Genome Polymorphism Detection Through Relaxed de Bruijn Graph Construction

### C. Guru Aswinidath

Computer science of Engineering

Artificial Intelligence   
Amritapuri

### [amenu4aie21023@am.students.amrita.edu](mailto:amenu4aie21023@am.students.amrita.edu)

### V. Praneeth

Computer science of Engineering

### Artificial Intelligence Amritapuri

### [amenu4aie21066@am.students.amrita.edu](mailto:amenu4aie21066@am.students.amrita.edu)

### G. Venkata Vignesh

Computer science of Engineering

### Artificial Intelligence Amritapuri

### [amenu4aie21074@am.students.amrita.edu](mailto:amenu4aie21074@am.students.amrita.edu)

*Abstract*— Comparing genomes to identify polymorphisms is a difficult task, especially beyond single nucleotide polymorphisms. Polymorphism detection is important in disease association studies as well as in phylogenetic tree reconstruction. We present a method for identifying polymorphisms in genomes by using a modified version de Bruijn graphs, data structures widely used in genome assembly from Next Generation Sequencing. we are able to identify polymorphisms that exist within a genome as well as well as see graph structures that form in the de Bruijn graph for particular types of polymorphisms.

Keywords— de Bruijn graph, graph, GWAS, phylogenetics, polymorphism

*I. Introduction*

Detecting polymorphisms in the genome is an important task for an individual specimen (disease association studies) and for a species as whole (phylogenetic tree reconstruction) . Genomic variation such as translocations and inversions have been shown to cause many human diseases. Translocations have been shown to be the cause of several different types of cancer, such as Burkitt’s lymphoma and acute promyelocytic leukemic . They have also been shown to be associated with schizophrenia . Studying and identifying different types of genomic polymorphisms could have impact on two very important fields in biology:

genome wide association studies

phylogenetic tree reconstruction.

## Genome wide Association studies:

## Genome wide association studies (GWAS), next-generation sequence (NGS) reads are mapped to a reference genome. Differences, commonly SNPs and indels, are then identified from the read mapping results. This method has helped identify and associate many mutations with different diseases Maintaining the Integrity of the Specifications.

Read mapping is a difficult task. More than 10% of reads were unmapped when mapping 12.2 million reads to the human genome using the popular Burrows-Wheeler Aligner. Some of the reads will be left unmapped due to errors generated during sequencing. Other reads are left unmapped for unknown reasons. It may be that some unmapped reads vary significantly from the reference genome making read mapping difficult. Mapped reads represent reads that are similar enough to the reference genome to be mapped with a given set of parameters.

Unmapped reads may contain more interesting and novel biological information than mapped reads because these reads diverge enough from the reference genome to remain unmapped. Harnessing unmapped reads enables more thorough analysis of how individuals within a species differ and how genomic rearrangements may affect phenotypes.

## Phylogenetic tree reconstruction:

Phylogenetic tree reconstruction is often completed through comparing homologous gene sequences in a group of species of interest. Identification of homologous genes is a difficult task and is often a conservative process, allowing for only gene sequences that are very similar to be clustered together . This approach is limited because it only allows for comparing gene sequences instead of comparing whole genomes . Comparing the entire genome of one species to another is valuable to see if genomic rearrangements or other structural variations occurred to the genome.

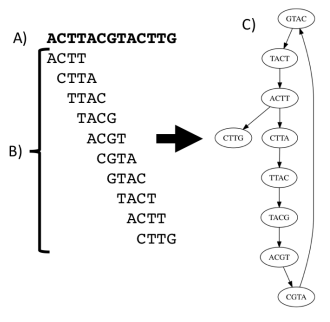
# *Methods*

Our method for utilizing unmapped reads and to compare whole genomes is to construct a relaxed de Bruijn graph that allows for more complex genomic variation.

## Standard de Bruijn Graph

De Bruijn graph is a directed graph that is used in bio informatics to represent the overlaps between sequences of nucleotides or amino acids. A standard de Bruijn graph is a specific type of de Bruijn graph where the sequences are represented by the edges of the graph, rather than the nodes. In a standard de Bruijn graph, the nodes represent k-1-mers, or substrings of length k-1, and the edges represent the nucleotides or amino acids that follow these substrings.

* The edges in a standard de Bruijn graph are directed, with the tail of the edge representing the k-1-mer and the head of the edge representing the next nucleotide or amino acid.
* It can be used to assemble a sequence from a set of reads, by finding a path through the graph that visits each k-mer exactly once.
* It can also be used to detect and correct errors in a sequence, by finding the path through the graph that has the highest coverage or the lowest number of errors.
* The standard de Bruijn graph can also be used to find the repeats in a sequence, by finding cycles in the graph.
* The standard de Bruijn graph can also be used to analyze the genetic variations between different sequences, by comparing the differences between the paths that represent the sequences in the graph.



The above figure shows the construction of a standard de Bruijn Graph. A) The original sequence. B) sequence broken into k mers (k=4) showing k mer overlap. C) A de Bruijn graph with edges formed from overlapping k mers.

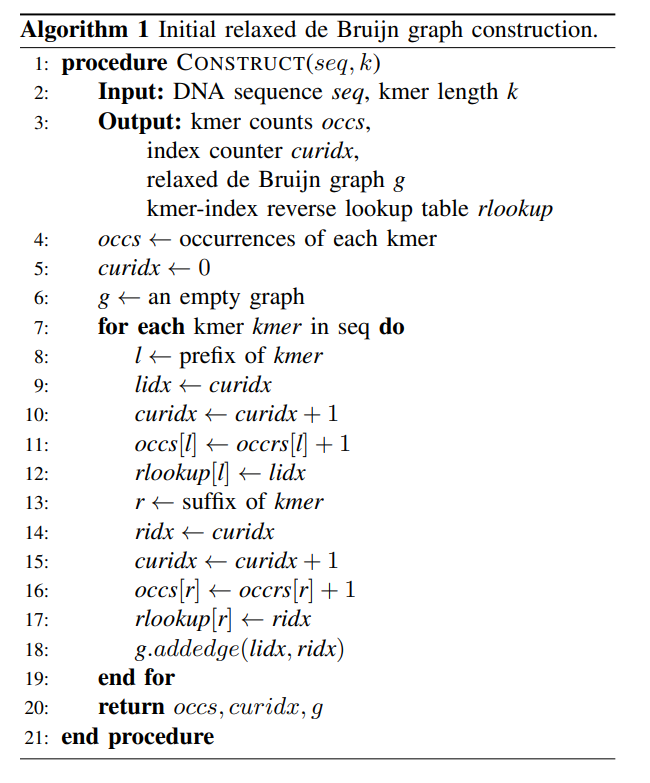
* 1. Relaxed de Bruijn graph

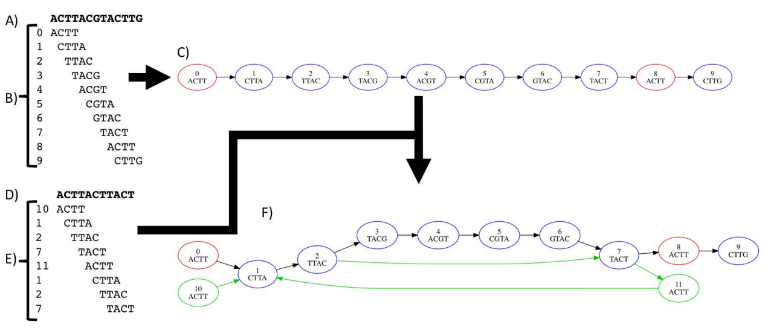
A relaxed de Bruijn graph is a variant of the de Bruijn graph that is used to represent sequences of nucleotides or amino acids in bioinformatics. The main difference between a regular de Bruijn graph and a relaxed de Bruijn graph

1) The graph contains sequence information for multiple species.

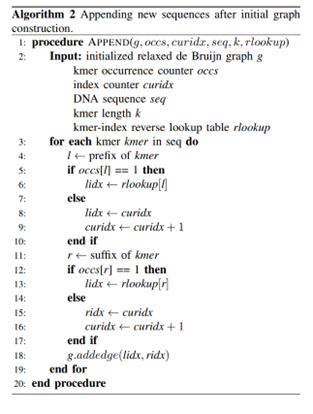
2) Kmers can occur multiple times in the graph

* They are an extension of the de Bruijn graph, where the edges are labeled with k-mers, but allows some errors or ambiguities in the k-mers.
* They are used to assemble DNA or protein sequences from short reads.
* It's more robust to sequencing errors, allowing the assembly of regions that would be difficult to assemble with a traditional de Bruijn graph.
* They are useful in assembling genomes with high levels of repetitive sequences or structural variations.
* They are computationally more complex than de Bruijn graphs and require more memory to store.





Construction method for our relaxed de Bruijn Graph for two reference genome sequences. A and D are two different sequences. B and E represent the sequence broken into kmers and the graph node IDs assigned to each kmer. C is the initial relaxed de Bruijn graph containing only A. Blue nodes are unique k mers and red nodes are non-unique kmers occurring in sequence A. F is the resulting relaxed de Bruijn graph once kmers from sequence D are added. Green nodes and edges are new nodes or edges that were added to the graph. See Algorithm 1 for construction of C and Algorithm 2 for F.



In the simplified graph, the mutation, insertion and inversion form simple bubble structures (graph structure where a node has multiple outgoing edges to other nodes that later merge as incoming edges into another node) in the graph while the translocation forms a much more complex structure. In these very ideal conditions (all kmers in the reference sequence are unique, reads with no errors), the generated graph shows structures that could be used to generated a phylogenetic signal or for phenotype association with additional generated graphs from other individuals.

# 

# *III. CONCLUSION AND FUTURE WORK*

In conclusion, the use of relaxed de Bruijn graphs for genome polymorphism detection is a powerful tool in bioinformatics. The ability to tolerate a certain degree of error or ambiguity in the sequences it represents allows for the assembly of regions that may be difficult to assemble with traditional de Bruijn graphs. This is particularly useful for genomes with high levels of repetitive sequences or structural variations. By constructing a relaxed de Bruijn graph, it is possible to identify variations in the genome, such as single nucleotide polymorphisms (SNPs) or structural variations. However, it should be noted that relaxed de Bruijn graph construction is computationally more complex and requires more memory than traditional de Bruijn graph construction.

*REFERENCES*

[1] R. Hoffman, E. J. Benz Jr, L. E. Silberstein, H. Heslop, J. Anastasi, and J. Weitz, Hematology: basic principles and practice. Elsevier Health Sciences, 2013.

[2] H. Li and R. Durbin, “Fast and accurate short read alignment with burrows–wheeler transform,” Bioinformatics, vol. 25, no. 14, pp. 1754–1760, 2009.

[3] M. S. Fujimoto, A. Suvorov, N. O. Jensen, M. J. Clement, Q. Snell, and S. M. Bybee, “The ogcleaner: filtering false positive homology clusters,” Bioinformatics, vol. 33, no. 1, pp. 125–127, 2016.

[4] L. Li, C. J. Stoeckert, and D. S. Roos, “Orthomcl: identification of ortholog groups for eukaryotic genomes,” Genome research, vol. 13, no. 9, pp. 2178–2189, 2003.

[5] <https://ieeexplore.ieee.org/document/8251291>

[6] A. Limasset, B. Cazaux, E. Rivals, and P. Peterlongo, “Read mapping on de bruijn graphs,” BMC bioinformatics, vol. 17, no. 1, p. 237, 2016. [8] A. Hagberg, P. Swart, and D. S Chult, “Exploring network structure, dynamics, and function using networkx,” Los Alamos National Laboratory (LANL), Tech. Rep., 2008.

[7] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An open source software for exploring and manipulating networks,” 2009. [Online]. Available: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>

[8] https://www.youtube.com/watch?v=lC7kZOOYFpo&ab\_channel=BiologyLectures

[9] https://www.youtube.com/watch?v=PDOMF6Uc3Is&ab\_channel=DrColHarrison

[10]

https://www.ncbi.nlm.nih.gov/gds/?term=E.+coli+K12%2FW3110