Scalable Visual Comparison of Biological Trees and Sequences

Tamara Munzner
Imager Lab
University of British Columbia CS

Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration (j P BIRS)
25 May 2004

Outline
Stirring up controversy

Comparing big phylogenetic trees
- Treejuxtaposer
  - phylogeny background
  - structural difference computation
  - guaranteed visibility

Browsing huge trees
- TJC, TJC-Q

Comparing many large gene sequences
- Sequencejuxtaposer

Collaborators
Treejuxtaposer joint work with
- Francois Guimbretiere, Maryland
- Serrdar Tasiran, Compaq SRC
- Li Zhang, Compaq SRC
- Yunhong Zhou, Compaq SRC
- James Slack, UBC

TJC, TJC-Q joint work with
- Dale Beerman, Virginia
- Greg Humphreys, Virginia

Sequencejuxtaposer joint work with
- James Slack, UBC
- Kristian Hildebrand, UBC
- Katherine St. John, CUNY/Lehman

Funding: NSF/DEB-0121682

Stirring up controversy
definitions and scope, infovis vs. scavis:
- spatialization chosen not given
- big parameter space, justify design decisions
- wider scope, mostly more shallowly explored

Stirring up controversy
definitions and scope, infovis vs. scavis:
- spatialization chosen not given
- big parameter space, justify design decisions
- wider scope, mostly more shallowly explored
- many algorithms and techniques span the border
Navigation

intimate relationship with spatial layout choices
- constrained
- nonliteral

Focus + Context
- overview and detail integrated into single view
- show features in context
- help users maintain their orientation

distortion-based navigation
- preserve topological order
- non-linearly compress/expose geometry

Outline

Stirring up controversy

Comparing big phylogenetic trees
- Treejuxtaposer
  - phylogeny background
  - structural difference computation
  - guaranteed visibility

Browsing huge trees
- TJC, TJC-Q

Comparing many large gene sequences
- Sequencejuxtaposer

Tree comparison

active area: hierarchy browsing
- previous work: browsing
- comparison still open problem

bioinformatics application
- phylogenetic trees reconstructed from DNA

Phylogeny background

tree describing evolutionary relationships
- leaves (taxa): species, genes, disease strains

Phylogenetic reconstruction

know leaves, infer interior nodes
- similarity:
  - parallel evolution or common ancestor?
  - siblings unordered

old: morphology
- observable similarities

new: molecular
- DNA sequences – nucleotides
- protein sequences – amino acids

Phylogeny uses

establish relationships
- understand species evolution
- track diseases genes evolve 1M x faster

predict characteristics
- design drugs
- reveal gene function
Phylogenetic/Evolutionary tree

Common tree size now

Tree of Life: 10M species

Phylogenetic reconstruction

multiple trees
- reconstruction algorithm returns many possibilities
- different biological assumptions or data

Phylogenetic reconstruction

Paper comparison

focus
context

visual pairwise comparison
- open problem

Will Fischer, UT-Austin, May 2003
TreeJuxtaposer video
platforms shown
- java 1.4, GL4Java 2.7 bindings for OpenGL

Windows
- 2.4 GHz P3, nVidia Quadro4 700XGL
  - 1.1GB java heap
  - window sizes 1280x1024, 3800x2400

Linux
- 3.1 GHz P4, nVidia GeForce FX 5800 Ultra
  - 1.7GB java heap
  - window size 800x600

Outline
Stirring up controversy
Comparing big phylogenetic trees
  - TreeJuxtaposer
    - phylogeny background
    - structural difference computation
    - guaranteed visibility
Browsing huge trees
  - TJC, TJC-Q
Comparing many large gene sequences
  - SequenceJuxtaposer

Previous work
tree comparison
- RF distance [Robinson and Foulds 81]
- perfect node matching [Day 85]
- creation/deletion [Chi and Card 99]
- leaves only [Graham and Kennedy 01]

Similarity score

\[ S(m,n) = \frac{|L(m) \cap L(n)|}{|L(m) \cup L(n)|} \]

\[ T_1 \]
\[ T_2 \]

\[ L(m) = \{E,F\} \]
\[ L(n) = \{D,E,F\} \]

\[ S(m,n) = \frac{2}{3} \]

Best corresponding node

\[ BCN(m) = n \]

\[ BCN(m) = \text{argmax}_{v \in T_1} (S(m,v)) \]
- computable in \( O(n \log^2 n) \)
- linked highlighting

Marking structural differences

\[ \text{Nodes for which } S(v,BCN(v)) = 1 \]
Structural difference algorithm
powerful and totally automatic
matches intuition
- UT-Austin biology lab
  - other biologists
  - other domains
leads users to important locations
efficient algorithms: 7s for 2 x 140K nodes

Outline
Stirring up controversy
Comparing big phylogenetic trees
  - TreeJuxtaposer
    - phylogeny background
    - structural difference computation
    - guaranteed visibility
Browsing huge trees
  - TIC, TIC-Q
Comparing many large gene sequences
  - SequenceJuxtaposer

Guaranteed mark visibility

Marks (Features)
regions of interest shown with color highlight
  - structural difference
  - search results
  - user-specified
purpose
  - guide navigation
  - provide landmarks
  - contiguity check for subtrees

How can a mark disappear?

moving outside viewport
- choose global Focus+Context navigation
  "tacked-down" borders

Focus+Context previous work
combine overview and detail into single view
Focus+Context
  - large tree browsing
    - Cone Trees [Robertson et al 91]
    - Hyperbolic Trees [Lamping et al 95, Munzner 97]
    - Space Tree [Plaisant et al 03]
    - DOI Tree [Card and Nation 02]
  - global
    - Document Lens [Robertson and Mackinlay 93]
    - Rubber Sheets [Sarker et al 93]
our contribution
  - scalability, guaranteed visibility
How can a mark disappear?

- moving outside viewport
  - choose global Focus+Context navigation
    - "tacked-down" borders
- occlusion
  - choose 2D++ layout
- culling at subpixel sizes
  - develop efficient check for marks when culling
  - cost depending on visible, not total, node count

Mark checking when culling

- does region of space enclose mark on this tree?
  - precompute range beneath subtree
  - correlate objects to spatial extent with quadtree
- does region of space enclose linked mark from other tree?
  - up to O(m) to look up best match for each node
  - solution: intersect node ranges between trees
    - reduces to point in polygon test
    - O(n log n) preprocess, O(log² n) lookup

Intersecting ranges between trees

- point in polygon
  - tuple of indices in N-dim range

Focus+Context quadtrees

- quadtree cells also "painted on rubber sheet"
  - geometry at fixed offset from cell boundary
  - opposite of kinetic data structures
  - must update boundary position when stretch/shrink
- hierarchical position encoding
  - absolute location for boundary
    - lookup: O(1), update: O(n)
  - relative distance between parent cell boundaries
    - lookup: O(log n), update: O(log n)

Guaranteed visibility

- infrastructure needed for efficient computation
- relief from exhaustive exploration
  - missed marks lead to false conclusions
  - hard to determine completion
  - tedious, error-prone
- compelling reason for Focus+Context
  - controversy: does distortion help or hurt?
  - strong rationale for comparison

TreeJuxtaposer contributions

- first interactive tree comparison system
  - automatic structural difference computation
  - guaranteed visibility of landmark areas
- scalable to large datasets
  - 250,000 to 500,000 total nodes
  - all preprocessing subquadratic
  - all realtime rendering sublinear
- techniques broadly applicable
  - not limited to biological trees
- overall winner: InfoVis Contest 2003
Outline

Stirring up controversy

Comparing big phylogenetic trees
- Treeluxtaper
  - phylogeny background
  - structural difference computation
  - guaranteed visibility

Browsing huge trees
- TJC, TJC-Q

Comparing many large gene sequences
- SequenceJuxtaposer

Scaling up

TreeJuxtaposer limits
- memory footprint
- rendering CPU bound, want graphics bound

goal: browse huge trees
- concentrate on browsing

TJC-Q: 5M nodes
- commodity platforms

TJC: 15M nodes
- leading-edge graphics hardware

Memory footprint reduction

TJ Focus + Context quadtrees
- navigating, culling, drawing, picking
- new data structures and algorithms instead

Quadtrees: navigating

navigating with stretch/shrink
- TJ: quadtree
- new: lightweight grid data structure

Quadtrees: culling and drawing

culling subpixel objects
- TJ: quadtree cell size test
- new: leaf overlap test

drawing
- TJ: progressive in order of importance
- new: from root
- new alg fast enough to ignore order

Quadtrees: picking

TJ: picking with spatial subdivision

TJC: multiple render target buffer
- encode object ID into offscreen buffer
- supported in hardware on latest ATI cards

TJC-Q: low-memory quadtrees
Outline
Comparing big phylogenetic trees
  - TreeJuxtaposer
    - phylogeny background
    - structural difference computation
    - guaranteed visibility

Browsing huge trees
  - TJC, TJC-Q

Comparing many large gene sequences
  - SequenceJuxtaposer

Accordion drawing
not just for trees!
general scalable visualization infrastructure
  - "rubber sheet" navigation
  - guaranteed visibility of marked areas
implementation: modular package
  - layer below TreeJuxtaposer

SequenceJuxtaposer
accordion drawing for DNA/RNA

previous work: web-based sequence browsers
  - Ensembl, UCSC Genome Browser, NCBI MapViewer
    - heavily used, huge server-side databases
  - zoom or pan in jumps
  - can’t see context

fluid Focus+Context navigation
guaranteed visibility
  - establish when these features useful
  - proof of concept prototype, eventually merge

SJ in action
shown on publicly available data
  - onion yellows phytoplasma: whole genome
    - 860 Kbp
  - Murphy: 22 genes
    - 44 mammals x 17000 bp each = 748 Kbp
  - Treezilla: single gene
    - 500 plants x 1428 bp each = 714 Kbp

scales to 1.7 Mbp with 1.7GB heap
[videos]

Expanding search results

Changing difference thresholds

25%
inspecting 1 of 22 genes
Changing difference thresholds

50%

Changing difference thresholds

60%

Changing difference thresholds

67%

sequences in phylogenetic order
  - phylogenetic signal visible

Work in progress

trees with weighted edges
protein sequences
linking tree and sequence navigation
accordion drawing for sets
  - data mining: transaction processing
open-source release
  - olduvai.sourceforge.net

Other projects in progress

dimensionality reduction
  - steerable MDS (multidimensional scaling)
  - with Matt Williams

Other projects in progress

perception experiments
  - quantifying cost of Focus+Context fisheye distortions
  - no-cost and low-cost regions for visual search task
  - with Keith Lau, Ron Rensink
More information

www.cs.ubc.ca/~tmm/papers.html
www.cs.ubc.ca/~tmm/talks.html

papers, slides, images, movies

software: olduvai.sourceforge.net