

The De Bruijn graph of non-sequential pattern repetitions in DNA strings

Wan Heng Fong, Ahmed Idrussi, Ahmad Firdaus Yosman

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Article Info

Article history:

Received Feb 1, 2024

Revised Mar 27, 2024

Accepted Apr 6, 2024

Keywords:

De Bruijn

DNA

Graphs

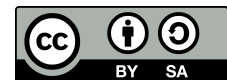
Non-sequential

Repetitions

ABSTRACT

In molecular biology, constructing a genome based on substantially many reads from multitudes of deoxyribonucleic acid (DNA) strings has become an insurmountable task; one which has been continuously addressed by the introduction of various assembly algorithms based on three steps called the overlap-layout-consensus strategy. In the overlap step, the De Bruijn graph is one of many graphs that illustrate the data of all the assembly algorithms. In this article, by using definitions and methods of mathematical induction, some properties of the De Bruijn graph of one time and two times non-sequential repetition of patterns in a DNA string are presented. Examples of these De Bruijn graphs are also given. From there, a generalisation of said properties for m times non-sequential pattern repetition in a DNA string is acquired by means of mathematical induction, as well. The theoretical work in this research is invaluable to develop algorithms that increase the computational efficiency of assembly algorithms.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Wan Heng Fong

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia

81310 UTM Johor Bahru, Johor, Malaysia

Email: fwh@utm.my

1. INTRODUCTION

The discovery of a deoxyribonucleic acid (DNA) structure by Watson and Crick [1] as a three-dimensional double helix was a revolutionary contribution to biology because it was an accurate description of a once complex molecule. However, studies of the DNA structure and its functions have been proven difficult without help from other fields of research because of the discrete nature of DNA. Hence, discrete mathematics (especially the theory of graphs) has a special value for the researchers in the area of molecular biology [2]. The intersection between molecular biology and mathematics has occurred often in history, as it did in 1987 [3] when Head proposed the DNA splicing system that modelled the recombination behaviour of DNA strings in the presence of enzymes by treating the nucleotide bases in DNA as letters in an alphabet.

Through the years, many other DNA recombinant models have since been introduced, such as sticker systems [4], weighted sticker systems [5], probabilistic sticker systems [6], and bonded insertion-deletion systems [7]–[9]. However, the concept of graphs was first combined with DNA recombinant models in 1995, called graph splicing systems [10]. This new model not only birthed numerous variants and results [11]–[14], but it also showed that graphs were viable and useful tools to depict the true molecular structure of DNA. In this paper, it is shown that De Bruijn graphs illustrate the non-sequential pattern repetition found in DNA strings during recombination, where these results will aid in the future development of efficient DNA computing devices.

2. METHOD

In molecular biology, the recognition, assembly, and recombination of a structure of human genome is challenging because of its massive size, where its DNA string is composed of 3×10^9 pairs of bases. One of the methods for reading these DNA strings is sequencing by hybridisation [15], [16], which consists of two stages. The first stage is a biochemical stage in which a set of all possible substrings that forms the DNA chain (considered as reads) is found. The second stage is the computational stage which is based on graph theory [15], where either the vertices or the arcs (directed edges) correspond to the DNA substrings (reads). In these graphs, the DNA strings correspond to either Hamiltonian cycles or Eulerian paths. The definitions of these concepts are presented in the following.

Definition 1 [17] (Hamiltonian cycle): a cycle passing through all the vertices of a graph is called a Hamiltonian cycle. A graph containing a Hamiltonian cycle is called a Hamiltonian graph.

Definition 2 [17] (Euler path): an Euler path is a path that travels through all edges of a connected graph. A graph containing an Euler path is called an Euler graph.

Many assembly algorithms have been used in the second stage of this method [18]–[20], where the overlap-layout-consensus strategy is featured most often. In the overlap step, many modern assembler algorithms use the De Bruijn graph because it has proven time and again to successfully solve the problem of substantially large amounts of information gained from the next-generation sequencers [21]–[26]. The De Bruijn graph, introduced in [27], gave a solution of the superstring problem, which was defined as finding a circular superstring that contains each possible substring of length k (k -mer) which occurs exactly once over a given alphabet.

In De Bruijn's solution, a directed graph where each vertex corresponds to a word of length $(k - 1)$ of the alphabet ($(k - 1)$ -mer) is constructed. Two vertices are linked by an arc (directed edge) if there is some k -mer whose $(k - 1)$ -suffix corresponds to the first vertex and its $(k - 1)$ -prefix corresponds to the other. Therefore, all the edges of the De Bruijn graph represent all possible k -mers, thus an Eulerian path (or a Hamiltonian cycle) in the De Bruijn graph represents the shortest superstring that contains each k -mer exactly once. In fact, there exist n^k k -mers in an alphabet containing n symbols. For example, given the alphabet consisting of the symbols α and β , there exist $2^3 = 8$ 3-mers given by $\alpha\alpha\alpha, \alpha\alpha\beta, \alpha\beta\alpha, \alpha\beta\beta, \beta\alpha\alpha, \beta\alpha\beta, \beta\beta\alpha, \beta\beta\beta$. These 3-mers form the circular superstring $\alpha\alpha\alpha\beta\beta\beta\alpha\beta\alpha$ which is the shortest superstring containing all the 3-mers because it contains each 3-mer exactly once. Thus, the representation of the overlap of words of $(k - 1)$ length in the DNA strings as a directed graph is achievable using De Bruijn's solution. The pattern of a DNA string refers to some group of alphabets in a DNA string that repeat in a specific way, whether sequential or non-sequential. Two types of patterns considered in this paper are namely n copies of the same element a , and n distinct elements. In the following, the definition of a De Bruijn graph is presented.

Definition 3 [28] (De Bruijn graph): a De Bruijn graph is defined as a directed multi graph in which each vertex corresponds to a $(k - 1)$ -mer and two $(k - 1)$ -mers are linked by a directed edge if and only if there exists a k -mer where one of the $(k - 1)$ -mers is its suffix and the other is its prefix.

Mathematically, let Γ be a read of length n . The symbol $\gamma[i, j]$ denotes the substring of Γ from the i -th to the j -th alphabets. The read Γ contains $n - k + 1$ overlapping k -mers which are $\gamma[1, k], \gamma[2, k + 1], \dots, \gamma[n - k, n - 1], \gamma[n - k + 1, n]$. Now, split each k -mer into left and right $(k - 1)$ -mers i.e. $\gamma_1[1, k - 1], \gamma_2[2, k], \gamma_3[3, k + 1], \dots, \gamma_{n-k+1}[n - k + 1, n - 1], \gamma_{n-k+2}[n - k + 2, n]$, these $(k - 1)$ -mers need not be distinct. Again, connect each left $(k - 1)$ -mer to its corresponding right $(k - 1)$ -mer by a directed edge. Hence, the De Bruijn graph of Γ is presented as:

$$\begin{aligned} \gamma_1[1, k - 1] &\xrightarrow{\gamma[1, k]} \gamma_2[2, k] \xrightarrow{\gamma[2, k + 1]} \gamma_3[3, k + 1] \xrightarrow{\gamma[3, k + 2]} \dots \gamma_{n-k}[n - k, n - 2] \xrightarrow{\gamma[n-k, n-1]} \\ &\gamma_{n-k+1}[n - k + 1, n - 1] \xrightarrow{\gamma[n-k+1, n]} \gamma_{n-k+2}[n - k + 2, n] \end{aligned}$$

where the $(k - 1)$ -mers correspond to the vertices of the De Bruijn graph and the k -mers correspond to the directed edges.

The concepts presented in this section, along with the method of mathematical induction, are utilised to obtain the results presented in section 3, where some properties of different De Bruijn graphs and its examples are presented. Following that, the generalisation of these De Bruijn graphs is presented. The results presented in this research serve as a theoretical basis to reducing the required storage allocation of numerous computations by the development of an algorithm that can provide the number of loops and multiple edges in the De Bruijn

graph. Note that the length of the patterns considered in this paper is $n \geq 3$ because in the case $n = 2$, the De Bruijn graph with the pattern of n copies of the same element α , does not contain self-loops and the De Bruijn graph with the pattern of n distinct elements does not contain multiple edges.

3. RESULTS AND DISCUSSION

In this paper, we consider De Bruijn graphs when the patterns for n copies of the same element α and n distinct elements of a DNA string are repeated once (one time) and twice (two times) non-sequentially. Note that a once repetition means that the pattern occurs twice, while a twice repetition means that the pattern occurs three times. Moving on, some properties of these graphs are presented in section 3.1 along with some examples in section 3.2. Lastly, a generalisation of these properties is presented in section 3.3.

3.1. The De Bruijn graph of once and twice repetition of a pattern

Firstly, the De Bruijn graphs of a DNA string containing n copies of the same element α occurring two times and three times non-sequentially, and n distinct elements occurring two times and three times non-sequentially, are found. Here, n copies of the same element α are indicated by $\underbrace{\alpha\alpha\dots\alpha}_n$ where $n \geq 3$; while n distinct elements are indicated by $\alpha_1\alpha_2\dots\alpha_n$ where $n \geq 3$. For both cases of De Bruijn graphs, a one time non-sequential repetition of a pattern is recognised when the sequences $(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_1(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_2$ or $(\alpha_1\alpha_2\dots\alpha_n)\omega_1(\alpha_1\alpha_2\dots\alpha_n)\omega_2$ appear in a DNA string, such that $\omega_i, i = 1, 2$ are arbitrary sequences of letters. Meanwhile, a two time non-sequential pattern repetition occurs when the sequences $(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_1(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_2(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_3$ or $(\alpha_1\alpha_2\dots\alpha_n)\omega_1(\alpha_1\alpha_2\dots\alpha_n)\omega_2(\alpha_1\alpha_2\dots\alpha_n)\omega_3$ appear in a DNA string, such that $\omega_i, i = 1, 2, 3$ are arbitrary sequences of letters.

The De Bruijn graph of one time non-sequential repetition of a pattern $\underbrace{\alpha\alpha\dots\alpha}_n$, where $n \geq 3$ is given in Lemma 1 while the De Bruijn graph of one time non-sequential repetition of a pattern $\alpha_1\alpha_2\dots\alpha_n$, where $n \geq 3$ is presented in Lemma 2.

Lemma 1: Given a DNA string R and a substring $\underbrace{\alpha\alpha\dots\alpha}_n \subset R$ with length n , where $n \geq 3$. For the one time non-sequential repetition of the substring $\underbrace{\alpha\alpha\dots\alpha}_n \subset R$, the string R contains $(2n - 2)$ 3-mers and the De Bruijn graph of R contains $(2n - 4)$ self-loops.

Lemma 2: Given a DNA string R and a substring $\alpha_1\alpha_2\dots\alpha_n \subset R$ with length n , where $n \geq 3$. For the one time non-sequential repetition of the substring $\alpha_1\alpha_2\dots\alpha_n \subset R$, the string R contains $(n - 2)$ 3-mers that occur two times and the De Bruijn graph of R contains $(n - 2)$ pairs of vertices connected by two directed edges.

In the following lemmas, the De Bruijn graph of two times non-sequential repetition of a pattern $\underbrace{\alpha\alpha\dots\alpha}_n$, where $n \geq 3$ and the De Bruijn graph of two times non-sequential repetition of a pattern $\alpha_1\alpha_2\dots\alpha_n$, where $n \geq 3$ are presented.

Lemma 3: Given a DNA string R and a substring $\underbrace{\alpha\alpha\dots\alpha}_n \subset R$ with length n , where $n \geq 3$. For the two times non-sequential repetition of the substring $\underbrace{\alpha\alpha\dots\alpha}_n \subset R$, the string R contains $(3n - 3)$ 3-mers and the De Bruijn graph of R has $(3n - 6)$ self-loops.

Lemma 4: Given a DNA string R and a substring $\alpha_1\alpha_2\dots\alpha_n \subset R$ with length n , where $n \geq 3$. For the two times non-sequential repetition of substring $\alpha_1\alpha_2\dots\alpha_n \subset R$, the string R contains $(n - 2)$ 3-mers that occur three times and the De Bruijn graph of R contains $(n - 2)$ pairs of vertices connected by three directed edges. In the next section, some examples are provided to illustrate the results in Lemma 3 and Lemma 4.

3.2. Examples of the De Bruijn graph of non-sequential pattern repetitions

In this section, the results in Lemma 3 and Lemma 4 are illustrated by Example 1 and Example 2, respectively:

Example 1: Let AGTTTTCTTTTGTTTTA be a DNA string with length 17. Notice that the substring TTTT appears three times non-sequentially, hence it is a two times repetition. Therefore, all the 3-mers of this string

are given by:

AGT, GTT, TTT, TTT, TTC, TCT, CTT, TTT, TTT, TTG, TGT, GTT, TTT, TTT, TTA.

To construct the De Bruijn graph of the DNA string, first split each 3-mer into left and right 2-mers. Then, connect each left 2-mer to its corresponding right 2-mer by a directed edge, given by:

$$\begin{aligned} &AG \xrightarrow{AGT} GT \xrightarrow{GTT} TT \xrightarrow{TTT} TT \xrightarrow{TTT} TT \xrightarrow{TTC} TC \xrightarrow{TCT} CT \xrightarrow{CTT} TT \xrightarrow{TTT} TT \xrightarrow{TTT} TT \xrightarrow{TTG} TG \xrightarrow{TGT} GT \xrightarrow{GTT} \\ &TT \xrightarrow{TTT} TT \xrightarrow{TTT} TT \xrightarrow{TTA} TA, \end{aligned}$$

as shown in Figure 1.

From Figure 1, the number of 3-mers which form the 2-mer TT and form six self-loops on the vertex TT in the De Bruijn graph of the DNA string is equal to nine, which are TTT, TTT, TTC, TTT, TTT, TTG, TTT, TTT, and TTA. This result coincides with Lemma 3, where, for $n = 4$, the number of 3-mers which form the 2-mer TT in the De Bruijn graph is given by $3(4) - 3 = 9$, and the number of self-loops on the vertex TT is given by $3(4) - 6 = 6$.

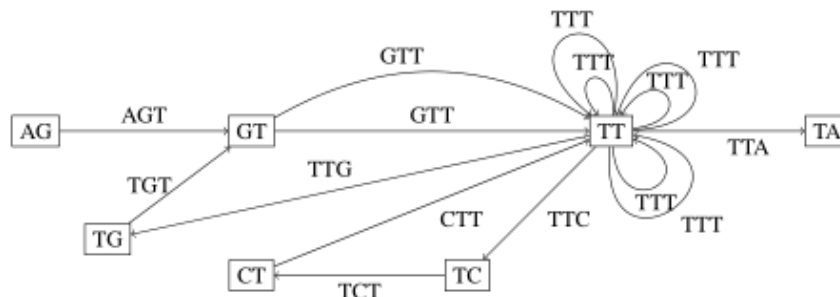


Figure 1. The De Bruijn graph of AGTTTTCTTTTGTTTAA with substring TTTT repeated two times non-sequentially

Example 2: Let AGTCTGCAGTCTACAGTCTTT be a DNA string with length 21. The substring AGTCT appears three times non-sequentially, thus making it a two times repetition. Therefore by Definition 3, all the 3-mers of this string are given by:

AGT, GTC, TCT, CTG, TGC, GCA, CAG, AGT, GTC, TCT, CTA, TAC, ACA, CAG, AGT, GTC, TCT, CTT, TTT.

Once again, the De Bruijn graph is constructed by the same procedure as in Example 1, in this case given by:

$$\begin{aligned} &AG \xrightarrow{AGT} GT \xrightarrow{GTC} TC \xrightarrow{TCT} CT \xrightarrow{CTG} TG \xrightarrow{TGC} GC \xrightarrow{GCA} CA \xrightarrow{CAG} AG \xrightarrow{AGT} GT \xrightarrow{GTC} TC \xrightarrow{TCT} CT \xrightarrow{CTA} TA \xrightarrow{TAC} \\ &AC \xrightarrow{ACA} CA \xrightarrow{CAG} AG \xrightarrow{AGT} GT \xrightarrow{GTC} TC \xrightarrow{TCT} CT \xrightarrow{CTT} TT \xrightarrow{TTT} TT, \end{aligned}$$

as shown in Figure 2.

Observe from Figure 2 that the string R contains three 3-mers which appear three times, which are AGT, GTC, and TCT. These 3-mers form three pairs of vertices, namely (AG, GT), (GT, TC), and (TC, CT), which are each connected by three directed edges in the graph. This result coincides with Lemma 4, whereby $n = 5$, then the string R contains $5 - 2 = 3$ 3-mers that occur three times. Moreover, for $n = 5$, the De Bruijn graph of R contains $5 - 2 = 3$ pairs of vertices connected by three directed edges.

In the next section, the generalisations of the results in Lemma 1, Lemma 2, Lemma 3, and Lemma 4 for m times non-sequential pattern repetition are obtained.

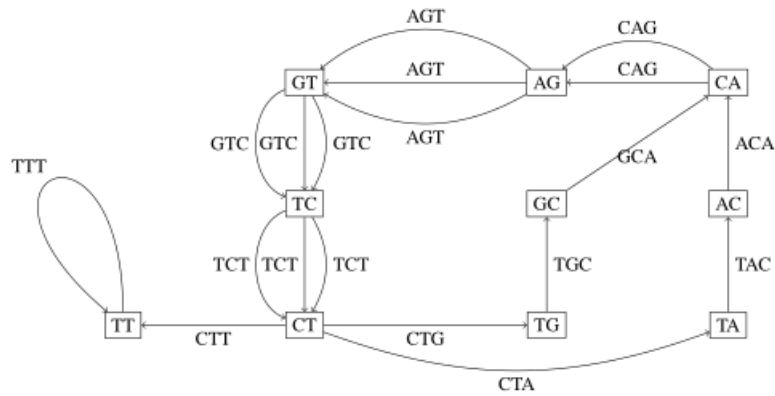


Figure 2. The De Bruijn graph of AGTCTGCAGTCTACAGTCTTT with substring AGTCT repeated two times non-sequentially

3.3. The De Bruijn graph of m times non-sequential pattern repetitions

Now, the results in section 3.1 are generalised for m times non-sequential repetition of substrings with n copies of the same element α and n distinct elements. Here, the structure of the substrings contained in the DNA string are presented as:

$$(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_1(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_2\dots\omega_{m-1}(\underbrace{\alpha\alpha\dots\alpha}_n)\omega_m(\underbrace{\alpha\alpha\dots\alpha}_n)$$

and,

$$(\alpha_1\alpha_2\dots\alpha_n)\omega_1(\alpha_1\alpha_2\dots\alpha_n)\omega_2\dots\omega_{m-1}(\alpha_1\alpha_2\dots\alpha_n)\omega_m(\alpha_1\alpha_2\dots\alpha_n)$$

respectively, where ω_i with $i = 1, 2, \dots, m$, are arbitrary sequences of letters.

The properties of the De Bruijn graph of m times non-sequential repetition of a pattern $\underbrace{\alpha\alpha\dots\alpha}_n$, where $n \geq 3$ is presented in Theorem 1.

Theorem 1: Given a DNA string R and a substring $\underbrace{\alpha\alpha\dots\alpha}_n \subset R$ with length n , where $n \geq 3$. For the m times non-sequential repetition of substring $\underbrace{\alpha\alpha\dots\alpha}_n \subset R$, where $m \geq 2$, the number of 3-mers in the string R is equal to $(m + 1)(n - 1)$ and the De Bruijn graph of R has $(m + 1)(n - 2)$ self-loops.

Proof: Suppose the pattern of m times non-sequential repetition of the substring $\underbrace{\alpha\alpha\dots\alpha}_n$, is written as:

$$\underbrace{\alpha\alpha\dots\alpha}_n\omega_1\underbrace{\alpha\alpha\dots\alpha}_n\omega_2\dots\omega_{m-1}\underbrace{\alpha\alpha\dots\alpha}_n\omega_m\underbrace{\alpha\alpha\dots\alpha}_n\omega_{m+1}$$

where ω_i for $i = 1, 2, \dots, (m + 1)$ are arbitrary sequences of letters.

The proof is done by mathematical induction. First, let $m = 2$. Then by Lemma 3, the result is true. Next, assume that the result is true for $m = k$, which means that for:

$$\underbrace{\alpha\alpha\dots\alpha}_n\omega_1\underbrace{\alpha\alpha\dots\alpha}_n\omega_2\dots\omega_{k-1}\underbrace{\alpha\alpha\dots\alpha}_n\omega_k\underbrace{\alpha\alpha\dots\alpha}_n\omega_{k+1}$$

there are $(k + 1)(n - 1)$ 3-mers in the string R , which give 2-mer of the form $\alpha\alpha$ and form $(k + 1)(n - 2)$ self-loops on the vertex $\alpha\alpha$ in the De Bruijn graph.

Now, let $m = k + 1$. Then the pattern is written as:

$$\underbrace{\alpha\alpha\dots\alpha}_n\omega_1\dots\omega_k\underbrace{\alpha\alpha\dots\alpha}_n\omega_{k+1} = \underbrace{\alpha\alpha\dots\alpha}_n\omega_1\dots\omega_{k-1}\underbrace{\alpha\alpha\dots\alpha}_n\omega_k\underbrace{\alpha\alpha\dots\alpha}_n\omega_{k+1}\underbrace{\alpha\alpha\dots\alpha}_n\omega_{k+2}$$

where ω_i for $i = 1, 2, \dots, (k+2)$ are arbitrary sequences of letters. It can be seen that this pattern increases the substring by $\underbrace{\alpha\alpha \dots \alpha}_n \omega_{k+2}$ as compared to the pattern in the case $m = k$. Therefore, the number of 3-mers increases by $n - 1$, i.e. the number of 3-mers in this case which give 2-mer of the form $\alpha\alpha$ is given by:

$$(k+1)(n-1) + n - 1 = (k+2)(n-1)$$

Also, the number of self-loops increases by $n - 2$, shown as follows:

$$(k+1)(n-2) + n - 2 = (k+2)(n-2)$$

Thus, the result is true for $m = k + 1$. Hence, the proof is complete.

In the sequel, the properties of the De Bruijn graph of m times non-sequential repetition of a pattern $\alpha_1\alpha_2 \dots \alpha_n$, where $n \geq 3$ are presented in Theorem 2.

Theorem 2: Given a DNA string R and a substring $\alpha_1\alpha_2 \dots \alpha_n \subset R$ with length n , where $n \geq 3$. For the m times non-sequential repetition of substring $\alpha_1\alpha_2 \dots \alpha_n \subset R$, where $m \geq 2$, the string R contains $(n-2)$ 3-mers that occur $m+1$ times and the De Bruijn graph of R contains $(n-2)$ pairs of vertices connected by $m+1$ directed edges.

Proof: Suppose the pattern of m times non-sequential repetition of the substring $\alpha_1\alpha_2 \dots \alpha_n$ is given by:

$$(\alpha_1\alpha_2 \dots \alpha_n)\omega_1(\alpha_1\alpha_2 \dots \alpha_n)\omega_2 \dots \omega_{m-1}(\alpha_1\alpha_2 \dots \alpha_n)\omega_m(\alpha_1\alpha_2 \dots \alpha_n)\omega_{m+1}$$

where ω_i for $i = 1, 2, \dots, (m+1)$ are arbitrary sequences of letters.

The proof is done by mathematical induction. First, let $m = 2$. Then, by Lemma 4, the result is true. Next, assume that the result is true for $m = k$, which means that for:

$$(\alpha_1\alpha_2 \dots \alpha_n)\omega_1(\alpha_1\alpha_2 \dots \alpha_n)\omega_2 \dots \omega_{k-1}(\alpha_1\alpha_2 \dots \alpha_n)\omega_k(\alpha_1\alpha_2 \dots \alpha_n)\omega_{k+1}$$

there are $(n-2)$ 3-mers in the string R occurring $k+1$ times, which are $\alpha_{i-2}\alpha_{i-1}\alpha_i$, $i = 3, 4, \dots, n$. These 3-mers form $(n-2)$ pairs of vertices of the form $(\alpha_{i-2}\alpha_{i-1}, \alpha_{i-1}\alpha_i)$, $i = 3, 4, \dots, n$, connected by $k+1$ directed edges in the De Bruijn graph.

Lastly, let $m = k + 1$. Then the pattern is written as:

$$\begin{aligned} & (\alpha_1\alpha_2 \dots \alpha_n)\omega_1 \dots \omega_k(\alpha_1\alpha_2 \dots \alpha_n)\omega_{k+1} \\ & = (\alpha_1\alpha_2 \dots \alpha_n)\omega_1 \dots \omega_{k-1}(\alpha_1\alpha_2 \dots \alpha_n)\omega_k(\alpha_1\alpha_2 \dots \alpha_n)\omega_{k+1}(\alpha_1\alpha_2 \dots \alpha_n)\omega_{k+2} \end{aligned}$$

where ω_i for $i = 1, 2, \dots, (k+2)$ are arbitrary sequences of letters. It can be seen that this pattern increases the substring by $(\alpha_1\alpha_2 \dots \alpha_n)\omega_{k+2}$ as compared to the case of $m = k$. Therefore, the 3-mers of the form $\alpha_{i-2}\alpha_{i-1}\alpha_i$, $i = 3, 4, \dots, n$ are repeated one more time which implies that the pairs of vertices of the form $(\alpha_{i-2}\alpha_{i-1}, \alpha_{i-1}\alpha_i)$, $i = 3, 4, \dots, n$, are connected by one more directed edge in the De Bruijn graph. Thus, in the case $m = k + 1$, there are $(n-2)$ 3-mers in the string R occurring $(k+1) + 1 = k+2$ times, and these 3-mers form $(n-2)$ pairs of vertices connected by $(k+1) + 1 = k+2$ directed edges in the De Bruijn graph. Hence, the result is true for $m = k + 1$ and the proof is complete.

4. CONCLUSION

In this paper, based on the definitions of a Hamiltonian cycle, Euler path, and De Bruijn graph as well as using mathematical induction, the De Bruijn graphs of once and twice non-sequential repetitions of n copies of the same element α and n distinct elements where $n \geq 3$ in DNA strings have been presented. It has been shown that these De Bruijn graphs with specific non-sequential repetitions of substrings produce strings with distinctive properties, which have been illustrated by examples provided. These results have been generalised for m times non-sequential repetition as follows: if the DNA string contains a pattern $\underbrace{(\alpha\alpha \dots \alpha)}_n$ repeated m times non-sequentially where $n \geq 3$ and $m \geq 2$, then the string has $(m+1)(n-1)$ 3-mers and its De Bruijn graph contains $(m+1)(n-2)$ directed self loops; while in the case of a pattern of the form $(\alpha_1\alpha_2 \dots \alpha_n)$ repeated m times non-sequentially where $n \geq 3$ and $m \geq 2$, then the string contains $(n-2)$

3-mers occurring $m + 1$ times and the De Bruijn graph contains $(n - 2)$ pairs of vertices linked by $m + 1$ directed edges. The results presented in this article have provided meaningful and helpful advancement in developing new algorithms to reduce the storage load and increase efficiency in the computation of assembly algorithms for DNA strings. In the future, researchers may also develop a graphical user interface (GUI) to illustrate the properties of various De Bruijn graphs which correspond to different pattern repetitions.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support from Universiti Teknologi Malaysia (UTM) for the funding under UTM Fundamental Research (UTMFR) Vote no. (Q.J130000.3854.22H45).





REFERENCES

- [1] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953, doi: 10.1038/171737a0.
- [2] J. Blazewicz, A. Hertz, D. Kobler, and D. De Werra, "On some properties of DNA graphs," *Discrete Applied Mathematics*, vol. 98, no. 1–2, pp. 1–19, 1999, doi: 10.1016/S0166-218X(99)00109-2.
- [3] T. Head, "Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors," *Bulletin of Mathematical Biology*, vol. 49, no. 6, pp. 737–759, 1987, doi: 10.1007/BF02481771.
- [4] L. Kari, G. Păun, G. Rozenberg, A. Salomaa, and S. Yu, "DNA computing, sticker systems, and universality," *Acta Informatica*, vol. 35, no. 5, pp. 401–420, 1998, doi: 10.1007/s002360050125.
- [5] W. H. Fong, Y. S. Gan, N. H. Sarmin, and S. Turaev, "Variants of weighted sticker systems with different weighting spaces," *ScienceAsia*, vol. 43S, no. 1, pp. 43–51, 2017, doi: 10.2306/scienceasia1513-1874.2017.43s.043.
- [6] M. Selvarajoo, F. W. Heng, N. H. Sarmin, and S. Turaev, "Probabilistic sticker systems," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 9, no. 3, 2014, doi: 10.11113/mjfas.v9n3.101.
- [7] M. Holzer, B. Truthe, and A. F. Yosman, "On bonded sequential and parallel insertion systems," *RAIRO - Theoretical Informatics and Applications*, vol. 52, no. 2–4, pp. 127–151, 2018, doi: 10.1051/ita/2018010.
- [8] A. F. Yosman, W. H. Fong, and H. I. M. Hassim, "On bonded sequential and parallel deletion systems," *Journal of Critical Reviews*, vol. 7, no. 16, pp. 902–909, 2020.
- [9] W. H. Fong, A. F. Yosman, and H. I. M. Hassim, "Closure properties of bonded sequential insertion-deletion systems," *Journal of Physics: Conference Series*, vol. 1988, no. 1, p. 012075, 2021, doi: 10.1088/1742-6596/1988/1/012075.
- [10] R. Freund, "Splicing systems on graphs," in *Proceedings First International Symposium on Intelligence in Neural and Biological Systems*. INBS'95, 1995, pp. 189–194, doi: 10.1109/inbs.1995.404262.
- [11] I. Aisah, P. R. E. Jayanti, and A. K. Supriatna, "2-cut splicing and 4-cut splicing on DNA molecule," *IOP Conference Series: Materials Science and Engineering*, vol. 567, no. 1, p. 012018, 2019, doi: 10.1088/1757-899X/567/1/012018.
- [12] Z. Ouyang, Y. Huang, and F. Dong, "The maximal 1-planarity and crossing numbers of graphs," *Graphs and Combinatorics*, vol. 37, no. 4, pp. 1333–1344, 2021, doi: 10.1007/s00373-021-02320-x.
- [13] M. N. S. A. Razak, W. H. Fong, and N. H. Sarmin, "Graph splicing rules with cycle graph and its complement on complete graphs," *Journal of Physics: Conference Series*, vol. 1988, no. 1, p. 012067, 2021, doi: 10.1088/1742-6596/1988/1/012067.
- [14] W. H. Fong, M. N. S. A. Razak, and N. H. Sarmin, "Planarity on spliced graphs by one splicing rule in graph splicing systems," *AIP Conference Proceedings*, vol. 2905, no. 1, p. 070006, 2024, doi: 10.1063/5.0171633.
- [15] P. A. Pevzner, "1-tuple DNA sequencing: computer analysis," *Journal of Biomolecular Structure and Dynamics*, vol. 7, no. 1, pp. 63–73, Aug. 1989, doi: 10.1080/07391102.1989.10507752.
- [16] P. A. Pevzner and R. J. Lipshutz, "Towards DNA sequencing chips," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 841 LNCS, pp. 143–158, 1994, doi: 10.1007/3-540-58338-6.64.
- [17] J. A. Bondy and U. S. R. Murty, *Graph theory with applications*. Macmillan London, 1976.
- [18] M. Kasahara and S. Morishita, *Large-scale genome sequence processing*. Imperial College Press, 2006.
- [19] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, no. 6, pp. 315–327, 2010, doi: 10.1016/j.ygeno.2010.03.001.
- [20] M. Pop, "Genome assembly reborn: recent computational challenges," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 354–366, 2009, doi: 10.1093/bib/bbp026.
- [21] R. Li *et al.*, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Research*, vol. 20, no. 2, pp. 265–272, 2010, doi: 10.1101/gr.097261.109.
- [22] I. MacCallum *et al.*, "ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads," *Genome Biology*, vol. 10, no. 10, p. R103, 2009, doi: 10.1186/gb-2009-10-10-r103.
- [23] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 17, pp. 9748–9753, 2001, doi: 10.1073/pnas.171285098.
- [24] M. Sahli and T. Shibuya, "Arapan-S: a fast and highly accurate whole-genome assembly software for viruses and small genomes," *BMC Research Notes*, vol. 5, p. 243, 2012, doi: 10.1186/1756-0500-5-243.
- [25] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABYSS: a parallel assembler for short read sequence data," *Genome Research*, vol. 19, no. 6, pp. 1117–1123, 2009.
- [26] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821–829, 2008, doi: 10.1101/gr.074492.107.





- [27] N. G. De Bruijn, "A combinatorial problem," in *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 1946, vol. 49, no. 7, pp. 758–764.
- [28] Y. Ben-Ari, D. Flomin, L. Pu, Y. Orenstein, and R. Shamir, "Improving the efficiency of de Bruijn graph construction using compact universal hitting sets," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2021*, 2021, pp. 1–9, doi: 10.1145/3459930.3469520.

BIOGRAPHIES OF AUTHORS







Wan Heng Fong     is Associate Professor at Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia (UTM), where she obtained her B.Sc (Industrial Mathematics) with Honours, M.Sc (Mathematics), and Ph.D. (Mathematics). Dr. Fong's illustrious career in mathematics and research has been studied by many awards, achieving 12 exhibition awards and 8 intellectual properties for her works in graphical user interface (GUI) development for visualisation of DNA splicing. She has amassed more than 120 publications, spanning across ISI and SCOPUS-indexed journals in DNA computing, formal language theory, group theory, and graph theory, which constitute her main focus areas of research. To date, she has supervised 18 Ph.D. and Master students and is currently the research group leader of Applied Algebra and Analysis Group (AAAG) at UTM. She can be contacted at email: fwh@utm.my.



Ahmed Ildrussi     received his Bachelor of Mathematics Sciences from Benghazi University, Libya and his Master of Science in Mathematics from Universiti Teknologi Malaysia (UTM), Malaysia with a dissertation under the supervision of Associate Professor Dr. Wan Heng Fong. His work in DNA computing further established the connection between DNA assembly patterns and mathematical modelling using graph theory, where De Bruijn graphs illustrate the pattern repetitions in DNA strings. He is currently serving as a mathematics teacher in his home country of Libya but is keen on continuing his work in the field of DNA computing. He can be contacted at email: hussin.mohamed.ahmed@graduate.utm.my.



Ahmad Firdaus Yosman     obtained his Bachelor of Science (Mathematics) with Honours, Master of Philosophy (Mathematics), and Doctor of Philosophy (Mathematics) from Universiti Teknologi Malaysia (UTM), each with full funding from the Ministry of Higher Education (MOHE) Malaysia. Through his work as a research assistant, he has successfully secured multiple grants, including in-campus and national research grants. His work in formal language theory and group theory revolve around the concept of bonded insertion-deletion systems, a new variant of insertion-deletion systems that possess the generative power equivalent to Turing machines. Currently, he is mainly working in the areas of DNA computing and formal language theory. He can be contacted at email: firdausyosman@yahoo.com.