Applying Information Visualization Principles to Biological Network Displays

Tamara Munzner
University of British Columbia

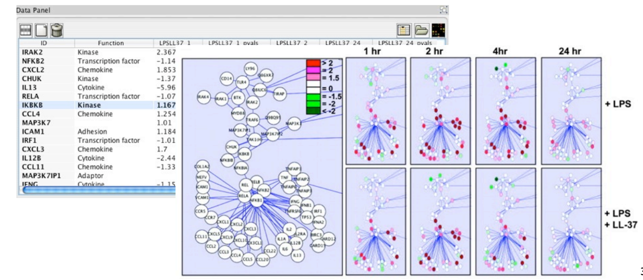
Human Vision and Electronic Imaging 2011
25 Jan 2011

Outline

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

Why do visualization?

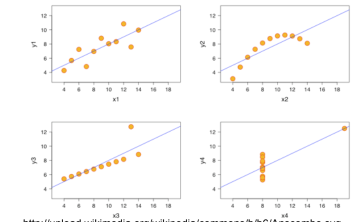
- pictures help us think
  - substitute perception for cognition
  - external memory: free up limited cognitive/memory resources for higher-level problems



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

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

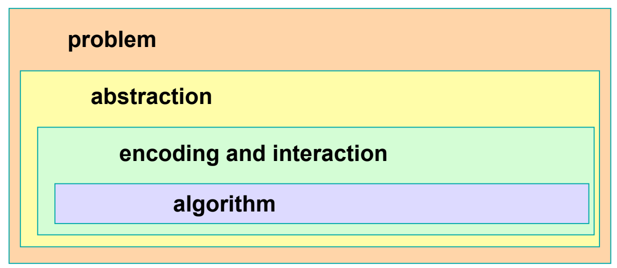


http://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg

What does visualization allow?

- discovering new things
  - hypothesis 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!)

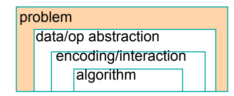
Separate visualization concerns into four levels



- different threats to validity at each level

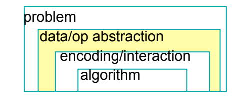
A Nested Model for Visualization Design and Validation
Munzner. IEEE InfoVis 2009.

Characterizing problems of real-world users



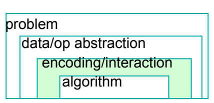
- understanding domain concepts and current workflow
- finding gaps, breakdowns, slowdowns
  - where conjecture that vis would help
- threat to validity: users don't do that

Abstracting into operations on data types

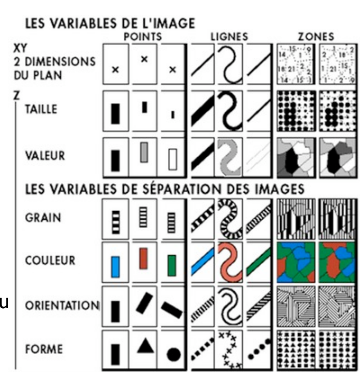


- operations
  - sorting, filtering, browsing, comparison, characterizing trends and distributions, finding anomalies and outliers, finding correlation...
- data types
  - number tables, relational networks, spatial
  - transform into useful configuration: derived data
- threat to validity: you're showing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

Designing visual encoding and interaction tech

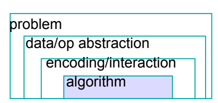


- visual encoding
  - marks: points, lines, areas
  - attributes: position, color, shape, size, orientation, ...
- interaction
  - selecting, navigating, ordering,...
- threat to validity: the way you show it doesn't work



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

Creating algorithms to execute techniques



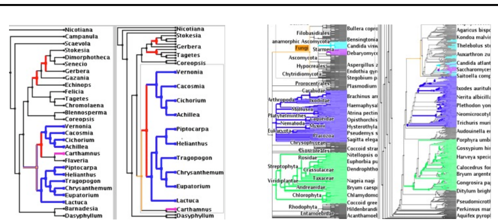
- classic computer science problem
  - create algorithm given clear specification
- threat to validity: your code is too slow

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
  - conflicting tradeoffs
    - iterative refinement often necessary

Principles in action: walk through examples

- vis work in many domains
  - topology
  - computer networking
  - computational linguistics
  - web logs
  - large-scale system administration
  - ...
  - **biology**



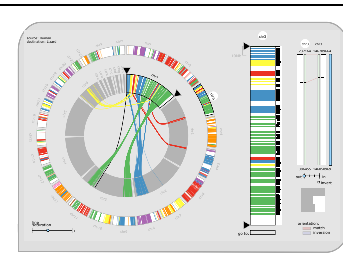
TreeJuxtaposer

Scalable Phylogenetic Tree Comparison

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

http://olduvai.sf.net/tj

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



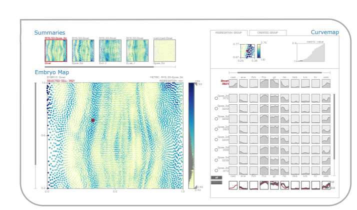
MizBee

A Browser for Comparative Genomics Data

joint work with:
Miriah Meyer, Hanspeter Pfister

http://www.mizbee.org

MizBee: A Multiscale Synteny Browser.
Meyer, Munzner, Pfister, IEEE InfoVis 2009.



MulteeSum

A Tool for Exploring Space-Time Expression Data

joint work with:
Miriah Meyer, Angela DePace, Hanspeter Pfister

http://www.multeesum.org

MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data.
Meyer, Munzner, DePace, Pfister. IEEE InfoVis 2010.

Cerebral

Comparing Multiple Experimental Conditions Within Biologically Meaningful Network Context

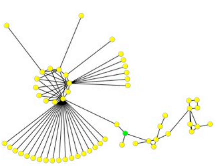
joint work with:
Aaron Barsky, Jennifer Gardy, Robert Kincaid

http://www.pathogenomics.ca/cerebral/

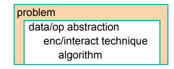
Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context.
Barsky, Munzner, Gardy, Kincaid. IEEE InfoVis 2008.

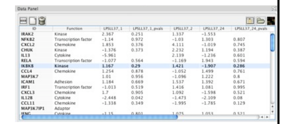
Systems biology model

- graph G = {V, E}
  - –V: proteins, genes, DNA, RNA, tRNA, etc.
    - –metadata: labels, biological attributes
  - –E: interacting molecules
    - –known from previous research



Cycle: model - experiment

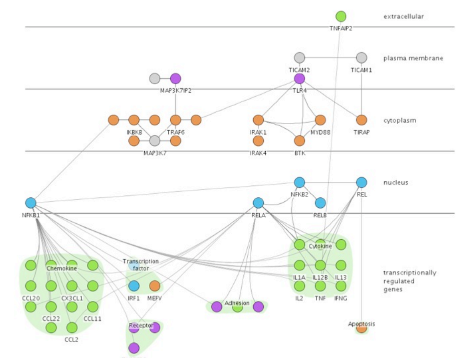




- conduct experiments on cells
  - –microarrays
  - –measurements for each vertex in graph
- interpret results in current graph model
- propose modifications to refine model
- vis tool to accelerate workflow
  - –integrated tool to see graph and measurements together
  - –choose scope for problem complexity

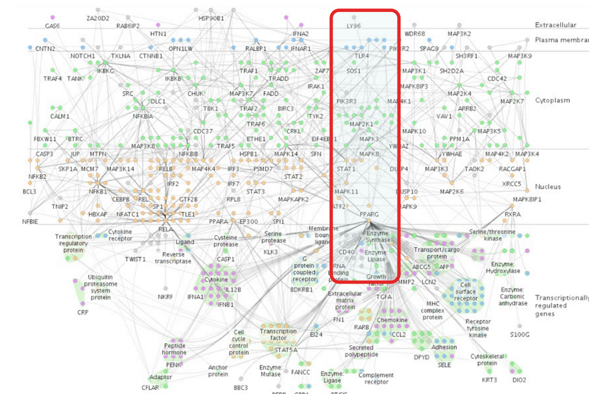
TLR4 biomolecule: E=74,V=54

- very local view

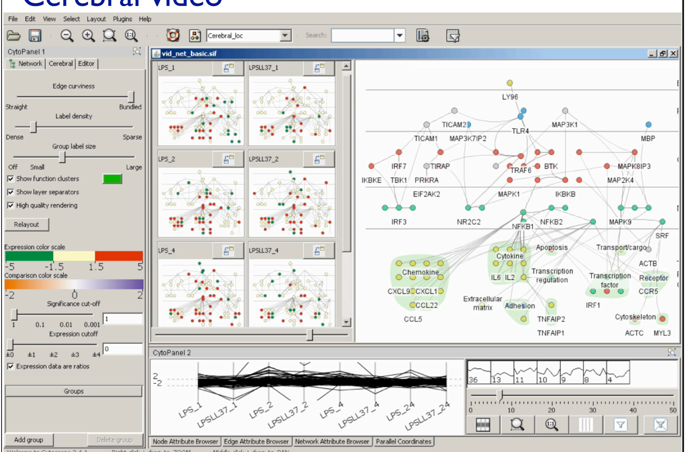


Immune system: E=1263,V=760

- bigger picture, target size for Cerebral



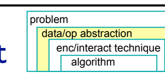
Cerebral video



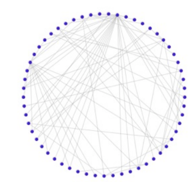
Encoding and interaction design decisions

- 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

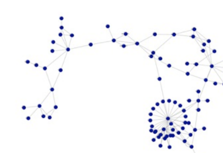
Choice: Create custom graph layout



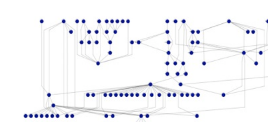
- graph layout heavily studied
  - –given graph G={V,E}, create layout in 2D/3D plane
  - –hundreds of papers
  - –annual Graph Drawing conf.



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
- interaction location often known as part of model

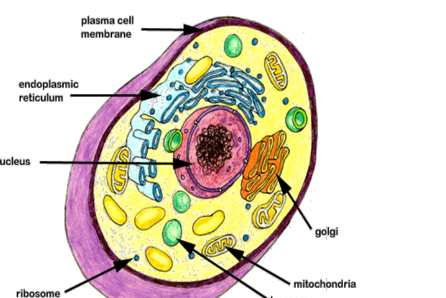
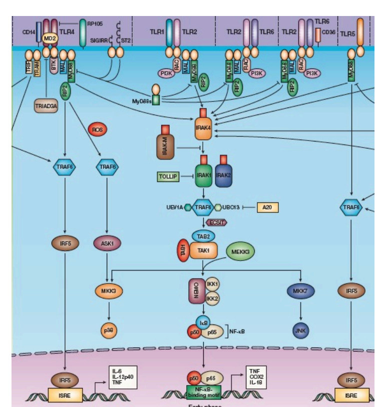


Image credit: Dr.G Weaver, Colorado University at Denver

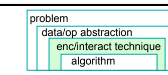
Hand-drawn diagrams

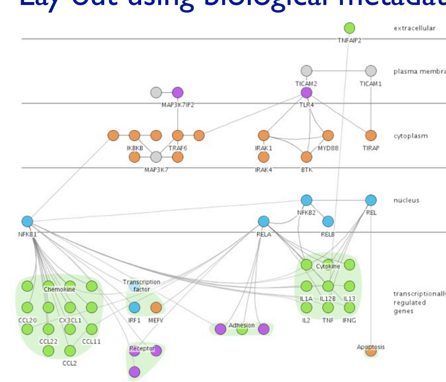


http://www.nature.com/nri/focus/tlr/nri1397.html

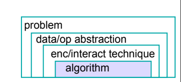
- cellular location spatially encoded vertically
- infeasible to create by hand in era of big data

Lay out using biological metadata



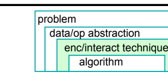


- similar to hand-drawn: spatial position reveals location in cell

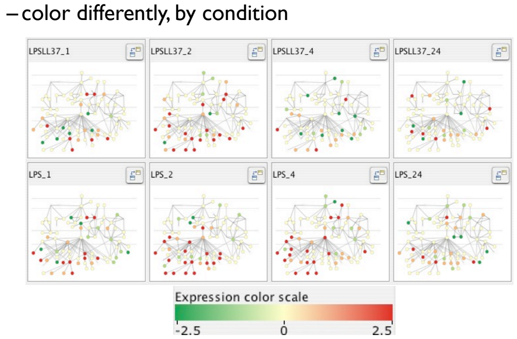


- simulated annealing in $O(E\sqrt{V})$ vs. $O(V^3)$ time

Choice 2: Use small multiple views

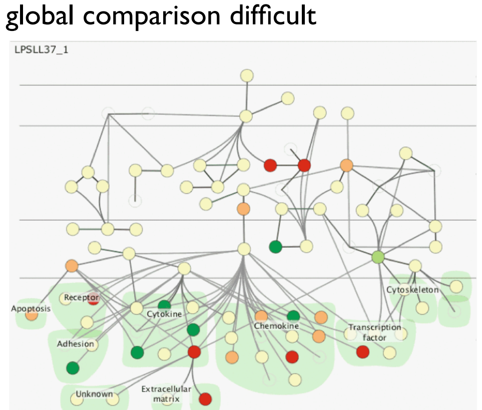


- one graph instance per experimental condition
  - –same spatial layout
  - –color differently, by condition



Why not animation?

- global comparison difficult



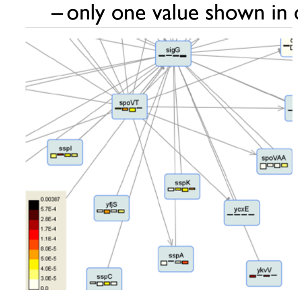
Why not animation?

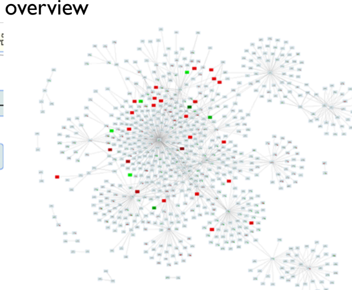
- limits of human visual memory
  - –compared to side by side visual comparison

- Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. Matthew Plumlee and Colin Ware. *ACM Trans. Computer-Human Interaction (ToCHI)*, 13(2):179-209, 2006.
- Animation: can it facilitate? Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. *International Journal of Human-Computer Studies*, 57(4):247-262, 2002.
- Effectiveness of Animation in Trend Visualization. George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, John Stasko. IEEE Trans. Visualization and Computer Graphics 14(6):1325-1332 (Proc. InfoVis 08), 2008.

Why not glyphs?

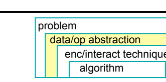
- embed multiple conditions as a chart inside node
- clearly visible when zoomed in
- but cannot see from global view
  - –only one value shown in overview



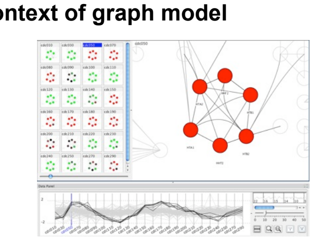


[M. A. Westenberg, S. A. F. T. van Hijum, O. P. Kuipers, J. B. T. M. Roerdink. Visualizing Genome Expression and Regulatory Network Dynamics in Genomic and Metabolic Context. Computer Graphics Forum, 27(3):887-894, 2008.]

Choice: Show measures and graph

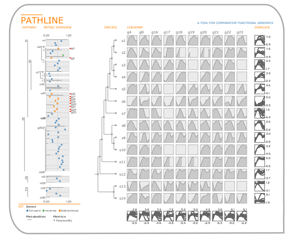


- why not measurements alone?
  - –data driven hypothesis: gene expression clusters indicate similar function in cell?
- clusters are often untrustworthy artifacts!
  - –noisy data: different clustering alg. → different results
  - –measured data alone potentially misleading
  - –**show in context of graph model**



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://innatedb.ca
  - –many more independent adopters
    - 12+ bio lit citations with Cerebral diagrams so far



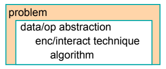
Pathline

A Tool for Comparative Functional Genomics Data

joint work with:
Miriah Meyer, Bang Wong, Mark Styczynski, Hanspeter Pfister

http://www.pathline.org

Pathline: A Tool for Comparative Functional Genomics
Meyer, Wong, Styczynski, Munzner, Pfister, IEEE/Eurographics EuroVis 2010.



problem: **functional genomics**

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

functional genomics data

gene expression
molecular pathways

gene expression is ...

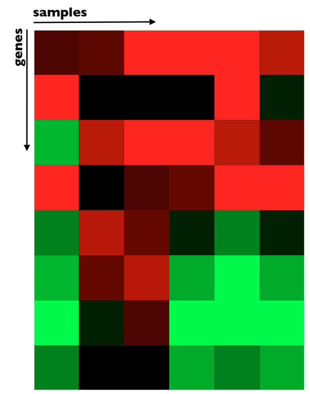
... the measured level of how much a gene is on or off
... a single quantitative value

biologists measure it ...

... for many genes
... in many samples (time points, tissue types, species)

visualized with heatmaps
[Wilkinson09] [Saldanha04] [Seo02] [Eisen98] [Gehlenborg10] [Weinstein08]

encode value with color



gene expression is ...

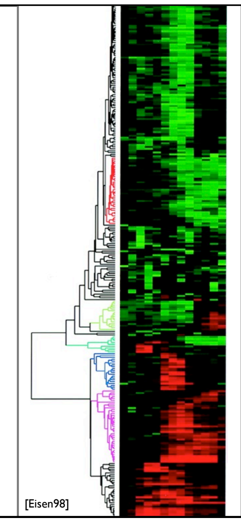
... the measured level of how much a gene is on or off
... a single quantitative value

biologists measure it ...

... for many genes
... in many samples (time points, tissue types, species)

visualized with heatmaps
[Wilkinson09] [Saldanha04] [Seo02] [Eisen98] [Gehlenborg10] [Weinstein08]

encode value with color
augmented with clustering

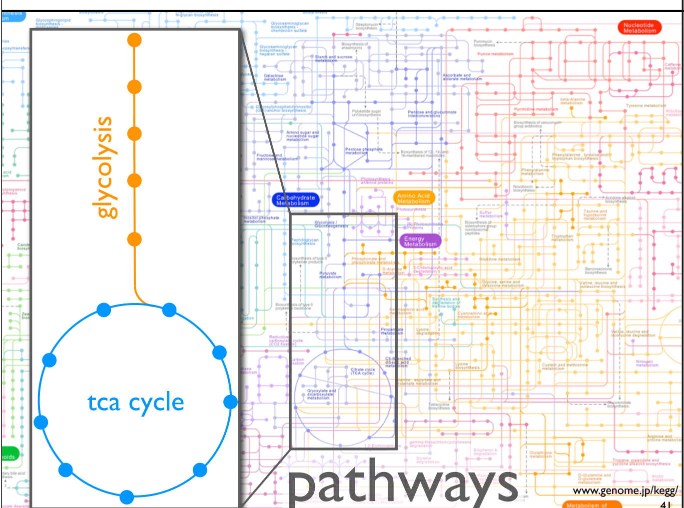


[Eisen98]

functional genomics data

gene expression
molecular pathways

the functioning of a cell is controlled by many interrelated chemical reactions performed by genes



functional genomics:

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

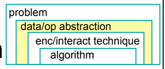
comparative functional genomics:

how do the gene interactions vary across different species?

collaborators: Regev 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*

comparative functional genomics

how do the gene interactions vary across different species?



metabolic pathways

gene expression

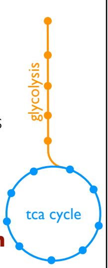
Data

similarity scores

phylogeny

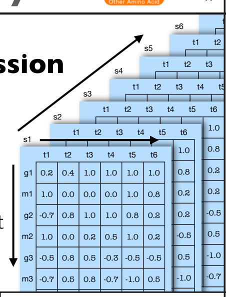
metabolic pathways

- 10 to 50 pathways of interest
- inputs/outputs called metabolites
- **directed graph**



gene expression

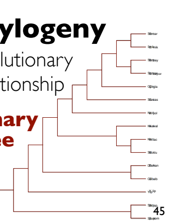
- 6000 genes and 140 metabolites
- 6 time points
- 14 species of yeast
- **3D table**



similarity scores

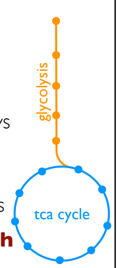
phylogeny

- evolutionary relationship
- **binary tree**



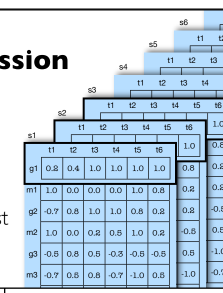
metabolic pathways

- 10 to 50 pathways of interest
- inputs/outputs called metabolites
- **directed graph**

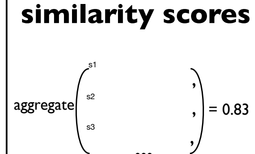


gene expression

- 6000 genes and 140 metabolites
- 6 time points
- 14 species of yeast
- **3D table**



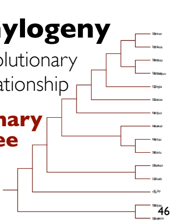
similarity scores

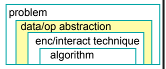


- aggregate time series for a gene/metabolite over species
- similarity of expression across species
- aggregate: Pearson, Spearman, others
- **quantitative value**

phylogeny

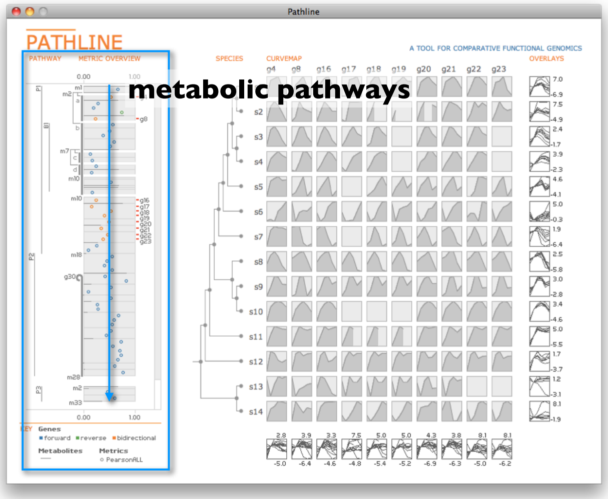
- evolutionary relationship
- **binary tree**

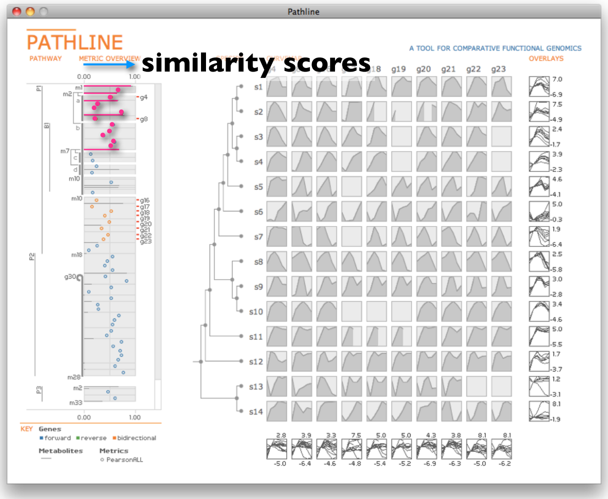


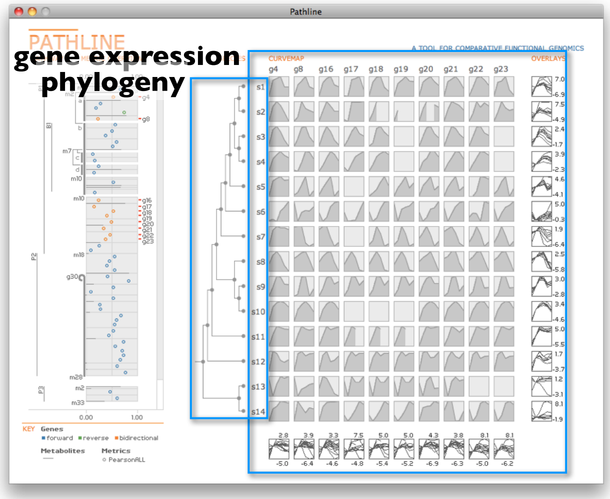


Tasks

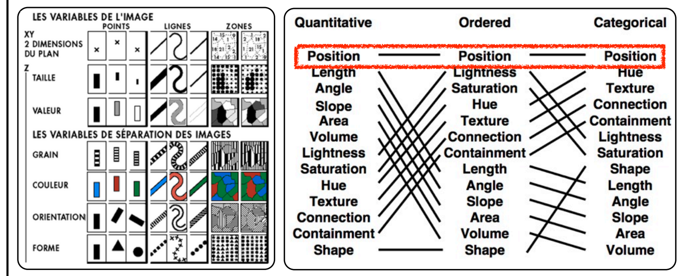
study expression data as a time series
compare a limited number of time series
compare similarity scores along a pathway(s)
comparison of multiple similarity scores







Principle: spatial position is visual channel most accurately perceived for all data types

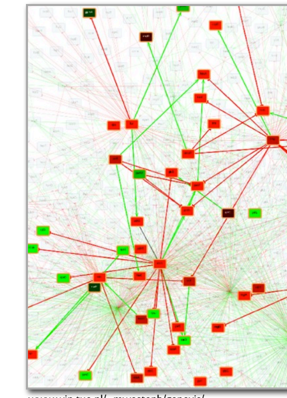


Semiology of Graphics
Bertin, 1967

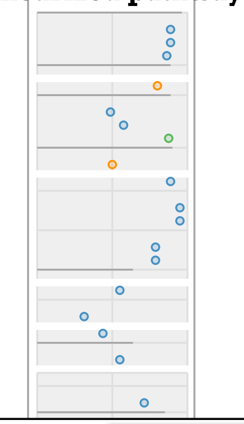
Automating the Design of Graphical Presentations of Relational Information
Mackinlay, 1986

Encode quantitative values with spatial position

topological layout **linearized pathway**

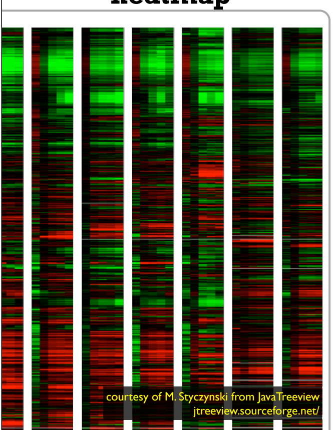


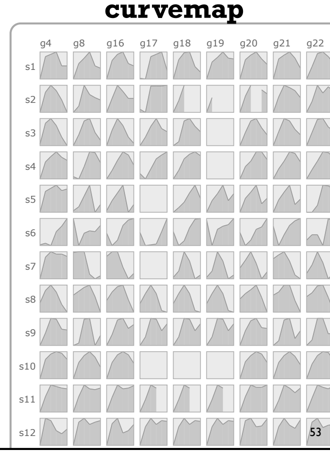
www.win.tue.nl/~mwestenb/genevis/



Encode quantitative values with spatial position

heatmap **curvemap**

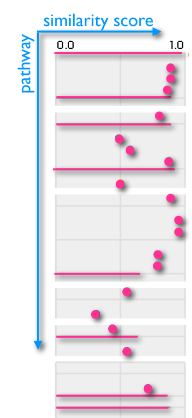




Linearized pathway

common axes to compare similarity scores

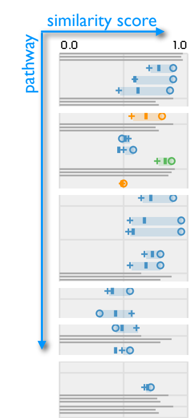
- bars and circles
  - visual layer for selective attention
  - color-code gene direction



Linearized pathway

common axes to compare similarity scores

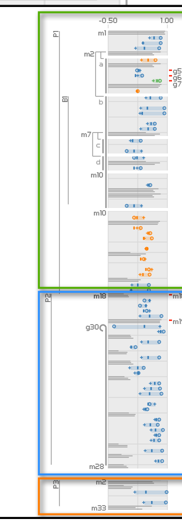
- bars and circles
  - visual layer for selective attention
  - color-code gene direction
- multiple similarity scores



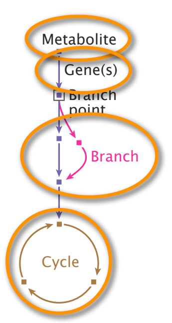
Linearized pathway

common axes to compare similarity scores

- bars and circles
  - visual layer for selective attention
  - color-code gene direction
- multiple similarity scores
- multiple pathways



Pathway to ordered list of nodes

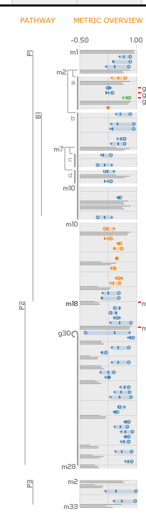


unroll and cut **reinsert** **shared coordinate frame and stylized marks**

Linearized pathway

putting it together . . .

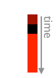
- use spatial position for similarity scores
- topology is secondary



Curvemap

alternative to heatmaps

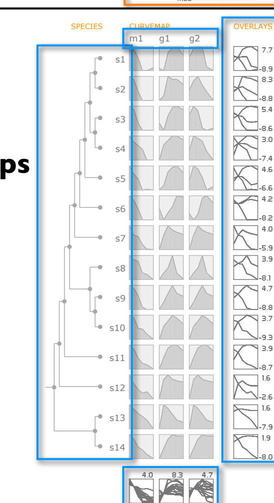
- base visual unit is a curve
- filled, framed line charts to enhance shape perception

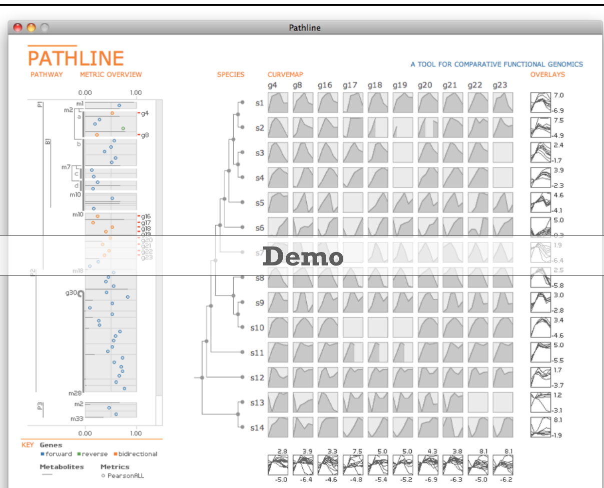


Curvemap

alternative to heatmaps

- base visual unit is a curve
- filled, framed line charts to enhance shape perception
- rows are species
- columns are genes/metabolites
- overlays to enhance trends





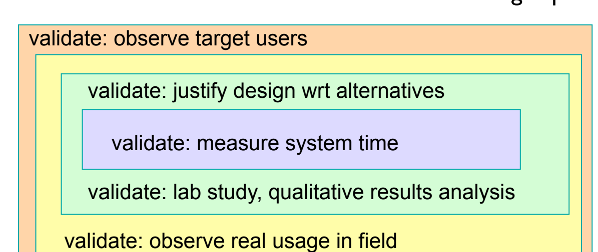
Contributions

- 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

Metabolomics and Integrative Systems Biology Analysis of the Evolution of the Diauxic Shift in Ascomycota fungi. M. Styczynski et al., in preparation.

Principle: use validation methods tuned to level

- is target problem really solved?
  - what have we learned about tradeoffs in design space?



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

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