

# Shortest Path Problem for a Graph of DNA String using Adenine and Guanine in Purine Bases as the Vertex

Chua Kee Yeong and Fong Wan Heng\*

Department of Mathematical Sciences, Universiti Teknologi Malaysia,  
81310 UTM Johor Bahru, Malaysia.

\*Corresponding author: fwh@utm.my

## Article history

Received: 16 November 2024

Received in revised form: 29 May 2025

Accepted: 25 June 2025

Published on line: 15 November 2025

---

**Abstract** In the field of research on Deoxyribonucleic acid (DNA), graph theory can be applied to model the structure of a DNA molecule. In particular, the shortest path problem in graph theory can be used to identify the shortest path between vertices of a graph. Hence, it is possible to apply the shortest path problem for minimization of a DNA string so that the time taken for computation of genome assembly can be reduced. This paper presents a method to represent a DNA string graphically where the shortest path is calculated for the graph generated from the DNA string, and the shortest path is then used to minimize the DNA string. In this research, a DNA string is presented in graphical form by using base pairs of length two as the vertices where the initial bases used are Adenine (A) and Guanine (G) which are the main bases in purine. The number of base pairs between adjacent vertices in the DNA string is represented by the edges. The graph is then reduced by following a given set of rules where the shortest path is calculated for all start and end vertices of the reduced graph. Next, the simplification of the graph is done based on the shortest paths obtained by removing all the untraversed paths where the Euler path for the simplified graph is used to form a minimized DNA string. The result shows that a minimized DNA string can be obtained by simplifying the graph of the DNA string using the shortest path problem.

**Keywords** Deoxyribonucleic Acid (DNA), graph theory, shortest path problem.

**Mathematics Subject Classification** 05C85, 05C90.

## 1 Introduction

In living organisms, Deoxyribonucleic Acid (DNA) has the function to store and transmit genetic information [1]. The main composition of DNA are the four bases which are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) where these bases form base pairs by chemically attaching opposite bases by using hydrogen bonds. For these base pairs, the opposite base for Adenine (A) is Thymine (T) while the opposite base for Cytosine (C) is Guanine (G). These base pairs, after being chemically bonded to sugar and phosphate molecules, are called a nucleotide. Many nucleotides then chain together into two long strands in a double helix shape which forms the structure of DNA [2].

As stated by Bernart and Prijith [3], graph theory has a strong relation with biological studies where the structure of a DNA molecule can be studied using existing concepts in graph theory. Some

examples are the enabling of long-read genome assembly by using minimizer-space de Bruijn graphs [4] and the classification of individuals according to ancestral lineages using adjacency matrix in graph representation [5]. Ashton [6] in 2023 has discussed that overlap graphs, also known as de Bruijn graphs, can give insights into overlapping regions in a DNA sequence. Furthermore, in his paper, it is found that Hamiltonian and Euler paths give information on sequence reconstruction and sequencing methods respectively. In addition, Hamiltonian paths, Euler paths and de Bruijn graphs have been shown to successfully sequence a genome as discussed by Chikomana and Hu [7]. It can be noticed that many concepts in graph theory have been used in research on DNA but not using the shortest path problem.

The shortest path problem in graph theory is used to obtain a path with the least edge weight from all the feasible paths between any start and end vertices in a graph [8]. Some algorithms by past mathematicians that can solve the shortest path problems are Dijkstra's algorithm [9] and the Bellman-Ford algorithm [10]. There are also some DNA algorithms that can be used for the computation of the shortest paths as presented by Narayanan and Zorbalas [11] and Ibrahim et al. [12] where the latter used Direct-Proportional Length-Based DNA computing to solve the shortest path problem.

Therefore, in this research, the shortest path problem is applied to graphs of DNA string where the graphs formed are only considering Purines (Adenine and Guanine) as the initial bases for the vertices. This is because Purines plays a significant role in the synthesis of nucleic acids and proteins [13]. The solution for the shortest paths then produces a minimized DNA string.

The graphs of DNA string used in this research have edges representing the bases between vertices which differ from de Bruijn graphs that have edges representing the overlap between vertices. Meanwhile, the application of shortest path problem in this research takes inspiration from the paper by Blazewicz et al. [14] which used Hamiltonian and Euler paths in DNA graphs to reconstruct DNA sequences.

## 2 Methodology

In this section, the methods used in this research are discussed, particularly on the selection of a DNA string, graph generation of the DNA string, calculation of the shortest paths for the graph generated and the formation of minimized DNA strings.

From the DNA of Bacteriophage lambda cI 857 Sam7 (Lambda) available from the New England Biolabs (NEB) website [15], a string with a length of 80 base pairs is selected for the research. From this DNA string, a linear graph is generated by using base pairs of length two with an initial base of Adenine (A) and Guanine (G) as the vertices respectively, forming the vertex sets of  $U = \{AA, AC, AT, AG\}$  and  $V = \{GA, GC, GT, GG\}$ . Next, the edges and edge weight of the graph is formed by the number of base pairs between the adjacent vertices. In mathematical terms, the edge weight for a start vertex  $u$  and end vertex  $v$  is denoted as  $W(\langle \overrightarrow{u(m)}, v(n) \rangle)$ ,  $m < n$  where  $m, n \in \mathbb{Z}^+$  is the corresponding position of the initial bases acting as the vertex in the DNA string. The formula to calculate the edge weight is given by  $W(\langle \overrightarrow{u(m)}, v(n) \rangle) = n - (m + 2)$ . For example, given the DNA sequence of ACCAGT and the vertex set of  $U = \{AA, AC, AT, AG\}$ , then the two vertices are AC at position one and AG at position four. Since the constraints are  $m < n$ , therefore, the start vertex is AC and the end vertex is AG. Next, applying the formula of edge weight calculation, the following edge

weight is obtained:

$$\begin{aligned} W(\overrightarrow{\langle u(m), v(n) \rangle}) &= W(\overrightarrow{\langle AC(1), AG(4) \rangle}) \\ &= 4 - (1 + 2) \\ &= 1. \end{aligned}$$

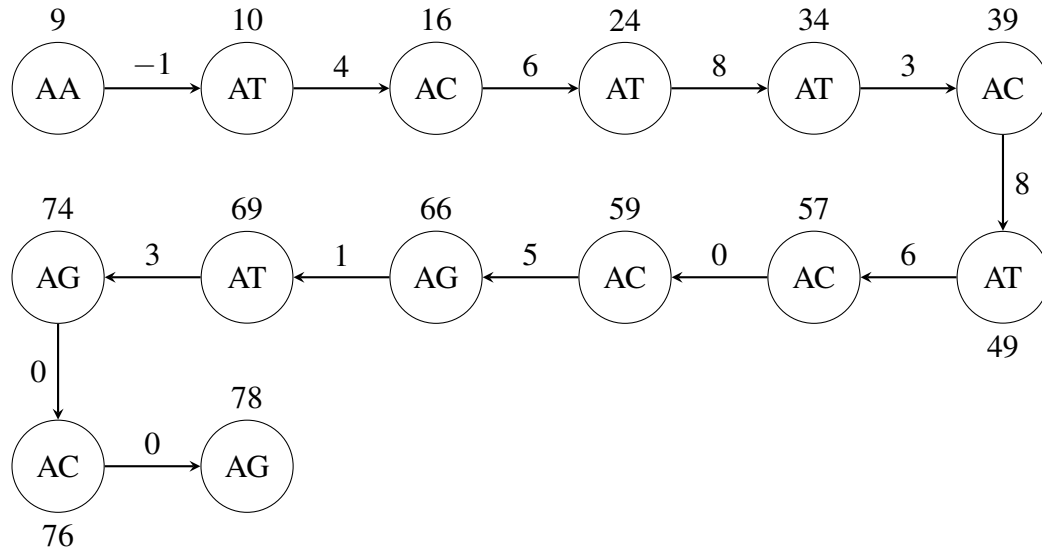
This edge weight of one corresponds to the C (colored yellow) between AC and AG (colored red) in the sequence of ACCAGAGT.

From this linear graph, all paths that can be formed for each vertex is calculated and tabulated. From the table of all paths formed, any start vertex which form a negative length is removed since negative length in a DNA string is not possible. Other than that, since there are many vertices which represent the same base pair of length two, these vertices as well as their edges are combined where only the edge weight(s) with the lowest numerical value is kept for further calculation. This forms the reduced graph for the DNA string.

From the reduced graph, the shortest path for all start and end vertices is calculated. Since not all edges are involved in the shortest path obtained, the paths that are not traversed are thus removed from the reduced graph in order to form the simplified graph for the DNA string. Next, an Euler path which is a route that traverse through all edges exactly once [16], is obtained from the simplified graph. This Euler path is used to minimize the DNA string by reconstructing the vertices and edge weights to their respective DNA strings, followed by arranging the strings obtained according to the traversed path.

### 3 Results and Discussion

In this section, the results for the selected DNA string and the cases for vertices of base pairs of length two with initial bases of Adenine (A) and Guanine (G) are shown respectively. The DNA string selected for this research is taken from the 12841st base to the 12920th base of Bacteriophage lambda cI 857 Sam7 (Lambda) which is 80 bases long. The 80 bases of the DNA string is GCGTGGGGAA TCTTTACCGG CTGATGCGCG GCTATGCCAC CGGCGGTTAT GTCGGTACAC CGGGCAGCAT GGCAGACAGC which is denoted as  $\alpha$  henceforth in this paper. For the case of A, the vertex set is  $U = \{AA, AC, AG, AT\}$  where the linear graph of all elements of the set for  $\alpha$  is shown in Figure 1.

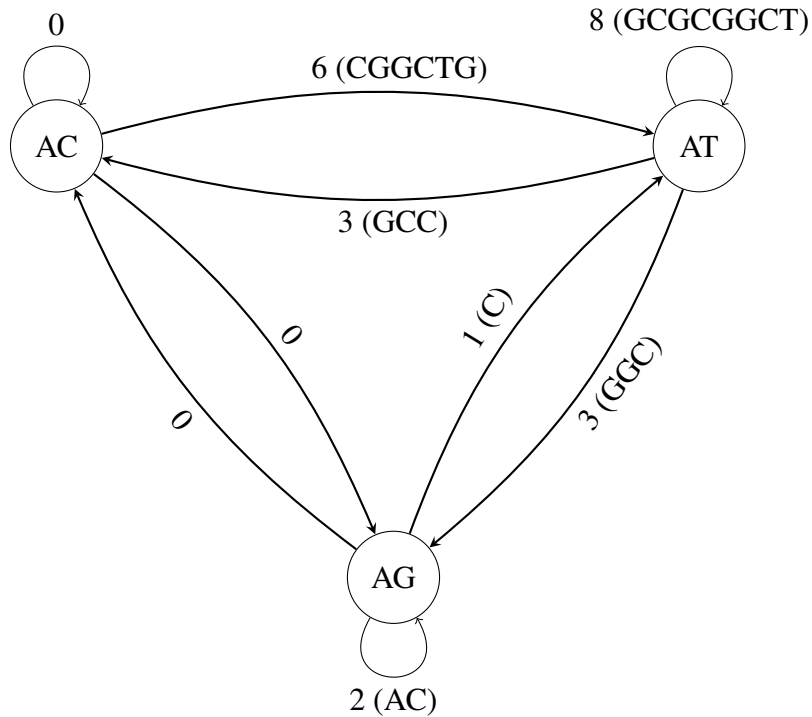
Figure 1: Linear graph for all elements of  $\alpha$  with vertex set  $U$ 

In Figure 1, the number above each vertex represents the corresponding position of the initial base acting as the vertices in the string  $\alpha$ . The edge weight represents the number of bases between each vertex which is obtained from  $\alpha$ .

The formula for calculating edge weight is  $W(\overrightarrow{\langle u(m), v(n) \rangle}) = n - (m + 2)$  where  $u$  and  $v$  denote the start and end vertex respectively and  $m < n$  for  $m, n \in \mathbb{Z}^+$  is the position of the bases which acts as the vertices in the graph. Thus, all possible paths for the graph can be tabulated as shown in Table 1. Note that the edges used in the reduced graph are represented by the cells colored in gray.

From Table 1, a negative edge weight is produced by the sole AA vertex. Therefore this vertex is removed, forming the vertex set of  $U' = \{AC, AG, AT\}$ . As mentioned in the methodology, since there are many vertices which represent the same bases, these vertices along with their edge weights are combined, leaving only the shortest edge weight which is represented by the cells colored in gray in Table 1. This results in the reduced graph as shown in Figure 2, where the brackets indicate the bases used in the formation of the edge weight.

[illegible]

Figure 2: Reduced graph of  $\alpha$  with vertex set  $U'$ 

From Figure 2, the shortest path for all possible start and end vertices are calculated and tabulated as shown in Table 2. From the calculations shown in Table 2, only the path with the lowest path length for all start and end vertices is kept and a minimized DNA string is formed for each of the kept paths. The minimized DNA string is formed by reconstructing the vertices and edge weights into the represented DNA string, followed by arranging the strings according to the paths traversed. If there are multiple paths that produce the lowest path length, then the path with lowest amount of vertices is kept as it produces a shorter DNA string. An example that can be shown is for the start and end vertex of AC where the paths of  $AC \rightarrow AC$  and  $AC \rightarrow AG \rightarrow AC$  has the same lowest path length of zero. Here, the path  $AC \rightarrow AC$  produces the string ACAC of length four which is shorter compared to the string ACAGAC of length six produced by the path of  $AC \rightarrow AG \rightarrow AC$ . The results of Shortest path taken, path length and minimized DNA string for the reduced graph are shown in Table 3 where the red colored bases in the minimized DNA string are obtained from the vertices.

From Table 3, it is noticed that the loop for the vertices AT and AG as well as the edge from the vertex AC to vertex AT in Figure 2 is not involved in any of the shortest path obtained. Thus these edges are removed, hence forming the simplified graph as shown in Figure 3.

From Figure 3, there are less than two vertices that are of odd degree which means that an Euler path exist for the simplified graph. One of the Euler paths that can be obtained is  $AT \rightarrow AC \rightarrow AC \rightarrow AG \rightarrow AT \rightarrow AG \rightarrow AC$ . By reconstructing the path into a DNA string, the string obtained is ATGCCACACAGCATGCCAGAC, which is a minimized DNA string of length 21 base pairs for  $\alpha$ . A minimized DNA string of the same length but with a different order of bases can be obtained by using other Euler paths. For example, another Euler path for Figure 3 is  $AT \rightarrow AG \rightarrow AT \rightarrow AC \rightarrow AC \rightarrow AG \rightarrow AC$  which forms the DNA string of ATGGCAGCATGCCACACAGAC which is also of length 21 base pairs.

Table 2: Shortest path calculation for the reduced graph of  $\alpha$  with vertex set  $U'$ 

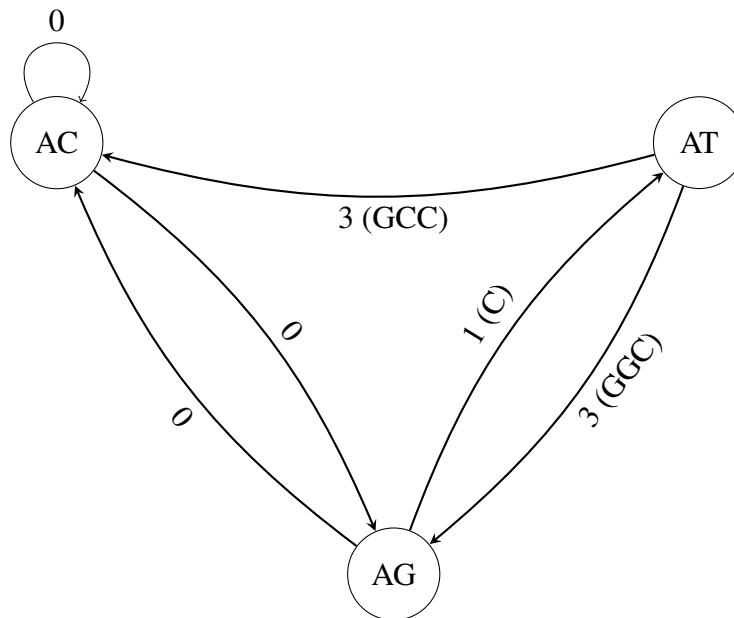
| Start vertex | End vertex | Path taken                                            | Path length   |
|--------------|------------|-------------------------------------------------------|---------------|
| AC           | AC         | AC $\rightarrow$ AC                                   | 0             |
|              |            | AC $\rightarrow$ AG $\rightarrow$ AC                  | 0 + 0 = 0     |
|              |            | AC $\rightarrow$ AG $\rightarrow$ AT $\rightarrow$ AC | 0 + 1 + 3 = 4 |
|              |            | AC $\rightarrow$ AT $\rightarrow$ AC                  | 6 + 3 = 9     |
|              |            | AC $\rightarrow$ AT $\rightarrow$ AG $\rightarrow$ AC | 6 + 3 + 0 = 9 |
|              | AG         | AC $\rightarrow$ AC $\rightarrow$ AG                  | 0 + 0 = 0     |
|              |            | AC $\rightarrow$ AG                                   | 0             |
|              |            | AC $\rightarrow$ AT $\rightarrow$ AG                  | 6 + 3 = 9     |
|              | AT         | AC $\rightarrow$ AC $\rightarrow$ AT                  | 0 + 6 = 6     |
|              |            | AC $\rightarrow$ AG $\rightarrow$ AT                  | 0 + 1 = 1     |
|              |            | AC $\rightarrow$ AT                                   | 6             |
| AG           | AC         | AG $\rightarrow$ AC                                   | 0             |
|              |            | AG $\rightarrow$ AG $\rightarrow$ AC                  | 2 + 0 = 2     |
|              |            | AG $\rightarrow$ AT $\rightarrow$ AC                  | 1 + 3 = 4     |
|              | AG         | AG $\rightarrow$ AC $\rightarrow$ AG                  | 0 + 0 = 0     |
|              |            | AG $\rightarrow$ AC $\rightarrow$ AT $\rightarrow$ AG | 0 + 6 + 3 = 9 |
|              |            | AG $\rightarrow$ AG                                   | 2             |
|              |            | AG $\rightarrow$ AT $\rightarrow$ AG                  | 1 + 3 = 4     |
|              |            | AG $\rightarrow$ AT $\rightarrow$ AC $\rightarrow$ AG | 1 + 3 + 0 = 4 |
|              | AT         | AG $\rightarrow$ AC $\rightarrow$ AT                  | 0 + 6 = 6     |
|              |            | AG $\rightarrow$ AG $\rightarrow$ AT                  | 2 + 1 = 3     |
|              |            | AG $\rightarrow$ AT                                   | 1             |
| AT           | AC         | AT $\rightarrow$ AC                                   | 3             |
|              |            | AT $\rightarrow$ AG $\rightarrow$ AC                  | 3 + 0 = 3     |
|              |            | AT $\rightarrow$ AT $\rightarrow$ AC                  | 8 + 0 = 8     |
|              | AG         | AT $\rightarrow$ AC $\rightarrow$ AG                  | 3 + 0 = 3     |
|              |            | AT $\rightarrow$ AG                                   | 3             |
|              |            | AT $\rightarrow$ AT $\rightarrow$ AG                  | 8 + 3 = 11    |
|              | AT         | AT $\rightarrow$ AC $\rightarrow$ AT                  | 3 + 6 = 9     |
|              |            | AT $\rightarrow$ AC $\rightarrow$ AG $\rightarrow$ AT | 3 + 0 + 1 = 4 |
|              |            | AT $\rightarrow$ AG $\rightarrow$ AT                  | 3 + 1 = 4     |
|              |            | AT $\rightarrow$ AG $\rightarrow$ AC $\rightarrow$ AT | 3 + 0 + 6 = 9 |
|              |            | AT $\rightarrow$ AT                                   | 8             |

Next, the results for the case of vertex set  $V = \{GA, GC, GT, GG\}$  is discussed. The linear graph obtained from the vertex set of  $V$  for the string  $\alpha$  is shown in Figure 4 where the number above each vertex represents the corresponding position of the initial base acting as the vertices in the string  $\alpha$  and the edge weight represents the number of bases between each vertex.

From Figure 4, all possible paths for the graph is calculated using the same formula of  $W(\langle u(m), v(n) \rangle) = n - (m + 2)$  and the results are tabulated in Table 4.

Table 3: Shortest path taken, path length and minimized DNA string for the reduced graph of  $\alpha$  with vertex set  $U'$ 

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

Figure 3: Simplified graph of  $\alpha$  with vertex set  $U'$ 

From Table 4, it is observed that all the negative lengths are produced by the GG vertices. Therefore, the GG vertex is removed from the vertex set of  $V$  to form the vertex set  $V' = \{GA, GC, GT\}$ . The reduced graph formed using the vertex set  $V'$  is shown in Figure 5.

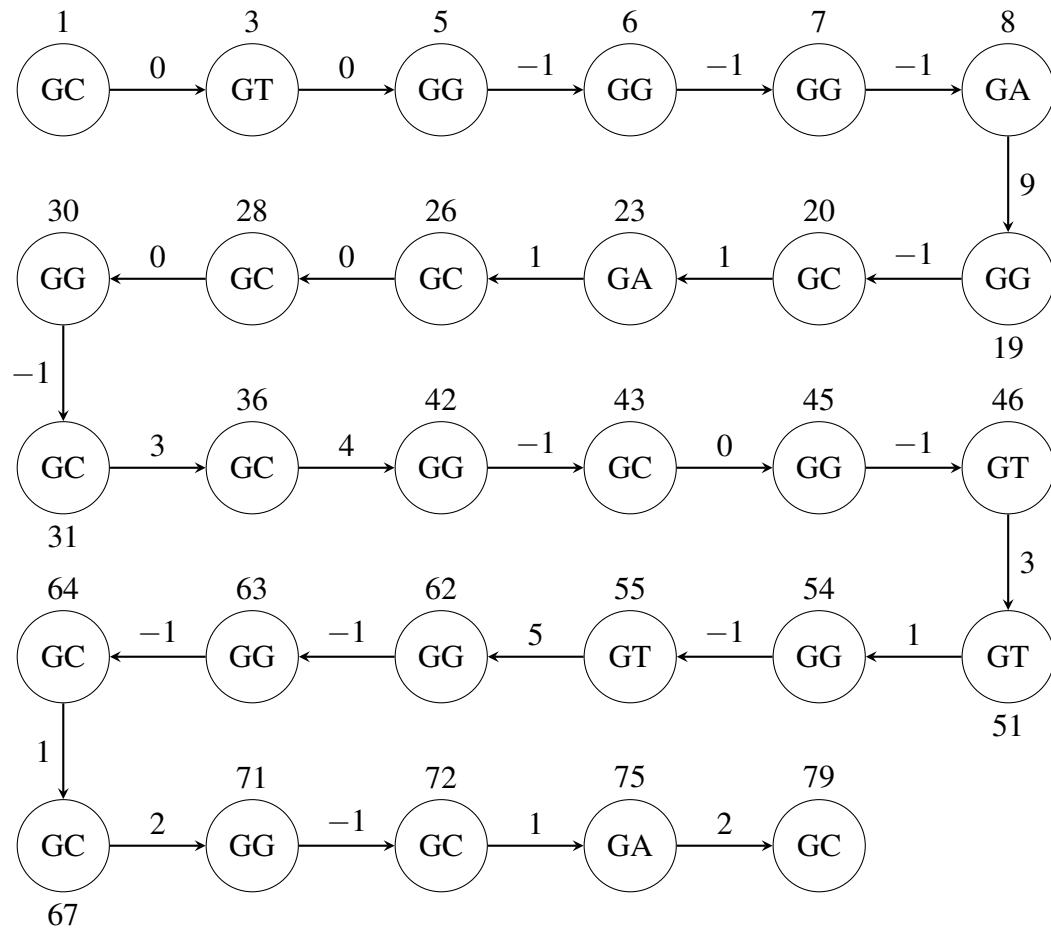
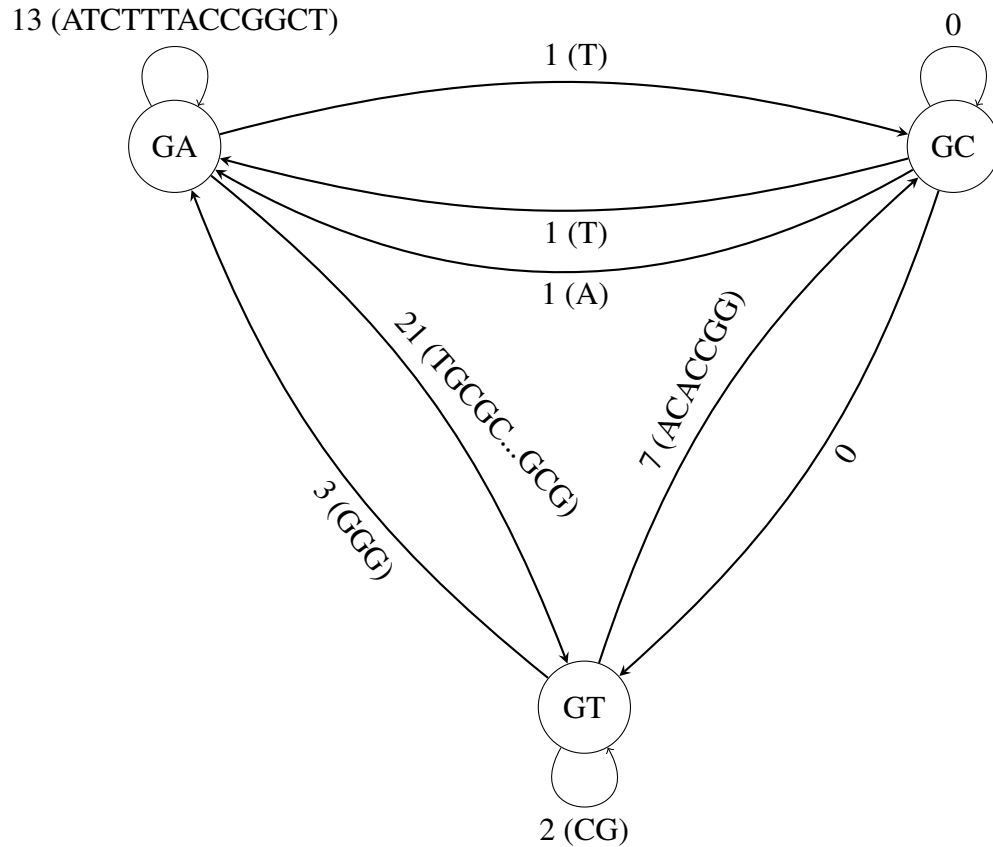


Figure 4: Linear graph for all elements of  $\alpha$  with vertex set  $V$

Table 4: All paths for each possible start and end vertex for the linear graph of  $\alpha$  with vertex set  $V$ 

|              | End vertex | GC | GT | GG | GG | GG | GA | GG | GC | GA | GC | GC | GG | GC | GC | GG | GC | GG | GT | GT | GG | GT | GG | GG | GC | GC | GG | GC | GA | GC |    |
|--------------|------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Start vertex | Position   | 1  | 3  | 5  | 6  | 7  | 8  | 19 | 20 | 23 | 26 | 28 | 30 | 31 | 36 | 42 | 43 | 45 | 46 | 51 | 54 | 55 | 62 | 63 | 64 | 67 | 71 | 72 | 75 | 79 |    |
| GC           | 1          | -  | 0  | 2  | 3  | 4  | 5  | 16 | 17 | 20 | 23 | 25 | 27 | 28 | 33 | 39 | 40 | 42 | 43 | 48 | 51 | 52 | 59 | 60 | 61 | 64 | 68 | 69 | 72 | 76 |    |
| GT           | 3          | -  | -  | 0  | 1  | 2  | 3  | 14 | 15 | 18 | 21 | 23 | 25 | 26 | 31 | 37 | 38 | 40 | 41 | 46 | 49 | 50 | 57 | 58 | 59 | 62 | 66 | 67 | 70 | 74 |    |
| GG           | 5          | -  | -  | -  | -  | -1 | 0  | 1  | 12 | 13 | 16 | 19 | 21 | 23 | 24 | 29 | 35 | 36 | 38 | 39 | 44 | 47 | 48 | 55 | 56 | 57 | 60 | 64 | 65 | 68 | 72 |
| GG           | 6          | -  | -  | -  | -  | -  | -1 | 0  | 11 | 12 | 15 | 18 | 20 | 22 | 23 | 28 | 34 | 35 | 37 | 38 | 43 | 46 | 47 | 54 | 55 | 56 | 59 | 63 | 64 | 67 | 71 |
| GG           | 7          | -  | -  | -  | -  | -  | -  | -1 | 10 | 11 | 14 | 17 | 19 | 21 | 22 | 27 | 33 | 34 | 36 | 37 | 42 | 45 | 46 | 53 | 54 | 55 | 58 | 62 | 63 | 66 | 70 |
| GA           | 8          | -  | -  | -  | -  | -  | -  | -  | 9  | 10 | 13 | 16 | 18 | 20 | 21 | 26 | 32 | 33 | 35 | 36 | 41 | 44 | 45 | 52 | 53 | 54 | 57 | 61 | 62 | 65 | 69 |
| GG           | 19         | -  | -  | -  | -  | -  | -  | -  | -1 | 2  | 5  | 7  | 9  | 10 | 15 | 21 | 22 | 24 | 25 | 30 | 33 | 34 | 41 | 42 | 43 | 46 | 50 | 51 | 54 | 58 |    |
| GC           | 20         | -  | -  | -  | -  | -  | -  | -  | -  | 1  | 4  | 6  | 8  | 9  | 14 | 20 | 21 | 23 | 24 | 29 | 32 | 33 | 40 | 41 | 42 | 45 | 49 | 50 | 53 | 57 |    |
| GA           | 23         | -  | -  | -  | -  | -  | -  | -  | -  | -  | 1  | 3  | 5  | 6  | 11 | 17 | 18 | 20 | 21 | 26 | 29 | 30 | 37 | 38 | 39 | 42 | 46 | 47 | 50 | 54 |    |
| GC           | 26         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 0  | 2  | 3  | 8  | 14 | 15 | 17 | 18 | 23 | 26 | 27 | 34 | 35 | 36 | 39 | 43 | 44 | 47 | 51 |    |
| GC           | 28         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 0  | 1  | 6  | 12 | 13 | 15 | 16 | 21 | 24 | 25 | 32 | 33 | 34 | 37 | 41 | 42 | 45 | 49 |    |
| GG           | 30         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 4  | 10 | 11 | 13 | 14 | 19 | 22 | 23 | 30 | 31 | 32 | 35 | 39 | 40 | 43 | 47 |    |
| GC           | 31         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 3  | 9  | 10 | 12 | 13 | 18 | 21 | 22 | 29 | 30 | 31 | 34 | 38 | 39 | 42 | 46 |    |
| GC           | 36         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 4  | 5  | 7  | 8  | 13 | 16 | 17 | 24 | 25 | 26 | 29 | 33 | 34 | 37 | 41 |    |
| GG           | 42         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 1  | 2  | 7  | 10 | 11 | 18 | 19 | 20 | 23 | 27 | 28 | 31 | 35 |    |
| GC           | 43         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 0  | 1  | 6  | 9  | 10 | 17 | 18 | 19 | 22 | 26 | 27 | 30 | 34 |    |
| GG           | 45         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 4  | 7  | 8  | 15 | 16 | 17 | 20 | 24 | 25 | 28 | 32 |    |
| GT           | 46         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 3  | 6  | 7  | 14 | 15 | 16 | 19 | 23 | 24 | 27 | 31 |    |
| GT           | 51         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 1  | 2  | 9  | 10 | 11 | 14 | 18 | 19 | 22 | 26 |    |    |
| GG           | 54         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 6  | 7  | 8  | 11 | 15 | 16 | 19 | 23 |    |    |
| GT           | 55         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 5  | 6  | 7  | 10 | 14 | 15 | 18 | 22 |    |    |
| GG           | 62         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 0  | 3  | 7  | 8  | 11 | 15 |    |    |
| GG           | 63         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 2  | 6  | 7  | 10 | 14 |    |    |
| GC           | 64         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 1  | 5  | 6  | 9  | 13 |    |    |
| GC           | 67         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 2  | 3  | 6  | 10 |    |    |
| GG           | 71         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -1 | 2  | 6  |    |
| GC           | 72         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 1  | 5  |    |
| GA           | 75         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | 2  |    |
| GC           | 79         | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |    |

Figure 5: Reduced graph of  $\alpha$  with vertex set  $V'$ 

The shortest paths are then calculated for Figure 5 and the results of the shortest paths taken, the path lengths and the minimized DNA strings are shown in Table 5 where the bases colored in red in the minimized DNA strings are obtained from the vertices.

Table 5: Shortest path taken, path length and minimized DNA string for the reduced graph of  $\alpha$  with vertex set  $V'$ 

| Start vertex | End vertex | Shortest path taken                  | Path length | Minimized DNA string |
|--------------|------------|--------------------------------------|-------------|----------------------|
| GA           | GA         | GA $\rightarrow$ GC $\rightarrow$ GA | 1 + 1 = 2   | <b>GATGCTGA</b>      |
|              |            | GA $\rightarrow$ GC $\rightarrow$ GA | 1 + 1 = 2   | <b>GATGCAGA</b>      |
|              | GC         | GA $\rightarrow$ GC                  | 1           | <b>GATGC</b>         |
|              | GT         | GA $\rightarrow$ GC $\rightarrow$ GT | 1 + 0 = 1   | <b>GATGCGT</b>       |
| GC           | GA         | GC $\rightarrow$ GA                  | 1           | <b>GCTGT</b>         |
|              |            | GC $\rightarrow$ GA                  | 1           | <b>GCAGA</b>         |
|              | GC         | GC $\rightarrow$ GC                  | 0           | <b>GCGC</b>          |
|              | GT         | GC $\rightarrow$ GT                  | 0           | <b>GCGT</b>          |
| GT           | GA         | GT $\rightarrow$ GA                  | 3           | <b>GTGGGGA</b>       |
|              | GC         | GT $\rightarrow$ GA $\rightarrow$ GC | 3 + 1 = 4   | <b>GTGGGGATGC</b>    |
|              | GT         | GT $\rightarrow$ GT                  | 2           | <b>GTCCGT</b>        |

It is observed from Table 5 that the edge of GA to GT, the edge of GT to GC and the loop of GA in Figure 5 is not used in any of the shortest paths. These edges are removed to form a simplified graph as shown in Figure 6.

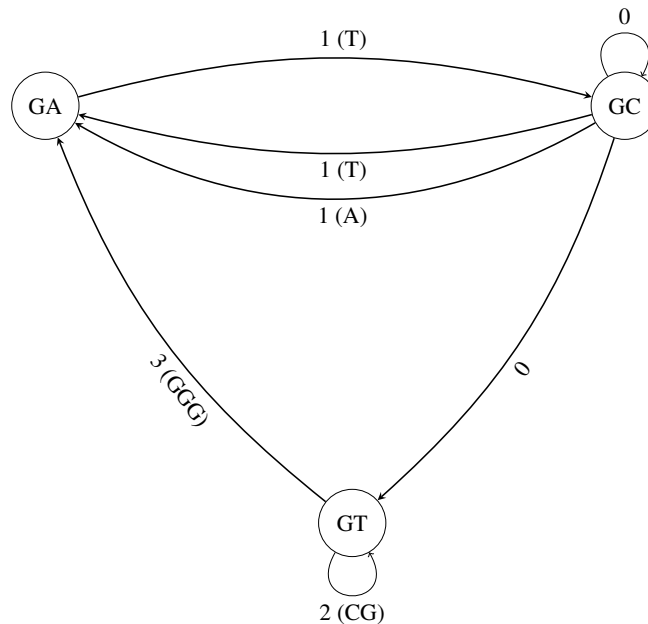


Figure 6: Simplified graph of  $\alpha$  with vertex set  $V'$

From Figure 6, there is no Euler path because there is one vertex of odd degree while all other vertices are of even degree. Therefore, only one path is chosen between the two paths with length of one base pair from vertex GC to vertex GA such that there exists an Euler path. Hence, there will be two minimized string formed for an Euler path. An Euler path that can be formed from Figure 6 above is  $GC \rightarrow GA \rightarrow GC \rightarrow GC \rightarrow GT \rightarrow GT \rightarrow GA$ . By reconstructing the path into a DNA string, two strings with a length of 21 bases are obtained, which are GCTGATGCGCGTCGGTGGGGA and GCAGATGCGCGTCGGTGGGGA. Similar to the case of Adenine (A), other Euler paths used will result in minimized DNA strings of the same length where the bases are ordered differently.

## 4 Conclusion

This research has shown that from a DNA string of length 80 base pairs, graphs of the DNA string can be formed where the vertices have base pairs of length two. The graphs formed are split into two cases, one with initial base of Adenine (A) and the other with initial base of Guanine (G) where A and G are the main bases in Purines. Next, the shortest paths are obtained for the graphs where any edge that is not traversed are removed, hence forming a simplified graph. The Euler paths of the simplified graphs are then reconstructed into a minimized DNA string. The length of the minimized DNA string obtained from both cases of A and G as the initial bases for the vertices is 21 base pairs where the original length of DNA string used is 80 base pairs.

Furthermore, the results of this research help in identifying some properties of a DNA string from a different perspective by using other graphs apart from de Bruijn graphs. This approach can contribute to future advancements regarding the properties of DNA strings.

Some suggestion for future works is to obtain the results of cases for graphs with vertices of initial bases of Cytosine (C) and Thymine (T) as the main bases for Pyrimidines to complete all the possible cases, hence a comparison of results for all the cases can be done.

## Acknowledgments

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, J.Y., Wang, J. and Jan, A. *Biochemistry, DNA structure*. StatPearls Publishing, Treasure Island (FL) 2023. <https://doi.org/10.1038/171737a0>
- [2] Watson, J. D. and Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. 1953. 171(4356): 737-738. <https://doi.org/10.1038/171737a0>
- [3] Bernard, N. M. and Prijith, G. S. Graph theory in other subjects. In *Sustainable Development for Society, Industrial Development, Material, Energy and Environment: Key Issues, Opportunities and Challenges*. Pune: Jyotikiran Publications. 2022. 175-181.
- [4] Ekim, B., Berger, B. and Chikhi, R. Minimizer-space de bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell systems*. 2021. 12(10): 958-968. <https://doi.org/10.1016/j.cels.2021.08.009>
- [5] Palleschi, V., Pagani, L., Pagnotta, S., Amato, G. and Tofanelli, S. Application of graph theory to the elaboration of personal genomic data for genealogical research. *PeerJ Computer Science*. 2015. 1: e27. <https://doi.org/10.7717/peerj-cs.27>
- [6] Ashton, B. Graph theory in DNA sequencing: Unveiling genetic patterns. *International Journal of Biology and Life Sciences*. 2023. 3(1): 9-13. <https://doi.org/10.54097/ijbls.v3i1.9593>
- [7] Chikomana, S. and Hu, X. Genome Sequencing Using Graph Theory Approach. *Open Journal of Discrete Mathematics*. 2023. 13(2): 39-48. <https://doi.org/10.4236/ojdm.2023.132004>
- [8] Ortega-Arranz, H., Gonzalez-Escribano, A. and Llanos, D. R. *The shortest-path problem: Analysis and comparison of methods*. Switzerland: Springer Nature. 2022. <https://doi.org/10.1007/978-3-031-02574-7>
- [9] Dijkstra, E. W. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy (1st edition)*. New York: Association for Computing Machinery. 2022. 287–290. <https://doi.org/10.1145/3544585.3544600>
- [10] Shimbel, A. Structure in communication nets. In: *Proceedings of the symposium on information networks*. Brooklyn: Polytechnic Institute. 1954. 119-203.

- [11] Narayanan, A. and Zorbalas, S. DNA algorithms for computing shortest paths. In: *Proceedings of genetic programming* Burlington, Massachusetts: Morgan Kaufmann Publishers. 1998. 718-723.
- [12] Ibrahim, Z., Tsuboi, Y., Ono, O. and Khalid, M. Direct-proportional length-based DNA computing for shortest path problem. *International Journal of Computer Science and Applications* 2004. 1: 46–60.
- [13] Coppens, I. Exploitation of auxotrophies and metabolic defects in Toxoplasma as therapeutic approaches. *International journal for parasitology*. 2014. 44(2): 109-120. <https://doi.org/10.1016/j.ijpara.2013.09.003>
- [14] Blazewicz, J., Hertz, A., Kobler, D. and de Werra, D. On some properties of DNA graphs. *Discrete Applied Mathematics*. 1999. 44(1-2): 1-19. <https://doi.org/10.1016/j.ijpara.2013.09.003>
- [15] New England Biolabs Website. DNA Sequence and Maps Tool. 2025. Retrieved from <https://www.neb.com/en/tools-and-resources/interactive-tools/dna-sequences-and-maps-tool>
- [16] Marcus, D. A. Graph theory. *American Mathematical Society*. 2020. 53: 1-2. <https://doi.org/10.1090/text/053>