International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 02 | February - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Computational Modelling: A Study of Different DNA Sequencing Using DNA Graphs

Dr C K Gomathy-Assistant Professor, Department of CSE, SCSVMV Deemed to be University, India

Nanda Pavan Kumar T-UG Scholar, Department of CSE, SCSVMV Deemed to be University, India

Abstract

Natural genetic material may help identify genetic abnormalities and provide insight into the workings of gene expression systems. Disorders associated with chromosomal abnormalities include single nucleotide polymorphisms (SNPs), minor insertions and deletions, and significant chromosomal aberrations. In order to analyse DNA sequences, one of the most important components of biological study, various techniques have been used. Thus, DNA analysis and computing have benefited greatly from a variety of mathematical and algorithmic advances. Sequencing systems are constructed on a quantitative framework, and their workings include cost minimization, deployment, and sensitivity analysis for various parameters. In order to analyse various diseases using DNA, this study will look into the role of DNA sequencing and how it is represented in graphs.

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

I. INTRODUCTION

All living things receive genetic information from DNA, a molecule. Its double helix structure is made up of two lengthy polynucleotide chains that are joined by covalent bonds. Deoxyribose, a sugar with five carbons, is the backbone of each of these chains and is joined to phosphate groups and bases that contain nitrogen. Adenine (A), thymine (T), guanine (G), or cytosine (C) are possible candidates for the nitrogenous base (C). A and T are constant partners, and G and C DNA sequence study is crucial for understanding cell structures and activities and solving biological mysteries since genomic DNA often plays a large role in determining individuals and species [1].

Molecular cloning, breeding, identifying diseases and genetic abnormalities, comparative and evolutionary study, and other uses of DNA sequencing technology may benefit scientists and medical professionals. Therefore, the ideal DNA sequencing method should be speedy, precise, easy to use, and affordable. In the past 30 years, DNA sequencing techniques and applications have evolved significantly, and today they are used to define genome ages using vast volumes of genomic data. There are several study topics and applications as a result. For the evaluation of next-generation sequencing systems (NGS), including pyrosequencing, genome analyzer and high-throughput sequencing, and SOLID sequencing, it is crucial to understand the development history of sequencing technology. Additionally, it aids in contrasting the technology's benefits and drawbacks and outlining the range of applications. Additionally, the capacity to assess personal genome machines (PGM) and third-generation technologies and their applications is crucial to understanding the history of sequencing development. The majority of the data and conclusions are

derived from independent users with extensive hands-on experience using the standard NGS tools from BGI (the Beijing Genomics Institute).

II. THE HISTORY OF DNA SEQUENCING

Friedrich Miescher made the initial discovery of deoxyribonucleic acid (DNA) in the year 1869. Using a location-specific primer extension strategy developed by Ray Wu at Cornell University in 1970, the first technique for identifying DNA sequences was used. By using RNA to degrade the virus's RNA, isolate oligonucleotides, and then separate them using electrophoresis and chromatography, Walter Fiers became the first person to sequence the DNA of a full gene (the gene encoding the coat protein of the bacteriophage MS2) in 1972. In 1977, Fredrick Sanger created the first DNA sequencing technique using radio labelled partially digested fragments, known as the "chain termination method," after continuing to develop an alternate DNA sequencing technique. In 1980, he received the Nobel Prizes additionally in 1977, Maxam and Gilbert presented a technique for DNA sequencing that relied on chemically altering DNA. At Applied Biosystems, Inc. (ABI), Leroy Hood and Michael Hunkapiller achieved success in automating the Sanger sequencing procedure in 1987. DNA sequencing has advanced to new levels and eventually reached the next generation as a result of ongoing inventive efforts.

III. DNA LIBRARY

One way to think of a DNA library is as a collection of DNA sequences. Like a book or any other kind of data library, a DNA library can be used to store and make information available. The data carrier in a similar kind of library is a DNA chain. These chains typically reflect real genetic information discovered through biochemical research. DNA can provide details on gene expression activities or help identify hereditary diseases. However, some applications demand DNA chains with unique properties that are rarely found in naturally occurring DNA molecules. Adleman [1] created DNA computing in 1994 as an example of one such use. The Hamiltonian route problem was solved by encoding graph vertices as randomly generated DNA chains of length l and arcs as a concatenation of two chains: one supplemental to the last 1/2 nucleotides of the preceding vertex and one extra to the first 1/2 nucleotides of the succeeding vertex. Vertices and arcs would hybridise in a perfect world, but random chain construction might prevent this from happening in the graph in question. An strategy for creating DNA libraries made of chains with a low inclination for hybridization is presented in the research. This type of library can therefore be used in the future to encode a group of vertices. This method has important mathematical implications. A generalised de Bruijn sequence [2] does not have complementarity as a result. First draughts of the method based on de Bruijn graphs were reported in [8] and [9], respectively. The strategy is based on graph theory, specifically on enlarged de Bruijn graphs called lexical graphs. Lexical graphs were first introduced in [10], and independently related concepts were discovered in [11,12]. They were referred to as alphabet overlap digraphs in [12] and word graphs in [13]. A de Bruijn graph is Eulerian cycles are correlated to de Bruijn sequences [2]. A lexical sequence, which is an enlarged de Bruijn sequence, corresponds to each Eulerian cycle in a lexical graph. The term "lexical sequences" refers to the presentation of De Bruijn sequences with multiple shifts in [9] and separately in [14]. De Bruijn sequences with numerous shifts were created in reference [14] in order to solve the Frobenius issue in a free monoid [15]. De Bruijn graphs with a connection to the DNA are subgraphs of labelled graphs, as explained in [4].

IV. DNA GRAPHS FOR DNA SEQUENCING

The DNA graphs fall within the category of labelled digraphs and were created using the methods described in Lysov et al. The Hamiltonian route problem can be solved polynomially for the labelled digraphs because one of their characteristics is that they are directed line graphs. In order to do this, a directed line graph is first transformed into its original graph, and then an Eulerian route is looked for in the original graph. A directed line graph G and its beginning (directed) graph H must adhere to the following rules: An arc (x, y) occurs in G if and only if the terminal endpoint of arc x in H is also the beginning endpoint of arc y in H. The vertices of G correspond to the arcs of H. A Hamiltonian track in G requires an Eulerian route in H, which is both a necessary and sufficient condition (not valid for undirected graphs). The initial graph for a DNA graph is a Pevzner graph for the same spectrum. Labeled digraphs called De Bruijn graphs are constructed using all possible labels for a given measurement throughout a specific alphabet.

V. LITERATURE REVIEW

Let's look at a quick overview of DNA sequencing history. In 1977, Frederick Sanger developed DNA sequencing using a chain-termination method (often referred to as Sanger sequencing). Walter Gilbert, however, created a different technique in 1978 that involved chemically altering DNA and then cleaving it at specific nucleotides. Sanger sequencing was chosen as the predominant method in the "first generation" of commercial and laboratory sequencing applications due to its low radioactive nature and high efficiency [2]. At the time, DNA sequencing required the use of radioactive chemicals and was labor-intensive [3]. The first automated sequencing machine (the AB370) was released by Applied Biosystems in 1987. It used capillary electrophoresis to speed up and enhance sequencing accuracy. The AB370 was capable of simultaneously detecting 96 bases at a daily rate of 500 K bases and 600 bases per reading.

Since 1995, the most recent model, the AB3730 xl, has had a daily base generation rate of 2.88 million and a base reading range of up to 900 bases. The "Human Genome Project" [HGP] was completed in 2001 [4] thanks primarily to automatic sequencing devices, Sanger sequencing technology, and related software based on capillary sequencing machines, which first debuted in 1998. This work contributed to the development of a potent new sequencing device that could boost efficiency and speed while reducing labour requirements and expenses. The X-prize has improved next-generation sequencing (NGS) technologies as well [5]. NGS technologies differ from the Sanger method in terms of decreased cost, massively parallel processing, and high throughput. NGS increases accessibility to genome sequences, but biological explanations and subsequent data analysis continue to be a barrier to understanding genomes.

The problem of formalising graph data models for genomic information representation was recently covered by Rusinova and Stroganov [6]. They want to employ the proposed model in a graph database for applications in comparative genomics. It is possible to compare genomes in a variety of ways, and the theory-based graph model aims to capture both the probabilistic and the definitive aspects of the process.

Weighted directed graph modelling, adjacency matrix generation, and representative vector translation theories were established by Karunasena and Wijesiri [7]. These vectors' similarity was determined using distance metrics like Euclidean, Cosine, and Correlation. They used molecular similarity coefficients as distance measurements and kept this method as the underlying method to determine whether this method

applies to any DNA fragments in the genomes under consideration. The results of the typical vector and graph spectrums are contrasted. The new method was tested on the mitochondrial DNA of humans, gorillas, and orangutans. The result remained the same even when DNA fragments had more nucleotides.

To protect our communication from outsiders, we currently employ a one-time pad including cryptography. The four nitrogenous bases A (adenine), T (thymidine), C (cytosine), and G are combined in a distinct way (guanine). An improved DNA structure that will be used with approaches for one-time graph labelling. Here, the DNA code and a binary value are both used to encrypt our message, and the result is a binary code. The decimal representation of the binary value is then changed. The last level of the encrypted message is displayed as a cypher graph using the OTP key. which employ our cutting-edge labelling processes for products. In this case, we use a private key to encrypt and decrypt the message.

A type of DNA damage known as a double-strand DNA break (DSB) can result in atypical chromosomal rearrangement. The expenses and technical difficulties of recent high-throughput experiment-based technologies are quite expensive. In order to forecast DSBs, they developed a graph neural network-based technique called GraphDSB, which makes use of data on chromosome shape and DNA sequence properties. They also introduced Jumping Knowledge architecture and many efficient structural encoding techniques to enhance the model's capacity for expressiveness. The investigations on datasets from the chronic myeloid leukaemia cell line and normal human epidermal keratinocytes (NHEK) confirm the role of structural information in the prediction of DSBs. (K562) and the ablation investigations provide additional evidence of the usefulness of the planned elements in the suggested GraphDSB structure. Finally, they demonstrated the significant contributions of the 5-mer DNA sequence features and the two chromatin interaction modes by using GNNExplainer to analyse the impact of node properties and topology to DSBs prediction.

The de Bruijn graph has also developed into a widely used graph model for biological data since it was first introduced in the late 1990s. In addition to genome assembly (Zerbino and Birney, 2008; Bankevich et al., 2012; Peng et al., 2012), variant discovery (Alipanahi et al., 2020b; Iqbal et al., 2012), and storage of assembled genomes, it has also been utilised for other applications (Chikhi et al., 2016). Because of this, there are more than a dozen techniques for quickly and efficiently creating and displaying the de Bruijn graph and its variations. [18]

DNA rearrangement is a common biological process that occurs at both the developmental and evolutionary levels. The rearrangement process can be explicitly modelled by four-regular graphs, as shown above, but numerous models have taken into account the abstract operations required. They discussed how double occurrence words and four-regular graphs could be used to explain DNA rearrangements. These concepts are demonstrated by the DNA recombination processes that occur during rearrangement in a well-researched ciliate species. The total number of molecules that are produced as well as any intermediate molecules can be tallied as terms in graph polynomials, which are closely connected to the well-known Tutte polynomial. [19]

Not everyone is aware of the underlying computational methodologies given the pervasiveness of next-generation sequencing in contemporary biological, genetic, pharmacological, and medical research. Even fewer scientists are aware of the history of the models used today to describe DNA assembly. They discussed the features and relationships between an original graph model that was employed in DNA

sequencing by hybridization. Additionally, they described how these graph models changed over time to accommodate the features of next-generation sequencing. We also show a useful comparison of graphs based on overlap and breakdown that represent these altered models and depict state-of-the-art DNA de novo assembly techniques. Even though there is fierce rivalry, certain assemblers outperform others, and significant variations in how hardware resources are used can be seen. Finally, they highlight the key developments in sequencing and provide predictions about how these may affect computational models in the future. [20].

Table 1: Summary of Graphs and recent developments in DNA sequencing.

Year	Author	Title	Contributions
2022	B. Deepa, and V. Maheswari [16]	An enhanced DNA structure for one-time pad together with graph labelling techniques	· Improved One-Time Pad DNA Structure · Used Graph Labelling Techniques
2022	XU Wang et al. [17]	Graph Neural Networks for Double-Strand DNA Breaks Prediction	· Analysis of Graph Neural Networks · Prediction model of DNA Double-Strand Breakage
2021	J Alanko et al. [18]	Buffering Updates Enables Efficient Dynamic de Bruijn Graphs	· Effective Dynamic de Bruijn Graphs · Made Possible by Buffering Updates
2019	Robert Brijder et al. [19]	Graphs Associated With DNA Rearrangements and Their Polynomials	· DNA Rearrangement Polynomials · Discussion on related Graphs
2018	Jacek Blazewicz et al. [20]	Graph algorithms for DNA sequencing–origins, current models and the future	· Graph algorithms for DNA sequencing · Studied on existing models, and predicted a model using graphs

VI. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

This review article examines the graph-theoretic method of DNA sequencing. Studies are also conducted to examine disorders caused by DNA sequencing issues. In this work, de-Bruijn graphs, overlap graphs, DNA libraries, and DNA graphs in DNA sequencing were examined. Additionally, various DNA sequencing methods were discussed in relation to various disorders. Defects in DNA sequencing can be visualised using a variety of graphing tools. The majority of situations in DNAgraph modelling happen haphazardly and uncertainly. Regarding overlapping or repeated areas, De-Bruijn graphs are not entirely clear. Thus, the fuzziness inherent in DNA sequencing will aid in a proper analysis of the traits. Based on the results of this investigation, the hazy and ambiguous areas surrounding diseases caused by DNA patterns can be targeted.

References

1. Godbole, A.; Knisley, D.; Norwood, R. Some properties of alphabet overlap graphs. *arXiv* **2005**, arXiv:math/0510094. [**Google Scholar**]
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. [**Google Scholar**] [**CrossRef**]
3. Gilbert, W. DNA sequencing and gene structure. *Science* **1981**, *214*, 1305–1312. [**Google Scholar**] [**CrossRef**] [**PubMed**][**Green Version**]
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. [**Google Scholar**]
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. [**Google Scholar**] [**CrossRef**][**Green Version**]
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. [**Google Scholar**]
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. [**Google Scholar**]
8. Berstel, J.; Perrin, D. The origins of combinatorics on words. *Eur. J. Comb.* **2007**, *28*, 996–1022. [**Google Scholar**] [**CrossRef**][**Green Version**]
9. Hutchison, C.A., III. DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Res.* **2007**, *35*, 6227–6237. [**Google Scholar**] [**CrossRef**]
10. Noual, M. Updating Automata Networks. Ph.D. Dissertation, Ecole Normale Supérieure de Lyon-ENS LYON, Lyon, France, 2012. [**Google Scholar**]
11. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194. [**Google Scholar**] [**CrossRef**][**Green Version**]
12. Blazewicz, J.; Hertz, A.; Kobler, D.; de Werra, D. On some properties of DNA graphs. *Discret. Appl. Math.* **1999**, *98*, 1–19. [**Google Scholar**] [**CrossRef**][**Green Version**]
13. Gresham, D.; Dunham, M.J.; Botstein, D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* **2008**, *9*, 291–302. [**Google Scholar**] [**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. [**Google Scholar**] [**CrossRef**] [**PubMed**][**Green Version**]
15. Healy, K. Nanopore-based single-molecule DNA analysis. *Nanomedicine* **2007**, *2*, 459–481. [**Google Scholar**] [**CrossRef**] [**PubMed**]
16. Deepa, B.; Maheswari, V. An enhanced DNA structure for one-time pad together with graph labeling techniques. *AIP Conf. Proc.* **2022**, *2385*, 130045. [**Google Scholar**]
17. 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. [**Google Scholar**]
18. 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. [**Google Scholar**] [**CrossRef**] [**PubMed**]
19. 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. [**Google Scholar**]
20. 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. [**Google Scholar**] [**CrossRef**]
21. Dr.C K Gomathy and et al, Machine Learning-Based Clinical Decision Support System, International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 06 Issue: 10 | October - 2022 Impact Factor: 7.185 ISSN: 2582-3930

22.Dr.C K Gomathy et al,Web Service Composition In A Digitalized Health Care Environment For Effective Communications, Published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 4, April 2016, ISSN: 2278 – 1323.
23. Vishnupriya C K and et al,Dimensional and Morphologic Variations of palatal Rugae-a hospital based study among Chennai populations, International Journal Of Science Research, ISSN No: 2277-8179 Volume 7, Issue 7,P.No-19-20,July ’2018
24. Dr.C K Gomathy et al, Machine Learning-Based Clinical Decision Support System, International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 06 Issue: 10 | October - 2022 Impact Factor: 7.185 ISSN: 2582-3930
25. Dr.C K Gomathy et al,A Review On IOT Based Covid-19 Patient Health Monitor In Quarantine, International Research Journal of Engineering and Technology (IRJET),e-ISSN: 2395-0056 Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072`
26. Dr.C K Gomathy, et al, A Medical Information Security Using Cryptosystem For Wireless Sensor Networks, International Journal Of Contemporary Research In Computer Science And Technology (Ijcrcst) E-Issn: 2395-5325 Volume3, Issue 4, P.No-1-5,April ’2017
27. Dr.C K Gomathy and et al, The Parkinson’s Disease Detection Using Machine Learning Techniques, International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 10 | Oct 2021, e-ISSN: 2395-0056, p-ISSN: 2395-0072.
28. C K Gomathy and et.all, Multi-Source Medical Data Integration And Mining For Healthcare Services, International Journal Of Early Childhood Special Education (Int-Jecse) Doi:10.9756/Intjecse/V14i5.66 Issn: 1308-5581 Vol 14, Issue 05 2022

29. C K Gomathy and et.all, An Efficient Way To Predict The Disease Using Machine Learning, International Journal Of Early Childhood Special Education (Int-Jecse) Doi:10.9756/Intjecse/V14i5.66 Issn: 1308-5581 Vol 14, Issue 05 2022