Scalable Solutions for DNA Sequence Analysis

Michael Schatz

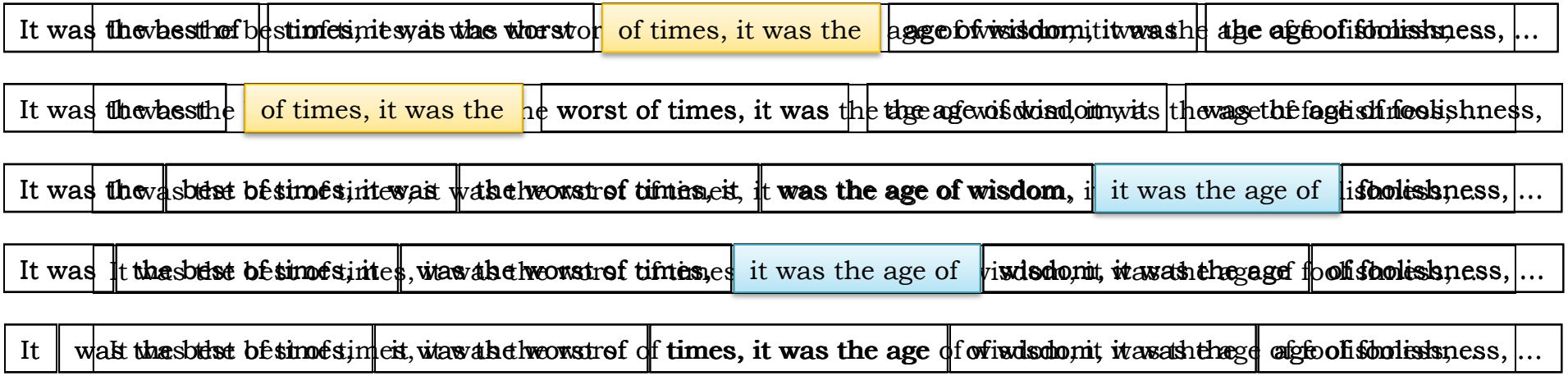
December 15, 2009

Hadoop User Group

Shredded Book Reconstruction

Dickens accidently shreds the only 5 copies of A Tale of Two Cities

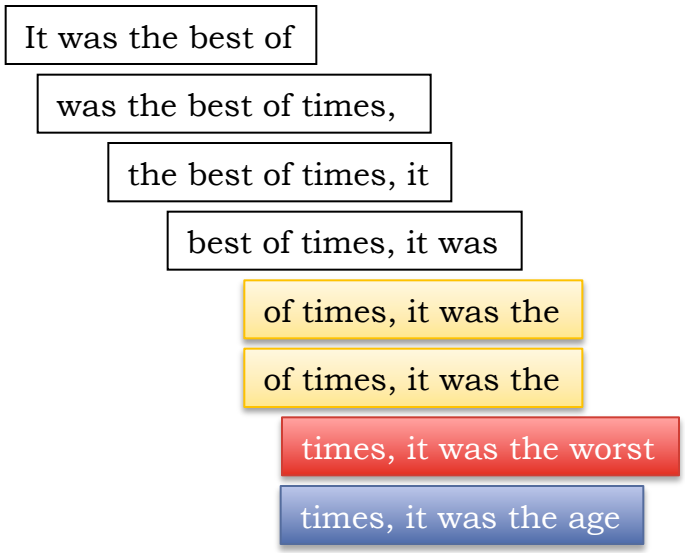
- Text printed on long spools



- 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



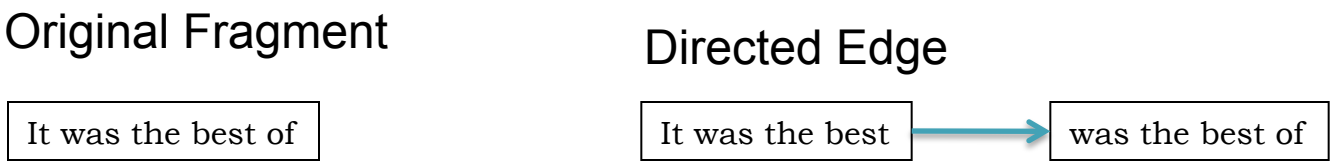
The repeated sequence makes the correct reconstruction ambiguous

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

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

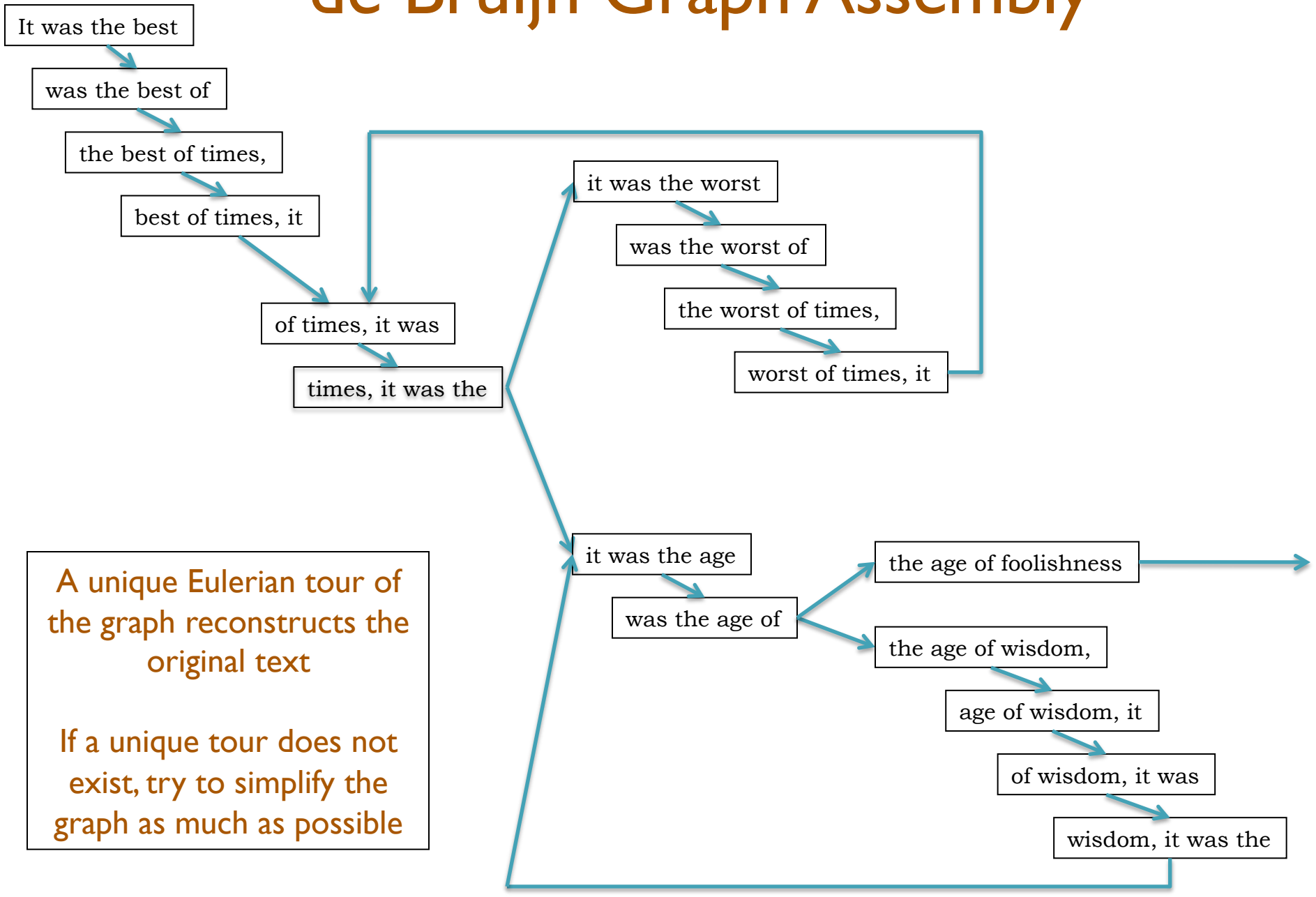
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

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

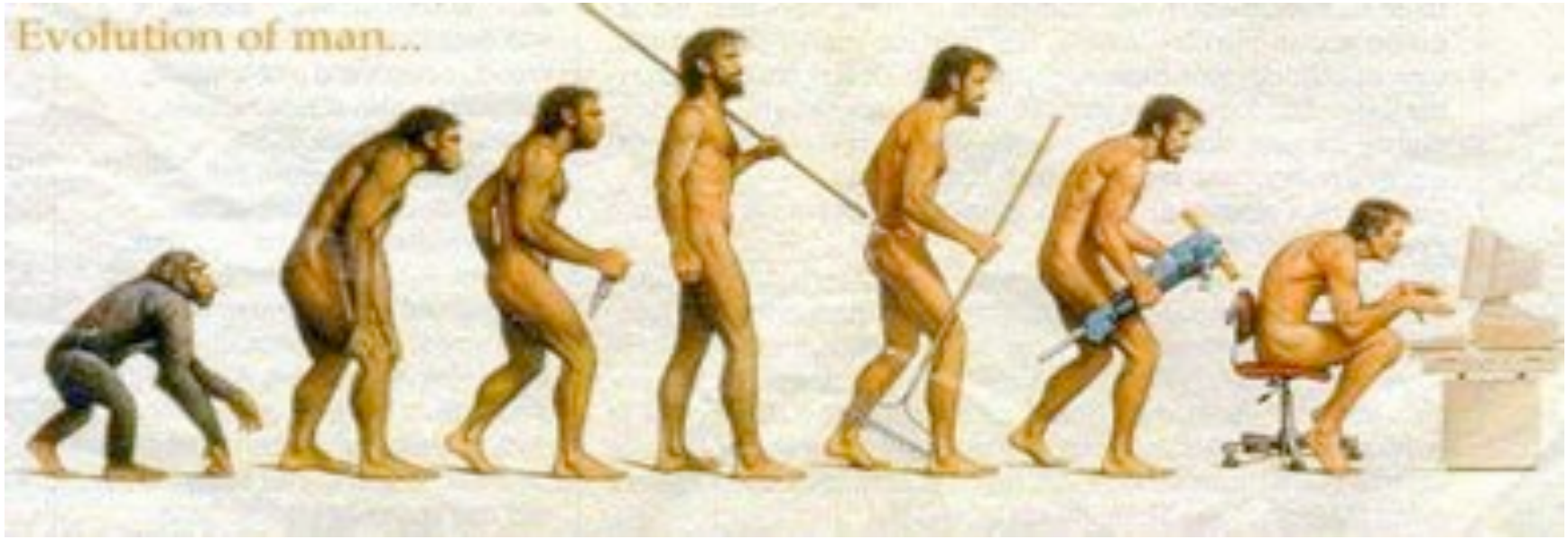
de Bruijn Graph Assembly



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

Genomics



Your genome influences (almost) all aspects of your life

- Anatomy & Physiology: 10 fingers & 10 toes, organs, neurons
- Diseases: Sickle Cell Anemia, Down Syndrome, Cancer
- Psychological: Intelligence, Personality, Bad Driving

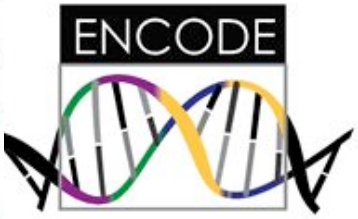
Your environment also plays a big role

- Recipe, not a blueprint

The Evolution of DNA Sequencing

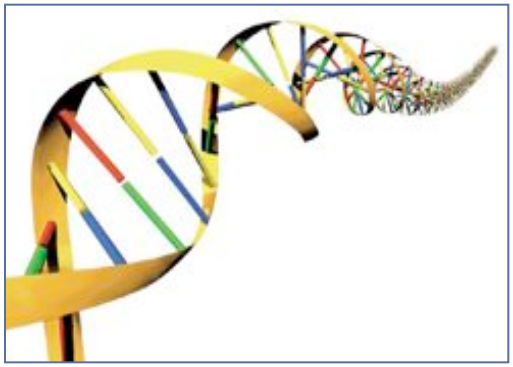
Year	Genome	Technology	Cost
2001	Venter *et al.*	Sanger (ABI)	$300,000,000
2007	Levy *et al.*	Sanger (ABI)	$10,000,000
2008	Wheeler *et al.*	Roche (454)	$2,000,000
2008	Ley *et al.*	Illumina	$1,000,000
2008	Bentley *et al.*	Illumina	$250,000
2009	Pushkarev *et al.*	Helicos	$48,000
2009	Drmanac *et al.*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

(Pushkarev *et al.*, 2009)



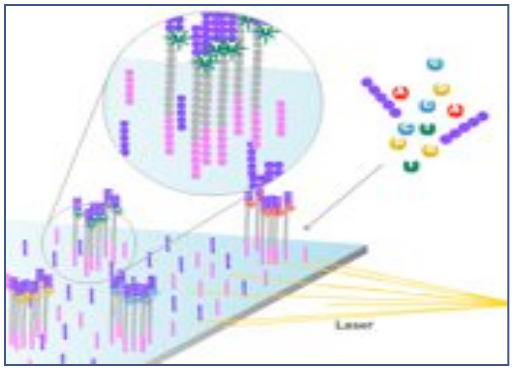
Critical Computational Challenges: Alignment and Assembly of Huge Datasets

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 generate 1-2 Gbp of sequence per day, in millions of short reads

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

ATCTGATAAGTCCCAGGACTTCAGT

GCAAGGCAAACCCGAGCCCAGTTT

TCCAGTTCTAGAGTTTCACATGATC

GGAGTTAGTAAAAGTCCACATTGAG

Recent studies of entire human genomes analyzed 3.3B (Wang, et al., 2008) & 4.0B (Bentley, et al., 2008) 36bp reads

- ~100 GB of compressed sequence data

DNA Resequencing

CloudBurst

Highly Sensitive Read Mapping
with MapReduce

(Schatz, 2009)

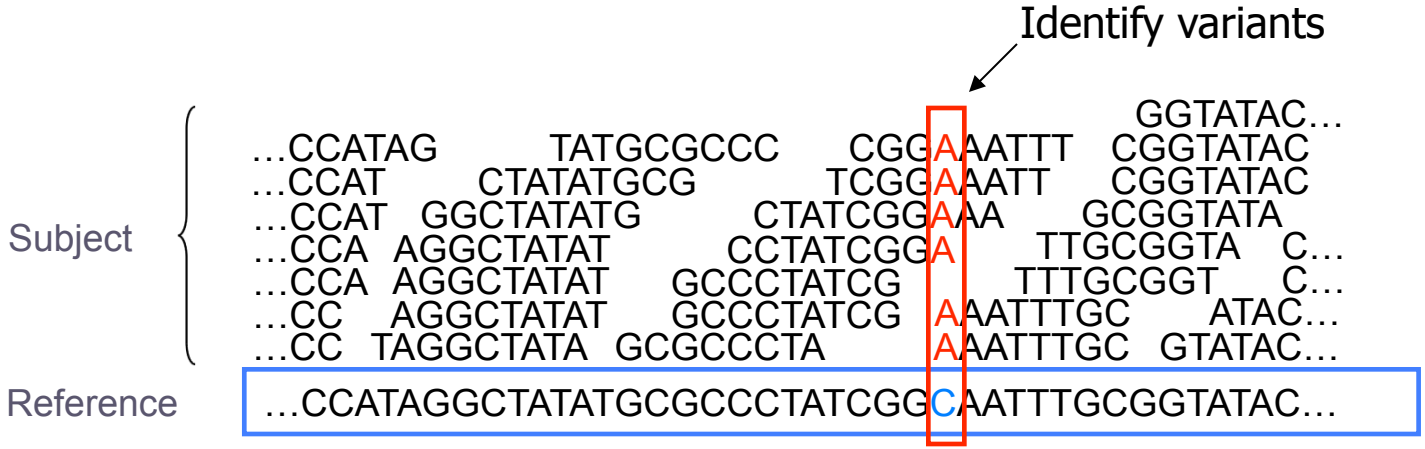
Parallel Distributed Hashing
100x speedup on 96 cores at EC2
http://cloudburst-bio.sf.net

Crossbow

Searching for SNPs
with Cloud Computing

(Langmead, Schatz, Lin, Pop, Salzberg, 2009)

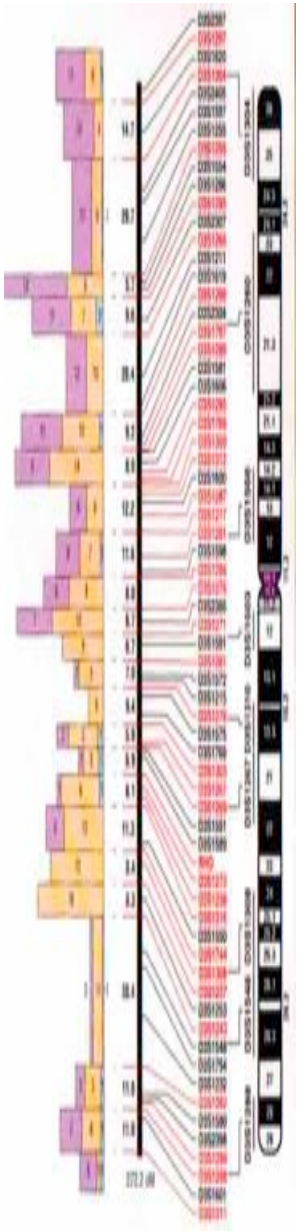
Scaling up mapping and genotyping
Reads to SNPs for <$100 in <3 hours
http://bowtie-bio.sf.net/crossbow



De novo assembly with MapReduce

Problem

- Current assemblers require tremendous computation
- Human genome requires TBs of RAM and many CPU years



Advantages

- Proven system for processing huge datasets
  - PageRank: Significance in web graph of >1 trillion pages
  - CloudBurst & Crossbow: Genome mapping
- Simple programming model
  - Reliability, redundancy, scalability built-in

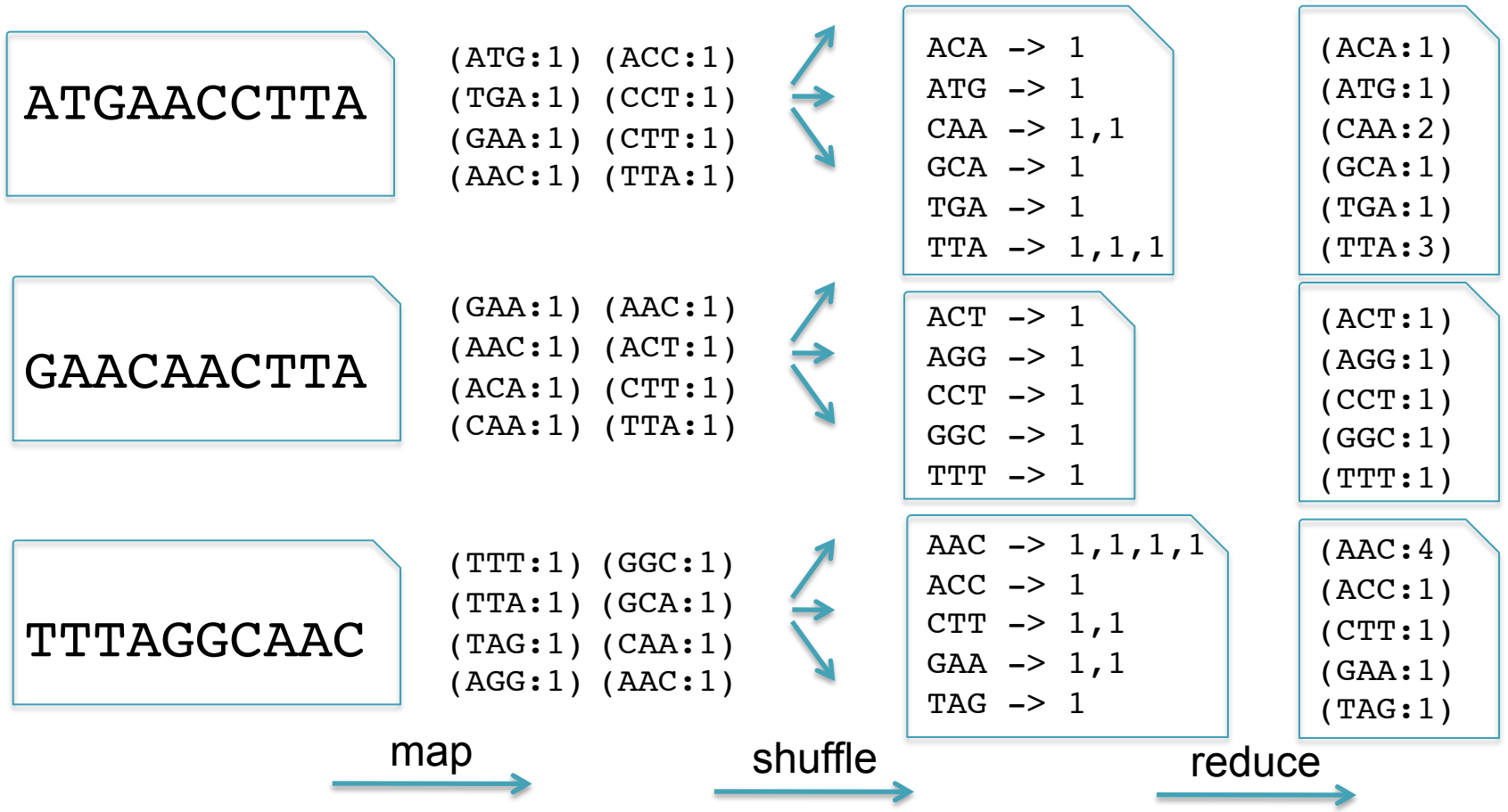
Challenges

- How to (efficiently) implement assembly graph algorithms?
  - Restricted programming model (not MPI, not shared memory)
  - Adjacent nodes may be stored on different machines

K-mer Counting

- Application developers focus on 2 (+1 internal) functions
  - Map: input ➔ key:value pairs
  - Shuffle: Group together pairs with same key
  - Reduce: key, value-lists ➔ output

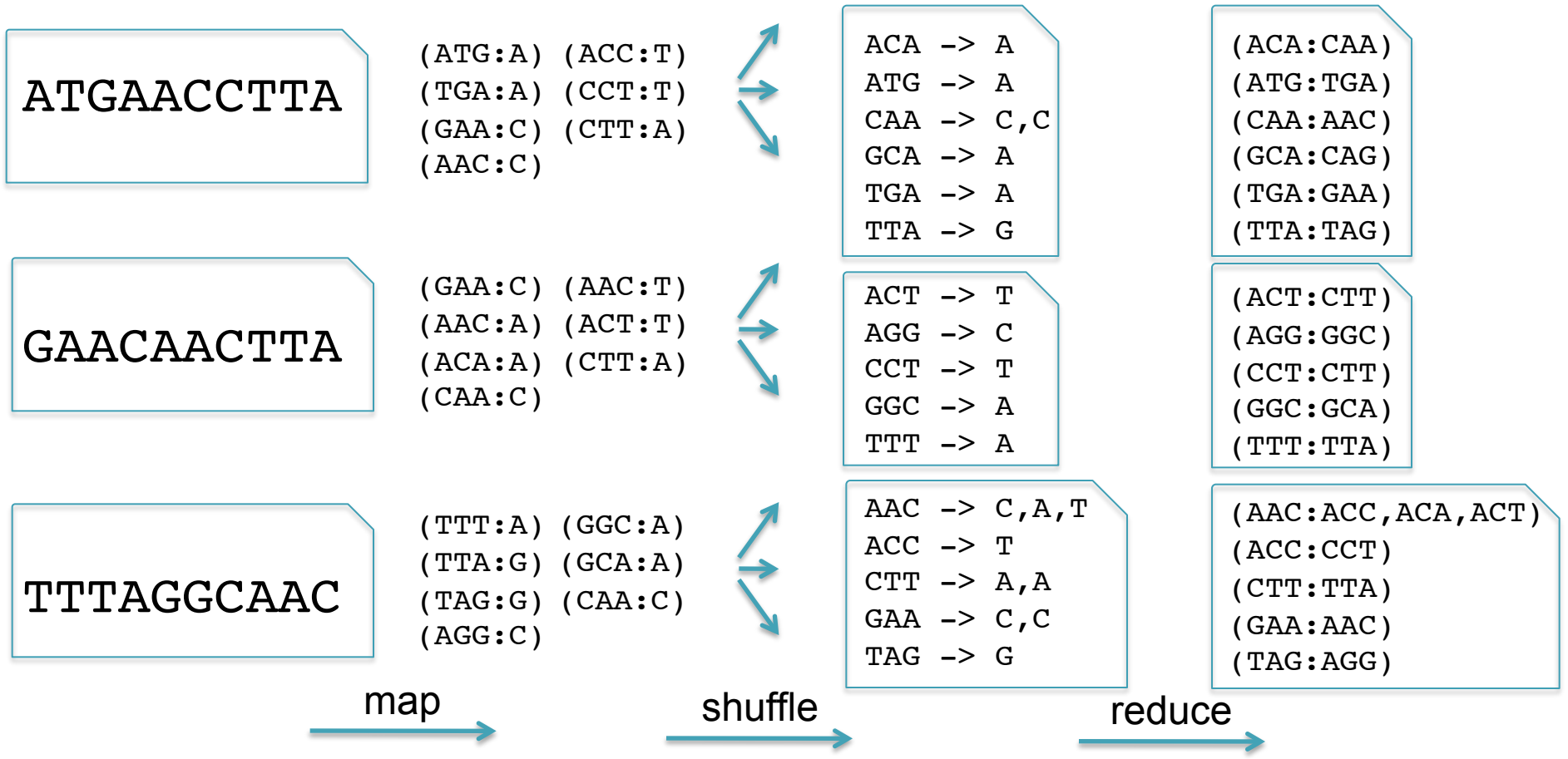
Map, Shuffle & Reduce
All Run in Parallel



Graph Construction

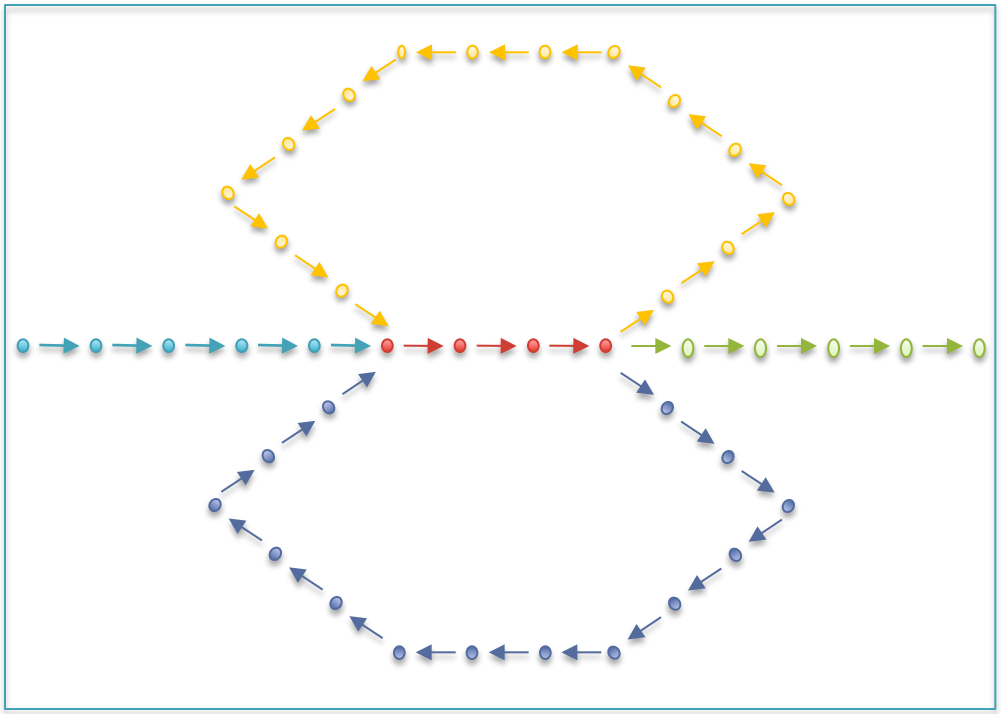
- Application developers focus on 2 (+1 internal) functions
  - Map: input ➔ key:value pairs
  - Shuffle: Group together pairs with same key
  - Reduce: key, value-lists ➔ output

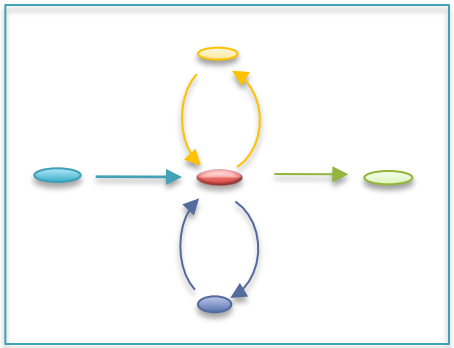
Map, Shuffle & Reduce
All Run in Parallel



Graph Compression

- After construction, many edges are unambiguous
  - Merge together compressible nodes





Find Compressible Nodes

Input: Graph stored as (n : (nodeinfo, ni))

Map:

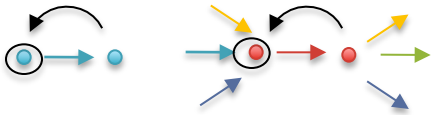
- For all nodes, emit (n : (nodeinfo, ni))
- If node n has unique predecessor p, emit (p : (unique-pred, n))

MapReduce Message Passing

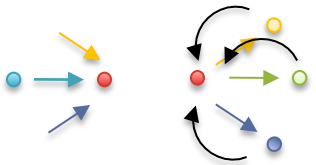
Reduce:

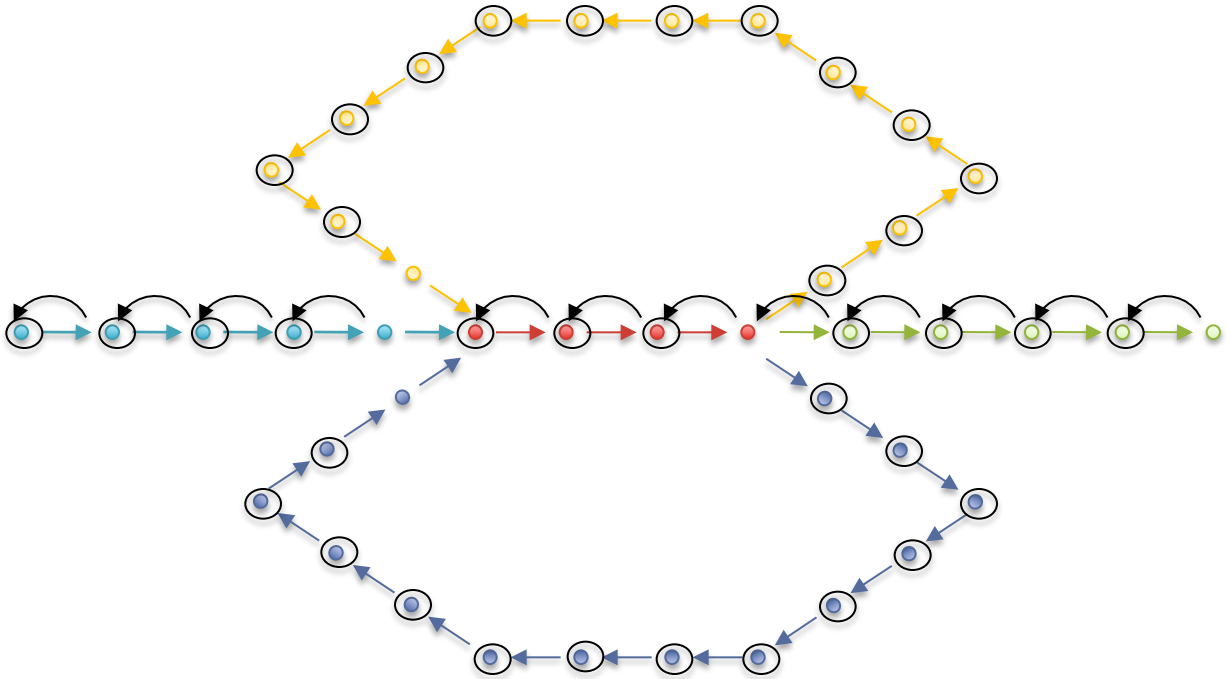
- If node n has unique successor s, and received (unique-pred, s),
  - Mark ni as compressible
- Save (n : (nodeinfo, ni))

Compressible



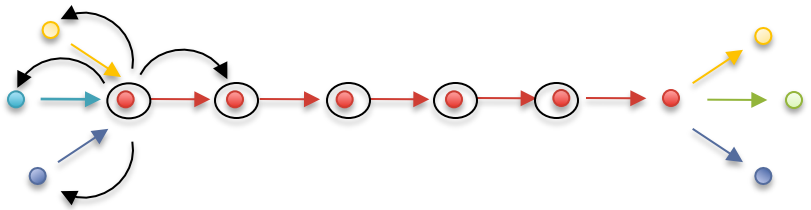
Not Compressible





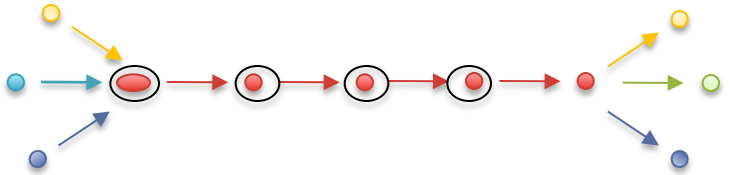
Linear Path Compression

Iteratively identify and collapse the beginning of each chain



Map:

- Emit messages to the neighbors of the head of each chain



Reduce:

- Update links, node label

Requires S MapReduce cycles, where S is the length of the longest simple path

- *B. anthracis*: L=5.2Mbp S=268,925 bp
- *H. sapiens* chr 22: L=49.6Mbp S=33,832 bp
- *H. sapiens* chr 1: L=247.2Mbp S=37,172 bp

Fast Path Compression

Challenges

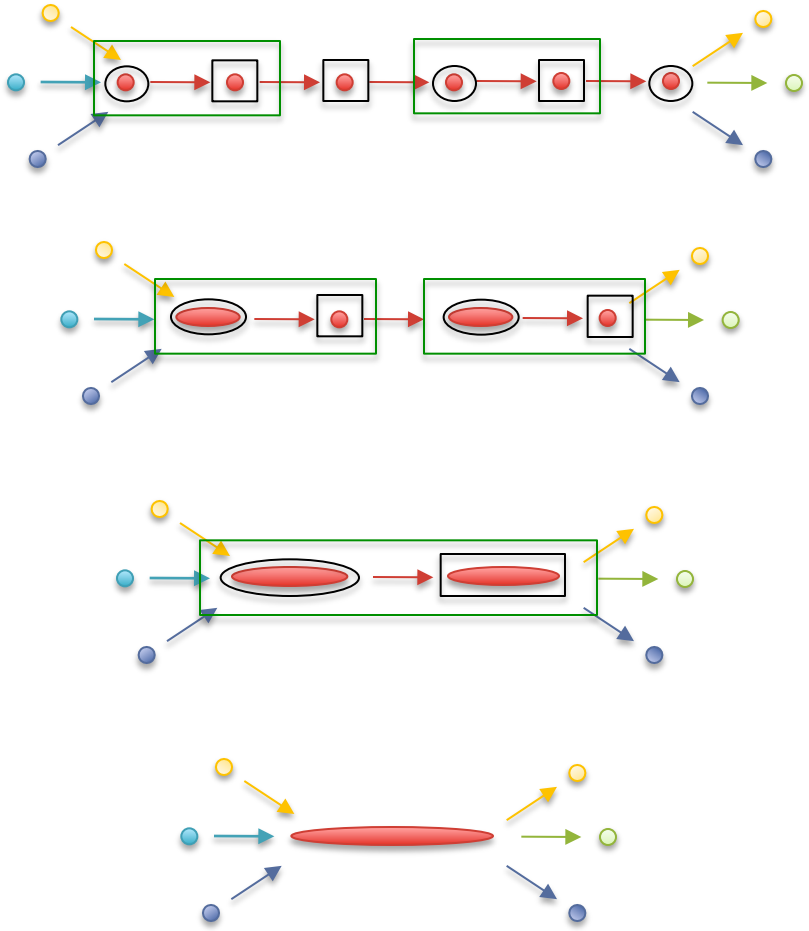
- Nodes stored on different computers
- Node only knows immediate neighbors

Randomized List Ranking

- Randomly assign Ⓗ / [T] to each compressible node
- Compress Ⓗ -> [T] links
- (Vishkin, 1984)

Optimizations

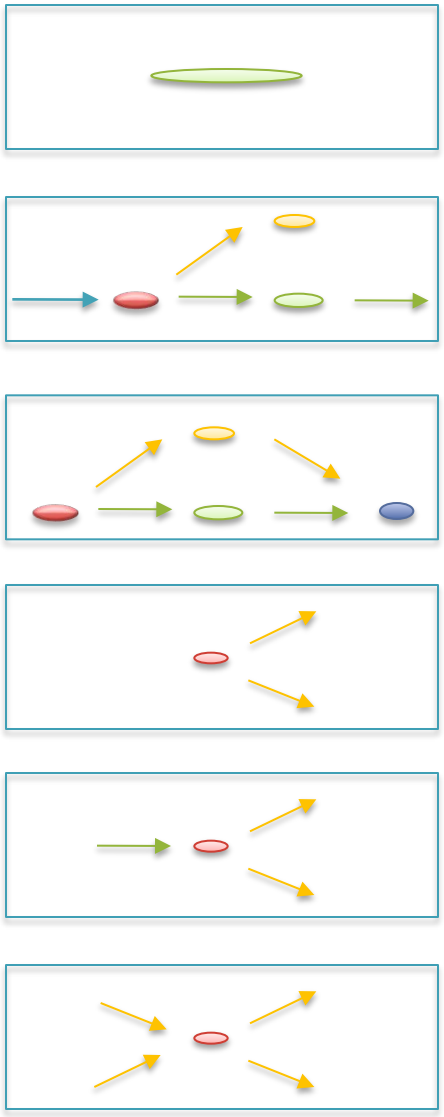
- Always compress ends of chains
- If <1000 nodes to compress, send them all to the same reducer



Parallel Randomized List Ranking

Node Types

(Chaisson, 2009)



Isolated nodes (10%)
- Contamination

Tips (46%)
- Clip short tips

Bubbles/Non-branch (9%)
- Pop bubbles

Dead Ends (.2%)
- Split forks

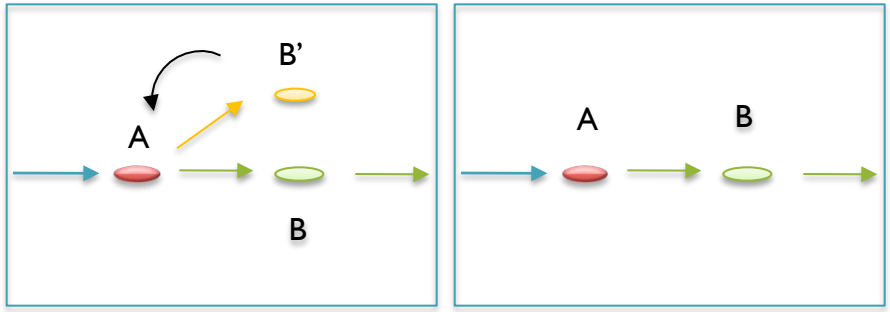
Half Branch (25%)
- Unzip

Full Branch (10%)
- Thread reads, cloud surfing

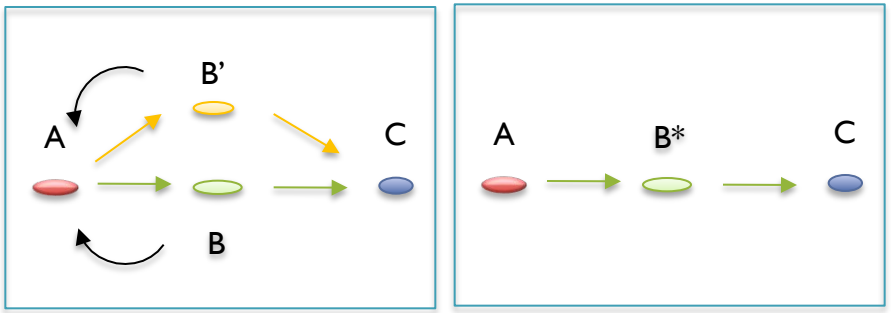
Error Correction

Sequencing error distorts graph structure

- Errors at end of read
  - Trim off 'dead-end' tips
  - B' passes *trim* message to A



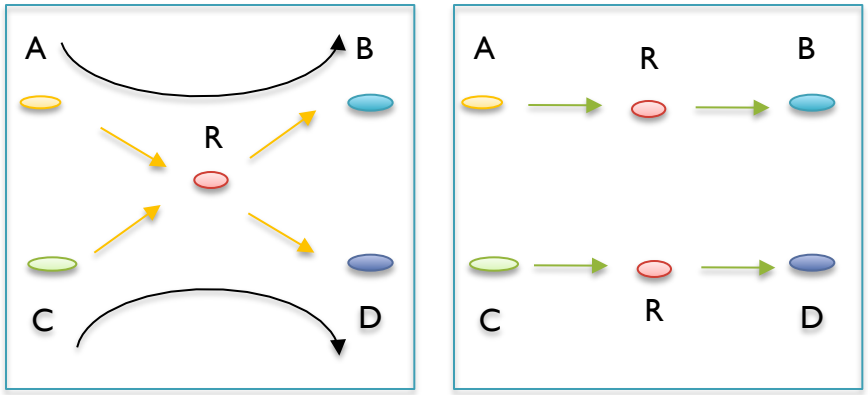
- Errors in middle of read
  - Pop Bubbles
  - B' and B pass *bubble* messages to A
    - A is lexicographically smaller than C



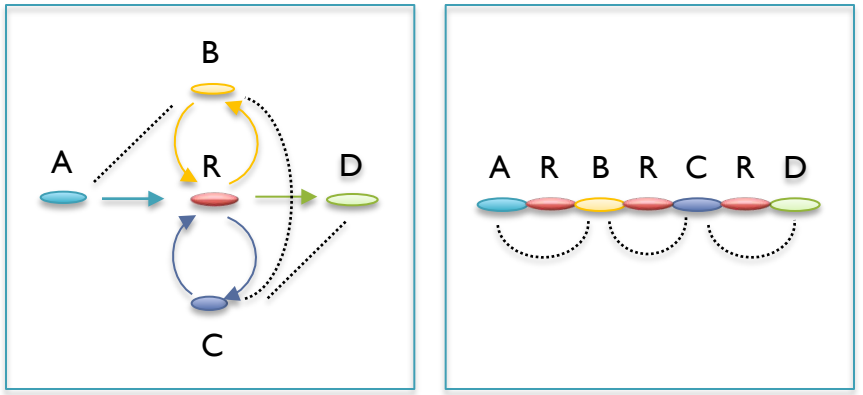
- Recursively apply, rerun path compression between each iteration

Graph Simplifications

- X-cut
  - Annotate edges with spanning reads
  - Separate fully spanned nodes
  - (Pevzner *et al.*, 2001)



- Scaffolding
  - If mate pairs are available search for a path consistent with mate distance
  - Use message passing to iteratively collect linked and neighboring nodes



- Other simplifications possible

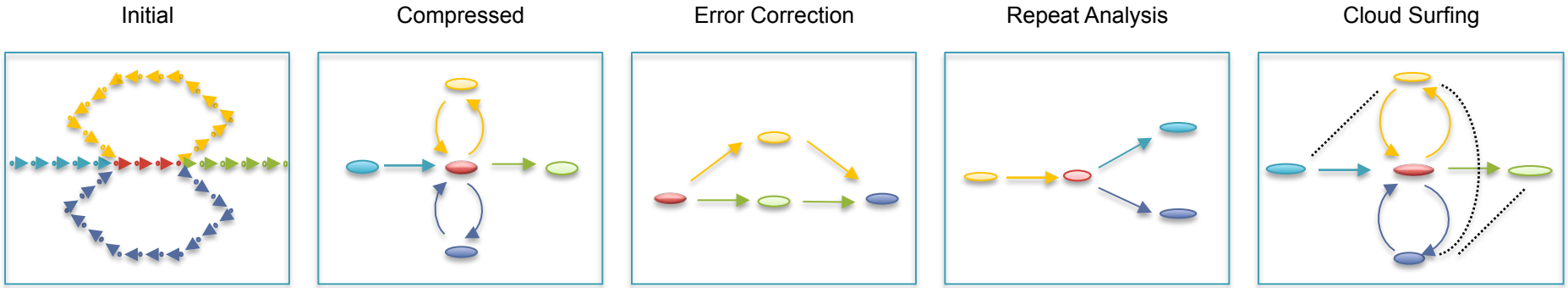
Parallel Frontier Search

Contrail

http://contrail-bio.sourceforge.net

Scalable Genome Assembly with MapReduce

- *Genome:* 4.6Mbp bacteria
- *Input: 4*M 36bp reads, 200bp insert
- *Coverage*: 31x



Results coming soon

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

Summary

1. Scaling up for the tidal wave of NextGen sequence data is a central challenge in biology

2. Hadoop & MapReduce may be the enabling technologies to stay afloat

3. Graph algorithms are challenging– PRAM algorithms may apply

Acknowledgements

Advisor

Steven Salzberg

UMD Faculty

Mihai Pop, Art Delcher, Amitabh Varshney,
Carl Kingsford, Ben Shneiderman,
James Yorke, Jimmy Lin, Dan Sommer

CBCB Students

Adam Phillippy, Cole Trapnell,
Saket Navlakha, Ben Langmead,
James White, David Kelley

Thank You!

http://www.cbcb.umd.edu/~mschatz
@mike_schatz