An Extension of DNA Splicing System

Yuhani Yusof, Nor Haniza Sarmin
Department of Mathematics,
Faculty of Science, Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
yuhani@ump.edu.my, nhs@utm.my

T. Elizabeth Goode
Department of Mathematics,
7800 York Road, Towson University,
Towson, Maryland, USA.
egoode@towson.edu

Mazri Mahmud
Emergency Department,
Hospital Tengku Ampuan Afzan,
Jalan Tanah Putih, 25100 Kuantan, Pahang, Malaysia.
ppp.mazri@yahoo.com

Fong Wan Heng
Ibnu Sina Institute for Fundamental Science Studies,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia.
fwh@utm.my

Abstract— **The first mathematical model of a splicing system that was analyzed in the framework of Formal Language Theory was developed in 1987 by Head. This model consists of a finite alphabet, a finite set of initial strings over the alphabet, and a finite set of rules that act upon the strings by iterated cutting and pasting, generating new strings. In this paper, a new notation for writing rules in a splicing system and a new extension of splicing systems is introduced in order to make the biological process transparent. These are called Yusof-Goode rules, and they are associated with Yusof-Goode splicing systems. Four different classes of splicing systems are discussed: null-context, uniform, simple and S_kH systems. Also, counterexamples are given to illustrate relationships between these splicing system classes.**

Keywords-DNA splicing systems; Formal Language Theory; DNA computing

I. Introduction

Every living organism has a unique deoxyribonucleic acid (DNA). The structure of DNA was firstly introduced in 1953 by Watson and Crick [1] in double-helical form. These structures of DNA are different from each other by the sequence of their bases namely: *Adenine* (*A*), *Guanine* (*G*), *Cytosine* (*C*) and *Thymine* (*T*). These bases are tied together by hydrogen bonds using base-complementary rules, where *A* pairs with *T*, *G* pairs with *C* and vice-versa [2]. These rules of pairing can simply be written as *a*, *g*, *c* and *t*, respectively.

In this research, a new extension of splicing system, called the Yusof-Goode splicing system (Y-G Model) is introduced. The idea of introducing this model is based on the characteristics of the restriction enzyme itself and in addition, it presents the transparent behavior of the biological process of DNA This model can be represented by $S = (A, I, R)$, where *A* is a set of alphabet *a*, *g*, *c* and *t*, *I* is an initial set of double-stranded DNA (dsDNA) and *R* is a set of rule that represents the existing restriction enzymes.

The rule *R* represents restriction enzymes that can cut the DNA molecules at specific places, resulting in molecules with sticky (5' or 3' overhang) or blunt ends based on their restriction sites. The 5' overhang and blunt-end are categorized as left-pattern rules while the 3' overhang is categorized as right-pattern rule. Hence, in Y-G Model, the rule $r \in R$ for left-pattern rule of restriction enzymes is presented as (*u*; *x*, *v*: *y*; *x*, *z*), whereas the rule for right-pattern restriction enzyme is presented as (*u*, *x*; *v*: *y*, *x*; *z*). Note that, *u*, *x*, *v*, *y* and *z* are strings over *A* and if the rule of Y-G Model is presented without semi-colon, i.e. (*u*, *x*, *v*: *y*, *x*, *z*), both left and right-patterns of restriction enzymes are applied. The resulting DNA molecules will religate with the existence of a ligase and non-dephosphorylated, for sticky and blunt-end respectively.

Note that, this Y-G model works as:
If $r = (u, x, v : y, x, z)$ *and* $s_1 = \alpha uxv\beta$ *and* $s_2 = \gamma yxz\delta$, *then a splice* s_1 and s_2 *using* r *produces* $\alpha uxz\delta$ *and* $\gamma yxv\beta$, *where* $\alpha, \beta, \gamma, \delta, u, x, v, y$ *and* $z \in A^*$, *the free monoid generated by A with the concatenation operation and* 1 *as the identity element.*

There are various types of splicing systems, including null-context, uniform, simple and S_kH system. The null-context and uniform splicing systems have been introduced by Head [3]. This paper shows that each null-context splicing system is persistent and if a language *L* is a persistent splicing language, *L* is also uniform. Mateescu et al. [4] introduced the notion of simple splicing systems in 1998. A decade after, some concepts involving simple splicing system using Formal Language Theory was done by Fong [5].

The sequence of language families S_kH was introduced in 1998 [6]. In 2008, Fong et al. [7] reduced the S_kH system to the simple splicing system using solid codes. Meanwhile in [8], Fong et al. introduced the concepts and examples of firmness and maximal firm subword (MFS) with their regular expressions and *SH*-automata applying on the reduction of splicing system.

978-0-7695-4514-1/11 $26.00 © 2011 IEEE
DOI 10.1109/BIC-TA.2011.67

This paper is structurally organized into four sections. The first section is the introduction, followed by Section II which includes some definitions used in this research. In Section III, some results and discussions are presented as theorems, corollaries and counterexamples. Finally, in the last section, the conclusion is given.

II. Preliminaries

The definition of splicing system will first be defined in this section.

Let A be defined as a fixed finite set to be used as an alphabet and A^* as a free monoid that consists of all strings of symbols in A, including the null string.

Definition 1: [3] **(Splicing System)**
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. □

Next, the definitions of four types of splicing system that will be discussed in this paper are presented.

Definition 2: [3] **(Null-Context Splicing System)**
A **null-context splicing system** is a splicing system $S = (A, I, B, C)$ for which each cleavage pattern in B and C has the form $(1, x, 1)$. □

Definition 3: [3] **(Uniform Splicing System)**
A **uniform splicing system** is a null-context splicing system $S = (A, I, X, X)$ for which there is a positive integer P such that $X = A^P$. □

Definition 4: [9] **(Simple Splicing System)**
Let $S = (A, I, R)$ be a splicing system in which all rules in R have the form $(a, 1; a, 1)$ where $a \in A$. Then S is called a **simple splicing system**. □

Definition 5: [6] **(S_k Splicing System, S_k Splicing Language)**
Let k be an integer ≥ -1 . An **S_k splicing system** (S_kH system) is a null-context splicing system $G = (A, I, R)$ for which, for each string r in R, length $r \leq k$. □

Note that when $k = 1$, the S_1H system is just the simple splicing system, denoted by SH.

III. Results and dicussions

In this section, the symmetric and reflexive of the rule $r \in R$ in Y-G Model are first presented.

Theorem 1
The rule $r \in R$ of Y-G splicing system is symmetric and reflexive. □

Proof
First, we show that $r \in R$ in Y-G splicing system is symmetric. By the preceding assumptions, the rule $r \in R$ can be presented as $(u, x, v$: $y, x, z)$ since both right and left patterns have to be considered.

Suppose that a finite initial set I of dsDNA contains two molecules s_1 and s_2 where $s_1 = \alpha uxv\beta$ and $s_2 = \gamma yxz\delta$ in either order. Thus, by applying Y-G Model, two new molecules will be generated besides the initial set I itself, namely $\alpha uxz\delta$ and $\gamma yxv\beta$. Note that, by Y-G model, the resulting strings can also be obtained by splicing the initial string I with the existing restriction enzymes in reverse order. Thus, $r \in R$ is symmetric.

Secondly, we show that $r \in R$ is reflexive. From previous assumptions, $r \in R$ can be presented as $(u, x, v$: $y, x, z)$ since both patterns apply to the initial string before ligation. As stated before, this system generates the initial strings, I which enable the pairing of rules with themselves. Hence, $r \in R$ is reflexive. ■

Next, some relations on different types of splicing system are analyzed and are presented as theorems and corollaries. Besides, some counterexamples are given to illustrate these relations. Since Y-G model is being used, the rules R will be presented in double-triple notation. Hence, the rule R of simple splicing system in Definition 4 can be rewritten as $R = (a,1,1:a,1,1)/(1,1,a:1,1,a)/(1,a,1:1,a,1)$, where $a \in A$ [10].

Theorem 2
Every simple splicing system S is a uniform splicing system where $S = (A, I, X, X)$. □

Proof
Suppose that t is not an element of a uniform splicing system for which each crossing site in X has the form of $(1, x, 1)$, with $X = A^P$. Thus, t is not an element of a simple splicing system since A is a subset of A^P. ■

However, there exists a uniform splicing system that is not simple as illustrated in Example 1 below.

Example 1
Let $S = (\{a, g, c, t\}, I(\text{unspecified}), \{1, catg; 1:1, acgt; 1\})$. The right pattern consists of two restriction enzymes namely, *Nla*III and *Tsc*I, with the crossing sites as follows:
Crossing site for the enzyme *Nla*III:

5'... $CATG^{\blacktriangledown}$...3'
3'...$_{\blacktriangle}GTAC$ ...5'.

Crossing site for the enzyme *Tsc*I:

5'... $ACGT^{\blacktriangledown}$...3'
3'...$_{\blacktriangle}TGCA$ ...5'.

Thus, S is a uniform splicing system since both restriction enzymes, *Nla*III and *Tsc*I, have the same length of crossing with $P = 4$. However, S is not a simple splicing system since the crossing sites of *Nla*III and *Tsc*I are *catg* and *acgt* respectively, which are two different elements that are not in A. ■

In the next theorem, the relation between uniform splicing system and S_kH system is presented.

Theorem 3
Every uniform splicing system is an S_kH system G where $G = (A, I, R)$. □

Proof
Assume that t is not an element of an S_kH system for which each crossing site in R has the form of $(1, x, 1)$ and $r \in R$, length $r \leq k$ for $k \geq -1$. Thus, t is not an element of a uniform splicing system since t itself is not an element of a null-context splicing system. ■

However, there exists an S_kH system that is not uniform as presented in the following counterexample.

Example 2
Let $S = (\{a, g, c, t\}, I(\text{unspecified}), \{1; gatc, 1:1; gtsac, 1\})$ be a splicing system where $s = c$ or g. The left pattern consists of two restriction enzymes, namely *Dpn*II and *Tsp*45I with the crossing sites as follows:
Crossing site for the enzyme *Dpn*II:

5'...▼*GATC* ...3'
3'... *CTAG*▲...5' .

Crossing site for the enzyme *Tsp*45I:

5'...▼*GTSAC* ...3'
3'... *CASTG*▲...5' .

Thus, S is an S_kH splicing system with $k = 5$ since five is the longest crossing site for this splicing system. However, S is not a uniform splicing system since restriction enzymes *Dpn*II and *Tsp*45I have two different lengths of crossings r, which is four and five, respectively. ■

Theorems 2 and 3 lead to Corollary 1.

Corollary 1
Every simple splicing system is an S_kH system. ■

By Definition 5, every S_kH system is a null-context splicing system. However, there exists a null-context splicing system that is not an S_kH system as given in Example 3 below.

Example 3
Let $S = (\{a, g, c, t\}, I(\text{unspecified}), \{1; gatc, 1:1; gatc, 1\})$ be a splicing system. The left pattern consists of two restriction enzymes**,** namely *Mbo*I and *Sau*3AI, with the crossing sites as follows:
Crossing site for the enzyme *Mbo*I:

5'...▼*GATC* ...3'
3'... *CTAG*▲...5' .

Crossing site for the enzyme *Sau*3AI:

5'...▼*GATC* ...3'
3'... *CTAG*▲...5' .

Thus, S is a null-context splicing system since both crossing for restriction enzymes *Mbo*I and *Sau*3AI, are in the form of $(1, x, 1)$. However, S is not an S_kH system for $k = -1, 0, n$ where n are elements of $\{\mathbb{Z}^+ \setminus 4\}$ since the longest length of crossing k for this splicing system is four. ■

Corollary 1 and Definition 5 lead to the following corollary.

Corollary 2
Every simple splicing system is null-context splicing system.■

IV. Conclusion

In this paper, the symmetric and reflexive rules $r \in R$ in Y-G Model are presented. In addition, some analysis on various types of splicing systems, namely null-context, uniform, simple and S_kH splicing systems are done and presented as Theorem 2, 3, 4, Corollaries 1, 2 and Example 1, 2, 3. These relations can be simplified as follows:

simple splicing system $\subset$ uniform splicing system
$\subset$ S_kH splicing system $\subset$ null-context splicing system.

Acknowledgment

We would like to thank the Ministry of Higher Education (MOHE) and RMC UTM for the GUP Vote No. 02J65. The first author would also like to acknowledge Universiti Malaysia PAHANG (UMP) for her study leave, JP&I, UMP for funding on BIC-TA 2011 registration and accommodation fees and to MOHE for her SLAI scholarship.

References

[1] R. H. Tamarin, Principle of Genetics. USA: The McGraw-Hill Companies, 2001.

[2] Research Biolabs Sdn. Bhd., New England Biolabs 2007-08 Catalog & Technical Reference, USA, 2007.

[3] T. Head, “Formal Language Theory and DNA: An Analysis of the Generative Capacity Specific Recombinant Behaviors,” Bull. Math. Biology, vol. 49, 1987, pp. 737-759.

[4] A. Mateescu, Gh. Paun, G. Rozenberg and A. Salomaa, “Simple Splicing Systems,” Discrete Applied Mathematics, vol. 84, 1998, pp. 145-163.

[5] W. H. Fong, “Modelling of Splicing Systems Using Formal Language Theory,” PhD Thesis, Universiti Teknologi Malaysia, 2008.

[6] T. Head, “Splicing Representations of Strictly Locally Testable Languages,” Discrete Applied Mathematics, vol. 87, 1998, pp. 139-147.

[7] W. H. Fong, N. H. Sarmin and Z. Ibrahim, “Reduction of Splicing Systems using Solid Codes,” Proceeding of the 16th Mathematical Sciences National Symposium, 2008, pp. 37-41.

[8] W. H. Fong, N. H. Sarmin and Z. Ibrahim, “Recognition of Simple Splicing Systems using *SH*-Automaton,” Journal of Fundamental Sciences, vol. 4(2), 2008, pp. 337- 342**.**

[9] E. G. Laun, “Constants and Splicing Systems,” Ph.D Thesis, State University of New York at Binghamton, 1999.

[10] Y. Yusof, N. H. Sarmin, T. E. Goode, M. Mahmud and W. H. Fong, “Hierarchy of Certain Types of DNA Splicing Systems,” International Conference on Mathematical and Computational Biology 2011. 2011.