Scalable Visual Comparison of Biological Trees and Sequences

Tamara Munzner

University of British Columbia

Stanford Computer Systems Lab Colloquium (EE380)
5 May 2004

Collaborators

TreeJuxtaposer joint work with
- Francois Guimbretiere, Maryland
- Serdar Tasiran, Compaq SRC
- Li Zhang, Compaq SRC
- Yunhong Zhou, Compaq SRC
- James Slack, UBC

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

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

Outline

Comparing big phylogenetic trees
- TreeJuxtaposer
  - phylogenetic 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?

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

Inferring species relationships

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

aligned DNA → tree reconstruction → many trees → pairwise comparison
Phylogenetic reconstruction

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

  visual filtering large sets of trees
  - TreeSetViewer
    - [Amema and Klingner, InfoVis 2002]

  pairwise comparison
  - open problem

Clades

- comparing contiguous groups
  - clade: ancestor + all descendants
  - is a clade in one tree also a clade in other?
  - is some group a clade?

Paper comparison

- focus

  context

Will Fischer, UT–Austin, May 2003

Biologists' requirements

- reliable detection of structural differences
- rapid identification of interesting spots
- analyses of differences in context
- mostly side by side comparison
- manipulation of increasingly larger trees
- support for multiple platforms

TreeJuxtaposer contributions

- interactive tree comparison system
  - automatic detection of structural differences
  - sub-quadratic preprocessing
  - efficient Focus+Context navigation and layout
    - merge overview and detail in single view
  - guaranteed visibility under extreme distortion
- scalable
  - dataset size: handles 280K–500K nodes
  - display size: handles 3800x2400 display

TreeJuxtaposer video

- platforms shown
  - Java 1.4, GL-Java 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
Comparing big phylogenetic trees
  - Treejuxtaposer
    - phylogeny background
    - structural difference computation
    - guaranteed visibility

Browsing huge trees
  - TJC, TJC-Q

Comparing many large gene sequences
  - Sequencejuxtaposer
Computing structural differences

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

Best corresponding node

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

- 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

- Comparing big phylogenetic trees
  - Treejuxtaposer
    - phylogeny background
    - structural difference computation
    - guaranteed visibility
- Browsing huge trees
  - TJC, TJC-Q
- Comparing many large gene sequences
  - Sequencejuxtaposer

Guaranteed mark visibility

Marks

- regions of interest shown with color highlight
  - structural difference
  - search results
  - user-specified

- purpose
  - guide navigation
  - provide landmarks
  - contiguity check

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

---

**Preserving marks while culling**
show mark at unculled node

![Diagram](image1.png)

**Mark preservation strategies**
compress large subtree to small spatial area

![Diagram](image2.png)

---

**Precompute subtree ranges**

- propagation: cost depends on total nodes
- precomputation: cost depends on visible nodes

![Diagram](image3.png)
Marks and linked highlighting
also check for linked marks from other tree
check if best match for node is marked
  up to O(n) to look up each node in range
intersect node ranges between trees
  reduces to point in polygon test, O(log n)

Storing spatial ranges
in each box, store range of objects inside

Spatial range solution
recursive spatial subdivision
  quadtree
  store range of objects enclosed for each cell
  quick check: spatial range vs. selection range
  extending quadtree to Focus+Context
  quadtree cells also "painted on rubber sheet"
  efficient O(log n) update when stretch/shrink

Focus+Context quadtree
sparse cell instantiation
map from cell boundary to object location
store absolute location of cell boundaries?
  lookup: O(1), update: O(n)
instead, store boundaries hierarchically
  relative "split" between parent cell boundaries

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
constraint to fit everything in viewport
  instead could show indirectly
  ideas: Halo [Baudisch 03]
Guaranteed visibility previous work

visibility of abstract information
- effective view navigation [Furnas 97]
- critical zones [Jul and Furnas 98]

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

Comparing big phylogenetic trees
- TreeJuxtaposer
  - 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

[video]

Memory footprint reduction

TJ quadtrees
- navigating, culling, drawing, picking
new algorithms for drawing/culling
new data structures
- TJC-Q: low-memory quadtrees
- TJC: no quadtrees, picking with hardware

Quadtrees: navigating

navigating with stretch/shrink
- instead, lightweight grid data structure
Quadtrees: culling and drawing

culling subpixel objects
  - leaf overlap test, not gridcell size test

drawing in order of importance
  - new alg fast enough to ignore order

Quadtrees: picking

TJ: picking with spatial subdivision
  - encode object ID into offscreen buffer
  - supported in hardware on latest ATI cards

TJC: multiple render target buffer

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

general scalable visualization infrastructure
  - "rubber sheet" navigation
  - guaranteed visibility of marked areas

modular package
  - layer below TreeJuxtaposer
  - not just for trees

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%

Changing difference thresholds
50%

Changing difference thresholds
60%

Changing difference thresholds
67%
phylogenetic signal visible inspecting 1 of 22 genes

Codon bias shown with visual patterns
Codon bias shown with visual patterns

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
  - stay tuned!

Other projects in progress
- dimensionality reduction
  - steerable MDS (multidimensional scaling)
    - (with Matt Williams)
- 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: beta now, public release very soon