Molecular Aspects on Generalisationss of Splicing Languages

Nurul Izzaty Ismail[1,a)], Wan Heng Fong[1,b)] and Nor Haniza Sarmin[1,c)]

[1]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia,*
81310 UTM Johor Bahru, Johor, Malaysia.

[a)]Corresponding author: iamnurulizzaty1112@gmail.com
[b)]fwh@utm.my
[c)]nhs@utm.my

Abstract. The mathematical modelling of DNA splicing systems is developed from the biological process of recombinant DNA where DNA molecules are cut and reassociated with the presence of a ligase and restriction enzymes. The molecules resulting from the splicing system generate a language which is known as a splicing language using formal language theory. In previous research, the splicing languages from different splicing systems have been generalised based on the sequences of restriction enzymes. In this research, the molecular aspects on the generalisations of splicing languages are discussed to validate the splicing languages through wet lab experiment. The initial string used in this model is taken from bacteriophage lambda. From the model, the predicted molecules resulting from the combination of the initial string and the chosen restriction enzymes are determined. The actual results from the experiment will be compared with the modelled results from the generalisations of splicing languages from DNA splicing system with palindromic and non-palindromic sequences of restriction enzymes with different crossings.

INTRODUCTION

A study relating between formal language theory and molecular biology is introduced namely deoxyribonucleic acid (DNA) splicing system in [1]. The splicing system is mathematically modelled by the biological process of recombinant DNA where DNA molecules react with restriction enzymes and a ligase through wet lab experiment. In the experiment, the restriction enzymes cut DNA at the specific sites and rejoin DNA by the ligase to form new molecules [2]. The mathematical modelling of splicing systems is called a dry model while the experimental design of splicing systems is known as a wet model [3]. By using formal language theory, DNA base pairings, DNA molecules and restriction enzymes act as double-stranded DNA (dsDNA) symbols, DNA strings and rules respectively. There are two DNA nucleobases where adenine (A) bonds with thymine (T), while cytosine (C) bonds with guanine (G) [4]. The dsDNA symbols are represented by four symbols a, c, g, and t where each symbol stands for [A/T], [C/G], [G/C] and [T/A] pairings respectively. Every rule of restriction enzyme has cleavage pattern in the form of a triple: left context, crossing and right context [1]. In this research, New England Biolabs (NEB) catalogue [5] is referred to identify all the restriction enzymes by name and sequence.

In splicing system, DNA strings are cut and reassociated with the presence of rules to generate new strings [1]. The strings resulting from the splicing system generate a language via formal language theory which is called a splicing language [1]. Formal language is a set of words or strings of symbols derived from an alphabet [6]. The symbols of empty string (λ), union (+), concatenation ($\cdot$), star closure (*) and brackets ({} or ()) are the notations for regular expressions in formal languages that are applied in this research [6]. For instance, the language L given by the expression $p^* \cdot (q + r)$ is shown in the following:

$$\begin{aligned} L(p^* \cdot (q+r)) &= L(p^*)L(q+r) \\ &= (L(p))^*(L(q)) \cup (L(r)) \\ &= p^n\{q, r\} \text{ where } n \geq 0 \\ &= \{\lambda, p, pp, ppp, \ldots\}\{q, r\} \\ &= \{q, pq, ppq, pppq, \ldots, r, pr, ppr, pppr, \ldots\} \end{aligned}$$

Proceedings of the 27th National Symposium on Mathematical Sciences (SKSM27)
AIP Conf. Proc. 2266, 060008-1–060008-6; https://doi.org/10.1063/5.0018377
Published by AIP Publishing. 978-0-7354-2029-8/$30.00

where p, q, and r denote symbols.

The splicing system is instigated by Head [1] in 1987. Over the past years, Head's splicing model had been improvised and named as Paun [7], Pixton [8], Goode-Pixton [9] and Yusof-Goode [10] splicing systems. The various splicing models evolved into different types of splicing languages such as persistent [1], adult [3], limit [9] and second order limit languages [11, 12].

The first experiment on the splicing system is carried out by Laun and Reddy [3] to validate the predicted results from splicing languages generated by Head's splicing model with restriction enzymes *Bgl*I and *Dra*III. Then, the behaviour of adult and limit languages from Head's splicing model is experimented by Fong [13] using restriction enzymes *Hpa*II and *Aci*I. Besides, Yusof et al. [14] also conducted an experiment on Yusof-Goode splicing system with restriction enzymes *Acl*I and *Aci*I to compare the predicted results from the splicing system using limit graph approach with the actual results from the experiment. In 2013, the verification of persistency properties of splicing systems involving restriction enzymes *CviQ*I and *Acc65*I through wet lab experiment is done by Karimi [15]. Furthermore, the experiment on second order limit language using restriction enzyme *Dpn*II is proposed by Ahmad et al. [16] in 2018.

Previously, research on the generalisations of splicing languages resulting from DNA splicing systems involving palindromic and non-palindromic rules has been discussed in [17, 18] where palindrome is a string that reads the same forwards and backwards. In this paper, the experimental designs on splicing systems with palindromic and non-palindromic rules are developed to verify the generalisations of splicing language through wet lab experiments. This research focuses on two cases of splicing systems with different palindromic and non-palindromic rules. The first case involves two cutting sites of the palindromic restriction enzyme *CviQ*I; while the second case involves one cutting site each of palindromic restriction enzyme *CviQ*I and non-palindromic restriction enzyme *Aci*I.

PRELIMINARIES

In this research, the generalisations of splicing languages from Head's splicing model are used. The definitions of Head's splicing model and splicing language are stated in the following.

Definition 1 [1] Splicing System and Splicing Language
A splicing system $S = (A, I, B, C)$ consists of a finite alphabet A, a finite set I of initial strings in A^*, and finite sets B and C of triples (c, x, d) with c, x and d in A^*. Each such triple in B or C is called a pattern. For each such triple the string cxd is called a site and the string x is called a crossing. Patterns in B are called left patterns and patterns in C are called right patterns. The language $L = L(S)$ generated by S consists of the strings in I and all strings that can be obtained by adjoining to $ucxfq$ and $pexdv$ whenever $ucxdv$ and $pexfq$ are in L and (c, x, d) and (e, x, f) are patterns of the same hand. A language, L is a splicing language if there exists a splicing system S for which $L = L(S)$.

Next, the definition of a palindromic string is presented.

Definition 2 [19] Palindromic String
A string I of a dsDNA is said to be palindromic if the sequence from the left to the right side of the upper single strand is equal to the sequence from the right to the left side of the lower single strand.

For instance, the enzyme *CviQ*I $\begin{array}{l}5' - \mathrm{G\,T\,A\,C} - 3' \\ 3' - \mathrm{C\,A\,T\,G} - 5'\end{array}$ is a palindrome when reading in $5'$ to $3'$ direction since the upper strand of enzyme *CviQ*I matches with the complementary strand; while the enzyme *Aci*I $\begin{array}{l}5' - \mathrm{C\,C\,G\,C} - 3' \\ 3' - \mathrm{G\,G\,C\,G} - 5'\end{array}$ is not a palindrome since the upper single strand and the complementary strand are not the same.

In the next section, the modellings of DNA splicing systems with palindromic and non-palindromic restriction rules are presented. From the models, the splicing language is generalised to form the set of all DNA strings resulting from the corresponding splicing systems.

METHODS

In this section, the DNA strings are predicted and obtained from the generalisations of splicing languages from the splicing systems with different rules. The generalisation of splicing languages from DNA splicing system with two non-overlapping cutting sites of a palindromic rule is presented in Theorem 1.

Theorem 1 [17]
Let $S = (A, I, B, C)$ be a DNA splicing system in which $A = (a, c, g, t)$ is the set of dsDNA symbols, $I = \{\alpha x_1 y x_2 \beta x_1 y x_2 \gamma\}$ is the set consisting of an initial string with two non-overlapping cutting sites of a palindromic rule $x_1 y x_2$ where α, x_1, y, x_2, β and γ are variables used to denote any arbitrary dsDNA which can be rotated 180°, represented as α', x_1', y', x_2', β' and γ' respectively, set $B = \{(x_1, y, x_2)\}$ is the set of cleavage pattern for the rule where y is the crossing and set C is the empty set, then the resulting splicing language consists of strings of the form

$$(\alpha + \gamma')x_1 y x_2((\beta + \beta')x_1 y x_2)^{n-1}(\gamma + \alpha')$$

where $n \in \mathbb{Z}^+$ represents multiple copies of strings and $x_1 y x_2 \in \{\alpha, \beta, \gamma\}$.

Next, the generalisation of splicing languages from DNA splicing system with one cutting site each of one palindromic and one non-palindromic rules with different crossings is presented in Theorem 2.

Theorem 2 [18]
Let $S = (A, I, B, C)$ be a DNA splicing system in which $A = (a, c, g, t)$ is the set of dsDNA symbols, $I = \{\alpha x_1 y x_2 \beta w_1 z w_2 \gamma\}$ is the set consisting of an initial string with two non-overlapping cutting sites of one palindromic and one non-palindromic rules $x_1 y x_2$ and $w_1 z w_2$ respectively where α, x_1, y, x_2, β, w_1, z, w_2 and γ are variables used to denote any arbitrary dsDNA which can be rotated 180°, represented as α', x_1', y', x_2', β', w_1', z', w_2' and γ' respectively, set $B = \{(x_1, y, x_2), (w_1, z, w_2)\}$ is the set of cleavage pattern for the rules where y and z are the crossings and set C is the empty set, then the resulting splicing language consists of strings of the form

$$(\alpha + \gamma' w_2' z' w_1' \beta')x_1 y x_2(\beta w_1 z w_2 \gamma + \alpha')$$

where $\{x_1 y x_2,\ w_1 z w_2,\ w_2' z' w_1'\} \in \{\alpha, \beta, \gamma\}$.

RESULTS AND DISCUSSION

In this research, the molecular aspects on generalisations of splicing languages from two different cases of splicing systems are designed to validate the splicing languages through wet lab experiments. First, the mathematical model of splicing systems S_1 involving two non-overlapping cutting sites of a palindromic restriction enzymes *CviQ*I (*g*, *ta*, *c*) is developed. From the generalisation of splicing languages in Theorem 1, the splicing language from this splicing system is shown in the following:

$$L(S_1) = (\alpha + \gamma')gtac((\beta + \beta')gtac)^{n-1}(\gamma + \alpha') \tag{1}$$

where $n \in \mathbb{Z}^+$. The initial string I_1 used in this model is taken from bacteriophage lambda between 12020 and 12209 (190 bp) which contains two cutting sites of enzymes *CviQ*I. The length of fragment for the initial string is presented as follows:

$$\begin{gathered}\text{Fragment: } \alpha - \textit{CviQ}\text{I site} - \beta - \textit{CviQ}\text{I site} - \gamma \\ |\alpha| = 52 \text{ bp} \\ |\textit{CviQ}\text{I site}| = 4 \text{ bp} \\ |\beta| = 76 \text{ bp} \\ |\gamma| = 54 \text{ bp.}\end{gathered}$$

The genome locations for the first and second cutting sites are found at 12072-12075 and 12152-12155 respectively. Table 1 shows the size (bp) of predicted molecules which are obtained from the splicing language $L(S_1)$.

This experiment involves a technique of gel electrophoresis to show the presence of DNA molecules based on the molecular sizes which are generated when the process of digestion and ligation of enzymes is carried out. Low Molecular Weight (LMW) DNA Ladder is used in this research to determine the molecular sizes on the gel with the size range from 25 bp to 766 bp. The predicted gel of *CviQ*I digestion and ligation towards I_1 is shown in Figure 1.

TABLE 1. The size (bp) of predicted molecules for $L(S_1)$

No.	Molecule	Size (bp)	No.	Molecule	Size (bp)
1.	α	52	9.	$\alpha-\beta-\alpha'$	188
2.	β	76	10.	$\alpha-\beta'-\alpha'$	188
3.	γ	54	11.	$\gamma'-\beta-\gamma$	192
4.	$\alpha-\beta-\gamma$	190	12.	$\gamma'-\beta'-\gamma$	192
5.	$\alpha-\gamma$	110	13.	$\alpha-(\beta+\beta')^{n-1}-\gamma$	$110+(n-1)(80)$
6.	$\alpha-\alpha'$	108	14.	$\alpha-(\beta+\beta')^{n-1}-\alpha'$	$108+(n-1)(80)$
7.	$\gamma'-\gamma$	112	15.	$\gamma'-(\beta+\beta')^{n-1}-\gamma$	$112+(n-1)(80)$
8.	$\alpha-\beta'-\gamma$	190			

where $n \in \mathbb{Z}^+$.

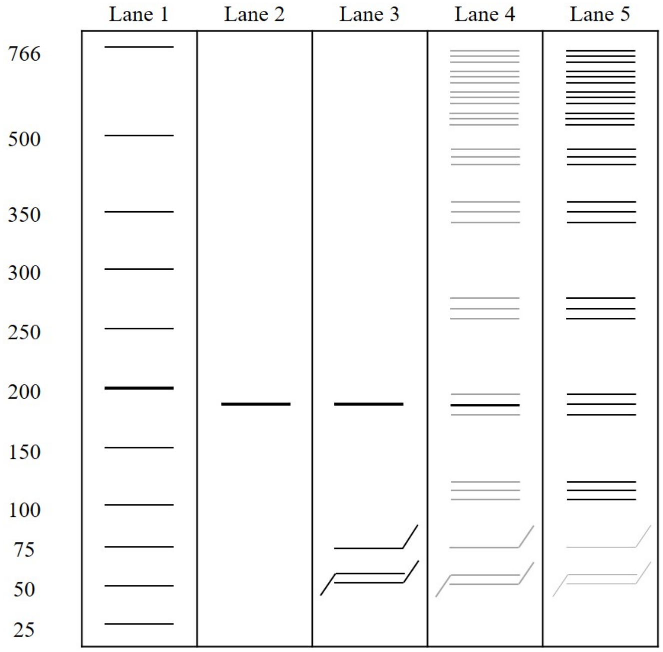

FIGURE 1. Predicted gel of *CviQ*I digestion and ligation towards I_1.

In Figure 1, the first lane indicates the bands from LMW DNA Ladder with the molecular sizes 25, 50, 75, 100 ,150, 200, 250, 300, 350, 500 and 766 bp. The second lane shows the presence of initial string I_1 which gives 190 bp. In the third lane, there exists the initial string and sticky ends of α, β and γ. Lanes 4 and 5 show the existence of the molecules resulting from the splicing language $L(S_1)$ in (1), where the sizes of molecules are given in Table 1.

Next, a mathematical model of splicing systems S_2 involving with one cutting site each of palindromic restriction enzyme *CviQ*I (g, ta, c) and non-palindromic restriction enzyme *Aci*I (c, cg, c) with different crossings is developed. From the generalisation of splicing languages in Theorem 2, the splicing language from this splicing system is shown in the following:

$$L(S_2) = (\alpha + \gamma' gcgg\beta')gtac(\beta ccgc\gamma + \alpha'). \tag{2}$$

The initial string I_2 used in this model is taken from bacteriophage lambda between 42958 and 43117 (160 bp) which contains one cutting site each of both the enzymes *CviQ*I and *Aci*I. The length of fragment for the initial string is presented as follows:

Fragment: $\alpha -$ *CviQ*I site $- \beta -$ *Aci*I site $- \gamma$
$|\alpha| = 34$ bp

$$|CviQ\text{I site}| = 4 \text{ bp}$$
$$|\beta| = 40 \text{ bp}$$
$$|Aci\text{I site}| = 4 \text{ bp}$$
$$|\gamma| = 78 \text{ bp}.$$

The genome locations for the first and second cutting sites are found at 42992-42995 and 43513-43516 respectively. Table 2 shows the size (bp) of predicted molecules which are obtained from the splicing language $L(S_2)$.

TABLE 2. The size (bp) of predicted molecules for $L(S_2)$

No.	Molecule	Size (bp)	No.	Molecule	Size (bp)
1.	α	34	4.	$\alpha - \beta - \gamma$	160
2.	β	40	5.	$\alpha - \alpha'$	72
3.	γ	78	6.	$\gamma' - \beta' - \beta - \gamma$	248

The predicted gel of *CviQ*I and *Aci*I digestion and ligation towards I_2 is shown in Figure 2.

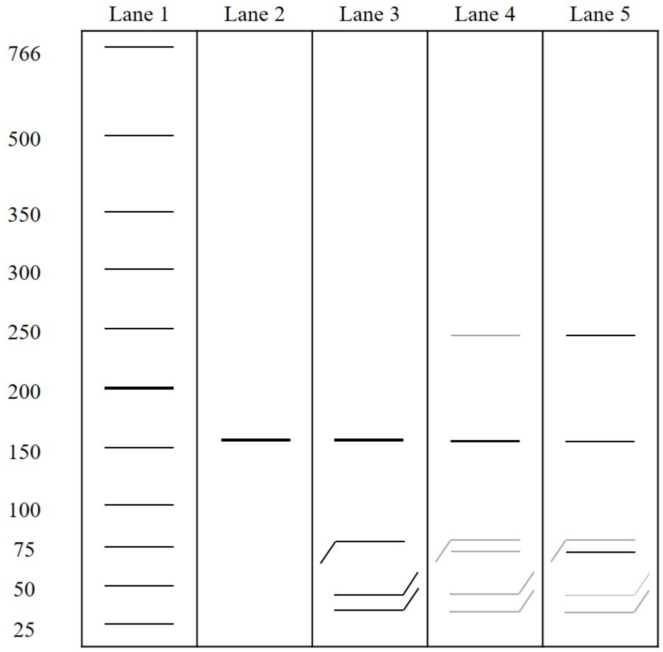

FIGURE 2. Predicted gel of *CviQ*I and *Aci*I digestion and ligation towards I_2

In Figure 2, the first lane indicates the bands from LMW DNA Ladder. The second lane shows the presence of initial string I_2 which gives 160 bp. In the third lane, there exists the initial string and sticky ends of α, β and γ. Lanes 4 and 5 show the existence of the molecules resulting from the splicing language $L(S_2)$ in (2), where the sizes of molecules are given in Table 2.

CONCLUSION

In this paper, the experimental designs of generalisations of splicing languages from the splicing systems with palindromic and non-palindromic restriction enzymes are developed where the initial strings used in this research are taken from amplified bacteriophage lambda. The first experiment is designed using two cutting sites of palindromic restriction enzyme *CviQ*I; while the second experiment is designed involving one cutting site each of palindromic restriction enzyme *CviQ*I and non-palindromic restriction enzyme *Aci*I. From the results, the predicted molecules are determined from the sets of string in the generalisation of splicing languages. For future research, the actual results from the experiment will be compared with the modelled results from the generalisations of splicing languages from DNA splicing system with palindromic and non-palindromic rules.

ACKNOWLEDGMENTS

The first author would like to thank Universiti Teknologi Malaysia (UTM) for supporting her study through Zamalah Scholarship. The second and third authors would also like to express their gratitude to the Ministry of Education (MOE) and Research Management Centre (RMC), Universiti Teknologi Malaysia (UTM) for the funding through Fundamental Research Grant Scheme Vote No. 5F022.

REFERENCES

[1] T. Head, Bulletin of Mathematical Biology **49**, 737–759 (1987).

[2] A. J. F. Griffiths, S. R. Wessler, S. Carroll, and J. Doebley, *An Introduction to Genetic Analysis.*, 11th ed. (Macmillan Learning, New York, 2015).

[3] E. Laun and K. J. Reddy, “Wet splicing systems,” in *3rd DIMACS Workshop on DNA Based Computers*, Vol. 48, edited by H. Rubin and D. H. Wood (American Mathematical Society, Rhode Island, USA, 1999), pp. 73–84.

[4] G. Paun, G. Rozenberg, and A. Salomaa, *DNA Computing: New Computing Paradigms* (Springer -Verlag Berlin Heidelberg, Germany, 1998).

[5] N. E. B. Inc, “Neb 2017-18 catalog & technical reference,” (2017).

[6] P. Linz, *An Introduction to Formal Languages and Automata*, 4th ed. (Jones and Bartlett Publisher, USA, 2006).

[7] G. Paun, Discrete Applied Mathematics **70**, 57–79 (1996).

[8] D. Pixton, Discrete Applied Mathematics **69**, 101–124 (1996).

[9] E. Goode and D. Pixton, “Splicing to the limit,” in *Aspects of Molecular Computing, Lecture Notes in Computer Science*, edited by N. Jonoska, G. Paun, and G. Rozenberg (Springer-Verlag, Germany, 2004), pp. 189–201.

[10] Y. Yusof, N. H. Sarmin, W. H. Fong, T. E. Goode, and M. A. Ahmad, “An analysis of four variants of splicing system,” in *20th National Symposium on Mathematical Sciences - Research in Mathematical Sciences: A Catalyst for Creativity and Innovation (SKSM 2012)*, Vol. 1522 (AIP Conference Proceedings, Melville, NY, 2013), pp. 888–895.

[11] M. A. Ahmad, N. H. Sarmin, W. H. Fong, and Y. Yusof, “An extension of first order limit language,” in *3rd International Conference on Mathematical Sciences (ICMS3)*, Vol. 1602 (AIP Conference Proceedings, Melville, NY, 2014), pp. 627–631.

[12] M. A. Ahmad, N. H. Sarmin, W. H. Fong, Y. Yusof, and N. Adzhar, International Journal of Innovative Technology and Exploring Engineering **8**, 367–372 (2019).

[13] W. H. Fong, “Modelling of splicing systems using formal language theory,” Thesis, Universiti Teknologi Malaysia 2008.

[14] Y. Yusof, W. L. Lim, T. E. Goode, N. H. Sarmin, W. H. Fong, and M. F. A. Wahab, “Molecular aspects of dna splicing system,” in *AIP Conference Proceedings*, Vol. 1660 (AIP Publishing, 2015), pp. 050045 1–8.

[15] F. Karimi, “Mathematical modelling of persistent splicing systems in dna computing,” Thesis, Universiti Teknologi Malaysia 2013.

[16] M. A. Ahmad, N. H. Sarmin, M. F. Abdul-Wahab, W. H. Fong, and Y. Yusof, Malaysian Journal of Fundamental and Applied Sciences **14**, 15–19 (2018).

[17] W. H. Fong and N. I. Ismail, Malaysian Journal of Industrial and Applied Mathematics **34**, 59–71 (2018).

[18] N. I. Ismail, W. H. Fong, and N. H. Sarmin, “The mathematical modelling of dna splicing system with palindromic and non-palindromic restriction enzymes,” in *International Conference on Applied Analysis and Mathematical Modelling* (ICAAMM 2018, Istanbul Gelism University, Istanbul, Turkey, 2018), pp. 127–138.

[19] Y. Yusof, “Dna splicing system inspired by bio molecular operations,” Thesis, Universiti Teknologi Malaysia 2012.