Modeling Limit Languages via Limit Adjacency Matrix and Yusof-Goode Approaches

Wen Li, Lim[a] and Yuhani Yusof[a]

[a]*Faculty of Industrial Sciences and Technology*
Universiti Malaysia Pahang, Tun Razak Highway, 26300 Kuantan, Pahang, MALAYSIA

Abstract. Limit language was introduced by Goode and Pixton in 2004 under the framework of formal language theory. It is a subset of splicing languages which is restricted to the molecules that will be presented in the splicing system after the reaction of biochemical has run to its completion. In this paper, limit adjacency matrix will be introduced to model the existence of limit languages from splicing languages. Besides, it can be used to characterize the splicing language in terms of active persistent, adult/inert and transient properties based on Yusof-Goode splicing system. In this paper, some examples and theorems that have been formulated via limit adjacency matrix approach will be presented too.

Keywords: limit language, splicing language, limit adjacency matrix
PACS: 87.14.gk, 87.14.ej

INTRODUCTION

The mathematical modeling of splicing system was initiated by Head [1] to model the enzymatic activities acting on deoxyribonucleic acid (DNA) molecules and that led to the existence of Yusof-Goode (Y-G) splicing system [2], a splicing model that is based on the characteristics of the restriction enzyme itself and also presents the transparent behavior of the DNA biological process. Splicing language is the language defined by the splicing system. In terms of the molecular biological aspect, it is a set of molecular types that are produced during the process of DNA splicing. Next, Goode and Pixton [3] defined limit language as a subset of splicing language, which is a set of molecules that are left after the completion of evolution of the system. The concept of limit graph, G_L^∞, was then introduced in [3] and this concept is further extended by Yusof [2] with theorems, examples and laboratory experiments in order to identify the limit languages. However, it is observed that limit adjacency matrices seem to be a more suitable tool for describing the complex DNA recombination process and determining the limit languages. Hence, in this paper, limit adjacency matrix is introduced and some theorems related to the concepts of limit adjacency matrix are presented. It is shown with some examples that the splicing operation can be simulated by limit adjacency matrix, which justifies the fitness of adjacency matrix in modeling the existence of limit languages.

PRELIMINARIES

In this section, some fundamental definitions used in this paper are viewed. The definitions of Y-G splicing system and splicing language are stated below:

Definition 1 [2]: (Y-G splicing system, splicing language)

A **Y-G splicing system** $S = (A, I, R)$ consists of a set of alphabet A (*a, g, c* and *t*), an initial set I of double stranded DNA over A and a set R of rules that represents the existing restriction enzymes. If $r \in R$, where $r = (u, x, v : y, x, z)$ and $s_1 = \alpha uxv\beta$ and $s_2 = \gamma uxv\delta$ are elements of I, then 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 $L = L(S)$.

Next, the definition of limit language is given.

Definition 2 [3]: Limit Language

Limit language is defined as $L_\infty = \bigcap_{k=1}^{\infty} L_k$.

Also, the definition of adult language is described below.

Definition 3 [5]: Adult/Inert Language

Adult strings, also called inert strings in a splicing language, are strings in a splicing system which cannot be used for splicing. Adult molecules show a steady increase in quantity throughout the reaction, and are not involved in further interactions with other molecules or enzymes. These molecules therefore lie in the **adult language**, which is denoted as L_A

Next, the definition of active persistent language is presented.

Definition 4 [2]: Active Persistent Language

An **active persistent language** is a set of strings that participate in further splicing and is also contained in the limit language, L_∞.

The definition of transient language is provided next.

Definition 5 [3]: Transient Language

A word w which is not a first-order limit is called **transient** in L; in other words, w is transient in L if and only if there is a word z in L so that $w \to {}^{+}_{L} z$ but $z \to {}^{*}_{L} w$ is false.

The definition of non-trivial splicing is given below.

Definition 6 [2]: Non-trivial Splicing

A non-trivial splicing is a splicing where there exists a string w in a language L, for all v_i, v_j elements of L, such that $(v_i, w) \to^{+}_{L} v_j$ and $v_i \neq v_j$ with w, v_i, v_j strings in the language L and $\to^{+}_{L}$ .is a transitive closure of $\to_L$.

Lastly, the definition of limit graph is given.

Definition 7 [3]: Limit Graph

A **limit graph**, G_L^∞, is a sequence of graphs obtained by the splicing process. The vertices and lines in G_L^∞ are represented for words or strings of double-stranded DNA and 'produced strings', respectively. The line exists between vertices of v_i and v_j if there is a word v_i in language L, such that by splicing v_j with another word v_k in L, produces the word $v_j \in L$, where v_j can be a different word or the same as v_i.

LIMIT ADJACENCY MATRIX, A_{ij}^{∞}

In this section, the definition of limit adjacency matrix is presented.

Definition 8: Limit Adjacency Matrix

A limit adjacency matrix, A_{ij}^{∞}, is a binary matrix to describe the splicing process. A Y-G splicing system $S=(A,I,R)$ consists of a finite alphabet A, a finite set I of strings in A^*, where A^* is denoted by the free monoid over A [1]. The limit adjacency matrix, A_{ij}^{∞} over A of size $p \times p$ has a square arrangement of positions (i, j) with the entries of the matrix representing the number of 'produced strings' from words or strings of double stranded deoxyribonucleic acid (dsDNA). The values of the elements in A_{ij}^{∞} are the number of words $w_j \in L$ that are produced by splicing a word w_i with another word w_k in L, where w_j can be a different word or the same as w_i. The column headers of limit adjacency matrix, A_{ij}^{∞} are denoted as w_j while the row numbers of limit adjacency matrix are denoted as w_i.

In the following subsection, some theorems regarding to limit adjacency matrix are given.

Some Theorems in Limit Adjacency Matrix, A_{ij}^{∞}

In this subsection some theorems in limit adjacency matrix which are applicable to the behavior of single stage splicing languages are presented. The first theorem presents the characteristics of the limit adjacency matrix which involves single stage limit language.

Theorem 1: The row number of A_{ij}^{∞}, w_i, is an adult/inert language if and only if every elements in a row of A_{ij}^{∞} is zero.

Proof:

Let A_{ij}^{∞} be a limit adjacency matrix. By contradiction, assume that at least one of the elements in a_j is 1. By definition of A_{ij}^{∞}, there exists $(w_i, w_k) \xrightarrow{r} w_j$, where $w_j, w_i, w_k \in L$. The situation is as follows:

Case 1: $w_j = w_i$

String w_i is spliced to itself by rule r to regenerate itself.

Case 2: $w_j \neq w_i$

String w_i is spliced by rule r to produce new string w_j.

Both cases above imply that string w_i can participate in further splicing, and thus, not in adult/inert language. Hence, the supposition is false and the 'if part' is proved. The other part of the proof is obtained by retracing the above steps. ■

Theorem 2: In a limit adjacency matrix of a splicing system, the languages w_i or w_j are active persistent when $\sum_r^n a_{i,r} \leq \sum_r^n a_{r,j}$, where $\sum_r^n a_{i,r} \neq 0$

Proof:

Suppose w_j are produced strings along with w_k, w_i words of double-stranded DNA. In limit adjacency matrix, column header are denoted as w_j, and hence, the sum of column, $\sum_r^n a_{r,j}$ in A_{ij}^{∞} implies the total number pattern of strings w_j that are produced by splicing a word w_i with another word w_k in L. Meanwhile, the sum of row

$\sum_r^n a_{i,r}$ represents the total number pattern of strings that are generated by w_i. If $\sum_r^n a_{i,r} \le \sum_r^n a_{r,j}$, the total number pattern of strings produced, w_j, is more than the total number pattern of strings w_j, that are generated. Consequently, none of the strings are adult/inert, (since from Theorem 1, $\sum_r^n a_{i,r} \neq 0$ implies non-adult/inert language), nor do any of them vanish. Hence, each languages w_i or w_j in L_∞ is in a reactive steady-state at equilibrium, which is named as active persistent language, the desired results. ■

Corollary 1: In a limit adjacency matrix of a splicing system, the languages w_i or w_j are transient when $\sum_r^n a_{i,r} > \sum_r^n a_{r,j}$.

Theorem 3: A non-null, symmetric limit adjacency matrix where $\sum_r^n a_{i,r} \neq 0$, contains all active persistent languages.

Proof:
Since limit adjacency matrix is a binary matrix, by properties of symmetric matrix, $A_{ij}^{\infty} = A_{ij}^{\infty T}$, thus, $\sum_r^n a_{i,r} = \sum_r^n a_{r,j}$. From Theorem 2, it is evident that strings w_i are all active persistent languages. ■

Theorem 4: If A_{ij}^{∞} is a limit adjacency matrix of size 1×1, then

$$a_{ij} = \begin{cases} 1 & iff\ w_i \text{ or } w_j \text{ is active persistent limit language} \\ 0 & iff\ w_i \text{ or } w_j \text{ is adult/inert limit language} \end{cases}.$$

Proof:
Let A_{ij}^{∞} be a limit adjacency matrix. Since it is a binary matrix of size 1×1, there are two cases to be considered:

Case 1: If a_{ij} =1, then $\sum_r^1 a_{i,r} = \sum_r^1 a_{r,j} = 1$. By Theorem 2, it is an active persistent limit language. Hence, the 'if part' is proven. The other part of the proof is obtained by retracing the above steps.

Case 2: If a_{ij} = 0, then $\sum_r^1 a_{i,r} = 0$. By Theorem 1, it is an adult/inert language. Hence, the 'if part' is proven. The other part of the proof is obtained by repeating the above steps.

Theorem 5: Any of the leading diagonal, a_{ii} or a_{jj} ,of a limit adjacency matrix is 1, if and only if there exists a trivial splicing of string w_i or w_j by itself.

Proof:
If there exists a trivial splicing in splicing system, then there exists a string w_k in language L, for all $w_i, w_j \in L$, such that $(w_i, w_k) \underset{L}{\overset{+}{\rightarrow}} w_i$ where w_k, w_i are strings in the language L. Hence, in limit adjacency matrix, the leading diagonal must be 1 ($(w_i, w_k) \xrightarrow{r} w_i$ is represented by the leading diagonal of limit adjacency matrix by the definition of A_{ij}^{∞}). If the leading diagonal a_{ii} or a_{jj}, of a limit adjacency matrix is 1, by definition of A_{ij}^{∞}, $(w_i, w_k) \xrightarrow{r} w_i$, which is a trivial splicing of string w_i and w_j by itself. Therefore the theorem is proved. ■

Theorem 6: Every splicing process can be replaced by limit adjacency matrix on creating the same limit languages.

Proof:
Let $S=(A,I,R)$ be a Y-G splicing system. Recall that any splicing process can be represented by limit graph G_L^∞ by Definition 7. Let A_{ij}^∞ be the limit adjacency matrix. Assume that limit languages generated from limit graph is the same as limit languages generated from limit adjacency matrix, where $L_\infty(A_{ij}^\infty)=L_\infty(G_L^\infty)$. This proof is presented by induction.

Suppose $I_1=\alpha abc\beta$ and $I_2=\gamma def\delta$ are two strings in I that can be spliced by using rules $R=(a,b,c:d,e,f)$. The splicing languages obtained are $m_1=\alpha abf\beta$ and $m_2=\gamma dec\delta$ where $L(S)=I\cup m_1,m_2$, by limit graph, m_1 and m_2 lies in limit languages since each of them lies in the terminal singleton Strongly Connected Component (SCC) [3]. There exist $m_1:\exists m_k\in L\ni(I_i,m_k)\to m_1$, but $\nexists(m_1,m_k)\to I_i$ for $i=1,2$ and $k=1,2$ and $m_2:\exists m_k\in L\ni(I_i,m_k)\to m_2$, but $\nexists(m_2,m_k)\to I_i$ for $i=1,2$ and $k=1,2$. Thus, m_1 and m_2 are categorized as limit languages by G_L^∞. Meanwhile, by limit adjacency matrix, m_1 and m_2 lies in limit languages from Theorem 1. Every element in a row of A_{ij}^∞ is zero from Definition 8. Hence, the same limit languages are obtained.

The induction hypothesis affirmed that the set of limit languages generated in A_{ij}^∞ up to the *k*-th iteration of splicing is the same as the set of limit languages generated in G_L^∞ up to the *k*-th iteration of splicing, where $k\geq 1$. Let x_1 and x_2 be two strings that are presented in both systems at the *k*-th iteration of splicing. If there exists G_L^∞ to present the sequence of splicing x_1 and x_2 in S in the $(k+1)^{th}$ iteration of splicing, then by an argument similar to that given in the basis, exist a corresponding A_{ij}^∞ to present the sequence of splicing x_1 and x_2 in S in the $(k+1)^{th}$ iteration of splicing , generating the same limit languages.
Therefore by induction, every splicing process, which represented by G_L^∞ can be replaced by A_{ij}^∞, since $L_\infty(A_{ij}^\infty)=L_\infty(G_L^\infty)$.■

Biological Example of Splicing Process via Limit Adjacency Matrix, A_{ij}^∞

In this subsection, two examples of DNA splicing process to determine active persistent language, adult language and transient language are elaborated with the limit adjacency matrix. According to [2], the restriction enzymes *Acl*I and *Aci*I yield the DNA Y-G splicing scheme, described as follows:

Example 1:
Let $S=(A,I,R)$ be a Y-G splicing scheme consisting $I_1=\alpha aacgtt\beta$ and $I_2=\gamma ccgc\delta$ with rule $r=(aa;cg,tt:c;cg,c)$. By using Y-G approach, the cutting and recombination process gives the resulting molecules as below.

$$w_1=\alpha aacgtt\alpha' \qquad w_6=\gamma ccgc\delta$$
$$w_2=\beta' aacgtt\beta \qquad w_7=\alpha aacgc\delta$$
$$w_3=\gamma ccgg\gamma' \qquad w_8=\gamma ccgtt\beta$$
$$w_4=\delta' gcgc\delta \qquad w_9=\alpha aacgg\gamma'$$
$$w_5=\alpha aacgtt\beta \qquad w_{10}=\delta' gcgtt\beta$$

The simulation of the DNA splicing process with the generated splicing languages above is demonstrated by limit adjacency matrix below.

$$
\begin{array}{c|cccccccccc|c}
 & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & w_{10} & \Sigma \\
\hline
w_1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 4 \\
w_2 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 4 \\
w_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_5 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 7 \\
w_6 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 7 \\
w_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
\Sigma & 2 & 2 & 1 & 1 & 3 & 1 & 3 & 3 & 3 & 3 &
\end{array}
$$

The limit adjacency matrix above clearly shows that the strings $w_3, w_4, w_7, w_8, w_9, w_{10}$ are adult/inert languages. Besides, strings w_1, w_2, w_5, w_6 are transient languages since $\sum_r^{10} a_{i,r} > \sum_r^{10} a_{r,j}$ by Corollary 1. There are no active persistent languages since the sum of column are all less than the sum of row excluding the row that contains all zero by Theorem 2. By looking at the diagonal matrix, there exists a trivial splicing of strings w_1, w_2, w_5, w_6 by themselves. To compare the results with the wet lab experiment done in [2] which shows adult/inert languages as $w_3, w_4, w_7, w_8, w_9, w_{10}$, limit adjacency matrix successfully predicted the result. Besides, the molecule w_6 is is shown in [2] as transient, which is the same as the result provided by A_{ij}^{∞}. However, the molecules w_1, w_2, w_5 are presented as active persistent language in [2], which is different from the predicted result above. From the conclusion in [2], it is due to the quantity of strings during the experiment yet the prediction in limit adjacency matrix above is ignoring the possibility of unbalanced numbers of molecules available for various reactions by the definition of limit language. Nevertheless, A_{ij}^{∞} shows that $\sum_r^{10} a_{1,r} = \sum_r^{10} a_{2,r} = 4 > \sum_r^{10} a_{r,1} = \sum_r^{10} a_{2,1} = 2, \sum_r^{10} a_{5,r} = 7 > \sum_r^{10} a_{r,5} = 3$, but $\sum_r^{10} a_{6,r} = 7 \sqcup \sum_r^{10} a_{r,6} = 1$. Hence, string w_6 is certainly transient but strings w_1, w_2, w_5 can be ambiguous case to stay in between active persistent and transient language depending on the quantity of initial strings. Therefore, the above example simulates the DNA splicing process in [2] in determining patterns of limit languages. The next example elaborates the difference between adult and limit languages in [4] by limit adjacency matrix.

Example 2:

Let $S = (A, I, R)$ be a Y-G splicing scheme consisting $I_1 = \alpha ccgc\beta$ and $I_2 = \gamma ccgg\delta$ with restriction enzyme *Aci*I, $r_1 = (c; cg, c)$ in stage one and *Hpa*II, $r_2 = (c; cg, g)$ in stage two. By using Y-G approach, the cutting and recombination process gives the resulting molecules as below.

Stage 1	Stage 2
$w_1 = \alpha ccgc\beta$	$w_2 = \alpha ccgg\alpha'$
$w_2 = \alpha ccgg\alpha'$	
$w_3 = \beta' gcgc\beta$	

The simulation of the DNA splicing process to equilibrium state with the generated splicing languages above is demonstrated by limit adjacency matrix below.

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ \Sigma \end{array} \begin{array}{c} \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \Sigma \\ \left[\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{array}\right] \begin{array}{c} 3 \\ 1 \\ 0 \\ \end{array} \end{array}$$

From A_{ij}^{∞}, it is apparent that w_3 is adult/inert language, w_2 is active persistent language and w_1 is transient language, which is parallel to the results produced by the mathematical model and wet lab experiment done in [4]. As in the examples, it is clear now that limit adjacency matrix is an easier approach to model the behavior of splicing languages.

CONCLUSION

Since matrices are better approach for representing the complex splicing process, limit adjacency matrix that can determine the number of patterns of limit languages are introduced in this paper. Besides, the characteristics of limit adjacency matrix in terms of the behaviour of splicing languages (adult, active persistent, transient languages) are explored and presented as theorems (from Theorem 1 to Theorem 6) and in terms of biological examples 1 and 2.

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: RDU131404.

REFERENCES

1. T.Head, *Bulletin of Mathematical Biology*, **49 (6)**,737-759, (1987)
2. Y. Yusof, Ph.D. Thesis, Universiti Teknologi Malaysia, (2012).
3. E. Goode and D. Pixton, *Lecture Notes in Computer Science*: Springer-Verlag. **189-201**; (2004)
4. W. H. Fong, Ph.D. Thesis, Universiti Teknologi Malaysia, (2008).
5. E. Laun, and K.J. Reddy, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. **48: 73-83** (1999)