Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 2, 2023

Efficient Genome Assembly Studies Using Overlap and Hamiltonian Graphs

Adebayo Femi Samuel[1, *], Diping Zhang[1]

[1] School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China

[*] Corresponding author: Adebayo Femi Samuel (Email: Joshycarter712@gmail.com)

Abstract: These DNA sequencing is the intricate process of deciphering the specific arrangement of nitrogenous bases, also known as nucleotides, within a DNA molecule. Each organism possesses a unique nucleotide sequence that dictates its genetic blueprint, or genome, influencing both physical traits (phenotypes) and hereditary characteristics (genotypes) at the cellular level. In the realm of mathematics, graph theory delves into the study of mathematical constructs called graphs, composed of vertices (nodes) interconnected by either directed or undirected edges. Determining the precise order in which these nucleotides are linked empowers scientists and researchers to compare DNA across organisms, shedding light on their evolutionary relationships. This research delves into the pivotal role of graph theory in genome sequencing, exploring the diverse types of graphs utilized in this process. We propose innovative methods for employing graph theory in DNA sequencing and investigate the application of graphs such as overlap graphs and Hamiltonian graphs in genome sequencing, along with their associated advantages and limitations.

Keywords: DNA, Directed graph, k-mer, Nucleotide, Sequence.

1. Introduction

According to the article "Graph Theory" by Robin J. Wilson [1], published in the Princeton Companion to Mathematics, graphs can be used to represent many different kinds of structures, such as networks of roads or computer connections, relationships between individuals or groups, chemical compounds, or even the structure of the Internet. Graph theory can be used to analyze these structures, study their properties, and solve problems related to them. A graph has points and lines between them. The length of the lines and the position of the points do not matter. Each object in a graph is called a node. A graph 'G' consists of vertices called nodes 'v' connected by edges called links 'e'. Therefore, G = (v, e). A vertex is an intersection point of a graph. It denotes a location such as a city, a road intersection, or a transport terminal (stations, harbors, and airports). An edge is a link that connects two nodes together.

A link denotes movement between nodes. A direction is generally represented by an arrow. Bidirectional links are indicated by the absence of an arrow. The purpose of this article is to present several methods used during genome sequencing and how graph theory concepts play a part in it. The article by Wilson, R. J. (2008) [2] explains that graphs can be classified into different types, such as directed or undirected graphs, weighted or unweighted graphs, and bipartite or planar graphs.

In graph theory, a directed graph (also called a digraph) (Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C, 2009)[3] and West, D. B. (2001)) [4] is a graph where each edge has a direction, pointing from one vertex to another. In contrast, an undirected graph is a graph where each edge is bidirectional, meaning it connects two vertices without any direction.

To illustrate the difference, consider the following two graphs(see Figure 1), in the directed graph, the edges have arrows that indicate their direction. For example, there is an edge from vertex 1 to vertex 2, but no edge from vertex 2 to vertex 1. In contrast, the edges in the undirected graph do not have arrows, and each edge can be traversed in both directions.

Directed and undirected graphs have different properties and applications in various fields, including computer science, mathematics, physics, and social sciences. For example, directed graphs are commonly used to model networks with a flow or directionality, such as transportation networks or the internet. In contrast, undirected graphs are often used to model symmetric relations, such as social networks or molecular structures. For the purposes of this study, we are sorely going to focus on directed graphs as DNA sequencing uses directed graphs for sequencing.

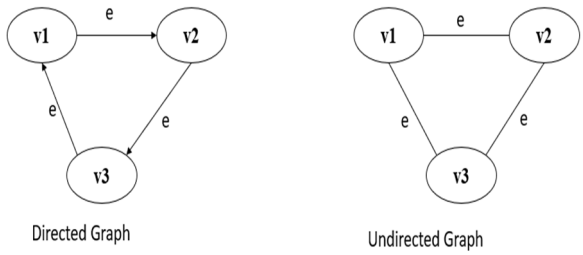

Figure 1. Directed and Undirected Graphs

The double helix structure of DNA was first proposed by James Watson and Francis Crick in 1953 , based on X-ray crystallography data obtained by Rosalind Franklin and Maurice Wilkins. On the diagram (see Figure 2), illustrates the molecular structure of DNA. The discovery of the structure of DNA was a landmark event in the history of molecular biology, and it paved the way for further research on the genetic code and the mechanisms of inheritance. Graph theory has become a fundamental (Pevzner, P. A., Tang, H., & Waterman, M. S, 2001) [5] tool in the analysis of genome sequencing data, as it allows researchers to represent and study the complex relationships between the sequences that make up a genome. The application of graph theory to genomics has led to the development of a range of powerful

algorithms and tools for DNA assembly, alignment, annotation, and visualization.

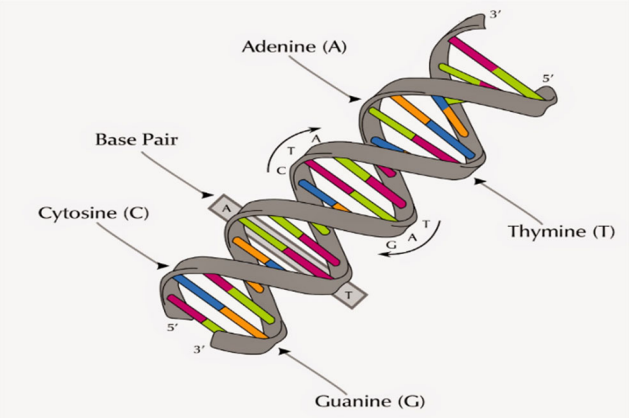

Figure 2. The molecular structure of DNA

2. Literature References

Format The year 1953, two scientists, J.D. Watson and F.H.C. Crick established the double-helix model for the DNA molecule after combining chemical and physical data. DNA is a short name for DeoxyriboNucleic Acid and according to this proposed model, the DNA molecule is made out of two antiparallel strands which are connected together by two or three hydrogen bonds and helically twisted. Within these nucleotides encode the genetic information of all living matter, the human beings included.

There are four different types of nucleotides bases in DNA which are guanine (G), thymine (T), adenine (A) and cytosine (C). Within these bases adenine bonds with thymine and guanine bonds with cytosine . DNA sequencing hence is the process of figuring out in which order are these nucleotides bases arranged in the genome. Rapid advancements in genome sequencing have made understanding genome sequencing essential for many biological studies, other research areas that use DNA sequencing and a variety of applied fields like biotechnology, forensic biology, and diagnostics.

In 1977, Frederick [6] and his colleagues proposed a method based on chain-termination inhibitors (Sanger, Nicklen, and Coulson, 1977), which marked the beginning of the DNA sequencing journey. For the best validation in the field of genetics, Sanger sequencing is known to provide 99.99 percent base accuracy. When it comes to understanding how genes carry and store information, it is regarded as the "gold standard" according to this article (The Genomic Services Company, 2020). The Human Genome Project used Sanger sequencing to identify the sequences of relatively short (900 bp or less) segments of human DNA. Larger DNA fragments and eventually entire chromosomes were put together using these pieces of DNA. A new DNA sequencing approach was introduced by Edwin Southern where the DNA sequencing is done by hybridization (SBH) (Southern, 1988) [7]. In this approach, an overlapping collection of oligonucleotide sequences is assembled together with the aim to determining an organism's DNA sequence.

With high efficiency of sequencing by hybridization (SBH), many scientists can obtain stored information on the genomes of variety of organism and species which will help advance the futuristic development of biological medicine, agriculture and sciences. We can distinguish Y.P. Lysov [8] who formulated the problem as finding a Hamiltonian path (Lysov et al., 1988) with his colleagues and Pevzner (1989) [9] who formulated the problem as finding a Eulerian path from the group of scientists of algorithmic approaches to sequencing by hybridization. . In 1999, Ludry and Waterman presented an algorithm for DNA sequencing by hybridization (SBH) by using concepts of graph theory [10].

Using Next Generation Sequencing is yet another method for genome sequencing, and it has proven to be a very powerful platform with the capability of sequencing thousands or millions of DNA molecules simultaneously (Margulies, Egholm, & Altman, 2005) [11]. It is clear that the methods of sequencing have become a game-changer in the field of biology and medicine in the modern era. In addition to accelerating biological research and discovery, DNA sequencing has also advanced medical diagnostics and the treatment of diseases, which is one of the key benefits of genome sequencing.

When it comes to advances in DNA Sequencing, the field of genome sequencing has undergone significant advancements in recent years, resulting in increased accuracy, reduced costs, and improved speed. The development of next-generation sequencing technologies has made it possible to sequence an entire genome in a matter of days, at a fraction of the cost of previous methods. This has led to a significant increase in the number of sequenced genomes, enabling researchers to study genetic variation on a large scale(Goodwin, S., McPherson, J. & McCombie, W., 2016) [12]. DNA sequencing has numerous applications in the fields of medicine, biology, and agriculture. In medicine, genome sequencing is used to diagnose genetic diseases, predict disease susceptibility, and develop personalized treatments (Ashley E. A. ,2015) [13]. In agriculture, DNA sequencing is used to improve crop yields, develop disease-resistant strains, and improve food security (Varshney RK, Terauchi R, McCouch SR , 2014) [14].

Despite the many advances in DNA sequencing, there are still some challenges and limitations that need to be addressed. One of the biggest challenges is the analysis of large amounts of sequencing data, which requires sophisticated computational tools and algorithms (Duan, J., Shi, J., & Ge, H., 2013) [15]. Another challenge is the interpretation of genetic variants, as many variants are of unknown significance and their effects on health are still not well understood (Richards, S., Aziz, N, 2015) [16]. DNA sequencing has revolutionized the field of genetics and has numerous applications in medicine, biology, and agriculture. With continued advancements in technology and analysis tools, DNA sequencing is poised to play an increasingly important role in personalized medicine, disease diagnosis, and treatment, and in improving food security.

3. Overlap Graph-based DNA Sequencing

An Overlap Graph also called OG, assuming we have a set of finite words can be described as a complete weighted digraph having each word as a node and the weight of each arc is equal to the length of the longest overlap of one word in the node to the other. This overlap is an asymmetric notation. In short, an OG of G will be a complete directed graph, weighted on its arcs, whose nodes are the words of G, and in which the weight of an arc (u, v) equals the length of the maximum overlap from string u to string v. The Overlap graph is used to reconstruct genome fragments or to compute

shortest superstrings, which are a compressed representation of the input.

The Overlap graph requires space that is quadratic in the number of words, which limits its scalability. The Hierarchical Overlap Graph (HOG) is an alternative graph that also encodes all maximal overlaps, but uses space that is linear in the sum of the lengths of the input words. Building this graph requires one to compute the weights of the arcs by solving the so-called All Pairs Suffix Prefix overlaps problem (APSP) on G. Overlap graphs can be constructed by aligning reads to a reference genome, or by using de novo assembly methods to create a graph from scratch (Myers E. W, 2005) [17]. Once the overlap graph is constructed, it can be traversed to generate a consensus sequence that represents the complete genome (Zerbino, D. R., & Birney, E., 2008) [18]. One of the challenges of overlap graph assembly is dealing with sequencing errors, which can introduce false overlaps and complicate the graph structure. To address this, several approaches have been developed to filter out low-quality reads and correct sequencing errors before constructing the overlap graph (Koren, S., & Phillippy, A. M. , 2015) [19].

Example 1: Let G ={TACGAT, GTACGT, ACGTAC, GTACGA, CGTACG, TACGTA} be a multiset of all 6-mers from long nucleotides of a DNA sequence and the edges be overlaps of length ≥4. For the threshold of this example, the suffix/prefix match will be an exact match and has a length of at least 4.

The diagrams (see Figure 3 to Figure 6) illustrate how prefix to suffix overlaps occurs through each node as long as overlaps of length ≥4. Each node is a distinct nucleotide sequence from a genome read and by using an edge connecting all distinct node that have overlaps will result in the below diagrams. From the diagrams above the weight on each node is shown above, some node has weight of 4 and most have of weight 5.

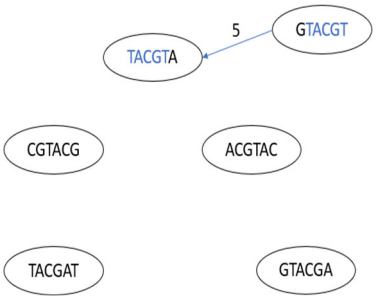

Figure 3. First Overlap

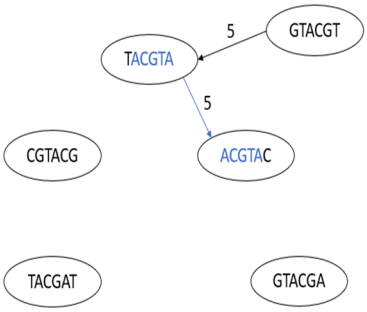

Figure 4. Second Overlap

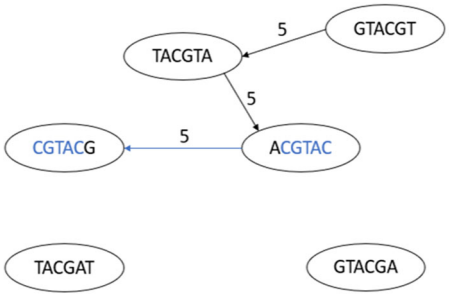

Figure 5. Third Overlap

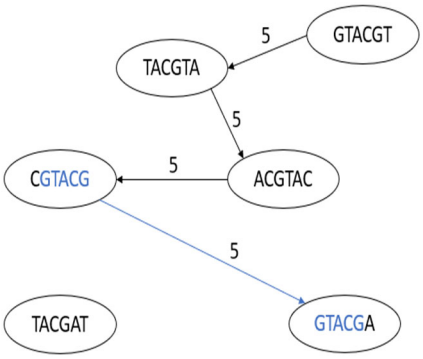

Figure 6. Forth Overlap

Overlap graphs have been used extensively in genome sequencing, with applications in both prokaryotic and eukaryotic genomes. In prokaryotic genomes, overlap graph assembly is often used to generate high-quality, closed genomes from short-read sequencing data (Rhoads, A., & Au, K. F., 2015) [21]. In eukaryotic genomes, long-read sequencing technologies such as PacBio and Oxford Nanopore have enabled the construction of more complex overlap graphs that can span repetitive regions and structural variations (Bondy, J.A. and Murty, U.S.R., 2008) [22].

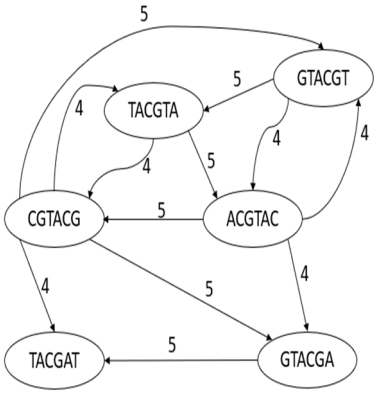

Figure 7. Completed OG of G

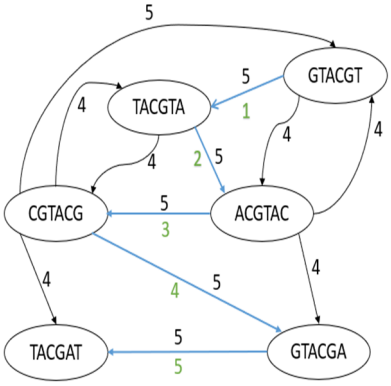

Figure 8. Completed OG of G with Path line

After completing connecting the 6-mers reads for this Overlap graph, at the end we'll come out with the diagram above (see Figure 7) clearly showing all the overlaps nodes joined together with a directed edge from prefix to suffix. The weight of this Overlap graph is shown on either the right side or the top part of the edge. Finally, for reconstructing the DNA genome, we walk through the overlap graph following each directed edge creating a path.

This walk is shown above(see Figure 8), where we can deduce the DNA genome sequence for our DNA. The walk will result in the table below(see Table 1). In conclusion, overlap graphs are a powerful tool in DNA sequencing, allowing the assembly of complete genomes from overlapping reads of DNA fragments. Their application extends beyond genome sequencing to other genomic applications such as metagenomics and transcriptomics, making them a versatile and valuable tool in genomic research.

Table 1. DNA sequence reconstruction of G

	G	T	C	T			
		T	C	T	A		
			C	T	A	G	
			C	T	A	G	
				T	A	G	T
				T	A		
DNA	***G***	***T***	***C***	***T***	***A***	***G***	***T***

From the table above (see Table 1), we can clearly see from the walk and reconstruct the original genome sequence from the overlaps as GTACGTACGAT. The OG has several drawbacks. First, it is not possible to know whether two distinct arcs represent the same overlap. Second, the OG has an inherently quadratic size since it contains an arc for each possible (directed) pairs of words. Overlap graphs have been a popular approach to genome assembly for many years, but they have limitations when it comes to handling repeat regions and errors in the data. To address these limitations, researchers have recently turned to Hamiltonian graph-based approaches. Subsequent research has explored different variations of the string graph algorithm and other Hamiltonian path-based approaches for DNA assembly.

4. A Hamiltonian Graph Approach to Genome Sequencing

The In this section, we show how the DNA is sequenced using Hamiltonian approach. If there is a closed walk in a connected graph that passes every vertex of the graph exactly once, with the exception of the root or starting vertex, the graph is referred to as a Hamiltonian graph. The Hamiltonian walk must not repeat any edges. Another definition of a Hamiltonian graph states that if a graph is connected and Hamiltonian circuit exists, the graph in question is said to be a Hamiltonian graph. A set of points called the vertex is what makes up a graph. These points are connected by a set of lines called the edges. An exact one-time pass through each vertex characterizes a Hamiltonian walk-in graph G.

We first show the very famous theorems for Hamiltonian graph:

Dirac's Theorem - If G is a simple graph with n vertices, where $n \geq 3$, If $\deg(v) \geq n$ for each vertex v, then the graph G is Hamiltonian graph .(Dirac, G.A., 1952) [23]

Ore's Theorem - If G is a simple graph with n vertices, where $n \geq 2$, if $\deg(x) + \deg(y) \geq n$ for each pair of non-adjacent vertices x and y, then the graph G is Hamiltonian graph. (Ore, O., 1960) [24]

Chvátal's Theorem: Let G be a simple graph with n vertices ($n \geq 3$) such that for every set S of k vertices ($1 \leq k \leq n/2$), the number of vertices adjacent to at least one vertex in S is at least k. Then G is Hamiltonian. (Chvátal, V., & Erdös, P., 1972) [25]

Bondy-Chvátal Theorem: Let G be a simple graph with n vertices ($n \geq 3$) such that for every non-empty proper subset S of vertices, the number of components in the subgraph induced by the vertices in V-S is at most |S|. Then G is Hamiltonian. (Bondy, J.A. and Chvátal, V., 1976) [26]

Dirac-Fan Theorem: If G is a simple graph with n vertices ($n \geq 3$) such that for every pair of non-adjacent vertices u and v, the sum of their degrees is at least n-1, then G is Hamiltonian. (Fan, K., 1964) [27]

These theorems provide conditions that guarantee the existence of a Hamiltonian cycle in a graph, which can be used to solve practical problems related to graph theory.

Objective: Use overlapping DNA reads in order to reconstruct the original genome sequence.

When having our fragments of the genome they often overlap. We are able to make use of this overlap and stitch them together. Assuming our fragments (often referred as mers) are 3 molecules long (3-mer). For instance, we could have fragments such as AAT, GCG, CAA. By also assuming they overlap with two molecules. This means the fragment AAT must be followed by a fragment beginning with AT e.g., ATT.

We create a Hamiltonian graph where each node is a fragment. And there is an edge going from a node to another when they only overlap by two nucleotides bases. So, the node AAT would have an edge connecting it to ATT.

Example 2: Let H= {TGC, TTC, GCT, TCC, CTA, CCA, TAG, CAA, AGT, GTT, AAT, TTT, ATA} be a multiset of all 3-long nucleotides of a DNA sequence. From given reads of genome above, lets reconstruct the original gene sequence using Hamiltonian cycle, constructing a network that represents the overlap information in our genome reads will be illustrated as the diagrams below (see Figure 9).

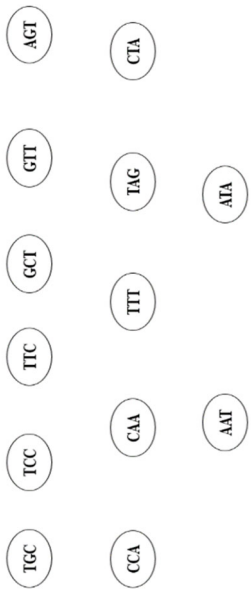

Figure 9. DNA reads as node.

From diagram above (see Figure 9) , as a first step again in using Hamiltonian approach, the 3-long nucleotides of a DNA sequence have been expressed as node and all the genome reads are distinct nodes without any repetition.

By completing connecting all the node (see Figure 12), as the first vertex overlaps with the k-1 leftmost nucleotides of the second vertex (see Figure 13), Complete Hamiltonian graph of H, is created. The graph can also be re-arranged into this graph.(see Figure 14).

On diagram (see Figure 10), a relationship between two nodes TGC & GCT is established and a directed edge is drawn from TGC to node GCT because it satisfies that the k-1 rightmost nucleotides from the first vertex overlap with the k-1 leftmost nucleotides of the second vertex.

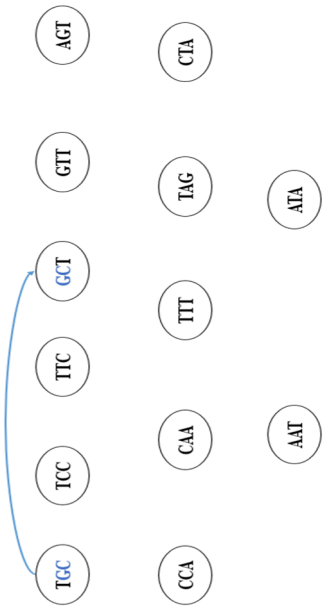

Figure 10. Directed edge between node TGC & GCT.

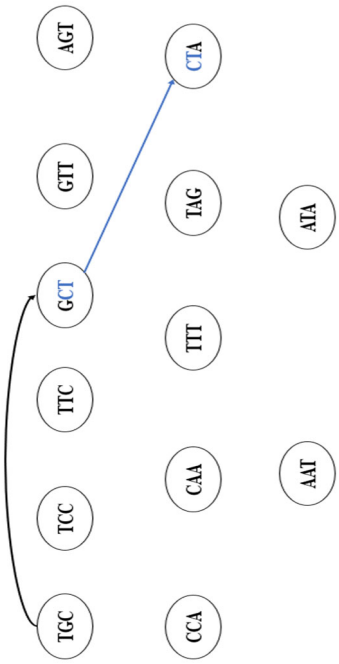

Figure 11. Directed edge between node TGC & GCT.

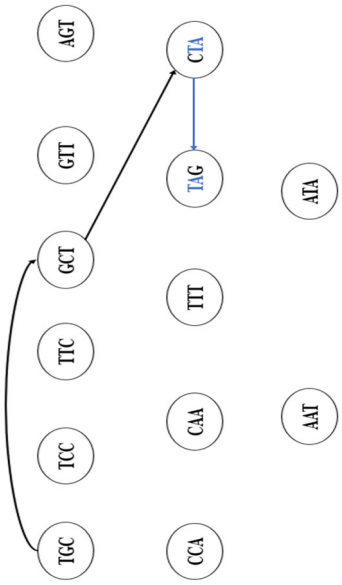

Figure 12. Continued connecting nodes together of graph of H.

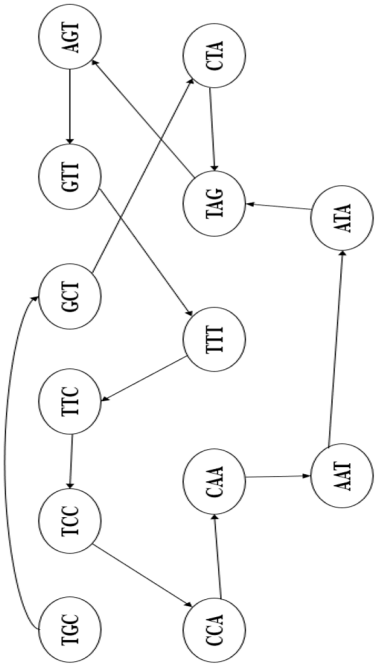

Figure 13. Complete Hamiltonian graph of H.

Therefore, from the re-arranged graph in Figure 14, Graph H has Hamiltonian path:
TGC→GCT→CTA→TAG→AGT→GTT→TTT→TTC→TCC→CCA→CAA→AAT. From reconstructing Graph H, our genome is TGCTAGTTTCCAAT. By finding a path that visits every node once (a Hamiltonian path) we have a found an ordering of the fragment that makes up the whole DNA sequence. Sadly, finding a Hamiltonian path isn't easy (it is classed as an NP-Complete problem).

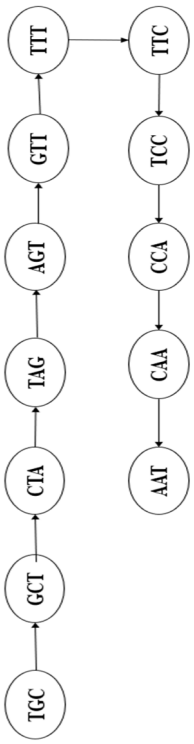

Figure 14. Re-arranged Hamiltonian path of H.

The goal of genome sequencing using this approach is to find a Hamiltonian path through the graph, which represents a complete sequence of the genome. One of the earliest applications of the Hamiltonian approach in genome sequencing was the work of Pevzner and Waterman in 1989. They proposed a combinatorial algorithm for genome sequencing that used the Hamiltonian path approach. The algorithm was based on the observation that the genome can be represented as a de Bruijn graph, which is a directed graph where each node represents a k-mer (a sequence of k nucleotides), and there is an edge between two nodes if they overlap by k-1 nucleotides.

5. Conclusion

DNA sequencing using overlap graphs and Hamiltonian graphs offers several advantages over traditional methods. Both approaches utilize the inherent relationships between overlapping DNA fragments to reconstruct the complete genome sequence. Overlap graphs, as the name suggests, represent DNA reads as nodes and their overlaps as edges. This visual representation allows for efficient identification of redundant information and assembly of overlapping fragments. However, overlap graphs can become quite complex for large genomes, making it computationally challenging to find the optimal path through the graph.

Hamiltonian graphs, on the other hand, provide a more streamlined approach by directly encoding the sequencing reads and their overlaps into the graph structure. This simplifies the task of finding the optimal path, which corresponds to the complete genome sequence. Additionally, Hamiltonian graphs can handle repetitive sequences more effectively, which is a common challenge in genome assembly. While both overlap graphs and Hamiltonian graphs offer advantages in genome sequencing, there are also some key differences between the two approaches. Overall, overlap graphs are more versatile and can be applied to a wider range of sequencing data. However, Hamiltonian graphs are more efficient for large genomes and can handle repetitive sequences more effectively.

The choice between overlap graphs and Hamiltonian graphs for genome sequencing depends on the specific characteristics of the sequencing data and the desired level of accuracy and efficiency. For small genomes with minimal repetitive sequences, overlap graphs may be sufficient. However, for large genomes with complex repetitive regions, Hamiltonian graphs offer a more powerful and efficient solution.

References

[1] Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 171(4356), 737-738.

[2] Wilson, R. J. (2008). Graph theory. In The Princeton companion to mathematics (pp. 602-609). Princeton University Press.

[3] Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to algorithms (3rd ed.). MIT Press. Chapter 22.

[4] West, D. B. (2001). Introduction to graph theory (2nd ed.). Prentice Hall. Chapter 2.

[5] Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America, 98(17), 9748–9753.

[6] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA Sequencing with Chain Terminating Inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74, 5463-5467.

[7] Southern, E. (1998) Analyzing Polynucleotide Sequences. International Patent Application PCT/GB89/00460.

[8] Khrapko KR, Lysov YuP, Khorlin AA, Ivanov IB, Yershov GM, Vasilenko SK, Florentiev VL, Mirzabekov AD. (1988) A method for DNA sequencing by hybridization with oligonucleotide matrix. DNA Seq. 1991;1(6):375-88.

[9] Pevzner P. A. (1989). 1-Tuple DNA sequencing: computer analysis. Journal of biomolecular structure & dynamics, 7(1), 63–73.

[10] Idury, R. M., & Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. Journal of computational biology : a journal of computational molecular cell biology,

[11] Margulies, M., Egholm, M., Altman, W. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380

[12] Goodwin, S., McPherson, J. & McCombie, W. (2016)Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351.

[13] Ashley E. A. (2015). The precision medicine initiative: a new national effort. JAMA, 313(21), 2119–2120.

[14] Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. PLoS Biol 12(6): e1001883.

[15] Duan, J., Shi, J., & Ge, H. (2013). Challenges in genome sequencing analysis. Genomics, Proteomics & Bioinformatics, 11(5), 317-323

[16] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., ... & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics

and Genomics and the Association for Molecular Pathology. Genetics in Medicine, 17(5)

[17] Myers E. W. (2005). The fragment assembly string graph. Bioinformatics (Oxford, England), 21 Suppl 2, ii79–ii85.

[18] Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research, 18(5), 821–829.

[19] Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current opinion in microbiology, 23, 110–120.

[20] Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION:

[21] Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. Genomics, proteomics & bioinformatics, 13(5), 278–289.

[22] Bondy, J.A. and Murty, U.S.R. (2008) Graph Theory. Springer, New York.

[23] Dirac, G.A. (1952), Some Theorems on Abstract Graphs. Proceedings of the London Mathematical Society, s3-2: 69-81.

[24] Ore, O. (1960). Note on Hamilton Circuits. The American Mathematical Monthly, 67(1), 55–55.

[25] Chvátal, V., & Erdös, P. (1972). A note on Hamiltonian circuits. Discret. Math., 2, 111-113.

[26] Bondy, J.A. and Chvátal, V. (1976) A Mothod in Graph Theory. Discrete Mathematics, 15, 111-135.

[27] Fan, K. (1964). Some theorems on the Hamiltonian circuits. Acta Mathematica Sinica, 14(2), 139-149.