Genomics: Genome Storage and Assembly

Sijia Huo[2] , Sean Kelly[1] , Gregory Raskind[1]
1. University of Michigan 2. University of Illinois
Big Data Summer Institute

Acknowledgments: Bhramar Mukherjee, Hyun Min Kang, Goncalo Abecasis, Greg Zajac, National Science Foundation

Introduction

- Data size and storage proves to be an obstacle for many genetic researchers
- 3 billion basepair-long genome is broken up into fragments to be sequenced and reassembled
- Fragments of DNA, called "kmers', are represented as k-character strings and stored in De Bruijn graph
- Slight variations of genome between people are responsible for traits can cause disease
- Single nucleotide polymorphisms (SNPs), insertions and deletions, and copy number variations are possible variations
- Determine variations by lack of presence in reference genome
- Variant Graph holds information for long sequences of kmers, helps to find complex variation
- Our project wants to improve upon current methods including Velvet[4] developed by Zerbino and Birney and Cortex[2][3] developed by Iqbal, McVean, and Turner

Methods

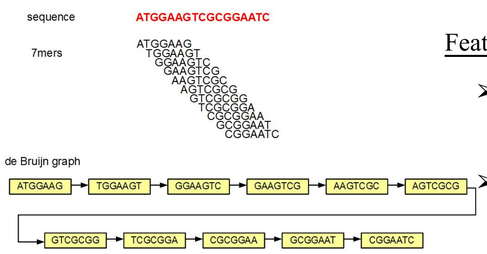


Figure 1: De Bruijn Graph for K = 7[7]

Features of the de Bruijn Graph:

- Large overlap helps with reconstruction of sequence
- Useful to handle repetitive regions

Features of the Variant Graph:

- Holds strings of kmer sequences
- Stores location of variation from reference genome, if found

Figure 2: Example of string variation[1]

Data Source/Format

```
>HSBGPG
GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAG
TGGGCACAGCCCAGAGGGTATAAACAGTGCTGGAGGC
AGTCCTGAGCAGCAGCCCAGCGCAGCCACCGAGACAC
CTCGCCCTATTGGCCCTGGCCGCACTTTGCATCGCTG
CCACCTCCCCTCAGGCCGCATTGCAGTGGGGGCTGAG
CACCTCTTCTCACCCCTTTGGCTGGCAGTCCCTTTGC
AGGCTCAATCCATTTGCCCCAGCTCTGCCCTTGCAGA
AGCTGCCCGAGACGCAGGGGAAGGAGGATGAGGGCCC
ACCAGGCTCCCTTTCCTTTGCAGGTGCGAAGCCCAGC
GTGCAGGTATGAGGATGGACCTGATGGGTTCCTGGAC
CCCTCAGTCTCATTCCCCCACTCCTGCCACCTCCTGT
GCCTGCTCCCCACCTGATCCTCCCAAACCCAGAGCCA
CTCCACAGCCTTTGTGTCCAAGCAGGAGGGCAGCGAG
GCTACCTGTATCAATGGCTGGGGTGAGAGAAAAGGCA
```

Figure 3: FASTA format for reference genome

```
@H3GFVCCXX150415:8:2224:9627:35467/1
GGGAATTTTAACTGGCAAAACTCAGAACTCCATCCAAAC
+
AAAFF<<FFAFAFAAAAFFFAFAAAAAAFAAAFAAFAFA
@H3GFVCCXX150415:8:2224:8957:23407/1
CATACTTGATGGTCTCAGATATGTGTGGATTTTGGAATT
+
<FAFAFFAAAA7AAFAAAAFAAAAAAFAFAFFFAAFFFF.
@H3GFVCCXX150415:8:2224:8907:25745/1
GTTAATTAAAAGCCCTTTACGAATGGACTAGATGTACCT
+
AAAFFAFAF/FFAAAA<FAA7FFAAAFAFAF7FAAFAAF
@H3GFVCCXX150415:8:2224:8825:55175/1
GCACCCTGTGTCAACAACCTGAGAGTGGCCTTGAGTTGC
+
AAFAAAFAAAAAFFAFFAAFAFAAAAFAAAFFFAFAFAA
```

Figure 4: FASTQ format for sequenced reads

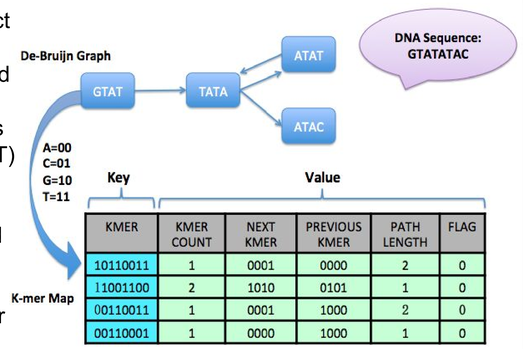
Data Source: Nation Center for Biotechnology Information[5]

Methods

❑Data Structure: KmerGraph

Features:

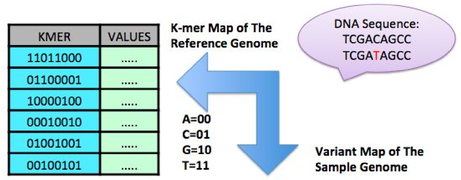
- Kmer sequences act as the keys of bitpacked unordered map
- Encoding four types nucleobases (ACGT) into 2 bit binary numbers.
- Struct can only hold 32 character sequence
- Stores various kmer statistics to assist with queries and assembly



KMER	KMER COUNT	NEXT KMER	PREVIOUS KMER	PATH LENGTH	FLAG
10110011	1	0001	0000	2	0
11001100	2	1010	0101	1	0
00110011	1	0001	1000	2	0
00110001	1	0000	1000	1	0

Figure 5: Data Structure of the Kmer Graph (Map)

❑Data Structure: VariantGraph



KMER	VALUES
11011000	.....
01100001	.....
10000100	.....
00010010	.....
01001001	.....
00100101	.....

KMER	NEXT KMER COUNT OF A	NEXT KMER COUNT OF C	NEXT KMER COUNT OF G	NEXT KMER COUNT OF T
01100011	1	0	0	0
10001100	0	0	1	0
00110010	0	1	0	0
11001001	0	1	0	0

Figure 6: Data Structure of the Variant Graph (Map)

Features:

- Variant Graph holds edge information needed to construct sequences
- Being able to store long sequences allows identification of complex variations

❑Implementation: Functions

Longest Path	Basic Query	Data Storage	String Complement
• Calculate the lengths of non-branching paths starting from every distinct kmers in the KmerGraph in O(N) time. • Print out the lengths and paths of the non-branching sequences starting from any kmer strings. • Store and print the length and the heads of the longest non-branching paths of the graph.	• Most common kmer • Number of nodes (can be split into unique and non-unique nodes) • Lookup of kmer by key	• Insert data from FASTA and FASTQ files into unordered map • Transfer data between binary file and hash table	• Takes in sequence input and returns sequence of complementary strand (can then be used to construct graph for string's complementary strand)

Results

❑Memory Footprint Reduction and Time Efficiency

- Use of bit packed structs reduces memory usage
- Hashing allows for fast access to stored data
- Storage in binary files for compactness
- *O*(n) run time for all functions

❑Constructed De Bruijn graph for reference genome
❑Stored reference graph with k = 32 in 56 GB binary file
❑Implemented function to find longest linear (i.e. non-branching) path through DeBruijn graph

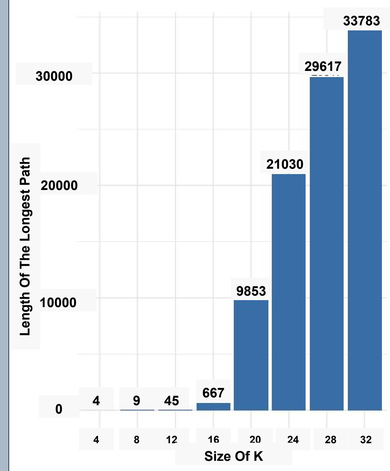


Figure 7: Influence of the size of K on the length of the longest non-branching path

Choosing the right size of K:

- Finding longest path assists in reassembly of genome sequence
- Choice of K has impact on longest path length
- Ideal value of K is uncertain, small k leads to ambiguity in assembly, large k has trouble with sequencing errors

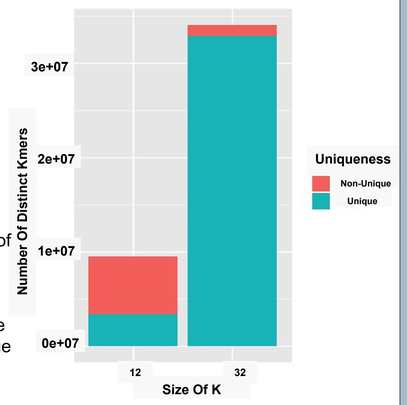


Figure 8: Influence of the size of K on the total number of nodes and the percentage of unique nodes in the de bruijn graph.

Future Work

❑Pruning and error cleaning of variant graph
❑Store 10,000+ genomes as variant graphs to find common variants
❑FM index graph using Burrows-Wheeler Transform to reduce storage size while allowing fast query lookups

References

1. Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and McVean Gil. "De Novo Assembly and Genotyping of Variants Using Colored De Bruijn Graphs." Nature Genetics (2012): n. pag. Nature Publishing Group, 08 Jan. 2012. Web. 11 July 2017.
2. Software For Genome Assembly And Variation Analysis. "Cortex." CORTEX Website. N.p., n.d. Web. 11 July 2017.
3. Iqbal, Zamin, Isaac Turner, and Gil McVean. "High-throughput Microbial Population Genomics Using the Cortex Variation Assembler." Bioinformatics. Oxford University Press, 19 Nov. 2012. Web. 11 July 2017.
4. Zerbino, Daniel R., and Ewan Birney. "Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs." Genome Research, n.d. Web.
5. National Center for Biotechnology Information
6. "The Genome Assembly Problem." Tutorials. N.p., n.d. Web. 11 July 2017.
7. "De Bruijn Graph of a Small Sequence." *Tutorials*. Homologus, 2011. Web. 11 July 2017