Next Generation Bioinformatics Tools

Fall 2012
Day 4 – Sequencing and Assembly

Hesham H. Ali
UNO Bioinformatics Research Group
College of Information Science and Technology

UNIVERSITY OF NEBRASKA AT OMAHA

UNO BIOINFORMATICS

Sequencing, Assembly and Genome-Wide Association Studies

- History of Sequencing Technologies
- Sequencing and Bioinformatics: The Computing analogy
- Key phases in ST: The Sequencing Generations
- From Microarray technology to whole genome studies
- Assembly
- Annotation
- What to do after assembly
  - Annotation
  - Genome-wide association
- Next Steps: Translational Bioinformatics

UNO BIOINFORMATICS

History of Assembly

UNO BIOINFORMATICS

Biomedical Research Timeline

1800s

Joseph Lister
- Antiseptics

Lewis Pasteur
- Germ Theory

Gregor Mendel
- Mendelian Inheritance

1900 – 1950

Frederick Griffith
- Transforming Factor

Avery, MacLeod, & McCarty
- DNA Discovery

1950 – 1990

Watson & Crick
- DNA Structure

Frederick Sanger
- DNA Sequencing

1990 - 2000

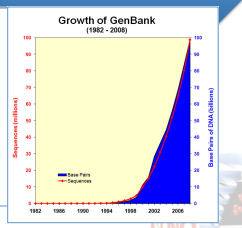
Genome Projects
- *H. influenzae*
- *C. elegans*
- *A. thaliana*
- Human

Bioinformatics

Biomedical Research Timeline

2000 - current

- Cancer Research
- Infectious Diseases
- Genetic Diseases
- Personalized Medicine
- Metagenomics
- Agriculture
- Future ???

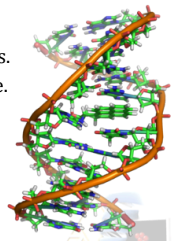


Biological Background

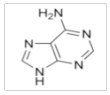
- An overview of the genome
- Why sequencing?
- Decoding the genome
- Whole genome sequencing
- Next generation sequencing

The Genome

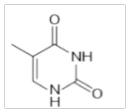
- Genome: Blueprint of all living things.
- **DNA** (deoxyribonucleic acid) is the genetic material of all living organisms.
- DNA consists of a four nucleotide code.
  - **Adenine** (A)
  - **Thymine** (T)
  - **Cytosine** (C)
  - **Guanine** (G)



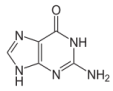
The Genome



Adenine (A)

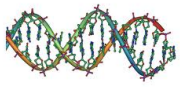


Thymine (T)



Guanine (G)

Cytosine (C)



DNA

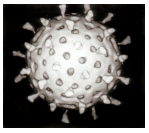
Contains Instructions

Transcription, Translation

Protein

Why Sequence Genomes?

- Infectious disease research
- Metagenomics
- Personalized medicine
- Cancer research
- Numerous other applications



Infectious Diseases

Various strains of HIV are now aggressively targeted by specifically tailored treatments

Why Sequence Genomes?

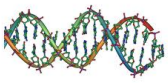
- Infectious disease research
- Metagenomics
- Personalized medicine
- Cancer research
- Numerous other applications

Environmental Genomics

Novel microorganisms were collected and sequenced in the global ocean sampling expedition (Craig Venter Institute).

Why Sequence Genomes?

- Infectious disease research
- Metagenomics
- Personalized medicine
- Cancer research
- Numerous other applications



Personalized Medicine

In the near future, medicine will be able to be personalized according to an individual's genetic makeup, resulting in better treatment response and fewer treatment side affects.

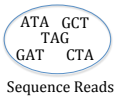
Decoding the Genome

Genomic Sequence
... AGCTAGCTGATAGCTAGCGCTACGATCGCGT...

Sequencing Technology

Chemistry
- 30-1000 bps
- Error rate
- Various technologies

Data



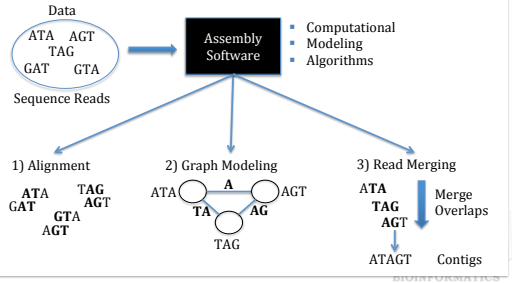
Sequence Reads

Sequence Assembly

Information

Computers
- Software
- Modeling
- Algorithms

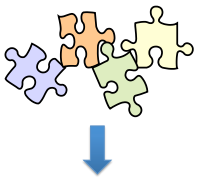
Sequence Assembly



Decoding the Genome

DNA Fragments

```
GGATCCATGGATAGGATAATGGA
GATAATGGATAGAGGATCCATGG
ATGGATAGAGGATCCATGGCTAG
ATGGATAGGATTATGGATAGAGG
GATAGAGGATCCATGGCTAGATC
```



Contig

```
GGATCCATGGATAGGATAATGGATAGAGGATCCATGGCTAGATC
GGATCCATGGATAGGATAATGGA
              GATAATGGATAGAGGATCCATGG
                  ATGGATAGAGGATCCATGGCTAG
     ATGGATAGGATTATGGATAGAGG
                    GATAGAGGATCCATGGCTAGATC
```

History of Sequencing

- 1953: Watson & Crick discover DNA helix
- 1968: First DNA sequencing
- 1974-1977: Modern day sequencing
  - Maxam & Gilbert
  - Sanger, Nicklen, Coulsen- dideoxy method
    - First complete seq: phage ΦX174
- 1992: First sequencing 'factory'
- 1995: First complete bacterial sequences
- 2001: Human Genome published
- 2005: Sequence data at NCBI passes 100Gb mark

Early Challenges of DNA Sequencing

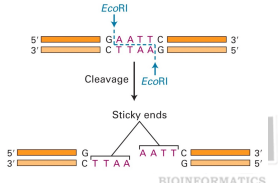
- Chemical properties of ACTG very similar
- Chain length of DNA unusually long
  - Compared to proteins
- Four bases total compared to 20 for proteins
- No base-specific DNAases known
  - Proteins have proteases

1959: First Genome Purification

- Bacteriophage ΦX174
- Purified to homogenity by Sinsheimer in 1959
- Single stranded circular molecule
- ~5000nt length

1970: Discovery of type II restriction enzymes

- Recognize and cleave DNA at short nucleotide sequences (palindromes)
- 4-6bp in length
- Provided a method for cutting large DNA sequences down into pieces
- Pieces separated by gel electrophoresis
- Endonuclease IV from *E. coli*
  - 10-20bp fragments



Sequencing Generations

Illumina Genome Analyzer IIx

Sanger sequencing

- Older platform 70s
- 400-600 bp reads
- 6 to 10x Coverage

Next generation sequencing

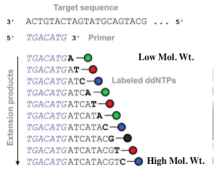
- Newer platform
- Introduced in 2005
- Illumina: 35 – 150 bp reads[4]
- Roche: up to 1000 bp reads[5]
- ABI SOLiD: 35 – 75 bp reads[6]
- Shorter read lengths
- Very high throughput

First Generation Sequencing

Sanger Sequencing

- Millions of copies of DNA purified and amplified
- Reverse strand synthesis is performed by adding dNTPs and small fraction of ddNTPs
- Labeled ddNTPS (A, G ,C and T) terminate the extension
- The resulting molecules are sorted by molecular weight
- Terminating ddNTPS are read sequentially

TGACATGATCATACGTC

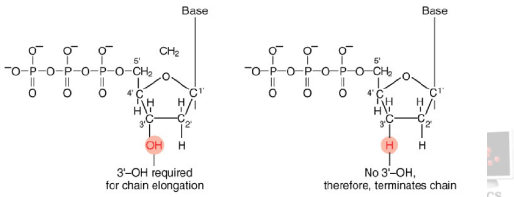


Sanger Technology

- Technically 384 sequences sequenced in parallel
- Read length = 600 to 1000nt
- Usually 96 sequences done in parallel
  - 6Mb of DNA sequence per day
  - $500 per 1Mb
- Errors:
  - amplification error, contamination, polymerage slippage at repeats, homopolymers (stretch of same nucleotide)
  - Base miscalls
- Average error is low (1 per 10,000 -100,000 nt)

1974: Sanger Dideoxy Method

- Shared Nobel Prize with Maxam & Gilbert
  - Maxam & Gilbert: cleavage method
  - Sanger method more practical
- Based on dideoxynucleotides (ddNTPs)
  - ddNTPs are same as nucleotides
  - Contain extra hydrogen group on 3' carbon instead of OH
  - DNA replication halts at a ddNTP



1974: Sanger Dideoxy Method

- Denature DNA strands with heat
- Anneal primer to template strands
  - Primer made so 3' end next to DNA sequence of interest
  - Primer labeled so it can be ID'ed on a gel
  - Mixture separated into four tubes and the following are added:

"G" tube: all four dNTP's, ddGTP and DNA polymerase

"A" tube: all four dNTP's, ddATP and DNA polymerase

"T" tube: all four dNTP's, ddTTP and DNA polymerase

"C" tube: all four dNTP's, ddCTP and DNA polymerase

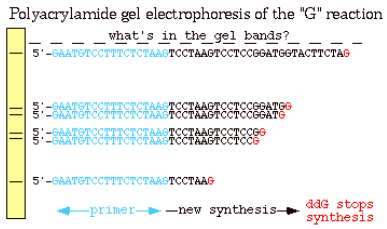
1974: Sanger Dideoxy Method

- With addition of polymerase, DNA replication begins
- ddNTPs are randomly integrated into newly made DNA

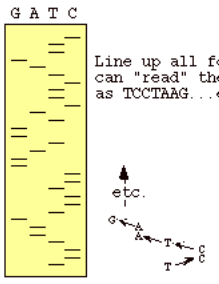
```
     5'-GAATGTCCTTTCTCTAAGTCCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
```


1974: Sanger Dideoxy Method

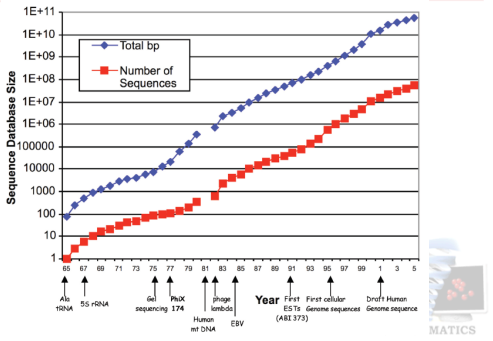
- Multiple length fragments made
- DNA denatured for electrophoresis



1974: Sanger Dideoxy Method



Evolution of Sequencing



1977: Birth of Bioinformatics

- Michael Smith – member of the Sanger group
  - Brother-in-law Duncan McCallum
  - Wrote algorithms in COBOL to compile and analyze DNA sequence data
    - Compiled and numbered the complete sequence
    - Allowed for editing of a previously compiled sequence
    - Searched for known restriction sites
    - Translated the sequence in all reading frames
- 1982: GenBank created by NIH as genetic repository

Second Generation Sequencing

454 FLX Titanium Sequencer (Roche)

- First high-throughput sequencing platform (October 2005)
- By pyrosequencing process complementary base is added iteratively
  - When there is an interaction of a dNTP and the base in the template, light signal is emitted and read
  - Sequencing during extension (nucleotide is read as it gets extended)

454 FLX Titanium sequencer

- Average read length = 400 nt
- 750 Mb of DNA per day
- $20/Mb

- Errors:
  - like in Sanger, amplification error in prep step
  - mostly indels and homopolymers

Illumina Genome Analyzer

- Like Sanger sequencing, once a base is added reaction is stopped
- The fluorescence is measured (base call) and the flurophore is removed
- Sequencing reaction is continued to add next base and the process is iterative.

Illumina Genome Analyzer

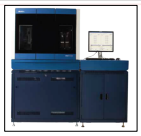
- Read length = 100nt
- 5,000Mb/day
- $0.50 /Mb

- Errors
  - During preparation, amplification errors
  - Image analysis may identify dust and lint particles
  - Higher substitution error between A/C and G/T
    - A/C and G/T have similar emission spectra

ABISOLiD

- Developed by Harvard Medical school and Howard Hughes Medical institute in 2005
- Cheaper than other high throughput system (< $200,000)
- SOLiD = sequencing by oligonucleotide ligation and detection
  - Sequencing is different than previous extension methods
  - A library of 8mer probes carrying different fluorescent labels ligate(join) to primer and is read out
  - Iterative cycle of ligation and detection

ABISOLiD



- Read Length between 25 to 75nt
- 5000Mb/day
- $0.5/Mb

- Works well for problematic homo-polymers
- Error
  - Like in previous technologies, in vitro amplification errors, library prep errors
  - In the absence of reference genome, error detection and correction can not be applied

Types of Sequence Reads

- Single end (In the past)
  - Reads are from one end of the DNA fragments
  - One fasta/fastq file with reads
- Paired end (At present)
  - reads are sequenced from both ends of a fragment
  - distance between two ends is called insert size
  - two separate fasta/fastq files, each end in separate file
  - double amount of data
  - gives better assembly as the position information is known between the pair of reads

Third Generation Sequencing

Sequencing versus Bioinformatics

- Computers versus Computing
- Direction of Bioinformatics has been influenced by sequencing technologies and models
- Third Generation Sequencing:
  - Going big
  - Going small

Third Generation Sequencing

- High throughput sequencing plus personal less expensive sequencing
- Analogy to Super computing and personal computing – Mobile computing
- New problems:
  - How to automate the process for the super sequencers
  - How to deal with regions with low coverage
  - How to distinguish errors versus variants
- The analysis part: how to identify/classify mutations Drive mutations and passenger mutations

Two Directions

- Very high-throughput sequencers
  - Longer reads
  - Lower percentage of sequencing errors
  - More expensive
  - Multi organisms sequencing
- Small "personal" sequencers
  - Small and inexpensive
  - Higher ratio of errors
  - Higher degree of availability

From Genome Sequencing to Clinical Diagnosis

- It is here – the possibility of using NGS to narrowing options in clinical diagnosis
- Science Translational Medicine: Cover Story
  - Fetal Genome in mother's blood with SWAS revealing genetic variants and related phenotypes
- Exome analysis will be as common as BLAST search
  - Exome sequencing (targeted exome capture) is an efficient strategy to selectively sequence the coding regions of the genome to identify novel mutations associated with rare and common disorders.
- Targeted therapy based on genomic alterations
- Transcriptome analysis by sequencing

Comparison among Sequence generations

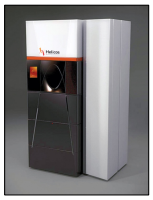
- Expression level
- False positive rate
- Alternative Splicing
- Mutation Detection
- Non Coding DNA
- Cost

Factors to consider for Sequencing

- What to Sequence
  - Something of interest
- Right time to sequence
- Which model to use
- Which team to get involved

Newer Technologies

Helicos Heliscope
- single molecule sequencing
- Least affected by prep and amplification errors
- Highly expensive



PacBIO
- Single Molecule sequencing
- Longer reads
  - 3000bp to 10000bp

Ion Torrent
- Semiconductor based sequencing
  - highly parallel
- Very less run time

What to do with Sequenced Data

- Assembly
  - Annotation
  - Genome wide association study
  - Comparative genomics
- Transcriptome study
  - RNA-Seq to find expression level of mRNAs

Assembly

Survey of Assembly Algorithms

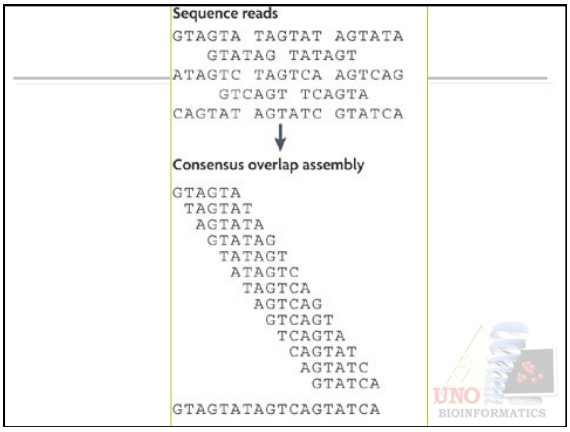
- Graph theoretic sequence assembly
- Overlap graph assembly
  - Cap3
  - Cabog
- De Bruijn graph assembly
  - Velvet
  - Abyss

Genome Assembly

- Types of assemblers
  - Assembly paradigms
    - Overlap-layout-consensus
    - De Bruijn Graphs
  - De novo vs reference based
  - Sequencing types
    - Single end vs Paired end

Assembly Paradigm 1

- Overlap, Layout and Consensus
  - Reads = nodes
  - Overlap between reads = edges
  - Layout = Finding Hamiltonian Path
  - Consensus = Multiple sequence Alignment
  - Examples:
    - Newbler, Celera, Edena



Assembly Paradigm 2

- De Bruijn Graph
  - Series of overlapping kmers = Node
  - K-1 bp overlap between last kmer of previous node and first kmer of current node = Edge
  - Find Eulerian Path
  - Example
    - ABySS, Velvet, Soap deNovo

The reads are the edges

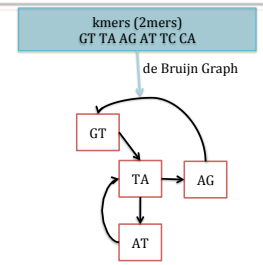
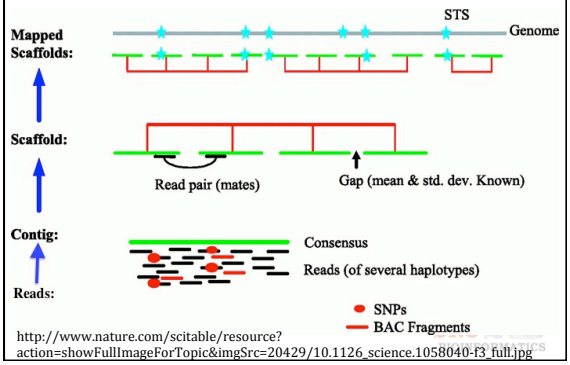


Figure from: Application of 'next-generation' sequencing technologies to microbial genetics
Daniel MacLean, Jonathan D. G. Jones & David J. Studholme

De novo vesus Reference Based

- In reference based assembly, all the reads are aligned to the reference and overlap and consensus is generated from the alignment
  - Example: AMOScmp, 454gsMapper
- De novo assemblers find overlap with in the reads and join and build consensus
  - Example: ABySS, Velvet, Celera

Assembly Hierarchy



http://www.nature.com/scitable/resource?action=showFullImageForTopic&imgSrc=20429/10.1126_science.1058040-f3_full.jpg

Assembly Challenges

Sequencing errors

```
        ATTAGGTAGGTTTGAT
TTACATTAGGCCG
```

Small overlaps

```
          ATTAGTTAGTTAGATTAC
...GGCATTA
```

Repetitive regions of the genome

Gap closure

Velvet Assembler (de Bruijn Graph)

Three stage assembly

- Graph construction
  - K-mers and reads are hashed.
  - Reads are indexed throughout the de Bruijn graph
- Graph simplification and error correction
  - Merge consecutive chains of k-mers
  - Trim branches and bubbles in the graph
- Repeat resolution
  - Read pairs are used to resolve repeats

Velvet (de Bruijn Graph)

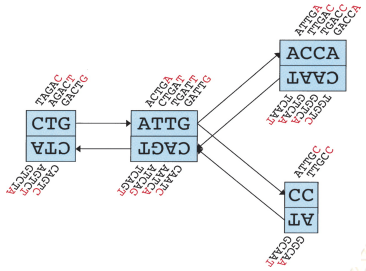


Figure from: Zerbino, D.R. & Birney, E. velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008)

Velvet (Bubble Correction)

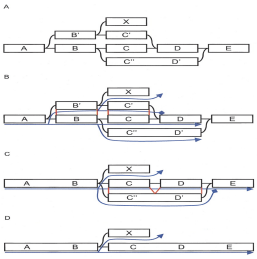


Figure from: Zerbino, D.R. & Birney, E. velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008)

Abyss (de Bruijn Graph)

- Utilizes a distributed de Bruijn graph for its assembly model.
- Hashing function to store kmers in a cluster.
- Efficient representation of edges in the de Bruijn graph.
  - Eight edges per k-mer.
  - Represent the presence or absence of the edges using eight bits.
- Successfully applied to 3.5 billion paired-end reads from the genome of an African male publicly released by Illumina.

Cap3 (overlap-layout-consensus)

- Developed in 1999, still used for 454 assembly.
- All overlap are computed between reads.
  - Five evaluation criterion.
  - Minimum length, percent identity, similarity score.
  - Number of differences at high quality values.
  - Difference of error rate of the two regions being aligned.
- Greedy layout of reads is used to construct contigs.
- Weighted multiple read alignment is used to derive the consensus.

CABOG (Overlap Graph)

- Modification of the Celera assembler.
- Overlap based trimming (preprocessing).
  - Overlaps between reads are used to confirm regions that are high quality.
- Seeded matching is used to detect overlapping fragments.
- Relies on a "Best overlap Graph"
  - No containment overlaps
  - Only one in and out edge per node
- Reduces memory consumption

ICD Assembly

New Proposed Approach: Merge and Traverse

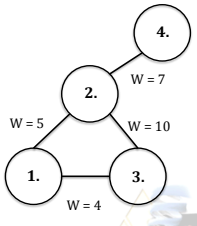
- Domain specific assembly
  - Graph theoretic model for assembly
  - Fragment overlap discovery
  - Graph construction and manipulation
  - Consensus sequence generation
- Assembly performance characteristics
  - Experimental setup
  - Coverage depth and assembly
  - Genome composition and assembly

First Generation Assembly Tools

- Next generation sequencing technologies arrived in the mid 2000's.
- Drastically different from previous sequencing technologies
- First generation short DNA fragment assemblers:
  - Velvet
  - SHARGCs
  - VCAKE
- Filled an important gap
- Developed in a data independent manner

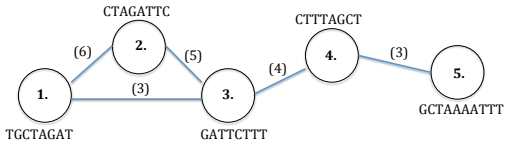
Graph Modeling

- Abstract Mathematical Model
  - Set of Objects
  - Object Relationships
- Nodes (Objects)
- Edges (Relationships)



Graph Theoretic Sequence Assembly

Overlap Graph Approach

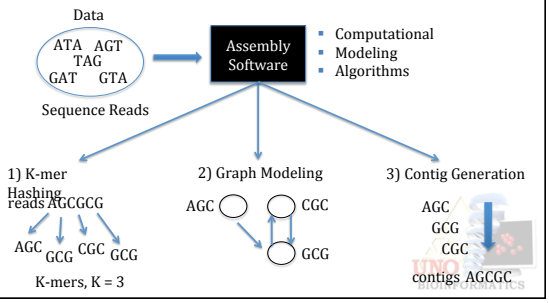


Reads: TGCTAGAT CTTTAGCT GATTCTTT CTAGATTC GCTAAAATTT

Min Ovl = 3 — Contig: TGCTAGAT TCTTTAGCT AAAATTT

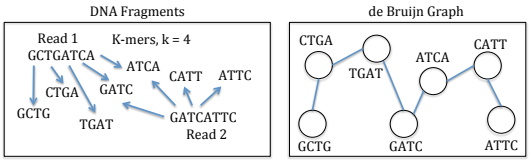
Graph Theoretic Sequence Assembly

De Bruijn Approach



Graph Theoretic Sequence Assembly

De Bruijn Approach



Domain-Specific Assembly

- Assembly should be a dynamic process - each sequencing project consists of multiple variables
  - Sequencing technology
  - Fragment length
  - Error rate
  - Genome coverage
  - Domain properties
- Some assembly methods may perform better in some assembly domains versus others
- **Key idea:** Domain specific assembler performance characterization

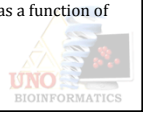
Project Objective

Merge and Traverse:
A graph theoretic approach for the assembly of next generation sequencing data

Goal:
Create a knowledge-based, data dependent approach for the assembly process

Two strategies:

1) Evaluate domain specific assembler performance
2) Characterize optimal parameter configurations as a function of dataset characteristics.

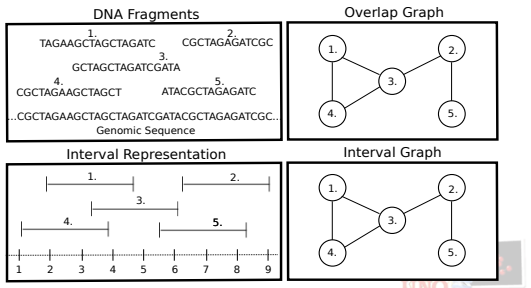


Merge and Traverse Assembler

Three module assembler

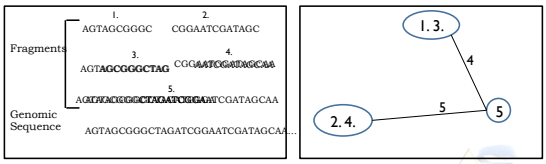
- Fragment clustering and overlap detection
- Tolerance graph construction and manipulation
  - Merging
  - Trimming
  - Traversal
- Consensus

Graph Theoretic Model

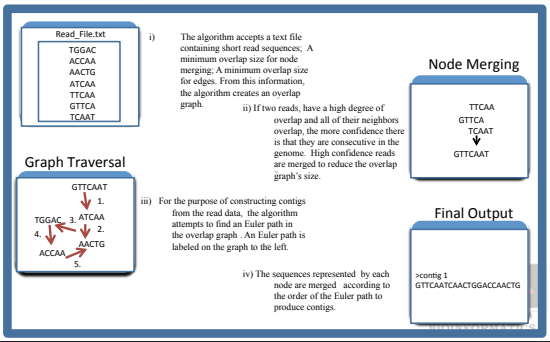


Merge and Traverse Assembly

Assembly through fragment merging and graph traversal



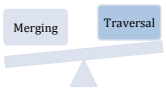
Assembly Algorithm



Balancing the Assembler

- The greedy node merging process may merge reads that have false-positive relationships
- A pure graph traversal approach may not be feasible in a highly complex and large overlap graph
- The proportions of graph traversal and node merging can be adjusted by changing the stringency of the merging parameter

What balance of graph traversal and node merging will produce the best assembly?



Assembly Challenges

- Errors and small overlaps
  - Sequencing errors

    ATTAGGTAGGTTTGAT
    TTACATTAGGCCG
  - Small overlaps

    ATTAGTTAGTTAGATTAC
    …GGCATTA
- Repeats
- Gaps

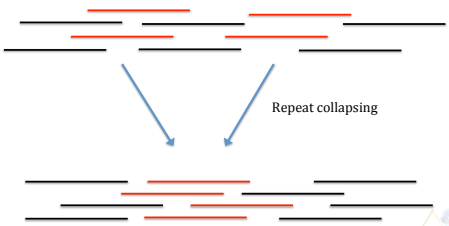
Main Challenge

Repetitive regions of the genome

Research Problem:
- Dynamic data characteristics
- Diverse applications
- General assembly algorithms
  - Limited flexibility
  - Data independent
  - Computational approach

Approach:
- Knowledge incorporation
- Domain specific
- Data-dependent assembly

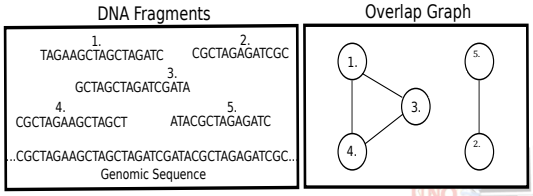
Repeats and mis-assembly



Tolerance Graph Model

Each fragment is assigned two tolerances

1) Minimum overlap length (how large is the overlap?)
2) Minimum overlap identity (what is the quality of the overlap?)



Tolerance Graph Model: Minimum overlap length = 5.

Fragment Clustering and Overlap Detection

Containment Overlap

AGCTAGATAGAGAGTCGCGCGCTAGAGGATCGCGGCTAGAGCTCG

GATAGAGAGTCGCGCGCTAGAGGATCGCGGCTAGA

Dove-tail Overlap

GATAGAGAGTCGCGCGCTAGAGGATCGCGGCTAGA

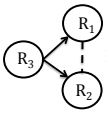
GCTAGAGGATCGCGGCTAGAACTTTATCTTA

Fragment Clustering and Overlap Detection

Overlap Tangles

AGCTAGATCGCGCGTAGATAGCCTAGAGAGGAGAGATCGCGCTAGATAGCT
CTAGAGCGCGTAGATAGCCTAGAGAG
AACTAGATCGCGCGTAGATAGCCTAGAGAGAGCTGGATAGCTCGATAGAGG

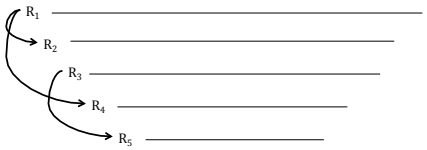
R_1 R_3 R_2



R_1 and R_2 do not overlap

Fragment Clustering and Overlap Detection

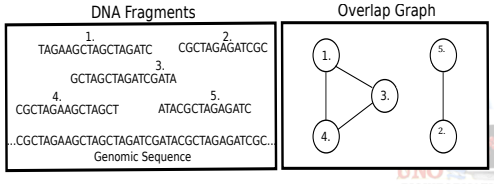
Greedy hierarchical clustering



R_2 and R_5 are clustered to R_1.
R_5 is clustered to R_3.

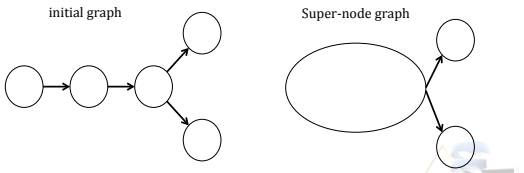
Graph Construction

- The construction of the tolerance graph is controlled by two parameters
  - Minimum overlap
  - Minimum identity



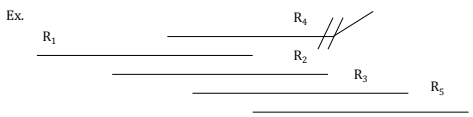
Node Merging

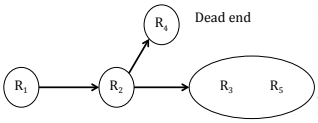
- Merging of non-branching dovetail overlap chains



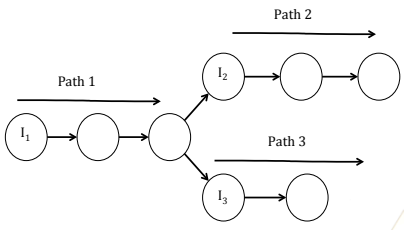
Graph Trimming

- Dead-end paths
- Bubbles





Graph Traversal



Initial starting nodes: I_1, I_2, I_3

Experimental Setup

Three total experiments

1) Coverage depth
2) Genome composition
   - GC rich bacterial genome
   - GC poor bacterial genome
3) Assembler comparative study

Experimental Setup

- Fine-tuning of tolerance graph parameters
- Initial minimum overlap parameter was set at 30 bps
  - Increased by a 20 bp step size until 290 bps
- Three assemblies were conducted at each iteration step
  - 90% alignment identity threshold
  - 93% alignment identity threshold
  - 96% alignment identity threshold
- A total of 42 assembly configurations per dataset.

Case Study1: Evaluating Assembly Quality

- Evaluation criteria

1. N50 statistic
   - Length N such that 50% of the assembly is contained in contigs that are longer than N.
2. Percentage of contigs mapped to reference genome
   - Alignment identity of at least 98%
   - Alignment coverage of at least 95%
3. Number of contigs

Case Study2: Coverage Depth and Assembly

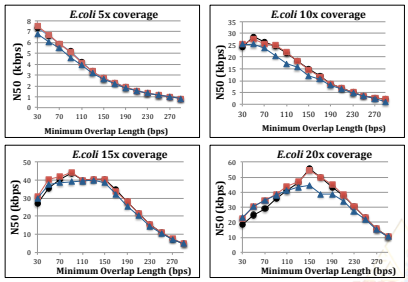
- *Escherichia coli* W datasets SRR060737 and SRR060736 were downloaded from NCBI's sequence read archive (SRA).
- These two datasets were combined and fragments were randomly select to form datasets at 5x, 10x, 15x, and 20x genome coverage.

***Escherichia coli* W dataset properties**

Dataset	*Coverage depth*	*Total Fragments*	*Avg. fragment length (bps)*
1	5x	56631	425
2	10x	113769	425
3	15x	174876	425
4	20x	232248	425

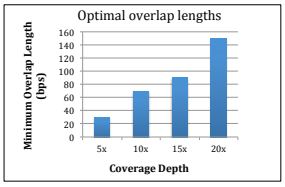
Coverage Depth and Assembly

N50 Statistic



■ = 93% overlap identity. ● = 90% overlap identity. ▲ = 96% overlap identity.

Coverage Depth and Assembly



Minimum overlap lengths that produce the greatest N50 lengths at different coverage depths

In all cases, over 88% of the contigs could be mapped to the reference genomes with an alignment identity of 98% and alignment coverage of 95%

CS3: Genome Composition and Assembly

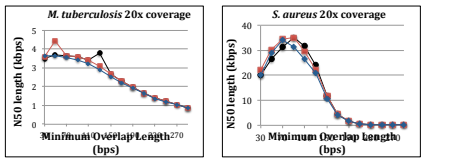
- *Staphylococcus aureus* dataset SRR014924
- *Mycobacterium tuberculosis* dataset SRR039113
- Fragments were randomly selected for 20x genome coverage

Assembly dataset properties

Dataset	*Genome GC content*	*Total Fragments*	*Avg. fragment length (bps)*
S. aureus	~32%	234385	238
M. tuberculosis	~66%	250998	350

Genome Composition and Assembly

N50 Statistic



■ = 93% overlap identity. ● = 90% overlap identity. ▲ = 96% overlap identity.

The N50 lengths of assemblies from the *Mycobacterium tuberculosis* genome are smaller than those of *Staphylococcus aureus.*

For all *Mycobacterium tuberculosis* assemblies, at least 90% of all contigs could be mapped back to the reference genome.

Assembly comparison

- Mira assembler
- 20x coverage *E.coli* dataset
  - Mira configuration: de novo, genome, accurate, and 454
  - Merge and Traverse configuration: 150 bp overlap and 90% identity

Comparative assembly results

Assembler	*Total Contigs*	*N50*	*Percent contigs aligned to reference*
Merge and Traverse (150 bp)	774	55469	89.1%
Mira	131	143578	87.0%

Summary

- We have introduced an algorithm for the assembly of next generation sequencing data
- Flexible enough to take input dataset characteristics into consideration
- Results demonstrate domain dependent assembly performance characteristics
- Results provide a knowledge base for the development of intelligent and customized approach for the assembly process
- Improved assembly tactics and better assembly results

Graph Enrichment

Graph Enrichment

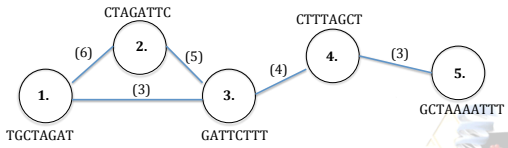
- Previous models
  - Static structure
  - Framework
  - General model
- *Knowledge-based parameter adjustment*
- Read characteristics
  - Length
  - Error rate
- Genome characteristics
  - Repeat content
  - Composition

Assembly is a Dynamic Process

- Each sequencing project consists of multiple variables
  - Sequencing technology
  - Fragment length
  - Error rate
  - Genome coverage
  - Domain properties
- Some assembly methods may perform better in some assembly domains versus others
- **Key idea:** Domain specific assembler performance characterization

Graph Enrichment: Structural Based

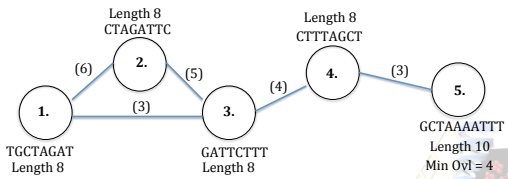
- Provides structure
- Skeletal framework



Min Ovl = 3

Graph Enrichment Example

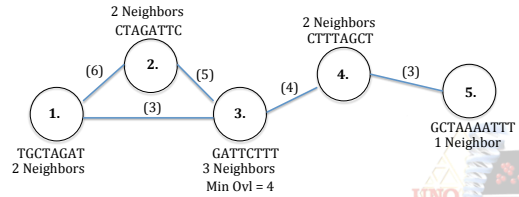
Read Specific Information
- Read length



Min Ovl = 3

Graph Enrichment Example

Node Specific Information
- Node degree



Graph Enrichment: How To

Normalizing graph and read information

$$Zscore\_degree(\mathrm{u}) = \frac{(\text{node\_degree(u)} - \text{average\_node\_degree})}{\text{node\_degree\_standard\_deviation}}$$

$$Zscore\_read\_length(u) = \frac{(\text{read\_length(u)} - \text{average\_read\_length})}{\text{read\_length\_standard\_distribution}}$$

$$Zscore\_minoverlap(\mathrm{u}) = \frac{(\text{minoverlap\_length} - \text{average\_overlap\_length})}{\text{overlap\_length\_standard\_deviation}}$$

$$Min\_Ovl(u) = a(Zscore\_degree(u)) + b(Zscore\_read\_length(u)) + c(Zscore\_minoverlap(u))$$

Experimental Results

- *Escherichia coli str. K-12 substr. MG1655 chromosome*
- 300 bp reads at 20x coverage, 1% error rate
  - Generated by MetaSim

N50 (bps)	Read Length Weight	Degree Weight	Domain Specific Weight
12978	0.1	0.1	0.8
13677	0.1	0.3	0.6
13406	0.1	0.5	0.4
12549	0.3	0.1	0.6
16761	0.3	0.3	0.4
16464	0.3	0.5	0.2
12095	0.5	0.1	0.4
15988	0.5	0.3	0.2
15427	0.5	0.5	0

Summary of Results

- Proposed ICD is Flexible
- Domain dependent assembly produces high performance
- Knowledge based approach
- Integrates data specific information
- Better assembly tactics
- Improved assembly results

Current Steps

- Incorporating additional information
- Iterative process on xarious stages/levels

Read Specific	Edge Specific	Node Specific
Read Length Quality Values Entropy GC Content Phylogeny Motifs	Min Identity Min Overlap Contig Specific Coverage Quality Statistics	Node Degree Path Length Clustering Coefficient Bridge Nodes Cut Nodes

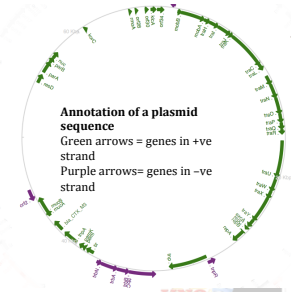
After The Assembly

After Assembly

- Annotation
- Genome Wide Association Studies (GWAS)
- Comparative Genomics
- Meta-genomic Studies (Eco-systems)

Annotation

- Finding genes and other features in the consensus sequence
- Study of insertion elements, gene disruption, mobile elements
- Project ENCODE



Annotation of a plasmid sequence
Green arrows = genes in +ve strand
Purple arrows= genes in –ve strand

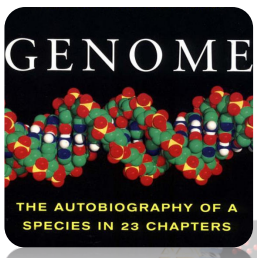
Visualization of Genome

Genome = A Book

Written in 4 letters of nucleotides – A T G C

23 Chromosomes = 23 Chapters

Genes = Stories in each chapter



GWAS

- Genome Wide Association Study
  - Rapid scanning of disease markers in genomes from a population
  - detect common genetic variants from a population and associate with a disease
- 2 key components
  - Human genome project
    - Not only DNA sequence, but all the genes have been identified
  - Hapmap project
    - Finding genes associated with diseases

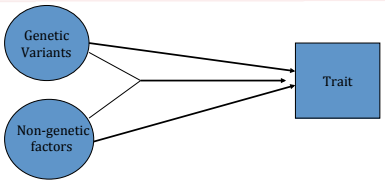
What is Genome Wide Association Study?

- A genome-wide association study (GWA study, or GWAS), is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases.
- Understanding the genetics of common and complex diseases.

DNA Variation

- >99.9 % of the sequence is identical between any two chromosomes.
  - Compare maternal and paternal chromosome 1 in single person
  - Compare Y chromosomes between two unrelated males
- Even though most of the sequence is identical between two chromosomes, since the genome sequence is so long (~3 billion base pairs), there are still many variations.
- Some DNA variations are responsible for biological changes, others have no known function.
- Alleles are the alternative forms of a DNA segment at a given genetic location.
- Genetic polymorphism: DNA segment with ≥ 2 common alleles.

Complex Traits - Multifactorial Inheritance



- Examples
  - Cancers
  - Type 1 diabetes
  - Hypertension
  - Rheumatoid arthritis
  - Schizophrenia
  - Type 2 diabetes
  - Alzheimer disease

Genetic Association Studies

- Short-term Goal: Identify genetic variants that explain differences in phenotype among individuals in a study population
  - Qualitative: disease status, presence/absence of congenital defect
  - Quantitative: blood glucose levels, % body fat
- If association found, then further study can follow to
  - Understand mechanism of action and disease etiology in individuals
  - Characterize relevance and/or impact in more general population
- Long-term goal: to inform process of identifying and delivering better prevention and treatment strategies

Single Nucleotide Polymorphisms: SNPs

- SNPs – DNA sequence variations that occur when a single nucleotide is altered

A	T	G	A	C	A	G	G	C
A	T	G	A	C	A	T	G	C

- Alleles at this SNP are "G" and "T"
- SNPs are the most common form of variation in the human genome
- SNPs catalogued in several databases

Genotypes and Haplotypes

- Genotype: pair of alleles (one paternal, one maternal) at a locus

Maternal	A	T	G	A	C	A	G	G	C
Paternal	A	T	G	A	C	A	T	G	C

Genotype for this individual is GT

- Haplotype: sequence of alleles along a single chromosome

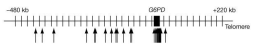
Maternal	A	T	G	C	C	A	T	G	C
Paternal	A	T	G	A	C	A	T	G	C

Genotypes for this individual (vertical) : CA and TT

Haplotypes (horizontal): CT and AT

Scope of a Genetic Association Study

- Candidate gene
  - Known functional variants
  - Variants with unknown function in exons, introns, regulatory regions



* Sabeti PC et al. (2002). *Nature* 419: 832-837

- Linkage candidate region
  - Functional variants, or those with unknown function in candidate genes
  - More general coverage of region using many markers
- Genome-wide
  - Test for association with hundreds of thousands (millions) of SNPs spread across the entire genome.
  - Many design strategies possible for distributing markers

Genome-Wide Association Studies

- Linkage analysis using families takes unbiased look at whole genome, but is underpowered for the size of genetic effects we expect to see for many complex genetic traits.
- Candidate gene association studies have greater power to identify smaller genetic effects, but rely on *a priori* knowledge about disease etiology.
- Genome-wide association studies combine the genomic coverage of linkage analysis with the power of association to have much better chance of finding complex trait susceptibility variants.

Why are They Possible Now?

Genotyping Technology:

- Now have ability to type hundreds of thousands (or millions) of SNPs in one reaction on a "SNP chip." The cost can be as low as $200-$300 per person
- Two primary platforms: Affymetrix and Illumina.

Design and analysis:

- Availability of SNP databases, HapMap, and other resources to identify the SNPs and design SNP chips.
- Faster computers to carry out the millions of calculations make implementation possible.

Design and Analysis Strategies: Moving Target

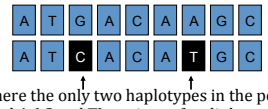
- A genetic factor is like any other potential risk factor and the same study design and analysis principles hold – in addition to those specific to GWAs.
- Standard case-control (matched or unmatched), cohort-based quantitative trait and longitudinal designs are common.
- Focus today is on case-control design, but many of the principles apply to other designs.

SNP Chips: Number and Placement of SNPs

- A "typical" SNP chip has at least 317,000 SNPs distributed across the genome. Newest: ~1 million.
- The newest chips can also measure (directly or indirectly) some types of copy number variation.
- We do not directly measure genotypes at all genetic polymorphisms, but rely on association between the polymorphisms we do assay and those which we do not assay.
- SNP-SNP association, or linkage disequilibrium, is fundamental to our ability to sample the whole genome with relatively few SNPs.

Linkage Disequilibrium (LD)

- Linkage disequilibrium: the non-random association of alleles at linked loci.
- A measure of the tendency of some alleles to be inherited together on haplotypes descended from ancestral chromosomes.



- If these where the only two haplotypes in the population, then alleles G and A (C and T) are in perfect linkage disequilibrium.
- If we genotype the first SNP, we know what the alleles are at the second SNP.

HapMap

- International collaboration of scientists
  - Identified the genetic similarities and differences in large population of human beings
- Hapmap is a catalog of genetic variants in human
  - What are the variants
  - Location of variants (which genes)
  - Distribution within population and among populations across the world

HapMap

- Multi-country effort to identify, catalog common human genetic variants.
- Developed to better understand and catalogue LD patterns across the genome in several populations.
- Genotyped ~4 million SNPs on samples of African, east Asian, European ancestry.
- All genotype data in a publicly available data base.
- Can download the genotype data
  - Able to examine LD patterns across genome
  - Can estimate approximate coverage of a given SNP chip
- Can represent 80-90% of common SNPs with
  - ~300,000 tag SNPs for European or Asian samples
  - ~500,000 tag SNPs for African samples

Recent studies

- A genome-wide association study in the Japanese population confirms 9p21 and 14q23 as susceptibility loci for primary open angle glaucoma, *Human Molecular Genetics 2012*
  - Primary open angle glaucoma (leading cause for adult blindness)
  - 1394 cases and 6599 controls
  - Identified 34 SNPs association
  - confirmed 9p21 (chromome9, p arm, region 21) and 14q23 as susceptible loci for the disease
- Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis, *Nature 2011*
  - *1841 cases and 5163 controls*
  - *12 new susceptible genes are identified for the diesase*

More Studies

- Genome-wide association study identifies three new melanoma susceptibility loci, Nature Genetics 2011
- A genome-wide association study identifies a potential novel gene locus for keratoconus, one of the commonest causes for corneal transplantation in developed countries, Human Molecular Genetics 2011
- A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease, Nature Genetics 2011
- A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants, Nature 2007

Case and Control Selection

- Case and control samples may be population-based
- Cases and controls may be chosen to increase magnitude of contrast

Case sample may be selected to be enriched for predisposing variant(s)

- Family history
- Early age of onset
- Increased severity of disease

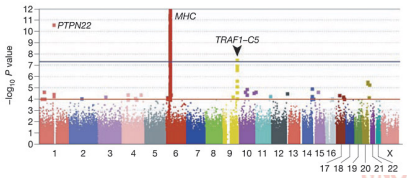
Control sample may be selected to be "very healthy" or "super controls"

- E.g. for type 2 diabetes, may select individuals who have normal response to glucose at age 70
- Control selection just as important (and tricky) as for any case-control study.

120

Testing for Genetic Association with Disease

- Question of interest: Are the alleles or genotypes at a genetic marker associated with disease status?
- Use usual statistical machinery get estimates of measures of association and to test for association for each of the SNPs.
- One typical approach: Test for association between having 0, 1 or 2 copies of rare allele at a SNP using Cochran-Armitage test for trend.



Pearson, T. A. et al. JAMA 2008;299:1335-1344.

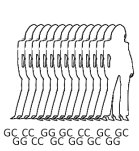
Interpreting the Statistical Results

- Testing for association at each of hundreds of thousands of markers dictates that traditional statistical significance thresholds (e.g. α=.05) not appropriate.
- That aside (more in a few minutes), if you identify a SNP that is significantly associated with disease, there are three possibilities:
  - There is a causal relationship between SNP and disease
  - The marker is in linkage disequilibrium with a causal locus
  - False positive
- Many potential sources of systematic errors that might lead to false positive results.
- Genotyping quality control issues particularly important.

GWAS and diseases

- Bipolar disorder
- Coronary artery disease
- Crohn's disease
- Hypertension
- Rheumatoid arthritis
- Type 1 diabetes, Type 2 diabetes
- Different cancer types

Finding significant SNPs

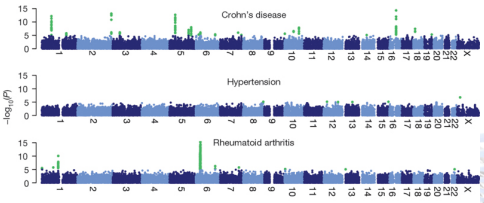


SNP1	*SNP2*	*SNP…*
Cases Count of G: 2104 of 4000	**Cases** Count of G: 1648 of 4000	*Repeat for all SNPs*
Frequency of G: 52.6%	Frequency of G: 41.2%	
Controls Count of G: 2676 of 6000	**Controls** Count of G: 2532 of 6000	
Frequency of G: 44.6%	Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

http://en.wikipedia.org/wiki/Genome-wide_association_study

Manhattan Plots

- *X-axis*
  - *Chromosome Number*
- *Y-axis*
  - *P* value for quality-control-positive SNPs
  - *P* values <1 10^{-5} highlighted in green



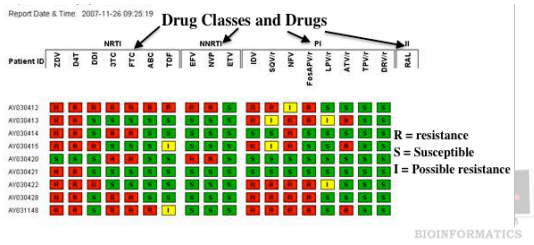
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007

HIV and drug resistance

- HIV mutates and reproduces in presence of antiretroviral drugs
  - Drug resistance
- Stanford University Data Base
  - Curated database to analyze HIV drug resistance
  - hivdb.stanford.edu
  - Mutation frequencies at each position of Protease or RT and HIV subtypes and the treatments

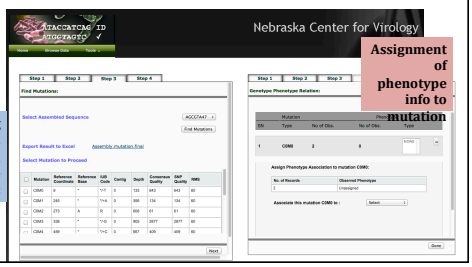
Commercial Software

- BioNumerics



At University of Nebraska at Omaha

- Collaboration with Nebraska Center for Virology (https://hivis.ist.unomaha.edu/)

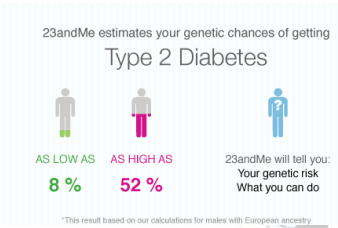


Personalized Medicine

- 23 and Me
  - Personal genetic information
  - Illumina Genotype Chip
  - DNA from cheek cells in saliva
- Approximately one million variants are tested using probes
- Set of markers associated with diseases and traits
  - Lactose intolerance, baldness,
  - Diabetes, Alzheimer's disease,

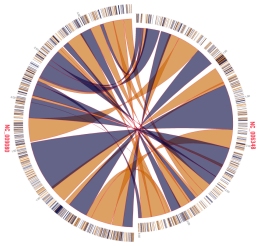
23 and Me

- Disease Risk
  - Type2 diabetes
  - Inheritable conditions like Cystic Fibrosis, etc.
- Sensitivity to certain drugs



Comparative Genomics

- Studying genomic structures among different organisms (species or strains)
- Evolutionary events in the history



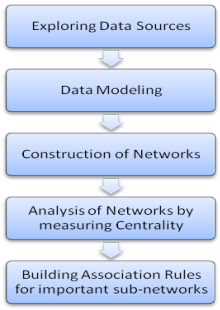
Many genomic regions appearing at different locations in Burkholderia sp.

A Graph Theoretical Approach for Integrating Association Rule Mining and GWA Database in the Analysis of Cancer Correlation Networks

Key Objectives

- Model the intricate relationship between Genes, Variants and Diseases using a Graph
- Visualize the associations and highlight the common features among the various relationships.
- Further explore the relationship to extract meaningful features describing the observed behavior of genetic variants for a phenotype.

Methodology



Exploring Data Sources

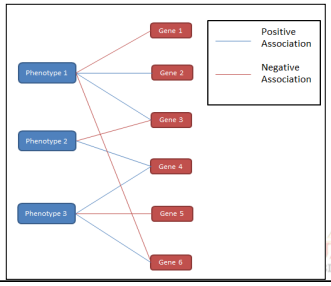
- Scientists have been able to explore the genetic similarity between diseases because of the availability of large-scale knowledge-bases such as the Online Mendelian Inheritance in Man (OMIM) and Genetic Association Database (GAD).
- Comprehensive GWAS have branded thousands of disease-variants association.
- The National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies (NHGRI GWAS catalog) conveniently lists significantly associated marker SNPs from these studies in a manually curated, online database.
- The catalog includes data on study designs, individual SNV P-values, odds ratios and links to the published studies.

Modeling

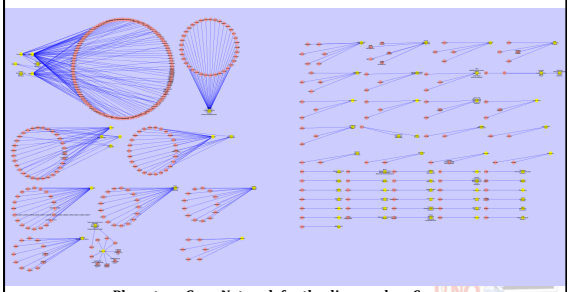
- We generated two types of networks:
  - Phenotype Variant Network
  - Phenotype Gene Network.
- A Phenotype-Variant network is such that: $G_{PV} = (P,V,E)$;
  - where P is the set of Phenotypes, V is the set of variants and E is an edge between P and V if there is a reported association between P and V.
- A Phenotype- Gene network is such that: $G_{PG} = (P,G,E)$;
  - where P is the set of phenotypes, G is the set of genes and E is the edge between P and G if there is a reported association between P and G.

Construction of Networks

A bi-partite graph with Phenotypes as one set and Genes as another connected by an edge if there is an association between the two. Edges are colored blue if there is a reported association and red if it is reported that no association exists between the phenotype and gene.



Phenotype- Gene Network



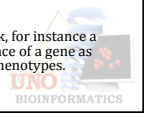
Phenotype Gene Network for the disease class Cancer

Network Analysis

- In many types of networks, including gene networks, one can assume that the importance of a node (for example the likelihood of causing a disease) is correlated with centrality.

Network Analysis

- We measure four different centrality quantities:
  - Degree
    - In our Phenotype-Gene network, phenotypes and genes with very high degree are interacting with several others, thus suggesting a central regulatory role, that is they are likely to be regulatory hubs.
  - Eccentricity
    - In our Phenotype-variant network, eccentricity can be interpreted as the easiness of a phenotype to be functionally reached by the variants in the network.
  - Closeness
    - In our Phenotype-gene network, closeness can be interpreted as the probability of a gene to be functionally relevant for several other phenotypes, but with the possibility to be irrelevant for few other phenotypes.
  - Betweenness
    - The Betweenness of a node in a biological network, for instance a Phenotype-gene network, can indicate the relevance of a gene as functionally capable of holding together related phenotypes.



Association Rule Mining

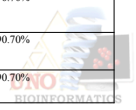
- Graph modeling and network visualization was used as an intermediate step in our final analysis. The graphs are used as pre-processors for the association rule building in our study as we filter the database on the basis of phenotypes with high betweenness centrality
- An association rule is an implication expression of the form X→Y, where X and Y are disjoint item sets. The strength of an association can be measured in terms of its support and confidence.

Support and Confidence Scores

- For the genome wide association study done on phenotypes (diseases) and genetic variants, support is an important measure because a rule that has very low confidence may occur simply by chance.
- A low support rule is also likely to be generally uninteresting however for phenotypes and diseases it usually is not the case as it may describe the exciting analysis of a small population which might be said to be a characteristic of that population.
- Confidence, on the other hand, measures the reliability of the inference made by the rule. For a given rule, X→ Y, the higher the confidence, the more likely it is for Y to be present in the result set that contain X. Confidence also provides an estimate of the conditional probability of Y given X.

Results

Association Rule	Accuracy
Broad Phenotype=Lung cancer ; Chromosome=15.0 ==> Association=Y	98.67%
Broad Phenotype=Prostate cancer; Chromosome=8.0 ==> Association=Y	98.32%
Broad Phenotype=Prostate cancer; Chromosome=17.0 ==> Association=Y	97.07%
Broad Phenotype=Prostate cancer; Chromosome=11.0 ==> Association=Y	96.07%
Broad Phenotype=Colorectal Cancer; Gene=SMAD7 ==> Association=Y	90.70%
Broad Phenotype=Breast cancer; Chromosome=10.0 ==> Association=Y Gene=FGFR2	90.70%
Broad Phenotype=Breast cancer Gene=FGFR2 ==> Association=Y Chromosome=10.0	90.70%
Broad Phenotype=Lung cancer Gene=CLPTM1L ==> Association=Y Chromosome=5.0	90.70%
Broad Phenotype=Lung cancer Gene=LOC123688 ==> Association=Y Chromosome=15.0	90.70%



Analysis of Association Rule Mining Results

- The rules shown in Table predict the accuracy of the associations formed by the attributes.
- For example, the accuracy of the association between Lung cancer and chromosome 15 is 98.67% which means that the genes and single nucleotide variants found on chromosome 15 which are linked with lung cancer, are strong candidates for being causative of the disease.
- Similarly, the accuracy of the gene SMAD7 being a causative for Colorectal Cancer is 90.7%.

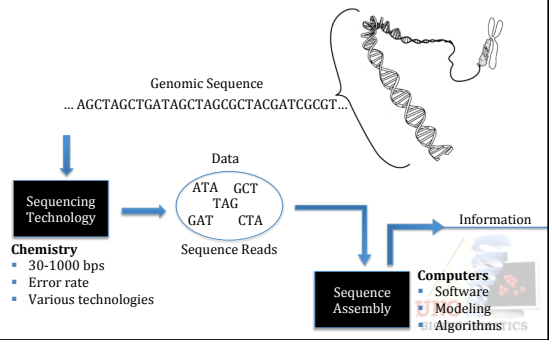
High Performance Computing (HPC)and Assembly

Biological Data and High Performance Computing

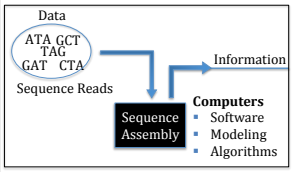
Biological Data
- Increasing exponentially
- Novel technologies
- Data Analysis

High Performance Computing
- Database search
- Phylogeny
- Sequence Analysis
- Assembly
- Network Analysis

Sequencing



Sequence Assembly



Assembly Algorithm Overview

Step 1: Read overlapping
- Suffix array seed and extend
- Needleman and Wunsch Alignment

Step 2: Graph Modeling
- Reads -> Nodes
- Overlaps -> Edges
- Transitive reduction
- Node merging
- Error correction
- Traversal

Step 3: Consensus Sequence
- Multiple sequence alignment

High Performance Computing

Why do we need high performance computing (HPC)?

Sequence Read Archive (SRA) at NCBI[7]
- Surpassed 100 terabases (2011)

SRA Composition in 2011

Massive amounts of data

Why HPC?

Roche/454

Approximate Output	
Technology	Output
GS FLX Titanium XL +	700 Mb
GS FLX Titanium XLR70	400 Mb

SOLiD

Approximate Output	
Technology	Output
5500	7-9 Gb
5500xl	10 – 15 Gb
5500xl (.75 µm nanobeads)	> 20 Gb

Illumina Genome Analyzer IIx

Approximate Output	
Read Length	Output
1 x 35 bp	10-12 Gb
2 x 50 bp	25-30 Gb
2 x 75 bp	37.5-45 Gb
2 x 100 bp	54-60 Gb
2 x 150 bp	85-95 Gb

High-throughput production of data

Energy-Aware HPC

Why do we need energy-aware HPC?

Next generation sequencing has become the backbone of many research applications.

- Infectious disease research
- Personalized medicine
- Metagenomics
- Cancer research
- Numerous other applications

The Assembly Algorithm

Step 1: Read overlapping
- Suffix array seed and extend
- Needleman and Wunsch Alignment

Step 2: Graph Modeling
- Reads -> Nodes
- Overlaps -> Edges
- Transitive reduction
- Node merging
- Error correction
- Traversal

Step 3: Consensus Sequence
- Multiple sequence alignment

Containment Overlap

r_1 AGCTAGCTAGAGGATCGCGGCTAGAAATCGAAA

r_2 GCTAGAGGATCGCGGCTAG

Dovetail Overlap

r_1 AGCTAGCTAGAGGATCGCGGCT

r_2 GGATCGCGGCTGGCAATTCCAAA

Two Step Approach
- Cluster reads contained in overlaps to parent reads
- Determine dovetail overlaps among the remaining reads

Reverse complements of reads are also used to find overlaps

The Assembly Algorithm

Algorithm

SORT reads in descending order of length

FOR each read R

IF not in_cluster(R)

R <- cluster_representative

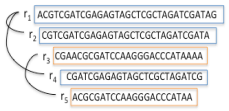
FOR each read S, length(S) < length (R)

IF is_contained(S, R) AND

IF align(S, R) >= min_align_score AND

IF align(S, R) >= previous_align_score

S <- cluster(R)



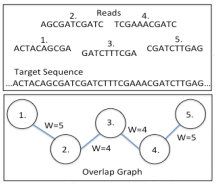
Containment clustering
- Reads are sorted in descending order.
- Reads two and four cluster to read one.
- Read five clusters to read three.

The Assembly Algorithm

Step 1: Read overlapping
- Suffix array seed and extend
- Needleman and Wunsch Alignment

Step 2: Graph Modeling
- Reads -> Nodes
- Overlaps -> Edges
- Transitive reduction
- Node merging
- Error correction
- Traversal

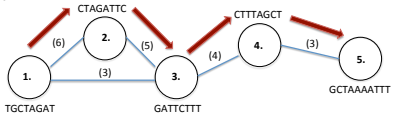
Step 3: Consensus Sequence
- Multiple sequence alignment



The overlap graph
- Reads map to nodes.
- Overlaps map to edges.

The Assembly Algorithm

Graph Traversal And Assembly



The Assembly Algorithm

Step 1: Read overlapping
- Suffix array seed and extend
- Needleman and Wunsch Alignment

Step 2: Graph Modeling
- Reads -> Nodes
- Overlaps -> Edges
- Transitive reduction
- Node merging
- Error correction
- Traversal

Step 3: Consensus Sequence
- Multiple sequence alignment

Iterative multiple alignment merges reads into contigs

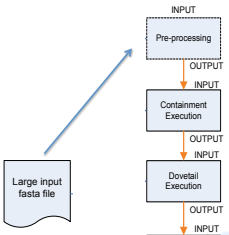
Parallel Read Overlapping

Read Overlapping Overview

Step 1: Read Preprocessing
- Sort in descending order of length
- Generate reverse complements
- Partition into *n* subsets
- Pairs of read subsets are tasks

Step 2: Containment Clustering
- Not naively parallel
- Task dependencies

Step 3: Dovetail Overlapping
- Naively parallel
- No task dependencies

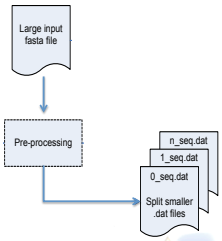


Parallel Read Overlapping

Step 1: Read Preprocessing
- Sort in descending order of length
- Generate reverse complements
- Partition into *n* subsets
- Pairs of read subsets are tasks

Step 2: Containment Clustering
- Not naively parallel
- Task dependencies

Step 3: Dovetail Overlapping
- Naively parallel
- No task dependencies



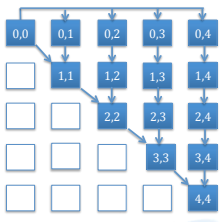
Subsets of set *S* are ordered such that *readLengths*(S_0) ≥ *readLengths*(S_1) ≥ … ≥ *readLengths*(S_{n-1}).

Parallel Read Overlapping

Step 1: Read Preprocessing
- Sort in descending order of length
- Generate reverse complements
- Partition into *n* subsets
- Pairs of read subsets are tasks

Step 2: Containment Clustering
- Not naively parallel
- Task dependencies

Step 3: Dovetail Overlapping
- Naively parallel
- No task dependencies



Containment clustering dependencies

Diagonal tasks (0, 0), (1, 1) (2, 2), (3, 3) and (4, 4) are considered to be higher priority tasks because they have a greater number of child/dependent tasks

Parallel Read Overlapping

Step 1: Read Preprocessing
- Sort in descending order of length
- Generate reverse complements
- Partition into *n* subsets
- Pairs of read subsets are tasks

Step 2: Containment Clustering
- Not naively parallel
- Task dependencies

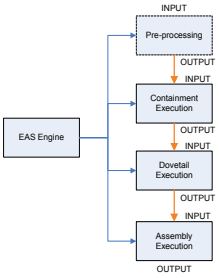
Step 3: Dovetail Overlapping
- Naively parallel
- No task dependencies

Total dovetail tasks T

$$T = \frac{n(n+1)}{2}$$

Dovetail tasks have no dependencies

Energy Aware Scheduling



Process flow diagram

Case Study: Experimental Setup

Dataset
- *Escherichia coli W* reads produced by the 454 Titanium technology
  - Obtained from NCBI's Sequence Read Archive (SRA)
  - SRR060736 and SRR060737, made public by JCVI
- 337,294 trimmed reads
  - 674,588 reads including reverse complements
- Split into 40 read subsets during preprocessing
  - Approximately 16,866 reads per subset

Case Study: Experimental Setup

Environment
- Firefly Cluster
- Holland Computing Center
- 1,151-node supercomputer of Dell SC1435 servers
- 800 MB/sec Infiniband interconnect

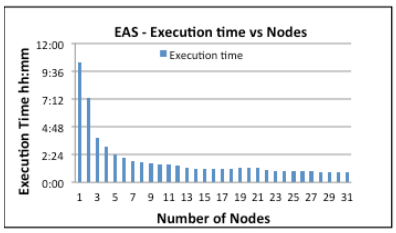
Node Architecture
- Two sockets
- Four 64-bit AMD Opteron 2.2 GHz processors per socket
- 8 GB of memory and 73 GB of disk space per node

Case Study: Experimental Setup

I. EAS – Execution time versus Nodes
II. Execution time/Overhead versus Nodes
III. Assembler Speedup Curve

Results

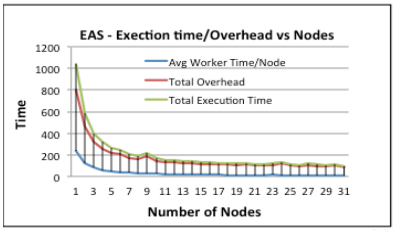
I. EAS – Execution time versus Nodes



After 11 to 12 nodes we do not see any significant performance gain

Results

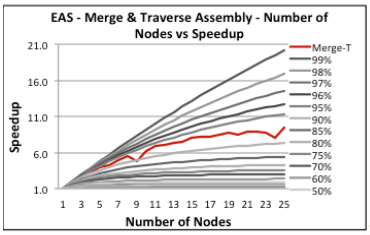
II. Execution time/Overhead versus Nodes



Overhead = Total Execution Time – Avg. Worker Time/Node

Results

III. Assembler Speedup Curve



Approximately a 20 - 25 times speedup (which is between 90% - 95% paralleliza

Experimental Setup

IV. Dynamic Node Adjustment
- Five different deadlines
- 30, 60, 90, 120, and 150 minutes
- Same environment as previous experiments

Table 1. Read subset groups used for analysis (*E.coli W*)

Group	Number of Files	Number of Reads
G1	5	84330
G2	10	168660
G3	15	337320
G4	20	674588

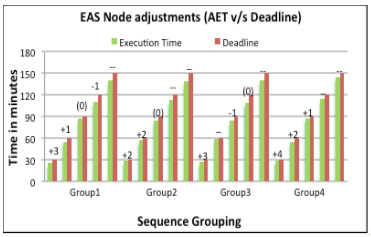
Experimental Setup

EAS Engine

- Profile/speedup curve based assignment of starting node number
- EET (Expected Execution Time)
- AET (Actual Execution Time)
- EET–AET variance
  - Number of nodes adjusted up (+N) or down (–N)

Results

IV. Dynamic Node Adjustment



The EAS engine was able to dynamically adjust nodes to minimize energy utilized while meeting the deadlines.

Storage and Multilevel Modeling

Short Read Storage and Multilevel Modeling

- Dataset size and complexity
- Metagenomics
- Graph Coarsening
  - Structures
  - Parameters
- Graph storage
- Efficient graph relabeling
- Experiments and Results

Massive Datasets

Illumina Genome Analyzer IIx

Illumina Genome Analyzer
- 1 x 35 bp reads (10 – 12 Gb)
- 2 x 50 bp reads (25 – 30 Gb)
- 2 x 75 bp reads (37.5 – 45 Gb)
- 2 x 100 bp reads (54 -60 Gb)
- 2 x 150 bp reads (85 – 95 Gb)

454 GS FLX+ System
- 700 bp reads (700 Mb)
- 450 bp reads (450 Mb)

Complex Datasets

- Multiple assembly sub-problems
  - RNA-seq applications
  - Single cell sequencing
  - Metagenomics
- Underlying genome complexity
  - Repeats
  - Sequencing error
  - Sequencing gaps

Metagenomics

- Analysis of communities of microbes
- Few metagenomics-specific assemblers
  - MetaVelvet
  - Genovo
  - ShoRAH
- Major metagenomics studies
  - Human microbiome
  - Sargasso sea

Metagenomics Analysis

- Clustering (benefits)
  - Feature discovery: Genes, OTUs, similar reads
  - Reduce redundancy
  - Reduce complexity
  - Improve assembly
- Clustering (negatives)
  - Information loss
  - Removal of most read overlap relationships
- Can we combine the benefits of clustering with the structure of the overlap graph?

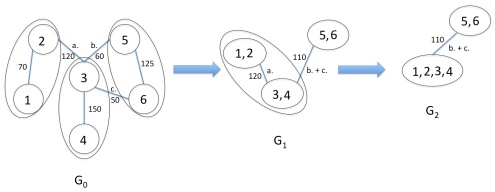
Multi-level Graph Coarsening

- Create the initial graph G_0 from read overlaps
- Find a maximal Heavy Edge Matching (HEM) M on G_0
- Merge nodes that are endpoints of edges in M to form G_1

Graph Coarsening

- Remove each edge with endpoints that are the same in G_1
- Duplicate edges are combined into a single edge and their weights are summed
- Repeat the matching/merging process on G_1, G_2, G_3 …. G_n

Graph Coarsening



Final result is a series of graphs G_0, G_1, G_2 … G_n

Graph Coarsening Structures

- *node_map*
  - Let (u, v) be an edge in a matching M on graph G_n, then *node_map*[u] = *node_map*[v] = z, where z is the label of the combined node in G_{n+1}.
- *node_map_inverse*
  - The inverse of *node_map*. Let z be a node in G_{n+1}, then *node_map_inverse*[2*z] = u and *node_map_inverse*[2*z +1] = v, where (u, v) is a matched node in G_n.

Graph Coarsening Structures

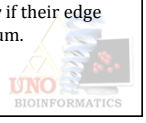
- *node_weights*
  - For each node in a graph G_n, the array *node_weights* stores the number of child nodes in G_0 descended from that node.
- *edge_weights*
  - For each node in a graph G_n, the array *node_weights* stores the combined edge weights of the edges induced by its child nodes in G_0.

Graph Coarsening Parameters

- Minimum cluster edge density
  - The edge density of a node *z* in G_{n+1}, formed by nodes *u* and *v* in G_n is given by:

$$\frac{2*(ew[u]+ew[v]+w(u,v))}{(nw[u]+nw[v])*((nw[u]+nw[v])-1)},$$

  where *ew*[*u*] = *edge_weights*[u] and *nw*[u] = *node_weights*[v]
- Minimum edge weight
  - A node *u* will not be matched with a node *v* if their edge weight is less than a user-provided minimum.



Graph Coarsening

- Clustering similarities
  - Feature discovery: similar reads
  - Reduce complexity
  - Reduce graph size
- Overlap graph similarities
  - Provide structure and organization for reads
  - No information loss

Graph Coarsening

- A local-to-global view
  - The graph is recorded at each level of merging
  - Captures multiple levels of information granularity
- How is this useful? (Feature extraction)
  - Reduced graph: A large node weight or edge density may indicate that a node is a part of a repeat.
  - Reduced graph: Cycles may indicate repeats.
  - Reduced graph: Different GC values across nodes may correspond to different species or genome regions.
  - Full graph: Small scale details like individual overlap information are retained.

Graph Storage

- Relies on succinct dictionary structures.
  - An indexed bit array with two functions.
  - *Rank* function – returns the rank of the *i*th position in the array.
  - *Select* function – returns the position of the *i*th bit set to one.

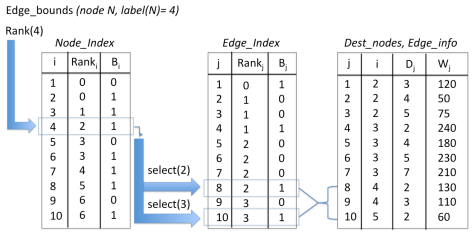
Graph Storage

- *node_index*
  - Contains a bit array of length *n*, where *n* is the number of nodes in the overlap graph.
  - If a node in the overlap graph has incident edges, then its corresponding bit is set to one.
  - If a node has no incident edges, then its corresponding bit remains set to zero.

Graph Storage

- *edge_index*
  - For each node that contains edges in the overlap graph, a bit in the array is set to one to index the start position of that node's edges in the edge list.
  - The last bit in the array is set to one.
- *dest_nodes*
  - Stores each edge's destination node's label in the order of the edge list.
- *edge_info*
  - Stores overlap length and identity information.

Graph Storage

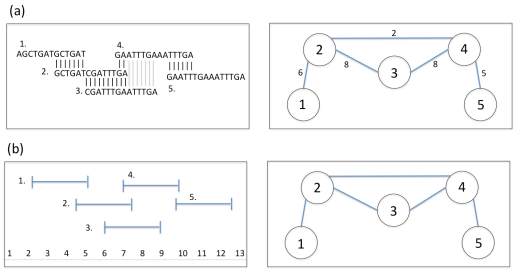


The *edge_bounds* function

Efficient Node Relabeling

- Re-label nodes such that they are close to one another in the succinct dictionary structures
- Allows for efficient access of nodes from the graph data structure.
- Takes advantage of the natural ordering of the interval graph to organize the nodes.

Efficient Node Relabeling



(a) The overlap graph (b) The interval graph. Notice that there is a natural left-to-right (right-to-left) ordering of the intervals in (b) that can be extended to nodes in the interval graph.

Experiments

- Metagenomics
  - Eight reference genomes were downloaded from NCBI.
  - 40 000 reads were generated from each genome by Metasim.
- Node relabeling
  - Three bacterial short read datasets were downloaded from NCBI.
  - *Escherichia coli, Staphylococcus aureus,* and *mycobacterium tuberculosis*

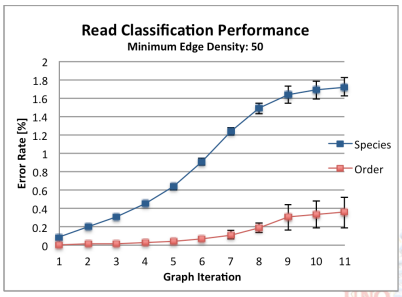
Experiments

- Metagenomics
  - After graph coarsening, clusters recovered at each graph level.
  - Classification assignment at the species, order, and phylum level was determined by majority read vote.
  - Error rate was defined to be the percentage of reads assigned incorrectly to a cluster.
- Node relabeling
  - The smaller the distance between two nodes' labels, the closer they are to one another in the graph data structure.
  - Calculation of the difference between the node labels of each edge.

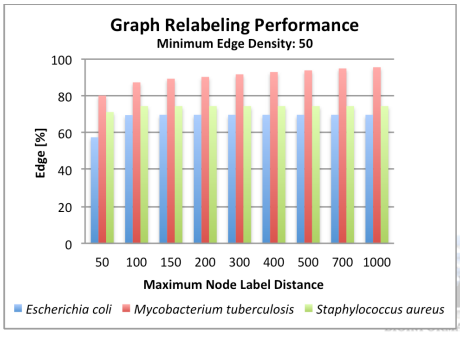
Efficient Node Relabeling

- Graph coarsening to produce a series of graphs G_0, G_1, G_2, … G_n.
- Extract paths in the final reduced graph G_n
- Relabel G_0 according to the clusters and path ordering obtained from G_n.
- Create a final graph G_{final} from the relabeling of G_0.

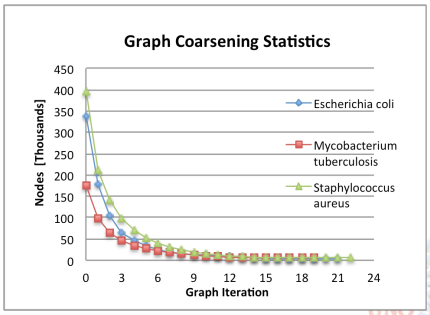
Results



Results



Results



Summary

- Graph coarsening reveals multiple levels of information granularity
- The reduced graph can reveal global features of the dataset, while small scale details are retained in the full overlap graph
- We have introduced a efficient method for storing our overlap graphs that takes advantage of the inherent properties of the interval graph

References

[1] AE Darling *et al*, "The Design, Implementation, and Evaluation of mpiBlast" *Computer*, pp. 13-15. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.3974

[2] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models", *Bioinformatics,* vol. 22, no. 21, pp. 2688-90, 2006

[3] MC. Schatz, "High-throughput sequence alignment using Graphics Processing Units", *BMC Bioinformatics*, vol. 8, no. 474, 2007

[4] Illumina. *Systems/genome analyzer iix.* Retrieved from http://www.illumina.com/systems/genome_analyzer_iix.ilmn

References

[5] Applied Biosystems. 5500 Series Genetic Analysis Systems. Retrieved http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html

[6] Roche/454. GS FLX + System. Retrieved from http://454.com/products/gs-flx-system/index.asp

[7] Y. Kodama et al, "The Sequence Read Archive: explosive growth of sequencing data", *Nucleic acids res.,* vol. 40, pp. 54 – 56, 2012

[8] Zerbino, D.R. & Birney, E. velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008)