.

Why not glyphs?

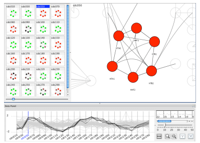
- embed multiple conditions as a chart inside node
- clearly visible when zoomed in
- but cannot see from global view
  - only one value shown in overview



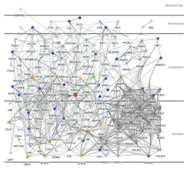
[M. A. Westenberg, S. A. F. T. van Hijum, O. P. Kuipers, J. B. T. M. Roerdink. Visualizing Genome Expression and Regulatory Network Dynamics in Genomic and Metabolic Context. Computer Graphics Forum, 27(3):887-894, 2008.]

Show measured data in graph context

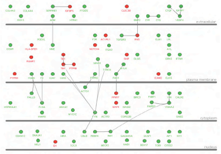
- data driven hypothesis
  - clusters indicate similar function?
  - same pattern of gene expression → same role in cell?
- clusters are often untrustworthy artifacts!
  - noisy data: different clustering alg. → different results
  - measured data alone potentially misleading
  - **show in context of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



Adoption by biologists



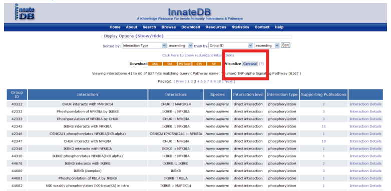
- Matthew D Dyer, T. M Murali, and Bruno W Sobral. The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathogens, 4(2):e32, 2008.



- Liqun He et al. The glomerular transcriptome and a predicted protein-protein interaction network. Journal of the American Society of Nephrology, 19(2):260-268, 2008.

InnateDB links to Cerebral

- InnateDB: facilitating systems-level analyses of the mammalian innate immune response
  - David J Lynn, Geoffrey L Winsor, Calvin Chan, Nicolas Richard, Matthew R Laird, Aaron Barsky, Jennifer L Gardy, Fiona M Roche, Timothy H W Chan, Naisha Shah, Raymond Lo, Misbah Naseer, Jaimmie Que, Melissa Yau, Michael Acab, Dan Tulpan, Matthew D Whiteside, Avinash Chikatamarla, Bernadette Mah, Tamara Munzner, Karsten Hokamp, Robert E W Hancock, Fiona S L Brinkman. Molecular Systems Biology 2008; 4:218
  - http://innatedb.ca



InnateDB

A Knowledge Resource For Innate Immunity Interactions & Pathways

Display Options (Show/Hide)

Sorted by: Interaction Type ascending then by Group ID ascending Sort

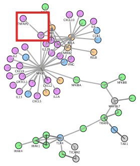
Click here to show redundant interactions

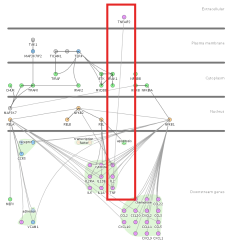
Download XML TAB MS Excel CSV SIF Visualize Cerebral (?)

Viewing interactions 41 to 60 of 857 hits matching query (Pathway name: '(human) TNF-alpha Signaling Pathway [816]')

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

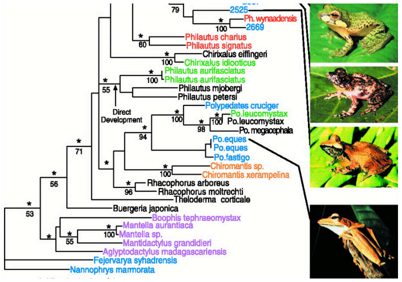
TreeJuxtaposer

collaboration with biologists at UT-Austin Hillis Lab

TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility.
Tamara Munzner, François Guimbretière, Serdar Tasiran, Li Zhang, Yunhong Zhou.
ACM Trans. Graphics 22(3): 453-462, 2003 (Proc. SIGGRAPH 2003).
http://www.cs.ubc.ca/labs/imager/tr/2003/tj

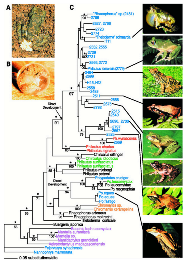
open-source software download
http://olduvai.sourceforge.net/tj

Phylogenetic (evolutionary) tree



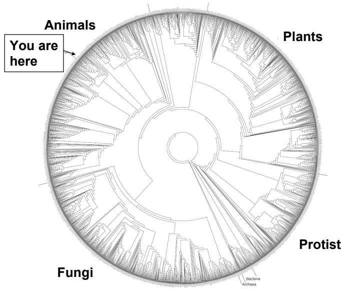
M Meegaskumbura et al., Science 298:379 (2002)

Common dataset size today



M Meegaskumbura et al., Science 298:379 (2002)

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



David Hillis, Science 300:1687 (2003)

Operation: Comparing multiple trees

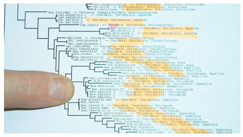
- presentation: single tree shown as final result

- exploration: determine true tree from many possibilities
  - different biological conjectures or data
  - different phylogenetic reconstruction algorithms
  - multiple alternatives from same reconstruction algorithm

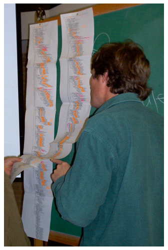
- most previous work on browsing
  - necessary but not sufficient for comparison

Limitations of paper: Scale and speed

- literal: actual paper
- figurative: interfaces with same semantics as paper



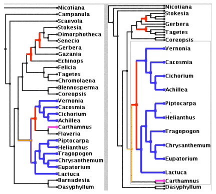
need to focus on details

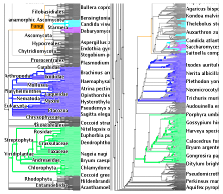


yet maintain context

TreeJuxtaposer video

- stretch and squish navigation
- linked side by side comparison



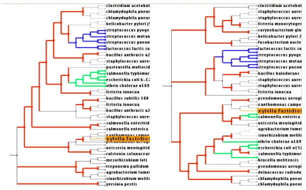


Encoding and interaction design decisions

- guaranteed visibility of small marks
  - scaling up to millions of nodes

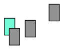
Guaranteed visibility

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



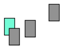
Guaranteed visibility challenges

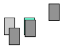
- hard with larger datasets
- reasons a mark could be invisible



Guaranteed visibility challenges

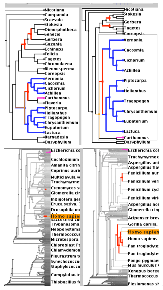
- hard with larger datasets
- reasons a mark could be invisible
  - mark outside the window
    - solution: constrained navigation





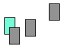
Constrained navigation for visibility

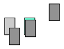
- stretch and squish navigation
  - stretch out part of surface, the rest squishes
  - borders nailed down
  - integrated focus and context
- items never fall outside camera
  - but squished regions can have many items per pixel

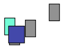


Guaranteed visibility challenges

- 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

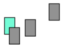


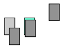


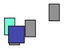


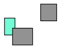
Guaranteed visibility challenges

- 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



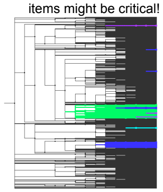




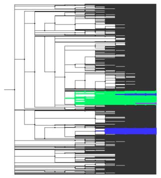


Smart culling for small item visibility

- naïve 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 mark visibility



no guaranteed visibility

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
    - Partitioned Rendering Infrastructure for Scalable Accordion Drawing (Extended Version). James Slack, Kristian Hildebrand, and Tamara Munzner. Information Visualization, 5(2), p. 137-151, 2006
    - Composite Rectilinear Deformation for Stretch and Squish Navigation. James Slack and Tamara Munzner. Proc. Visualization 2006, published as Transactions on Visualization and Computer Graphics 12(5), September 2006.

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
- papers, videos
  http://www.cs.ubc.ca/~tmm
- software
  http://olduvai.sourceforge.net/tj

  http://www.pathogenomics.ca/cerebral

  http://www.innatedb.ca