Scalable Solutions for DNA Sequence Analysis

Michael Schatz, Ph.D.

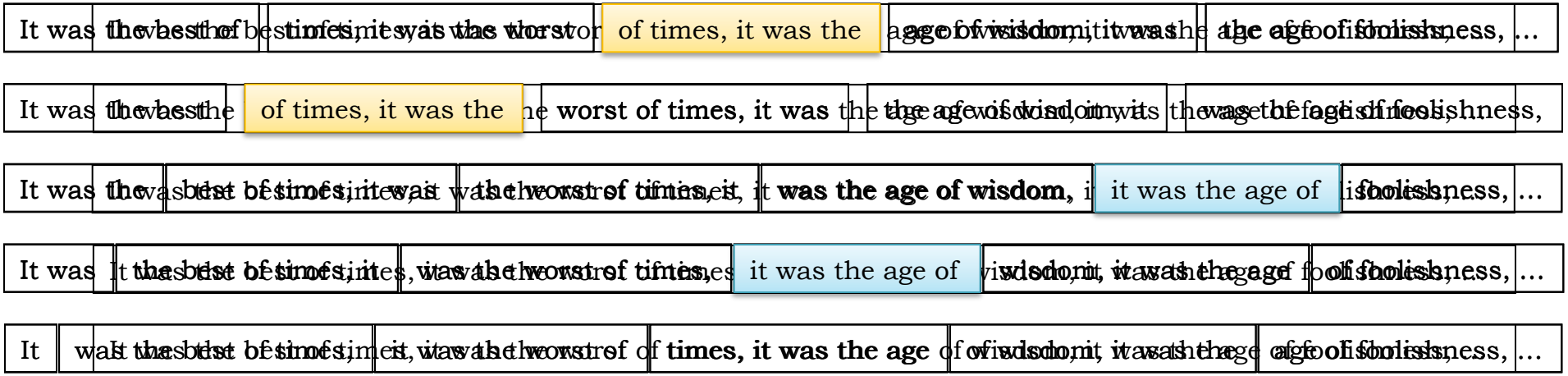
September 15, 2010
Research Topics in Biology

Outline

1. Genome Assembly by Analogy
2. DNA Sequencing and Genomics
3. Large Scale Sequence Analysis
    1. Mapping & Genotyping
    2. Genome Assembly

Shredded Book Reconstruction

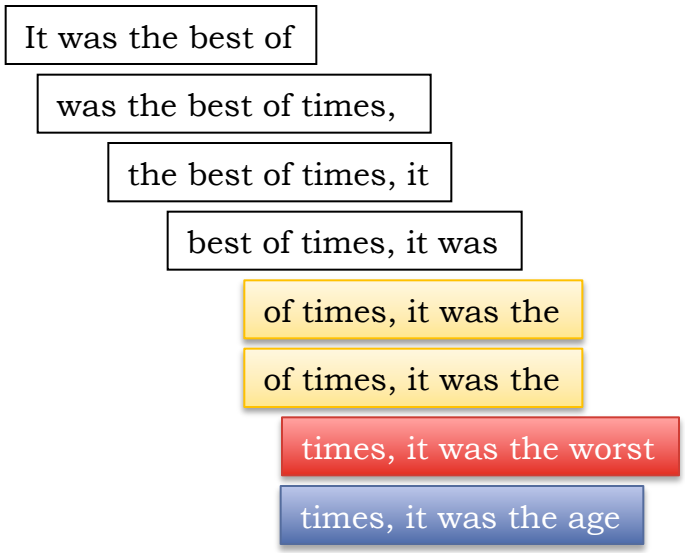
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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



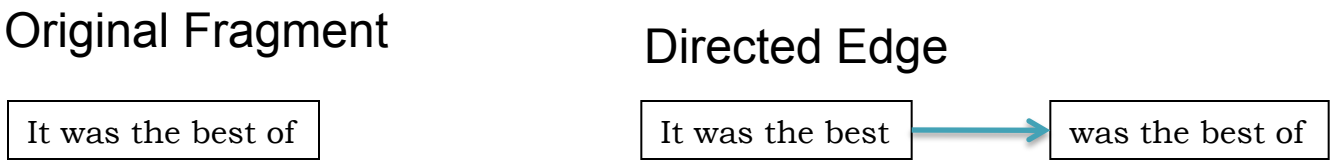
The repeated sequence make the correct reconstruction ambiguous

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

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

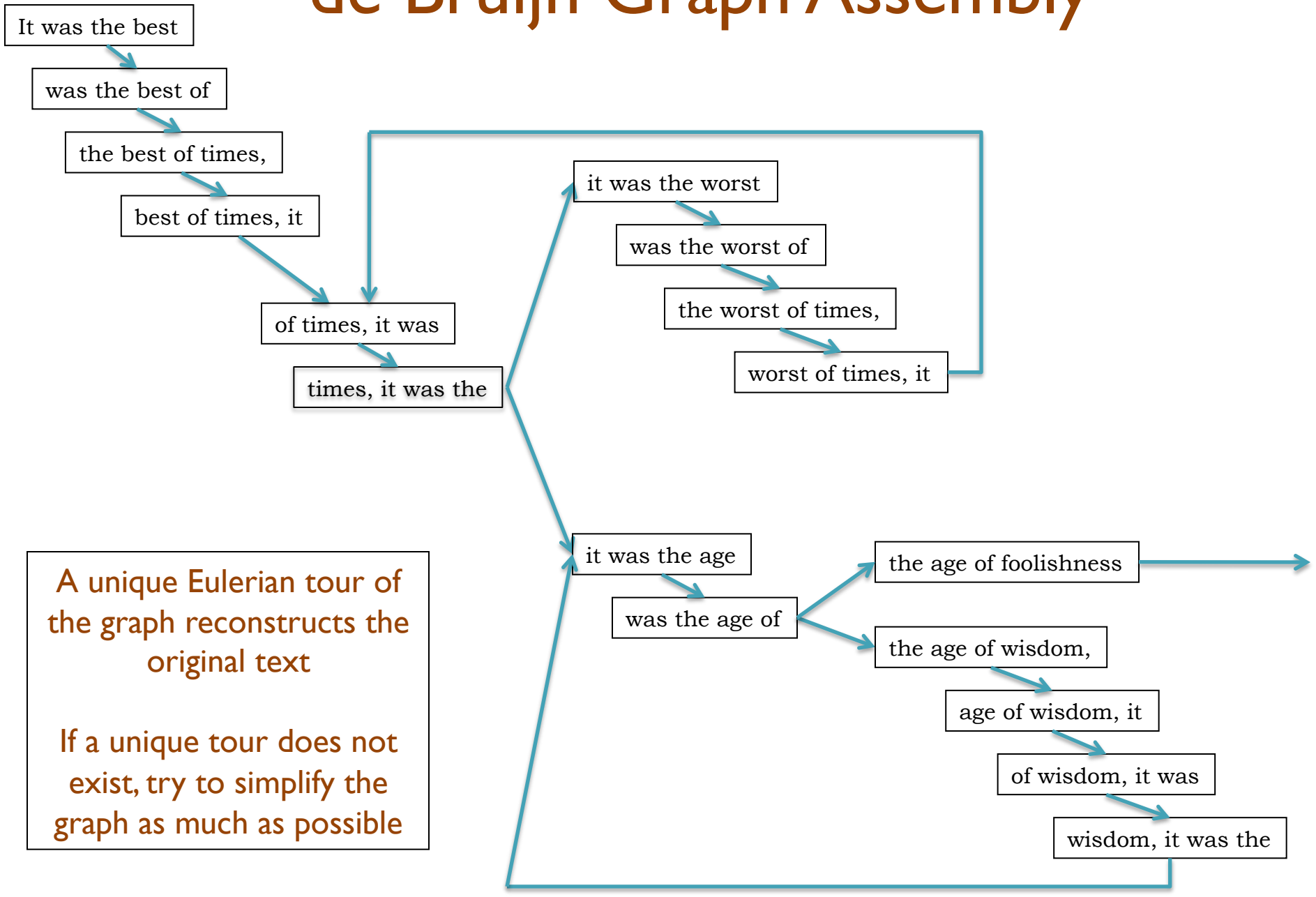
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

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

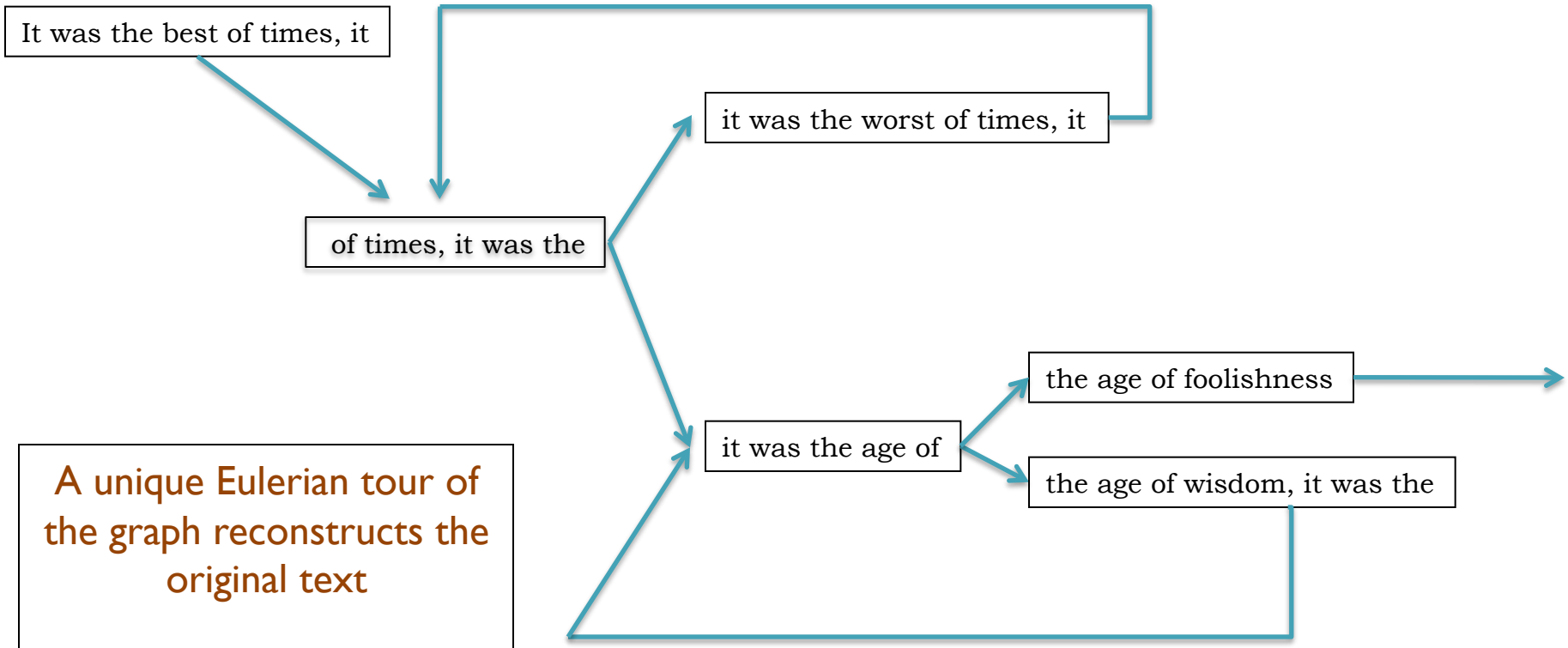
de Bruijn Graph Assembly



A unique Eulerian tour of the graph reconstructs the original text

If a unique tour does not exist, try to simplify the graph as much as possible

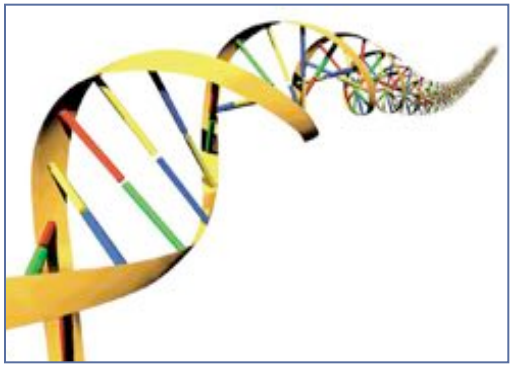
de Bruijn Graph Assembly



A unique Eulerian tour of the graph reconstructs the original text

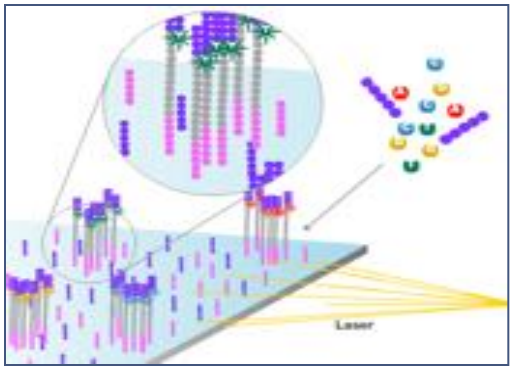
If a unique tour does not exist, try to simplify the graph as much as possible

Molecular Biology & DNA Sequencing



Genome of an organism encodes the genetic information in long sequence of 4 DNA nucleotides: ACGT

- Bacteria: ~3 million bp
- Humans: ~3 billion bp



Current DNA sequencing machines can sequence millions of short (25-500bp) reads from random positions of the genome

- Per-base error rate estimated at 1-2% (Simpson *et al*, 2009)

ATCTGATAAGTCCCAGGACTTCAGT

GCAAGGCAAACCCGAGCCCAGTTT

TCCAGTTCTAGAGTTTCACATGATC

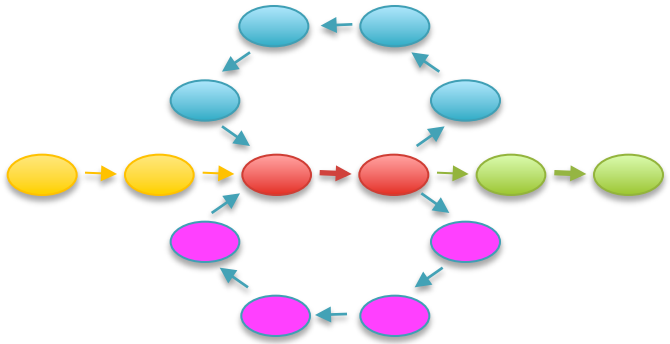
GGAGTTAGTAAAAGTCCACATTGAG

Like Dickens, we can only sequence small fragments of the genome at once.

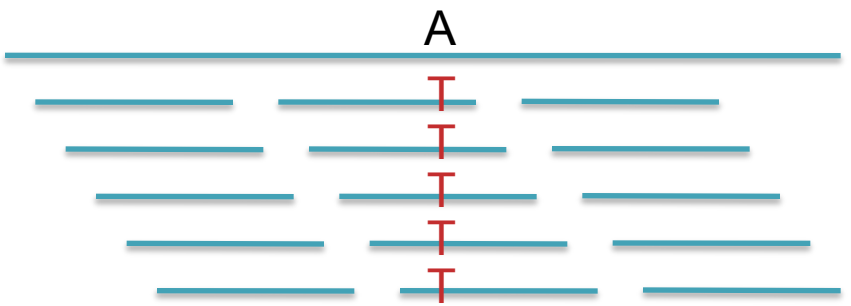
- Use software to analyze the sequences
- Modern Biology requires Computational Biology

DNA Sequence Analysis

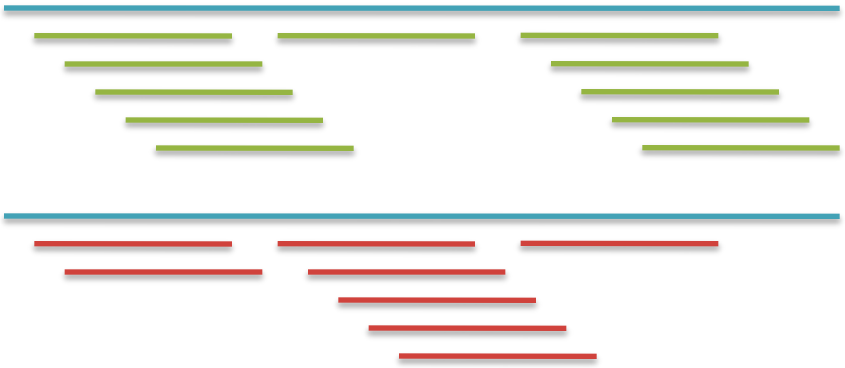
Assembly



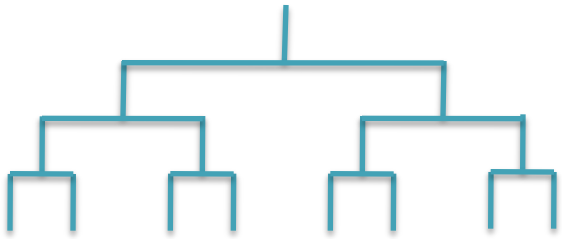
Alignment & Variations



Coverage Analysis



Phylogeny & Evolution



Quantitative Biology Class: Oct 12 – Nov 5

The DNA Data Race

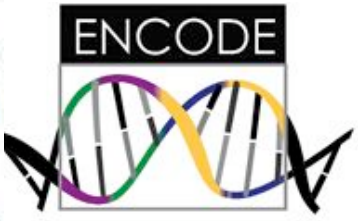
Year	Genome	Technology	Cost
2001	Venter *et al.*	Sanger (ABI)	$300,000,000
2007	Levy *et al.*	Sanger (ABI)	$10,000,000
2008	Wheeler *et al.*	Roche (454)	$2,000,000
2008	Ley *et al.*	Illumina	$1,000,000
2008	Bentley *et al.*	Illumina	$250,000
2009	Pushkarev *et al.*	Helicos	$48,000
2009	Drmanac *et al.*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

The DNA Data Tsunami



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

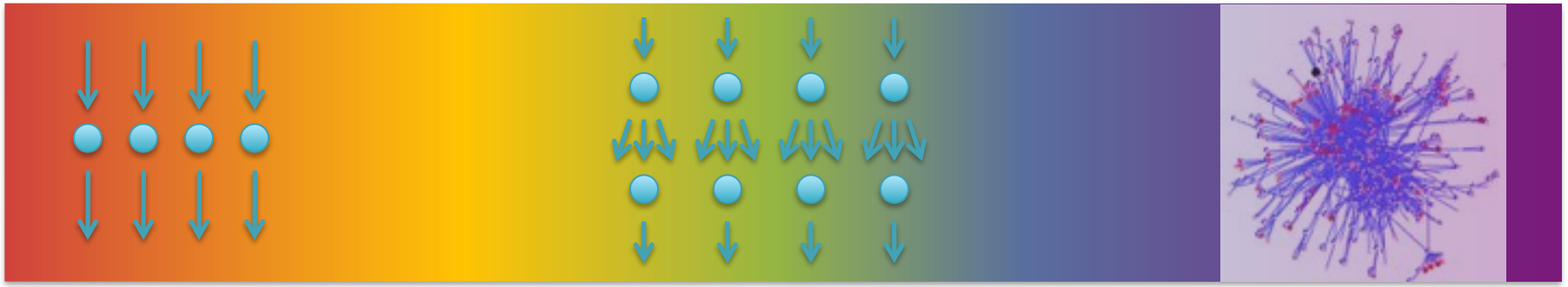
Our best (only) hope is to use many computers:

- Parallel Computing aka Cloud Computing

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

Parallel Computing Spectrum

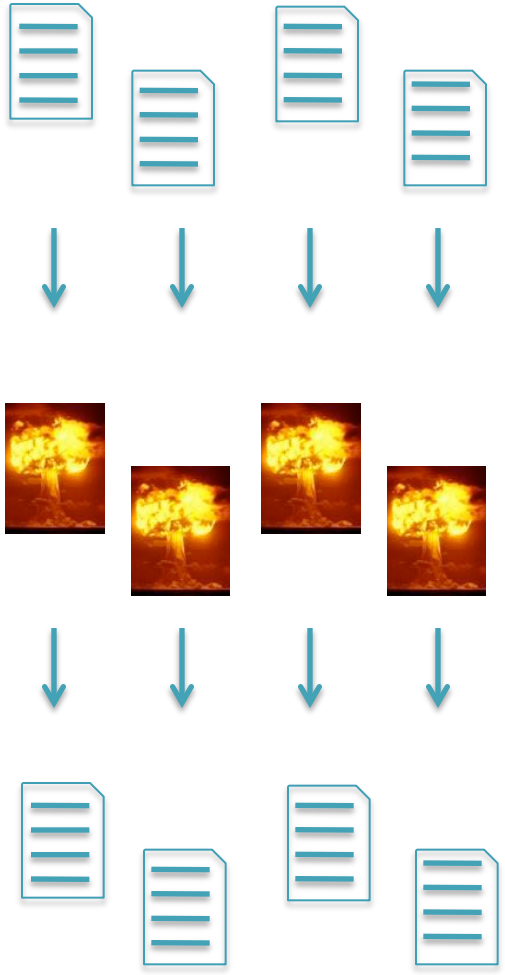
Embarrassingly Parallel	Loosely Coupled	Tightly Coupled



Batch Computing	MapReduce/DryadLINQ	MPI/MapReduce/Pregel
Alignment HMM Scoring	Genotyping K-mer Counting	Graph Analysis Genome Assembly
Scheduling + Load Balance	Embarrassingly Parallel + Parallel Communication	Loosely Coupled + Parallel Algorithm Design

Embarrassingly Parallel

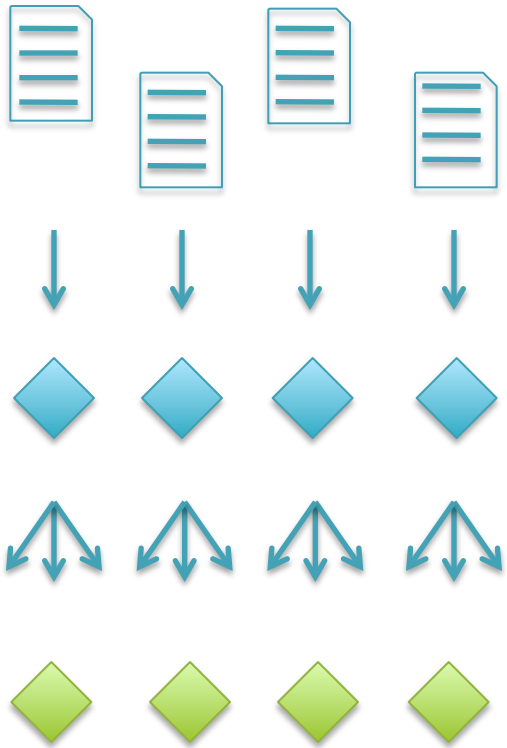
- Batch computing
  - Each item is independent
  - Split input into many chunks
  - Process each chunk separately on a different computer
- Challenges
  - Distributing work, load balancing, monitoring & restart
- Technologies
  - Condor, Sun Grid Engine
  - Amazon Simple Queue



Elementary School Dance

Loosely Coupled

- Divide and conquer
  - Independently process many items
  - Group partial results
  - Scan partial results into final answer
- Challenges
  - Batch computing challenges
  - + Shuffling of huge datasets
- Technologies
  - Hadoop, Elastic MapReduce, Dryad
  - Parallel Databases



Junior High Dance

Short Read Mapping

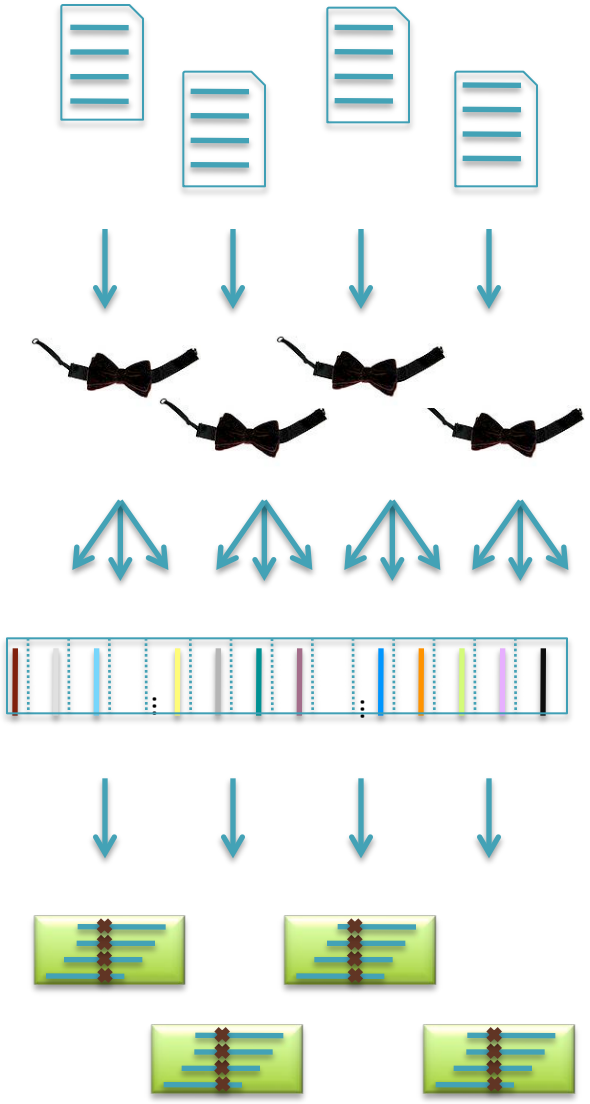
- Given a reference and many subject reads, report one or more "good" end-to-end alignments per alignable read
  - Find where the read most likely originated
  - Fundamental computation for many assays
    - Genotyping
    - Structural Variations
    - RNA-Seq
    - Chip-Seq
    - Methyl-Seq
    - Hi-C-Seq

- 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
- Map: Bowtie (Langmead *et al.*, 2009)
  - Find best alignment for each read
  - Uses a concise index of the genome
- Shuffle: Hadoop
  - Group and sort alignments by region
- Scan: 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

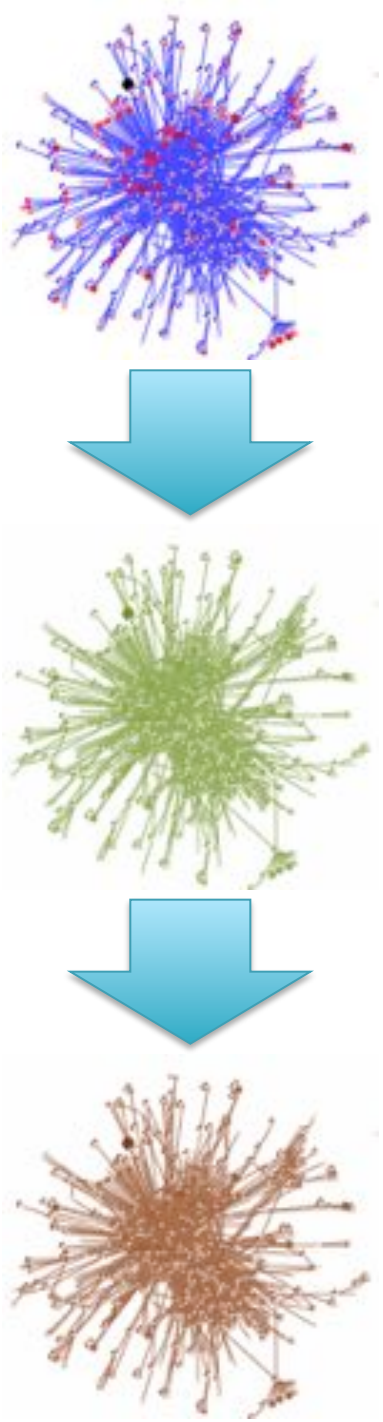
	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	$10.65
Data Transfer	1h :15m	40 cores	$3.40
Setup	0h : 15m	320 cores	$13.94
Alignment	1h : 30m	320 cores	$41.82
Variant Calling	1h : 00m	320 cores	$27.88
End-to-end	4h : 00m		$97.69

Analyze an entire 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

Tightly Coupled

- Computation that cannot be partitioned
  - Graph Analysis
  - Molecular Dynamics
  - Population simulations
- Challenges
  - Loosely coupled challenges
  - + Parallel algorithms design
- Technologies
  - MPI
  - MapReduce, Dryad, Pregel

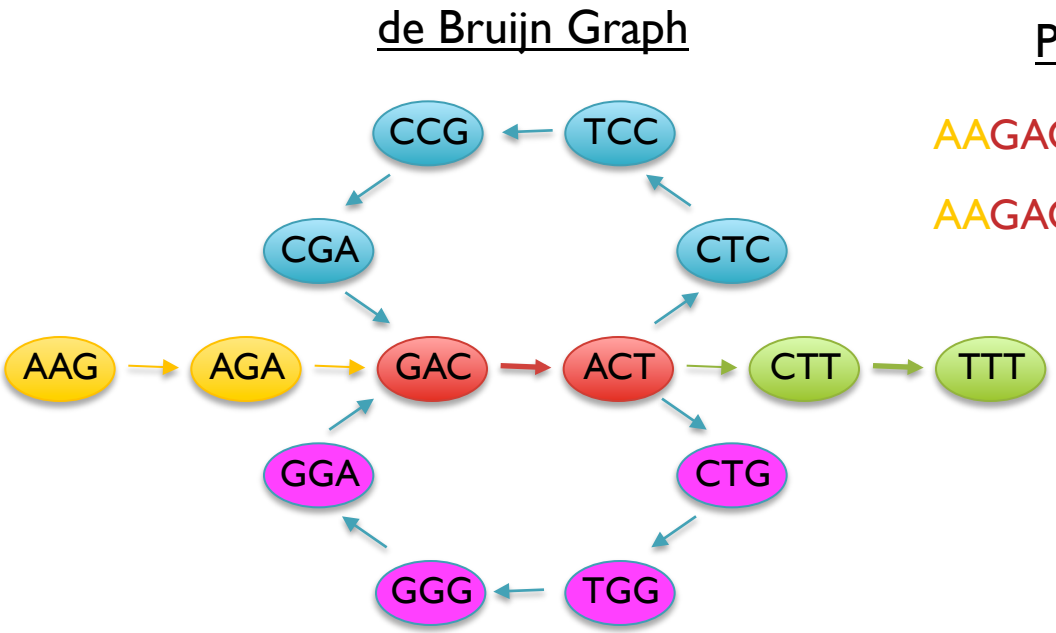


High School Dance

Short Read Assembly

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…



Potential Genomes

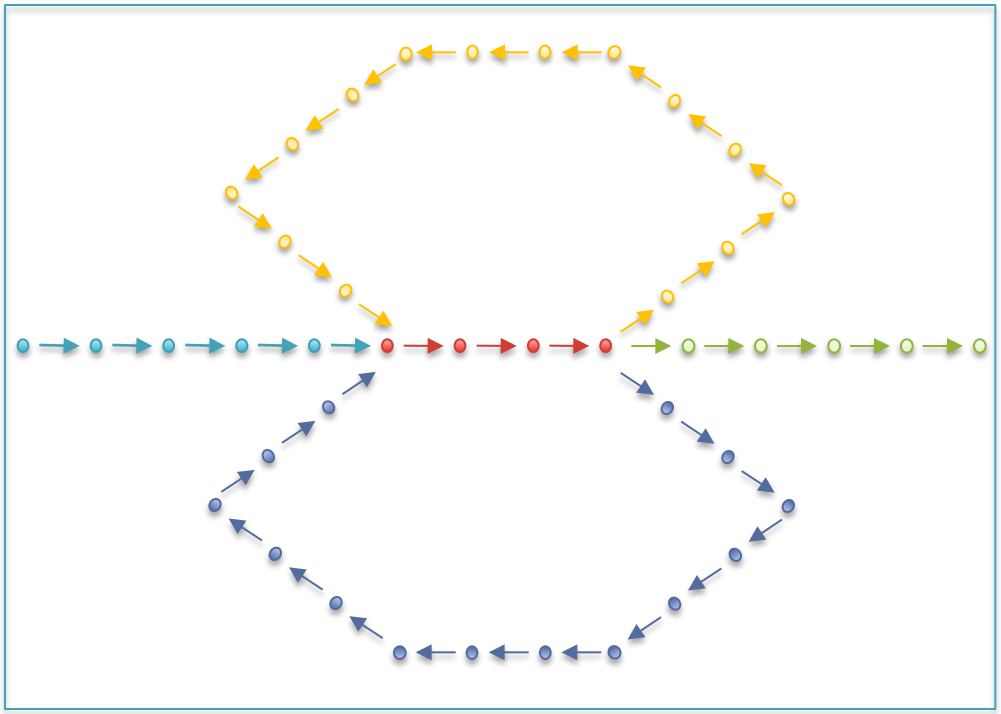
AAGACTCCGACTGGGACTTT

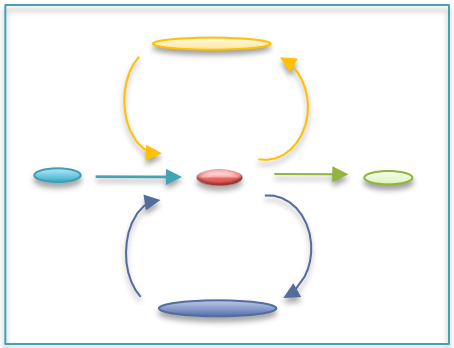
AAGACTGGGACTCCGACTTT

- 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

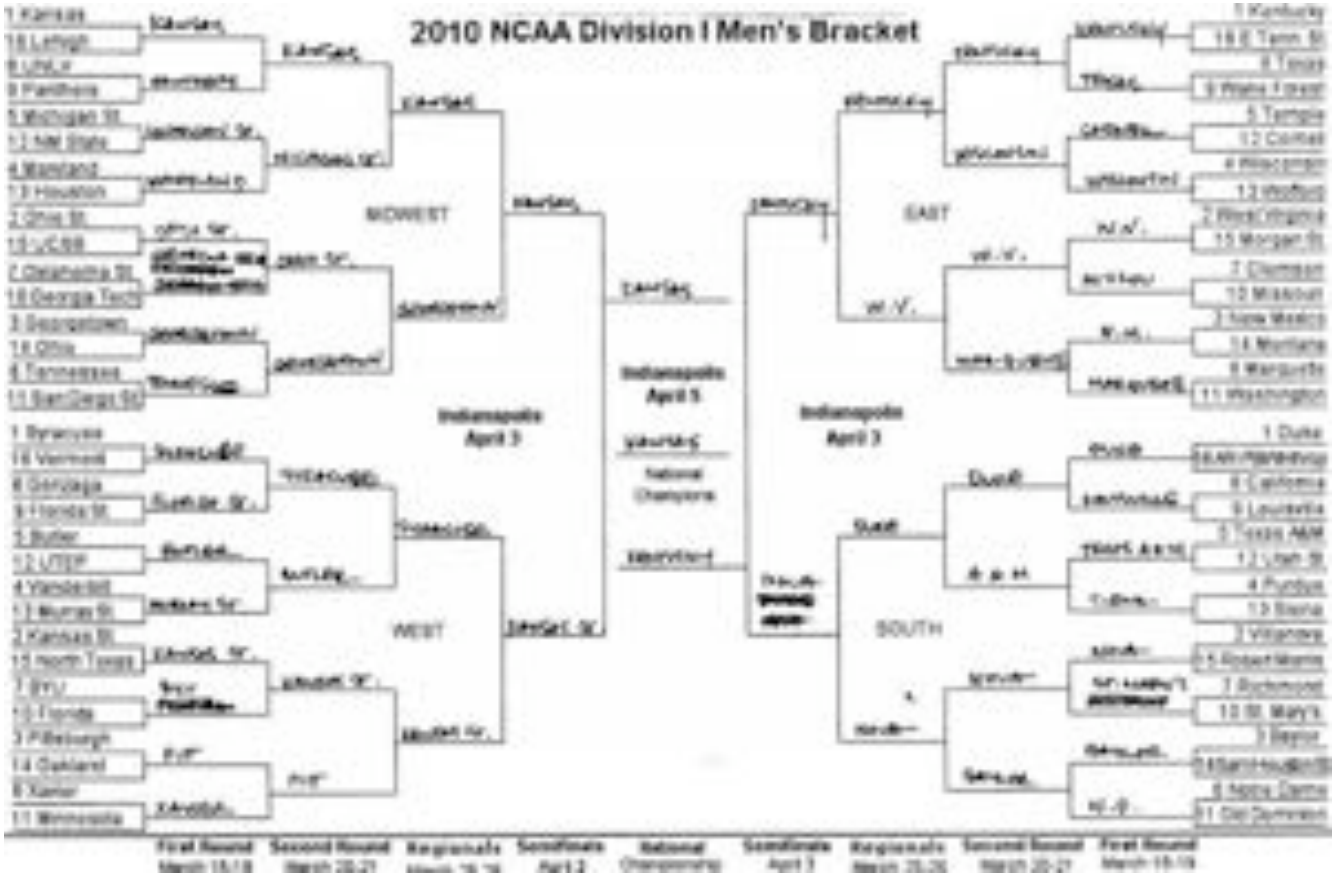
- After construction, many edges are unambiguous
  - Merge together compressible nodes
  - Graph physically distributed over hundreds of computers





Warmup Exercise

- Who here was born closest to September 15?
  - You can only compare to 1 other person at a time



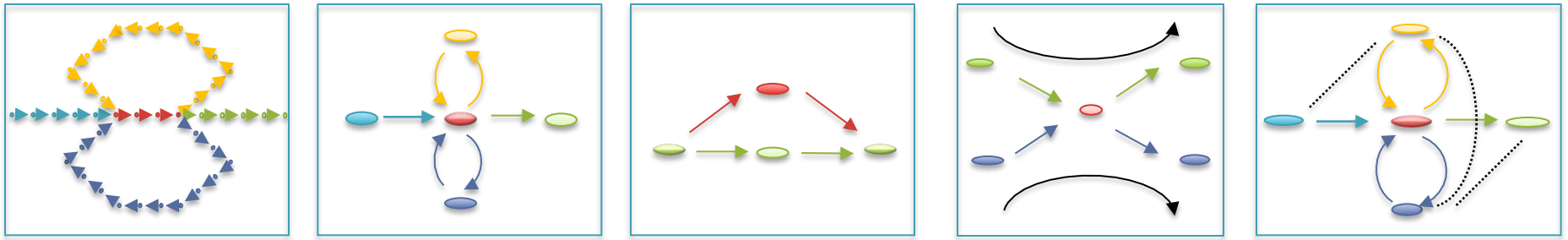
Find winner among 64 teams in just 6 rounds

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	
Max	27 bp	303 bp	20,594 bp	20,594 bp	In progress
N50	27 bp	< 100 bp	995 bp	1,050 bp	

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

Research Directions

- How do we survive the tsunami of sequences?
- How do we gain insight?
  - Unique opportunities at CSHL to partner wet & dry research
- Applications
  - Recurrent de novo mutations in autism spectrum disorder through exome sequencing
  - Genetic origins of ovarian cancer
  - Tracing tumor development through microsatellite mutations (Mitch Bekritsky)
  - De novo assembly of bacteria, plants, animals, fungi
  - …
- There is more exciting work than hours in the day
  - I need your help!
  - I will guide you to become an expert sequence analyst & data detective

Thank You!

http://schatzlab.cshl.edu