Parallel Graph Genome Assembly

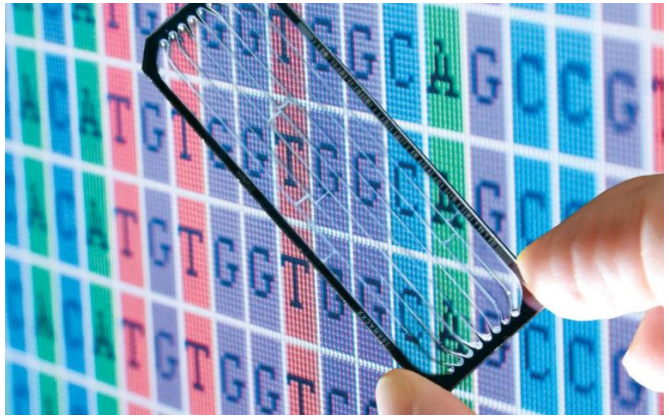
Aaron Solomon

Genomics - what is it?

- DNA - life instructions
- Four bases: A, T, G, C
- 6,000,000,000 base pairs/cell
- Uniquely identifies
  - Individuals
  - Species

Genomics - why do we care?

- Disease
  - Cancer
  - Hereditary Disease
  - Therapeutic/Drug Targeting
- Ancestry
  - Paternity
  - Heritage
- Organism Identification
  - **Pathogen Identification/Monitoring**
  - **Epidemic Tracking**
  - **Conservation/Speciation**



Sequencing Changes

Process

GATTC TAGGGCC

ATGCC CCCGGG

AATTTAA GCGCGC

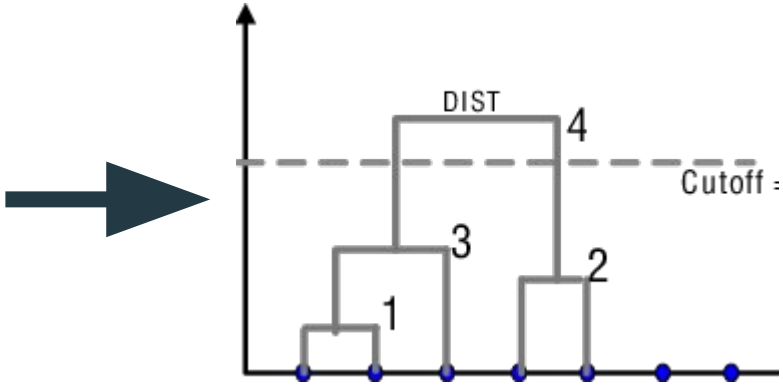
AAGGGCC

ATGCCGGGCGCGATTC...

Analysis

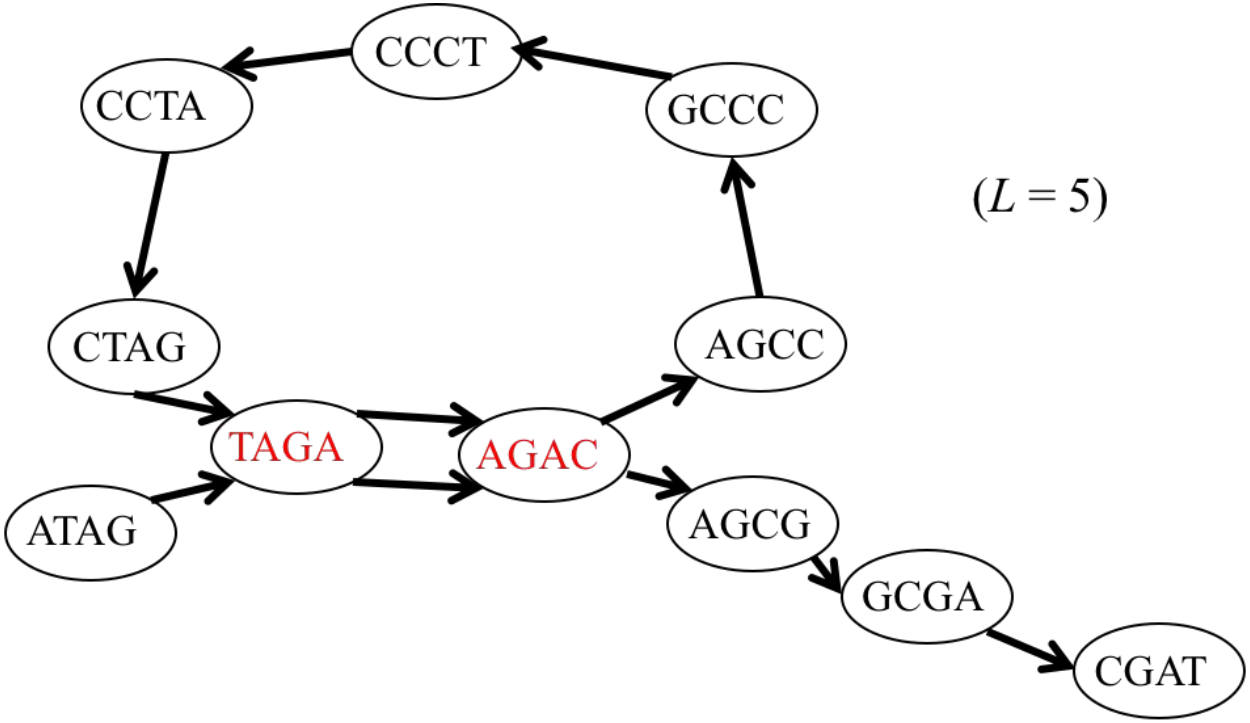
Assembly: Greedy

	ATGCC	AATTTAA	GATTC	TAGGGCC
ATGCC	X	X	X	X
AATTTAA	X	X	X	X
GATTC	X	X	X	X
TAGGGCC	X	X	X	X



Cluster

Assembly 2.0 - Euler Paths



$(L = 5)$

Current Technology

- No or weak parallelism
  - Examples
    - Velvet
    - ABySS
    - SOAPdenovo
- Parallel graph construction
  - Examples
    - ABySS 2.0
    - Ray
- Streaming graph construction
    - Faucet

Proposed Integrations

- Streaming Graph Generation (**Faucet**)
  - Minimize memory utilization
- Parallel Eulerian Tour Selection (**ABySS 2.0**)
  - Reduce computation time
- Probabilistic Error Avoidance (**LightAssembler**)
  - Eliminate fitting error on bad reads

Challenges

- Sequencing Errors
  - Bubbles
  - Dead-End Branches
- Graph Parallelism
  - High connectivity
  - Genomic repeats

Naiad vs GiRaph Comparisons

- Timing
  - Graph construction
  - Tour selection
- Memory utilization
- Parallel efficiency
- Constructed sequence accuracy

Timeline

- Naiad implementation of streaming graph generation, parallel tour selection, and Bloom filters
- GiRaph implementation of streaming graph generation, parallel tour selection, and Bloom filters
- Benchmarking on current standards (ABySS 2.0, GiGA, Faucet)
- Benchmarking on Illumina HiSeq Data
- Benchmarking on Oxford Nanopore Data
- Write Report