MATEMATIKA: MJIAM, Special Issue, December 2019, 1–14

Automata for DNA Splicing Languages with Palindromic and Non-Palindromic Restriction Enzymes using Grammars

Wan Heng Fong, Nurul Izzaty Ismail and Nor Haniza Sarmin

Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: fwh@utm.my

Article history
Received: 10 November 2019
Received in revised form: 12 December 2019
Accepted: 20 December 2019
Published online: 31 December 2019

Abstract In DNA splicing system, DNA molecules are cut and recombined with the presence of restriction enzymes and a ligase. The splicing system is analyzed via formal language theory where the molecules resulting from the splicing system generate a language which is called a splicing language. In nature, DNA molecules can be read in two ways; forward and backward. A sequence of string that reads the same forward and backward is known as a palindrome. Palindromic and non-palindromic sequences can also be recognized in restriction enzymes. Research on splicing languages from DNA splicing systems with palindromic and non-palindromic restriction enzymes have been done previously. This research is motivated by the problem of DNA assembly to read millions of long DNA sequences where the concepts of automata and grammars are applied in DNA splicing systems to simplify the assembly in short-read sequences. The splicing languages generated from DNA splicing systems with palindromic and non-palindromic restriction enzymes are deduced from the grammars which are visualised as automata diagrams, and presented by transition graphs where transition labels represent the language of DNA molecules resulting from the respective DNA splicing systems.

Keywords Automata; DNA; splicing language; palindromic; restriction enzyme.

2010 Mathematics Subject Classification68Q45; 03D05; 92B05.

1 Introduction

Deoxyribonucleic acid (DNA) splicing system, also known as Head's splicing model, was modelled by Head [1] in 1987 based on a study relating informational macromolecules and formal language theory. In the splicing system, DNA molecules are cut and reassociated by restriction enzymes and a ligase to produce new molecules [1]. The splicing system is analyzed via formal language theory where the resulting molecules from the splicing process generate a language known as splicing language [1].

In formal language theory, a language is a set of strings of symbols where each string is called a word [2]. The operations on formal languages are applied in this research such us concatenation, union and star-closure [2]. Formal languages are generated by grammars to study languages mathematically and provide mechanisms for natural languages and programming languages.The theory of grammar is first introduced by Chomsky in [3] where grammar is used to generate strings of a language. The grammar contains the set of rules to transform a string to another string in the formation of language [2].

By using formal language theory, the molecules from splicing system act as strings which come from some portions in gene. The strings consist of double stranded DNA (dsDNA) symbols structured from two base pairings where adenine (A) forms with thymine (T), while cytosine (C) forms with guanine (G) [4]. The restriction enzyme acts as the rule in splicing system where the rule is formed as a triple: left context, crossing and right context [5].

Motivated by the concept from the process of recombinant DNA in Head's splicing model, variants of splicing models had been studied. For instance, Paun [6], Pixton [7], Goode-Pixton [8] and Yusof-Goode [9] splicing systems had been introduced with different notations of rule. There are different types of splicing languages such as simple [10], limit [8], second order, single and two stage splicing language which can be generated by different splicing models. In DNA splicing systems, splicing languages resulting from the splicing systems can be generalized based on palindromic and non-palindromic rules.

Palindrome is a string that reads the same forwards and backwards [11]. Research on the generalisations of splicing languages resulting from DNA splicing systems involving palindromic and non-palindromic rules has been done previously [12]. The name and sequence for all palindromic and non-palindromic rules used in this research are taken from [13].

In this research, the concepts of automata are applied in DNA splicing systems. In automata theory, finite state machine is a type of automata. Finite state machine is designed with a set of states, where every state accepts or rejects the input which produces output while moving to another state [14]. Various kinds of finite state machines had been designed such as Mealy machine and Moore machine. Mealy machine is a finite-state machine with output, where the output is determined by the current state and input [15],while Moore machine gives no output, has a set of final states and recognizes language from the inputs accepted from every path moving through the states.[16]. The finite state machine without output is also known as finite state automaton [14]. This research uses the concept of finite state automaton since the language generated by the automaton depicts the splicing language from the splicing system. The relation between splicing systems, maximal firm sub-words and automata is studied by Fong *et al.* [17]. In 2013, splicing languages from splicing systems are investigated using automata and grammars [18]. The automata diagrams for the general splicing language and the second order limit language are also presented in [19].

In this paper, the generalizations of splicing languages from DNA splicing system with palindromic and non-palindromic rules for the same and different crossings are given as deterministic finite automata diagrams using grammars, where the dsDNA strings are visualised as the inputs to depict the language generated by the grammars.

2 Preliminaries

In this research, the generalizations of splicing languages are generated from Head's splicing model. The splicing language is a language resulting from process of recombinant DNA in the splicing system in which both definitions are given in the following.

Definition 1 [1] Splicing System and Splicing Language
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. The language $L = L(S)$ generated by S consists of the strings in I and all strings that can be obtained by adjoining to $ucxfq$ and $pexdv$ whenever $ucxdv$ and $pexfq$ are in L and (c, x, d) and (e, x, f) are patterns of the same hand. A language, L, is a splicing language if there exists a splicing system S for which $L = L(S)$.

The sequences of enzymes can be analyzed as palindromic or non-palindromic rules. Next, the definition of a palindromic string is presented.

Definition 2 [20] Palindromic String
A string I of a dsDNA is said to be palindromic if the sequence from the left to the right side of the upper single strand is equal to the sequence from the right to the left side of the lower single strand.

For instance, the enzyme *CviA*II $\begin{matrix}5'-\text{CATG}-3'\\ 3'-\text{GTAC}-5'\end{matrix}$ is a palindromic rule since the upper single strand of enzyme *CviA*II is exactly the same with the lower single strand when reading $5'$ to $3'$ direction. Similarly, the enzyme *BseY*I $\begin{matrix}5'-\text{CCCAGC}-3'\\ 3'-\text{GGGTCG}-5'\end{matrix}$ is not palindrome since the upper single strand of enzyme *BseY*I, $5'-\text{CCCAGC}-3'$, does not match with the lower single strand $3'-\text{GGGTCG}-5'$when reading the same direction.

In automata theory, finite state automata can be deterministic since there is a unique move in each transition. This research applies concepts of deterministic finite automaton in the splicing systems. The definition of deterministic finite automaton is presented next.

Definition 3 [2] Deterministic Finite Automaton
A deterministic finite automaton M is a 5-tuple, $(Q, \Sigma, \delta, q_0, F)$ consisting of a finite set of states Q, a finite set of input symbols called the alphabet Σ, a transition function ($\delta : Q \times \Sigma \rightarrow \quad Q$), an initial state $q_0 \in Q$ and a set of final states $F \subseteq Q$.

The concepts of automata and grammars are used in this research to visualise the splicing language as automata diagrams using grammars. The definition of a grammar is shown next.

Definition 4 [2] Grammar
A grammar G is defined as a quadruple $G \quad = (V, T, S, P)$, where V is a finite set of objects called variables, T is a finite set of objects called terminal symbols, $S \in V$ is a special symbol called the start variable and P is a finite set of productions.

The set $L(G) = \left\{ w \in T^* : S \overset{*}{\Rightarrow} w \right\}$ is the language generated by G, where $\overset{*}{\Rightarrow}$ denotes zero or more steps of sequence of productions.

Figure 1 shows an example of a deterministic finite automaton that accepts the language $L = a \cdot ((b + c)\, a)^* \cdot c$ generated by the grammar with P consisting of the productions

$$S_0 \rightarrow aS_1$$

$$S_1 \rightarrow bS_0|cS_2$$

$$S_2 \rightarrow aS_1|\lambda$$

where $Q = \{S_0, S_1, S_2\}$, $\Sigma = \{a, b, c\}$, S_0 is the initial state, $F = \{S_2\}$ and δ is given by

$$\begin{aligned}\delta(S_0, a) &= S_1,\\ \delta(S_1, b) &= S_0,\\ \delta(S_1, c) &= S_2,\\ \delta(S_2, a) &= S_1.\end{aligned}$$

From the current state, the automaton makes a move to another state according to the transition function and accepts the input symbols. In the automaton, each state, final state, transitions, and input symbols (transition labels) are represented in single circle, double circle, arrows and arrow labels respectively.

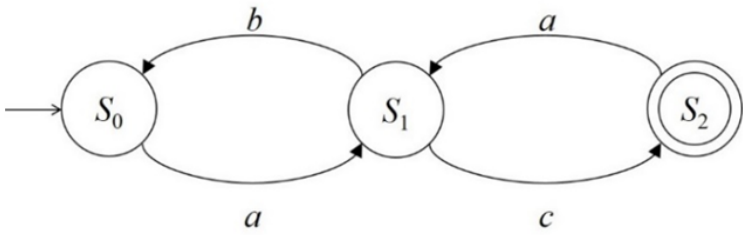

Figure 1: A Deterministic Finite Automaton

3 Research Methodology

The generalizations of splicing languages from DNA splicing systems with one cutting site each of palindromic and non-palindromic rules for the same and different crossings are used in this research for constructing the automata diagrams. The generalization of splicing languages from DNA splicing system with one cutting site each of palindromic and non-palindromic rules and the same crossing is presented in Theorem 1.

Theorem 1 [12] *Let* $S = (A, I, B, C)$ *be a DNA splicing system in which*

$$A = \left\{ \begin{matrix} \mathrm{A} \\ \mathrm{T} \end{matrix}, \begin{matrix} \mathrm{C} \\ \mathrm{G} \end{matrix}, \begin{matrix} \mathrm{G} \\ \mathrm{C} \end{matrix}, \begin{matrix} \mathrm{T} \\ \mathrm{A} \end{matrix} \right\}$$

is the set of dsDNA symbols,

$$I = \left\{ \begin{matrix} N_1N_1...N_1 & X_1 & Y & X_2 & M\,M\,...M & W_1 & Y & W_2 & N_2N_2...N_2 \\ N_1'N_1'...N_1' & X_1' & Y' & X_2' & M'M'....M' & W_1' & Y' & W_2' & N_2'N_2'...N_2' \end{matrix} \right\}$$

is the set consisting of an initial string with one cutting site each of palindromic and non-palindromic rules

$$\begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix} \quad and \quad \begin{matrix} W_1 & Y & W_2 \\ W_1' & Y' & W_2' \end{matrix}$$

where

$$\begin{matrix} N_1 \\ N_1' \end{matrix}, \begin{matrix} X_1 \\ X_1' \end{matrix}, \begin{matrix} Y \\ Y' \end{matrix}, \begin{matrix} X_2 \\ X_2' \end{matrix}, \begin{matrix} M \\ M' \end{matrix}, \begin{matrix} W_1 \\ W_1' \end{matrix}, \begin{matrix} W_2 \\ W_2' \end{matrix} \quad and \quad \begin{matrix} N_2 \\ N_2' \end{matrix}$$

are variables used to denote any arbitrary dsDNA and $N_1', X_1', Y', X_2', M', W_1', W_2'$ *and* N_2' *are complementaries for* $N_1, X_1, Y, X_2, M, W_1, W_2$ *and* N_2*, respectively, set*

$$B = \left\{ \left(\begin{matrix} X_1 \\ X_1' \end{matrix}, \begin{matrix} Y \\ Y' \end{matrix}, \begin{matrix} X_2 \\ X_2' \end{matrix} \right) \left(\begin{matrix} W_1 \\ W_1' \end{matrix}, \begin{matrix} Y \\ Y' \end{matrix}, \begin{matrix} W_2 \\ W_2' \end{matrix} \right) \right\}$$

is the set of rules where $\begin{matrix} Y \\ Y' \end{matrix}$ *is the crossing and set* C *is the empty set. The resulting splicing language consists of strings of the form*

$$\begin{pmatrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{pmatrix} + \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M \end{matrix} \begin{matrix} X_1 & Y \\ X_1' & Y' \end{matrix} \begin{pmatrix} X_2 & M\,M\,...M & W_1 & Y \\ X_2' & M'M'....M' & W_1' & Y' \end{pmatrix}^{n-1}$$

$$\left(\begin{matrix} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{matrix} + \begin{matrix} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{matrix} \right)$$

where $n \in \mathrm{Z}^+$ *represents multiple copies of strings and*

$$\left\{ \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix}, \begin{matrix} W_1 & Y & W_2 \\ W_1' & Y' & W_2' \end{matrix}, \begin{matrix} W_1' & Y' & W_2' \\ W_1 & Y & W_2 \end{matrix} \right\} \notin \left\{ \begin{matrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{matrix}, \begin{matrix} M\,M\,...M \\ M'M'....M' \end{matrix}, \begin{matrix} N_2N_2...N_2 \\ N_2'N_2'...N_2' \end{matrix} \right\}$$

which indicates no other cutting site is present in strings

$$\begin{matrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{matrix}, \begin{matrix} M\,M\,...M \\ M'M'....M' \end{matrix} \quad and \quad \begin{matrix} N_2N_2...N_2 \\ N_2'N_2'...N_2' \end{matrix} .$$

The generalization of splicing languages from DNA splicing system with one cutting site each of palindromic and non-palindromic rules and different crossings is presented in Theorem 2.

Theorem 2 [**12**] *Let* $S = (A, I, B, C)$ *be a DNA splicing system in which*

$$A = \left\{ \begin{matrix} \mathrm{A} \\ \mathrm{T} \end{matrix}, \begin{matrix} \mathrm{C} \\ \mathrm{G} \end{matrix}, \begin{matrix} \mathrm{G} \\ \mathrm{C} \end{matrix}, \begin{matrix} \mathrm{T} \\ \mathrm{A} \end{matrix} \right\}$$

is the set of dsDNA symbols,

$$I = \left\{ \begin{matrix} N_1N_1...N_1 & X_1 & Y & X_2 & M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 \\ N_1'N_1'...N_1' & X_1' & Y' & X_2' & M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' \end{matrix} \right\}$$

is the set consisting of an initial string with one cutting site each of palindromic and non-palindromic rules

$$\begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix}$$

and

$$\begin{matrix} W_1 & Z & W_2 \\ W_1' & Z' & W_2' \end{matrix}$$

where

$$\begin{matrix} N_1 \\ N_1' \end{matrix}, \begin{matrix} X_1 \\ X_1' \end{matrix}, \begin{matrix} Y \\ Y' \end{matrix}, \begin{matrix} X_2 \\ X_2' \end{matrix}, \begin{matrix} M \\ M' \end{matrix}, \begin{matrix} W_1 \\ W_1' \end{matrix}, \begin{matrix} Z \\ Z' \end{matrix}, \begin{matrix} W_2 \\ W_2' \end{matrix} \text{ and } \begin{matrix} N_2 \\ N_2' \end{matrix}$$

are variables used to denote any arbitrary dsDNA and $N_1', X_1', Y', X_2', M', W_1', Z', W_2'$ *and* N_2' *are complementaries for* $N_1, X_1, Y, X_2, M, W_1, Z, W_2$ *and* N_2*, respectively, set*

$$B = \left\{ \begin{pmatrix} X_1 \\ X_1' \end{pmatrix}, \begin{pmatrix} Y \\ Y' \end{pmatrix}, \begin{pmatrix} X_2 \\ X_2' \end{pmatrix} \begin{pmatrix} W_1 \\ W_1' \end{pmatrix}, \begin{pmatrix} Z \\ Z' \end{pmatrix}, \begin{pmatrix} W_2 \\ W_2' \end{pmatrix} \right\}$$

is the set of rules where $\begin{matrix} Y \\ Y' \end{matrix}$ *and* $\begin{matrix} Z \\ Z' \end{matrix}$ *are the different crossings and set* C*is the empty set. The resulting splicing language consists of strings of the form*

$$\begin{pmatrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{pmatrix} + \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M \end{matrix} \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix}$$

$$\begin{pmatrix} M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 \\ M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' \end{pmatrix} + \begin{matrix} N_1'N_1'...N_1' \\ N_1N_1...N_1 \end{matrix}$$

where

$$\left\{ \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix}, \begin{matrix} W_1 & Z & W_2 \\ W_1' & Z' & W_2' \end{matrix}, \begin{matrix} W_1' & Z' & W_2' \\ W_1 & Z & W_2 \end{matrix} \right\} \notin \left\{ \begin{matrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{matrix}, \begin{matrix} M\,M\,...M \\ M'M'....M' \end{matrix}, \begin{matrix} N_2N_2...N_2 \\ N_2'N_2'...N_2' \end{matrix} \right\}$$

which indicates no other cutting site is present in strings

$$\begin{matrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{matrix}, \begin{matrix} M\,M\,...M \\ M'M'....M' \end{matrix} \text{ and } \begin{matrix} N_2N_2...N_2 \\ N_2'N_2'...N_2' \end{matrix}.$$

4 Results and Discussion

In this paper, the generalized splicing languages from DNA splicing systems with one cutting site each of palindromic and non-palindromic rules are given as automata diagrams. In order to construct the automata diagrams, the generalized splicing languages from the corresponding DNA splicing systems are deduced from grammars and the automata diagrams are presented as theorems.

The automaton for DNA splicing system with one cutting site each of palindromic and non-palindromic rules and the same crossing is presented in Theorem 3.

Theorem 3 *Given*

$$S = \left\{ \left\{ \begin{matrix} \mathrm{A} \\ \mathrm{T} \end{matrix}, \begin{matrix} \mathrm{C} \\ \mathrm{G} \end{matrix}, \begin{matrix} \mathrm{G} \\ \mathrm{C} \end{matrix}, \begin{matrix} \mathrm{T} \\ \mathrm{A} \end{matrix} \right\}, \begin{matrix} N_1N_1...N_1 & X_1 & Y & X_2 & M\,M\,...M & W_1 & Y & W_2 & N_2N_2...N_2 \\ N_1'N_1'...N_1' & X_1 & Y' & X_2' & M'M'....M' & W_1' & Y' & W_2' & N_2'N_2'...N_2' \end{matrix} \right\},$$

$$\left\{\left(\begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array}\right)\left(\begin{array}{ccc} W_1 & Y & W_2 \\ W_1' & Y' & W_2' \end{array}\right)\right\}, \emptyset\Bigg\}$$

is a splicing system involving one cutting site each of palindromic and non-palindromic rules

$$\begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} \quad and \quad \begin{array}{ccc} W_1 & Y & W_2 \\ W_1' & Y' & W_2' \end{array}$$

with the same crossing

$$\begin{array}{c} Y \\ Y' \end{array} \quad where \quad \begin{array}{c} N_1 \\ N_1' \end{array}, \begin{array}{c} X_1 \\ X_1' \end{array}, \begin{array}{c} Y \\ Y' \end{array}, \begin{array}{c} X_2 \\ X_2' \end{array}, \begin{array}{c} M \\ M' \end{array}, \begin{array}{c} W_1 \\ W_1' \end{array}, \begin{array}{c} W_2 \\ W_2' \end{array} \quad and \quad \begin{array}{c} N_2 \\ N_2' \end{array}$$

are variables used to denote any arbitrary dsDNA and $N_1', X_1', Y', X_2', M', W_1', W_2'$ *and* N_2 *' are complementaries for* $N_1, X_1, Y, X_2, M, W_1, W$ *and* N_2*, respectively,* $M_1 = (Q, \Sigma, \delta, q_0, F)$*is a deterministic finite automaton for the DNA splicing system that accepts the language* $L(S)$ *in which* $Q = \{q_0, q_1, q_2, q_3, q_4, q_5\}$ *is the set of states where* q_0*is the initial state and* $F = \{q_4, q_5\}$ *is the set of final states,*

$$\Sigma = \left\{ \begin{array}{cccccccc} N_1N_1...N_1 & X_1 & N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_1'N_1'...N_1' & X_1' & , N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{array} \right.,$$

$$\left. \begin{array}{cccccccc} Y & X_2 & M\,M\,...M & W_1 & W_2 & N_2N_2...N_2 & X_2 & N_1'N_1'...N_1' \\ Y' & X_2' & M'M'....M' & W_1' & W_2' & N_2'N_2'...N_2' & X_2' & N_1N_1...N_1 \end{array} \right\}$$

is the set of inputs and δ *is given by*

$$\delta\left(q_0, \begin{array}{ccc} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{array}\right) = q_1,$$

$$\delta\left(q_0, \begin{array}{cccccc} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{array}\right) = q_2,$$

$$\delta\left(q_1, \begin{array}{c} Y \\ Y' \end{array}\right) = q_3, \quad \delta\left(q_2, \begin{array}{c} Y \\ Y' \end{array}\right) = q_3,$$

$$\delta\left(q_3, \begin{array}{ccc} X_2 & M\,M\,...M & W_1 \\ X_2' & M'M'....M' & W_1' \end{array}\right) = q_1, \quad \delta\left(q_3, \begin{array}{ccc} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{array}\right) = q_4$$

and

$$\delta\left(q_3, \begin{array}{ccc} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{array}\right) = q_5.$$

The automaton diagram for M_1 *is shown in Figure 2.*

Proof

In this proof, S_0, S_1, S_2, S_3, S_4 and S_5represent states q_0, q_1, q_2, q_3, q_4 and q_5, respectively, where q_4 and q_5 are the final states. The splicing language $L(S)$ from the splicing system can be written as a language generated by a grammar G_1 where

$$G_1 = \left(\{S_0, S_1, S_2, S_3, S_4, S_5\}, \left\{ \begin{array}{c} N_1 \\ N_1' \end{array}, \begin{array}{c} X_1 \\ X_1' \end{array}, \begin{array}{c} Y \\ Y' \end{array}, \begin{array}{c} X_2 \\ X_2' \end{array}, \begin{array}{c} M \\ M' \end{array}, \begin{array}{c} W_1 \\ W_1' \end{array}, \begin{array}{c} W_2 \\ W_2' \end{array}, \begin{array}{c} N_2 \\ N_2' \end{array} \right\}, S_0P_1\right)$$

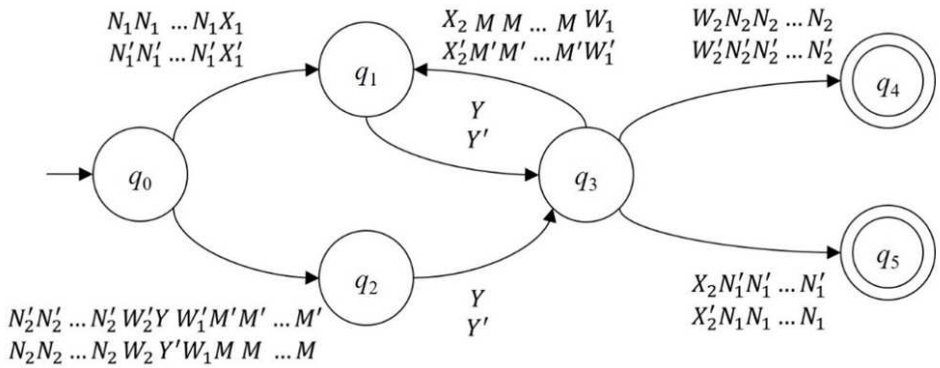

Figure 2: An Automaton Diagram for M_1

with P_1 consisting of the productions,

$$S_0 \rightarrow \begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix} S_1 \Big| \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix} S_2$$

$$S_1 \rightarrow \begin{matrix} Y \\ Y' \end{matrix} S_3$$

$$S_2 \rightarrow \begin{matrix} Y \\ Y' \end{matrix} S_3$$

$$S_3 \rightarrow \begin{matrix} X_2 & M\,M\,...M & W_1 \\ X_2' & M'M'....M' & W_1' \end{matrix} S_1 \Big| \begin{matrix} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{matrix} S_4 \Big| \begin{matrix} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{matrix} S_5$$

$S_4 \rightarrow \lambda$ and $S_5 \rightarrow \lambda$. Then, a sequence for the language generated by the grammar G_1, $L(G_1)$ is

$$\begin{aligned}
S_0 \quad &\Rightarrow \quad \left(\begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix} S_1 + \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix} S_2 \right) \\
&\Rightarrow \quad \left(\begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix} + \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix} \right) \begin{matrix} Y \\ Y' \end{matrix} S_3 \\
&\Rightarrow \quad \left(\begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix} + \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix} \right) \begin{matrix} Y \\ Y' \end{matrix} \\
&\qquad \left(\begin{matrix} X_2 & M\,M\,...M & W_1 \\ X_2' & M'M'....M' & W_1' \end{matrix} S_1 \right) \left(\begin{matrix} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{matrix} S_4 + \begin{matrix} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{matrix} S_5 \right) \\
&\Rightarrow \quad \left(\begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix} + \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix} \right) \begin{matrix} Y \\ Y' \end{matrix} \\
&\qquad \left(\begin{matrix} X_2 & M\,M\,...M & W_1 & Y \\ X_2' & M'M'....M' & W_1' & Y' \end{matrix} \right)^* \left(\begin{matrix} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{matrix} + \begin{matrix} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{matrix} \right)
\end{aligned}$$

which depicts the splicing language $L(S)$ from Theorem 1.

Based on G_1, the automaton for the splicing system is constructed using productions in G_1. The relation between productions in G_1 and transition functions, $\delta$for $M_1$are given in Table 1. Thus, Theorem 3 is proved. □

Table 1: Productions in G_1 and Transition Functions for M_1

Production in G_1	**Transition Function, δ**
$S_0 \rightarrow \begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix} S_1$	$\delta\left(q_0, \begin{matrix} N_1N_1...N_1 & X_1 \\ N_1'N_1'...N_1' & X_1' \end{matrix}\right) = q_1$
$S_0 \rightarrow \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix} S_2$	$\delta\left(q_0, \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & X_1 \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & X_1' \end{matrix}\right) = q_2$
$S_1 \rightarrow \begin{matrix} Y \\ Y' \end{matrix} S_3$	$\delta\left(q_1, \begin{matrix} Y \\ Y' \end{matrix}\right) = q_3$
$S_2 \rightarrow \begin{matrix} Y \\ Y' \end{matrix} S_3$	$\delta\left(q_2, \begin{matrix} Y \\ Y' \end{matrix}\right) = q_3$
$S_3 \rightarrow \begin{matrix} X_2 & M\,M\,...M & W_1 \\ X_2' & M'M'....M' & W_1' \end{matrix} S_1$	$\delta\left(q_3, \begin{matrix} X_2 & M\,M\,...M & W_1 \\ X_2' & M'M'....M' & W_1' \end{matrix}\right) = q_1$
$S_3 \rightarrow \begin{matrix} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{matrix} S_4$	$\delta\left(q_3, \begin{matrix} W_2 & N_2N_2...N_2 \\ W_2' & N_2'N_2'...N_2' \end{matrix}\right) = q_4$
$S_3 \rightarrow \begin{matrix} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{matrix} S_5$	$\delta\left(q_3, \begin{matrix} X_2 & N_1'N_1'...N_1' \\ X_2' & N_1N_1...N_1 \end{matrix}\right) = q_5$

The automaton for DNA splicing system with one cutting site each of palindromic and non-palindromic rules and different crossings is presented in Theorem 4.

Theorem 4 *Given*

$$S = \left\{ \left\{ \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \\ \mathrm{T} & \mathrm{G} & \mathrm{C} & \mathrm{A} \end{array} \right\}, \right.$$
$$\begin{array}{cccccccccc} N_1N_1...N_1 & X_1 & Y & X_2 & M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 \\ N_1'N_1'...N_1' & X_1' & Y' & X_2' & M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' \end{array},$$
$$\left. \left\{ \left(\begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} \right) \left(\begin{array}{ccc} W_1 & Z & W_2 \\ W_1' & Z' & W_2' \end{array} \right) \right\}, \emptyset \right\}$$

is a splicing system involving one cutting site each of palindromic and non-palindromic rules

$$\begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} \quad \textit{and} \quad \begin{array}{ccc} W_1 & Z & W_2 \\ W_1' & Z' & W_2' \end{array}$$

with different crossings

$$\begin{array}{c} Y \\ Y' \end{array} \quad \textit{and} \quad \begin{array}{c} Z \\ Z' \end{array}$$

where

$$\begin{array}{ccccccccccc} N_1 & X_1 & Y & X_2 & M & W_1 & Z & W_2 & \textit{and} & N_2 \\ N_1' & X_1' & Y' & X_2' & M' & W_1' & Z' & W_2' & & N_2' \end{array}$$

are variables used to denote any arbitrary dsDNA and $N_1', X_1', Y', X_2', M', W_1', Z', W_2'$ *and* N_2' *are complementaries for* $N_1, X_1, Y, X_2, M, W_1, Z, W_2$ *and* N_2 *respectively,* $M_2 = (Q, \Sigma, \delta, q_0, F)$*is a deterministic finite automaton for the DNA splicing system that accepts the language* $L(S)$ *in which* $Q = \{q_0, q_1, q_2, q_3, q_4, q_5\}$ *is the set of states where* q *is the initial state and* $F = \{q_4, q_5\}$ *is the set of final states,*

$$\Sigma = \left\{ \begin{array}{cccccccc} N_1N_1...N_1 & N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' & & \\ N_1'N_1'...N_1' & N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M & , & \end{array} \right.$$
$$\left. \begin{array}{ccccccccc} X_1 & Y & X_2 & M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 & N_1'N_1'...N_1' \\ X_1' & Y' & X_2' & M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' & N_1N_1...N_1 \end{array} \right\}$$

is the set of inputs and δ *is given by*

$$\delta \left(q_0, \begin{array}{c} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{array} \right) = q_1,$$
$$\delta \left(q_0, \begin{array}{ccccc} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M \end{array} \right) = q_2,$$
$$\delta \left(q_1, \begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} \right) = q_3, \quad \delta \left(q_2, \begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} \right) = q_3,$$
$$\delta \left(q_3, \begin{array}{ccccc} M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 \\ M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' \end{array} \right) = q_4 \textit{ and } \delta \left(q_3, \begin{array}{c} N_1'N_1'...N_1' \\ N_1N_1...N_1 \end{array} \right) = q_5.$$

The automaton diagram for M_2 *is shown in Figure 3.*

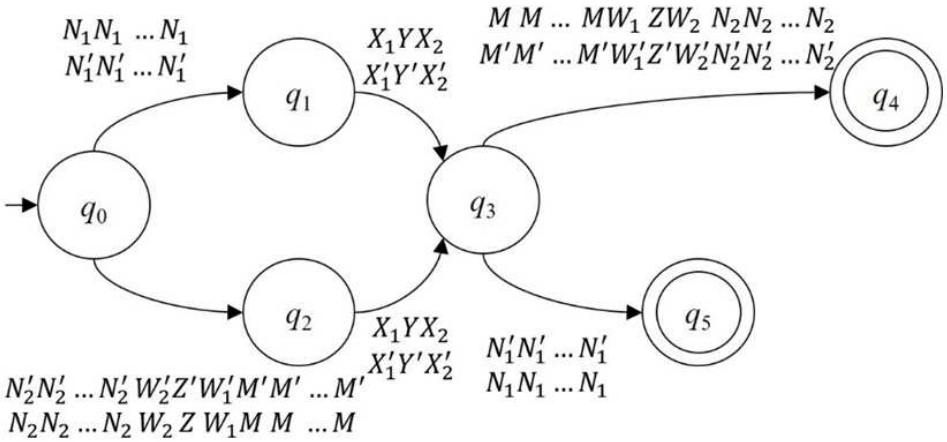

Figure 3: An Automaton Diagram for M_2

Proof

In this proof, S_0, S_1, S_2, S_3, S_4 and S_5 represent states q_0, q_1, q_2, q_3, q_4 and q_5,respectively, where q_4 and q_5 are the final states. The splicing language $L(S)$ from the splicing system can be written as a language generated by a grammar G_2 where

$$G_2 = \left(\{S_0, S_1, S_2, S_3, S_4, S_5\}, \left\{ \begin{array}{c} N_1 \\ N_1' \end{array}, \begin{array}{c} X_1 \\ X_1' \end{array}, \begin{array}{c} Y \\ Y' \end{array}, \begin{array}{c} X_2 \\ X_2' \end{array}, \begin{array}{c} M \\ M' \end{array}, \begin{array}{c} W_1 \\ W_1' \end{array}, \begin{array}{c} Z \\ Z' \end{array}, \begin{array}{c} W_2 \\ W_2' \end{array}, \begin{array}{c} N_2 \\ N_2' \end{array} \right\}, S_0P_2\right)$$

with P_2 consisting of the productions,

$$S_0 \rightarrow \begin{array}{c} N_1N_1 \ldots N_1 \\ N_1'N_1' \ldots N_1' \end{array} S_1 \Big| \begin{array}{cccccc} N_2'N_2' \ldots N_2' & W_2' & Y & W_1' & M'M' \ldots M' \\ N_2N_2 \ldots N_2 & W_2 & Y' & W_1 & M\,M \ldots M \end{array} S_2,$$

$$S_1 \rightarrow \begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} S_3,$$

$$S_2 \rightarrow \begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} S_3,$$

$$S_3 \rightarrow \begin{array}{cccccc} M\,M \ldots M & W_1 & Z & W_2 & N_2N_2 \ldots N_2 \\ M'M' \ldots M' & W_1' & Z' & W_2' & N_2'N_2' \ldots N_2' \end{array} S_4 \Big| \begin{array}{c} N_1'N_1' \ldots N_1' \\ N_1N_1 \ldots N_1 \end{array} S_5,$$

$S_4 \rightarrow \lambda$ and $S_5 \rightarrow \lambda$.

Then, a sequence for the language generated by the grammar G_2, $L(G_2)$ is

$$\begin{aligned} S_0 &\Rightarrow \left(\begin{array}{c} N_1N_1 \ldots N_1 \\ N_1'N_1' \ldots N_1' \end{array} S_1 + \begin{array}{cccccc} N_2'N_2' \ldots N_2' & W_2' & Y & W_1' & M'M' \ldots M' \\ N_2N_2 \ldots N_2 & W_2 & Y' & W_1 & M\,M \ldots M \end{array} S_2 \right) \\ &\Rightarrow \left(\begin{array}{c} N_1N_1 \ldots N_1 \\ N_1'N_1' \ldots N_1' \end{array} + \begin{array}{cccccc} N_2'N_2' \ldots N_2' & W_2' & Y & W_1' & M'M' \ldots M' \\ N_2N_2 \ldots N_2 & W_2 & Y' & W_1 & M\,M \ldots M \end{array} \right) \begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} S_3 \\ &\Rightarrow \left(\begin{array}{c} N_1N_1 \ldots N_1 \\ N_1'N_1' \ldots N_1' \end{array} + \begin{array}{cccccc} N_2'N_2' \ldots N_2' & W_2' & Y & W_1' & M'M' \ldots M' \\ N_2N_2 \ldots N_2 & W_2 & Y' & W_1 & M\,M \ldots M \end{array} \right) \begin{array}{ccc} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{array} \\ &\quad \left(\begin{array}{cccccc} M\,M \ldots M & W_1 & Z & W_2 & N_2N_2 \ldots N_2 \\ M'M' \ldots M' & W_1' & Z' & W_2' & N_2'N_2' \ldots N_2' \end{array} S_4 + \begin{array}{c} N_1'N_1' \ldots N_1' \\ N_1N_1 \ldots N_1 \end{array} S_5 \right), \end{aligned}$$

which depicts the splicing language $L(S)$ from Theorem 2.

Based on G_2, the automaton for the splicing system is constructed using productions in G_2. The relation between productions in G_2 and transition functions, δ for M_2 are given in Table 2. Thus, Theorem 4 is proved. □

Table 2: Productions in G_2 and transition functions for M_2

Production in G_2	**Transition Function, δ**
$S_0 \rightarrow \begin{matrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{matrix} S_1$	$\delta\left(q_0, \begin{matrix} N_1N_1...N_1 \\ N_1'N_1'...N_1' \end{matrix}\right) = q_1$
$S_0 \rightarrow \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M \end{matrix} S_2$	$\delta\left(q_0, \begin{matrix} N_2'N_2'...N_2' & W_2' & Y & W_1' & M'M'....M' \\ N_2N_2...N_2 & W_2 & Y' & W_1 & M\,M\,...M \end{matrix}\right) = q_2$
$S_1 \rightarrow \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix} S_3$	$\delta\left(q_1, \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix}\right) = q_3$
$S_2 \rightarrow \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix} S_3$	$\delta\left(q_2, \begin{matrix} X_1 & Y & X_2 \\ X_1' & Y' & X_2' \end{matrix}\right) = q_3$
$S_3 \rightarrow \begin{matrix} M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 \\ M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' \end{matrix} S_4$	$\delta\left(q_3, \begin{matrix} M\,M\,...M & W_1 & Z & W_2 & N_2N_2...N_2 \\ M'M'....M' & W_1' & Z' & W_2' & N_2'N_2'...N_2' \end{matrix}\right) = q_4$
$S_3 \rightarrow \begin{matrix} N_1'N_1'...N_1' \\ N_1N_1...N_1 \end{matrix} S_5$	$\delta\left(q_3, \begin{matrix} N_1'N_1'...N_1' \\ N_1N_1...N_1 \end{matrix}\right) = q_5$

Therefore, automata diagrams can be used to visualize the generalized splicing languages from DNA splicing systems with palindromic and non-palindromic rules using the grammars.

5 Conclusion

In this research, the concepts in automata theory and grammar are applied in DNA splicing systems with one cutting site each of palindromic and non-palindromic rules for the same and different crossings using deterministic finite automata. The automata diagrams for the DNA splicing systems are given as Theorems 3 and 4 and constructed using the grammars for the generalizations of splicing languages from the corresponding DNA splicing systems, which are given in Theorems 1 and 2. The languages generated by the grammars depict the generalized splicing languages consisting of the dsDNA strings.

Acknowledgments

The first and third authors would like to express their gratitude to the Ministry of Education (MOE) and Research Management Centre (RMC), Universiti Teknologi Malaysia (UTM) for the funding through Fundamental Research Grant Scheme Vote No. 5F022. The second author would also like to thank Universiti Teknologi Malaysia (UTM) for supporting her study through Zamalah Scholarship.

References

[1] Head, T. Formal Language Theory and DNA: An analysis of the generative capacity of specific recombinant behaviors. *B. Math. Biol.* 1987. 49(6): 737-759.

[2] Linz, P. *An Introduction to Formal Languages and Automata*. 4th ed. USA: Jones and Bartlett Publisher. 2006.

[3] Chomsky, N. Three models for the description of language. *IRA Transactions on Information Theory*. 1956. 2(3): 113-124.

[4] Paun, G., Rozenberg, G. and Salomaa, A. *DNA Computing: New Computing Paradigms*. Germany: Springer -Verlag Berlin Heidelberg. 1998.

[5] Head, T. Splicing representations of strictly locally testable languages. *Discrete. Appl. Math.* 1998. 87(1): 139-147.

[6] Pun, G. On the splicing operation. *Discrete. Appl. Math.* 1996. 70(1): 57-79.

[7] Pixton, D. Regularity of splicing languages. *Discrete. Appl. Math.* 1996. 69(1-2): 101-124.

[8] Goode, E. and Pixton, D. Splicing to the limit. In Jonoska, N., Pun, G., and Rozenberg, G. (Ed.). *Aspects of Molecular Computing, Lecture Notes in Computer Science*. Germany: Springer-Verlag. 2004. 189-201.

[9] Yusof, Y., Sarmin, N. H., Fong, W. H., Goode, T. E. and Ahmad, M. A. An analysis of four variants of splicing system. In *Proceeding of the 20th National Symposium on Mathematical Sciences - Research in Mathematical Sciences: A Catalyst for Creativity and Innovation (SKSM 2012), December 18-20, 2012*. Melville, NY: AIP Conference Proceedings. 2013. 888-895.

[10] Laun, T. E. G. *Constants and Splicing Systems*. Ph.D. Thesis. State University of New York. 1999.

[11] Tomohiro, I., Inenaga, S. and Takeda, M. Palindrome pattern matching. *Theor. Comput. Sci.* 2013. 483: 162-170.

[12] Ismail, N. I., Fong, W. H. and Sarmin, N. H. The mathematical modelling of dna splicing system with palindromic and non-palindromic restriction enzymes. In *Proceeding of the International Conference on Applied Analysis and Mathematical Modeling, June 20-24, 2018*. Istanbul, Turkey: ICAAMM 2018, Istanbul Gelism University. 2018. 127-138.

[13] New England Biolabs Inc 2017. NEB 2017-18 Catalog & Technical Reference. Ipswich, United States. Catalogue.

[14] Rosen, K. H. *Discrete Mathematics & Applications*. 8th ed. New York: McGraw-Hill Education. 2019.

[15] Mealy, G. H. A method for synthesizing sequential circuits. *The Ball System Tecnical Journal*. 1955. 34(5): 1045-1079.

[16] Moore, E. F. Gedanken-Experiments on Sequential Machines. In Shannon, C. E. and McCarthy, J. (Ed.). *Automata studies: Annals of Mathematical Studies*. Princeton, NJ: Princeton University Press. 1956. 129-153.

[17] Fong, W. H., Sarmin, N. H. and Ibrahim, Z. Recognition of simple splicing systems using SH automaton. *MJFAS*. 2008. 4(2): 337-342.

[18] Mohamad Jan, N., Fong, W. H. and Sarmin, N. H. Regular languages, regular grammars and automata in splicing systems. In *Proceeding of the 20th National Symposium on Mathematical Sciences: Research in Mathematical Sciences: A Catalyst for Creativity and Innovation, December 18–20, 2012*. Melville, NY: AIP Conference Proceedings. 2013. 856-863.

[19] Ahmad, M. A. *Second Order Limit Language and Its Properties in Yusof-Goode Splicing System*. Ph.D. Thesis. Universiti Teknologi Malaysia. 2016.

[20] Yusof, Y. *DNA Splicing System Inspired by Bio Molecular Operations*. Ph.D. Thesis. Universiti Teknologi Malaysia. 2012.