KEMENTERIAN PENDIDIKAN MALAYSIA

UNIVERSITI MALAYSIA TERENGGANU

PROCEEDINGS

The 10th IMT-GT International Conference on Mathematics, Statistics and its Application (ICMSA) 2014

Knowing Nature Through Mathematical Science

October 14th - 16th, 2014
Kuala Terengganu

Prediction of Single Stage Yusof-Goode Splicing Language involving Two Strings

Lim Wen Li[a], Yuhani Yusof[a] & Mohammad Hassan Mudaber[a]

[a]*Faculty of Industrial Sciences and Technology,*
Universiti Malaysia Pahang, 26300 UMP Gambang, Pahang

Abstract. Recombinant deoxyribonucleic acid (DNA) technologies play a pivotal role in attempts to recombine sets of double-stranded DNA molecules when acted on restriction enzymes and a ligase. It is money and time consuming if to conduct a laboratory experiment in predicting the number of resulted molecules. Hence, Yusof-Goode (Y-G) Splicing System, a formal Mathematics characterization of the generative capacity of specified enzymatic activities acting on DNA molecules with new symbolization of representing rule is used to formulate theorems in this paper. In addition, the number pattern of single stage splicing languages with respect to two strings and at most two rules with one cutting site are predicted with biological examples.

Keywords: Y-G Splicing System; Single Stage Splicing Languages; Strings
PACS: 87.14.gk, 87.14.ej

INTRODUCTION

Recombinant double stranded deoxyribonucleic acid (DNA) molecules are a sequence of molecules that does not exists in nature which has been created by laboratory methods [1]. Through gene manipulation, human can appreciate the result of higher yield crop, a lifestyle free of illness and the possibilities of living longer. In 1987 [2], Head has introduced the formalism of splicing system that describes the recombination behavior of DNA molecules via Formal Language Theory [3]. New extension in splicing system, Yusof-Goode (Y-G) splicing system, was presented in [4] to illustrate the translucent behavior of the biological process in splicing DNA. To date, many laboratory experiments have been conducted to verify the model of splicing system in either one stage or two stages. Only single stage splicing languages are considered here in this paper since letting all the restriction enzymes, double stranded deoxyribonucleic acid strings and ligases act simultaneously in a single buffer can optimize the time and money during experiment with minimum difference in the resulted strings generated. In this paper, the production of Y-G splicing system namely splicing languages involving two strings with one cutting site are predicted based on three theorems formulated.

PRELIMINARIES

In this section, the fundamental definitions of this paper are given. First and foremost, the definition of Y-G splicing system that will be used all through this paper is reviewed.

Definition 1 [5]: Y-G Splicing System. If $r \in R$, where $r = (u,x,v:y,x,z)$ and $s_1 = \alpha uxv\beta$ and $s_2 = \gamma yxz\delta$ are elements of I, then splicing s_1 and s_2 using r produces the initial string I together with $\alpha uxz\beta$ and $\gamma yxz\delta$, presented in either order where $\alpha,\beta,\gamma,\delta,u,x,v,y$ and $z \in A^*$ are the free monoid generated by A with the concatenation operation and 1 as the identity element.

The set of molecules generated during the evolution of the splicing system are known as splicing language. Next, the existing definition of splicing language introduced by Head [2] is further defined as single stage splicing language to model the set of all molecule types that exists in a test-tube environment with restriction enzymes and ligases all act simultaneously.

Definition 2: Single stage splicing language is defined as

$$[L_1 = L_1(S)] \cong \sum_{r=1}^{n} (R_r + I_r + l) \qquad (1)$$

R_r = set of rules, $1 \le r \le n$

I_r = set of initial strings, $1 \le r \le n$

l = ligases

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 the restriction enzymes, double stranded deoxyribonucleic acid strings and ligases act simultaneously in a single buffer.

Next, the definitions of the main characteristics of restriction enzymes, palindromic, are stated.

Definition 3 [3]: Palindromic Rule. A string I of double stranded DNA (dsDNA) is said to be palindromic if the sequence from the left side of the upper single strand is equal with the sequence from the right side of the lower single strand.

Besides palindromic rule, the inverse complement is a common behaviour of restriction enzyme as well, where the definition is presented as below.

Definition 4: Inverse Complement

A string x is an inverse complement to another string y if $x = \overleftarrow{y'}$, where $x, y \in A^*$. Two strings I_1 and I_2 of dsDNA are said to be **inverse complement** to each other if the sequence from the left side of the upper single strand in I_1 is equal to the lower single strand with the sequence from the right side in I_2.

Based on the characteristics of the restriction enzyme above, in the next section, three theorems are formulated in order to predict the number pattern of single stage splicing languages generated using Y-G splicing system.

THE NUMBER PATTERNS OF SINGLE STAGE SPLICING LANGUAGES INVOLVING TWO STRINGS

In this section, four theorems are presented to discuss the number of single stage splicing languages generated involving two strings. The first and third theorem proved the number pattern of single stage splicing languages generated when the rule(s) are palindromic, while the second and forth theorems proved the production of single stage splicing language based on non palindromic rules. In biological point of view, the theorems are formulated to predict the number set of molecules that will present in the system during the biochemical reaction.

Theorem 1. If the crossing sites of rule(s) in a two strings splicing system with one recognition site each is itself (are themselves) same and palindromic, then $n[L_1(S)] = \{10\}$. □

Proof: Suppose $S = (A, I, R)$ is a two strings Y-G splicing system, $I = \{I_1, I_2\}$ and the crossing sites of rule(s), $r \in R$ is same and palindromic. Assume that each string contains one recognition site and there will be at most two rules. Hence, three cases need to be considered:

I. One rule
II. Two rules
III. $I_1 \subset I_2$

Case I. Let us consider a case where $I_1 = \alpha abaa\beta, I_2 = \gamma abaa\delta$ with a given rule $r = (a, ba, b)$ for $a, b, \alpha, \beta, \gamma, \delta \in A^*$ where a and b are complement of each other. The rule is then acting on both

strings producing four fragments with sticky ends. Due to the palindromic properties, all the fragments can relegate among each other. The number of splicing languages that can be formed by ligating any two fragments is ${}^4C_2 = 6$. Besides, the sticky-ends have a chance to religate with the 180 degree rotation of itself to generate another four splicing languages. Thus, there will be ten single stage splicing languages in total. The splicing occurs as below,

$$\alpha abab\beta, \gamma abab\delta \xrightarrow{r} \{\alpha abab\beta, \gamma abab\delta, \alpha abab\delta, \gamma abab\beta, \alpha abab\gamma',$$
$$\beta' abab\delta, \alpha ababa\alpha', \gamma abab\gamma', \beta' abab\beta, \delta' abab\delta\}$$

Case II. Let us consider a case where $I_1 = \alpha babb\beta, I_2 = \gamma cabc\delta$ with a given rule $r = (b, ab, b:c, ab, c)$ for $a, b, c, \alpha, \beta, \gamma, \delta \in A^*$ where a and b are complement of each other. These rules are then acting on both strings producing four fragments with sticky ends. Due to the palindromic properties, all the fragments can relegate among each other. The number of splicing languages that can be formed by ligating any two fragments is ${}^4C_2 = 6$. Besides, the sticky-ends have a chance to religate with the 180 degree rotation of itself to generate another four splicing languages. Thus, there will be ten single stage splicing languages in total. The splicing occurs as below,

$$\alpha bbab\beta, \gamma cbac\delta \xrightarrow{r} \{\alpha bbab\beta, \gamma cbac\delta, \alpha abac\delta, \gamma cbab\beta, \alpha bbac'\gamma',$$
$$\beta' abac\delta, \alpha bbaa\alpha', \gamma cbac'\gamma', \beta' abab\beta, \delta' c' bac\delta\}$$

Case III. Let us consider a case where $I_1 = \alpha babb\beta, I_2 = \gamma\alpha babb\beta\delta$ with a given rule $r = (b, ab, b:b, ab, b)$ for $a, b, \alpha, \beta, \gamma, \delta \in A^*$ where a and b are complement of each other. Note that the first string is the subset of second string. These rules are then acting on both strings producing four fragments with sticky ends. Due to the palindromic properties, all the fragments can relegate among each other. The number of splicing languages that can be formed by ligating any two fragments is ${}^4C_2 = 6$. Besides, the sticky-ends have a chance to religate with the 180 degree rotation of itself to generate another four splicing languages. Thus, there will be ten single stage splicing languages in total. The splicing occurs as below,

$$\alpha bbab\beta, \gamma\alpha bbab\beta\delta \xrightarrow{r} \{\alpha bbab\beta, \gamma\alpha bbab\beta\delta, \alpha bbab\beta\delta, \gamma\alpha bbab\beta, \alpha bbaa\alpha'\gamma',$$
$$\beta' abab\beta\delta, \alpha bbaa\alpha', \gamma\alpha bbaa\alpha'\gamma', \beta' abab\beta, \delta'\beta' abab\beta\delta\}$$

The above three cases lead to the desired results. ■

The second and third theorem proves the importance of inverse complement in non palindromic rules to determine the number of splicing languages.

Theorem 2: Let $S = (A, I, R)$ be a Y-G splicing system such that $I = \{I_1, I_2\}$ and $R = \{(r_1, r_2), (r_1)\}$. If elements of R have the same or inverse complement in non palindromic crossing with one recognition site in each string, then $n[L_1(S)] = \{4\}$. □

Proof: Let *S= (A, I, R)* be a Y-G splicing system. Suppose I_1 and I_2 are two strings that contains only one cutting site can be spliced by using non palindromic rules. The rule(s) must fulfil the conditions below:

i. Two same crossing site.
ii. Two Inverse complement crossing site
iii. One rule

Condition i: Assume that Y-G splicing system contains two non palindromic rules with same crossing sites. Since two different strings are involved, two cases need to be considered:

Case I: $I_1 \neq I_2$

Assume that $(u,v,w:r,v,t) \in R$ and $I_1 = \alpha uvw\beta$ and $I_2 = \gamma rvt\delta$ for some I_1 and $I_2 \in A^*$. Applying the rules to I_1 and I_2 in S, four splicing languages are obtained as follow:

$$(\alpha uvw\beta, \gamma rvt\delta) \mapsto_{r_1,r_2} \{\alpha uvw\beta, \gamma rvt\delta, \alpha uvt\delta, \gamma rvw\beta\}$$

Case II: $I_1 \subset I_2$

Assume that $(u,v,w:u,v,w) \in R$ and $I_1 = \alpha uvw\beta, I_2 = \gamma\alpha uvw\beta\delta$ for I_1 and $I_2 \in A^*$.Applying the rules to I_1 and I_2 in S, four splicing languages are obtained as follow:

$$(\alpha uvw\beta, \gamma\alpha uvw\beta\delta) \mapsto_{r_1,r_2} \{\alpha uvw\beta, \gamma\alpha uvw\beta\delta, \alpha uvw\beta\delta, \gamma\alpha uvw\beta\}$$

Condition ii: Assume that Y-G spli cing system contains two non palindromic rules with inverse co mplement crossing sites. Since two different strings are involved, two cases need to be considered.

Case I: $I_1 \neq I_2$

Assume that $(u,v,w:r,s,t) \in R$ where the crossing sites are inverse complement to each other, $v' = s$. That means $I_1 = \alpha uvw\beta$ and $I_2 = \gamma rst\delta$ for some I_1 and $I_2 \in A^*$. Applying the rules to I_1 and I_2 in S, four splicing languages are obtained as follow:

$$(\alpha uvw\beta, \gamma rst\delta) \mapsto_{r_1,r_2} \{\alpha uvw\beta, \gamma rst\delta, \alpha uvr'\gamma', \beta'w'st\delta\}$$

Case II: $I_1 \subset I_2$

Assume that $(u,v,w:r,s,t) \in R$ and $I_1 = \alpha uvw\beta, I_2 = \gamma\alpha uvw\beta rst\delta$ for I_1 and $I_2 \in A^*$. Note that I_2 contains recognition sites from both rules, which is contradicting with the one recognition site condition. Hence, this case cannot be considered.

Condition iii: Assume that Y-G splicing system contains one non palindromic rule, $(u,v,w) \in R$. Since two different strings are involved, two cases need to be considered.

Case I: $I_1 \neq I_2$

Assume there exist $I_1 = \alpha uvw\beta$ and $I_2 = \gamma uvw\delta$ for some I_1 and $I_2 \in A^*$. Applying the rules to I_1 and I_2 in S, four splicing languages are obtained as follow:

$$(\alpha uvw\beta, \gamma uvw\delta) \mapsto_{r_1,r_2} \{\alpha uvw\beta, \gamma uvw\delta, \alpha uvw\delta, \gamma uvw\beta\}$$

Case II: $I_1 \subset I_2$

Assume there exist $I_1 = \alpha uvw\beta$ and $I_2 = \gamma\alpha uvw\beta\delta$ for some I_1 and $I_2 \in A^*$. Applying the rules to I_1 and I_2 in S, four splicing languages are obtained as follow:

$$(\alpha uvw\beta, \gamma\alpha uvw\beta\delta) \mapsto_{r_1,r_2} \{\alpha uvw\beta, \gamma\alpha uvw\beta\delta, \alpha uvw\beta\delta, \gamma\alpha uvw\beta\}$$

All of the above cases lead to desired result. ■

Theorem 3: If two initial strings and two different palindromic crossing site rule is involved in a splicing system, then there exist $n[L_1(S)] = \{6\}$.□

Proof: Suppose $S = (A, I, R)$ is a two strings Y-G splicing system and the crossing sites of two rules, $r \in R$ are different and palindromic. Assume that each string contains one recognition site. Consider $I_1 = \alpha abaa\beta, I_2 = \gamma aaba\delta$ with a given rule $r = (a, ba, b : a, ab, b)$ for $a, b \in A$ where a and b are complement of each other. The rule is then acting on both strings producing four fragments with sticky ends. Since the crossing sites of two rules are different, the strings I_1 and I_2 themselves are obtained. Besides, the sticky-ends have a chance t o

religate with the 180 degree rotation of itself to generate another four splicing languages due to the palindromic properties. The splicing occurs as below,

$$\alpha abab\beta, \gamma aabb\delta \xrightarrow{r} \{\alpha abab\beta, \gamma aabb\delta, \alpha aabb\gamma',$$
$$\alpha abab\alpha', \gamma aabb\gamma', \beta' abab\beta, \delta' aabb\delta\}$$

Thus, there will be six single stage splicing languages in total.■

Theorem 4: If two initial strings and two non inverse complement crossing site non palindromic rule is involved in a splicing system, then $n[L_1(S)] = \{2\}$.□

Proof: Suppose m_1 and m_2 are again two strings that can be spliced by using non palindromic rules with different crossing sites, $(u,v,w:r,s,t) \in R$ where the crossing sites are non inverse complement to each other, $v' \neq s$. That means $m_1 = \alpha uvw\beta$ and $m_2 = \gamma rst\delta$ for some m_1 and $m_2 \in A^*$. Applying the rules to m_1 and m_2 in *S*, the languages are obtained as follow:

$$(\alpha uvw\beta, \gamma rst\delta) \mapsto_{r_1, r_2} I$$

Two single stage splicing languages, which are the strings I_1 and I_2 themselves, are obtained. ■

EXAMPLES OF SINGLE STAGE SPLICING LANGUAGE IN TWO STRINGS SPLICING SYSTEM

In this section, two examples of single stage splicing language in two strings are provided. The first example shows that ten splicing languages are produced with the presence of two initial strand of double stranded DNA (dsDNA) and two same palindromic crossing site restriction enzymes in which both rules are being cut at the single stage. The second example shows a splicing system which consists of two initial strand of dsDNA and a restriction enzyme.

In the first example, two initial dsDNA and two restriction enzymes from actual DNA process are chosen to explain the theorem one. The restriction enzymes are selected from [6].

Example 1. Let $S = (A, I, R)$ be a Y-G splicing system consisting of two restriction enzymes namely *Mbo*I and *Aci*I, where $A = \{a, c, g, t\}$, $I = \{\alpha ccgc\beta, \gamma gatc\delta\}$ such that $\alpha, \beta \in A^*$, and $R = \{(r_1 : r_2)\}$ where $r_1 = (1; gatc, 1)$ and $r_2 = (c; cg, c)$. Let PCR generates thousands copies of $\alpha - \beta$ and $\gamma - \delta$ strands, *I*. One Taq GC Reaction Buffer and OneTaq Standard Reaction Buffer are chosen to be used as the reaction buffer for DNA strands based on the percentage of G and C bases in the forward and reverse primer. By theorem 3, six splicing languages should be generated. By calculation, when splicing occurs, the following splicing languages are generated:

$$L_1 = \begin{matrix} \alpha\ GATC\beta \\ \alpha' CTAG\beta' \end{matrix}$$
$$L_2 = \begin{matrix} \alpha\ GATC\alpha' \\ \alpha' CTAG\alpha \end{matrix}$$
$$L_3 = \begin{matrix} \beta'\ GATC\beta \\ \beta CTAG\beta' \end{matrix}$$
$$L_4 = \begin{matrix} \gamma CCGC\delta \\ \gamma' GGCG\delta' \end{matrix}$$
$$L_5 = \begin{matrix} \gamma CCGG\gamma' \\ \gamma' GGCC\gamma \end{matrix}$$
$$L_6 = \begin{matrix} \delta' GCGC\delta \\ \delta CGCG\delta' \end{matrix}$$

Six types of resulted molecules are produced, as stated in the theorem 3.

In the next example, a splicing system consists of two initial strand of dsDNA and a restriction enzyme is illustrated.

Example 2. Let $S=(A,I,R)$ be a Y-G splicing system consisting of a restriction enzyme namely *Hpy*99I, where $A=(a,c,g,t)$, $I=\{\alpha cgwcg\beta,\gamma cgwcg\delta\}$ which each consists of one recognition site of the restriction enzyme and $R=(1,cgwcg;1)$. The nucleotide represented by w can be a or t, $W=(a,t)$. By Theorem 2, four splicing languages should be generated. When splicing occurs in the presence of ligase and NEBuffer4, the following splicing languages are generated:

For the case $w=a$,

$$L_1=\begin{matrix}\alpha\ \text{CGACG}\beta\\ \alpha' GCTGC\beta'\end{matrix}$$
$$L_2=\begin{matrix}\gamma\ \text{CGACG}\delta\\ \gamma' GCTGC\delta'\end{matrix}$$
$$L_3=\begin{matrix}\alpha\ \text{CGACG}\delta\\ \alpha' GCTGC\delta'\end{matrix}$$
$$L_4=\begin{matrix}\gamma\ \text{CGACG}\beta\\ \gamma' GCTGC\beta'\end{matrix}$$

For the case $w=t$,

$$L_1=\begin{matrix}\alpha\ \text{CGTCG}\beta\\ \alpha' GCAGC\beta'\end{matrix}$$
$$L_2=\begin{matrix}\gamma\ \text{CGTCG}\delta\\ \gamma' GCAGC\delta'\end{matrix}$$
$$L_3=\begin{matrix}\alpha\ \text{CGTCG}\delta\\ \alpha' GCAGC\delta'\end{matrix}$$
$$L_4=\begin{matrix}\gamma\ \text{CGTCG}\beta\\ \gamma' GCAGC\beta'\end{matrix}$$

Four types of splicing languages are produced for both cases, as stated in theorem 2.

CONCLUSION

Four theorems are presented to predict the number pattern of single stage Y-G splicing languages. This research can be extended to predict the outcomes of mixing unbounded DNA string with unbounded number of restriction enzymes for unbounded stages in order to reduce the number of experiment conducted.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Ministry of Education (MOE) and Research and Innovation Department, Universiti Malaysia Pahang (UMP) for the financial funding through UMP Research Grant Vote No: RDU 130354 and RAGS Grant Vote No: RDU 131404.

REFERENCES

1. Campbell, Neil A. & Reece, Jane B., *Biology(6th Edition)*, San Francisco: Addison Wesley. 2002
2. T. Head, *Bulletin of Mathematical Biology* **49**, 737-759 (1987).
3. P. Linz, *An Introduction to Formal Languages and Automata*, USA: Jones and Barlett Publisher, 2006.
4. Y. Yusof, N. H. Sarmin, M. Mahmud, T. E. Goode and W. H. Fong, "An Extension of DNA Splicing System" in 6th International Conference on Bio-Inspired Computing: Theories and Applications, Pulau Pinang, Malaysia, 2011, pp. 246-248.
5. Y. Yusof, "DNA Sp licing System Inspired by Bio Molecular Operations", Ph.D. T hesis, Universiti Teknologi Malaysia, 2011.
6. Research Biolabs Sdn. Bhd. *New England Biolabs 2011-12 Catalogue & Technical Reference*. USA: Catalogue. 2011.