PAPER • OPEN ACCESS

n-cutting site of DNA splicing language for single string and palindromic rule

n-cutting site of DNA splicing language for single string and palindromic rule

N M Ruslim[1], Y Yusof[1] and N Adzhar[1]

[1]Centre for Mathematical Sciences, College of Computing and Applied Sciences
Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan,
Pahang, Malaysia

E-mail: PSE20003@stdmail.ump.edu.my, yuhani@ump.edu.my

Abstract. A new symbolization of Yusof-Goode (Y-G) splicing system was introduced by Yusof in 2012, is inspired by the framework of Formal Language Theory introduced by Head in 1987. Y-G splicing system is intended to present the biological process of DNA splicing in a translucent way. In this paper, starting with some relevant preliminaries, one theorem is proposed via Y-G approach using one initial string and one rule with different characteristics of the restriction enzyme. Additionally, the theorem showed the behavior of the splicing languages generated at single stage splicing. Two cases are considered in the theorem by conducting splicing using palindromic rule and palindromic recognition site with same left and right context for Case I and different left and right context for Case II. Furthermore, two molecular examples are discussed to validate two cases proposed in the theorem, which shows the real meaning of the theorem in biological aspect. From the proposed theorem, based on the splicing language generated, the type of splicing language can be determined. It is discovered that, the generated languages are in the form of limit and transient.

1. Introduction

The DNA recombinant technology is being used in animal and plant breeding, to produce antibiotics, hormones and other medically important agents [1]. This laboratory experiment definitely will cause a huge amount of money and high time investment, without promising result. Thus, with Formal Language Theory [2], a new strategy in biomolecular field is founded. The introduction of Formal Language Theory has sparked interest among researchers to explore the DNA splicing system with utilization of various mathematical modelling, for instance probability and automata [3–5]. By adopting probabilistic approach, some properties of probabilistic semi-simple splicing systems are investigated and proven that the generative power of the splicing languages is increase [3]. Other than probabilistic approach, automata has also been chosen by many researchers to be applied in solving DNA splicing problems. Fong *et al.* [4] studied on the relation between automata and subgroups. By using modified automata, the conditions for the recognition of subgroups were established. Following, in 2019, Fong *et al.* [5] applied the automata diagrams to visualize the splicing languages generated from DNA splicing system. The automata diagrams are presented by transition graphs where the representation of the languages is from the respective DNA splicing systems. These has proven that via mathematical modelling, DNA splicing system can be better explored.

In 2012, some modification on the rule notation proposed by Yusof [6] to the splicing system initiated by Head, has encouraged more researchers to explore this field via Yusof-Goode (Y-G) splicing system [7–9]. Most recent researches focused on the language generated from splicing system, called splicing language, namely second order limit language [7], single stage splicing language [8] and two stages splicing language [9]. Other than employing the Y-G splicing system in their researches, authors applied variety of mathematical method in order to prove and validate the proposed languages, that is, by using automata, limit adjacency matrix and de Bruijn graph, respectively.

In 2004, three types of splicing languages are classified that is limit, adult or inert, and transient language [10]. Then, after eight years, Yusof has redefined and renamed the inert language to inert persistent language. A new type of splicing language called active persistent language is also introduced [6]. In this paper, one theorem is presented, which employed the new definition founded in 2015 [8]. The theorem is modelled via Y-G splicing system, which aim to investigate the effect of n-cutting to the generated languages. Two molecular examples are then predicted by considering two cases disclosed in the theorem.

This paper comprises of four segments. The first segment is the introduction, followed by the second segment which stated some important definitions used in this paper. The main results and discussion are then presented in the third segment, with the proposition of theorem and disclosure of molecular examples. The generalized splicing languages from Case I and Case II are obtained and discussed. Finally, the results are summarized in the final segment.

2. Methodology

This segment itemizes some fundamental definitions used in this paper. Since this paper is based on Yusof-Goode approach, therefore the definition of Y-G splicing system and splicing language is first introduced:

Definition 1 [6] *A Y-G splicing system $S = (A, I, R)$ consists of a set of alphabets A, a set of initial strings I in A^* and a set of rules, $r \in R$ where $r = (u; x, v: y; x, z)$ for left pattern, $r = (u, x; v: y, x; z)$ for right pattern or $r = (u, x, v: y, x, z)$ for both patterns of rules, applied on DNA string. For $s_1 = \alpha uxv\beta$ and $s_2 = \gamma yxz\delta$ elements of I, splicing s_1 and s_2 using r produces the initial string I together with $\alpha uxz\delta$ and $\gamma yxv\beta$, presented in either order where $\alpha, \beta, \gamma, \delta, u, x, v, y$ and $z \in A^*$ are the free monoids generated by A with the concatenation operation and 1 as the identity element. A language L is a splicing language if there exists a splicing system S for which $L = L(S)$.*

Then, definition of single stage splicing language is stated.

Definition 2 [11] *Let $S = (A, I, R)$ be the Y-G splicing system. The set of single stage splicing language, $L_1 = L_1(S)$, models the set of all molecule types which appear when all restriction enzymes, double stranded deoxyribonucleic acid (dsDNA) strings and ligases act simultaneously in a single buffer.*

Next, definition of limit language is provided.

Definition 3 [10] *A limit language is the set of words that are predicted to appear if some amount of each initial molecule is present, and sufficient time has passed for the reaction to reach equilibrium state, regardless of the balance of the reactants in a particular experimental run of the reaction.*

Lastly, definition of transient language is given as follows.

Definition 4 [6] *A transient splicing language is a set of strings that will ultimately be used up and disappear.*

3. Results and Discussion

In this segment, to investigate the effect of *n*-cutting to the splicing language, one theorem is proposed. The theorem which is modelled via Y-G splicing system shows the presence of different type of generated languages. This is significant to the behavior of DNA splicing for one rule on a single string [8]. Through Theorem 1, prediction of single stage splicing languages based on palindromic rule and palindromic crossing site is presented.

Theorem 1. *Let $S = (A, I, R)$ be a Y-G splicing system. If $I = \{s\}$ is a set of non-palindromic initial string and the element of set $R = \{r\}$ is a palindromic rule which contains palindromic crossing site of more than one cutting site, then $n\big(L_1(S)\big) > 3$, in the sequence of $\alpha - \beta, \alpha - \alpha', \beta' - \beta$ with infinitely long molecules.*

Proof. Suppose $S = (A, I, R)$ be a Y-G splicing system, consists of a set of non-palindromic initial string, $I \in A^*$ and n palindromic crossing sites of a palindromic rule, $r \in R$. Hence, two cases need to be considered, which are:

Case I: same left and right contexts in a rule

For the same left and right contexts, a rule $r \in R$ for $\forall a, b, u, v \in A^*$ is presented in the form of $(ab, uv, ab\colon ab, uv, ab)$. Let $s = \alpha abuvab \ldots abuvab\beta$ where $a = b'$ and $u = v'$, $\forall \alpha, \beta, a, b, u, v \in A^*$. Due to the palindromic properties in the restriction enzyme, there is possibility for the sticky ends to religate with its 180° rotation molecule. Hence, the generated splicing languages are: $I \cup \{\alpha ab(uv)^k ab\alpha', \beta' ab(uv)^k ab\beta\ \}$ where $k \geq 0$ with $(uv)^k$ can be infinitely long fragments either in $(uv)^k$ or $(u'v')^k$ order. Therefore, three patterns of splicing languages exist.

Case II: different left and right contexts in a rule

For different left and right contexts, let $r \in R$ be $(a, uv, b\colon a, uv, b)$. Let $s = \alpha auvb \ldots auvb\beta$ be a string in I where a is the complement of b and u is the complement of v, $\forall \alpha, \beta, a, b, u, v \in A^*$. Due to the palindromic properties in the restriction enzyme, the sticky ends have chance to religate with its 180° rotation of itself to generate:
$I \cup \{\alpha a(uv)^k b\alpha', \beta' a(uv)^k b\beta\ \}$ where $k \geq 0$ with $(uv)^k$ can be infinitely long fragments. Hence, three patterns of splicing languages exist.

It is apparent that, both cases, I and II lead to the same outcomes, which produced splicing languages in the sequence of $\alpha - \beta, \alpha - \alpha', \beta' - \beta$. Hence, the theorem proved.

3.1 Some molecular examples of single string with one palindromic rule in Y-G splicing system

A splicing model proposed in [6] has shown that single string with one palindromic rule will generate more than 3 infinitely long splicing languages in the sequence of $\alpha - \beta, \alpha - \alpha', \beta' - \beta$, which is parallel to Theorem 1 above. The patterns produced in both cases are consistent with theorem proposed by Lim which suggests that, for one recognition site in an initial string, with palindromic characteristics of the crossing site, 3 patterns of splicing languages are generated [8]. However, from the suggested theorem above, it is observed that, if an initial string consists of more than one recognition site, the generated languages will be in the sequence of $\alpha - \beta, \alpha - \alpha', \beta' - \beta$, but with more than 3 infinitely long molecules. Thus, this result shows the effect of *n*-cutting to the generated languages in a splicing system. Therefore, to validate Case I and Case II respectively, two molecular examples are provided.

Example 1: Let $S = (A, I, R)$ be a Y-G splicing system with $A = \{a, g, c, t\}$ where $\alpha, \beta \in A^*$. Initial string $I = \alpha cgatcgcgatcg\beta$ and a palindromic restriction enzyme, *Pvu*I, with palindromic recognition site of right cleavage pattern on 3' overhang, $r = (cg, at; cg: cg, at; cg)$. With the presence of the enzyme, string I will cleave as follows:

$$I_0 = \begin{matrix} \alpha CGAT^{\blacktriangledown}CGCGAT^{\blacktriangledown}CG\beta \\ \alpha' GCTAGC_{\blacktriangle}GCTAGC_{\blacktriangle}\beta' \end{matrix}$$

Consequently, the string can split to its 180° rotation as follows:

$$I_{180} = \begin{matrix} \beta' CGAT^{\blacktriangledown}CGCGAT^{\blacktriangledown}CG\alpha' \\ \beta GCTAGC_{\blacktriangle}GCTAGC_{\blacktriangle}\alpha \end{matrix}$$

When splicing occurs at two cutting sites, with the reaction of ligase, the above molecules, I_0 and I_{180} can religate to form new molecules. Given $k \geq 0$, the following splicing languages are generated:

$$\{\alpha cgat(cgcgat)^k cg\beta, \alpha cgat(cgcgat)^k cg\alpha', \\ \beta' cgat(cgcgat)^k cg\beta\}$$

By induction, it is proved that, $n\big(L_1(S)\big) > 3$,that is:

Table 1. Conceivable strings for $0 \leq k \leq 2$

k	$\alpha - \beta$	$\alpha - \alpha'$	$\beta' - \beta$
$k = 0$	$\alpha cgatcg\beta$	$\alpha cgatcg\alpha'$	$\beta' cgatcg\beta$
$k = 1$	$\alpha cgatcgcgatcg\beta$	$\alpha cgatcgcgatcg\alpha'$	$\beta' cgatcgcgatcg\beta$
$k = 2$	$\alpha cgatcgcgatcgcgatcg\beta$	$\alpha cgatcgcgatcgcgatcg\alpha'$	$\beta' cgatcgcgatcgcgatcg\beta$

From Table 1, the conceivable strings are considered for $0 \leq k \leq 2$. When the restriction enzyme, *Pvu*I which is taken from the NEB catalogue reacted to cleave the DNA molecules, then, with the existence of ligase, the conceivable molecules are produced. For $\alpha - \beta$, $\alpha - \alpha'$ and $\beta' - \beta$ sequences, the value of k is substituted in the generalised splicing language obtained in the example. It shows that, the number of splicing languages produced by 2-cutting sites in a splicing system, will be more than 3 and infinitely long molecules, as k increases.

Example 2: Let $S = (A, I, R)$ be a Y-G splicing system with $A = \{a, g, c, t\}$ where $\alpha, \beta \in A^*$. Initial string $I = \alpha gtacgtac\beta$ and a palindromic restriction enzyme, *CviQ*I, with left cutting pattern on 5' overhang, $(g; ta, c: g; ta, c)$ and palindromic recognition site. With the existence of restriction enzyme, the string I will cleave as follows:

$$I_0 = \begin{matrix} \alpha G^{\blacktriangledown}TACG^{\blacktriangledown}TAC\beta \\ \alpha' CAT_{\blacktriangle}GCAT_{\blacktriangle}G\beta' \end{matrix}$$

Consequently, the string can split to its 180° rotation as follows:

$$I_{180} = \begin{matrix} \beta' G^{\blacktriangledown}TACG^{\blacktriangledown}TAC\alpha' \\ \beta CAT_{\blacktriangle}GCAT_{\blacktriangle}G\alpha \end{matrix}$$

When splicing occurs at two cutting sites, with the reaction of ligase, I_0 and I_{180} can religate to form new molecules of splicing languages, given $k \geq 0$:

$$\{\alpha g(tacg)^k tac\beta, \alpha g(tacg)^k tac\alpha', \beta' g(tacg)^k tac\beta\}$$

By induction, it is proved that, $n\big(L_1(S)\big) > 3$, that is:

Table 2. Conceivable strings for $0 \leq k \leq 2$.

k	$\alpha - \beta$	$\alpha - \alpha'$	$\beta' - \beta$
$k = 0$	$\alpha gtac\beta$	$\alpha gtac\alpha'$	$\beta' gtac\beta$
$k = 1$	$\alpha gtacgtac\beta$	$\alpha gtacgtac\alpha'$	$\beta' gtacgtac\beta$
$k = 2$	$\alpha gtacgtacgtac\beta$	$\alpha gtacgtacgtac\alpha'$	$\beta' gtacgtacgtac\beta$

From Table 2, the conceivable strings are shown for $0 \leq k \leq 2$. The 5' overhang restriction enzyme, *CviQ*I taken from the NEB catalogue cleaved the DNA molecules. Then, with the existence of ligase, the imaginable fragments are considered for $k = 0, 1, 2$. By substituting the value of k to the $\alpha - \beta$, $\alpha - \alpha'$ and $\beta' - \beta$ sequences, it is observed that, the middle segment will duplicate according to the value of k. This shows that, the number of splicing languages produced by 2-cutting sites in a splicing system, will produce more than 3 and infinitely long molecules, as k increases.

From previous study in [7], to generate language and to preserve the biological characteristics, two types of approaches can be applied which are laboratory experiment and splicing system. In this segment, two molecular examples are given to show the existence of splicing language in the type of limit and transient. This is supported by findings in [12], which suggests that a language resulted from the splicing system is proven to exist experimentally such as limit, adult or inert persistent, transient and active persistent language [6,12].

4. Conclusions

In this paper, one theorem of a single string with a palindromic rule and palindromic crossing site is presented with number of cutting sites of more than one. Theorem 1 proved that, either with same or different left and right contexts, the number of splicing languages in the sequence of $\alpha - \beta$, $\alpha - \alpha'$ and $\beta' - \beta$ will be more than 3. Consequently, as k increases, the molecules will be longer and infinitely long due to the duplications of the middle segment in the generalized splicing language. It can be concluded that, an *n*-crossing site in a string will produce more and longer languages, in the form of limit and transient languages.

Patterns generated in the theorem is correlated to theorem proposed by Lim in 2015 [8]. However, according to Goode, these three patterns that generates infinitely long molecules will vanish and resulted in infinite set of transient languages, while the pattern in the form of initial string, I will remain as a type of limit language [10]. The conclusion of single stage splicing language for a single string with a palindromic rule for two cases of palindromic crossing site is summarized in the table below:

Table 3. Palindromic rule with palindromic crossing site.

Left and Right Contexts	Number of Cutting Sites	Number of Splicing Language Generated	Type of Splicing Language
Same	2	$> 3^*$	Limit and transient language
Different	2	$> 3^*$	

$*$ in the sequence of $\alpha - \beta, \alpha - \alpha', \beta' - \beta$

Hence, Table 3 summarising the cutting and pasting process on certain DNA molecules. When a molecule being cut by a palindromic restriction enzyme with palindromic crossing site at two cutting sites, either for same or different left and right contexts of the chosen restriction enzymes, there will be more than 3 limit or transient languages produced. This finding hence supports the aim of this research where we want to see the effect of the number of cutting sites on the type of splicing languages. In conducting this research, the chosen restriction enzyme is only limited to the palindromic characteristics. Nevertheless, for future work, other characteristics of restriction enzyme can be considered. Additionally, the application of graph theory in presenting the generated splicing languages can be proposed.

Acknowledgement

The authors would like to express their special thanks to University College of Yayasan Pahang (UCYP) for supporting this research and opportunities to attend the conference.

References

[1] Russell P J 2014 *iGenetics A Molecular Approach,* 3rd Ed. (USA: Pearson Education Limited) chapter 10 pp 296 - 349

[2] Head T 1987 Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviors *Bulletin of Mathematical Biology* **49**(6) 737-759

[3] Selvarajoo M, Fong W H, Sarmin N H and Turaev S 2013 Probabilistic Semi-Simple Splicing System and Its Characteristics *Jurnal Teknologi* **62**(3) 21-26

[4] Fong W H, Gan Y S, Sarmin N H and Turaev S 2014 Automata for subgroups *AIP Conference Proceedings* **1602** pp 632-639

[5] Fong W H, Ismail N I and Sarmin N H 2019 Automata for DNA splicing languages with palindromic and non-palindromic restriction enzymes using grammars *Matematika* Special Issue (December) pp 1-14

[6] Yusof Y 2012 DNA Splicing System Inspired by Bio Molecular Operations (Universiti Teknologi Malaysia, Malaysia) chapter 3-6 pp 25-83

[7] Ahmad M A 2016 Second Order Limit Language and Its Properties in Yusof-Goode Splicing System (Universiti Teknologi Malaysia, Malaysia) chapter 1 p 2

[8] Lim W L 2015 Single Stage DNA Splicing System via Yusof-Goode Approach (Universiti Malaysia Pahang, Malaysia) chapter 3-5 pp 22-82

[9] Mudaber M H 2015 Persistency and Permanency of Two Stages Splicing Languages Based on DNA Recombination Process by using Yusof-Goode (Y-G) Approach (Universiti Malaysia Pahang, Malaysia) chapter 1 p 1

[10] Goode T E and Pixton D 2004 Splicing to the Limit *Aspects of Molecular Computing* **2950** pp 189-201

[11] Lim W L, Yusof Y and Mudaber M H 2015 Modelling of DNA Single Stage Splicing Language via Yusof-Goode Approach: One String with Two Rules *AIP Conference Proceedings* **1643** pp 695-699

[12] Laun E and Reddy K 1999 Wet Splicing Systems *DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences* **48** 73-83