Constructing a De Bruijn Graph for De Novo Assembly

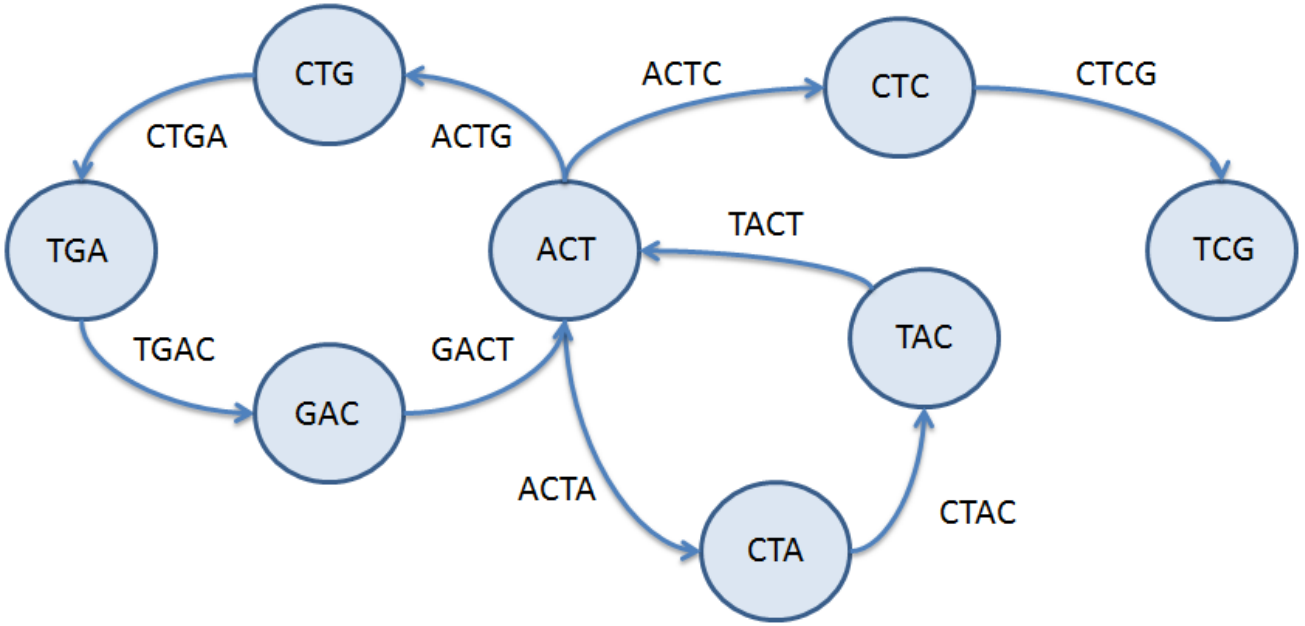
John Tyler
Seung-Jong Park

LSU

LSU | Center for Computation & Technology

The Project

- Genome Assembly
- Construct a De Bruijn Graph
- Intel GraphBuilder and Hadoop



The Algorithm

- Focused on Edges

The Algorithm

- Focused on Edges
- K-mer Length:

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
  - GTTA TTAC TACA

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
  - GTTA TTAC TACA
- Each k+1-mer represents two vertices and the edge between them:

The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
  - GTTA TTAC TACA
- Each k+1-mer represents two vertices and the edge between them:
  - GTT → TTA → TAC → ACA

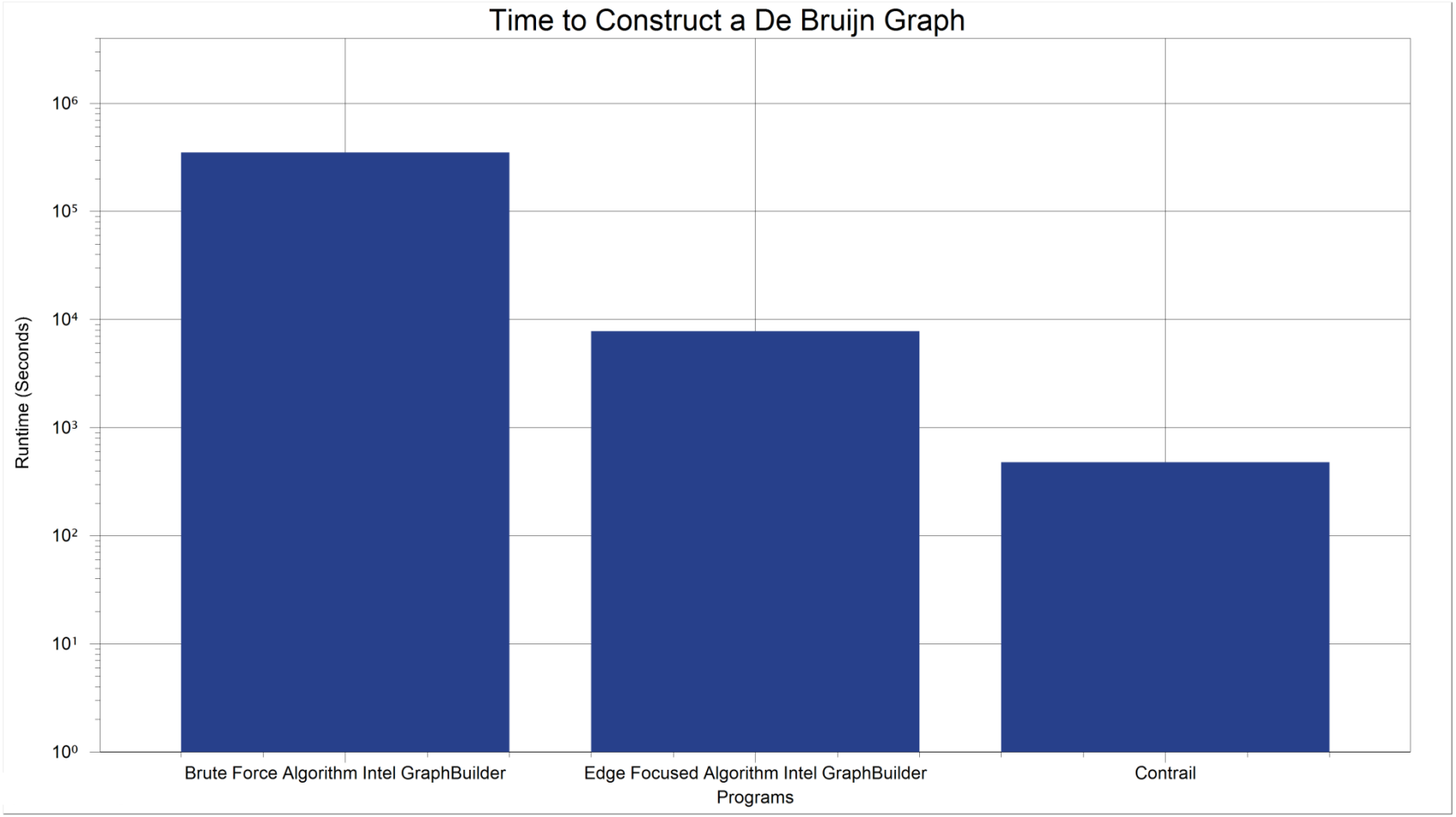
The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
  - GTTA TTAC TACA
- Each k+1-mer represents two vertices and the edge between them:
  - GTT → TTA → TAC → ACA
- Converted to a form GraphBuilder understands

Results

- Brute Force Algorithm
- Edge Focused Algorithm
- Contrail
- 10,000 Short Reads

Results


Time to Construct a De Bruijn Graph
Runtime (Seconds)
10⁶
10⁵
10⁴
10³
10²
10¹
10⁰
Brute Force Algorithm Intel GraphBuilder
Edge Focused Algorithm Intel GraphBuilder
Contrail
Programs

Conclusions

Performance

- Brute Force Algorithm ~ 4 Days
- Edge Focused Algorithm ~ 2 Hours
- Contrail ~ 8 Minutes

Intel GraphBuilder

- Some Significant Bugs
- Documentation Incomplete
- Runtime
- Early in Development

References

• "Contrail – A de Bruijn Genome Assembler that uses Hadoop." *Frontier in Bioinformatics*. Homologus, 8 Sept 2011. Web. 30 Jul 2013. http://www.homolog.us/blogs/blog/2011/09/08/contrail-a-de-bruijn-genome-assembler-that-uses-hadoop/.
• Craven, Mark. "Sequence Assembly." BMI/CS 576. University of Wisconsin. 2011. Web. 11 Jun 2013. http://www.biostat.wisc.edu/bmi576/lectures/assembly.pdf.
• Datta, Kushal, and Ted Willke. "GraphBuilder." *01.org Intel Open Source Technology Center*. Intel, 28 Jun 2013. Web. 25 Jul 2013. https://01.org/graphbuilder/.
• "De Bruijn graph." *Wikipedia*. Wikimedia Foundation, Inc.. Web. 11 Jun 2013. http://en.wikipedia.org/wiki/De_Bruijn_graph.
• "Graph Algorithms in Bioinformatics." *An Introduction to Bioinformatics Algorithms*. bioalgorithms. Web. 11 Jun 2013. http://bix.ucsd.edu/bioalgorithms/presentations/Ch08_GraphsDNAseq.pdf.
• *Hadoop*. Apache, 11 Jun 2013. Web. 11 Jun 2013. http://hadoop.apache.org/.
• Roy , Sushmita. "Sequence Assembly: Concepts." BMI/CS 576. University of Wisconsin. Sept 2012. Web. 11 Jun 2013. http://www.biostat.wisc.edu/bmi576/2012-lectures/SequenceAssembly_v2.pdf.
• Schatz, Michael, , et al, and et al. "Contrail." *SourceForge*. Cold Spring Harbor Laboratory, 19 Sept 2012. Web. 30 Jul 2013. http://sourceforge.net/apps/mediawiki/contrail-bio/index.php?title=Contrail.
• Swami, Nikhil, Nikolaos Frangiadakis, and Konstantinos Bitsakos. "A Distributed Algorithm for Constructing a Generalization of de Bruijn Graphs." . Department of Computer Science University of Maryland. Web. 11 Jun 2013. http://research.microsoft.com/en-us/um/people/nswamy/papers/halo-tr.pdf.
• White, Tom. *Hadoop: The Definitive Guide*. 1st ed. Sebastopol: O'Reilly Media, Inc., 2009. eBook.
• Willke, Theodore, Nilesh Jain, and Haijie Gu. "GraphBuilder - A Scalable Graph Construction Library for Hadoop." . Intel. Web. 11 Jun 2013. https://01.org/graphbuilder/sites/default/files/documentation/graphbuilder-whitepaper.pdf