PERSISTENCY AND PERMANENCY OF TWO STAGES SPLICING LANGUAGES BASED ON DNA RECOMBINATION PROCESS BY USING YUSOF-GOODE (Y-G) APPROACH

MOHAMMAD HASSAN MUDABER

Thesis submitted in fulfilment of the requirements
for the awards of the degree of
Master of Science (Mathematics)

Faculty of Industrial Sciences and Technology
UNIVERSITI MALAYSIA PAHANG

MARCH 2015

ABSTRACT

The study of the biological process of deoxyribonucleic acid (DNA) splicing was investigated by Yusof in 2012 as a translucent approach under framework of formal languages theory. The investigation was achieved by proposing a new symbolization representing the rule in splicing system called Yusof-Goode (Y-G) rule associated with Y-G splicing system. A laboratory experiment is usually performed to show the DNA splicing process under the existence of restriction enzyme and appropriate ligase. This laboratory experiment is time consuming process and it can incur a lot of expenses. In addition, to accomplish the reaction up to two stages the generating recombining DNA molecules must have persistency concept. Thus, to overcome this problem, mathematical approach via Y-G model is applied to introduce the new concept of splicing system at two stages with respect to two initial strings and two rules. Furthermore, the existence relation between the families of stage one and stage two splicing languages are investigated. In this investigation, some sufficient conditions for persistency and permanency of two stages DNA splicing languages according to the number of cutting sites of initial strings and properties of rules are provided. The de Bruijn graph is used to predict the persistency and permanency of two stages splicing languages. A user friendly interface as an alternative of wet-lab experiment is coded using Microsoft Visual C Sharp (C#) to predict the persistency and permanency, and the relations between two stages (stage one and stage two) splicing languages.

ABSTRAK

Kajian proses biologi hiris-cantum asid deoksiribonukleik (DNA) telah dikaji oleh Yusof pada 2012 sebagai satu pendekatan telus di bawah rangka kerja teori bahasa formal. Kajian itu telah dicapai dengan mencadangkan simbol baru untuk mewakili peraturan dalam sistem hiris cantum yang dipanggil peraturan Yusof-Goode (Y-G) yang berkaitan dengan sistem hiris-cantum Y-G. Satu ujikaji makmal biasanya dijalankan untuk menunjukkan proses hiris-cantum DNA dengan kehadiran enzim pembatas dan ligase yang sesuai. Ujikaji makmal ini merupakan proses yang memakan masa dan boleh menelan perbelanjaan yang tinggi. Tambahan pula, untuk mencapai tindak balas sehingga peringkat kedua, gabungan molekul-molekul DNA yang terhasil harus mempunyai konsep berterusan. Oleh itu, bagi mengatasi masalah ini, pendekatan matematik melalui model Y-G diaplikasikan untuk memperkenalkan konsep baharu sistem hiris-cantum pada peringkat kedua dengan dua jujukan awal dan dua peraturan. Tambahan pula, hubungan antara peringkat satu dan peringkat kedua bahasa hiris-cantum telah dikaji. Dalam kajian ini, beberapa syarat cukup untuk konsep berterusan dan kekal dua peringkat bahasa hiris-cantum DNA serta beberapa ciri bahasa hiris-cantum juga turut diberikan dari segi berterusan dan kekal berdasarkan bilangan potongan dalam jujukan awal dan ciri-ciri peraturan. Graf de Bruijn telah digunakan untuk meramal bahasa hiris-cantum kekal dan berterusan. Satu antara muka pengguna sebagai altenatif bagi ujikaji makmal dikod menggunakan Microsoft Visual C Sharp (C#) untuk meramalkan konsep berterusan dan kekal, dan juga hubungan antara dua peringkat (peringkat satu dan peringkat kedua) bahasa-bahasa hiris-cantum.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

a	$[A/T]$: Base pairing of adenine and thymine
c	$[C/G]$: Base pairing of cytosine and guanine
g	$[G/C]$: Base pairing of guanine and cytosine
t	$[T/A]$: Base pairing of thymine and adenine
▼... ...▲	Crossing of recognition site of restriction enzymes
S	Splicing system
$S_k H$	Splicing languages families
$L(S)$	Language generated by splicing system at stage one
$L'(S)$	Language generated by splicing system at stage two
A	Set of four alphabets
I	Set of initial strings
R	Set of splicing rules
A^*	Set of strings that consists of zero or more symbols concatenated from A
$\{\ \}$	Set
$\subseteq$	Subset
$\cup$	Union
$\neq$	Not equal to
$\in$	Element of
$\alpha, \beta, \gamma, \delta$	Strings in A^*
C#	C sharp

□ End of definition/ theorem

■ End of proof/ example

LIST OF ABBREVIATIONS

bp	Base pair
DNA	Deoxyribonucleic acid
dsDNA	Double- stranded DNA
NEB	New England Biolab
RE	Restriction enzyme
SL	Splicing language
SLT	Strictly locally testable
SS	Splicing system
UI	User friendly interface
Y-G	Yusof-Goode

CHAPTER 1

INTRODUCTION

1.1 AN OVERVIEW

Manipulating operation of deoxyribonucleic acid (DNA) molecules is biochemically accomplished by a recombination process. The double-stranded DNA is cleaved by restriction enzyme into fragments; then the pasting operation can be achieved by the existence of appropriate ligase to form new DNA molecules. The study of the biological process of DNA splicing system in a translucent approach was investigated by Yusof (2012) under the framework of formal language theory. This novelty mathematical model which is known as Yusof-Goode (Y-G) splicing system is considered as a new contribution in terms of biology, particularly in the field of molecular DNA. According to its formulation, Y-G splicing system consists of three main sets: set of finite alphabets, set of initial strings and set of splicing rules. In this formalism, the DNA molecules, restriction enzymes and recombination process are considered as initial strings, splicing rules and splicing operation, respectively. The language which is generated by Y-G splicing system is called Y-G splicing language. In this thesis, the splicing languages that are produced by Y-G splicing system are defined as two stages splicing languages. Nevertheless, the existence relations between stage one and stage two splicing languages as well as the persistency and permanency of two stages splicing languages are investigated. Additionally, the de Bruijn graph of the above mentioned two stages DNA splicing languages are developed.

1.2 RESEARCH BACKGROUND

DNA is a double-stranded molecule consisting of two single strands over four alphabets *A*, *G*, *C*, and *T* which are abbreviated from adenine, guanine, cytosine and thymine, respectively. Based on Watson-Crick proposal, the two strands of DNA are antiparallel and have complementary property, where *A* is only paired with *T* and *C* is paired with *G* and vice-versa (Kratz, 2009).

To study the connectivity between recombination process and formal language theory, Head (1987) formulated splicing system mathematically based on the biological problem of cutting and pasting phenomenon. In addition in terms of mathematics the recombinant DNA molecules which will arise by the recombinant process are called splicing languages.

Yusof (2012) investigated on three classes of splicing system namely: non- semi simple splicing system, free splicing system and non-semi simple splicing system restricted to two rules and presented new definitions for these three types of splicing systems. Besides, a new symbolization representing the splicing system namely Y-G splicing system which is associated with Y-G rule was proposed. The language which is produced by Y-G splicing system is called Y-G splicing language. Moreover, the concepts of persistent and permanent properties of splicing system, in both aspects are theoretically and biologically investigated and their relations with non- semi simple splicing system restricted to two rules were proven via Y-G approach.

Fong (2008) conducted a wet-lab experiment up to stage two based on one initial string and two rules in order to generate the splicing languages via experiment as well as to differentiate between adult and limit languages. In this research, according to the theory of wet-lab that was proposed by Fong (2008) , the concept of splicing system is generalized up to stage two and its splicing languages are defined as two stages splicing languages.

1.3 PROBLEM STATEMENT

Splicing system was developed in different senses by mathematicians under framework of a formal language theory. As it is known the languages which are produced by splicing system is called splicing languages (Head, 1987). In real situation, these splicing languages represent the recombinant DNA molecules. DNA recombination is a part of gene manipulation, which can be achieved up to two stages reaction based on some necessities such as removing some parts of DNA sequences or generating the hybrid DNA molecules in the field of molecular biology. Thus, for this purpose conducting laboratory experiments for predicting the above two stages splicing languages as well as presenting its characteristics is time-consuming and needs a huge expenditure. Therefore, to overcome this problem, the characteristics of two stages DNA splicing languages with respect to the most two initial strings and rules must be considered through this research by providing some mathematical proofs. Hence, the following questions will be addressed and answered in this research.

1. What is the definition of two stages DNA splicing languages?
2. What are the relations between two stages (stage one and stage two) DNA splicing languages?
3. What are the sufficient conditions for persistency and permanency of two stages DNA splicing languages?
4. How to model the two stages DNA splicing languages based on crossing sites and contexts of restriction enzymes factors at the existing two initial strings and two rules on de Bruijn graph?
5. How to construct a user friendly interface to predict the persistency and permanency of two stages DNA splicing languages as well as the existence relations between them.

1.4 RESEARCH OBJECTIVES

This research embarks on the following objectives:

1. To introduce the concept of two stages in splicing system.

2. To present the relations between two stages (stage one and stage two) DNA splicing languages.
3. To provide some sufficient conditions for persistency and permanency of two stages DNA splicing languages, given as lemmas, theorems and corollaries.
4. To model the two stages DNA splicing languages based on crossing sites and contexts of restriction enzymes factors at the existing two initial strings and two rules on de Bruijn graph.
5. To construct a user friendly interface to predict the persistency and permanency of two stages DNA splicing languages as well as the existence relations between them.

1.5 RESEARCH SCOPE

This research will be focused on persistent and permanent characterizations of splicing languages mostly at the existence of two initial strings (with one or two cutting sites) and two rules at two stages via Y-G approach.

1.6 RESEARCH METHODOLOGY

This research is conducted according to six phases as follows:

Phase 1: Literature review

To study various concepts and theories regarding splicing systems, particularly Y-G splicing system and to see how this system works biologically. In addition, studying the persistent and permanent characteristics of splicing system, as well as splicing languages which have been introduced by previous researchers.

Phase 2: Presenting the concept of two stages in splicing system

To present the concepts of two stages in splicing system in both mathematical and biological perspectives. Furthermore, to provide some mathematical theorems and lemmas on showing the relations between two stages DNA splicing languages. Additionally, to predict the number of generating DNA splicing languages at stage one and stage two.

Phase 3: Providing some sufficient conditions for persistency and permanency of two stages DNA splicing languages.

To provide and prove some theorems, lemmas and corollaries which illustrate the sufficient conditions for persistency and permanency of two stages DNA splicing languages with respect to one initial string (with two cutting sites) and two rules, and two initial strings (with one or two cutting sites) and two rules.

Phases 4: Developing model using de Bruijn graph

To develop a model in order to represent the two stages DNA splicing languages on de Bruijn graph. Additionally, to show the persistency and permanency of two stages DNA splicing languages.

Phase 5: Constructing a user friendly interface

To construct a user friendly interface using Microsoft Visual C Sharp (C#) for predicting the persistency and permanency of two stages splicing languages as well as relations between two stages splicing languages.

Phase 6: Thesis writing and presentation

To write up, submit and present master thesis.

1.7 RESEARCH SIGNIFICANCE

Biomathematics, which studies the connectivity between biology and mathematics, is almost a new field in Malaysia. The findings of this investigation will contribute to both mathematicians and biologists. Furthermore, this research will have some benefits as categorized below.

1. New Findings/ Knowledge

The new concept, lemmas, theorems and corollaries on studying the characterization of two stages DNA splicing languages will be presented which may have some significance in the fields of mathematics and molecular biology.

2. Specific or Potential Applications

This research will benefit mathematical and bimolecular scientific communities in the world, particularly in Afghanistan, because this research is totally new in Afghanistan. The theoretical finding of this research will contribute to certain organization such as New England BioLab (NEB) to predict the persistency and permanency of the recombinant DNA molecules. Besides, the results obtained in this research also can be applied in a DNA recombination process for the formation of hybrid DNA molecules.

1.8 THESIS ORGANIZATION

This thesis is structurally organized into seven chapters. The first chapter discusses on general introduction for this research. It consists of an overview, research background, problem statement, research objectives, research scope, research methodology and research significance.

In Chapter 2, the molecular background related to splicing system is reviewed and DNA is structurally discussed. In addition, restriction enzymes and its behaviours on DNA molecules are included. Y-G splicing system, which is used in this research, is reviewed and also its biological aspect is discussed. This chapter ends with some historical reviews of splicing system and its properties and proper cases.

In Chapter 3, the concept of two stages splicing system is introduced and its biological and mathematical aspects are presented. In addition, the need for doing stages in splicing system with respect to recognition sites of initial DNA strand is justified and proven mathematically. In addition, some relation between two stages (stage one and stage two) splicing languages are provided, which show the set of stages one DNA splicing languages is a subset of stage two DNA splicing languages.

In Chapter 4, the persistent and permanent point of views of two stages DNA splicing languages are presented. Some sufficient conditions are provided based on cutting sites of initial strings and crossing and contexts of splicing rules, which show the persistency and permanency of two stages DNA splicing languages.

In Chapter 5, three different cases based on crossing sites properties of splicing rules are considered in order to generate the splicing languages. Then, the splicing languages which are produced by Y-G splicing system at two stages are described on de Bruijn graph. It is an important topic in the field of molecular biology, because the graph gives different paths of recombinant DNA molecules (splicing languages) as well as the persistency and permanency of splicing languages can be easily determined.

In chapter 6, based on the provided theorems and corollaries, a user friendly interface is constructed using Microsoft Visual C Sharp (C#). This software, which is replaced with wet-lab experiment in order to optimize time and money, is able to predict the persistency and permanency of two stages DNA splicing languages as well relations between two stages DNA splicing languages.

In Chapter 7, the summary of the whole thesis is presented. Furthermore, some suggestions and recommendations for future research are proposed.

In the next chapter, literature review related to this research is stated and discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Splicing system was first mathematically formulated by Head (1987) to model the recombination process of deoxyribonucleic acid (DNA) molecules when the DNA molecules were cut by restriction enzyme into fragments and by re-joining the fragments by ligase to form new DNA molecules. Therefore, this chapter reviews on the molecular background of splicing system, which includes the structure of DNA and restriction enzymes. Additionally, Yusof-Goode (Y-G) splicing system is reviewed and an example is provided to illustrate how this system works biologically. This chapter is ended by a historical review of splicing system.

2.2 MOLECULAR BACKGROUND OF SPLICING SYSTEM

In this section, some related biological concepts, which are a foundation on formulating of splicing system, are studied. There are the DNA, its structure and restriction enzyme which are discussed in the following two subsections.

2.2.1 The Structure of DNA

The fundamental chemical structure of DNA is a sequence of nucleotides, which consists of three main components: nitrogenous bases, phosphate group and deoxyribose sugar. The chemical structure of sugar consists of five carbons, which are numbered 1' to 5' so that the phosphate, base and hydroxyl group are bonded to 5', 1'

and 3' carbons, respectively. The single strand of DNA can be formed by joining nucleotides together when the $3'-OH$ of one nucleotide is attached to the 5' phosphate of other nucleotide to form a strong covalent bond named Phosphodiester (Weaver, 2012). DNA was structurally investigated by Watson and Crick where the DNA has form like a double helix so that the sugar and phosphate group are stacked outside and the bases are inside the double helix (Fu and Sven, 2009). The bases are categorized into purines (adenine and guanine) and pyrimidines (cytosine and thymine). In a DNA, adenine always bonds with thymine and guanine bonds with cytosine and vice-versa. A DNA molecule consists of two single strands joining together by hydrogen bonding between nucleotides bases. There are two hydrogen bonds between adenine and thymine and three hydrogen bonds between guanine and cytosine. These base pairings can be shown as [*A*/ *T*], [*G*/ *C*], [*C*/ *G*] and [*T*/ *A*]. The two bases adenine and guanine (purines) have two rings (double ring) in their chemical structures while cytosine and thymine (pyrimidines) have only single ring. Since the two strands of DNA are complemented of each other, therefore, the lower strand of DNA can be determined by its upper strand and vice-versa (Kratz, 2009). There are three different models of DNA structure named double helix, twisted ladder and space filling model as given in Figures 2.1 (a), (b) and (c), respectively. In the helical form of DNA, the helix represents the sugar-phosphate, while the stair represents the bases. In the form of a twisted ladder, the sides represent the sugar-phosphate, while the rungs represent the base-pairs. However, in the form of space filling, the chains of dark gray, red, light gray and yellow spares represent the sugar-phosphate, while the horizontal flat plates in the form of blue spheres show the base-pairs. These three models of DNA structure are shown in three different forms in Figure 2.1.

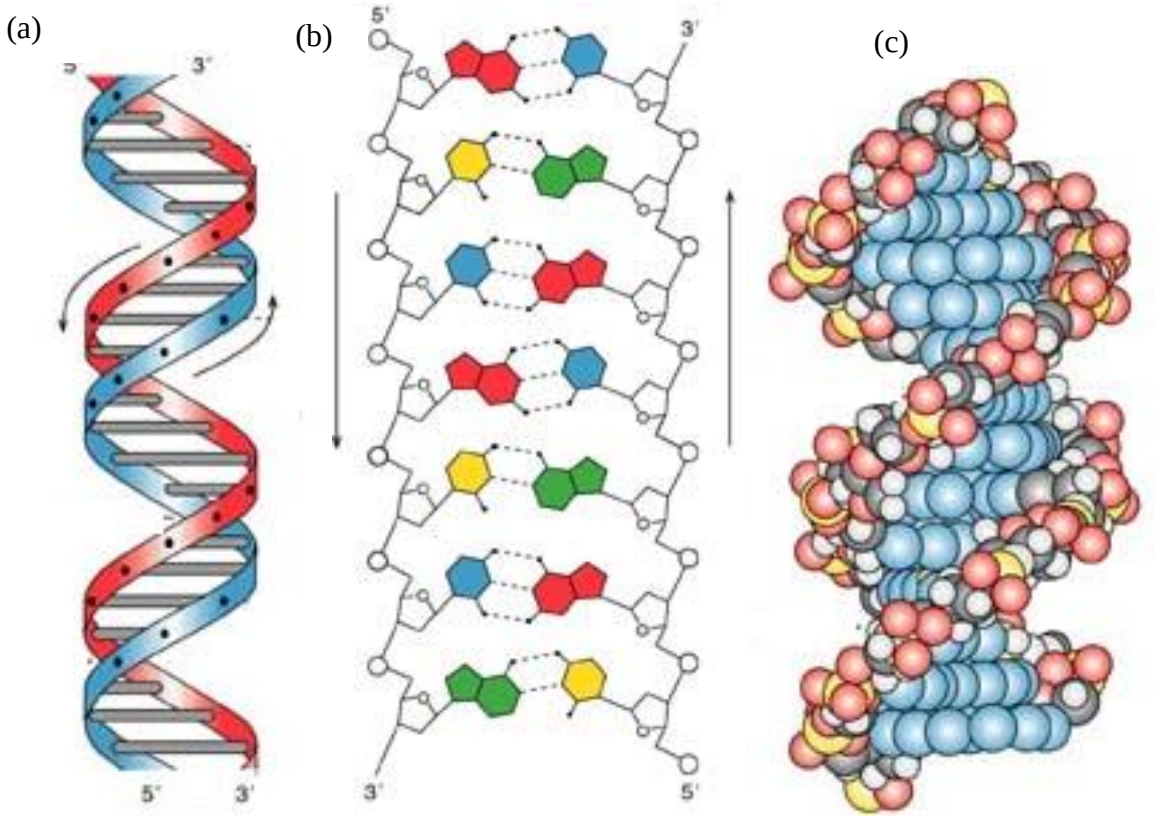

Figure 2.1: Three models of DNA structure (a) Double helix (b) Twisted Ladder (c) Space filling

Source: Weaver (2012)

Since DNA is a double-stranded molecule, the two strands of DNA are antiparallel and have two opposite directions as stated in Figure 2.1 (a) and (b). The standard convention of writing double stranded DNA (dsDNA) molecule is presented in the following so that the upper single strand form $5' - 3'$ direction and the lower single strand form $3' - 5'$ direction.

5′...*CGAGCTCG*...3′
3′...*GCTCGAGC*...5′

2.2.2 Restriction Enzyme (RE)

Restriction enzymes (RE) are enzymes that split the sugar-phosphate backbone of the double-stranded DNA (dsDNA). The vast majority of RE have been isolated from bacteria, where they accomplish a host-defence function for the cell. Naturally, restriction enzymes recognize specific DNA sequences usually 4 to 6 base-pairs in length and then cut them in a particular manner (sticky ends or blunt ends) (Walker and Rapley, 2009). A very important characteristic of sticky ends is being complementary of the fragments of DNA molecules that are cut by different restriction enzymes having identical crossing sites. Then the fragments of DNA re-join with their complementary ends by DNA ligase and forming new DNA molecules. There is a convention approach for naming the restriction enzymes according to genus, species, strain and order of discovery. Thus, the first letter of genus' name and the first two letter of species name are used to derive the basic enzyme name. For example, from *Escherichia coli* obtains (*Eco*) and the fourth letter is the particular strain *R* and finally a Roman number (I) is added to depict the order of discovery. Therefore, the first enzyme from *E. coli* strain *R* is written as *EcoR*I. During restriction, the restriction enzymes must split the strands of DNA to produce the dsDNA cut. For example, three different cutting sites of restrictions enzymes are given in the bellow.

$$\textit{EcoR}\text{I}: \begin{array}{l} 5'...G\blacktriangledown AATTC...3' \\ 3'...CTTAA\blacktriangle G...5' \end{array}$$

$$\textit{BstKT}\text{I}: \begin{array}{l} 5'...GAT\blacktriangledown C...3' \\ 3'...C\blacktriangle TAG...5' \end{array}$$

$$\textit{BsuR}\text{I}: \begin{array}{l} 5'...GG\blacktriangledown CC...3' \\ 3'...CC\blacktriangle GG...5' \end{array}$$

The first two restriction enzymes *EcoR*I and *BstKT*I produce sticky ends. Sticky ends contain of two different dsDNA cut, those produce $5'$ and $3'$ overhang, respectively. However, the restriction enzyme *BsuR*I produces a blunt end. The two symbols ▼ and

▲determine the cutting sites of restriction enzymes as mentioned in the above notations. Thus, a restriction enzyme consists of three main parts: left context, splicing site and right context. Therefore, the fragments of DNA molecules after cutting by restriction enzymes can only re-join if the restriction enzymes are having same pattern and identical crossing sites.

Since this research is based on Y-G approach, therefore this splicing system is viewed below.

2.3 AN OVERVIEW OF Y-G SPLICING SYSTEM

A mathematical model named Y-G splicing system associated with Y-G splicing rule was recently proposed and introduced by Yusof (2012) to investigate the biological process of DNA splicing in a translucent way. This new contribution which is considered as a useful model in the world of splicing system has been formulated from Head and Goode-Pixton splicing system with a slight innovation. Y-G model can be written mathematically as $S=(A,I,R)$, where A is the set of four alphabets, I is the set of initial string of double- stranded DNA and R is the set of rules. The rule R in this model is presented as $(u;x,v:y;x,z)$ and $(u,x;v:y,x;z)$, which shows the left pattern and right pattern of DNA molecules, respectively. However, the notation $(u,x,v:y,x,z)$ indicates that both patterns of rules are applied on DNA molecule.

Definition 2.1: Yusof-Goode (Y-G) Splicing System (Yusof, 2012)

Let $S=(A,I,R)$ is a Y-G splicing system. If $r=(u,x,v:y,x,z)\in R$ 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 produce the initial string I together with $\alpha uxz\delta$ and $\gamma yxv\beta$, where α,β,γ and δ are strings in the set free monoid, A^*. □

In the above definition, a free monoid is the set of strings obtained by concatenation operation of zero or more symbols from A and detonated A^*.

CHAPTER 1

INTRODUCTION

1.1 AN OVERVIEW

Manipulating operation of deoxyribonucleic acid (DNA) molecules is biochemically accomplished by a recombination process. The double-stranded DNA is cleaved by restriction enzyme into fragments; then the pasting operation can be achieved by the existence of appropriate ligase to form new DNA molecules. The study of the biological process of DNA splicing system in a translucent approach was investigated by Yusof (2012) under the framework of formal language theory. This novelty mathematical model which is known as Yusof-Goode (Y-G) splicing system is considered as a new contribution in terms of biology, particularly in the field of molecular DNA. According to its formulation, Y-G splicing system consists of three main sets: set of finite alphabets, set of initial strings and set of splicing rules. In this formalism, the DNA molecules, restriction enzymes and recombination process are considered as initial strings, splicing rules and splicing operation, respectively. The language which is generated by Y-G splicing system is called Y-G splicing language. In this thesis, the splicing languages that are produced by Y-G splicing system are defined as two stages splicing languages. Nevertheless, the existence relations between stage one and stage two splicing languages as well as the persistency and permanency of two stages splicing languages are investigated. Additionally, the de Bruijn graph of the above mentioned two stages DNA splicing languages are developed.

1.2 RESEARCH BACKGROUND

DNA is a double-stranded molecule consisting of two single strands over four alphabets *A*, *G*, *C*, and *T* which are abbreviated from adenine, guanine, cytosine and thymine, respectively. Based on Watson-Crick proposal, the two strands of DNA are antiparallel and have complementary property, where *A* is only paired with *T* and *C* is paired with *G* and vice-versa (Kratz, 2009).

To study the connectivity between recombination process and formal language theory, Head (1987) formulated splicing system mathematically based on the biological problem of cutting and pasting phenomenon. In addition in terms of mathematics the recombinant DNA molecules which will arise by the recombinant process are called splicing languages.

Yusof (2012) investigated on three classes of splicing system namely: non- semi simple splicing system, free splicing system and non-semi simple splicing system restricted to two rules and presented new definitions for these three types of splicing systems. Besides, a new symbolization representing the splicing system namely Y-G splicing system which is associated with Y-G rule was proposed. The language which is produced by Y-G splicing system is called Y-G splicing language. Moreover, the concepts of persistent and permanent properties of splicing system, in both aspects are theoretically and biologically investigated and their relations with non- semi simple splicing system restricted to two rules were proven via Y-G approach.

Fong (2008) conducted a wet-lab experiment up to stage two based on one initial string and two rules in order to generate the splicing languages via experiment as well as to differentiate between adult and limit languages. In this research, according to the theory of wet-lab that was proposed by Fong (2008) , the concept of splicing system is generalized up to stage two and its splicing languages are defined as two stages splicing languages.

1.3 PROBLEM STATEMENT

Splicing system was developed in different senses by mathematicians under framework of a formal language theory. As it is known the languages which are produced by splicing system is called splicing languages (Head, 1987). In real situation, these splicing languages represent the recombinant DNA molecules. DNA recombination is a part of gene manipulation, which can be achieved up to two stages reaction based on some necessities such as removing some parts of DNA sequences or generating the hybrid DNA molecules in the field of molecular biology. Thus, for this purpose conducting laboratory experiments for predicting the above two stages splicing languages as well as presenting its characteristics is time-consuming and needs a huge expenditure. Therefore, to overcome this problem, the characteristics of two stages DNA splicing languages with respect to the most two initial strings and rules must be considered through this research by providing some mathematical proofs. Hence, the following questions will be addressed and answered in this research.

1. What is the definition of two stages DNA splicing languages?
2. What are the relations between two stages (stage one and stage two) DNA splicing languages?
3. What are the sufficient conditions for persistency and permanency of two stages DNA splicing languages?
4. How to model the two stages DNA splicing languages based on crossing sites and contexts of restriction enzymes factors at the existing two initial strings and two rules on de Bruijn graph?
5. How to construct a user friendly interface to predict the persistency and permanency of two stages DNA splicing languages as well as the existence relations between them.

1.4 RESEARCH OBJECTIVES

This research embarks on the following objectives:

1. To introduce the concept of two stages in splicing system.

CHAPTER 3

TWO STAGES SPLICING SYSTEM

3.1 INTRODUCTION

In this chapter, the concept of splicing system is mathematically and biologically investigated and generalized up to two stages. The aim of this chapter is to introduce the concepts of splicing system at two stages as well as presenting the existence relations between the generated splicing languages at stage one and stage two. Therefore, to achieve this objective some mathematical theorems and lemmas are provided and proven. Furthermore, an example is provided to show the biological process of DNA splicing at two stages. First of all, some preliminaries related to this study are stated in the following section.

3.2 PRELIMINARIES

In this section, some definitions and concepts related to this research are presented. The definition of palindromic string which was introduced by Yusof (2012) is stated.

Definition 3.1: Palindromic (Yusof, 2012)

A string I of 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.□

For example, the dsDNA $\begin{matrix} 5'...CGAGCTCG...3' \\ 3'...GCTCGAGC...5' \end{matrix}$ is palindromic, since the sequence from left to the right side of the upper single strand is equal to the sequence of the lower single strand from right to the left side.

The rule in splicing system has three parts: left context, crossing site and right context. For example, in the rule of (a,x,b), where $a,x,b \in A^*$, the three strings a, x and b are called left context, crossing site and right context, respectively. Therefore, being disjointed the crossing of the splicing rules in a splicing system has an effect on the number of generating splicing languages as well as on persistency and permanency of splicing system and splicing languages which will be discussed in the next chapter. Hence, this concept is defined below.

Definition 3.2: Crossing Disjoint (Gatterdam, 1989)

A splicing system $S=(A, I, B, C)$ is a crossing disjoint if there exist no patterns (a,x,b) in B and (c,x,d) in C with same crossing x .□

For example, the splicing system $S=\left(\{a,g,c,t\}, I, B=\{(c,cg,g),(t,ta,a)\}, C=\varnothing\right)$ is crossing disjoint due to its rules which have different crossing sites.

In the next section the concept of two stages in splicing system is introduced and discussed.

3.3 THE CONCEPT OF TWO STAGES IN SPLICING SYSTEM

In real sense, sometimes it is not sufficient to manipulate the genome sequences to produce the new genome sequences only at stage one and it will need another manipulating operation to remove the parts of genome sequences and generate new paths of genomes. This work can be achieved by applying a restriction enzyme on the initial strand of DNA. The fragments of DNA molecules at the existence of ligase can be re-joined with their complementary ends to form new DNA molecules. For example, if there are two DNA molecules which represents two different plants with different colours, then after cutting DNA by restriction enzymes and re-ligating the fragments by

ligase, the hybrid DNA will be formed, which shows a plant with unique colour. Naturally, the recombinant DNA molecules can be split by the existence of RE to produce new DNA molecules at stage two. Therefore, generating new DNA molecules at stage two depend on the recognition sites of initial DNA strand. As a result, it is sufficient to achieve the manipulating action for a DNA strand by having one recognition site only at stage one, since it does not bring any changes to the sequence of DNA molecules at stage two. However, if the initial DNA strand has two recognition sequences, the manipulation process can be accomplished at the second stage, and new DNA molecules will be produced. Since running the wet-lab experiment up to *n*-stages ($n \geq 1$) can incur a lot of expenses and time consuming, therefore, the recombinant DNA strands of initial DNA molecules are predicted mathematically so that the biologists can use this mathematical consideration as their pre-processing or hypothesis of conducting wet-lab experiments.

Mathematically, a splicing system is the form $S = (A, I, R)$, where A is the set of four alphabets a, g, c and t, I is the set of initial strings of double- stranded DNA and R is the set of splicing rules that indicates the set of enzymatic operation. The languages that are produced by splicing system are called splicing languages. In other words, this set of splicing languages can be produced by splicing process (cutting and pasting) that is considered as stage one splicing languages. It consists of initial string(s) and all new splicing languages which can be resulted by splicing. The second- stage recombinant process among the resulted splicing languages at stage one, can also be carried out based on the above biological consideration. Therefore, the set of stage two splicing languages contain the set of stage one splicing languages. In this study, the concept of splicing language is generalized up to stage two based on Y-G splicing system. If $L = L(S)$ represents the set of splicing languages that will arise by splicing of the initial strings at the existence of appropriate rules and $L' = L'(S)$ represents the set of those splicing languages which are produced by splicing of the resulted splicing languages, then the union of these two sets of splicing languages is presented in the form of $L(S) \cup L'(S)$, which forms the two stages splicing languages. Hence, the process of DNA recombination up to stage two is considered by splicing operation and the splicing languages is defined as two stages splicing languages.