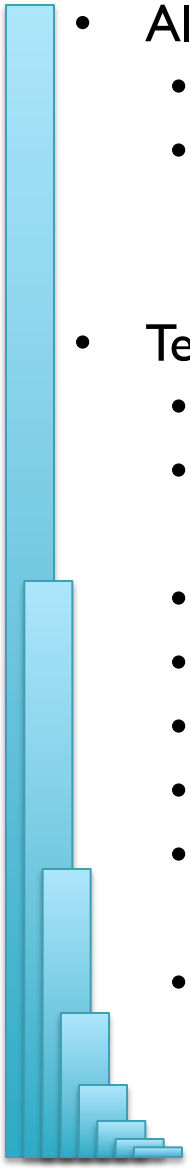
Graphs and Genomes

Michael Schatz

Bioinformatics Lecture 3
Undergraduate Research Program 2011

Recap



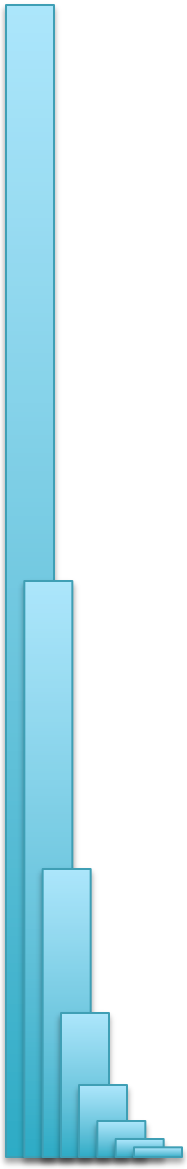
- Algorithms choreograph the dance of data inside the machine
  - Algorithms add provable precision to your method
  - A smarter algorithm can solve the same problem with much less work

- Techniques
  - Analysis: Characterize performance, correctness
  - Modeling: Characterize what you expect to see

  - Binary search: Fast lookup in any sorted list
  - Divide-and-conquer: Split a hard problem into an easier problem
  - Recursion: Solve a problem using a function of itself
  - Indexing: Focus on just the important parts
  - Seed-and-extend: Anchor the problem using a portion of it

  - Brute Force, Suffix Arrays, Binary Search, Quicksort, Bowtie

Challenge Question



Using Bowtie (bowtie -v 0 –a --norc) or your own implementation of the brute force algorithm, scan the E. coli K12/MG1655 genome for GATTACA:

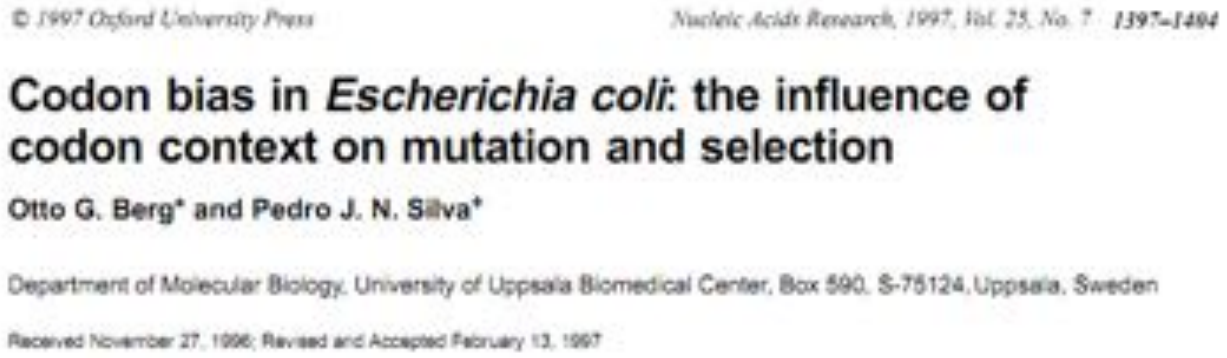
http://schatzlab.cshl.edu/teaching/2011/Ecoli.fa
http://schatzlab.cshl.edu/teaching/2011/GATTACA.fq

Compute the number of occurrences for each of the following queries, and the degree to which the empirical number of matches is consistent with the theoretical e-value. Point out any particularly significant deviations from the theoretical model.

Gattaca:	GATTACA
Gattaca^2:	GATTACAGATTACA
Gattaca^3:	GATTACAGATTACAGATTACA
Start Codon:	ATG
Stop Codons:	TAG, TAA, TGA

Challenge Response

Sequence	Observed	Expected	Difference
GATTACA	230	283	-19%
$GATTACA^2$	0	0.01	-
$GATTACA^3$	0	0	-
Start: ATG	76238	72494	+5%
Stop: TAG	27243	72494	-62%
Stop: TAA	68838	72494	-5%
Stop: TGA	83491	72494	+14%



Nucleic Acids Research, 1997, Vol. 25, No. 7 1397–1404

Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection

Otto G. Berg* and Pedro J. N. Silva*

Department of Molecular Biology, University of Uppsala Biomedical Center, Box 590, S-75124, Uppsala, Sweden

Received November 27, 1996; Revised and Accepted February 13, 1997

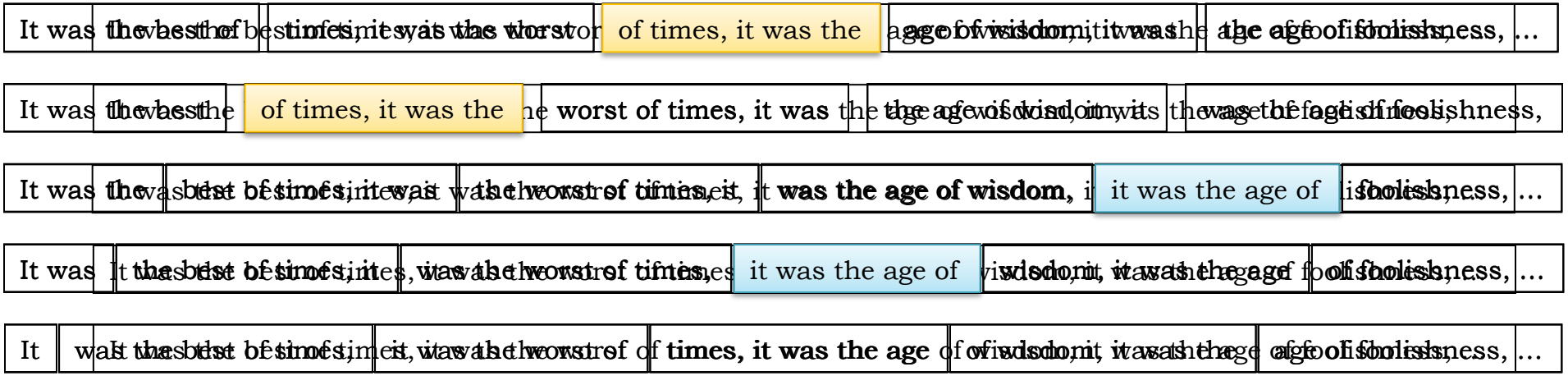
Outline

1. Part 1: Graphs
    1. Genome Assembly by Analogy
    2. Graph Searching

2. Part 2: Schatz Lab
    1. A little about me
    2. Projects

Shredded Book Reconstruction

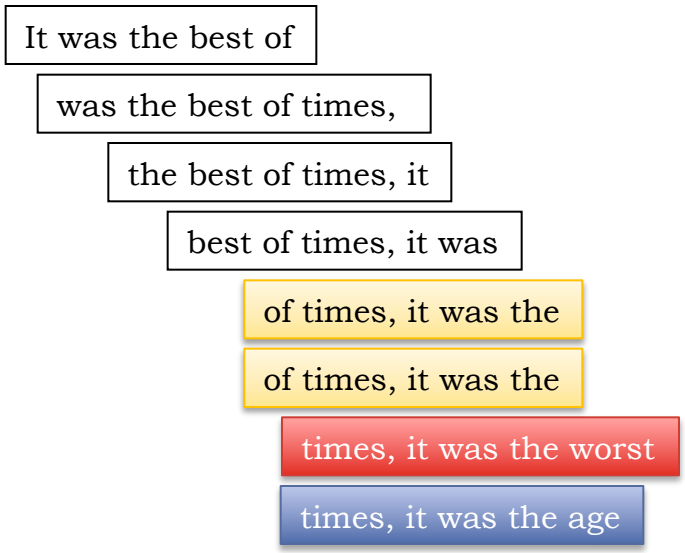
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was



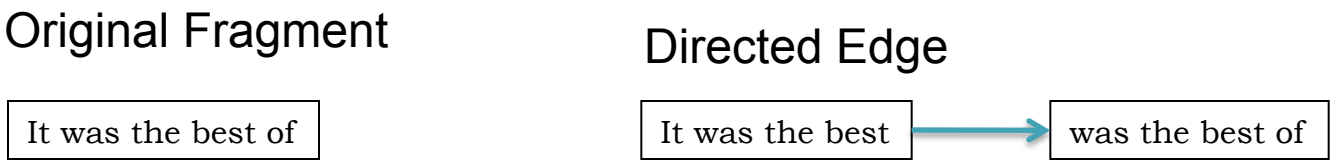
The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

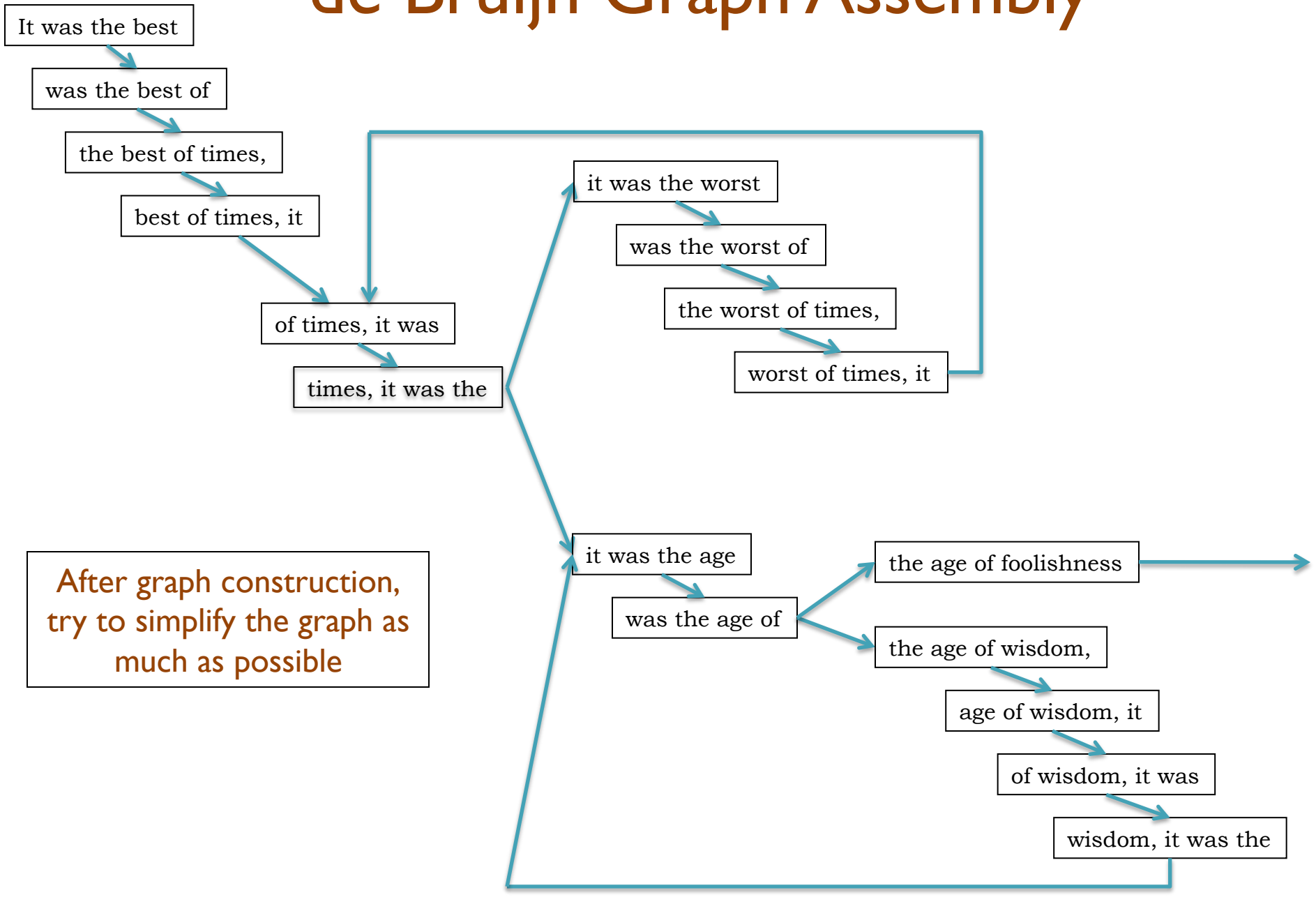
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

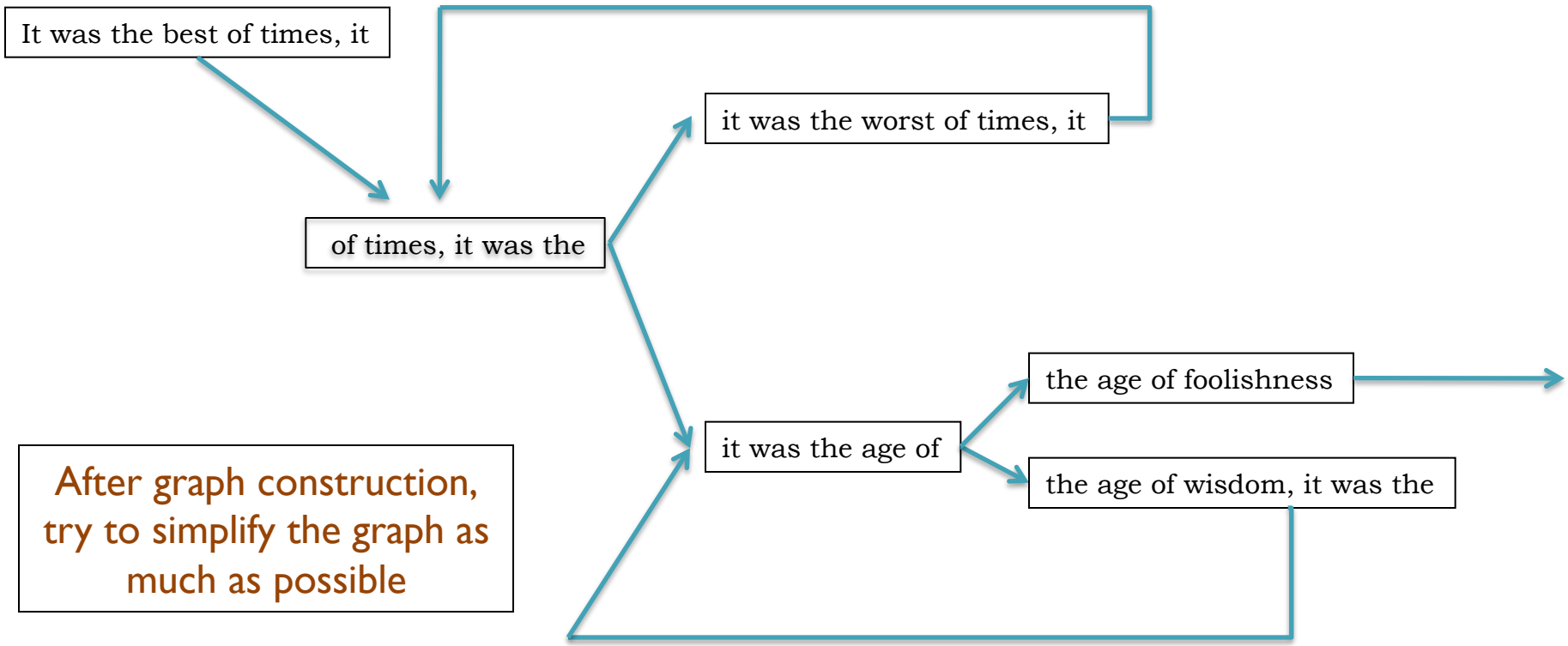
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



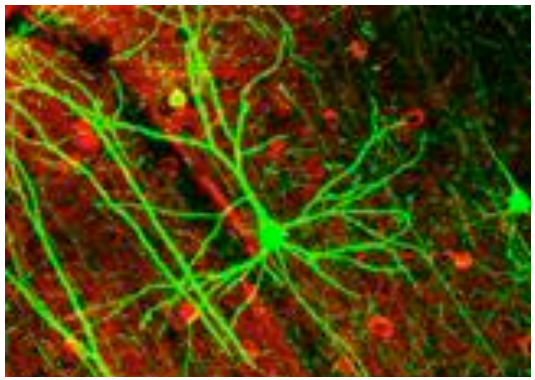
After graph construction, try to simplify the graph as much as possible

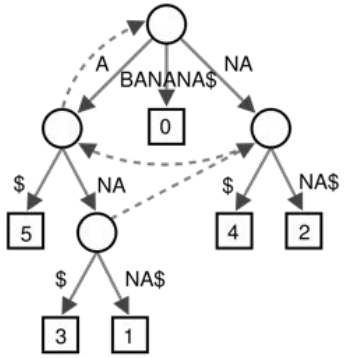
de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

Biological Networks





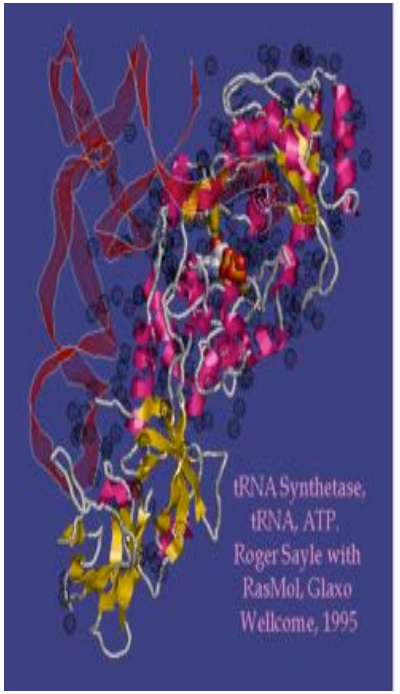
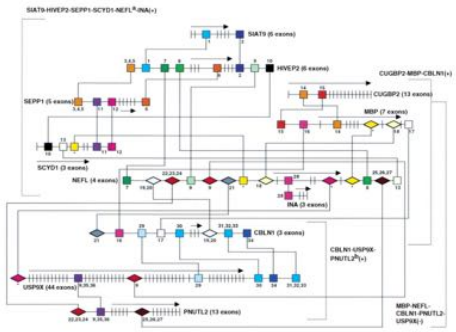
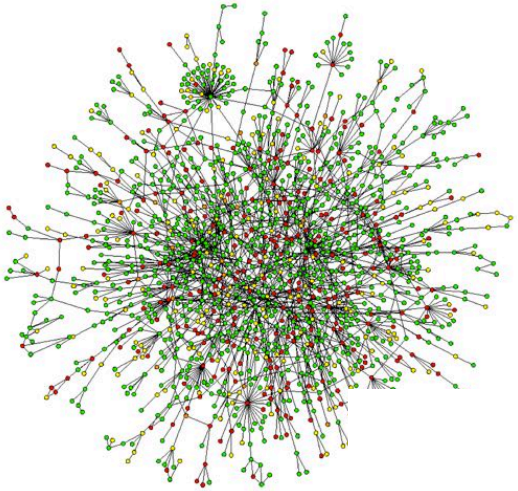


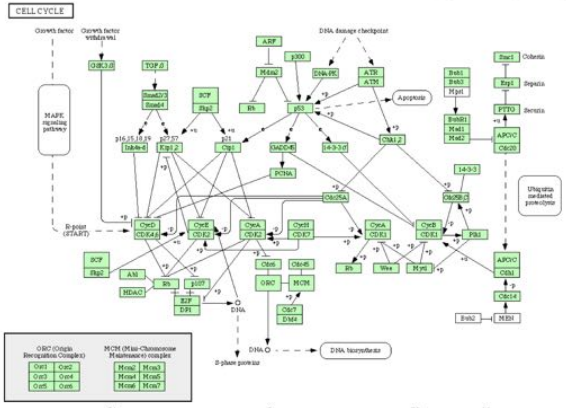
Figure 5 Putative regulatory elements shared between groups of correlated and anticorrelated genes

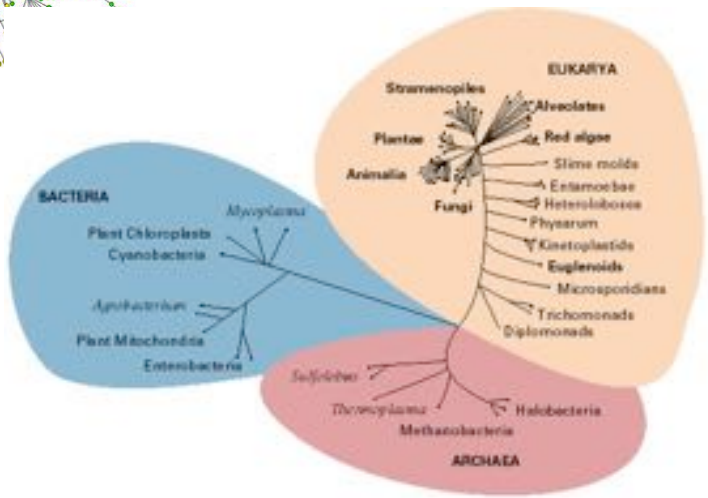


Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

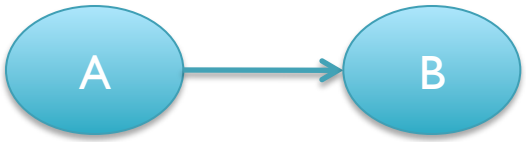
Cold Spring Harbor Laboratory Press





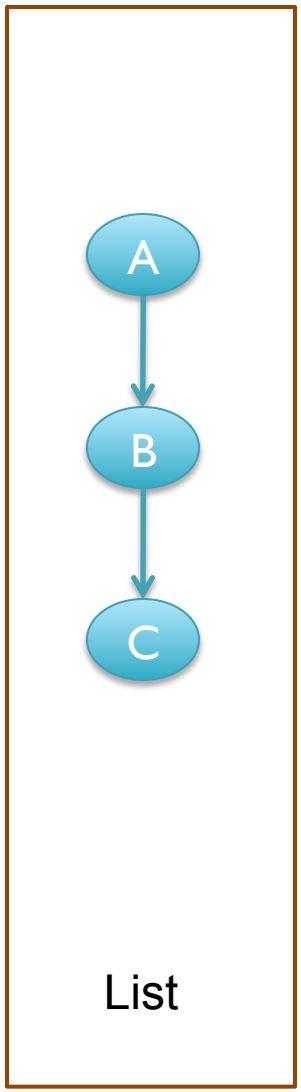


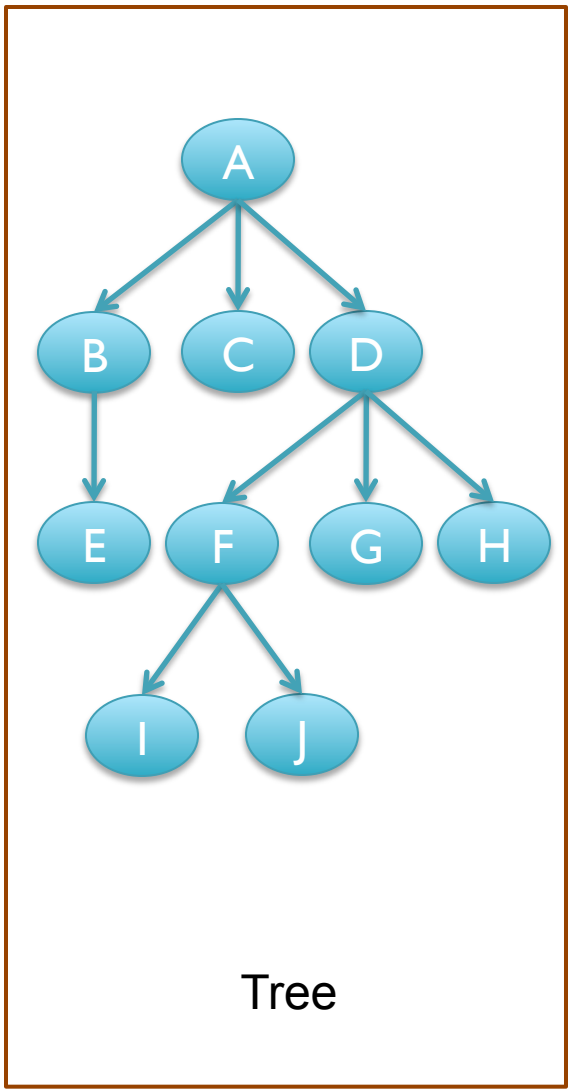
Graphs

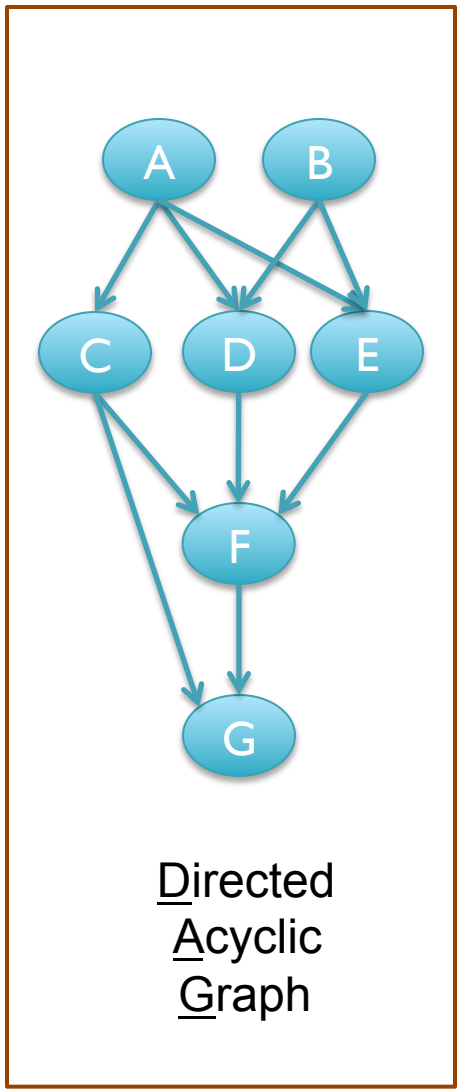


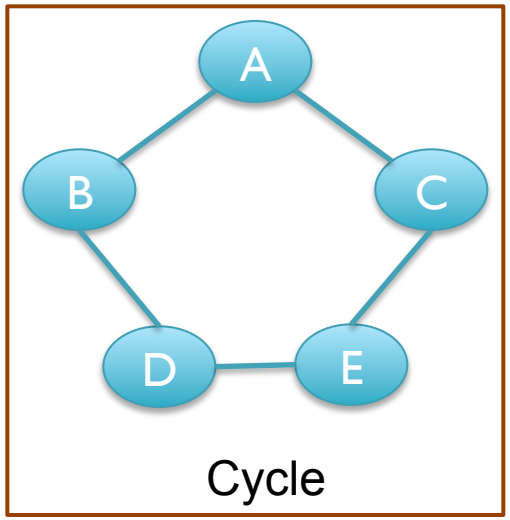
- Nodes
  - People, Proteins, Genes, Neurons, Sequences, Numbers, …
- Edges
  - A is connected to B
  - A is related to B
  - A regulates B
  - A precedes B
  - A interacts with B
  - A activates B
  - …

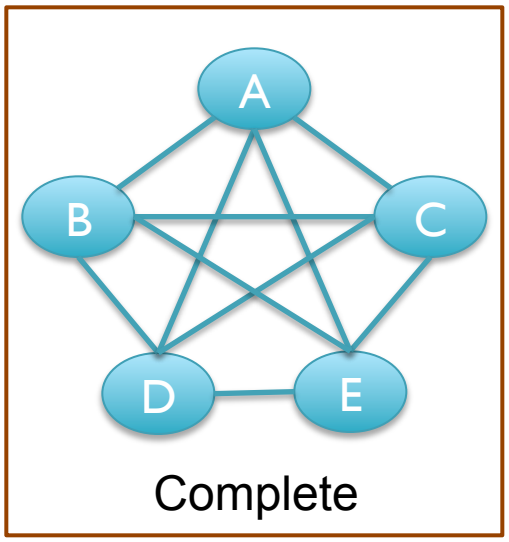
Graph Types


A
B
C
List


A
B
C
D
E
F
G
H
I
J
Tree


A
B
C
D
E
F
G
Directed
Acyclic
Graph


A
B
C
D
E
Cycle

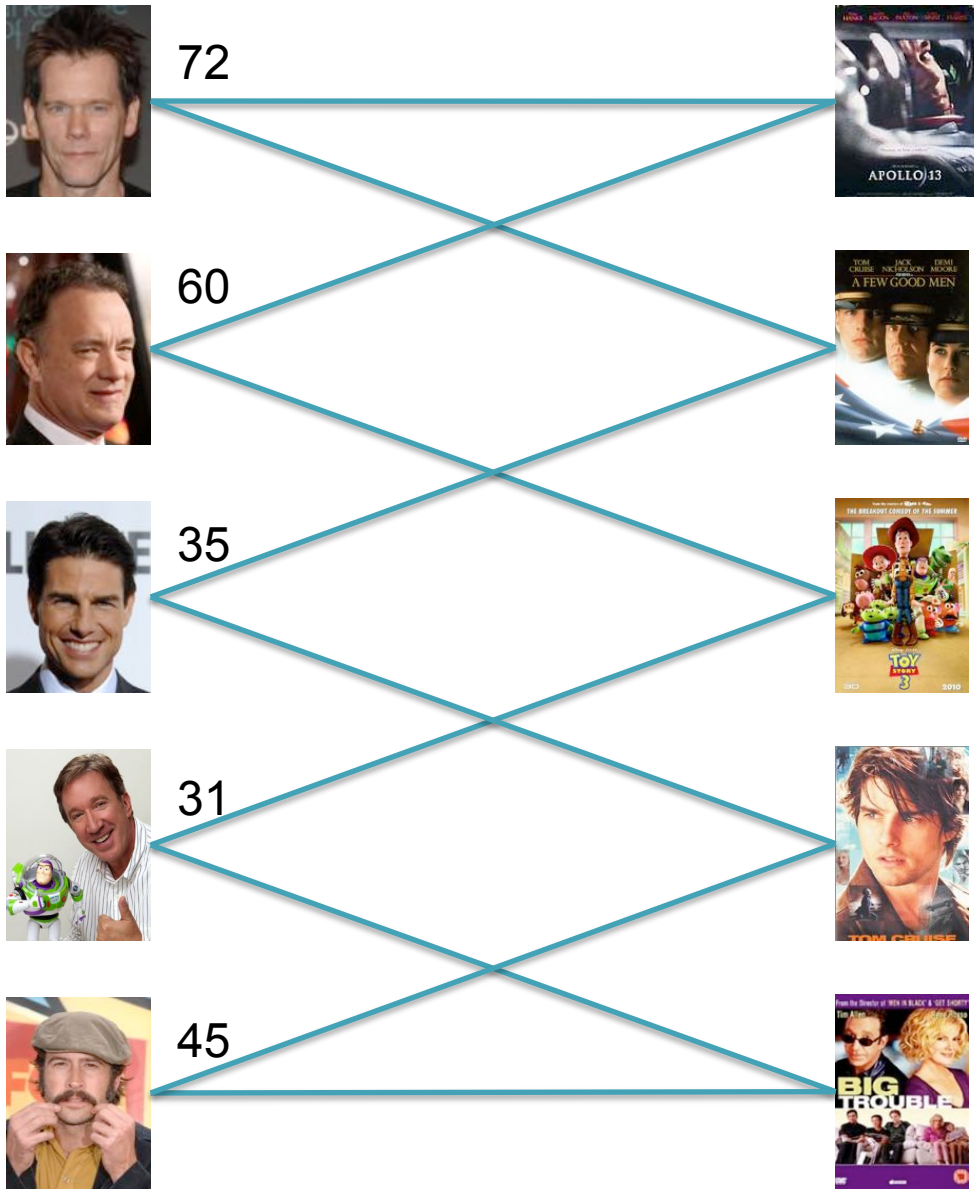

A
B
C
D
E
Complete

Kevin Bacon and Bipartite Graphs

Find the ***shortest*** path from Kevin Bacon to Jason Lee

Breadth First Search: 4 hops

Bacon Distance: 2



BFS and TSP

- BFS computes the shortest path between a pair of nodes in $O(|E|) = O(|N|^2)$

- What if we wanted to compute the shortest path visiting every node once?
  - Traveling Salesman Problem

ABDCA: 4+2+5+3 = 14

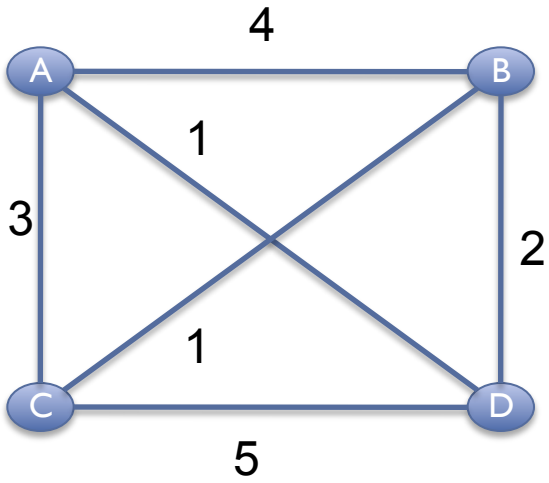
ACDBA: 3+5+2+4 = 14*

ABCDA: 4+1+5+1 = 11

ADCBA: 1+5+1+4 = 11*

ACBDA: 3+1+2+1 = 7

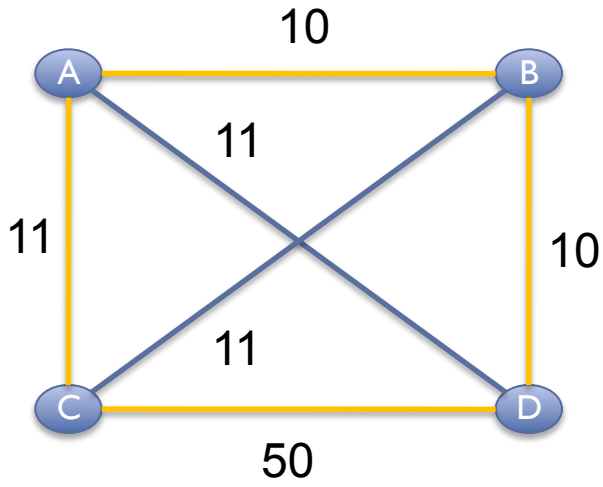
ADBCA: 1+2+1+3= 7 *



Greedy Search

Greedy Search

```
cur=graph.randNode()
while (!done)
    next=cur.getNextClosest()
```



Greedy: ABDCA = 10+10+50+11= 81

Optimal: ACBDA = 11+11+10+11 = 43

Greedy finds the global optimum only when

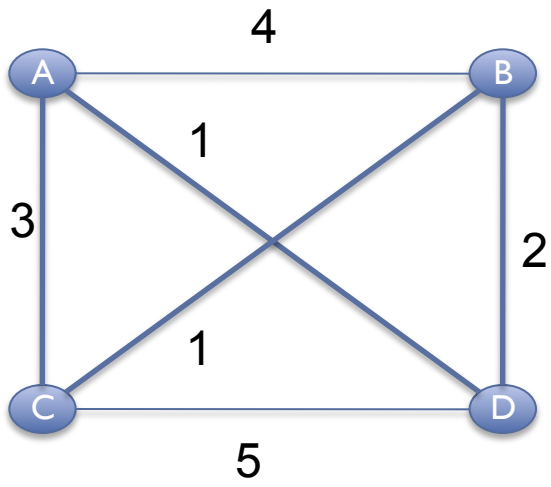
1. Greedy Choice: Local is correct without reconsideration
2. Optimal Substructure: Problem can be split into subproblems

Optimal Greedy: Making change with the fewest number of coins

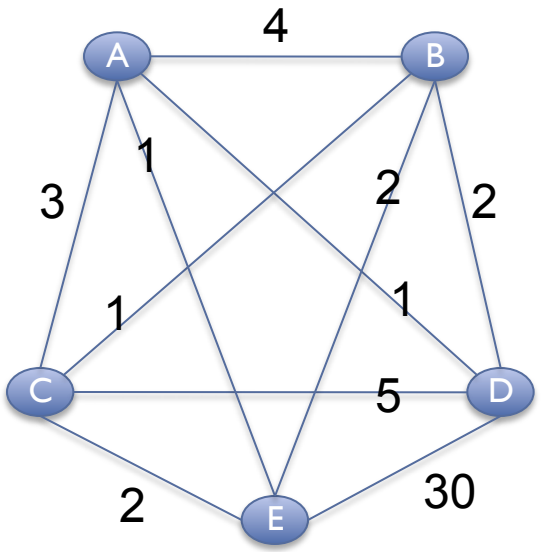
TSP Complexity

- No fast solution
  - Knowing optimal tour through n cities doesn't seem to help much for n+1 cities

[How many possible tours for n cities?]

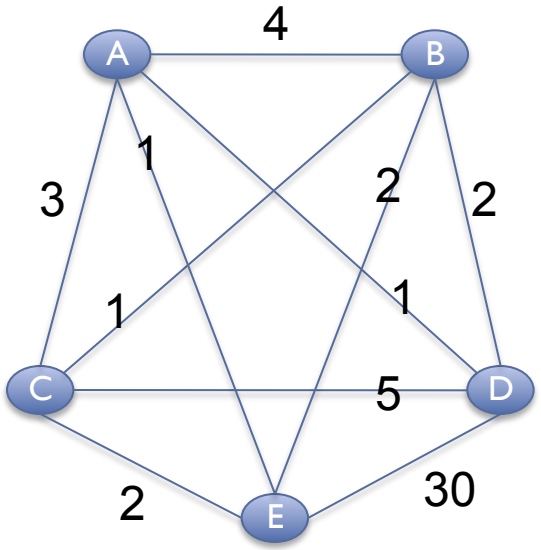


- Extensive searching is the only provably correct algorithm
  - Brute Force: O(n!)
    - ~20 cities max
    - $20! = 2.4 \times 10^{18}$



Branch-and-Bound

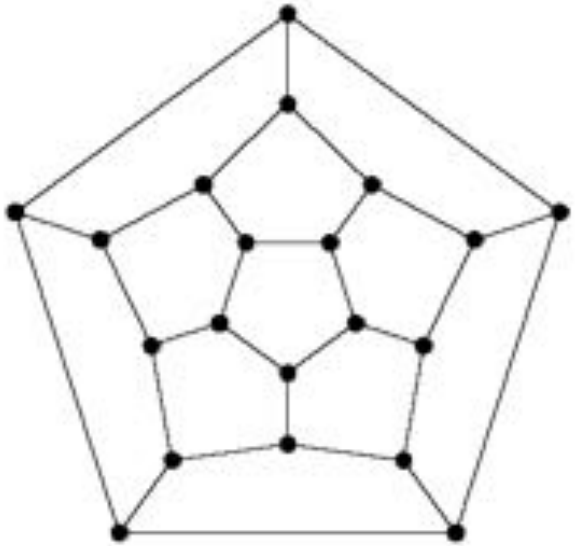
- Abort on suboptimal solutions as soon as possible
  - ADBECA = 1+2+2+2+3 = 10
  - ABDE = 4+2+30 > 10
  - ADE = 1+30 > 10
  - AED = 1+30 > 10
  - …



- Performance Heuristic
  - Always gives the optimal answer
  - Doesn't always help performance, but often does
  - Current TSP record holder:
    - 85,900 cities
    - 85900! = 10^{386526}

[When not?]

TSP and NP-complete

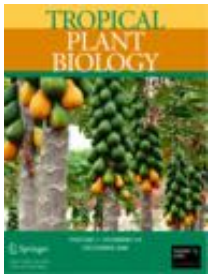


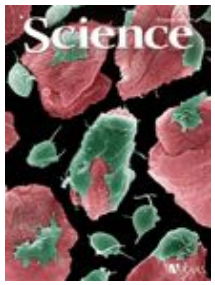
- TSP is one of many extremely hard problems of the class NP-complete
  - Extensive searching is the only way to find an exact solution
  - Often have to settle for approx. solution

- WARNING: Many biological problems are in this class
  - Find a tour the visits every node once (Genome Assembly)
  - Find the smallest set of vertices covering the edges (Essential Genes)
  - Find the largest clique in the graph (Protein Complexes)
  - Find the highest mutual information encoding scheme (Neurobiology)
  - Find the best set of moves in tetris
  - …
  - http://en.wikipedia.org/wiki/List_of_NP-complete_problems

2 minute break

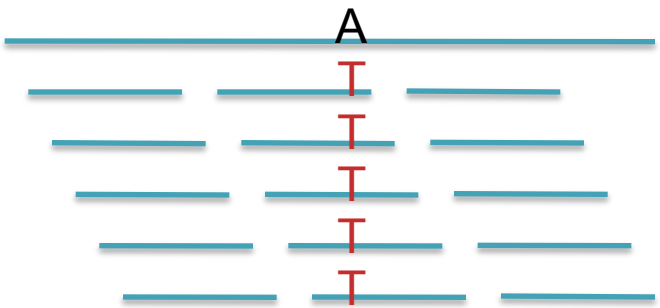
A Little About Me





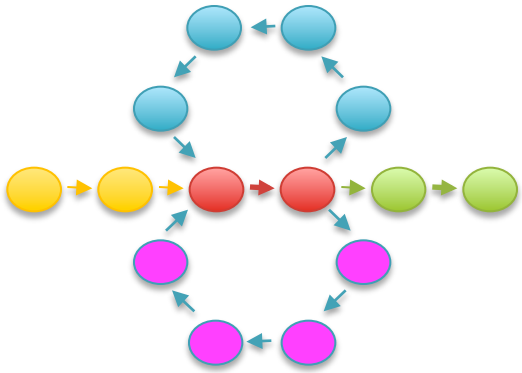
Sequencing Applications

Alignment & Variations

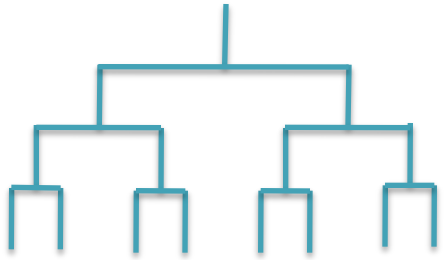


Differential Analysis

De novo Assembly



Phylogeny & Evolution



The DNA Data Race

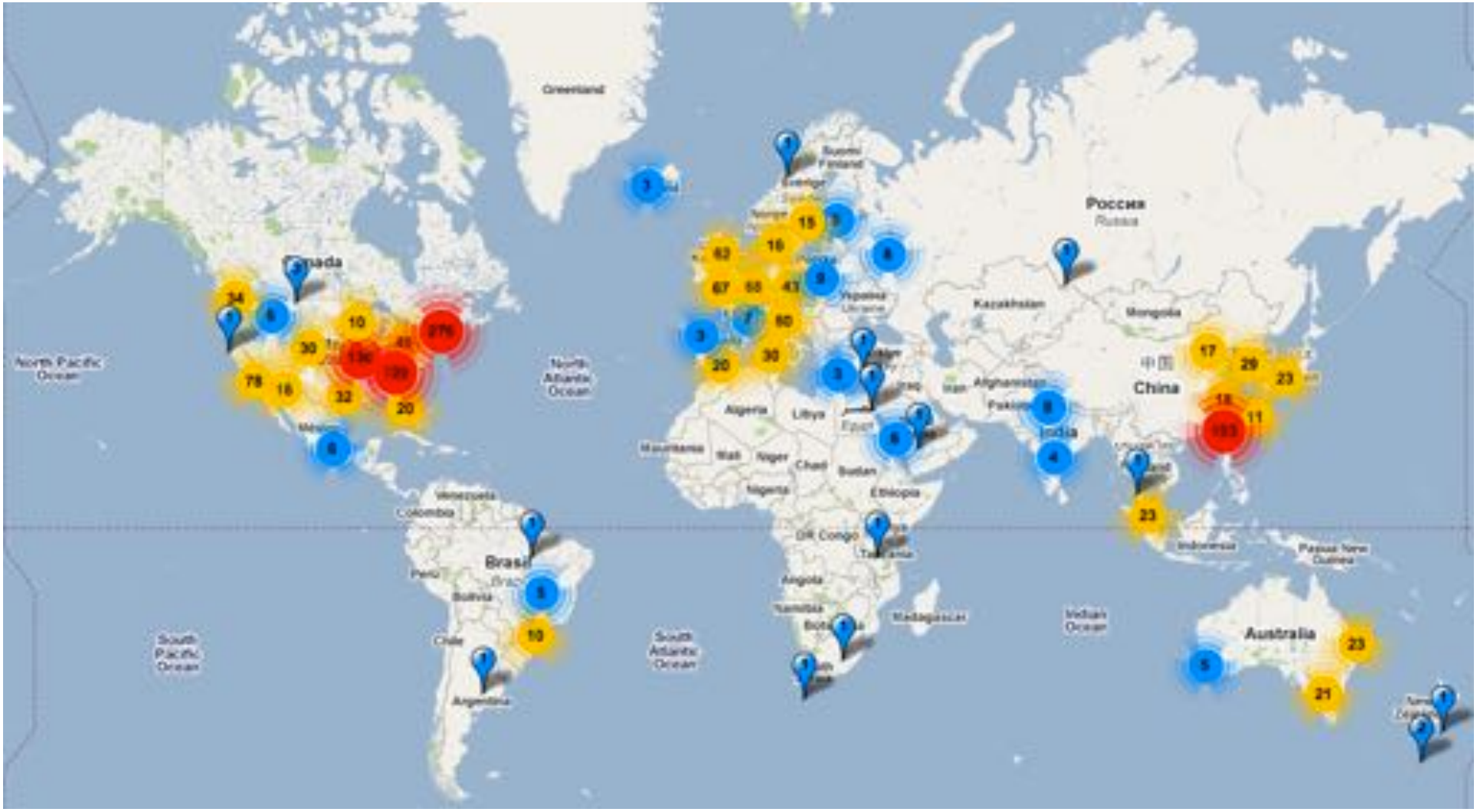
Year	Genome	Technology	Cost
2001	Venter *et al.*	Sanger (ABI)	$300,000,000
2007	Levy *et al.*	Sanger (ABI)	$10,000,000
2008	Wheeler *et al.*	Roche (454)	$2,000,000
2008	Ley *et al.*	Illumina	$1,000,000
2008	Bentley *et al.*	Illumina	$250,000
2009	Pushkarev *et al.*	Helicos	$48,000
2009	Drmanac *et al.*	Complete Genomics	$4,400

(Pushkarev *et al.*, 2009)

Sequencing a single human genome uses ~100 GB of compressed sequence data in billions of short reads.

~20 DVDs / genome

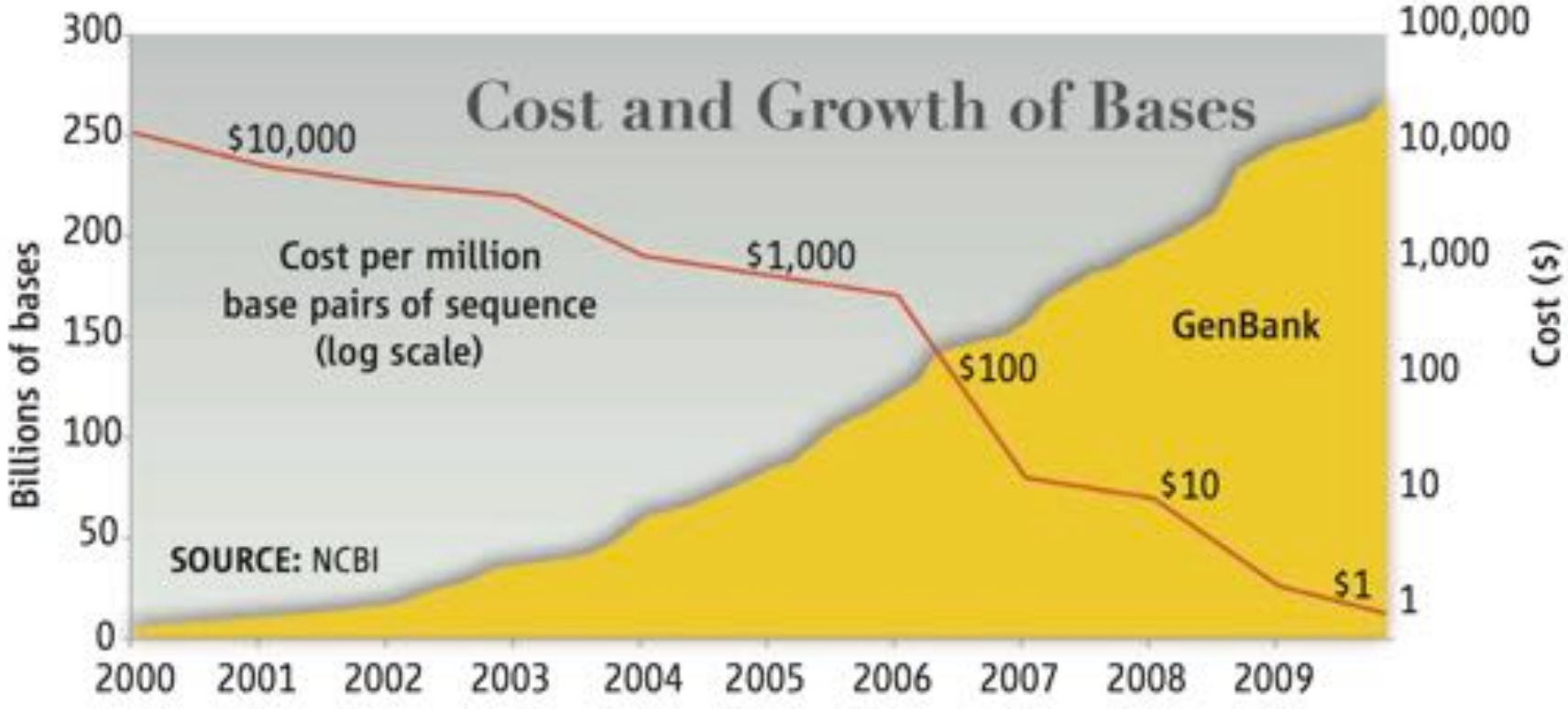
Sequencing Centers



Next Generation Genomics: World Map of High-throughput Sequencers
http://pathogenomics.bham.ac.uk/hts/

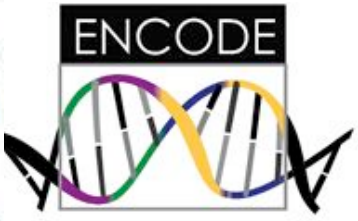
The DNA Data Tsunami

Current world-wide sequencing capacity exceeds 33Tbp/day (12Pbp/year) and is growing at 5x per year!



"Will Computers Crash Genomics?"
Elizabeth Pennisi (2011) *Science*. 331(6018): 666-668.

The DNA Data Tsunami

Use massive amounts of sequencing to explore the genetic origins of life

Our best (only) hope is to use many computers:

- Parallel Computing aka Cloud Computing

- Now your programs will crash on 1000 computers instead of just 1 ☺

Hadoop MapReduce

http://hadoop.apache.org

- MapReduce is Google's framework for large data computations
  - Data and computations are spread over thousands of computers
    - Indexing the Internet, PageRank, Machine Learning, etc… (Dean and Ghemawat, 2004)
    - 946 PB processed in May 2010 (Jeff Dean at Stanford, 11.10.2010)
  - Hadoop is the leading open source implementation
    - Developed and used by Yahoo, Facebook, Twitter, Amazon, etc
    - GATK is an alternative implementation specifically for NGS

- Benefits
  - Scalable, Efficient, Reliable
  - Easy to Program
  - Runs on commodity computers

- Challenges
  - Redesigning / Retooling applications
    - Not Condor, Not MPI
    - Everything in MapReduce

Parallel Algorithm Spectrum

Embarrassingly Parallel

Map-only

Each item is Independent

Loosely Coupled

MapReduce

Independent-Sync-Independent

Tightly Coupled

Iterative MapReduce

Constant Sync

Short Read Mapping

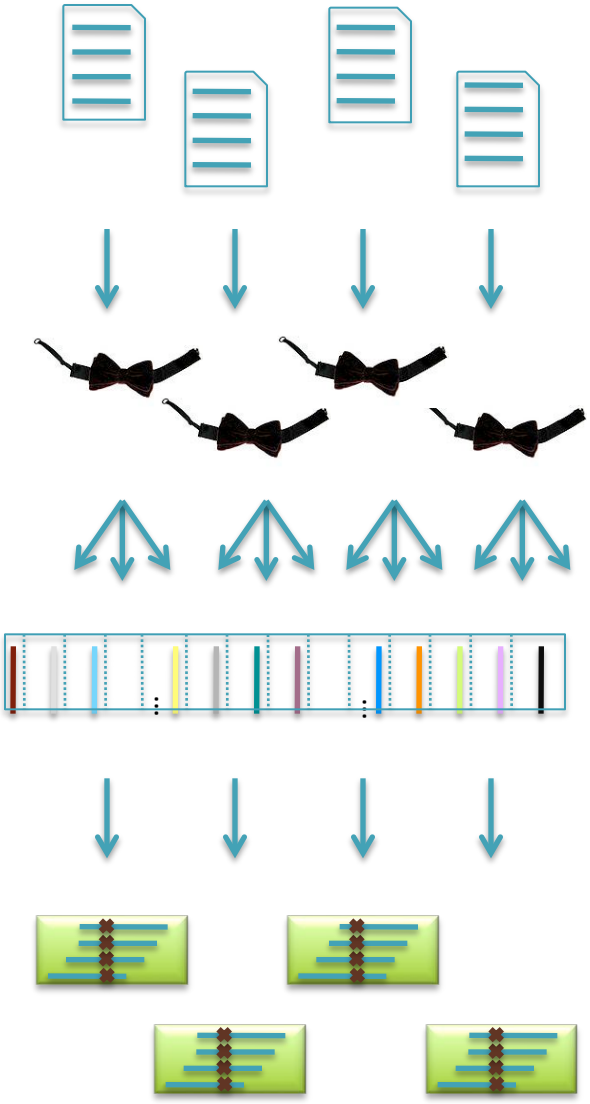
- Given a reference and many subject reads, report one or more "good" end-to-end alignments per alignable read
  - Find where the read most likely originated
  - Fundamental computation for many assays
    - Genotyping
    - Structural Variations
    - RNA-Seq
    - Chip-Seq
    - Methyl-Seq
    - Hi-C-Seq

- Desperate need for scalable solutions
  - Single human requires >1,000 CPU hours / genome

Crossbow

http://bowtie-bio.sourceforge.net/crossbow

- Align billions of reads and find SNPs
  - Reuse software components: Hadoop Streaming
- Map: Bowtie (Langmead *et al.*, 2009)
  - Find best alignment for each read
  - Emit (chromosome region, alignment)
- Shuffle: Hadoop
  - Group and sort alignments by region
- Reduce: SOAPsnp (Li *et al.*, 2009)
  - Scan alignments for divergent columns
  - Accounts for sequencing error, known SNPs



Performance in Amazon EC2

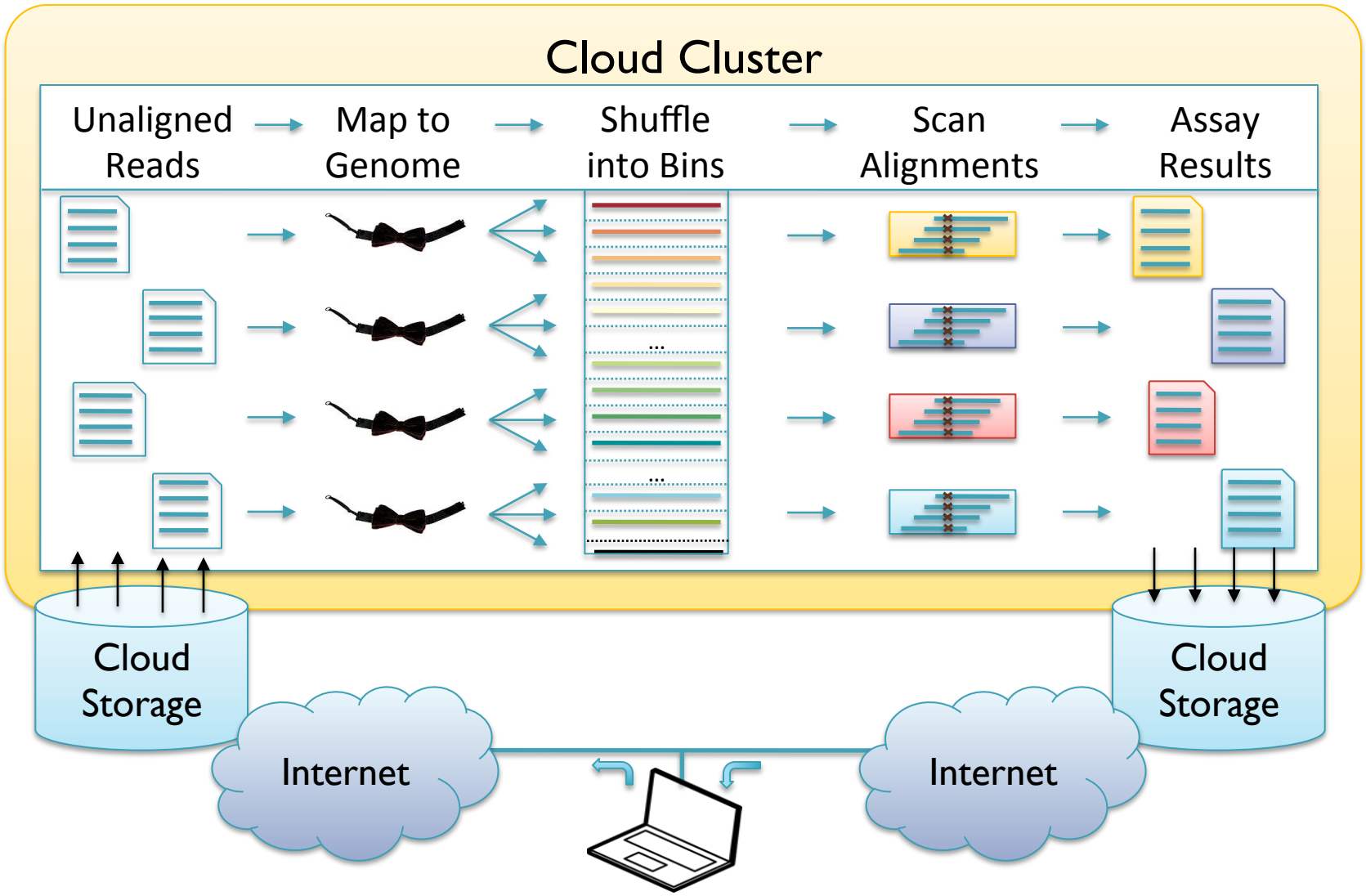
http://bowtie-bio.sourceforge.net/crossbow

	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	$10.65
Data Transfer	1h :15m	40 cores	$3.40
Setup	0h : 15m	320 cores	$13.94
Alignment	1h : 30m	320 cores	$41.82
Variant Calling	1h : 00m	320 cores	$27.88
End-to-end	4h : 00m		$97.69

Discovered 3.7M SNPs in one human genome for ~$100 in an afternoon.
Accuracy validated at >99%

Searching for SNPs with Cloud Computing.
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology*. **10**:R134

Map-Shuffle-Scan for Genomics

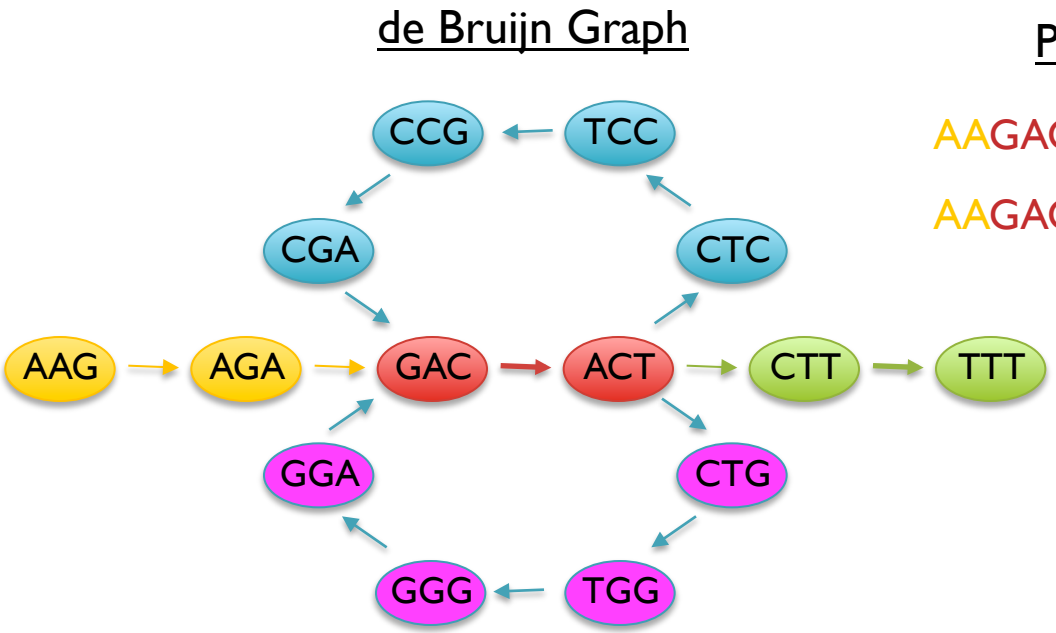


Cloud Computing and the DNA Data Race.
Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology.* **28**:691-693

Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…



Potential Genomes

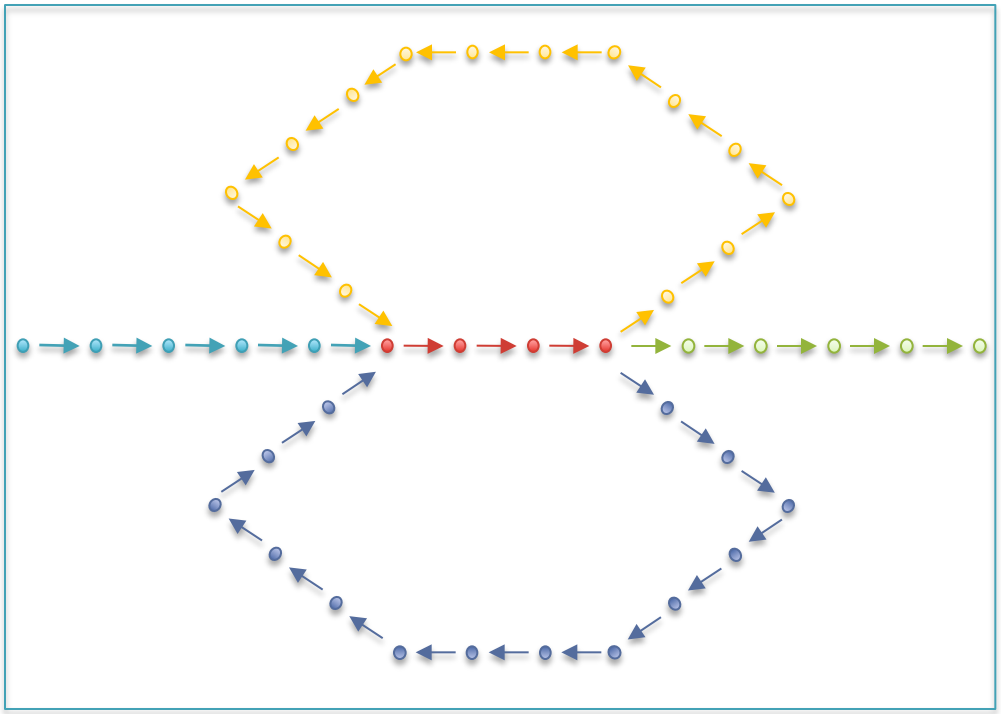
AAGACTCCGACTGGGACTTT

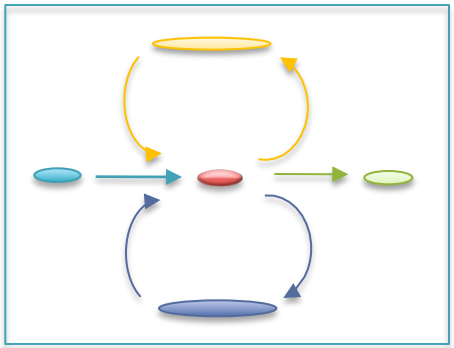
AAGACTGGGACTCCGACTTT

- Genome assembly as finding an Eulerian tour of the de Bruijn graph
  - Human genome: >3B nodes, >10B edges
- The new short read assemblers require tremendous computation
  - Velvet (Zerbino & Birney, 2008) serial: > 2TB of RAM
  - ABySS (Simpson *et al.*, 2009) MPI: 168 cores x ~96 hours
  - SOAPdenovo (Li *et al.*, 2010) pthreads: 40 cores x 40 hours, >140 GB RAM

Graph Compression

- After construction, many edges are unambiguous
  - Merge together compressible nodes
  - Graph physically distributed over hundreds of computers





Warmup Exercise

- Who here was born closest to July 8?
  - You can only compare to 1 other person at a time

Find winner among 64 teams in just 6 rounds

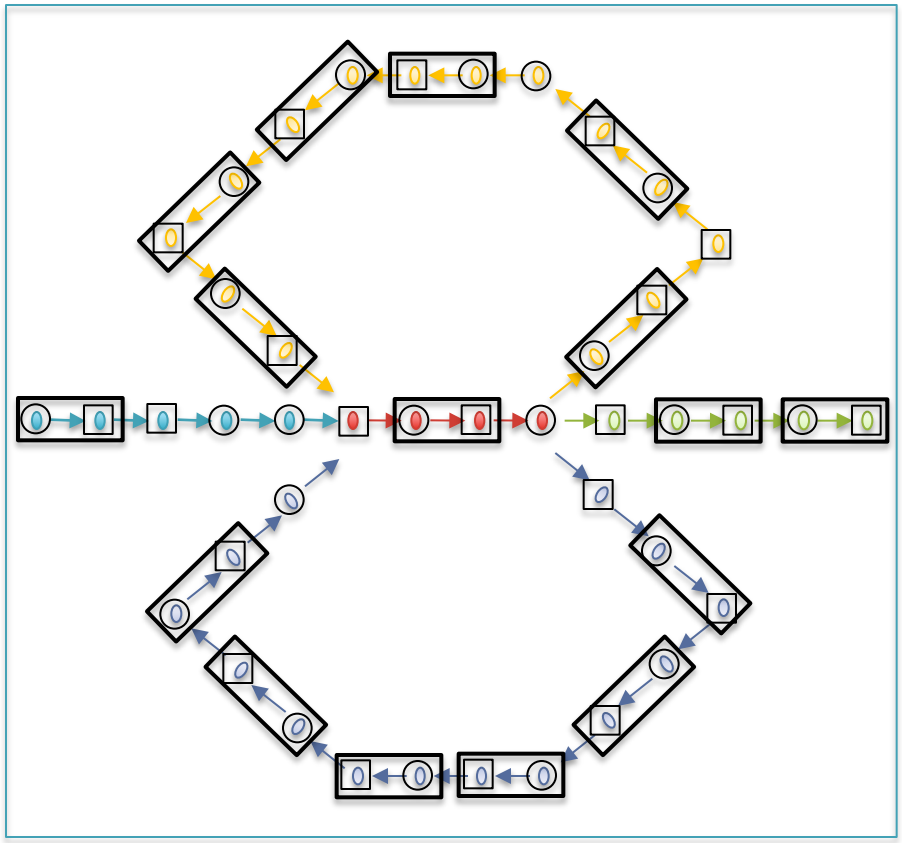
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign (H) / [T] to each compressible node
- Compress (H) → [T] links



Initial Graph: 42 nodes

Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

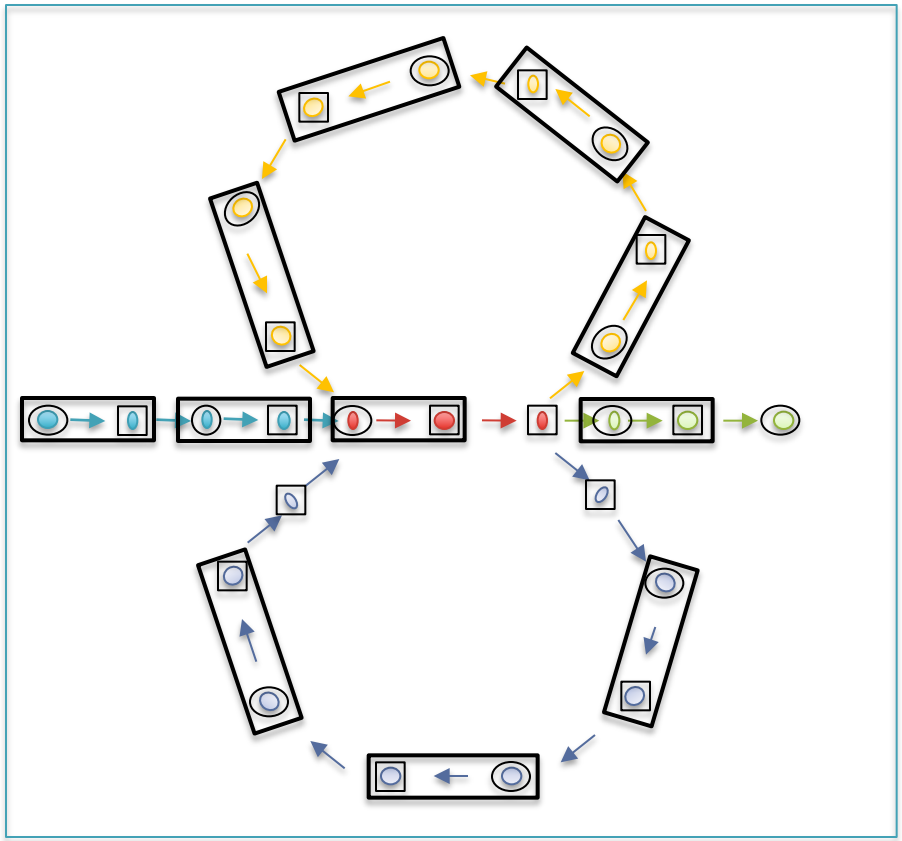
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [T] links



Round 1: 26 nodes (38% savings)

Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

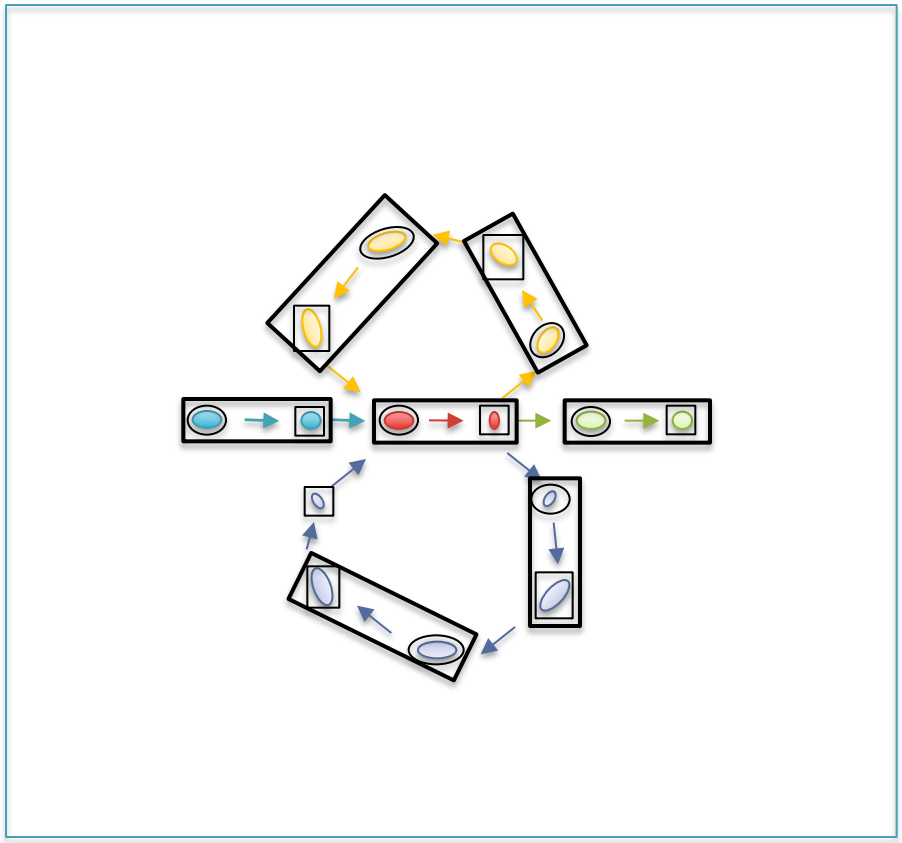
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [T] links



Round 2: 15 nodes (64% savings)

Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

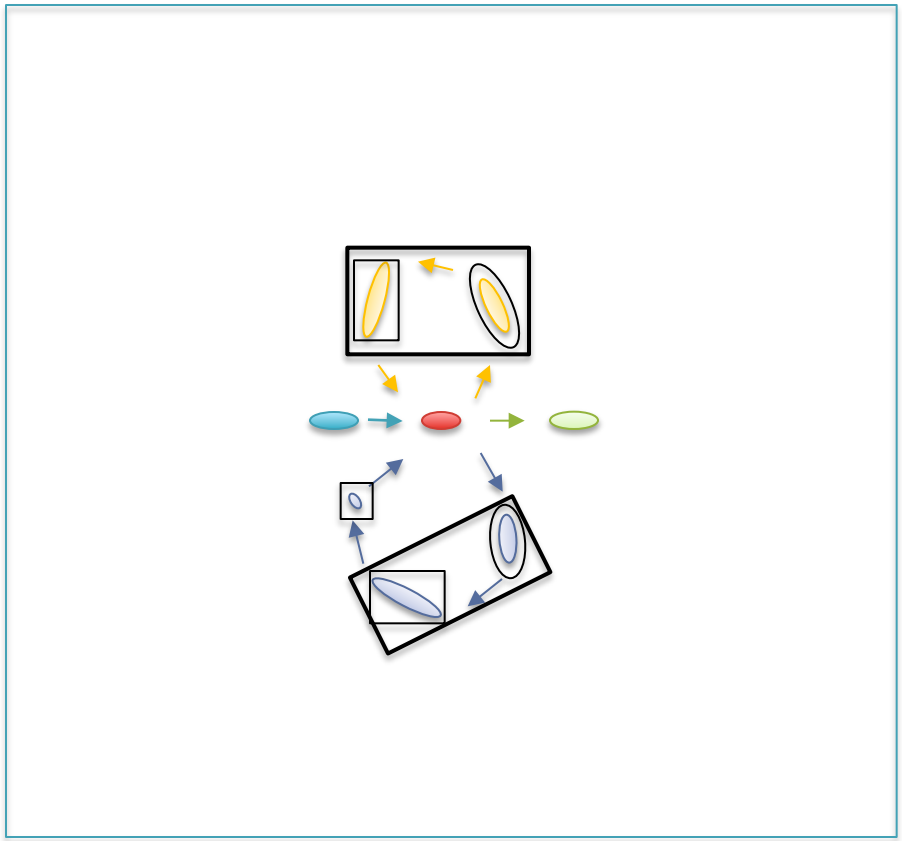
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [T] links



Round 2: 8 nodes (81% savings)

Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

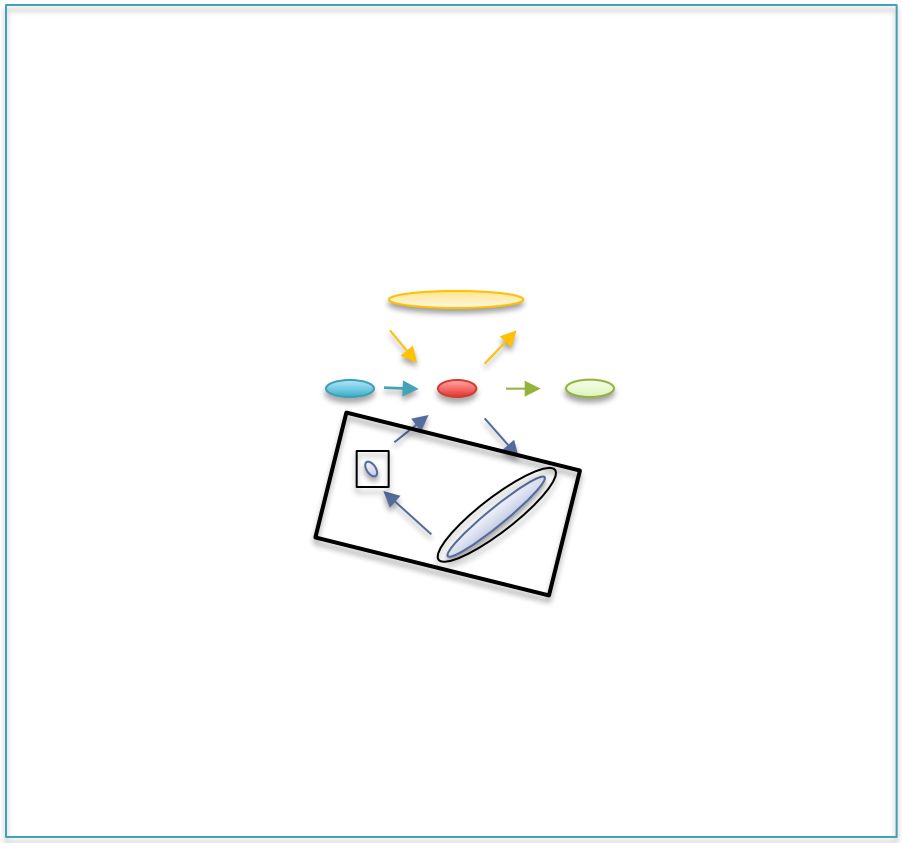
Fast Path Compression

Challenges

- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [T] links



Round 3: 6 nodes (86% savings)

Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

Fast Path Compression

Challenges

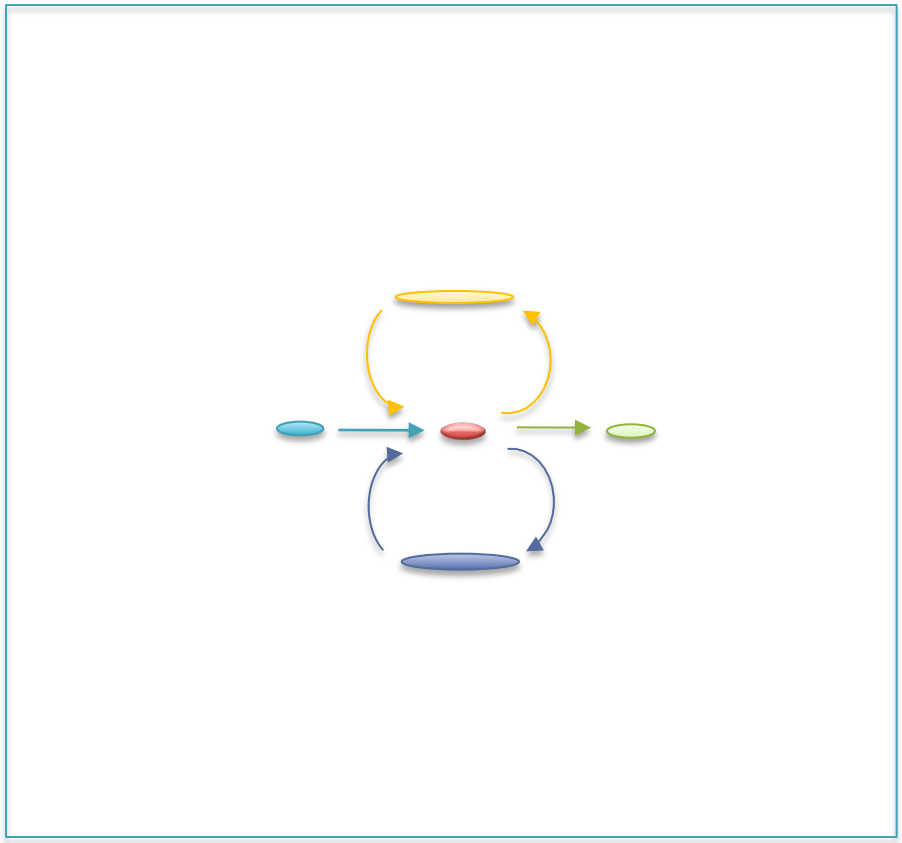
- Nodes stored on different computers
- Nodes can only access direct neighbors

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ → [T] links

Performance

- Compress all chains in log(S) rounds



Round 4: 5 nodes (88% savings)

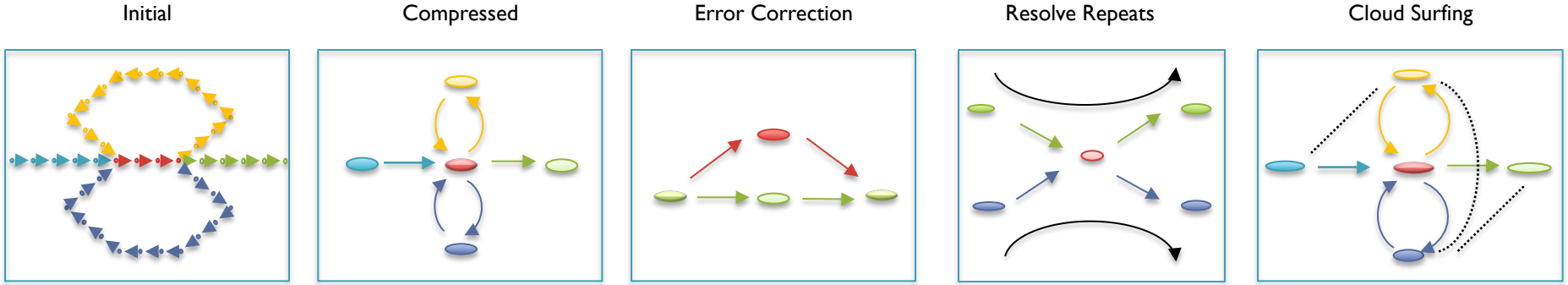
Randomized Speed-ups in Parallel Computation.
Vishkin U. (1984) *ACM Symposium on Theory of Computation. 230-239.*

Contrail

http://contrail-bio.sourceforge.net

De novo bacterial assembly

- *Genome: E. coli* K12 MG1655, 4.6Mbp
- *Input:* 20.8M 36bp reads, 200bp insert (~150x coverage)
- *Preprocessor*: Quake Error Correction



N	5.1 M	245,131	2,769	1,909	300
Max	27 bp	1,079 bp	70,725 bp	90,088 bp	149,006 bp
N50	27 bp	156 bp	15,023 bp	20,062 bp	54,807 bp

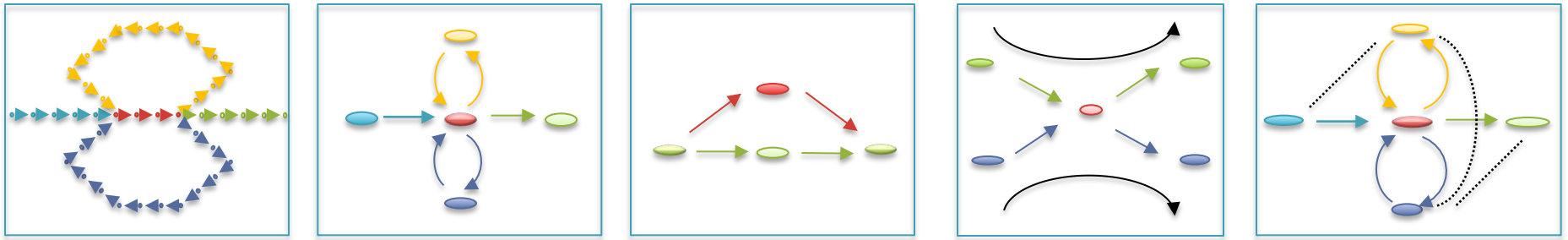
Assembly of Large Genomes with Cloud Computing.
Schatz MC, Sommer D, Kelley D, Pop M, *et al. In Preparation.*

Contrail

http://contrail-bio.sourceforge.net

De novo Assembly of the Human Genome

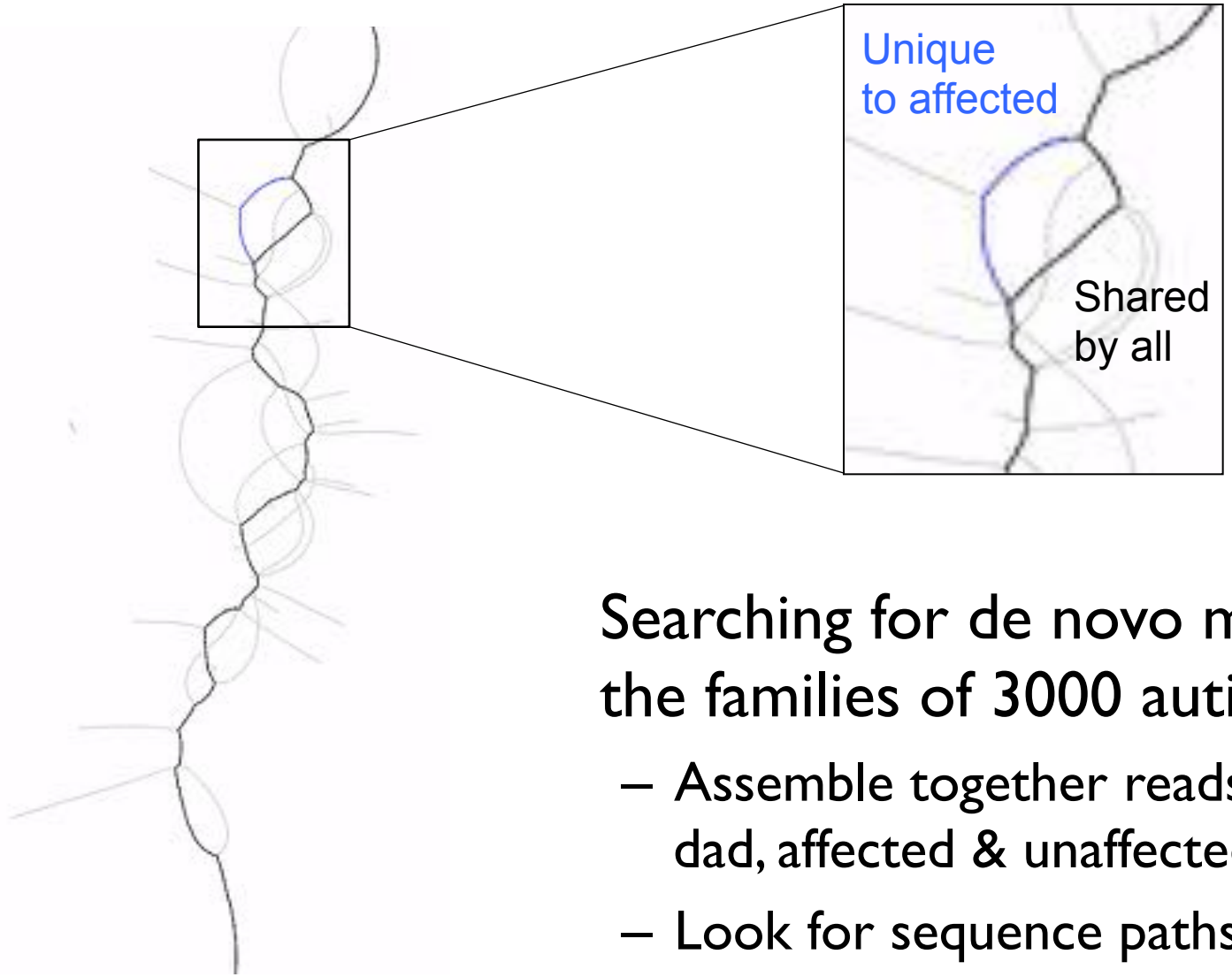
- *Genome:* African male NA18507 (SRA000271, Bentley *et al.*, 2008)
- *Input: 3.5*B 36bp reads, 210bp insert (~40x coverage)



N	>7 B	>1 B	4.2 M	4.1 M	3.3 M
Max	27 bp	303 bp	20,594 bp	20,594 bp	20,594 bp
N50	27 bp	< 100 bp	995 bp	1,050 bp	1,427 bp*

Assembly of Large Genomes with Cloud Computing.
Schatz MC, Sommer D, Kelley D, Pop M, *et al. In Preparation.*

De novo mutations and de Bruijn Graphs

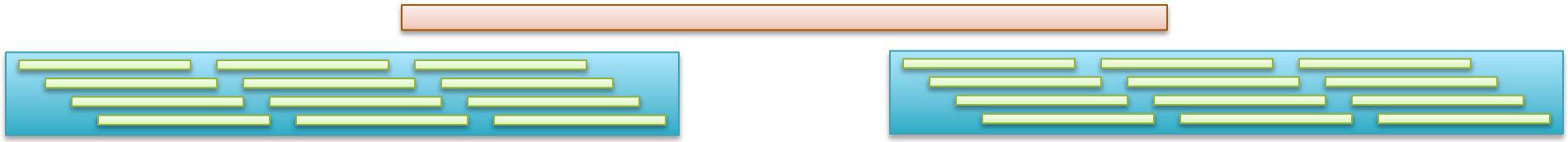


COLEC12
C->A

Searching for de novo mutations in the families of 3000 autistic children.

- Assemble together reads from mom, dad, affected & unaffected children
- Look for sequence paths unique to affected child

Illumina/PacBio Hybrid Assembly



Find long reads that align well to the ends of the contigs/scaffolds

- Require >80% sequence identity
- Require <100bp overhang
- Require >1 read spans gap

Require >50bp match length
Require >-50bp gap span
Require >500bp contig length

Yeast

151 linked scaffold pairs
Gap sizes: 289 +/- 270bp
Max Gap: 1582bp

Scaffold N50: 125kbp (+54%)
Scaffolds >500bp: 242 (-36%)
Scaffolds >1kbp : 210 (-28%)

Rice

14890 linked scaffold pairs
Gap sizes: 240.5 +/- 269.4
Max Gap: 2680bp

Scaffold N50: ----
(4000 CPU hours until failure)

Structural Variations in Cancer

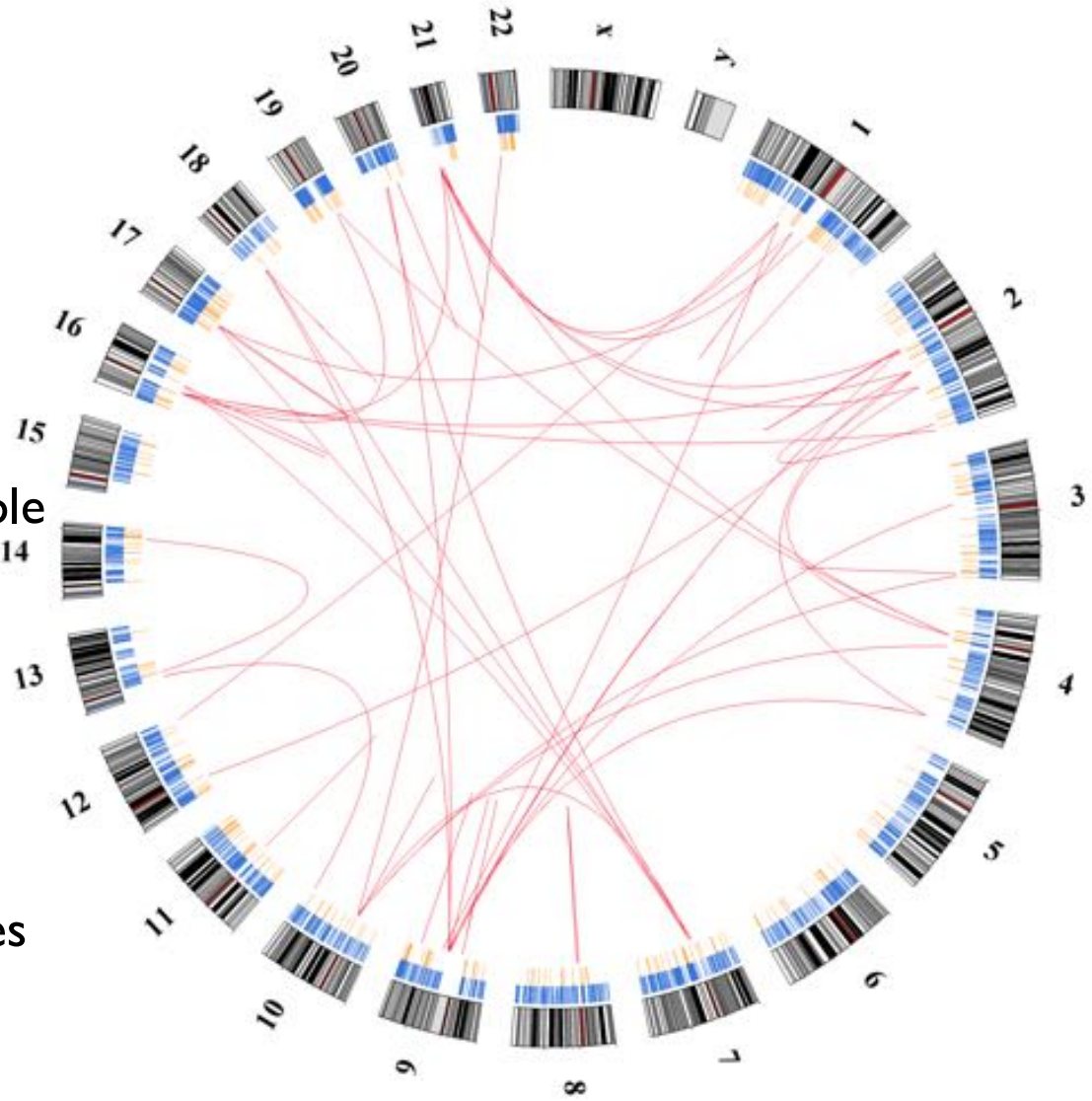
Use short reads to discover large scale variations

- Discordant Pairs Analysis with Hydra (Quinlan et al. 2010)

Circos plot of high confidence SVs specific to esophageal cancer sample

- Red: SV links
- Orange: 375 cancer genes
- Blue: 4950 disease genes

Detailed analysis of disrupted genes and fusion genes in progress

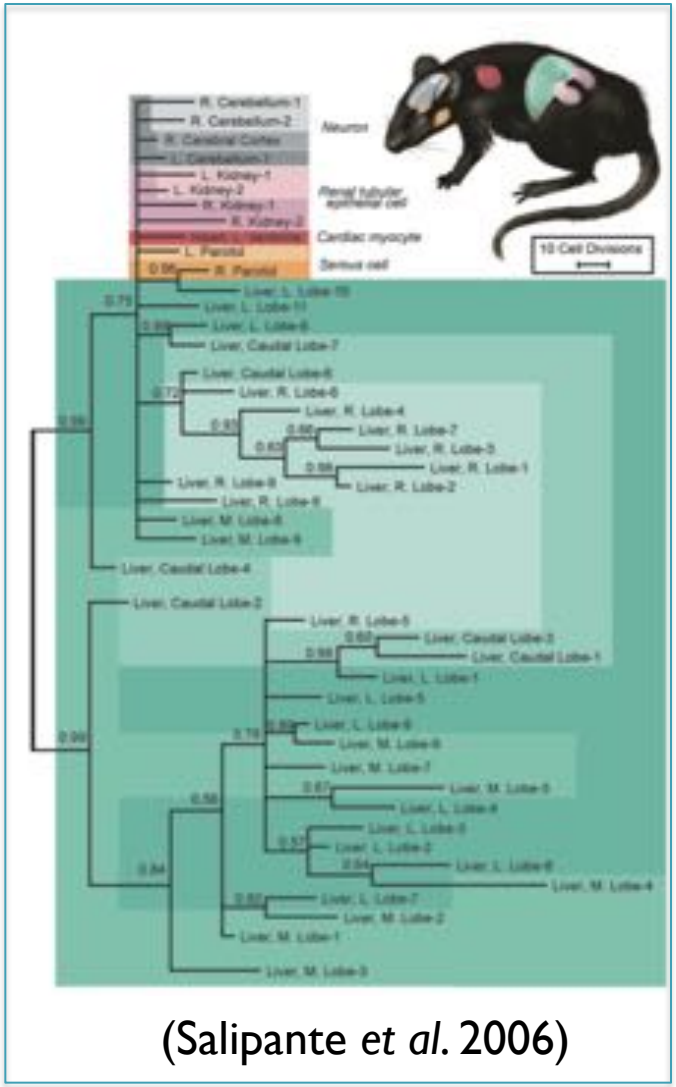


MicroSeq: NextGen Microsatellite Profiling

Mitchell Bekritsky, WSBS

- Class of simple sequence repeats
  - …GCACACACACAT… = …G$(CA)_5$T…
  - Created and mutate primarily through slippage during replication
  - Highly variable & ubiquitous

- Genotyping with SeqMS
  - Rapidly detect MS sequences
  - Map reads using a new MS-mapper
  - Analyze profiles in cells, across cells, & across populations
    - Loss of heterozygosity
    - Development of somatic & cancer cells
    - Relations across strains, across species
    - etc…



(Salipante *et al.* 2006)

Summary

- We are witnessing the dawn of the digital age of biology
  - Next generation sequencing, microarrays, mass spectrometry, microscopy, ecology, etc
- Modern biology requires (is) quantitative biology
  - Computational, mathematical, and statistical techniques applied to analyze, integrate, and interpret biological sensor data
- Don't let the data tsunami crash on you
  - Study, practice, collaborate with quantitative techniques

Acknowledgements

Schatzlab
Matt Titmus
Hayan Lee
Mitch Bekritsky
Paul Baranay
Rohith Menon
Goutham Bhat
James Gurtowski

CSHL
Dick McCombie
Melissa Kramer
Laura Gelley
Sneh Lata Fnu
Stephanie Muller

Wigler Lab

NBACC
Adam Phillippy
Sergey Koren

UMD
Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell

Thank You