Scalable Visual Comparison of Biological Trees and Sequences

Tamara Munzner University of British Columbia Department of Computer Science



Outline

Accordion Drawing

- information visualization technique

- TreeJuxtaposer
 - tree comparison
- SequenceJuxtaposer
 - sequence comparison
- PRISAD

– generic accordion drawing framework

Accordion Drawing

- rubber-sheet navigation
 - stretch out part of surface, the rest squishes
 - borders nailed down
 - Focus+Context technique
 - integrated overview, details
 - old idea
 - [Sarkar et al 93],
 [Robertson et al 91]
- guaranteed visibility
 - marks always visible
 - important for scalability
 - new idea
 - [Munzner et al 03]



Guaranteed Visibility

- marks are always visible
- easy with small datasets



hard with larger datasets



• reasons a mark could be invisible

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





- hard with larger datasets
- · reasons a mark could be invisible
 - outside the window
 - AD solution: constrained navigation
 - underneath other marks
 - AD solution: avoid 3D



|--|



- hard with larger datasets
- · reasons a mark could be invisible
 - outside the window
 - AD solution: constrained navigation
 - underneath other marks
 - AD solution: avoid 3D
 - smaller than a pixel
 - AD solution: smart culling









Guaranteed Visibility: Small Items

• Naïve culling may not draw all marked items



Guaranteed visibility of marks



No guaranteed visibility

Guaranteed Visibility: Small Items

• Naïve culling may not draw all marked items



Guaranteed visibility of marks



No guaranteed visibility

Outline

- Accordion Drawing
 - information visualization technique
- TreeJuxtaposer

-tree comparison

- SequenceJuxtaposer
 - sequence comparison
- PRISAD

– generic accordion drawing framework

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: 10M node Tree of Life



<section-header>

David Hillis, Science 300:1687 (2003)

Paper Comparison: Multiple Trees

focus



context



TreeJuxtaposer

- side by side comparison of evolutionary trees
- [video]
 - video/software downloadable from http://olduvai.sf.net/tj



TJ Contributions

- first interactive tree comparison system
 - automatic structural difference computation
 - guaranteed visibility of marked areas
- scalable to large datasets
 - 250,000 to 500,000 total nodes
 - all preprocessing subquadratic
 - all realtime rendering sublinear
- scalable to large displays (4000 x 2000)
- introduced
 - guaranteed visibility, accordion drawing

Structural Comparison



Matching Leaf Nodes



Matching Leaf Nodes



Matching Leaf Nodes











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]





- BCN(m) = argmax_{$v \in T_2$} (S(m, v))
 - computable in O(n log² n)
 - linked highlighting

Marking Structural Differences



- Nodes for which $S(v, BCN(v)) \neq 1$
 - Matches intuition

Outline

- Accordion Drawing
 - information visualization technique
- TreeJuxtaposer
 - tree comparison
- SequenceJuxtaposer
 - sequence comparison
- PRISAD
 - generic accordion drawing framework

Genomic Sequences

- multiple aligned sequences of DNA
- now commonly browsed with web apps
 - zoom and pan with abrupt jumps
 - previous work
 - Ensembl [Hubbard 02], UCSC Genome Browser [Kent 02], NCBI [Wheeler 02]
- investigate benefits of accordion drawing
 - showing focus areas in context
 - smooth transitions between states
 - guaranteed visibility for globally visible landmarks

SequenceJuxtaposer

- comparing multiple aligned gene sequences
- provides searching, difference calculation
- [video]
 - video/software downloadable from http://olduvai.sf.net/tj

Open Tools		Ann	otati	on	He	ip				_									1					Qui
								_						Ì										
Sloth	Т	A	T٦	ΓA	G	т	с	С	т						т	т	A	с	A	с	A	т		
Armadillo	т	G	T٦	ΓA	А	т	т	С	A	т					т	т	A	с	A	с	A	т		
Anteater	т	A	ТI	r G	A	т	т	A	с	т		г		т	т	т	A	с	A	с	A	т		
Hedgehog	Т	A	T٦	гΤ	т	с	т	A	т	т		Г			т	т	А	с	А	с	A	т		
Mole	т	A	ΤI	ΓA	G	с	т	G	т			г			т	т	А	с	А	с	A	т		
Shrew	Т	A	ТI	ΓA	G	т	т	G	т	т					т	т	A	С	A	с	A	т		
Tenrecid	т	A	ΤI	ΓA	G	т	т	С	т	т	-	г			т	т	А	т	А	с	A	т		
								Ξ																
			-		~	~	-	_	-	-	-	-	-		-	-		~	<u> </u>	~		-		
Human		A		A	G	U	1	L	-	-						÷	A	C	A	C	A			
Whale	т	A	ТI	ΓA	G	т	т	с	т	т					т	т	A	с	A	с	A	т		
Dolphin	Т	A	T٦	ΓA	G	т	т	C	т	т					т	т	А	с	А	с	A	т		
Hippo	Т	G	T٦	ΓA	A	т	т	T	т	т	-	Г			т	т	А	с	А	с	А	т		
Llama	Т	A	T٦	٢A	G	Т	т	F	с	Т		Г		т	т	Т	А	С	А	С	А	Т		
Ruminant	Т	G	ТI	ΓA	A	С	т	С	т	т	-	Г			т	т	А	С	А	С	А	т		
Pig	Т	A	ΤI	ΓA	A	т	Т	С	т	т		Г			т	т	А	С	А	С	А	т		
								Ξ																
Human 1/D	Ξ.		-				_	-																Searc
rianian, r(r)																							_	ocart

Searching

- search for motifs
 - protein/codon search
 - regular expressions supported
- results marked with guaranteed visibility



Differences

- explore differences between aligned pairs
 slider controls difference threshold in realtime
- results marked with guaranteed visibility



SJ Contributions

- fluid tree comparison system
 - showing multiple focus areas in context
 - guaranteed visibility of marked areas
 - thresholded differences, search results
- scalable to large datasets
 - 2M nucleotides
 - all realtime rendering sublinear

Outline

- Accordion Drawing
 - information visualization technique
- TreeJuxtaposer
 - tree comparison
- SequenceJuxtaposer
 - sequence comparison
- PRISAD

– generic accordion drawing framework
Goals of PRISAD

- generic AD infrastructure
 - tree and sequence applications
 - PRITree is TreeJuxtaposer using PRISAD
 - PRISeq is SequenceJuxtaposer using PRISAD
- efficiency
 - faster rendering: minimize overdrawing
 - smaller memory footprint
- correctness

- rendering with no gaps: eliminate overculling

PRISAD Navigation

- generic navigation infrastructure
 - application independent
 - uses deformable grid
 - split lines
 - Grid lines define object boundaries
 - horizontal and vertical separate
 - Independently movable



Split line hierarchy

- data structure supports navigation, picking, drawing
- two interpretations
 - linear ordering

– hierarchical subdivision

PRISAD Architecture

world-space discretization

- preprocessing
 - initializing data structures
 - placing geometry



screen-space rendering

- frame updating
 - analyzing navigation state
 - drawing geometry



World-space Discretization

interplay between infrastructure and application



Laying Out & Initializing

- application-specific layout of dataset
 - non-overlapping objects
- initialize PRISAD split line hierarchies
 - objects aligned by split lines



Α	Α	С	С
Α	т	т	т

Laying Out & Initializing

application-specific layout of dataset

non-overlapping objects

initialize PRISAD split line hierarchies
 – objects aligned by split lines





Gridding

 each geometric object assigned its four encompassing split line boundaries



Mapping

- PRITree mapping initializes leaf references
 - bidirectional O(1) reference between leaves and split lines





Screen-space Rendering

control flow to draw each frame



Partitioning

- partition object set into bite-sized ranges
 - using current split line screen-space positions
 - required for every frame
 - subdivision stops if region smaller than 1 pixel
 - or if range contains only 1 object



Seeding

- reordering range queue result from partition

 marked regions get priority in queue
 - drawn first to provide landmarks



Drawing Single Range

- each enqueued object range drawn according to application geometry
 - selection for trees
 - aggregation for sequences

PRITree Range Drawing

- select suitable leaf in each range
- draw path from leaf to the root
 - -ascent-based tree drawing
 - -efficiency: minimize overdrawing
 - only draw one path per range



Rendering Dense Regions

- correctness: eliminate overculling
 - bad leaf choices would result in misleading gaps
- efficiency: maximize partition size to reduce rendering
 - too much reduction would result in gaps



Intended rendering



Partition size too big 51

Rendering Dense Regions

- correctness: eliminate overculling
 - bad leaf choices would result in misleading gaps
- efficiency: maximize partition size to reduce rendering
 - too much reduction would result in gaps





Intended rendering

Partition size too big 52

PRITree Skeleton

 guaranteed visibility of marked subtrees during progressive rendering

first frame: one path per marked group



full scene: entire marked subtrees



53

PRISeq Range Drawing: Aggregation

- aggregate range to select box color for each sequence
 - random select to break ties





PRISeq Range Drawing

- collect identical nucleotides in column
 - form single box to represent identical objects
 - attach to split line hierarchy cache
 - lazy evaluation
- draw vertical column

{ A:[1,1], T:[2,3] }





PRISAD Performance

- PRITree vs. TreeJuxtaposer (TJ)
- synthetic and real datasets
 - complete binary trees
 - lowest branching factor
 - regular structure
 - star trees
 - highest possible branching factor



InfoVis Contest Benchmarks

- two 190K node trees
- directly compare TJ and PT



OpenDirectory benchmarks

- two 480K node trees
- too large for TJ, PT results only

Genres Romance A Arts Composers G C Sw Photography Europe	M	Genres Romance Free	
Arts Composers G C Sw Photography Europe	Inchirational		Inspirational
Photography Europe	The les	Photography Groups A	_
	Ely,_Joe	Visual Arts	Europe
Markating and Advantiging Cuides	History	Asia	Regional
Marketing_and_Advertising Guides	Hardware	-consultants	Hardware
Developers and Publishers I B X	Infinity Ward	Hardware Research	CPLD
Cids_and_Teens Medical Chefs	Men s	Internet Weblogs	Directories
Reference Occasion Obio	By Breed	EPOC_Devices	Hardwara
South Koroo	Society and Cultu	Computers lawa laby/IDW	ITATUWATE
Souch_Korea Parks	Organizations	Ruid Management	Hardware_Driver
Europe London Poetry	Magazines and F-	<u> </u>	Training
Middle Fast Josdan - M	Business and Eco	Hardware_Support	
Middle_East Jordan D Movies	Recreation and St	Speech Technology	Keyboard
Blount Blount	Recreation and St	Board Games People	Hardware
Ca <u>itornia</u> M C Health	Chiropractic	Casinos Dice	Elemental
<u>Counties</u> Lodging	Mans and Views	Live_Action Classic	8
Georgia Atlanta Bacon	Maps_and_views	Developers and Publishers I A 8	Shadowrun
Regional Command	Government		Titles
lotth America	Education	Reference Chron_X	Came Neverendi
Nabraska M Music F	Services	Regional North Lemon	Car Rentals
Now York N Parks		Biology Description	our_neuro
A B F Polk		Math Companies	Hardware_and_Ir
N L Health	Travel_and_louris	Aviation Carope	T 605 Joint Con
Tappassaa 2004 Moody	Floris	Organizations Education	Literacy
Varmont H J C	Liberalism	Mensa Events	Europa
D B Ployer	Health	People 4–H	Caribbaan
Australia Ryde_City	By_County	Philosophers P C	Cioran Emila
Biology Chile Companies	Science_and_Envire	Fillosophers	Ciuran,_Linne
CraftsFresh_Cut	Hardware_and_lr	Society Philosophy_of_Mind	Valini,_Lucino
Society Aviation	Switchplates		Mentalese
Pentecostalism B Aircoft	I-60S_Joint_Con	Anarchism	Anarcha-Feminis
Sport's American Europe	Events	Delities	Resources
Dansk Erhverv Rassen	Grand_Kapids_Ran	Pointics People	Delaware
World Europa Afrika Politik	A	Libertarianism	Polling
Musica Dalmazia	Moosleerau	Socialism States	Theorists
Neuerianus Mazowsze Nysa I	Pubblica_Amminis	Christianity Basel-Stadt F	Christian_Critique

TreeJuxtaposer renders **all** nodes for star trees

• branching factor k leads to O(k) performance



TreeJuxtaposer renders **all** nodes for star trees

• branching factor k leads to O(k) performance



InfoVis 2003 Contest dataset

• 5x rendering speedup



a closer look at the fastest rendering times





Detailed Rendering Time Performance

PRITree handles 4 million nodes in under 0.4 secondsTreeJuxtaposer takes twice as long to render 1 million nodes



Detailed Rendering Time Performance

TreeJuxtaposer valley from overculling



Memory Performance

linear memory usage for both applications4-5x more efficient for synthetic datasets



Memory Performance

1GB difference for InfoVis contest comparison

• marked range storage changes improve scalability



Performance Comparison

- PRITree vs. TreeJuxtaposer
 - detailed benchmarks against identical TJ functionality
 - 5x faster, 8x smaller footprint
 - handles over 4M node trees
- PRISeq vs. SequenceJuxtaposer
 - 15x faster rendering, 20x smaller memory size
 - 44 species * 17K nucleotides = 770K items
 - 6400 species * 6400 nucleotides = 40M items

Future Work

- future work
 - editing and annotating datasets
 - PRISAD support for application specific actions
 - logging, replay, undo, other user actions
 - develop process or template for building applications

PRISAD Contributions

- infrastructure for efficient, correct, and generic accordion drawing
- efficient and correct rendering
 - screen-space partitioning tightly bounds overdrawing and eliminates overculling
- first generic AD infrastructure
 - PRITree renders 5x faster than TJ
 - PRISeq renders 20x larger datasets than SJ

Joint Work

- TreeJuxtaposer
 - François Guimbretière, Serdar Ta_iran, Li Zhang, Yunhong Zhou
 - SIGGRAPH 2003
- SequenceJuxtaposer
 - James Slack, Kristian Hildebrand, Katherine St.John
 - German Conference on Bioinformatics 2004
- TJC/TJC-Q
 - Dale Beermann, Greg Humphreys
 - EuroVis 2005
- PRISAD
 - James Slack, Kristian Hildebrand
 - IEEE InfoVis Symposium 2005
 - Information Visualization journal, to appear

Open Source

- software freely available from <u>http://olduvai.sourceforge.net</u>
 - SequenceJuxtaposer
 olduvai.sf.net/sj
 - TreeJuxtaposer
 olduvai.sf.net/tj
 - requires Java and OpenGL
 - JOGL bindings for TJ, GL4Java for SJ (JOGL coming soon)
- papers, talks, videos also from http://www.cs.ubc.ca/~tmm
PRITree Rendering Time Performance

a closer look at the fastest rendering times



PRITree Rendering Time Performance

a closer look at the fastest rendering times

