GREEDY ALGORITHMS FOR HAPLOTYPE ASSEMBLY

Rachel Ha
Mentored by Dr. Derek Aguiar Ph.D., Assistant Professor of Computer Science and Engineering
University of Connecticut

ABSTRACT

Current experimental methods for determining genetic variation only determine the set of alleles on both chromosomes at each variant site. However, variations that co-occur on a single DNA strand in a chromosome, known as haplotypes, are important for many genetic analyses. Due to the challenging and expensive costs associated with determining haplotypes, the field of bioinformatics has produced computational methods to resolve these obstacles. Current computational methods for determining haplotypes from genome sequence data, called haplotype assembly, are unrealistic for large amounts of data and use restricted optimization that does not produce accurate haplotypes. We present a Python implementation for the HapCompass algorithm that optimizes the greedy minimum weighted edge removal (MWER) problem. The HapCompass algorithm works as a graph in which a node represents a genetic variant and an edge represents supporting evidence from sequence reads for co-occurrence of alleles in a haplotype between the two nodes it connects. We optimized the accuracy of our haplotypes by minimally removing edges and targeting those that are of greatest ambiguity and least confidence. To evaluate our method, we compare haplotype reconstruction against the Levy algorithm and demonstrate encouraging results.

DEFINITIONS:

- **Haplotype:** A set of variations that co-occur on a strand of DNA in a chromosome
- **Haplotype Assembly:** Using small DNA fragments to assemble respective haplotypes

Proper Sequence
Chromosome 1A: C G T C G A A C T G A T
Chromosome 1B: C G T C G A A C T G A T

Variated Sequence
Chromosome 1A: G G T C G A A C T G G T
Chromosome 1B: C G C G G A A C T G A T

HAPCOMPASS ALGORITHM FOR MINIMUM WEIGHTED EDGE REMOVAL:

- A graph in which a node represents variant positions and an edge represents the most likely phasing between the two variants
- A positive edge indicates a **0 0** phasing **1 1**
- A negative edge indicates a **0 1** phasing **1 0**
- Avoid a conflicting graph in which there exists an odd number of negative edges
- Must create a cycle basis by building a maximum spanning tree

RESULTS:

After implementing HapCompass, the algorithm was tested on human genome data. In addition, the Levy algorithm, a previously established haplotype assembly method, was tested.

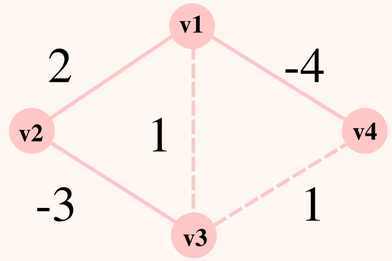
DISCUSSION:

Preliminary results are encouraging as HapCompass is fast and performs well. The HapCompass algorithm focuses on identifying ambiguity (edges of low weight), calculating the optimal phasing at each location, and removing only what is least invasive.

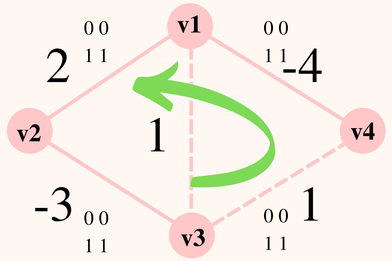
LEVY ALGORITHM:

- There exist 2 essential iterative parts:
 1) Assign reads to temporary haplotypes
 2) Check each read at each location on a majority rules basis

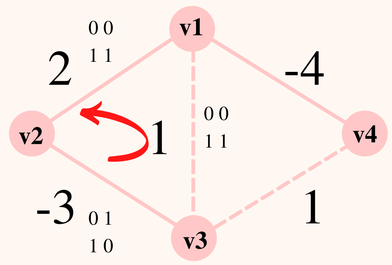
Maximum Spanning Tree:



Happy Graph:



Conflicting Graph:



ACKNOWLEDGEMENTS:

I thank my mentor, Dr. Aguiar, for his endless help and encouragement throughout the process. I also thank Ms. Marjan Hosseini for her engagement in many discussions. Lastly, my deepest gratitude goes to Ms. Pintavalle for leading a class that has taught me much more than I could have ever imagined.

REFERENCES:

Aguiar, D., & Istrail, S. (2012). HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. Journal of Computational Biology, 19(6), 577-590.
Aguiar, D., & Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics, 29(13), i352-i360.
Bishop, C. M. (2013). Model-based machine learning. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20120222.
Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. Annual Review of Statistics and Its Application, 1, 203-232.
Edge, P., Bafna, V., & Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome research, 27(5), 801-812.
Schwartz, R. (2010). Theory and algorithms for the haplotype assembly problem. Communications in Information & Systems, 10(1), 23-38.