On the New Relation of Second Order Limit Language and Other Different Types of Splicing System

Muhammad Azrin Ahmad[1,a)], Nor Haniza Sarmin[2,b)], Wan Heng Fong[2,c)], Yuhani Yusof[1,d)] and Noraziah Adzhar[1,e)]

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

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

[a)]Corresponding author: azrin@ump.edu.my
[b)]nhs@utm.my
[c)]fwh@utm.my
[d)]yuhani@ump.edu.my
[e)]noraziah@ump.edu.my

Abstract. Mathematical modelling of splicing system has been introduced to initiate a linkage between the study of informational macromolecules that includes DNA and formal language theory. The ability to present the nitrogenous base which is a component in a nucleotide of DNA, as a series of alphabet, ignites this interdisciplinary study. Over the years, researchers have developed models to match their need. In addition, product of splicing system is called splicing language. Through some development, second order limit language is derived from other type of languages known as limit language. Its existence and characteristics have been vastly discussed. Beside, some types of splicing system can produce second order limit language. In this research, the characteristics of varieties of splicing system are studied and their relations with second order limit language are established.

INTRODUCTION

Deoxyribonucleic acid (DNA) is an important molecule that exist in living organism [1]. It plays important roles that involve proteins production and also helps pass the particulars from one generation to another [2]. A DNA molecule is a composition of nucleotide that made up of three distinct components such as phosphate group, a sugar group (deoxyribose) and a nitogenous base [3]. Nitrogenous bases can be classified as Cytosine (C), Thymine (T), Guanine (G) and Adenine (A). Based on Watson-Crick complementarity, only certain pairings are allowed such as T with A, G with C and vice versa. Double-stranded DNA (dsDNA) molecules can be cut with the presence of restriction endonuclease and pasted with the existence of ligase. Then, either a new or hybrid DNA or the same DNA molecules is formed [4]. These pairings and situations then inspired researchers to study the mathematical modelling of DNA splicing system where a set of alphabets represents pairings between the nitrogenous bases, initial strings represent dsDNA and restriction enzymes can be represented by a set of rules [5].

Besides, some other interesting findings on varieties of splicing system were discovered by Head around 1987 [5], namely persistent, null context, uniform splicing system and also strictly locally testable (SLT) language. These types of splicing systems have contributed to the exploring of new properties to the existing splicing systems. As an example, an in-depth study on the characterisation of persistent splicing system has shown a relation with strictly locally testable language [5].

In addition, Mateescu *et al.* in 1998 [6] has defined simple splicing system which was then further studied by Laun in 1999 [7] based on the continuity aspect. An extension of a simple splicing system namely semi-simple

Proceedings of the 27th National Symposium on Mathematical Sciences (SKSM27)
AIP Conf. Proc. 2266, 060003-1–060003-5; https://doi.org/10.1063/5.0018075
Published by AIP Publishing. 978-0-7354-2029-8/$30.00

splicing system and also null context splicing system have been introduced by Goode and Pixton in 2001 [8] while a semi-null splicing system has been introduced by Laun in 1999 [7]. A null context splicing system is determined when the left and right context of the rule is set to any string. Then, it is proven that S_kH system which is a type of null context splicing system is a simple splicing system if $k = 1$ which proves that the union of S_kH families is a family of SLT languages. Furthermore, a solid code is applied to the S_kH system to reduce it to the simple splicing system by Fong in 2008 [9]. Solid codes are identified when two conditions are literally fulfilled: no word in a set $S \in A^*$ is a subword of any other word in S and no two distinct words in S overlap non-trivially. When R is a solid code, S_kH may be reduced to S_1H=SH splicing system given that the concept of maximal firm subword of the SH splicing system and words in the solid code R are also viewed.

In addition, some relations are established where simple splicing system is a subset of semi-simple splicing system and also a subset of semi-null splicing system. In addition, simple splicing system is a subset of uniform splicing system which is a subset of S_kH system and also a subset of null context splicing system [10]. Fong [9] conducted an intensive study to explore more properties and also some sufficient conditions of SLT language. Some examples with different number of restriction enzymes were illustrated to show languages which are not strictly locally testable. The varieties of the splicing system have been used to explore some properties in two different directions: either on the generation of languages or the observation of biological traits in the splicing process.

In this paper, we will explore the formation of the second order limit language in the semi-null splicing system and the uniform splicing system. Then, the relation between these splicing systems with the presence of the second order limit language are then established.

PRELIMINARIES

The relations between the second order limit language and some varieties of the splicing system are also pursued. Firstly, the definition of Y-G splicing system is given as the model is used throughout this research. After that, the definitions of three varieties of splicing system are given.

Definition 1 ***[8] Yusof-Goode (Y-G) Splicing System***
Let splicing system $S = (A, I, R)$ be a set of alphabets A, a set of initial strings I in A^* and a set of rules, $r \in R$ where $r = (y, x, z : u, x, v)$. For $s_1 = \alpha uxv\beta$ and $s_2 = \gamma yxz\delta \in I$, $\alpha uxz\delta$ and $\gamma yxv\beta$ can be obtained by splicing s_1 and s_2 using r, presented in either order where $\gamma, \delta, \alpha, \beta, x, v, u, y$ and *$z \in A^*$ are the free monoids generated by A with the concatenation operation.*

The following definitions are transformed in [10] to fulfill the Y-G rule in the Y-G splicing system.

Definition 2 ***[8] Semi-Simple Splicing System***
$S = (A, I, R)$ is a splicing system in which I and R are finite and every rule $r \in R$ has the form $(a, 1, 1 : b, 1, 1)$ where $a, b \in A$. Then S is called a semi-simple splicing system. A language L is called a semi-simple splicing language if there exists a semi-simple splicing system S for which $L = L(S)$.

Definition 3 ***[7] Semi-Null Splicing System***
$S = (A, I, R)$ is a splicing system in which I and R are finite and every rule $r \in R$ has the form $(u, 1, 1 : v, 1, 1)$ where $u, v \in A^+$. Then S is a semi-null splicing system. A language L is called a semi-null splicing language if there exists a semi-null splicing system S for which $L = L(S)$.

Definition 4 ***[5] Uniform Splicing System***
A null context splicing system is also a uniform splicing system $S = (A, I, X, X)$ in which there is a positive integer P such that $X = A^p$ i.e. the length of crossing site is equal to the set of all strings over the alphabet A that have length exactly P where $(1, x, 1 : 1, x, 1)$ and $x \in A^p$. A language L is called a uniform splicing language if there exists a uniform splicing system S for which $L = L(S)$.

A string assigned with the numeral 1 is called a null string and also acts as the identity element.

Some relations on the aforementioned types have been established by Yusof in 2012 [10]. The relations are presented as propositions in the following.

Proposition 1 *Every semi-simple splicing system is a semi-null splicing system.*

Proposition 2 *Every simple splicing system is a semi-null splicing system.*

Proposition 3 *Every simple splicing system is a uniform splicing system.*

The next two theorems which are results in [12] will be used in proving the theorems that will be discussed later in the Main Results section.

Theorem 1 *If S is simple where I has two crossing sites, then the second order limit language exists.*

Theorem 2 *If S is semi-simple where I has two crossing sites, then the second order limit language exists.*

MAIN RESULTS

The formation of the second order limit language in those types of splicing system such as semi-null splicing system and uniform splicing system will be presented respectively in the following. In addition, the relations between these types of splicing systems emanated from the presence of the second order limit language are presented as corollaries.

The proof to show the formation of the second order limit language in the semi-null splicing system is similar to the proving of the formation of the second order limit language in the semi-simple splicing system [12]. The difference is that, the strings in the semi-null splicing system, u and v belong to A^+ instead of a, b in semi-simple splicing system which belong to a set of alphabets, A.

Theorem 3 *If S is semi-null where I has two crossing sites, then the second order limit language exists.*

Proof. Let A be a set of alphabets, $I = \{\alpha mumum\beta, \gamma pvpvp\delta\}$ is a set of initial strings such that u with u', v with v', m with m', p with p', α with α', β with β', γ with γ' and δ with δ' are complement to each other and $\gamma, \delta, \alpha, \beta, \alpha', \beta', \gamma', \delta' \in A^*$. Meanwhile, $R = \{r\}$, where $r = (v, 1, 1 : u, 1, 1)$ such that $v, u \in A^+$. In the semi-simple splicing system, since the crossing site is written in a null form, consider the splicing action in two patterns, namely the left pattern and right pattern. These patterns describe the type of cut involve in a splicing process. The following shows the resulted splicing language:

i. The splicing language of left pattern, $L_{(\text{left pattern})}$:

$$\{\alpha mumum\beta, \gamma pvpvp\delta\} \mapsto^R I \cup \{\alpha mum\beta, \alpha muu'm'\alpha', \beta'm'm\beta, \gamma pvp\delta, \gamma pvv'p'\gamma', \delta'p'p\delta\}.$$

ii. The splicing language of right pattern, $L_{(\text{right pattern})}$:

$$\{\alpha mumum\beta, \gamma pvpvp\delta\} \mapsto^R I \cup \{\alpha mum\beta, \alpha muu'm'\alpha', \beta'm'm\beta, \gamma pvp\delta, \gamma pvv'p'\gamma', \delta'p'p\delta\}.$$

When the splicing process occurs once again among the resulted splicing language obtained from the first splicing, the following new languages are formed:

$$L_2(S) = \left\{\alpha ma(mu)^k m\beta, \alpha ma(mu)^k u'm'\alpha', \beta'm'(mu)^k m\beta, \gamma pv(pv)^k p\delta \gamma pv(pv)^k v'p'\gamma', \delta'p'(pv)^k p\delta \mid \text{for } k \geq 2 \text{ and } k \in \mathbb{Z}^+\right\}.$$

Remember, $L_{2(\text{left pattern})}(S) = L_{2(\text{right pattern})}(S) = L_2(S)$. Since the splicing language is the same for both patterns, then the second order limit language for both patterns is also the same. Hence, the second order limit language, $L_2(S)$ exists. □

Using Proposition 1, Theorem 2 & 3, the next corollary is obtained.

Corollary 1 *If S is semi-simple and I has two crossing sites, then S is semi-null in which I has two crossing sites and forms second order limit language.*

Proof. From Theorem 2, since S is semi-simple and I has two crossing sites, thus S produces a second order limit language. By using Proposition 1, a semi-simple splicing system is indeed a semi-null splicing system. This, together with Theorem 3, implies that S is semi-null in which I has two crossing sites and produces second order limit language. □

Using Proposition 2, Theorem 1 and Theorem 3, the following corollary is obtained.

Corollary 2 *If S is simple and I has two crossing sites, then S is semi-null in which I has two crossing sites and forms second order limit language.*

Proof. From Theorem 1, since S is simple and I has two crossing sites, thus S produces a second order limit language. By using Proposition 2, a simple splicing system is indeed a semi-null splicing system. This, together with Theorem 3, implies that S is semi-null in which I has two crossing sites and produces second order limit language. □

The next theorem discusses the formation of the second order limit language in the uniform splicing system.

Theorem 4 *If S is uniform where I has two crossing sites and there is a positive integer P such that $X = A^P$, then a second order limit language exists.*

Proof. Let I be a set of initial strings that contains two crossing sites, $I = \{\alpha abxcdxab\beta\}$ and $R = \{r\}$ is a set of rules such that $r = (1; x, 1 : 1; x, 1)$ where a with b, c with d, α with α' and β with β' are complement to each other and $\beta, \alpha, a, b, x, c, d, \beta', \alpha' \in A^*$. There exists a positive integer, P such that $P = N$ where P indicates the length of the crossing site with length N. Therefore, the second order limit language, $L_2(S)$ is given below.

$$L_2(S) = \{\alpha ab(xcd)^* xab\alpha', \alpha ab(xcd)^* xab\beta, \beta' ab(xcd)^* xab\beta\}.$$

Therefore, the uniform splicing language produces second order limit language. □

Using Proposition 3, Theorem 1 & 4, the next corollary is obtained.

Corollary 3 *If S is simple and I has two crossing sites, then S is uniform in which I has two crossing sites and forms second order limit language.*

Proof. From Theorem 1, since S is simple and I has two crossing sites, thus S produces a second order limit language. By using Proposition 3, a simple splicing system is indeed a uniform splicing system. This, together with Theorem 4, implies that S is uniform in which I has two crossing sites and produces second order limit language. □

Next, some examples are discussed. Semi-null splicing system is inspired from the study related to the generation of language. Hence, a general example is given. Meanwhile, molecular example is provided to establish the formation of the second order limit language in the uniform splicing system as it is inspired from biological perspective. In the examples below, the initial strings and the dsDNA molecules with two crossing sites are chosen.

Example 1 *Suppose $I = \{\alpha wwumwwmvww\beta\}$ is a set of initial strings such that u with u', v with v' and m with w are complement to each other and $\gamma, \delta, m, w, \gamma', \delta' \in A^*$. Meanwhile, $R = \{r\}$ is a set of rules where $r = (v, 1, 1 : u, 1, 1)$ such that $v, u \in A^+$. The following second order limit language is obtained:*

$$L_2(S) = \left\{\gamma wwu\,(mwwmv)^k\, ww\delta, \gamma wwu\,(mwwmv)^k\, u'mm\gamma', \delta' mm\,(mppmv)^k\, pp\delta \mid k \geq 2 \text{ and } k \in \mathbb{Z}^+\right\}.$$

Example 2 *The initial dsDNA molecule is represented as $I = \{\alpha xxxxgatcxyxygatcyyyy\beta\}$ and the restriction enzyme namely FatI can be written as $R = \{r\}$ such that $r = (1; catg, 1 : 1; catg, 1)$ where β' with β and α' with α are complement to each other and $\alpha,\ \beta, x, y, \alpha', \beta' \in A^*$. A suitable buffer namely CutSmart^{TM} is added to the reaction for robust production. Besides, there is a positive integer, P such that $P = 4$ where P indicates the length of the crossing site with length 4. When the splicing process occurs again, the following second order limit language is given below.*

$$\begin{aligned} L_2(S) = \{&\alpha xxxx(catgxyxy \cup catgxyxy)^* catgxxxx\alpha', \\ &\beta' yyyy(catgxyxy \cup catgxyxy)^* catgyyyy\beta, \\ &\alpha xxxx(catgxyxy \cup catgxyxy)^* catgyyyy\beta\}. \end{aligned}$$

CONCLUSION

Some varieties of Y-G splicing system namely semi-null splicing system and uniform splicing system have been considered in this study. The formation of the second order limit language in these types of splicing system is explored and their relations are established. New hierarchies that focus on the formation of the second order limit language in some types of the splicing system are established such as simple splicing system, semi-simple splicing system ⊂ semi-null splicing system and simple splicing system ⊂ uniform splicing system. Two types of examples inspired from the language-based and molecular problems are given to give deep insights on the formation of the second order limit language in these types of the splicing system.

ACKNOWLEDGMENTS

All authors are indebted to Minister of Education (MOE) Malaysia for the financial funding through Fundamental Research Grant Scheme (FRGS) FRGS/1/2018/STG06/UMP/03/1, RDU190118.

REFERENCES

[1] D. Roy, *Biotechnology (Cytogenetics, Biotechnology and Bioinformatics)* (Alpha Science International Limited, Oxford, United Kingdom, 2010).

[2] S. S. Purohit, *Biotechnology Fundamentals and Applications* (AGROBIOS (INDIA), Jodhpur, India, 2005).

[3] J. D. Watson and F. H. C. Crick, Nature **171**, 737–738 (1953).

[4] M. Wink, *An Introduction to Molecular Biotechnology* (WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2006).

[5] T. Head, Bull. Math. Biol **49**, 737–759 (1987).

[6] A. Mateescu, G. Paun, G. Rozenberg, and A. Salomaa, Discrete Appl. Math **84**, 145–163 (1998).

[7] T. E. G. Laun, "Constants and splicing systems," Ph.D. thesis, State University of New York 1999.

[8] E. Goode and D. Pixton, "Semi-simple splicing systems," (Springer Netherlands, Berlin, 2001) Chap. Where Mathematics, Computer Science, Linguistic and Biology Meet,, pp. 343–352.

[9] W. H. Fong, "Modelling of splicing systems using formal language theory," Ph.D. thesis, Universiti Teknologi Malaysia 2008.

[10] Y. Yusof, N. H. Sarmin, T. E. Goode, M. Mahmud, and W. H. Fong, International Journal of Modern Physics: Conference Series **9**, 271–277 (2012).

[11] Y. Yusof, "Dna splicing system inspired by bio-molecular operations," Ph.D. thesis, Universiti Teknologi Malaysia 2011.

[12] M. A. Ahmad, N. H. Sarmin, W. H. Fong, and Y. Yusof, Jurnal Teknologi **71**, 13–15 (2015).