Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
www.akademiabaru.com/ard.html
ISSN: 2289-7984

Minimization of DNA String Using Shortest Path Problem

Kee Yeong Chua[1], Wan Heng Fong[1,*], Sherzod Turaev[2]

[1] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bharu, Johor, Malaysia
[2] Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, P.O. Box 15551, Al Ain, United Arab Emirates

ARTICLE INFO

Article history:
Received 3 June 2024
Received in revised form 15 August 2024
Accepted 19 August 2024
Available online 30 August 2024

Keywords:
Deoxyribonucleic graph theory; shortest path problem; mathematical model

ABSTRACT

The structure of deoxyribonucleic acid (DNA) consists of two nucleotides chained together where nucleotides are long strands of DNA bases, namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), linked together by sugar and phosphates molecules. DNA molecule can be represented by a graph as a mathematical model. In graph theory, finding the shortest path between two vertices in a graph is called the shortest path problem. In this research, a DNA string is represented in graphical form so that the shortest path problem can be applied to minimize the DNA string. The vertex of the graphical form of a DNA string is formed by base pairs of length two, while the shortest length of bases between the vertices acts as the edges of the graph. In this process, four graphs are formed, classified by four initial bases used for the vertices. From the graphs generated, the shortest path for all vertices is calculated where any untraversed path is removed from each of the graphs, forming a simplified graph for each case. Next, by finding the Euler path for the simplified graph and converting the path taken back into a DNA string, a minimized DNA string is formed. From the results, it is shown that, the graph with initial base of T produces the shortest minimized DNA string while the graph with initial base of C produces the longest minimized DNA string.

1. Introduction

Deoxyribonucleic acid (DNA) is found in all living organisms and it has many functions. One of the functions of DNA is storing and transmitting genetic information from one generation to the next [1]. Other than that, DNA also contains the genetic information responsible for the development, functioning and reproduction of all living organisms [2]. Watson *et al.,* [3] found that the structure of DNA is a double helix shape formed by numerous nucleotides that are chained together into two long strands, where the term nucleotide refers to a base pair that is chemically attached to a sugar and a phosphate molecule [3]. These base pairs are two opposite bases linked together chemically by hydrogen bonds where Adenine (A) and Thymine (T) are bases opposing to each other; also, Cytosine (C) and Guanine (G) are bases opposing to each other [4]. This structure proposed by Watson and Crick marks an important milestone in molecular biology.

* *Corresponding author.*
E-mail address: fwh@utm.my

https://doi.org/10.37934/ard.119.1.2735

In graph theory, a graph is a mathematical structure that models pairwise relations between objects [5]. There are two sets in a graph which is the set V for vertices and the set E for edges [6]. The elements in the set V is denoted as $v \in V$ while elements in the set E is denoted as $e \in E$ [7]. Other than that, a directed edge, e from a vertex u to a vertex v is denoted as $e = \overrightarrow{\langle u, v \rangle}$ [8] and the edge weight is denoted as $W(E)$ [9]. In addition, a path in a graph is an alternately placed series of edges and vertices that starts and ends with a vertex where all its vertices are distinct [10].

From Bernart and Prijith [11], graph theory is used in many areas of biology where the concepts of graph theory can be used to study the structures of DNA. An example of graph theory applied to DNA is the research done by Ekim *et al.,* [12] who used minimizer-space de Brujin graphs to enable long-read genome assembly. Another example of graph theory applied in DNA is the use of adjacency matrix in graph representation to classify individuals according to their different ancestral lineages [13]. Graph theory provides a rich set of mathematical tools for analysing the properties of graphs, including the degree distribution, clustering coefficient, shortest path and many others [14]. One example of utilizing analytical methods of graph theory is using graph colouring to optimize diabetes cupping point [15]. The shortest path problem in graph theory is used to find the path with minimum weight among all possible paths between two nodes or vertices in a graph [16]. This shortest path problem can be solved using multiple algorithms such as Dijkstra's algorithm [17] and the Bellman-Ford algorithm [18]. The use of DNA algorithms for computation of shortest path problems is also possible [19,20]. However, the use of shortest path problem on research regarding DNA is rare, if any. Thus, this research aims to apply the shortest path problem in graphs of DNA strings to minimize the length of a DNA string.

2. Methodology

This section discussed on the methods used in the selection of a DNA string as well as the generation of graphs for the selected string. The calculation of shortest path for the generated graphs and the minimized DNA string for each path calculated was also discussed. Furthermore, the graphs are simplified based on the shortest paths calculated and hence the minimized DNA string is obtained.

For the selection of DNA string, a random number generator is used to get a starting point for obtaining a DNA string of 80 base pairs from the DNA of Bacteriophage lambda cI 857 Sam7 (Lambda). From the DNA string of lambda, a base of length two is used as the vertices of the graph. There are four bases in DNA, namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), thus there are 16 possible pairs for the vertices which forms the vertex set $V = \{\mathrm{AA}, \mathrm{AC}\ \mathrm{AG}, \mathrm{AT}, \mathrm{CA}, \mathrm{CC}, \mathrm{CG}, \mathrm{CT}, \mathrm{GA}, \mathrm{GC}, \mathrm{GG}, \mathrm{GT}, \mathrm{TA}, \mathrm{TC}, \mathrm{TG}, \mathrm{TT}\}$. Then, the vertex set V is split into four cases, namely V_a, V_c, V_g and V_t where each case is classified by the initial base. The vertex sets formed are as in Eq. (1),

$$\begin{aligned} V_a &= \{\mathrm{AA}, \mathrm{AC}, \mathrm{AG}, \mathrm{AT}\}, \\ V_c &= \{\mathrm{CA}, \mathrm{CC}, \mathrm{CG}, \mathrm{CT}\}, \\ V_g &= \{\mathrm{GA}, \mathrm{GC}, \mathrm{GG}, \mathrm{GT}\}, \\ V_t &= \{\mathrm{TA}, \mathrm{TC}, \mathrm{TG}, \mathrm{TT}\} \end{aligned} \tag{1}$$

Following that, a path graph for all elements in the vertex sets for the DNA string was drawn in which the vertices were enumerated with the position of the bases in the string while the edge weight is the number of bases between the base acting as the vertices. More specifically, let $m < n$ where $m, n \in \mathbb{Z}^+$, be the label for the positions of the bases of the vertex in the DNA string. Then, the edge

weight from a start vertex u to an end vertex v, denoted as $W\left(\langle\overrightarrow{u(m),v(n)}\rangle\right)$, can be calculated using the Eq. (2);

$$W\left(\langle\overrightarrow{u(m),v(n)}\rangle\right) = n - (m+2) \quad (2)$$

Using the path graph for each vertex set above, the graph for all paths between all possible start and end vertices was drawn. Since this graph is quite complex, it is represented in tabular form. Next, any start vertex (including the edges) that has a negative length was removed to avoid an infinitely small path length when the shortest path is calculated. Furthermore, the graphs formed above have multiple vertices representing the same DNA bases. Hence, these graphs were reduced by combining all similar vertices into one where only the edges with the shortest edge weight were kept. The vertex set of the reduced graph is represented by V'_a, V'_c, V'_g and V'_t.

The shortest path of each graph was then calculated for all possible start and end vertices, where the calculation was done by going through all possible paths. Each of these calculated paths was then converted to a minimized string by converting the DNA bases represented by the vertex and edge weight back into a DNA string and then ordering the DNA string according to the traversed path.

After the calculation of the shortest path, some edges are not involved in the calculation. As such, each of the graph can be simplified such that these edges are removed. A Euler path is a route that covers every edge exactly once [21]. Therefore, the Euler path of the simplified graphs is then used to form a minimized DNA string by converting the path back into a string using the same method as stated above.

3. Results

This section presents the results obtained where a DNA string of 80 bases long is represented in a graphical form and minimized using the shortest path problem.

A DNA string was taken from the 12841st base to the 12920th base of Lambda which is GCGTGGGGAA TCTTTACCGG CTGATGCGCG GCTATGCCAC CGGCGGTTAT GTCGGTACAC CGGGCAGCAT GGCAGACAGC. This string of 80 base pairs is denoted as α. For the case of vertex with base of length two and initial base of A, the path graph for all elements in the vertex set, $V_a = \{\mathrm{AA}, \mathrm{AC}, \mathrm{AG}, \mathrm{AT}\}$ for α is shown in Figure 1.

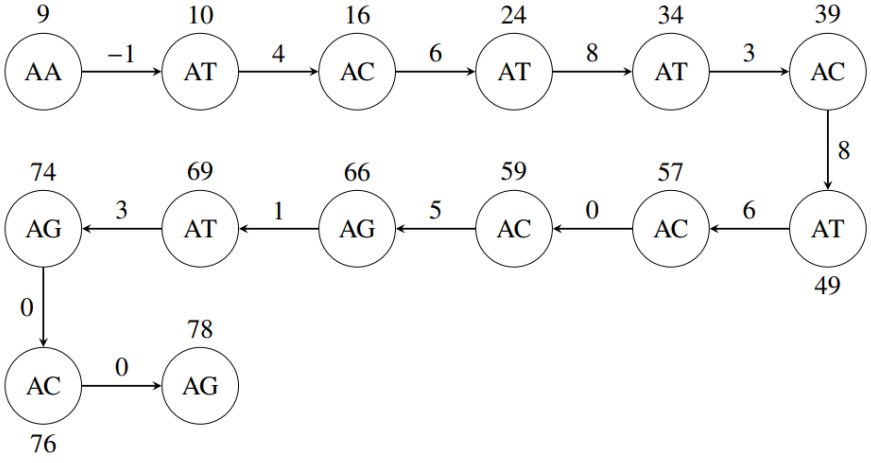

Fig. 1. Graph for all elements in V_a for α

From this graph, using the formula of calculating the edge weight, $W\left(\langle\overrightarrow{u(m),v(n)}\rangle\right) = n - (m+2)$, the table for all paths for the graph for all possible start and end vertex is shown in Table 1. Note that the grey colored cells are the edges of the reduced graph.

Table 1
All paths for each possible start and end vertex in the graph with vertex set V_a

Start vertex \ End vertex	Position	AA	AT	AC	AT	AT	AC	AT	AC	AC	AG	AT	AG	AC	AG
		9	10	16	24	34	39	49	57	59	66	69	74	76	78
AA	9	-	-1	5	13	23	28	38	46	48	55	58	63	65	67
AT	10	-	-	4	12	22	27	37	45	47	54	57	62	64	66
AC	16	-	-	-	6	16	21	31	39	41	48	51	56	58	60
AT	24	-	-	-	-	8	13	23	31	33	40	43	48	50	52
AT	34	-	-	-	-	-	3	13	21	23	30	33	38	40	42
AC	39	-	-	-	-	-	-	8	16	18	25	28	33	35	37
AT	49	-	-	-	-	-	-	-	6	8	15	19	23	25	27
AC	57	-	-	-	-	-	-	-	-	0	7	10	15	17	19
AC	59	-	-	-	-	-	-	-	-	-	5	8	13	15	17
AG	66	-	-	-	-	-	-	-	-	-	-	1	6	8	10
AT	69	-	-	-	-	-	-	-	-	-	-	-	3	5	7
AG	74	-	-	-	-	-	-	-	-	-	-	-	-	0	2
AC	76	-	-	-	-	-	-	-	-	-	-	-	-	-	0
AG	78	-	-	-	-	-	-	-	-	-	-	-	-	-	-

From Table 1, it is noticed that only the AA vertex produces a negative length, hence the vertex AA was removed, forming the vertex set $V'_a = \{\mathrm{AC},\ \mathrm{AG},\ \mathrm{AT}\}$. Other than that, since there were multiple vertices representing the same bases, these vertices were combined. As for the many edges, only the vertices highlighted in grey were used as they are the shortest edge(s) after the combination of the vertices. As a result, the graph for vertex set of $V'_a = \{\mathrm{AC}, \mathrm{AG}, \mathrm{AT}\}$ is shown in Figure 2. Here, the bases used to form the edge weight is indicated in brackets.

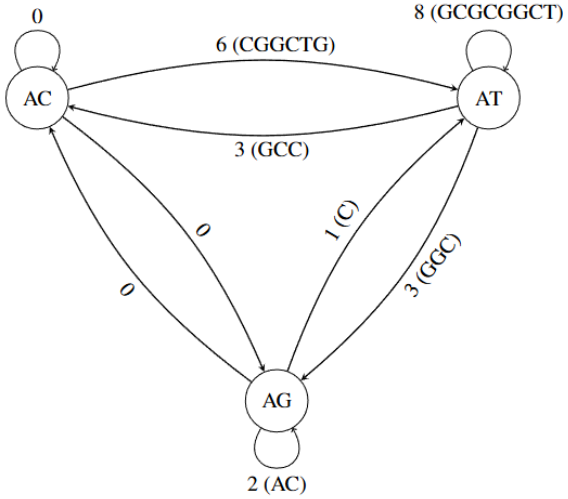

Fig. 2. Reduced graph of vertex set V'_a for α

The shortest path is now calculated for Figure 2 for all possible start and end vertices. From the calculations, the shortest path for each start and end vertex was selected where the paths were then mapped back into a DNA string by converting the DNA bases represented by the vertex and edge

weight back into a DNA string and then ordering the DNA string obtained according to the traversed path. The results are tabulated in Table 2.

Table 2
Shortest path taken, path length and minimized DNA string for vertex set of V'_a of α

Start vertex	End vertex	Shortest path taken	Path length	Minimized DNA string
	AC	$AC \rightarrow AC$	0	ACAC
AC	AG	$AC \rightarrow AG$	0	ACAG
	AT	$AC \rightarrow AG \rightarrow AT$	$0+1=1$	ACAGCAT
	AC	$AG \rightarrow AC$	0	AGAC
AG	AG	$AG \rightarrow AC \rightarrow AG$	$0+0=0$	ACCAT
	AT	$AG \rightarrow AT$	1	ATGCCAC
	AC	$AT \rightarrow AC$	1	ATGCCAC
AT	AG	$AT \rightarrow AG$	3	ATGGCAG
	AT	$AT \rightarrow AG \rightarrow AT$	$3+1=4$	ATGGCAGCAT

Note that the bases in the minimized DNA string obtained from the vertices are colored in red. From Table 2, it is noticed that in the shortest path, not all the edges from Figure 2 were traversed through. Therefore, to simplify the graph in Figure 2, only the path used in the calculation was kept while the other paths were considered obsolete and removed. This simplified graph is shown in Figure 3.

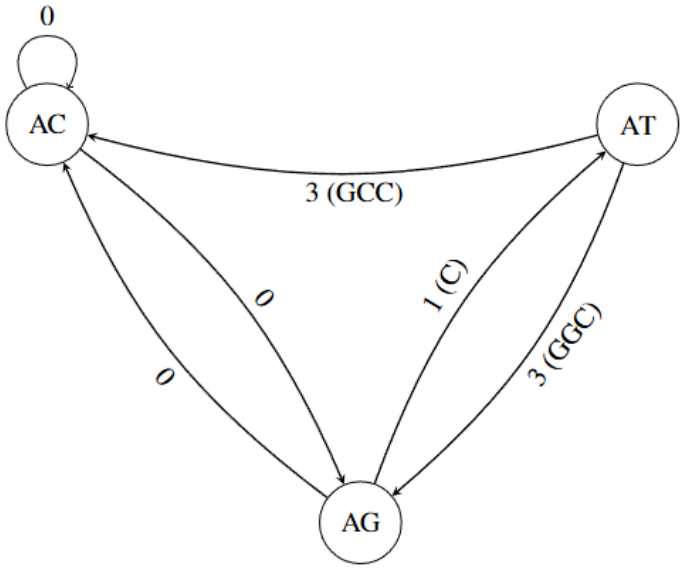

Fig. 3. Simplified graph of vertex set V'_a for α

Subsequently, the Euler path is now obtained for the graph in Figure 3 which is $AT \rightarrow AC \rightarrow AC \rightarrow AG \rightarrow AT \rightarrow AG \rightarrow AC$. By mapping this path back to a DNA string, the string obtained was ATGCCACACAGCATGCCAGAC, which is a minimized DNA string for α. Thus, this concludes the results for the case of vertex with base of length two and initial base of A.

For the other cases, since the calculation process is lengthy, only the results for the shortest path was taken, the simplified graph and the minimized DNA string were presented. The shortest paths and minimized string for each path for the vertex sets V'_c, V'_g and V'_t are shown in Tables 3 to 5, respectively. This was followed with the simplified graphs for the vertex sets of V'_c, V'_g and V'_t are shown in Figures 4 to 6, respectively.

Table 3
Shortest path taken, path length and minimized DNA string for vertex set of V'_c of α

Start vertex	End vertex	Shortest path taken	Path length	Minimized DNA string
CA	CA	$CA \rightarrow CA$	1	CAGCA
	CG	$CA \rightarrow CG$	1	CACCG
	CT	$CA \rightarrow CG \rightarrow CT$	$1+1=2$	CACCGGCT
CG	CA	$CG \rightarrow CA$	2	CGGGCA
	CG	$CG \rightarrow CG$	0	CGCG
	CT	$CG \rightarrow CT$	1	CGGCT
CT	CA	$CT \rightarrow CA$	4	CTATGCCA
	CG	$CT \rightarrow CG$	4	CTTTACCG
		$CT \rightarrow CG$	4	CTGATGCG
	CT	$CT \rightarrow CG \rightarrow CT$	$4+1=5$	CTTTACCGGCT
		$CT \rightarrow CG \rightarrow CT$	$4+1=5$	CTGATGCGGCT

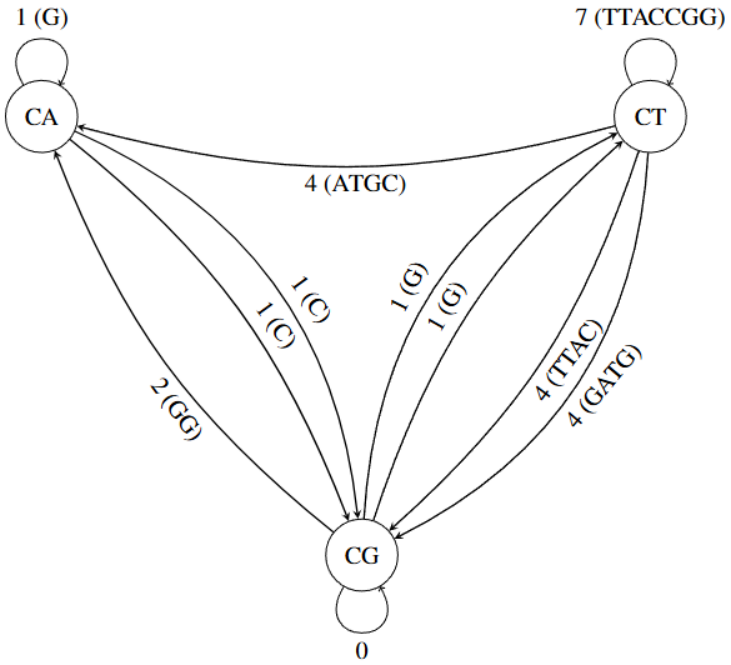

Fig. 4. Simplified graph of vertex set V'_c for α

Table 4
Shortest path taken, path length and minimized DNA string for vertex set of V'_g of α

Start vertex	End vertex	Shortest path taken	Path length	Minimized DNA string
GA	GA	$GA \rightarrow GC \rightarrow GA$	$1+1=2$	GATGCTGA
		$GA \rightarrow GC \rightarrow GA$	$1+1=2$	GATGCAGA
	GC	$GA \rightarrow GC$	1	GATGC
	GT	$GA \rightarrow GC \rightarrow GT$	$1+0=1$	GATGCGT
GC	GA	$GC \rightarrow GA$	1	GCTGA
		$GC \rightarrow GA$	1	GCAGA
	GC	$GC \rightarrow GC$	0	GCGC
	GT	$GC \rightarrow GT$	0	GCGT
GT	GA	$GT \rightarrow GA$	3	GTGGGGA
	GC	$GT \rightarrow GA \rightarrow GC$	$3+1=4$	GTGGGGATGC
	GT	$GT \rightarrow GT$	2	GTCGGT

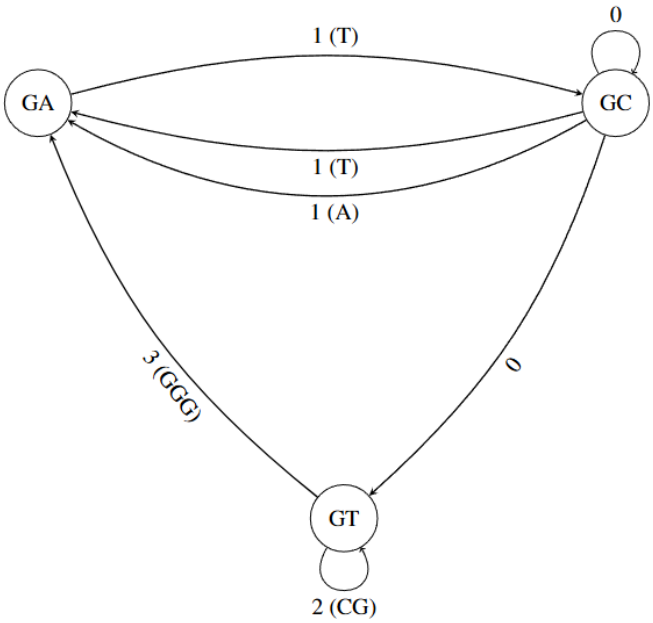

Fig. 5. Simplified graph of vertex set V'_g for α

Table 5
Shortest path taken, path length and minimized DNA string for vertex set of V'_t of α

Start vertex	End vertex	Shortest path taken	Path length	Minimized DNA string
TA	TA	$TA \rightarrow TG \rightarrow TC \rightarrow TA$	$0+0+2=2$	TATGTCTTTA
		$TA \rightarrow TG \rightarrow TC \rightarrow TA$	$0+0+2=2$	TATGTCGGTA
	TC	$TA \rightarrow TG \rightarrow TC$	$0+0=0$	TATGTC
	TG	$TA \rightarrow TG$	0	TATG
TC	TA	$TC \rightarrow TA$	2	TCTTTA
		$TC \rightarrow TA$	2	TCGGTA
	TC	$TC \rightarrow TA \rightarrow TG \rightarrow TC$	$2+0+0=2$	TCTTTATGTC
		$TC \rightarrow TA \rightarrow TG \rightarrow TC$	$2+0+0=2$	TCGGTATGTC
	TG	$TC \rightarrow TA \rightarrow TG$	$2+0=2$	TCTTTATG
		$TC \rightarrow TA \rightarrow TG$	$2+0=2$	TCGGTATG
TG	TA	$TG \rightarrow TC \rightarrow TA$	$0+2=2$	TGTCTTTA
		$TG \rightarrow TC \rightarrow TA$	$0+2=2$	TGTCGGTA
	TC	$TG \rightarrow TC$	0	TGTC
	TG	$TG \rightarrow TG$	1	TGATG

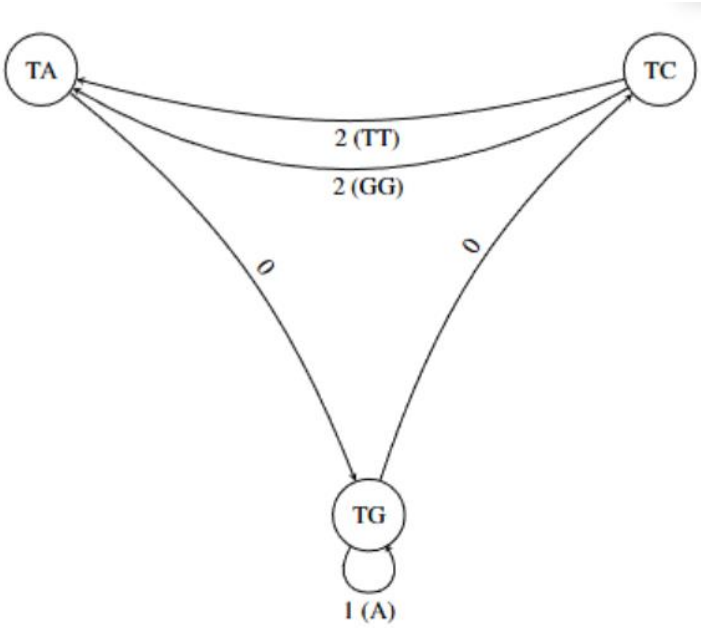

Fig. 6. Simplified graph of vertex set V'_t for α

Then, the minimized string for each of the simplified graphs is tabulated into Table 6. Note that the bases obtained from the vertices are colored in red.

Table 6
Minimized DNA string for V'_a, V'_c, V'_g and V'_t

Vertex set	Minimized DNA string
V'_a	ATGCCACACAGCATGCCAGAC
V'_c	CTATGCCAGCACCGCGGCTTTACCGGCTGATGCGGGCACCG
V'_g	GCTGATGCGCGTCGGTGGGGA
	GCAGATGCGCGTCGGTGGGGA
V'_t	TCTTTATGATGTCGGTA

It can be observed that there were two different minimized DNA string for the vertex set V'_g. This is because the Euler path for the simplified graph for V'_g does not exist. Hence in this situation, only one path was chosen from the two paths of length, one from start vertex GC to end vertex GA so that the Euler path exist which produces two different cases.

4. Conclusions

The shortest path problem in graph theory was applied into graphs of a DNA string. The results showed that calculations of the shortest path can be used in simplification of the graphs. Furthermore, minimized DNA strings have been formed for each graph where it is concluded that the shortest path problem can be used to minimize the DNA string. Afterwards, the difference in the results obtained from the different graphs of the DNA string was discussed.

This research can be further advanced using three base pairs to represent the vertices instead of two base pairs. This will increase the number of vertices which might yield different results in terms of the minimized string. Other than that, since it is known that DNA can be read in two directions, hence further research can be done for the same string by taking into account the other direction.

Acknowledgement

The authors would like to acknowledge Universiti Teknologi Malaysia (UTM) and Research Management Centre for the financial support through Universiti Teknologi Malaysia Fundamental Research (UTMFR) Vote Number QJ130000.3854.22H45.

References

[1] Ghannam, Jack Y., Jason Wang, and Arif Jan. "Biochemistry, DNA Structure." StatPearls Publishing, 2023.
[2] Siti Hajar Mohd Khairuddin, Muhammad Azrin Ahmad, and Mohd Sham Mohamad. 2024. "Classification of N-Th Order Limit Language in Formal Language Classes". *Journal of Advanced Research in Applied Sciences and Engineering Technology* 44 (2):1-10. https://doi.org/10.37934/araset.44.2.110
[3] Crick, Francis Harry Compton, and James Dewey Watson. "The complementary structure of deoxyribonucleic acid." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 223, no. 1152 (1954): 80-96. https://doi.org/10.1098/rspa.1954.0101
[4] Nooradelena Mohd Ruslim, Yuhani Yusof, Mohd Sham Mohamad, Mohd Firdaus Abdul-Wahab, and Faisal. 2024. "Characterize Type of Splicing Languages via Directed Splicing Graph". *Journal of Advanced Research in Applied Sciences and Engineering Technology* 45 (1):129-36. https://doi.org/10.37934/araset.45.1.129136
[5] Bender, Edward A., and S. Gill Williamson. *Lists, decisions and graphs*. S. Gill Williamson, 2010.
[6] Wan Nor Munirah Ariffin, Raveena Subramaniam, Erni Puspanantasari Putri, Muhammad Shahar Jusoh, Muhamad Hafiz Masran, Muhammad Nur Khairul Hafizi Rohani, Yussof Hussin, Noormaizatul Akmar Ishak, Emy Aizat Azimi, and Siti Sharina Mohd Shukri. 2023. "Product Pairing Selection for Promotion Using Partitioning Method". *Journal of Advanced Research in Applied Sciences and Engineering Technology* 30 (3):91-103. https://doi.org/10.37934/araset.30.3.91103.
[7] Chartrand, Gary, and Ping Zhang. *A first course in graph theory*. Courier Corporation, 2013.
[8] Guichard, David. "An introduction to combinatorics and graph theory."*Whitman College-Creative Commons*(2017).
[9] Strang, Gilbert. Linear Algebra and Its Applications. *India: Thomson, Brooks/Cole*, 2006.
[10] Balakrishnan, Rangaswami, and Kanna Ranganathan. *A textbook of graph theory*. Springer Science & Business Media, 2012. https://doi.org/10.1007/978-1-4614-4529-6
[11] Bernard, N. Mary, and G. Siva Prijith. "Graph theory in other subjects." *Sustainable Development for Society, Industrial* 175 (2022).
[12] Ekim, Barış, Bonnie Berger, and Rayan Chikhi. "Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer." *Cell systems* 12, no. 10 (2021): 958-968. https://doi.org/10.1016/j.cels.2021.08.009
[13] Palleschi, Vincenzo, Luca Pagani, Stefano Pagnotta, Giuseppe Amato, and Sergio Tofanelli. "Application of Graph Theory to the elaboration of personal genomic data for genealogical research." *PeerJ Computer Science* 1 (2015): e27.https://doi.org/10.7717/peerj-cs.27
[14] Salih Khasraw, Sanhan Muhammad, Nor Haniza Sarmin, Nur Idayu Alimon, Nabilah Najmuddin, and Ghazali Semil@Ismail. 2024. "Sombor Index and Sombor Polynomial of the Noncommuting Graph Associated to Some Finite Groups". *Journal of Advanced Research in Applied Sciences and Engineering Technology* 42 (2):112-21. https://doi.org/10.37934/araset.42.2.112121
[15] Nur Atikah Aziz, Yuhani Yusof, Hazulin Mohd Radzuan, and Viska Noviantri. 2024. "Optimizing Diabetes Cupping Point from Graph Colouring Perspective". *Journal of Advanced Research in Applied Sciences and Engineering Technology* 48 (1):205-12. https://doi.org/10.37934/araset.48.1.205212
[16] Ortega-Arranz, Hector, Arturo Gonzalez-Escribano, and Diego R. Llanos. *The shortest-path problem: Analysis and comparison of methods*. Springer Nature, 2022. https://doi.org/10.1007/978-3-031-02574-7
[17] Dijkstra, Edsger W. "A note on two problems in connexion with graphs." In *Edsger Wybe Dijkstra: his life, work, and legacy*, pp. 287-290. 2022. https://doi.org/10.1145/3544585.3544600
[18] Shimbel, Alfonso. "Structure in communication nets." In *Proceedings of the symposium on information networks*, pp. 119-203. Polytechnic Institute of Brooklyn, 1954.
[19] Narayanan, Ajit, and Spiridon Zorbalas. "DNA algorithms for computing shortest paths." *Proceedings of genetic programming* 718 (1998): 723.
[20] Ibrahim, Zuwairie, Yusei Tsuboi, Osamu Ono, and Marzuki Khalid. "Direct-Proportional Length-Based DNA Computing for Shortest Path Problem." *Int. J. Comput. Sci. Appl.* 1, no. 1 (2004): 46-60.
[21] Marcus, Daniel A. *Graph theory*. Vol. 53. American Mathematical Soc., 2020. https://doi.org/10.1090/text/053