Review

Different DNA Sequencing Using DNA Graphs: A Study

Abdulaziz M. Alanazi [1], G. Muhiuddin [1,*], Doha A. Al-Balawi [2] and Sovan Samanta [3]

1 Department of Mathematics, Faculty of Science, University of Tabuk,
   P.O. Box 741, Tabuk 71491, Saudi Arabia; am.alenezi@ut.edu.sa
2 Department of Biology, Faculty of Science, University of Tabuk,
   P.O. Box 741, Tabuk 71491, Saudi Arabia; d.albalawi@ut.edu.sa
3 Department of Mathematics, Tamralipta Mahavidyalaya, Tamluk 721636, West Bengal, India;
   ssamanta@tmv.ac.in
* Correspondence: gmuhiuddin@ut.edu.sa

Abstract: Natural genetic material may shed light on gene expression mechanisms and aid in the detection of genetic disorders. Single Nucleotide Polymorphism (SNP), small insertions and deletions (indels), and major chromosomal anomalies are all chromosomal abnormality-related disorders. As a result, several methods have been applied to analyze DNA sequences, which constitutes one of the most critical aspects of biological research. Thus, numerous mathematical and algorithmic contributions have been made to DNA analysis and computing. Cost minimization, deployment, and sensitivity analysis to many factors are all components of sequencing platforms built on a quantitative framework and their operating mechanisms. This study aims to investigate the role of DNA sequencing and its representation in the form of graphs in the analysis of different diseases by means of DNA sequencing.

Keywords: DNA bar graphs; DNA sequencing; DNA library; genetic diseases; graph theory

1. Introduction

DNA is a molecule that transfers genetic info to all living creatures. It is composed of two long polynucleotide chains that are bonded together to form a double helix structure. Each of these chains has a backbone made of a five-carbon sugar known as deoxyribose, which is attached to phosphate groups and a nitrogen-containing base. The nitrogenous base might be either adenine (A), thymine (T), guanine (G), or cytosine (C). A always pairs with T, and G with C. Genomic DNA has a significant role in the determination of individuals and species in most cases, making DNA sequence study important for understanding cell structures and activities and unlocking biological mysteries [1].

DNA sequencing technology might be useful for biologists and health care practitioners for various applications, including molecular cloning, breeding, diagnosing diseases and genetic disorders, and comparative and evolutionary research. Therefore, DNA sequencing technology should ideally be quick, accurate, simple to use, and inexpensive. DNA sequencing methods and applications have advanced tremendously in the last thirty years, and they now serve as an engine for genome age, which is defined by large amounts of genomic data. As a result, there is a diverse variety of study topics and applications. Studying the history of sequencing technology development is very important for evaluating next-generation sequencing systems (NGS), such as pyrosequencing (454), genome analyzer and high-throughput sequencing, and SOLiD sequencing. In addition, it helps for comparing the advantages and disadvantages of the technology and explaining the variety of applications. Moreover, the importance of the history of sequencing development also lies in the ability to evaluate the personal genome machines (PGM) and third-generation technologies and their applications. The majority of the information and conclusions come from independent users with considerable first-hand experience with BGI's (the Beijing Genomics Institute's) standard NGS equipment.

Prior to entering into NGS systems, let us take a look at the brief history of DNA sequencing. Frederick Sanger invented DNA sequencing in 1977 using the chain-termination approach (often referred to as Sanger sequencing). In contrast, Walter Gilbert developed another method in 1978 using a DNA chemical modification and subsequent cleavage at certain bases. Due to its low radioactive nature and great efficiency, Sanger sequencing was selected as the dominant approach in the "first generation" of commercial and laboratory sequencing applications [2]. DNA sequencing was time intensive and required the use of radioactive materials at the time [3]. In 1987, Applied Biosystems introduced the world's first automated sequencing machine (the AB370), which accelerated and improved sequencing accuracy via the use of capillary electrophoresis. The AB370 was competent in concurrently detecting 96 bases at a 500 K basis rate every day and 600 bases per reading. Since 1995, the latest model, AB3730 xl, has been able to generate 2.88 million bases per day and read lengths of up to 900 bases. Automatic sequencing devices and Sanger sequencing technology and related software based on capillary sequencing machines, which first appeared in 1998, became the primary instrument for completing the "Human Genome Project" [HGP] in the year 2001 [4]. This effort aided in the creation of a powerful new sequencing apparatus that was able to increase accuracy and speed while lowering costs and decreasing labor. Not only that, but the X-prize has strengthened next-generation sequencing (NGS) technology [5]. In terms of lower cost, massively parallel analysis, and high throughput, NGS technologies differ from the Sanger approach. Despite the fact that NGS makes genome sequences more accessible, the biological explanations and subsequent data analysis remain a bottleneck in comprehending genomes.

Recently, Rusinova and Stroganov [6] discussed the topic of formalizing graph data models for genomic information representation. The model outlined by them is intended for use in a graph database for comparative genomics applications. There are many different ways of comparing genomes, and the theory-based graph model is intended to reflect both the probabilistically and the definitive parts of the process.

Karunasena and Wijesiri [7] developed theories on modeling weighted directed graphs, generating their adjacency matrixes, and translating them to their representative vector. Distance metrics such as Euclidean, Cosine, and Correlation were used to establish how similar these vectors were. To determine whether this method applies to any DNA fragments in the considered genomes, they employed molecular similarity coefficients as distance measurements and kept this method as the underlying method. The representative vector and graph spectrum findings are compared. The mitochondrial DNA of humans, gorillas, and orangutans was used to test the improved approach. Even with more nucleotides in DNA fragments, the outcome was still the same.

In this review article, the graph-theoretic approach of DNA sequencing is studied. Additionally, studies are carried out to analyze the diseases due to sequencing problems of DNA. This study will direct researchers to identify the ambiguity of existing sequencing techniques and to identify the deficiency of patterns of DNA sequencing. The history of DNA sequencing is shown in Figure 1.
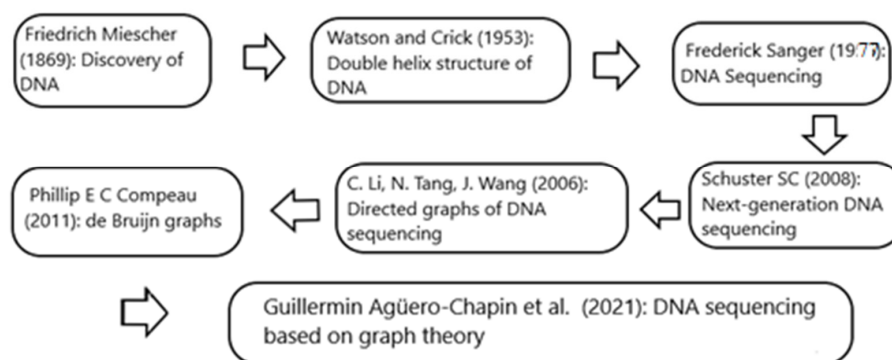
Figure 1. The history of DNA sequencing and double helix discovery.

2. DNA Library

A DNA library may be thought of as a collection of DNA sequences. A DNA library, like a book or any other type of data library, may be used to store and distribute information. A DNA chain serves as the data carrier in a similar type of library. Typically, these chains reflect natural genetic data derived from biochemical investigations. DNA can provide information about gene expression events or aid in the detection of genetic illnesses. However, some applications necessitate DNA chains with special characteristics that are seldom seen in natural DNA molecules. DNA computing, which was developed in 1994 by Adleman [1], is an example of such an application. Graph vertices are encoded as randomly produced DNA chains of length l, while arcs are encoded as a concatenation of two chains: one supplementary to the last 1/2 nucleotides of the preceding vertex and one additional to the first 1/2 nucleotides of the succeeding vertex, the Hamiltonian path problem was solved. In a perfect scenario, vertices and arcs would hybridize; nevertheless, random chain creation might result in hybridization between arcs and vertices non-existent in that graph. The research provides an approach for building DNA libraries composed of chains with a low proclivity for hybridization. As a result, this kind of library may later be utilized for encoding a collection of vertices. This approach is significant from a mathematical standpoint. As a result, a generalized de Bruijn sequence [2] lacks complementarity. The algorithm's first draught, on the basis of de Bruijn graphs, was published in [8], and it was modified further in [9]. The approach relies on graph theory, more precisely on lexical graphs, which are de Bruijn graphs that have been expanded. In [10], lexical graphs were introduced, and independently similar ideas were found in [11,12]. In [13], they were named word graphs; in [12], they were referred to as alphabet overlap digraphs. Each Eulerian cycle in a de Bruijn graph correlates to a de Bruijn sequence [2]. Every Eulerian cycle in a lexical graph corresponds to a lexical sequence, which is a de Bruijn sequence that has been expanded. De Bruijn sequences with several shifts were presented in [9] and independently in [14], where they are referred to as lexical sequences. Ref. [14] developed de Bruijn sequences with multiple shifts to solve the Frobenius problem in a free monoid [15]. Labeled graphs are subgraphs of de Bruijn graphs that relate to the DNA and labeled graphs described in [4]. Induced subgraphs of lexical graphs are analogically base labeled graphs [16]. DNA graphs are employed in DNA sequencing, for example [4,9] (see Figure 2). All types of graphs discussed above have the attribute of having labels given to their vertices [12,15]. These labels are words with a fixed measurement that is determined by an alphabet. Any DNA chain may be regarded as a sequence using the letters A, C, T, and G. The DNA chains being directed, they contain two separate endpoints, one designated 3′ and the other 5′ (for example, 5′-ACTG-3′ or, equivalently, 3′-GTCA-5′). Because the direction of the DNA chain is critical for the algorithm, we shall assume one fixed orientation, such as 5′ to 3′ (thus, ACTG will always imply 5′-ACTG-3′). In DNA, hybridization happens only between chains that run in reverse directions; hence, chain AATCCG is complementary to chain CGGATT. The main aim of this study is to analyze the visualization of DNA sequencing using graph theoretic approaches.
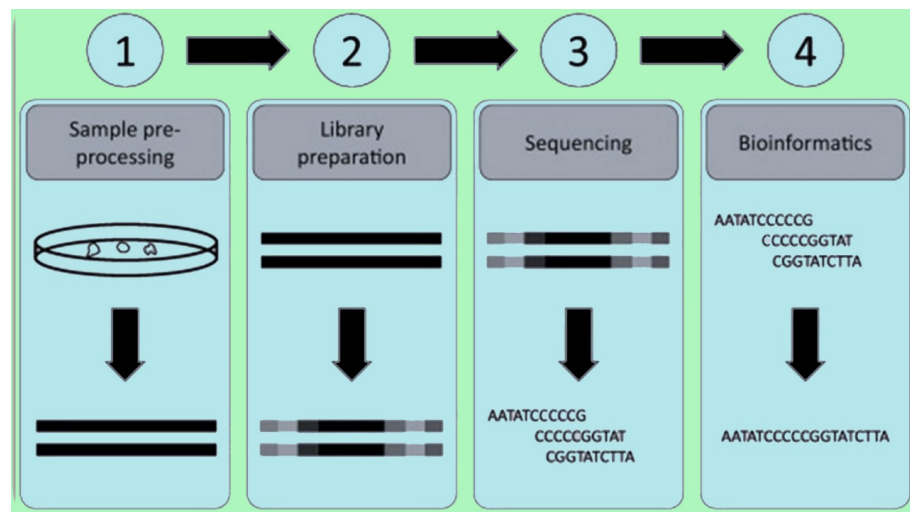
Figure 2. Next-generation sequencing workflow steps [5,9,17,18].

3. DNA Graphs

Studying DNA sequences has been a subject of supreme importance in biological research, and there have been many contributions regarding computing from an algorithmic and mathematical point of view and DNA analysis [19,20]. A sequencing platform built on a mathematical framework and its underlying mechanism provides several benefits, including cost optimization, implementation, and sensitivity analysis to various circumstances [17,21,22]. The read process was modeled using profile vectors and the DNA storage channel. They provided novel asymmetric coding techniques to resist sequencing noise and synthesis and an asymptotic study of the quantity of profile vectors. Finally, two code families for this channel model were created. A customized compressor design was devised to successfully store FASTQ data generated by large DNA sequences. The equation of motion linked cluster trick was used to determine the ionization capability with single and double excitations, and VIEs were calculated using density functional theory with dispersion adjusted omega B97x-D [11]. An investigation into how to construct independent spanning trees on hypercubes and use them to predict mitochondrial DNA sequence sections using hypercube routes [8]. Depending on the graph theory principles and genetic codes, an alignment-free technique for DNA sequence resemblance investigation was developed (see Figure 3). A novel method for testing DNA sequences was developed [2,8]. The electron propagator hypothesis [23] was used to study transverse electron transportation via all four N2 bases.
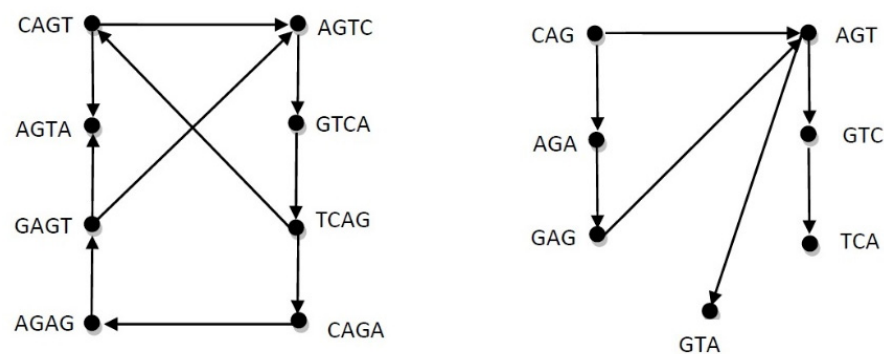

Figure 3. DNA sequencing using graph theory [24–26].

4. DNA Graphs for DNA Sequencing

The DNA graphs, which were generated using the approaches presented in Lysov et al., belong to the class of labeled digraphs. The fact that the labeled digraphs are directed

line graphs is one of their features; hence, the Hamiltonian route problem can be solved polynomially for them. This is performed by transforming a directed line graph into its original graph and then searching for a Eulerian route in the original graph. The following rules apply to a directed line graph G and its initial (directed) graph H: vertices of G correspond to arcs of H, and an arc (x, y) exists in G if and only if the terminal endpoint of arc x in H is also the beginning endpoint of arc y in H. A Eulerian route in H is a necessary and sufficient condition for a Hamiltonian track in G (not valid for undirected graphs). A Pevzner graph for the same spectrum is the original graph for a DNA graph. De Bruijn graphs [27,28] are labeled digraphs that are formed with all probable labels of a certain measurement across a particular alphabet. In the context of the next section's subject, they are especially important. A de Bruijn graph (see Figure 4) has k vertices, each of which is labeled with a unique word from the letters of the alphabet, given an alphabet of length k and size a. If the predecessor's suffix transcends the successor's prefix by a length of k 1, an arc links two vertices). DNA graphs are vortex-induced subgraphs of de Bruijn graphs with a = 4. DNA graphs and other de Bruijn graphs with vertex-induced subgraphs are not Pevsner graphs. They are de Bruijn graph subgraphs because they are subgraphs of DNA graphs. Those are not labeled/DNA graphs in general since the existence of arcs and the overlaps of vertex labels do not equate. See, for example, a series of publications on this issue that looked into the properties of labeled graphs in further depth [29–31]. On the other hand, there have been studies conducted on the links between different digraph classes and directed line graphs and a polynomial-time remedy to the Hamiltonian cycle/path problem [24,25,32]. The systematization depicted in Figure 2 may pique the reader's curiosity. The connection encompasses the graphs stated above and other classes outside the scope of this work. Directed line graphs are formed at the intersection of adjoints [2] and partially directed line graphs (PDLG, see [1]). (DLG). Apart from the scheme, quasi adjoints graphs [4] are a subclass of graphs that simulate the difficulty of isothermic DNA sequencing by hybridization (the graphs stated in [26], the issue presented in [33]). Non-directed adjoints, and directed line graphs are all instances of the latter. There is, however, always one directed graph, that is, a digraph with no multiple arcs connecting any two vertices.
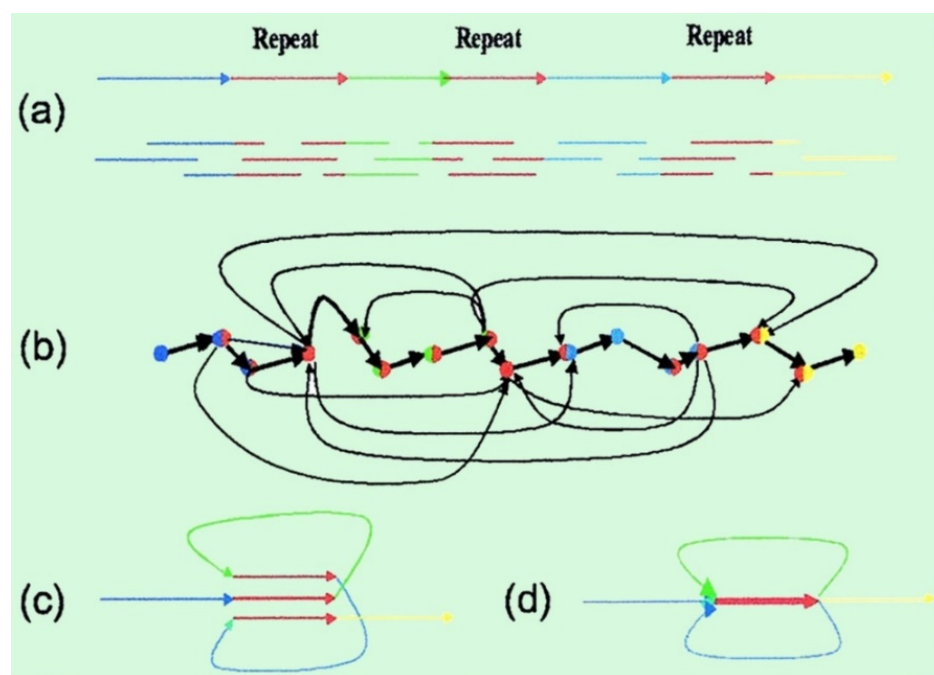

Figure 4. Steps of construction of de Bruijn graphs [24,32,34].

5. Overlap Graphs

Overlap graphs are a modified model of the Lysov version that has been adjusted to NGS data requirements. Like the original model, a vertex of directed graph G represents each read and arcs connected overlapping sequences showing the overlapping direction. One of the most significant distinctions is how individual read overlaps are determined. The overlying regions often only consider a small percentage of the genome, because the reads are substantially longer than SBH. Furthermore, the overlapping areas do not have to match completely due to the misreading. As a result, certain arcs are surer of themselves than others. As a result, they are frequently weighted, with the weight correlating, for example, with alignment score and length (the arrangement of two sequences in which related areas are joined together; for a comprehensive description). Furthermore, because reads start from both strands of the DNA double helix, each vertex often includes a double vertex corresponding to the read's opposite complement form. This would not be required if an algorithm could determine the original strand for each read. In practice, this is more difficult to do. Consequently, most traversal algorithms employ the dual vertex approach, which indicates that both vertices have been visited when one of them has been reached. Likewise, in the original model, identifying a Hamiltonian route might theoretically be utilized to produce a result in a defined graph G. Due to the weighting of the graph at this point, the route should ideally satisfy the criterion of greatest profit or least cost, which is a variant of the Traveling Salesman Problem. However, this method presents a variety of practical difficulties. First, because of data inaccuracies and changes in the amount of coverage, there is no certainty that a Hamiltonian route exists. For example, Graph G may be unconnected due to genome sequencing gaps. In this example, each linked component of graph G may be scanned for a Hamiltonian route, yielding a collection of disconnected contigs or continuous regions of a genome. However, the second challenge is the graph structure's vastness, which precludes the adoption of any exponential technique. As a result, only heuristic algorithms that use somewhat long pathways to cover the network topology are considered in practice. McPherson et al. [35] disclose an intriguing exception in which the authors offer a plan to resolve the Minimum Path Covering Problem in human genome, which is NP-hard in general, but straightforward for acyclic digraphs [5]. The central challenge was to develop a heuristic technique for converting the overlap graph (see Figure 5) to an acyclic one. However, most overlap graphs are not acyclic, which presents a significant challenge with this technique. The recent developments of DNA sequencing based on graph theory is shown in Table 1.
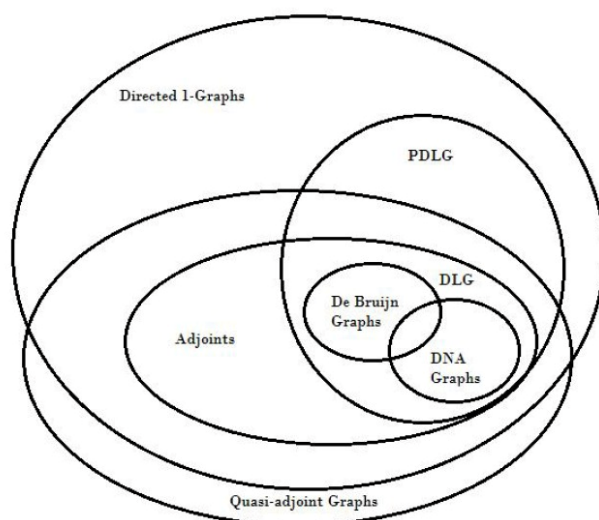

Figure 5. Belongingness of DNA graphs [16].

Table 1. Graphs and recent developments in DNA sequencing.

Year	Contributions	Authors
2022	• Improved One-Time Pad DNA Structure • Used Graph Labeling Techniques	B. Deepa, and V. Maheswari [34]
2022	• Analysis of Graph Neural Networks • Prediction model of DNA Double-Strand Breakage	XU Wang et al. [36]
2021	• Effective Dynamic de Bruijn Graphs • Made Possible by Buffering Updates	J Alanko et al. [37]
2019	• DNA Rearrangement Polynomials • Discussion on related Graphs	Robert Brijder et al. [38]
2018	• Graph algorithms for DNA sequencing • Studied on existing models, and predicted a model using graphs	Jacek Blazewicz et al. [39]

6. Conclusions and Future Research Directions

In this study, the topics of DNA Library, DNA Graphs, de-Bruijn graphs, and Overlap graphs in DNA sequencing were analyzed. Further, different DNA sequencing techniques were described in connection with different diseases. Different graphing techniques are useful for visualizing deficiencies in DNA sequencing, as well. In the modeling of DNA graphs, most cases occur randomly and uncertainly. De-Bruijn graphs have some ambiguity in terms of overlapping or repeating areas. Thus, the fuzziness involved in DNA sequencing will help analyze the characteristics properly. The fuzzy and uncertain sites regarding diseases due to DNA patterns can be targeted based on this study. This area will be covered in our next articles.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Godbole, A.; Knisley, D.; Norwood, R. Some properties of alphabet overlap graphs. *arXiv* **2005**, arXiv:math/0510094.
2. Bhavadharani, R.K.; Nagarajan, V.; Chandiramouli, R. Density functional study on the binding properties of nucleobases to stanane nanosheet. *Appl. Surf. Sci.* **2018**, *462*, 831–839. [CrossRef]
3. Gilbert, W. DNA sequencing and gene structure. *Science* **1981**, *214*, 1305–1312. [CrossRef] [PubMed]
4. Formanowicz, P.; Kasprzak, M.; Wawrzyniak, P. Labeled Graphs in Life Sciences—Two Important Applications. In *Graph-Based Modelling in Science, Technology and Art*; Springer: Cham, Switzerland, 2022; pp. 201–217.
5. Blazej, R.G.; Kumaresan, P.; Mathies, R.A. Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 7240–7245. [CrossRef]
6. Rusinova, D.E.; Stroganov, Y.V. Model Formalization for Genomes Comparative Analysis Using a Graph Database. In Proceedings of the 2022 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), Saint Petersburg, Russia, 25–28 January 2022; pp. 1593–1596.
7. Karunasena, W.W.P.M.T.M.; Wijesiri, G.S. Application of Graph Theory in DNA similarity analysis of Evolutionary Closed Species. *Psychol. Educ.* **2021**, *58*, 3428–3434.
8. Berstel, J.; Perrin, D. The origins of combinatorics on words. *Eur. J. Comb.* **2007**, *28*, 996–1022. [CrossRef]
9. Hutchison, C.A., III. DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Res.* **2007**, *35*, 6227–6237. [CrossRef]
10. Noual, M. Updating Automata Networks. Ph.D. Dissertation, Ecole Normale Supérieure de Lyon-ENS LYON, Lyon, France, 2012.

11. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194. [CrossRef]
12. Blazewicz, J.; Hertz, A.; Kobler, D.; de Werra, D. On some properties of DNA graphs. *Discret. Appl. Math.* **1999**, *98*, 1–19. [CrossRef]
13. Gresham, D.; Dunham, M.J.; Botstein, D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* **2008**, *9*, 291–302. [CrossRef]
14. Ewing, B.; Hillier, L.; Wendl, M.C.; Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **1998**, *8*, 175–185. [CrossRef] [PubMed]
15. Healy, K. Nanopore-based single-molecule DNA analysis. *Nanomedicine* **2007**, *2*, 459–481. [CrossRef] [PubMed]
16. Stevens, B.; Williams, A. The coolest way to generate binary strings. *Theory Comput. Syst.* **2014**, *54*, 551–577. [CrossRef]
17. Kari, L.; Xu, Z. de Bruijn sequences revisited. *Int. J. Found. Comput. Sci.* **2012**, *23*, 1307–1321. [CrossRef]
18. Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467. [CrossRef]
19. Kao, J.Y.; Shallit, J.; Xu, Z. The Frobenius problem in a free monoid. *arXiv* **2007**, arXiv:0708.3224.
20. Pevzner, P.A. l-tuple DNA sequencing: Computer analysis. *J. Biomol. Struct. Dyn.* **1989**, *7*, 63–73. [CrossRef]
21. Shendure, J.; Aiden, E.L. The expanding scope of DNA sequencing. *Nat. Biotechnol.* **2012**, *30*, 1084–1094. [CrossRef]
22. Adleman, L.M. Molecular computation of solutions to combinatorial problems. *Science* **1994**, *266*, 1021–1024. [CrossRef]
23. Lipshutz, R.J. Likelihood DNA sequencing by hybridization. *J. Biomol. Struct. Dyn.* **1993**, *11*, 637–653. [CrossRef]
24. Kasprzak, M. On the link between DNA sequencing and graph theory. *Comput. Methods Sci. Technol.* **2004**, *10*, 39–47. [CrossRef]
25. Pirzada, S. Applications of graph theory. In *PAMM: Proceedings in Applied Mathematics and Mechanics*; WILEY-VCH Verlag: Berlin, Germany, 2007; Volume 7, p. 2070013.
26. Margulies, M.; Egholm, M.; Altman, W.E.; Attiya, S.; Bader, J.S. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376–380. [CrossRef] [PubMed]
27. Agüero-Chapin, G.; Galpert, D.; Molina-Ruiz, R.; Ancede-Gallardo, E.; Pérez-Machado, G.; De la Riva, G.A.; Antunes, A. Graph Theory-Based Sequence Descriptors as Remote Homology Predictors. *Biomolecules* **2020**, *10*, 26. [CrossRef] [PubMed]
28. Alanko, J.; Slizovskiy, I.; Lokshtanov, D.; Gagie, T.; Noyes, N.; Boucher, C. Syotti: Scalable Bait Design for DNA Enrichment. *bioRxiv* **2021**. [CrossRef]
29. Ley, R.E.; Bäckhed, F.; Turnbaugh, P.; Lozupone, C.A.; Knight, R.D.; Gordon, J.I. Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11070–11075. [CrossRef]
30. Lu, C.; Meyers, B.C.; Green, P.J. Construction of small RNA cDNA libraries for deep sequencing. *Methods* **2007**, *43*, 110–117. [CrossRef]
31. Kasprzak, M. Classification of de Bruijn-based labeled digraphs. *Discret. Appl. Math.* **2018**, *234*, 86–92. [CrossRef]
32. Kumar, A. De-Bruijn Sequence-And Application in Graph theory. *Int. J. Progress. Sci. Technol.* **2016**, *3*, 4–17.
33. Kumar, A.; Pruthi, M. Application of eulerian graph. *Arya Bhatta J. Math. Inform.* **2018**, *10*, 295–298.
34. Deepa, B.; Maheswari, V. An enhanced DNA structure for one-time pad together with graph labeling techniques. *AIP Conf. Proc.* **2022**, *2385*, 130045.
35. McPherson, J.D.; Marra, M.; Hillier, L.D.; Waterston, R.H.; Chinwalla, A.; Wallis, J.; Sekhon, M.; Wylie, K.; Mardis, E.R.; Wilson, R.K.; et al. A physical map of the human genome. *Nature* **2001**, *409*, 934–941. [PubMed]
36. Wang, X.U.; Zhao, H.; Tu, W.; Li, H.; Sun, Y.; Bo, X. Graph Neural Networks for Double-Strand DNA Breaks Prediction. *arXiv* **2022**, arXiv:2201.01855.
37. Alanko, J.; Alipanahi, B.; Settle, J.; Boucher, C.; Gagie, T. Buffering Updates Enables Efficient Dynamic de Bruijn Graphs. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4067–4078. [CrossRef] [PubMed]
38. Brijder, R.; Hoogeboom, H.J.; Jonoska, N.; Saito, M. Graphs Associated With DNA Rearrangements and Their Polynomials. In *Algebraic and Combinatorial Computational Biology*; Academic Press: Cambridge, MA, USA, 2019; pp. 61–87.
39. Blazewicz, J.; Kasprzak, M.; Kierzynka, M.; Frohmberg, W.; Swiercz, A.; Wojciechowski, P.; Zurkowski, P. Graph algorithms for DNA sequencing–origins, current models and the future. *Eur. J. Oper. Res.* **2018**, *264*, 799–812. [CrossRef]