International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 3, No. 1, 2023

Graph Theory in DNA Sequencing: Unveiling Genetic Patterns

Banda Ashton

School of Science, Zhejiang University of Science and Technology, Hangzhou, 310023, China

Abstract: Graph theory, a branch of mathematics that studies the properties and relationships of graphs, has emerged as a powerful tool in addressing the complexities of DNA sequencing. This paper highlights the application of graph theory in DNA sequencing and its implications in various aspects of genomics research. One fundamental concept in graph theory applied to DNA sequencing is the construction of the de Bruijn graph. This graph represents overlapping k-mers, subsequences of length k, as nodes, with edges connecting adjacent k-mers. By constructing and analyzing the de Bruijn graph, researchers can infer the underlying DNA sequence, detect errors, resolve repetitive regions, and identify structural variations in the genome. Graph algorithms such as Overlap path, Eulerian path and Hamiltonian path have been adapted to reconstruct complete DNA sequences from fragmented reads obtained through sequencing. These algorithms leverage the connectivity information present in the de Bruijn graph to traverse and assemble the reads, enabling the reconstruction of long DNA sequences accurately. the application of graph theory in DNA sequencing has revolutionized the field of genomics by providing powerful computational tools for DNA assembly, sequence analysis, and functional annotation. By leveraging graph theory concepts and algorithms, researchers can unravel the intricate information embedded within DNA sequences, leading to deeper insights into the genetic basis of life and its applications in various fields, including medicine, agriculture, and evolutionary biology. Future developments in graph-based algorithms and computational techniques hold promise for further enhancing our ability to unlock the secrets encoded within the vast realm of DNA sequences.

Keywords: Overlap Graph; K-mer; de bruijn; Nucleotide; Reconstruction; Sequence.

1. Literature Review

In 1953, two scientists, J.D. Watson and F.H.C. Crick [1], formulated the double-helix model for the DNA molecule by combining chemical and physical data. DNA, an abbreviation for DeoxyriboNucleic Acid, consists of two antiparallel strands connected by two or three hydrogen bonds, forming a helical structure. These strands encode the genetic information of all living organisms, including humans. DNA is composed of four nucleotide bases: guanine (G), thymine (T), adenine (A), and cytosine (C). In this model, adenine pairs with thymine, and guanine pairs with cytosine. Genome sequencing, therefore, entails determining the arrangement of these nucleotide bases within the genome. The rapid advancements in genome sequencing have made it crucial for various biological studies, as well as research fields such as biotechnology, forensic biology, and diagnostics. Understanding genome sequencing has become indispensable in these applied disciplines. Edwin Southern introduced a novel method for genome sequencing called hybridization-based sequencing (SBH) in 1988 [2].

This technique involves assembling a set of overlapping oligonucleotide sequences with the objective of determining the genome sequence of an organism. With the highly efficient technique of sequencing by hybridization (SBH), scientists can now access stored genetic information from a wide range of organisms and species. This breakthrough has tremendous potential to advance the fields of biological medicine, agriculture, and sciences in the future. Noteworthy contributions in this area include the work of Y.P. Lysov and his colleagues, who in 1988 formulated the problem of finding a Hamiltonian path [3], as well as Pevzner's formulation of the problem as finding an Eulerian path in 1989 [4]. These scientists, along with other researchers in the field, have developed algorithmic approaches to sequencing by hybridization. In 1999, Ludry and Waterman further contributed to the progress by presenting an algorithm for DNA sequencing using the concepts of graph theory [5].

Next Generation Sequencing (NGS) emerges as a highly effective approach for genome sequencing, offering immense power to simultaneously sequence thousands or even millions of DNA molecules (Margulies, Egholm, & Altman, 2005) [6]. Undoubtedly, sequencing methods have revolutionized the realms of biology and medicine in the contemporary era. Not only does DNA sequencing expedite biological research and exploration, but it has also greatly enhanced medical diagnostics and disease treatment, showcasing one of the pivotal advantages of genome sequencing. "Genome Assembly Using Graph Algorithms" by Pevzner et al. (2001) (7), this seminal paper introduced the concept of using graph algorithms, particularly the De Bruijn graph, for genome assembly from short DNA sequencing reads. It laid the foundation for subsequent research in graph-based assembly methods. "ABySS: A parallel assembler for short read sequence data" by Simpson et al. (2009) (8), this paper presented ABySS, a popular graph-based genome assembly algorithm that uses De Bruijn graphs to reconstruct genomes from short DNA sequencing reads. It introduced efficient parallelization techniques to handle large-scale sequencing datasets. "Genome-scale algorithm design: Biological sequence analysis in the era of high-throughput sequencing" by Berger et al. (2011) (9), this review article provides an overview of various graph-based algorithms and their applications in DNA sequencing, including sequence assembly, variant detection, and comparative genomics. It highlights the challenges and opportunities associated with graph-based approaches in the context of high-throughput sequencing technologies.

"Genome Graphs and the Evolution of Genome Inference" by Paten et al. (2017) (9), this paper discusses the concept of genome graphs, which represent genetic variation across individuals or populations as graphs. It explores the advantages of using graph-based methods in variant detection and genotyping, and highlights the potential of graph genome reference structures for capturing complex genomic variation. "GraphAligner: Rapid and Versatile Sequence-to-Graph Alignment" by Siren et al. (2019) (10), this paper introduces GraphAligner, a fast and flexible tool for aligning DNA sequences to genome graphs. It enables accurate mapping of sequencing reads to complex graph structures, allowing for efficient variant detection and genotyping. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype" by Kim et al. (2019) (11), this paper presents HISAT2 and HISAT-genotype, a pair of graph-based tools for genome alignment and variant calling. They leverage a graph-based index structure to efficiently align sequencing reads to a reference graph, enabling accurate genotyping and variant detection. Genome 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 [12]. In agriculture, genome sequencing is used to improve crop yields, develop disease-resistant strains, and improve food security [13]

2. Utilizing an Overlap Graph for DNA Sequencing

In DNA sequencing, an overlap graph is a graphical representation of the relationship between different DNA fragments. It is used to determine the order and orientation of these fragments in order to reconstruct the original DNA sequence. The process of DNA sequencing typically involves breaking down the DNA molecule into smaller fragments and sequencing each fragment individually. These fragments are then assembled to reconstruct the complete sequence. However, the challenge lies in determining the correct order and orientation of the fragments. An overlap graph helps address this challenge by identifying overlapping regions between the fragments. If two fragments have overlapping regions, it suggests that they are adjacent in the original DNA sequence. The overlap graph represents this relationship by connecting the fragments with an edge. 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. After the correct path is identified, the fragments can be assembled into the original DNA sequence based on their order and orientation. The resulting sequence represents an approximation of the original DNA molecule. In Overlap graph each node is a read e.g., GCTCTAGCCCCTCATTT. Therefore, we draw a directed edge A → B only when the suffix on A overlaps prefix of B.

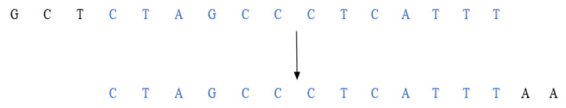

Figure 1. Illustration of prefix/suffix in Overlap graph

Example 1: Consider the multiset G, which consists of the following 6-mers derived from long nucleotides in a DNA sequence: {TACGAT, GTACGT, ACGTAC, GTACGA, CGTACG, TACGTA}. The edges in this example represent overlaps of length 4. In terms of the threshold for this scenario, a suffix/prefix match is defined as an exact match with a minimum length of 4.

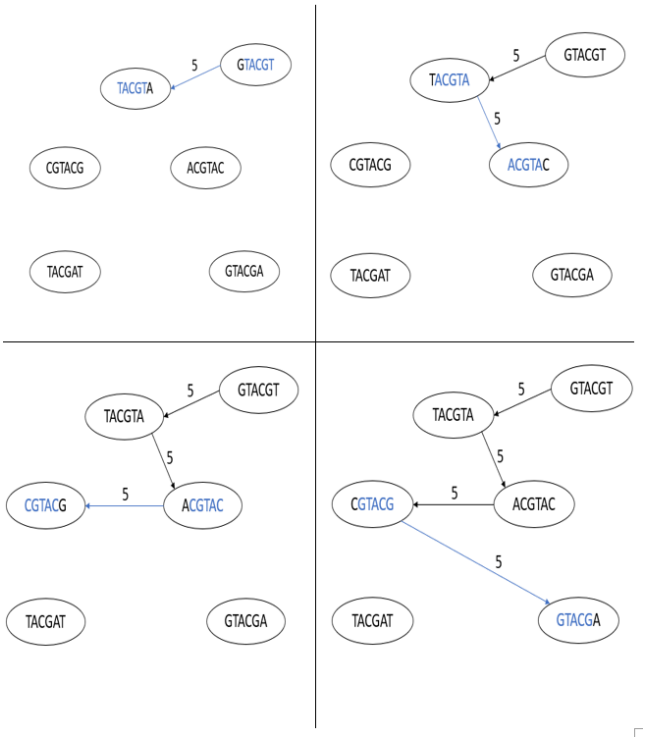

Figure 2. Reconstruction of Overlap Graph Steps

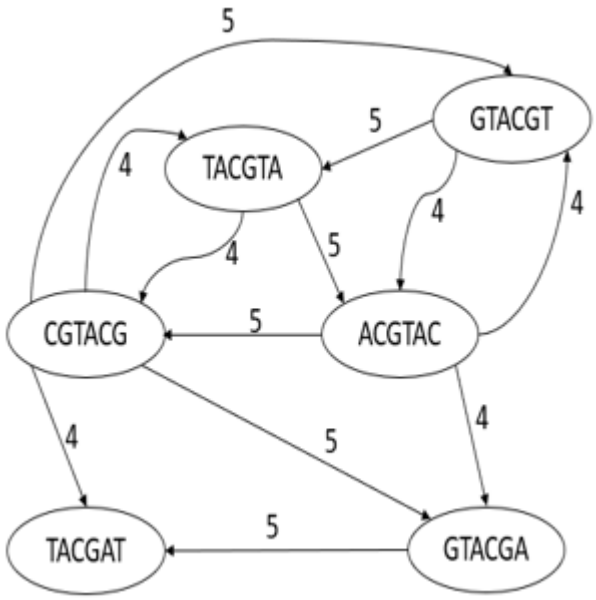

Figure 3. Completed Overlap Graph of Set G

The above diagrams (refer Figure 2) depict the occurrence of prefix to suffix overlaps across nodes, provided that the overlaps are of a minimum length of 4. Each node represents a unique nucleotide sequence derived from a genome read, and connecting edges between nodes indicate the presence of overlaps. The resulting diagrams are presented below. In these diagrams, the weight assigned to each node is displayed above it, with some nodes having a weight of 4 and most having a weight of 5. Overlap graphs have gained widespread utilization in genome sequencing, finding practical

applications in the examination of both prokaryotic and eukaryotic genomes. In the realm of prokaryotic genomes, overlap graph assembly has emerged as a prevalent technique for producing superior, complete genomes using short-read sequencing data [14]. Conversely, in the domain of eukaryotic genomes, the advent of long-read sequencing technologies like PacBio and Oxford Nanopore has facilitated the construction of intricate overlap graphs capable of encompassing repetitive regions and structural variations [15].

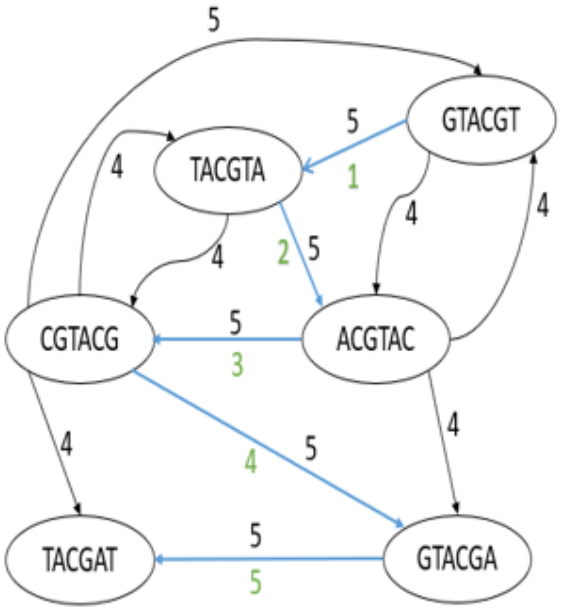

Figure 4. Completed Overlap Graph of G with Path line

Once the connection of the 6-mers reads in the Overlap graph is complete, we will obtain the diagram depicted above (refer to Figure 3), which clearly illustrates the linkage of all overlapping nodes through directed edges from the prefix to the suffix. The weight of this Overlap graph is indicated either on the right side or the top portion of each edge. To reconstruct the original genome, we traverse the overlap graph by following each directed edge, creating a path. The specific walk is illustrated above (refer to Figure 4), providing us with valuable insights into the sequence of the original genome. The resulting path is presented in the table below (refer to Table 1). In summary, overlap graphs serve as a potent tool in genome sequencing, enabling the assembly of complete genomes from overlapping DNA fragment reads. Their utility extends beyond genome sequencing and finds application in various genomic fields such as metagenomics and transcriptomics, making them a versatile and invaluable asset in genomic research.

In the provided table (see Table 1), a clear observation can be made by analyzing the walk and reconstructing the original DNA sequence using the overlaps, resulting in GTACGTACGAT. However, the Overlap Graph (OG) method has certain drawbacks. Firstly, it is challenging to determine if two distinct arcs represent the same overlap. Secondly, the OG has a quadratic size due to the inclusion of an arc for every possible directed word pair. Although Overlap Graphs have been widely used in genome assembly, they have limitations in handling repeat regions and data errors. To overcome these limitations, researchers have recently shifted their focus towards Hamiltonian graph-based approaches. Further studies have explored various variations of the string graph algorithm and other Hamiltonian path-based methods for genome assembly. For instance, the "de Bruijn graph" algorithm can also be regarded as a form of Hamiltonian graph.

Table 1. Genome sequence reconstruction of G

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

3. Utilizing an Hamiltonian Graph for DNA Sequencing

DNA sequencing using the Hamiltonian approach refers to a computational method that utilizes concepts from physics and optimization theory, specifically the Hamiltonian formalism, to solve the problem of determining the nucleotide sequence of a DNA molecule. The Hamiltonian approach in DNA sequencing involves modeling the DNA sequencing problem as an optimization problem, where the goal is to find the most likely sequence of nucleotides that explains a set of experimental data, typically obtained through high-throughput sequencing technologies.

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: Consider the multiset H = {TGC, TTC, GCT, TCC, CTA, CCA, TAG, CAA, AGT, GTT, AAT, TTT, ATA}, which comprises all 3-long nucleotides obtained from a genome sequence. To reconstruct the original gene sequence using a Hamiltonian cycle, we will utilize the given reads of the genome mentioned above. Following the steps outlined above, we construct a network that represents the overlap information present in our DNA reads. The resulting diagrams are depicted below.

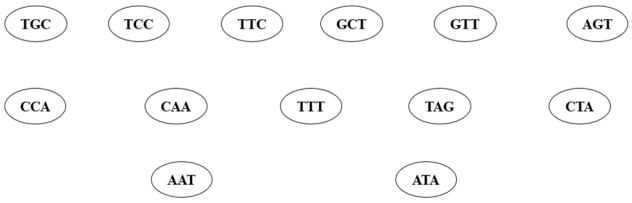

Figure 5. DNA Reads as Nodes

From Figure 5, as a first step again in using Hamiltonian approach, the 3-long nucleotides of a genome sequence have been expressed as node and all the genome reads are distinct

nodes without any repetition.

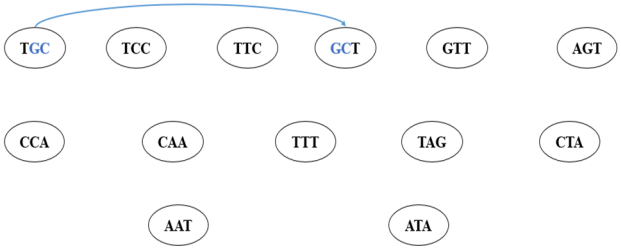

Figure 6. Directed edge between node TGC & GCT

On Figure 6, 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.

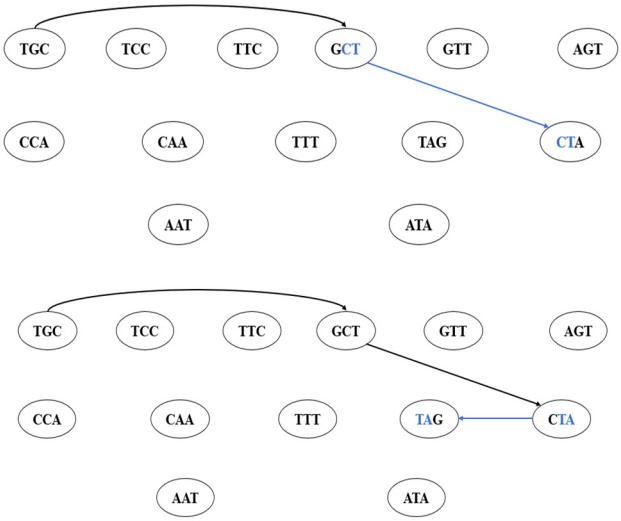

Figure 7. Continued connecting nodes together of graph of H

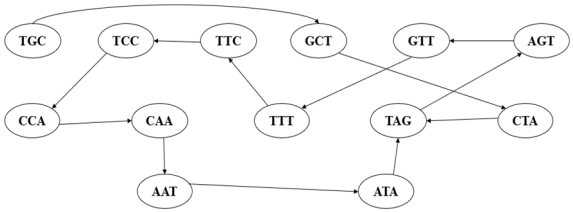

Figure 8. Completed Hamiltonian graph of set H

By completing connecting all the node (see Figure 7), as the first vertex overlaps with the k-1 leftmost nucleotides of the second vertex (see Figure 8), Complete Hamiltonian graph of H, is created. The graph can also be re-arranged into this graph (see Figure 9).

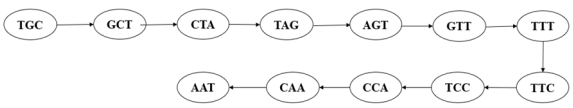

Figure 9. Re-arranged Hamiltonian path of H

Therefore, from the re-arranged graph above, 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 discovering a path that visits each node exactly once, known as a Hamiltonian path, we can determine an ordered arrangement of the fragments comprising the complete DNA sequence. However, it is unfortunate that finding a Hamiltonian path is a challenging task, classified as an NP-Complete problem.

4. Utilizing an Eulerian Graph for DNA Sequencing

In this section, we present the sequencing of the DNA using an Eulerian approach. When a closed trail traverses through every edge of a connected graph G, the graph is referred to as an Euler graph. To qualify as an Euler path, each edge of a graph must be used exactly once, with the starting and ending vertices being distinct. Similarly, an Euler circuit utilizes every edge of a graph exactly once, with the same vertex serving as both the beginning and end. For a connected graph G, it is considered an Euler graph if and only if all its vertices have an even degree. Additionally, a connected graph G is deemed Eulerian if and only if its edge set can be decomposed into cycles.

To create a graph using genome reads, we establish a node for each distinct prefix or suffix observed. The vertices of this graph are formed from the set of (l-1)-mers, which are substrings of some of the l-mers in our set S. Whenever a node with prefix v and suffix w is encountered, we connect node v to node w. For nodes v and w to be connected by a directed edge, two conditions must be met: the last l-2 elements of node v and the first l-2 elements of node w must match, and the concatenation of node v and node w must exist in set S. By satisfying these criteria, a directed edge is established between node v and node w. To reconstruct the shortest sequence string using the Eulerian path, we consider a set of (l-1)-mer strings. These strings have a length one less than the given strings and play a crucial role in the reconstruction process.

Example 3: Let H be a multiset consisting of all 3-long nucleotides from a genome sequence, represented by the set H= {AAT, TGC, CAA, GCT, CCA, CTA, TCC, TAG, AGT, TCC, TTT, TTC}. By constructing nodes for each distinct prefix/suffix, such as CTA, we can identify the prefix CT and the suffix TA. By exhaustively identifying the distinct prefixes and suffixes, we obtain the following results:

V= {AT, GC, CT, CC, AA, TG, TT, CA, AG, GT, TC, TA}

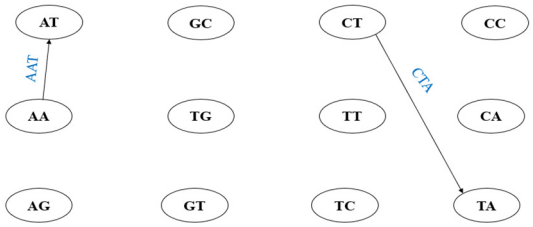

Figure 10. Multigraph with AAT and CTA

As depicted above (see Figure 10), prefix AA connects to the suffix AT with an edge AAT as the DNA read also prefix CT connects to the suffix TA with an edge CTA as the DNA read. By completing the diagram connecting these prefixes to suffix we can the following graph (see Figure 11).

From the diagram above (see Figure 11), the numbers mark the Eulerian path that we will be followed when reconstructing this genome from k-mer DNA reads and by using the overlaps DNA reads we can produce the following path table. (See Table 2)

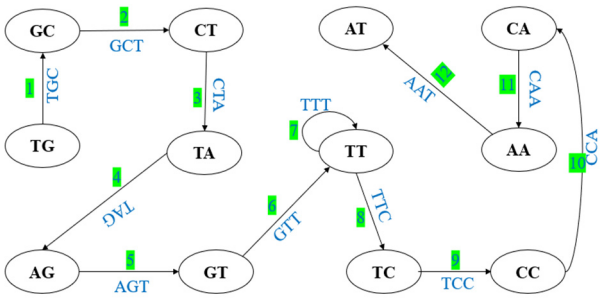

Figure 11. Complete Eulerian path of H

Table 2. Eulerian path table of H

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

The table above illustrates how we can reconstruct our original genome using DNA reads overlaps and we got our genome using Eulerian path TGCTAGTTTCCAAT. Comparing between Hamiltonian approach and Eulerian approach, the only difference is when using a computer, it can easily find the Eulerian cycle very fast compared to when using Hamiltonian cycle.

5. Conclusion

About Overlap Graph, DNA sequencing can be approached using an overlap graph, where nodes represent sequences and edges indicate overlaps between sequences. By analyzing the graph structure, we can identify potential paths or sequences that align with overlapping regions. This method is useful for identifying contiguous regions in a genome sequence and can help in reconstructing the original sequence. In the context of DNA sequencing, a Hamiltonian graph represents a graph where every node represents a DNA sequence and every edge connects sequences that overlap. However, finding a Hamiltonian path or cycle in a graph is computationally complex and may not always be feasible for large-scale DNA sequencing projects. Therefore, while Hamiltonian graphs provide a theoretical framework for sequence reconstruction, practical applications may require alternative approaches. Euler graphs, or Eulerian paths/cycles, have been extensively studied in the context of DNA sequencing. In a Eulerian graph, every edge represents a DNA sequence, and a path or cycle exists that visits each edge exactly once.

Eulerian graphs are particularly useful for de Bruijn graph-based sequencing methods, where sequences are represented as k-mers (short subsequences) and overlaps between k-mers are analyzed. This approach allows for efficient assembly of genome sequences. In summary, overlap graphs provide insights into overlapping regions in DNA sequences, Hamiltonian graphs offer theoretical frameworks for sequence reconstruction (though not always practical), and Eulerian graphs play a crucial role in de Bruijn graph-based DNA sequencing methods. Each graph type has its own strengths and limitations, and the choice of method depends on the specific requirements and scale of the DNA sequencing project.

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. https://doi.org/10.1038/171737a0.

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

[3] Khrapko KR, Lysov YuP, Khorlin AA, Ivanov IB, Yershov GM, Vasilenko SK, Florentiev VL, Mirzabekov AD. A method for DNA sequencing by hybridization with oligonucleotide matrix. DNA Seq. 1991;1(6):375-88. https://doi.org/ 10.3109/ 10425179109020793. PMID: 1768861.

[4] Pevzner P. A. (1989). 1-Tuple DNA sequencing: computer analysis. Journal of biomolecular structure & dynamics, 7(1), 63–73. https://doi.org/10.1080/07391102.1989.10507752.

[5] 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, 2(2), 291–306. https://doi.org/10.1089/cmb.1995.2.291.

[6] Margulies, M., Egholm, M., Altman, W. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380 (2005). https://doi.org/ 10.1038/ nature 03959.

[7] Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). Genome assembly using DNA sequencing reads. Proceedings of the National Academy of Sciences, 98(17), 9748-9753.

[8] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. Genome research, 19(6), 1117-1123.

[9] Berger, E., Yorukoglu, D., Peng, J., & Berger, B. (2011). Genome-scale algorithm design: Biological sequence analysis in the era of high-throughput sequencing. Proceedings of the National Academy of Sciences, 108(12), 5690-5695.

[10] Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome research, 27(5), 665-676.

[11] Siren, J., Almujayyaz, S., & Ukkonen, E. (2019). GraphAligner: Rapid and versatile sequence-to-graph alignment. Bioinformatics, 35(22), 4724-4732.

[12] Ashley E. A. (2015). The precision medicine initiative: a new national effort. JAMA, 313(21), 2119–2120. https://doi. org/10. 1001/ jama.2015.3595.

[13] 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. https:// doi.org/10.1371/journal.pbio.1001883.

[14] Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome biology, 17(1), 239. https:// doi.org/ 10.1186/s13059-016-1103-0.

[15] Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. Genomics, proteomics & bioinformatics, 13(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002.