Scientific Research Publishing

Open Journal of Discrete Mathematics, 2023, 13, 39-48
https://www.scirp.org/journal/ojdm
ISSN Online: 2161-7643
ISSN Print: 2161-7635

Genome Sequencing Using Graph Theory Approach

Shepherd Chikomana, Xiaoxue Hu*

School of Science, Zhejiang University of Science and Technology, Hangzhou, China
Email: *xxhu@zjnu.edu.cn

How to cite this paper: Chikomana, S. and Hu, X.X. (2023) Genome Sequencing Using Graph Theory Approach. *Open Journal of Discrete Mathematics*, **13**, 39-48.
https://doi.org/10.4236/ojdm.2023.132004

Received: February 23, 2023
Accepted: April 16, 2023
Published: April 19, 2023

Abstract

Genome sequencing is the process of determining in which order the nitrogenous bases also known as nucleotides within a DNA molecule are arranged. Every organism's genome consists of a unique sequence of nucleotides. These nucleotides bases provide the phenotypes and genotypes of a cell. In mathematics, Graph theory is the study of mathematical objects known as graphs which are made of vertices (or nodes) connected by either directed edges or indirect edges. Determining the sequence in which these nucleotides are bonded can help scientists and researchers to compare DNA between organisms, which can help show how the organisms are related. In this research, we study how graph theory plays a vital part in genome sequencing and different types of graphs used during DNA sequencing. We are going to propose several ways graph theory is used to sequence the genome. We are as well, going to explore how the graphs like Hamiltonian graph, Euler graph, and de Bruijn graphs are used to sequence the genome and advantages and disadvantages associated with each graph.

Keywords

DNA Sequencing, Hamiltonian Graph, Euler Graph, de Bruijn Graph, Nucleotide

1. Introduction

In 1953, two scientists, J.D. Watson and F.H.C. Crick [1] 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. *Genome 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 genome sequencing and a variety of applied fields like biotechnology, forensic biology, and diagnostics. The journey with genome sequencing began in 1977 [2] when Frederick with his colleagues proposed a method on chain-termination inhibitors. Sanger sequencing is known to deliver 99.99% base accuracy that is crucial for optimum validation in the field of genetics. It is considered the gold standard when the job is to understand how the genes carry out information (The Genomic Services Company, 2020).

Sanger sequencing was used in the Human Genome Project to determine the sequences of relatively small fragments of human DNA (900 bp or less). These fragments were used to assemble larger DNA fragments and, eventually, entire chromosomes. Edwin Southern [3] introduced a new genome sequencing approach where the genome is sequenced by hybridization (SBH). SBH is an approach whereby a collection of overlapping oligonucleotide sequences is assembled together to determine an organism's DNA sequence. Through the efficient method of SBH, scientists are able to gather information on the genomes of different species and organisms for the future development of biological sciences, medicine, and agriculture. Among the scientists of algorithmic approaches to SBH we can distinguish Y.P. Lysov with his colleagues [4] and Pevzner [5], who formulated the problem as finding a Hamiltonian path and an Eulerian path, respectively. Next Generation Sequencing is yet another genome sequencing method and is a powerful platform that has enabled the sequencing of thousands to millions of DNA molecules simultaneously (Margulies, Egholm, & Altman [6]).

The methods of sequencing have become a game-changer in modern biological and medical fields. DNA sequencing has accelerated not only biological research and discovery but also enhanced medical diagnostics and treatment of diseases. This article will focus more on methods for DNA sequencing which use concepts of graph theory.

Before we dive into the graph theory approaches in genome sequencing, let's briefly understand what graph theory is and definitions of key words used in graph theory.

Graph Theory is a mathematical representation of a network and it describes the relationship between lines and points. A graph consists of some points and lines between them. The length of the lines and position of the points do not matter. Each object in a graph is called a node. A graph G is a set of vertices, called nodes v which are connected by edges, called links e. Thus G = (v, e).

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). Edge is a link between two nodes. An edge denotes movements between nodes. It has a direction that is generally represented as an arrow. If an arrow is not used, it means the link is bi-directional.

2. Genome Sequencing Using Hamiltonian Graph

In this section, we show the genome sequencing by using Hamiltonian graph. A connected graph G is called Hamiltonian graph if there is a cycle which includes every vertex of G and the cycle is called Hamiltonian cycle. Hamiltonian walk-through graph G is a walk that passes through each vertex exactly once. We first show two very famous Theorems for Hamiltonian graph, which the proofs can be found in [7].

Theorem 2.1. (Dirac's Theorem) *states that if G is a simple graph with n vertices, where* $n \geq 3$, *If* $\deg(v) \geq \frac{n}{2}$ *for each vertex v, then the graph G is Hamiltonian graph.*

Theorem 2.2. (Ore's Theorem) *states 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.*

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 1: Let S = {AAT, GCG, GCA, ATG, TGG, TGC, GGC, GTG, CGT, CAA} be a multiset of all 3-long nucleotides of a DNA sequence. Let's construct a network that represents the overlap information in our reads. Each k-mer nucleotide from the multiset becomes a vertex (as depicted in **Figure 1**); two vertices are connected by a directed vertex if the $k-1$ rightmost nucleotides of first vertex overlap with the $k-1$ leftmost nucleotides of the second one.

First, we create a node for each read. e.g., GTG. Prefix: First two nucleotide of a read (GTG). Suffix: Last two nucleotide of a read (GTG). Note: Different 3-mers may share a prefix/suffix: ATG, TGA, CTG. As shown from **Figure 1** DNA reads are aligned and ready to be joined using overlap reads from prefix to suffix.

Figure 2 shows how to connect these DNA nodes based on the prefix and suffix. As illustrated from **Figure 2** with nodes ATG and GTG connecting to nodes TGC and TGG based on the overlapping part of the nucleotide.

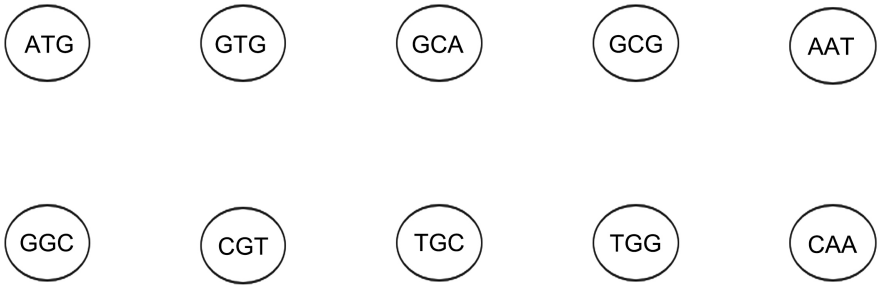

Figure 1. Aligning the 3-mer nodes.

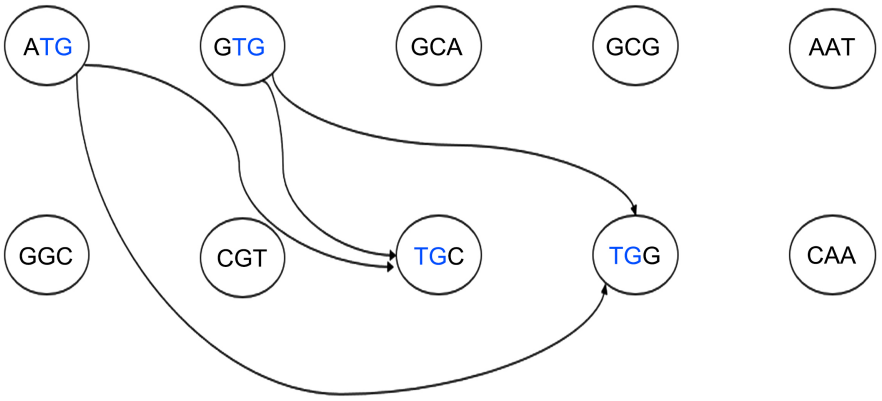

Figure 2. Connecting DNA nodes based on prefix and suffix.

From the diagram in Figure 3, we can clearly see a completed graph when connecting all nodes with the same prefix to the node with the same suffix and shows how the genome sequencing is done using the Hamiltonian approach. Now we need to deduce the order of the DNA and for us to do so we need to follow the path on the diagram that takes us from where we started. Our Hamiltonian cycle will be: ATG → TGG → GGC → GCG → CGT → GTG → TGC → GCA → CAA → AAT → ATG. Therefore, our genome from this reconstruction is ATGGCGTGCAAT.

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 the given reads of DNA above, let's reconstruct the original gene sequence using Hamiltonian cycle.

Using the steps in example 1, constructing a network that represents the overlap information in our DNA reads will give the diagram above, as depicted in Figure 4. Therefore, from the re-arranged graph above, as depicted in Figure 5, 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.

If we find 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 isnot easy (it is classed as an NP-Complete problem).

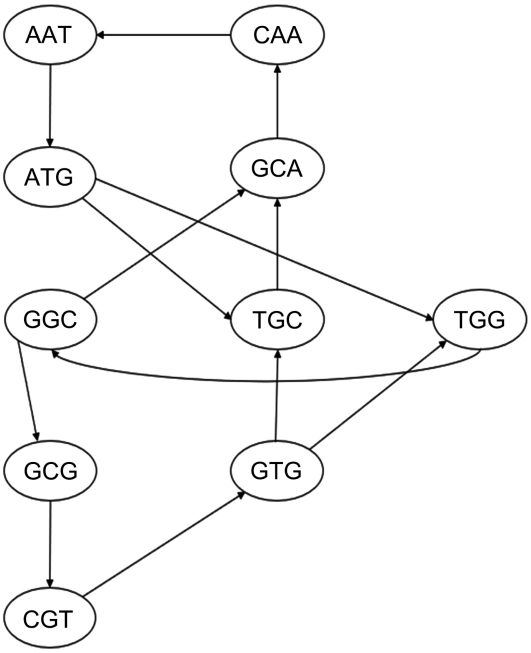

Figure 3. Complete Hamiltonian Graph of example 1.

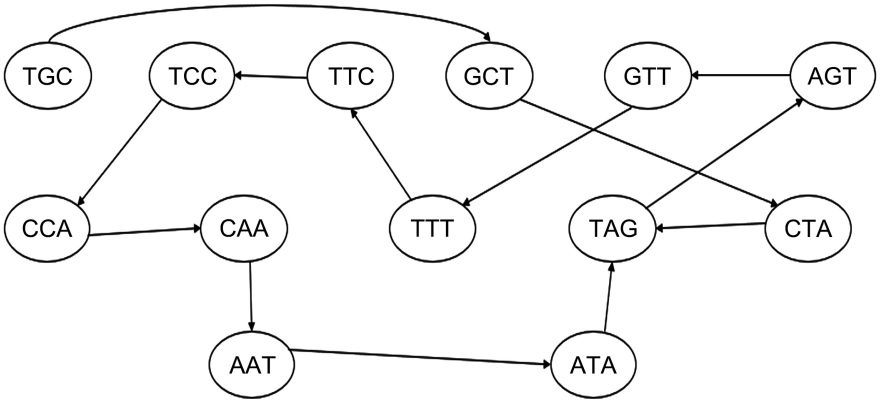

Figure 4. Hamiltonian graph of H.

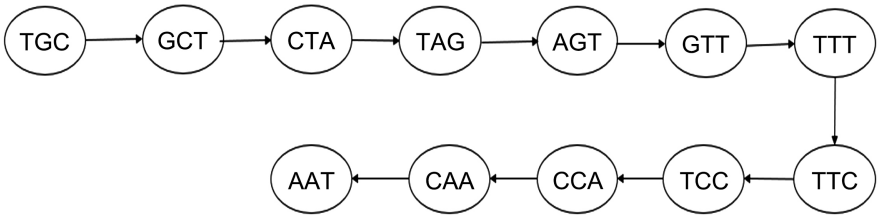

Figure 5. Re-arranged Hamiltonian path of H.

3. Genome Sequencing Using Eulerian Graph

In this section, we show the genome sequencing by using Eulerian graph.

A connected graph G is called a Euler graph, if there is a closed trail which includes every edge of the graph G. A Euler path is a path that uses every edge of a graph exactly once. A Euler path starts and ends at different vertices. A Euler

circuit is a circuit that uses every edge of a graph exactly once. A Euler circuit always starts and ends at the same vertex.

A connected graph G is a Euler graph if and only if all vertices of G are of even degree, and a connected graph G is Eulerian if and only if its edge set can be decomposed into cycles. Objective: From a given set, S, in reads, use Eulerian approach to reconstruct the genome sequence.

Using genome reads, make a node for each unique prefix or suffix. From the set of (l-1)-mers, which are substrings of some of the l-mers in our set S, will make up the vertices. Whenever there is a node which has a prefix $\boldsymbol{v}$ and suffix is $\boldsymbol{w}$, connect the node $\boldsymbol{v}$ to node $\boldsymbol{w}$. If the final l-1 elements of node $\mathbf{v}$ and first l-2 elements of node $\mathbf{w}$ match and the union of node $\boldsymbol{v}$ and node $\boldsymbol{w}$ is in set S, then node $\boldsymbol{v}$ and node $\boldsymbol{w}$ are connected by a directed edge.

In order to reconstruct the shortest sequence string using the Eulerian path, a set of (l-1) mer strings (*i.e.*, strings having length less by one from given strings) are taken into account.

Example 3: Let H = {AAT, TGC, CAA, GCT, CCA, CTA, TCC, TAG, AGT, TCC, TTT, TTC} be a multiset of all 3-long nucleotides of a DNA sequence. We create a node for each distinct prefix/suffix *i.e.*, CTA we get the prefix CT and the suffix TA. By completing finding the distinct prefix and suffix we get the following V = {AT, GC, CT, CC, AA, TG, TT, CA, AG, GT, TC, TA}.

As illustrated above,(as depicted in Figure 6), 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 show in Figure 7.

From the diagram above (as depicted in Figure 7), 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 path table (as depicted in Figure 8).

The path table above (as depicted in Figure 8) illustrates how we can reconstruct our original genome using DNA reads overlaps and we got our genome using Eulerian path TGCTAGTTTCCAAT.

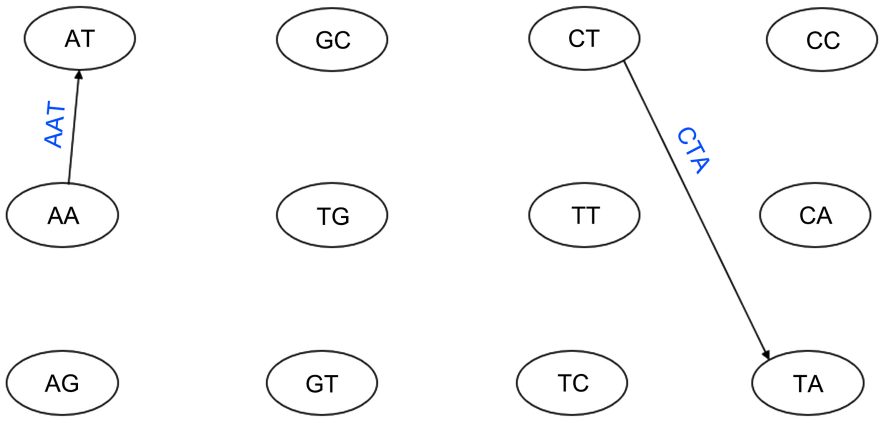

Figure 6. Multigraph with AAT and CTA.

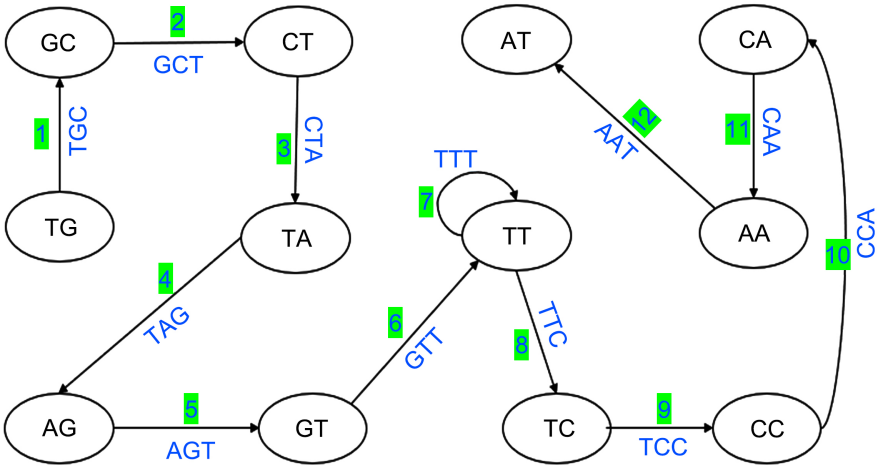

Figure 7. Complete Eulerian path of H.

	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
genome	T	G	C	T	A	G	T	T	T	C	C	A	A	T

Figure 8. Eulerian path table of H.

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.

4. Genome Sequencing Using de Bruijn Graph

A sequence's k-mer components can be efficiently represented using a de Bruijn graph. Despite the fact that de Bruijn graphs can be applied to a variety of issues, we will focus on nucleotide sequences in this article. Around the 1940s, Nicolaas de Bruijn, a Dutch mathematician, became interested in finding the shortest circular string of characters that encompasses all conceivable substrings of the same length in a particular alphabet. He came up with a solution that entailed creating a graph with all of the possible $(k-1)$-mers as the nodes. If the $(k-1)$-mer in node A is a prefix and that in node B is a suffix of the k-mer, then each

k-mer was an edge directed from node A to node B. Finding a path through the graph that passes over each edge exactly once, or an Eulerian trail, was the suggested answer. Our genome reads are fragmented into smaller fragments of a given size *k*. A node for each $(k-1)$-mer from *k*-mers for each *k*-mer in *k*-mers is formed and an edge is used to connect its prefix node with its suffix node.

Let's illustrate with an example below.

Example 4: Let M = {TGT, AAT, TGG, ATG, TGC, ATG, TAA, ATG, GTT, CAT, CCA, GGA, GCC, GAT, GGG} be a multiset of all 3-long nucleotides of a DNA sequence. Take all distinct $(k-1)$-mers from the set of *k*-mers, here $k=3$. *i.e.*, TGC, TGG → TG, GC, GG. Construct a multi-graph with nodes being $(k-1)$-mers; draw an edge between two $(k-1)$-mers only if the two $(k-1)$-mers are taken from the same read. *i.e.*, AAT & ATG (as depicted in Figure 9).

This method guarantees that the graph will have a Eulerian trail, by following the Eulerian trail and joining the nodes will thereby reconstruct our original genome sequence. A graph similar to this will be displayed (as depicted in Figure 10).

As illustrated in Figure 10, each edge is this graph corresponds to a length-3 length input string. Through this network, a Eulerian path is traced, and therefore as a result, we are able to reconstruct our original genome sequence table (as depicted in Figure 11).

Using a Eulerian walk crossing each edge exactly once gives a reconstruction of the genome. After funding the walk, our genome will be TAATGCCATGGGATGTT.

Figure 9. Multigraph with AAT and ATG.

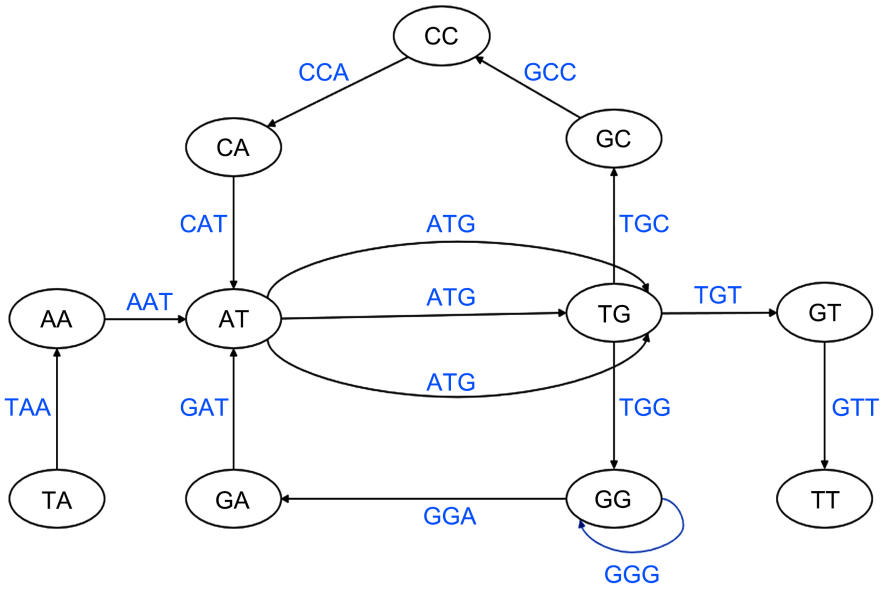

Figure 10. Complete de Bruijn graph of M.

	T	A	A													
		A	A	T												
			A	T	G											
				T	G	C										
					G	C	C									
						C	C	A								
							C	A	T							
								A	T	G						
									T	G	G					
										G	G	G				
											G	G	A			
												G	A	T		
													A	T	G	
														T	G	T
															G	T
																T
genome	T	A	A	T	G	C	C	A	T	G	G	G	A	T	G	T

Figure 11. The reconstruction of the original genome sequence table of M.

5. Conclusions

When it comes to solving biological problems and making medicine, graph theory plays a vital role and it's necessary for this generation to understand it. DNA sequencing was critical in mapping out the human genome, which was finished in 2003, and is now a crucial tool for many fundamental and practical research applications.

This paper discussed how graph theory is used in genome sequencing and showed some of the graph theory graphs that are used and showed some graphical representation of those methods and some step by step on how the methods are used.

When using Hamiltonian approach, as the DNA reads increase, finding a Hamiltonian path is not easy (it is classed as an NP-Complete problem). Euler approach is a better approach than Hamiltonian approach in genome sequencing because nowadays with massive growth in genetics finding the Euler cycle does not take a long time. Sadly, in real life there are some other problems that make this process harder. One example is if a fragment occurs multiple times in a sequence.

The de Bruijn graph approach has proven to be a better method in genome reconstruction compared to Euler approach and Hamiltonian approach.

True, assembly approaches based on de Bruijn graphs start rather counter-intuitively, by replacing each read with a collection of all-overlapping sequences of a shorter, fixed length, but this is a popular way for genome assembly. Although the de Bruijn assembler is a famous way to perform assembling, there are significant obstacles for de Bruijn genome assembly, including sequence error, unequal sequencing depth, repetitive parts, and processing expense.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Watson, J.D. and Crick, F.H.C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737-738. https://doi.org/10.1038/171737a0

[2] 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. https://doi.org/10.1073/pnas.74.12.5463

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

[4] Khrapko, K.R., Lysov, Yu.P., Khorlin, A.A., Ivanov, I.B., Yershov, G.M., Vasilenko, S.K., Florentiev, V.L. and Mirzabekov, A.D. (1991) A Method for DNA Sequencing by Hybridization with Oligonucleotide Matrix. *DNA Sequence*, **1**, 375-388. https://doi.org/10.3109/10425179109020793

[5] Pevzner, P.A. (1989) I-Tuple DNA Sequencing: Computer Analysis. *Journal of Biomolecular Structure and Dynamics*, **7**, 63-73. https://doi.org/10.1080/07391102.1989.10507752

[6] Margulies, M., Egholm, M., Altman, W., *et al.* (2005) Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature*, **437**, 376-380. https://doi.org/10.1038/nature03959

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