Scalable Visual Comparison of Biological Trees and Sequences

Tamara Munzner

University of British Columbia

Asilomar Microcomputer Workshop, 28 Apr 2004

Outline

TreeJuxtaposer
- comparing big trees

TJC, TJC-Q
- browsing huge trees

SequenceJuxtaposer
- comparing many large gene sequences

Collaborators

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

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

Tree comparison

active area: hierarchy browsing

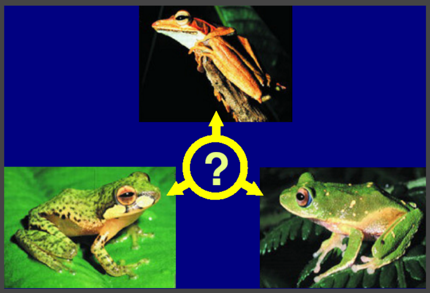
- previous work: browsing

- comparison still open problem

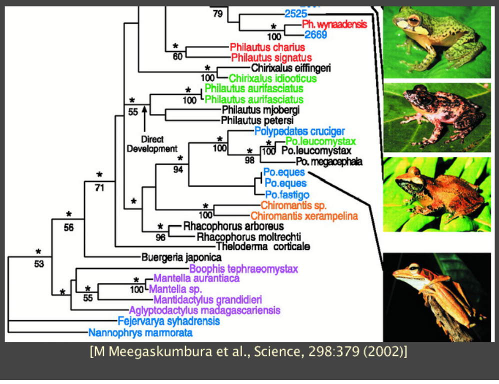
bioinformatics applicationn

- phylogenetic trees reconstructed from DNA

Inferring species relationships

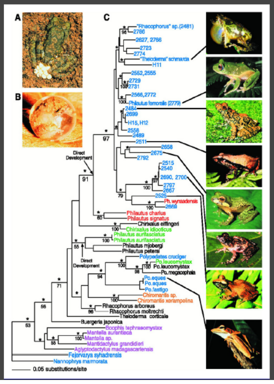


Phylogenetic/Evolutionary tree

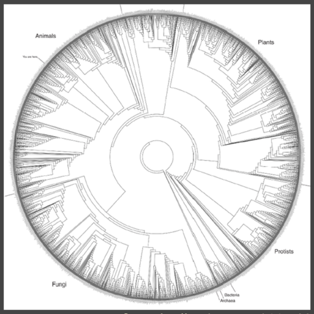


[M Meegaskumbura et al., Science, 298:379 (2002)]

Tree size, common case



Tree of Life: 10M species



[David Hillis, Science, 300:1687, 2003]

TreeJuxtaposer video

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 preprecessing subquadratic
  - all realtime rendering sublinear
- techniques broadly applicable
  - not limited to biological trees

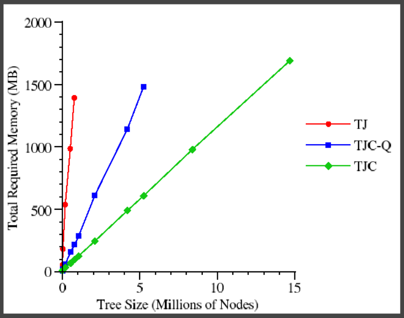
Scaling up

- TreeJuxtaposer limits
  - memory footprint
  - rendering CPU bound, want graphics bound
- goal: browse huge trees
  - browse, not compare
- TJC-Q: 5M nodes
  - commodity platforms
- TJC: 15M nodes
  - leading-edge graphics hardware
- [video]

Quadtree use in TJ

- navigating with stretch/shrink
  - lightweight grid data structure
- culling subpixel objects
  - leaf overlap test, not gridcell size test
- drawing in order of importance
  - new alg fast enough to ignore order
- picking with spatial subdivision
  - TJC: multiple render target buffer
  - TJC-Q: low-memory quadtrees

Memory footprint reduction



SequenceJuxtaposer

accordion drawing for DNA

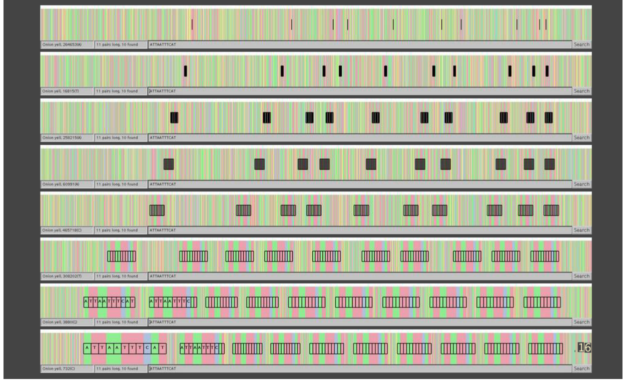
shown on publicly available data

- 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

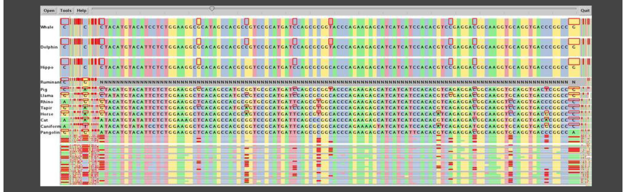
scales to 1.7 Mbp with 1.7GB heap

SequenceJuxtaposer video

Expanding search results

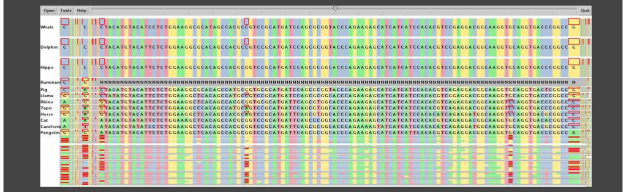


Changing difference thresholds



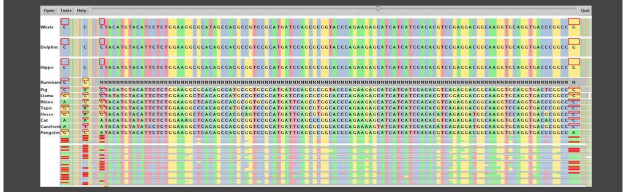
25%

Changing difference thresholds



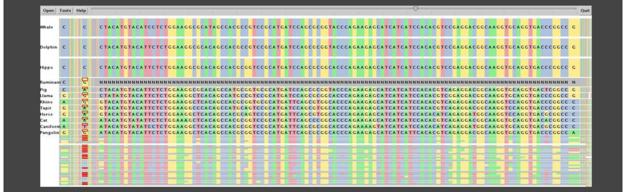
50%

Changing difference thresholds



60%

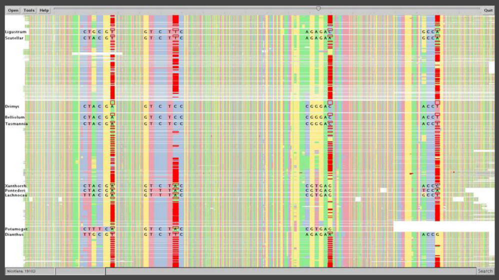
Changing difference thresholds



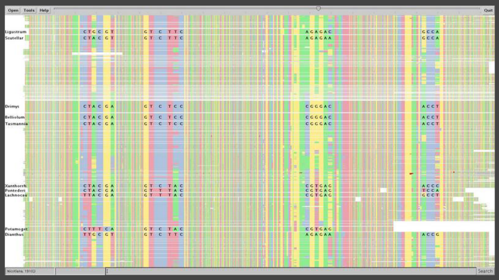
67%

phylogenetic signal visible
inspecting 1 of 22 genes

Codon bias shown with visual patterns



Codon bias shown with visual patterns



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 soon