SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Characterize Type of Splicing Languages via Directed Splicing Graph

Nooradelena Mohd Ruslim[1], Yuhani Yusof[1,*], Mohd Sham Mohamad[1], Mohd Firdaus Abdul-Wahab[2], Faisal[3]

[1] Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, 26300 Kuantan, Pahang, Malaysia
[2] Department of Biosciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[3] Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

ARTICLE INFO

Article history:
Received 4 September 2023
Received in revised form 5 March 2024
Accepted 11 April 2024
Available online 12 May 2024

Keywords:

Y-G splicing system; Splicing language; Splicing graph

ABSTRACT

A splicing system is a formal characterization of the ability to generate certain enzymatic activities acting on deoxyribonucleic acid (DNA) molecules. In this paper, the results from Laun's experiment are used in characterizing the type of splicing languages. In the experiments, two initial strings are involved with different features on the selected restriction enzymes. Case I and Case II discussed in this paper show that the splicing languages obtained from these experiments are in adult and limit languages. Nevertheless, the result obtained in this paper is more precise in showing the type of splicing languages which is beyond adult and limit languages when presented via a directed splicing graph. The features of the restriction enzyme that affect the formation of active persistent language are investigated based on the results proposed by Yusof.

1. Introduction

More than a century ago, a German biologist named Frederich Miescher began a thorough investigation on deoxyribonucleic acid (DNA) and revealed that DNA is the smallest element in living cells, as discussed by Dahm [1]. In the form of a deoxyribonucleic acid chemical compound, which was then recognized to be distinct from protein chemical compounds, Frederich Miescher's research at the time discovered the chemical structure of DNA. A new genetic sequence is then created when two DNA molecules from distinct sources unite in a process known as DNA recombination [2]. Genetic variety, evolution, and the repair of damaged DNA all depend heavily on this mechanism. Gartner *et al.,* [3] stated that, DNA repair and DNA recombination are crucial in combating DNA damage that results from both internal and external sources.

The genetic makeup of an organism is made up of DNA which consists of nucleotides [4]. Each molecule of the DNA subunit is made up of three parts; nitrogenous bases, deoxyribose sugar, and a phosphate group. The nitrogenous bases are classified into purines (adenine and guanine) and pyrimidines (cytosine and thymine) [5]. According to Crick and Watson [6], by following the Watson-

* *Corresponding author.*
E-mail address: yuhani@umpsa.edu.my

https://doi.org/10.37934/araset.45.1.129136

Crick complementarity, adenine (*A*) bonds to thymine (*T*), while guanine (*G*) bonds to cytosine (*C*). The coupling can simply be inscribed as [*A/T*] or [*T/A*] and [*C/G*] or [*G/C*], as clearly deliberated in Russell [7]. Furthermore, in Tamarin [8], the double-stranded DNA (dsDNA) that is linked by hydrogen bonds between the strands can be cleaved if the sequence of specific restriction enzymes recognition site matches the sequence at a certain restriction site. For example, the restriction site of *Eco*RI, $(g; aatt, c)$ be a substring in an initial string and can be represented in the splicing graph as follows:

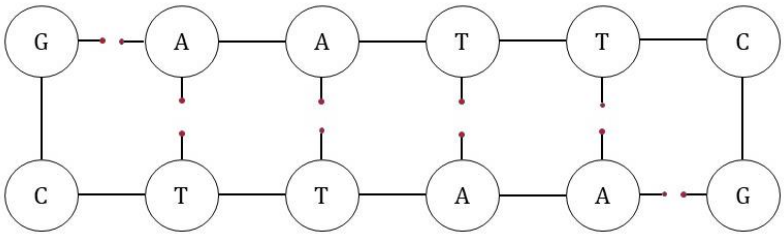

Fig. 1. Splicing graph representation of *Eco*RI

According to Head [9], in the context of DNA recombinant process, initial strings that are spliced by certain rules in a splicing system will produce a set of words, assuming a well-formed dsDNA to be considered as splicing languages. Some authors [9-11] agreed that, the resulting languages can be in the context of transient, adult, or limit languages. The formation of each type of splicing language depends on the number of DNA strings involved in a system and the sequence of restriction enzymes that existed in the string. For instance, the reaction of a non-palindromic rule on a single string will produce adult and limit languages as discussed by Sarmin and Fong [11] and Fong [12].

Previously, multiple researches were conducted by adopting various types of graphs either to present the splicing system or the splicing language. This can be seen from an exhaustive review of DNA splicing system from graph perspective conducted by Mohd Ruslim *et al.,* [13], for instance, limit graph [14,15], de Bruijn graph [16] and cycle graph [17]. In this paper, the formation of splicing languages from the wet model is represented in directed splicing graph. The characterization of splicing languages by using a directed splicing graph is an improvised method in showing more precise splicing languages which had taken from Laun and Reddy [18] who had first verified the result of the dry model presented in Head [9].

2. Preliminaries

Several definitions associated with this research are provided in this section. In this paper, the splicing system inspired from experiments which were conducted by Laun and Reddy [18], is rewritten in the Y-G splicing system, which is the most effective approach for transparently examining the biological aspect of the DNA splicing system. The Y-G splicing system is as described below.

2.1 Definition 1 [19]: Yusof-Goode Splicing System

Yusof-Goode (Y-G) splicing system, $S = (A, I, R)$ is a system that consists of a set of alphabets A, a set of initial strings $I \in A^*$ and a set of rules R. The rules R which refers to the specific restriction enzyme can be either in left cutting, $r = (a; z, b: c; z, d)$, right cutting, $r = (a, z; b: c, z; d)$ or both cuttings $r = (a, z, b: c, z, d)$. For $s_1 = \alpha azb\beta$ and $s_2 = \gamma czd\delta$, where $s_1, s_2 \in I$, then $L(S) = I \cup \{\alpha azd\delta, \gamma czb\beta\}$, $\forall \alpha, \beta, \gamma, \delta, a, b, c, d, z \in A^*$. L is a set of splicing language generated if there is a splicing system S for which $L = L(S)$.

The recombination process of DNA splicing system will result in various types of splicing languages, as given in Definition 2 to Definition 5.

2.2 Definition 2 [10]: Limit Language

A set of words that appears after sufficient time has passed and reach equilibrium state, regardless of the balance of the reactants in a particular experimental run of the reaction is known as a limit language.

Limit language is also known as inert language. It is then renamed to inert persistent language [19].

2.3 Definition 3 [11]: Adult Language

A set of words that cannot be used for further splicing is called an adult language. There is a steady increase in quantity of adult language throughout the reaction, and are not involved in further interactions with other molecules or enzymes.

2.4 Definition 4 [14]: Transient Language

A set of words that will eventually be used up and disappear is called a transient language.

2.5 Definition 5 [14]: Active Persistent Language

A set of words that will participate in further splicing and contained in the limit language is known as an active persistent language.

Next, some important definitions related to graphs are given as follows:

2.6 Definition 6 [20] : Directed Graph

A directed graph consists of a set of vertices, V and a set of edges, E. The vertex p is called the initial vertex of the edge (p, q) while the vertex q is called the terminal vertex of the edge.

In the context of splicing graph of L, a directed graph G_L is when the vertices are the words of L, so that there is an edge from w to z if and only if $w \rightarrow_L z$ [10].

2.7 Definition 7 [20] : Strongly Connected Component (SCC)

The subgraph C of a directed graph G that are strongly connected but not contained in larger strongly connected subgraphs, that is, the maximal strongly connected subgraphs, are called the strongly connected components of G.

In maximal subgraph C, for any vertices w, z of C, $w \rightarrow_{\overset{*}{L}} z$ where $\rightarrow_{\overset{*}{L}}$ is a reflexive transitive closure. If there is no edge $w \rightarrow_L z$ with $w \in C$ and $z \notin C$then it is a terminal component, which is parallel to the definition of SCC given by Rosen [20]. In the context of directed limit graph, the existence of terminal strongly connected component denoted as limit and transient languages were explored in Yusof [14]. It is observed that, 'in' and 'out' edges between vertices represent the existence of an active persistent language in a limit graph.

Apart from the given definition, the directed limit graph which was presented in Yusof [14] is included in Example 1 below.

2.7.1 Example 1

Let $S = (A, I, R)$ be a Y-G splicing system consisting of $I_1, I_2 \in I$ where $I_1 = \alpha aacgtt\beta$ and $I_2 = \gamma ccgc\delta$. The rule $r_1, r_2 \in R$ are *Acl*I and *Aci*I with recognition site of $(aa; cg, tt: c; cg, c)$, respectively. Thus, the generated splicing languages from the system are presented as follows:

$$(\alpha aacgtt\beta, \gamma ccgc\delta) \mapsto^{(r_1, r_2)} \{\alpha aacgtt\alpha', \beta' aacgtt\beta, \gamma ccgg\gamma', \delta' gcgc\delta, \alpha aacgtt\beta, \gamma ccgc\delta, \alpha aacgc\delta, \gamma ccgtt\beta, \alpha aacgg\gamma', \delta' gcgtt\beta\}$$

These splicing languages are then presented in the directed splicing graph as shown in Figure 2.

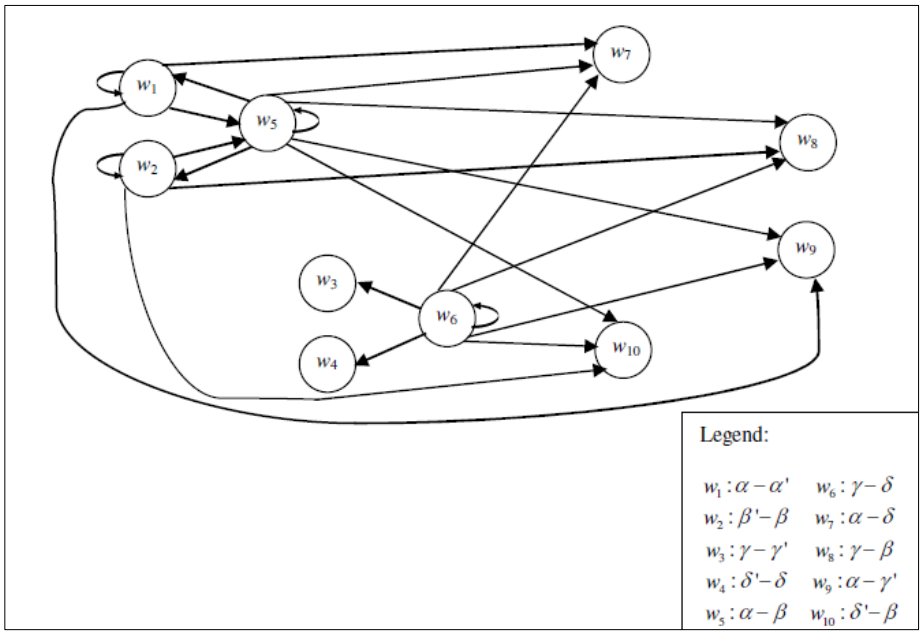

Fig. 2. Directed splicing graph of the generated splicing languages

The directed splicing graph in Figure 2 shows the existence of active persistent languages (w_1, w_2 and w_5), inert persistent languages (w_3, w_4, w_7, w_8, w_9 and w_{10}), while w_6 is the only transient language exist in the system. The determination of active persistent language in the directed splicing graph is depend on the Definition 7, which is related to the SCC. The characteristics of vertices discussed in directed splicing graph in Figure 2 are then used in the discussion of Case 1 and Case 2 in the following part.

3. Results and Discussion

In this section, the Laun's experiment results are discussed in two cases, motivated by the characteristics of restriction enzymes which are in palindromic and non-palindromic. This is to see if the directed splicing graph can predict the formation of other types of languages hinge on the features of the restriction enzymes, beyond the one that gained in the previous conducted experiment. Case 1 is represented from Laun and Reddy [18], where the experiment was carried out and the outcomes were gained. Meanwhile, Case 2 is also taken from Laun and Reddy [18], however, the experiment was not done yet but the results were predicted.

3.1 Case 1: Two Initial Strings with Two Non-Palindromic Rules

In this case, the characterization of splicing language is based on the experiment conducted. In the experiment, two initial strings were involved, with two different restriction enzymes, which are rewritten in Y-G splicing system as follows:

Let $S = (A, I, R)$ be a Y-G splicing system comprising of two initial strings, $I_1, I_2 \in I$ where $I_1 = \alpha gccgcaccggc\beta$ and $I_2 = \gamma caccacgtg\delta$. The associated restriction enzymes in the system, $r_1, r_2 \in R$ are *Bgl*I and *Dra*III, with restriction sites $(gccg, cac; cggc: cac, cac; gtg)$, respectively. The initial strings were spliced with the influence of the respective rules, producing the following languages, $L(S)$:

$I \cup \{\alpha gccgcacgtg\delta, \gamma caccaccggc\beta\}$.

The verified experiment assumed the splicing languages, $L(S)$ to be in the form of adult languages, other than the initial strings itself [18]. The splicing languages are then presented in the form of directed splicing graph. From the graph, the existence of active persistent language as defined by Yusof [14] is explored. The splicing languages are shown in the directed splicing graph given below:

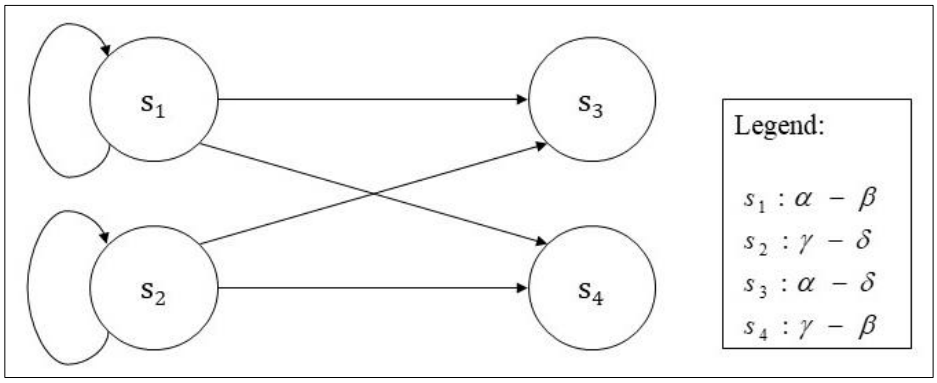

Fig. 3. Directed splicing graph for Case 1

In Figure 3, the vertices of s_1, s_2, s_3 and s_4 represent $\alpha gccgcaccggc\beta$, $\gamma caccacgtg\delta$, $\alpha gccgcacgtg\delta$ and $\gamma caccaccggc\beta$, respectively. From the figure, transient language, s_1 and s_2 cannot be classified as SCC because it does not contain 'in' edges from other vertices. Thus, with comparison made to Yusof [14], s_1 and s_2 are transient language but not active persistent language. Meanwhile, s_3 and s_4 which are the inert persistent language, can be classified as adult language as there is no restriction enzyme exist from the New England Biolabs (NEB) catalogue [21], to further cut the strings.

3.2 Case 2: Two Initial Strings with Two Palindromic Rules

In this case, the characterization of splicing language is based on the suggestion made by Laun and Reddy [18]. The author proposed an experiment by using two restriction enzymes. The splicing languages are expected to be in 10 intermediate dsDNA with no sticky ends and four adult languages. Meanwhile, Yusof [14] suggested that this attempt of an experiment will produce inert persistent language and transient language. The Y-G splicing system for this proposed of experiment is written as follows:

Given $S = (A, I, R)$ is a Y-G splicing system, with $I_1 = \alpha agatct\beta$ and $I_2 = \gamma ggatcc\delta$, where $I_1, I_2 \in I$. The associated restriction enzymes, $r_1, r_2 \in R$ are *Bgl*II and *BamH*III, with restriction sites

$(a; gatc, t: g; gatc, c)$ respectively. The initial strings spliced when the restriction enzymes were added to the system, giving the following splicing languages, $L(S)$:

$I \cup \{\alpha agatct\alpha', \beta' agatct\beta, \gamma ggatcc\gamma', \delta' ggatcc\delta,$
$\alpha agatcc\delta, \alpha agatcc\gamma', \beta' agatcc\delta, \gamma ggatct\beta\}$

The splicing languages are presented in the directed splicing graph as follows:

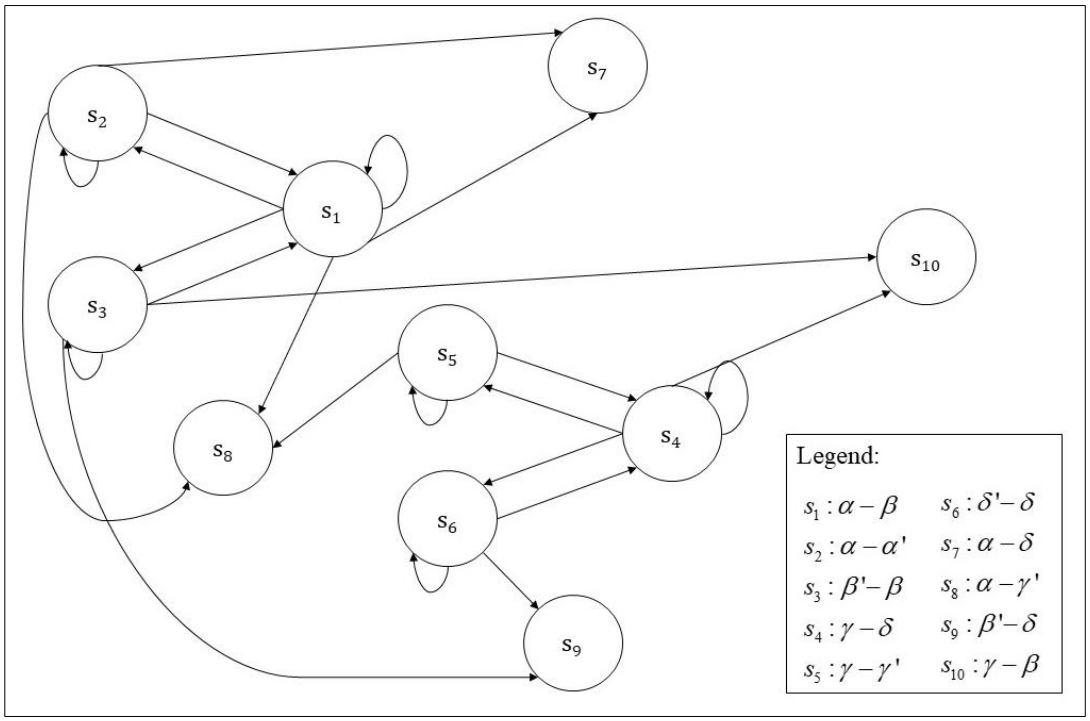

Fig. 4. Directed splicing graph for Case 2

Based on Figure 4, the active persistent language can be assumed to exist at vertices s_1, s_2, s_3, s_4, s_5 and s_6. This is due to the existence of 'in' and 'out' edges from these vertices. Thus, this graph improves the suggestion made by Yusof [14] and Laun and Reddy [18], where the transient language for the six vertices can be reclassified as active persistent language. Besides, s_7, s_8, s_9 and s_{10} are presumed to be in the form of inert persistent language as it appears to be the terminal component of SCC.

From both cases, the comparison in the type of splicing languages is depicted in Table 1. Vertex that classified as active persistent language is words that reacted with a palindromic rule, while vertex that remain as transient language is a string that reacted with non-palindromic rule [14]. In this paper, from Case 1, it can be shown that both strings were reacted with non-palindromic rules, hence the vertices are classified as transient and inert persistent languages. Nevertheless, Case 2 shows the existence of active persistent languages at certain vertex when both strings reacted with two palindromic rules.

Table 1
Comparison in the type of splicing languages generated from three sources

Results	String	Palindromic Rule	Non-Palindromic Rule	Type of Splicing Languages
[14]	I_1	r_1		Active persistent, inert persistent, transient
	I_2		r_2	
[18](Case 1)	I_1		r_1	Transient, inert persistent
	I_2		r_2	
[18](Case 2)	I_1	r_1		Active persistent, inert persistent
	I_2	r_2		

Additionally, Table 2 shows the type of splicing languages suggested from the wet model and directed splicing graph. For both cases taken from Laun and Reddy [18], the more precise type of splicing languages can be shown by using directed splicing graph.

Table 2
Type of splicing languages in wet model and directed splicing graph

Results	Laun's results	Directed splicing graph
Case 1	Transient language	Transient language
	Adult language	Inert persistent language
Case 2	Transient language	Active persistent language
	Adult language	Inert persistent language

4. Conclusions

In this paper, two cases are discussed. Case 1 and Case 2 which has taken from Laun and Reddy [18] are represented in directed splicing graph. The splicing languages obtained from Case 1 and Case 2, which contained two initial strings with two non-palindromic rules and two palindromic rules, respectively were illustrated in the form of directed splicing graphs. The results show a more precise type of splicing languages that is beyond the transient, adult and limit languages. The splicing languages are characterized with comparison to the outcome from Yusof [14]. From the previous section, it is shown that the characterization of the type of splicing languages also depends on the features of the restriction enzyme where the palindromic restriction enzyme will suggest the presence of active persistent language when the splicing languages are presented in a directed splicing graph. In the future, other types of graphs can be considered to characterize the type of splicing language. Besides, a more effective approach can be considered in determining the type or behaviour of splicing languages for instance by using graphical user interface. The simulator will benefit more users from various fields as it may reduce time, cost and give more accurate output as proposed by Syed Abdul Nasir *et al.,* [22].

Acknowledgement
The authors would like to express their gratitude to Universiti Malaysia Pahang Al-Sultan Abdullah for financial support under UMP Fundamental Research Grant RDU220344.

References
[1] Dahm, Ralf. "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research." *Human genetics* 122 (2008): 565-581. https://doi.org/10.1007/s00439-007-0433-0
[2] Glick, Bernard R., and Cheryl L. Patten. *Molecular biotechnology: principles and applications of recombinant DNA*. John Wiley & Sons, 2022.

[3] Gartner, Anton, and JoAnne Engebrecht. "DNA repair, recombination, and damage signaling." *Genetics* 220, no. 2 (2022): iyab178. https://doi.org/10.1093/genetics/iyab178
[4] Sutantyo, T. E. P., A. Ripai, Z. Abdullah, W. Hidayat, and Freddy P. Zen. "Soliton-like solution on the dynamics of modified Peyrard-bishop DNA model in the thermostat as a bio-fluid." *Emerg. Sci. J.(ISSN: 2610-9182)* 6, no. 4 (2022). https://doi.org/10.28991/ESJ-2022-06-04-01
[5] Weaver, Robert. *EBOOK: Molecular Biology*. McGraw Hill, 2011.
[6] Crick, Francis, and James Watson. "A structure for deoxyribose nucleic acid." *Nature* 171, no. 737-738 (1953): 3. https://doi.org/10.1038/171737a0
[7] Russell, Peter J. *IGenetics*. 2006.
[8] Tamarin, Robert H. *Principles of genetics*. Not Available, 2022.
[9] Head, Tom. "Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors." *Bulletin of mathematical biology* 49 (1987): 737-759. https://doi.org/10.1016/S0092-8240(87)90018-8
[10] Goode, Elizabeth, and Dennis Pixton. "Splicing to the Limit." In *Aspects of Molecular Computing: Essays Dedicated to Tom Head, on the Occasion of His 70th Birthday*, pp. 189-201. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.
[11] Sarmin, Nor Haniza, and W. H. Fong. "Mathematical modelling of splicing systems." In *Proceedings of the 1st International Conference on Natural Resources Engineering & Technology*, pp. 524-527. 2006.
[12] Fong, Wan Heng. "Modelling of splicing systems using formal language theory." PhD diss., Universiti Teknologi Malaysia, 2008.
[13] Ruslim, Nooradelena Mohd, Marta Elizabeth, Yuhani Yusof, Mohd Sham Mohamad, and Noraziah Adzhar. "Deoxyribonucleic Acid (DNA) Splicing System from Graph Theoretic Perspective." In *Journal of Physics: Conference Series*, vol. 1988, no. 1, p. 012081. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/1988/1/012081
[14] Yusof, Yuhani. "Dna splicing system inspired by bio molecular operations." PhD diss., Universiti Teknologi Malaysia, 2012.
[15] Yusof, Yuhani, Wen Li Lim, T. Elizabeth Goode, Nor Haniza Sarmin, Fong Wan Heng, and Mohd Firdaus Abd Wahab. "Molecular aspects of DNA splicing system." In *AIP Conference Proceedings*, vol. 1660, no. 1. AIP Publishing, 2015. https://doi.org/10.1063/1.4915678
[16] Mudaber, Mohammad Hassan, Yuhani Yusof, Mohd Sham Mohamad, Aizi Nor Mazila Ramli, and Wen Li Lim. "Modelling of Two Stages DNA Splicing Languages on de Bruijn Graph." *Jurnal Teknologi* 78, no. 1 (2016): 73-78. https://doi.org/10.11113/jt.v78.4236
[17] Razak, MNS Abdul, W. H. Fong, and N. H. Sarmin. "Graph splicing rules with cycle graph and its complement on complete graphs." In *Journal of Physics: Conference Series*, vol. 1988, no. 1, p. 012067. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/1988/1/012067
[18] Laun, Elizabeth, and Kalluru J. Reddy. "Wet splicing systems." In *DNA Based Computers*, pp. 73-83. 1997. https://doi.org/10.1090/dimacs/048/06
[19] Yusof, Yuhani, Nor Haniza Sarmin, T. Elizabeth Goode, Mazri Mahmud, and Fong Wan Heng. "An extension of DNA splicing system." In *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 246-248. IEEE, 2011. https://doi.org/10.1109/BIC-TA.2011.67
[20] Rosen, Kenneth H. *Discrete mathematics and its applications*. The McGraw Hill Companies,, 2007.
[21] Oldham, Ryanne C., and Michael A. Held. "Methods for detection and identification of beer-spoilage microbes." *Frontiers in Microbiology* 14 (2023): 1217704. https://doi.org/10.3389/fmicb.2023.1217704
[22] Nasir, Sharifah Noha Zahirah Syed Abdul, Nurul Ain Ab Wahab, and Mohd Agos Salim Nasir. "Graphical User Interface for Solving Non-Linear Equations for Undergraduate Students." *International Journal of Advanced Research in Future Ready Learning and Education* 30, no. 1 (2023): 25-34.