SINGLE-STAGE DNA SPLICING SYSTEM VIA YUSOF-GOODE APPROACH

LIM WEN LI

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Master of Science in Mathematics

Faculty of Industrial Sciences and Technology
UNIVERSITI MALAYSIA PAHANG

APRIL 2015

ABSTRACT

Yusof-Goode (Y-G) rule is the new translucent way representing the rule in splicing system under the framework of formal language theory. The motivation behind using the Y-G approach is to simulate the actual recombinant behaviours of deoxyribonucleic acid (DNA) molecules. The current laboratory experimental approaches to obtain the product of the DNA splicing process involve time and cost with high probability of not getting the desired results. Hence, in this research, a mathematical concept of single-stage splicing language restricted to at most two non-palindromic initial strings and two rules with one recognition site is introduced via Y-G model. In addition, the characteristics of rule are investigated by providing mathematical proofs. Based on the characteristics of rule, some theorems and lemmas have been formulated to predict the number types of single-stage splicing language of Y-G splicing system. Furthermore, in order to determine the characterizations of single-stage splicing language, a model is developed via limit adjacency matrix approach. The limit adjacency matrix can also be used to predict the number types of transient, inert persistent and active persistent limit language. Finally, a graphic user interface and a computational model of single-stage limit language are developed using Microsoft Visual Basic (VB) programming code focusing on prediction of the conceivable resulted molecules, the number types of single-stage splicing language as well as the behaviours of single-stage splicing language.

ABSTRAK

Peraturan Yusof-Goode (Y-G) adalah satu cara baru yang telus mempersembahkan peraturan dalam sistem hiris-cantum di bawah rangka kerja teori bahasa formal. Motivasi disebalik penggunaan pendekatan Y-G adalah untuk mensimulasikan kelakuan sebenar rekombinan molekul asid deoksiribonukleik (DNA). Pendekatan ujikaji makmal semasa untuk mendapatkan hasil proses hiris-cantum DNA melibatkan kekangan masa dan perbelanjaan dengan kemungkinan besar tidak dapat hasil yang diinginkan. Oleh itu, dalam kajian ini, satu konsep matematik bagi bahasa hiris-cantum peringkat tunggal terhad kepada paling banyak dua jujukan awal yang tidak palindromik dan dua peraturan dengan satu belah potongan telah diperkenalkan melalui model Y-G. Tambahan lagi, ciri-ciri peraturan telah dikaji dengan memberikan pembuktian matematik. Berdasarkan ciri-ciri peraturan, beberapa teorem dan lema telah diformulasikan untuk meramal bilangan jenis bahasa hiris-cantum peringkat tunggal dalam sistem hiris-cantum Y-G. Tambahan pula, untuk menentukan ciri-ciri bahasa hiris-cantum peringkat tunggal, satu model telah dibangunkan melalui pendekatan matriks had bersebelahan. Matrik had bersebelahan juga boleh digunakan untuk meramal bilangan jenis bahasa sementara, bahasa lengai dan bahasa berterusan aktif. Akhir sekali, antara muka pengguna grafik dan model komputasi bagi bahasa had peringkat tunggal telah dibangunkan menggunakan kod pengaturcaraan Microsoft Visual Basic (VB) dengan memberi tumpuan kepada ramalan molekul-molekul terhasil, bilangan jenis bahasa hiris-cantum peringkat tunggal dan juga kelakuan bahasa hiris-cantum peringkat tunggal.

TABLE OF CONTENTS

	Page
SUPERVISOR'S DECLARATION	ii
STUDENT'S DECLARATION	iii
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
ABSTRAK	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF SYMBOLS	xiv
LIST OF ABBREVIATIONS	xvi

CHAPTER 1 INTRODUCTION

1.1	An Overview	1
1.2	Research Background	2
1.3	Problem Statement	3
1.4	Research Objectives	4
1.5	Research Scope	4
1.6	Research Significance	5
1.7	Research Methodology	5
1.8	Thesis Organization	6

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction 8

2.2 The Basic Concepts of DNA Recombinant Behaviour 8

2.2.1 DNA and Its Structure 8
2.2.2 Restriction Enzyme 10
2.2.3 The DNA Splicing Process 11

2.3 The Linkage of Mathematics and Biology 12

2.3.1 The Development of Splicing System and Language 13

2.3.1.1 Mathematical Approach 13
2.3.1.2 Biological Approach 17
2.3.1.3 Relationship between the Mathematical and Biological Approach 18

2.4 Limit Language 19

2.5 Conclusion 20

CHAPTER 3 SINGLE-STAGE SPLICING LANGUAGE

3.1 Introduction 21

3.2 Non Semi-Null Splicing System 21

3.3 The Concept of Single-Stage in Splicing Language 22

3.4 Prediction of Single-Stage Splicing Language 24

3.4.1 Preliminary 25
3.4.2 One String 26
3.4.3 Two Strings 33

3.5 Conclusion 52

CHAPTER 4 THE CHARACTERIZATION OF RESTRICTION ENZYME

4.1 Introduction 53

4.2 Preliminaries 53

4.3	Concept of Inverse Complement		54
4.4	Some Characteristics of Rules		55
	4.4.1	Palindromic Rules	55
	4.4.2	Inverse Complement Rules	56
4.5	Prediction of Single-Stage Splicing Language involving Inverse Complement		58
4.6	Conclusion		71

CHAPTER 5 SINGLE-STAGE LIMIT LANGUAGE

5.1	Introduction	73
5.2	Concept of Single-Stage Limit Language	73
5.3	Prediction of Single-Stage Limit Language involving One String	75
5.4	Subset of Single-Stage Limit Language	80
5.5	Conclusion	81

CHAPTER 6 THE BEHAVIOUR OF SINGLE-STAGE SPLICING LANGUAGE

6.1	Introduction	83
6.2	Preliminaries	84
6.3	Limit Adjacency Matrix	84
6.4	Some Properties of Limit Adjacency Matrix	85
6.5	Biological Examples of Splicing Process	88
6.6	Conclusion	91

CHAPTER 7 COMPUTATIONAL MODELLING AND GRAPHIC USER INTERFACE (GUI)

7.1 Introduction 92

7.2 Theorems and Assumptions 92

7.3 Pre-processing and Importing into Visual Basic 93

7.3.1 Input Parameters 93
7.3.2 Computing Splicing Language 96
7.3.3 Analyzing the Behaviour of Splicing Language 96

7.4 Results and Discussions 98

7.5 Conclusion 101

CHAPTER 8 CONCLUSION AND RECOMMENDATIONS

8.1 Conclusion 103

8.2 Recommendations for Future Research 105

REFERENCES 106

APPENDICES 110

A Main Frame of Computational Modelling using Visual Basic 110

B List of Publications/Presentations in Conferences 113

LIST OF TABLES

Table No.	Title	Page
5.1	The prediction of single-stage limit language with one initial string and one rule based on the properties of crossing site	81
7.1	Parameters used for generating the number of splicing language	94
7.2	Comparison of the results from wet lab experiments and software	98
7.3	A comparison of the number of splicing language generated by software and theorems	101

LIST OF FIGURES

Figure No.	Title	Page
2.1	Structure of DNA	9
2.2	Blunt ends and sticky ends	12
2.3	Relationships between family of regular language generated from Head, Paun and Pixton splicing systems	14
2.4	Relation between four types of splicing system	17
5.1	The relationship between single-stage splicing language, single-stage limit language, active persistent language and inert persistent language	74
5.2	Directed graph of generated splicing language for non-palindromic rule	76
5.3	Directed graph of generated splicing language for palindromic rule	78
5.4	Directed graph of generated splicing language for non-palindromic rule with palindromic crossing site	79
7.1	Output of graphic user interface (GUI)	94
7.2	Flowchart on generating number patterns of splicing language	95
7.3	Flowchart on determining the behaviour of splicing language	97
7.4	Results obtained from Visual Basic: Modelling the behaviour of splicing language with two DNA and two restriction enzymes, *Bgl*I and *Dra*III	99
7.5	Results obtained from Visual Basic: Modelling the behaviour of splicing language with one DNA and one restriction enzyme, *Aci*I	100

LIST OF SYMBOLS

a	The base pairing of adenine and thymine
c	The base pairing of cytosine and guanine
g	The base pairing of guanine and cytosine
iff	If and only if
$n(L_1(S))$	The number types of single-stage splicing language
t	The base pairing of thymine and adenine
A	Set of four alphabets
A'	Complement of A
A^{-1}	inverse on A in direction
$A*$	Strings obtained by concatenating operation of one or more symbols from A
A^+	Set of strings that consists of zero or more symbols concatenated from A
A_{ij}^{∞}	Limit Adjacency Matrix
I	Set of initial strings
$L(S)$	Splicing Language generated by splicing system at single-stage
$L_A(S)$	Inert persistent language generated by splicing system at single-stage
$L_\infty(S)$	Limit language generated by splicing system at single-stage
R	Set of rules
S	Splicing system
$\mathbb{Z}^+$	Positive integer

$\alpha, \beta, \gamma, \delta$	Strings in A^*
■	End of proof/example
□	End of definition/theorem/lemma/corollary
▼... ...▲	Crossing of recognition site of restriction enzymes
$\cup$	Union
$\in$	Elements of
{ }	Set
$\neq$	Not equal to
$\ni$	Such that
$\exists$	Exist
$\nexists$	Does not exist
$\leq$	Less or equal to
$\geq$	Greater or equal to
$\forall$	For all
$\subset$	Proper subset
$\subseteq$	Subset or equals to
$\mapsto^r$	Derived by splicing using r with Y-G approach

LIST OF ABBREVIATIONS

bp	base pair
DNA	Deoxyribonucleic acid
dsDNA	Double- stranded DNA
RE	Restriction enzyme
Y-G	Yusof-Goode
NEB	New England Biolab

CHAPTER 1

INTRODUCTION

1.1 AN OVERVIEW

Framework of Formal Language Theory, a branch of Applied Discrete Mathematics and Theoretical Computer Science illustrated the formalism of Y-G splicing system to model the recombinant behaviours of deoxyribonucleic acid (DNA) molecules under the existence of restriction enzymes and ligase based upon the characteristics of restriction enzyme itself. Y-G splicing system is made up of a finite set of initial strings and a finite set of rules over an alphabet. The left-pattern and right-pattern of these rules can be determined clearly in Y-G splicing system and therefore it resembles the translucent behaviour of DNA biological process.

The finite set of initial strings can be spliced by the finite set of rules forming new strings. These new strings, which is the new set of dsDNA molecules that arise with the existence of specified enzyme activities is represented as a splicing language over the four-symbol alphabets of deoxyribonucleotide pairs. Meanwhile, active persistent limit language, inert persistent limit language and transient language have been determined as the subset of splicing language.

This thesis focuses on introducing a new concept of single-stage splicing language and limit language in non semi-null splicing system with respects to at most two non-palindromic initial strings and two rules with one cutting site, presenting new definition namely inverse complement where splicing language can be predicted based on its

properties. Besides, the limit adjacency matrix is introduced to model the existence of single-stage limit language from splicing language. Lastly, a software programming on modelling the behaviour of splicing language in terms of active persistent, inert persistent and transient has been developed based on Y-G splicing system.

1.2 RESEARCH BACKGROUND

The DNA molecules have four distinct bases, namely adenine (*A*), guanine (*G*), thymine (*T*) and cytosine (*C*) where the bases are classified into purines (A and G) and pyrimidines (C and T) (Dwyer and Lebeck, 2008). The cut and paste phenomenon of two double-stranded DNA (dsDNA) molecules act on the bases based on Watson-Crick complementary (Lamm and Unger, 2011), where *A* hydrogen bonds to *T*, *G* hydrogen bonds to *C*, *C* hydrogen bonds to *G* and *T* hydrogen bonds to *A*. These rules of pairing can be denoted as [*A*/*T*], [*G*/*C*], [*C*/*G*] and [*T*/*A*] respectively. Restriction enzyme is an enzyme found in bacteria that is usually used in lab experiments. It can cut the DNA molecules at recognition sites, producing molecules with sticky or blunt ends (Weaver, 2012). These molecules will then undergo recombination reaction.

Head (1987) was the first person who initiated the modelling of this biological recombinant behaviour of DNA in a mathematical abstraction, namely splicing system to decode the language of the biological behaviour. Splicing system can be classified into many classes, which are simple, semi-simple, semi-null and non semi-simple splicing system.

Since DNA strings can only involve in non semi-simple splicing language, Yusof (2012) introduced the concept of non semi-simple splicing system restricted to two rules. Besides, Y-G notation was introduced as a new notation of writing rules in splicing system which is associated with Y-G splicing system. This new splicing system was based upon the characteristics of the restriction enzyme itself, with some modifications from Head's and Pixton's splicing systems. It has been proven that Y-G rule can replace Pixton rule for any given initial set *I*.

In this study, the research on Y-G splicing system is extended by focusing on non semi-null splicing system with at most two initial strings, one cutting site and two rules in single-stage. Subsequently, the characteristics and behaviours of the resulting strings from Y-G splicing system are investigated.

1.3 PROBLEM STATEMENT

Splicing system underpins a vast array of recombinant DNA technologies. It plays a pivotal role in attempts to recombine sets of double-stranded DNA molecules when acted on by restriction enzymes and a ligase, which is currently estimated to cost around USD$300 per restriction endonuclease if to conduct laboratory experiments (NEB, 2014). By focusing on this problem, the objective of this research is to develop a mathematical and computational model in predicting splicing language, which is the language produced by splicing system, based upon the context and crossing of rules with at most two rules and two non-palindromic initial strings with one recognition site in a non semi-null splicing system. In particular, this study is to seek methods that are computationally feasible even for limit language, which are the molecules that will be present on the system after the reaction is complete. Moreover, the following questions will be addressed and answered:

(i) What is non semi-null splicing system?
(ii) What is the definition of single-stage splicing and limit language?
(iii) What are the conditions and methods to predict the number types of single-stage splicing and limit language?
(iv) What is the characteristics in terms of transient, active persistent, inert persistent of single-stage splicing language?
(v) How to develop a mathematical and computational model that can validate the existence of single-stage limit language based on crossing and contexts of restriction enzymes factors in order to optimize time and money?

1.4 RESEARCH OBJECTIVES

The aims of this research are as follows:

(i) To investigate a non semi-null splicing system restricted to two rules.

(ii) To introduce a new concept of single-stage splicing and limit language and their characteristics.

(iii) To predict the single-stage splicing and limit language based on crossing and context of restriction enzymes factors of at most two non-palindromic initial strings and two rules with one recognition site by providing mathematical proofs in terms of theorems, corollaries and counterexamples.

(iv) To develop a mathematical model that can validate the existence of single-stage limit language to optimize money and time via limit adjacency matrix.

(v) To construct a computational model on the behaviour of single-stage splicing language and also graphical user interface for the theorems formulated via programming.

1.5 RESEARCH SCOPE

This research involves the study of non semi-null splicing system restricted to at most two non-palindromic DNAs, two same patterns of restricted enzymes and one recognition site in each DNA. Y-G splicing systems, which can model the recombinant action of restriction enzymes and a ligase on DNA molecules are used to predict the different types of splicing languages that can result from a single-stage splicing system. New concepts related to inverse complement rules, single-stage splicing language and single-stage limit language are also introduced and theorems, lemmas, propositions and corollaries related to them are proved. This research also includes a mathematical model and a computational model to simulate the behaviour of single-stage splicing language.

1.6 RESEARCH SIGNIFICANCE

As a research area with promising potential to biomathematical field, this study will provide significant contribution to the mathematical and biomolecular scientist communities in terms of:

(i) New findings: New theories in the single-stage splicing and limit language where the existence and the characteristics are investigated by the factor of crossing sites, left and right context of the rules and others are discovered. Besides, new theorems on the characterization of single-stage splicing language are formulated. Also, new knowledge on the mathematical and computational models from theories developed is obtained.

(ii) Specific or potential application: New techniques of DNA splicing system are beneficial to restriction enzyme companies such as New England Biolabs (NEB), plant research institutes and Ministry of Agriculture. Not only the experimental outcomes of plant DNA recombination are more predictable, the cost and time spent on the research can be reduced as well.

1.7 RESEARCH METHODOLOGY

This research is first carried out theoretically by introducing the concept of non semi-null splicing system, single-stage splicing language and a new rule factors namely inverse complement. Some propositions and examples are given to illustrate these concepts. Next, theorems are stated with their proofs regarding to the predictions on single-stage splicing language in Y-G splicing systems. The concept of single-stage limit language are also introduced. Then, by investigating the concept of limit graph, a mathematical model is developed to simulate single-stage splicing language in terms of their behaviour such as inert persistent limit language, transient language and active persistent limit language by using limit adjacency matrix. Lastly, a computational model of single-stage limit language is constructed based on the limit adjacency matrix model and the algorithm in Visual Basic

programming is coded to stipulate the number types of inert persistent language, transient language and active persistent language.

1.8 THESIS ORGANIZATION

This thesis comprises of eight chapters all together. The overview of each chapter is as follows:

Chapter 1: Introduction

The first chapter introduce the research background, problem statement, research objectives, research scope and research significance. Research methodology employed in conducting this research is also included. Lastly, the overview of each chapter is given.

Chapter 2: Literature review

The second chapter consists of the literature review of the dsDNA structure and the characteristic of restriction enzymes. This chapter also discusses on DNA splicing process, followed by the linkage of mathematics and biology. This chapter ends with the development of splicing system, limit language and limit graph.

Chapter 3: Single-stage splicing language

The concept of non semi-null splicing system is introduced. New definition of single-stage splicing language applying on non semi-null splicing system is given. Predictions of single-stage splicing language with at most two non-palindromic DNA, two palindromic crossing restriction enzymes are presented as theorems.

Chapter 4: The characterization of restriction enzymes

Chapter four starts off with a new characteristic in restriction enzymes named as inverse complement. Besides, some theorems are given regarding the role of inverse complement in predicting the DNA splicing language in Y-G splicing system restricted to two non-palindromic crossing rules and also the relationship of palindromic and inverse complement. Some examples are included.

Chapter 5: Single-stage limit language

The concept of single-stage limit language is given. This chapter focuses on predicting the number types of single-stage limit language involving one non-palindromic string based on the crossing site of rules.

Chapter 6: The behaviour of single-stage splicing language

Limit adjacency matrix is constructed to form single-stage limit language and its behaviour. Theorems of the limit adjacency matrix are proven in this chapter.

Chapter 7: Computational modelling and graphic user interface (GUI)

Computational models of single-stage limit language and its behaviour in terms of inert persistent, transient, active persistent language and non trivial splicing are developed. The similarities as well as the differences between the computational model, theorems in Chapter 4 and some wet lab experiments that have been done are analysed.

Chapter 8: Conclusion and Recommendations

This chapter summarizes the thesis. It provides conclusion to the entire research, as well as some suggestion on further works that can be done.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

The structure of DNA, restriction enzyme and the relationship between them and Formal Language Theory are discussed mathematically and biologically in this chapter. In addition, the development of splicing systems and language through variants and classes of splicing system, biological approach and splicing language are presented. Lastly, Y-G model is reviewed.

2.2 THE BASIC CONCEPTS OF DNA RECOMBINANT BEHAVIOUR

This section consists of three sub-sections, which are a brief overview of DNA and its structure, the characteristics of restriction enzyme and the DNA splicing process.

2.2.1 DNA and Its Structure

DNA is the units of inheritance that pass on information in terms of genetic substance from parents to offspring. In terms of biology, DNA is a polymer that belongs in the class of compounds known as nucleic acids, which are the macromolecules that exist as polymers called polynucleotide. A nucleotide consists of three parts: a nitrogenous base, a five-carbon sugar and a phosphate group as shown in Figure 2.1.

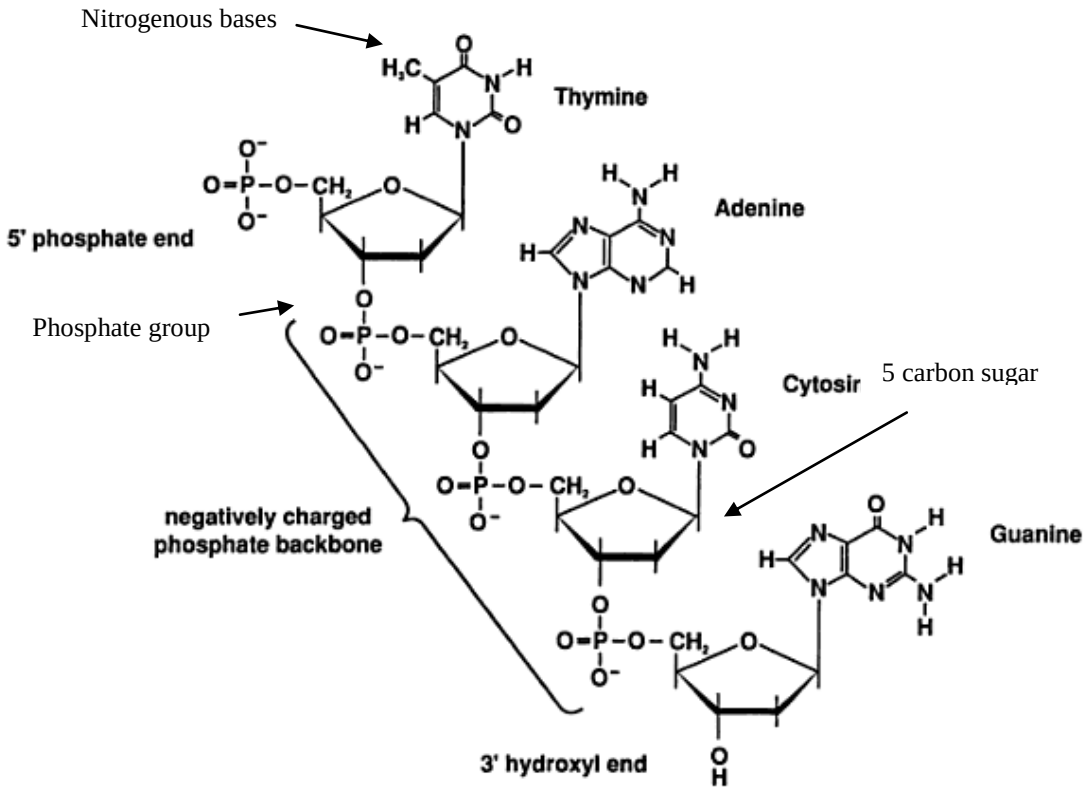

Figure 2.1: Structure of DNA

Source: Sinden 1994

Deoxyribose in DNA is the sugar connected to the nitrogenous base. It has 5 carbon atoms where the phosphate group is attached at 5’ carbon, base is attached to the 1’ carbon and for sugar structure, there is a hydroxyl group (OH) that is attached to the 3’ carbon (Weaver, 2012).

There are two groups of nitrogenous bases, namely pyrimidines and purines. The constituents of the pyrimidine group are cytosine (*C*) and thymine (*T*). The purines are adenine (*A*) and guanine (*G*). The two polynucleotides are connected together by hydrogen bonds between the paired bases. Cellular DNA molecules consist of two polynucleotides that spiral around an imaginary axis, forming a double helix. The molecular double-helix shape is formed by the two sugar phosphate backbones that run in anti-parallel arrangement.

The sugar-phosphate backbones are located externally of the helix, while the nitrogenous bases are paired internally of the helix. These two strands in the double helical structure are complementary and this makes the precise copying of genes responsible for inheritance possible (Fu and Sven, 2009).

Watson and Crick (1953) proposed only certain bases in the double helix which are compatible with each other. Adenine (*A*) always connected with thymine (*T*), and guanine (*G*) always connected with cytosine (*C*). *C*-*G* pairing is much stronger compare to *A*-*T* pairing because *C*-*G* pairing has three hydrogen bonds while *A*-*T* pairing has two hydrogen bonds between two nucleotides. These rules of pairing can be written and denoted as [*A**T*], [*G**C*], [*C**G*] and [*T**A*] (Lamm and Unger, 2011).

There are seven operations on DNA molecules. It consists of DNA strand recombination, DNA elongation, DNA shortening, DNA cutting, DNA linking, modification of nucleotides of DNA, and DNA multiplication (Weaver, 2012).

One of the special features of DNA sequences is a palindrome. A palindrome is a sequence of reversed repeats where reading from the 5' to 3' strand matches the reading from 5' to 3' on the complementary strand. Palindromes come in different length and it is often up to 50 bp (Bryce et. al., 2004). Only non-palindromic initial strands of dsDNA are considered in this thesis since the initial strands chosen for laboratory experiments are usually more than 50 bp long.

Next in the following sub-section, the characteristics of restriction enzyme is discussed.

2.2.2 Restriction Enzyme

Restriction enzyme is a type of enzyme that recognizes and cuts DNA molecules foreign to a bacterium. They only cuts at specific nucleotide sequences known as crossing sites or restriction sites and thus it can generate specific DNA fragments (Walker and

Rapley, 2009). This discovery enables scientists to identify, isolate and study specific DNA sequences (Klug et al., 2011). Hence, to produce DNA molecules that are not found together in nature, restriction enzyme allows researchers to choose specific cleavage site in a single-copy sequence that present in the DNA molecules during a lab experiment. Most restriction enzymes recognize sequences containing four to eight nucleotides. The frequency of occurrence of restriction sites in DNA depends on nucleotides. Based on the probability principles, the frequency of a short nucleotide pair sequence is greater than the frequency of a long nucleotide pair sequence (Russell, 2006). However, the number of cuts in DNA of an organism by a particular enzyme can be limited. Small plasmids contain only a single cleavage site for restriction enzyme. A certain combination of four bases can only occur in a random manner once every few hundred bases. For every few thousand bases, only a specific sequence of six bases will occur randomly. Hence it is possible that a DNA molecule contains no restriction site for a given enzyme (Hartl, 2011). Due to the reasons above, only one recognition site in each DNA is considered in this thesis.

Lastly, the process of DNA recombination with the existence of restriction enzymes, ligase and DNA molecules is presented in the next sub-section.

2.2.3 The DNA Splicing Process

DNA molecules can be cut by some restriction enzymes from their phosphodiester bond or at cutting sites, yielding molecules with sticky or blunt ends (Tamarin, 2001) as shown in Figure 2.2. After that, new molecules are produced when the molecules with sticky and blunt ends are pasted together by a specific type of enzyme, ligase, by catalyzing the formation of a phosphodiester bond (Russell, 2006).

Only sticky ends are considered for this study since any blunt ended DNA can be ligated to any other blunt ended DNA regardless of the molecular sequence, hence producing infinite splicing language.

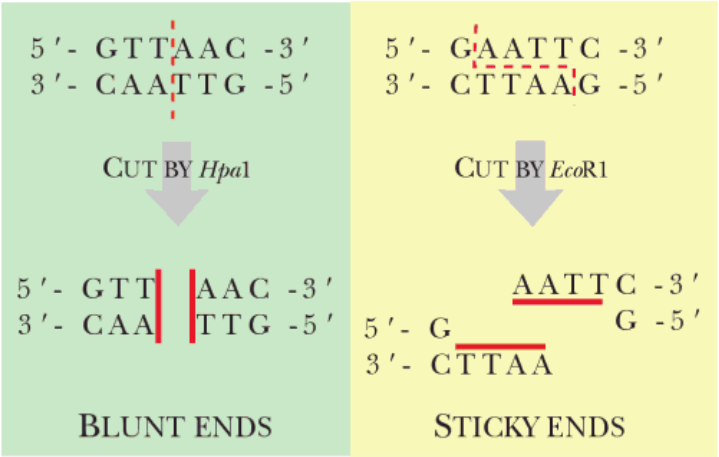

Figure 2.2: Blunt ends and sticky ends

Source: Clark and Pazdernik 2011

In reality, sticky ends can be ligated much more readily than blunt ended fragments, presumably due to hybridization between the single stranded regions connect the fragments together in the proper position for ligation (Hartl, 2011). Also, in terms of cloning, restriction enzymes that produce sticky ends are more valuable compared to blunt ends because every DNA fragment produced by cutting a segment of DNA with the same restriction enzyme has the similar base sequence at the two sticky ends (Russell, 2006).

Due to the complicated characteristic of DNA recombinant process, there is a need to represent them in terms of mathematical notation. Hence, in the next section, the relationship between biology and mathematics is shown.

2.3 THE LINKAGE OF MATHEMATICS AND BIOLOGY

In the context of formal language theory, four bases of DNA molecules which are known as *A*, *G*, *C* and *T* bases, enzymatic operation and recombinant behaviour are presented as initial set of strings, alphabet, rules and splicing language, respectively. The

CHAPTER 1

INTRODUCTION

1.1 AN OVERVIEW

Framework of Formal Language Theory, a branch of Applied Discrete Mathematics and Theoretical Computer Science illustrated the formalism of Y-G splicing system to model the recombinant behaviours of deoxyribonucleic acid (DNA) molecules under the existence of restriction enzymes and ligase based upon the characteristics of restriction enzyme itself. Y-G splicing system is made up of a finite set of initial strings and a finite set of rules over an alphabet. The left-pattern and right-pattern of these rules can be determined clearly in Y-G splicing system and therefore it resembles the translucent behaviour of DNA biological process.

The finite set of initial strings can be spliced by the finite set of rules forming new strings. These new strings, which is the new set of dsDNA molecules that arise with the existence of specified enzyme activities is represented as a splicing language over the four-symbol alphabets of deoxyribonucleotide pairs. Meanwhile, active persistent limit language, inert persistent limit language and transient language have been determined as the subset of splicing language.

This thesis focuses on introducing a new concept of single-stage splicing language and limit language in non semi-null splicing system with respects to at most two non-palindromic initial strings and two rules with one cutting site, presenting new definition namely inverse complement where splicing language can be predicted based on its

properties. Besides, the limit adjacency matrix is introduced to model the existence of single-stage limit language from splicing language. Lastly, a software programming on modelling the behaviour of splicing language in terms of active persistent, inert persistent and transient has been developed based on Y-G splicing system.

1.2 RESEARCH BACKGROUND

The DNA molecules have four distinct bases, namely adenine (*A*), guanine (*G*), thymine (*T*) and cytosine (*C*) where the bases are classified into purines (A and G) and pyrimidines (C and T) (Dwyer and Lebeck, 2008). The cut and paste phenomenon of two double-stranded DNA (dsDNA) molecules act on the bases based on Watson-Crick complementary (Lamm and Unger, 2011), where *A* hydrogen bonds to *T*, *G* hydrogen bonds to *C*, *C* hydrogen bonds to *G* and *T* hydrogen bonds to *A*. These rules of pairing can be denoted as [*A*/*T*], [*G*/*C*], [*C*/*G*] and [*T*/*A*] respectively. Restriction enzyme is an enzyme found in bacteria that is usually used in lab experiments. It can cut the DNA molecules at recognition sites, producing molecules with sticky or blunt ends (Weaver, 2012). These molecules will then undergo recombination reaction.

Head (1987) was the first person who initiated the modelling of this biological recombinant behaviour of DNA in a mathematical abstraction, namely splicing system to decode the language of the biological behaviour. Splicing system can be classified into many classes, which are simple, semi-simple, semi-null and non semi-simple splicing system.

Since DNA strings can only involve in non semi-simple splicing language, Yusof (2012) introduced the concept of non semi-simple splicing system restricted to two rules. Besides, Y-G notation was introduced as a new notation of writing rules in splicing system which is associated with Y-G splicing system. This new splicing system was based upon the characteristics of the restriction enzyme itself, with some modifications from Head's and Pixton's splicing systems. It has been proven that Y-G rule can replace Pixton rule for any given initial set *I*.

In this study, the research on Y-G splicing system is extended by focusing on non semi-null splicing system with at most two initial strings, one cutting site and two rules in single-stage. Subsequently, the characteristics and behaviours of the resulting strings from Y-G splicing system are investigated.

1.3 PROBLEM STATEMENT

Splicing system underpins a vast array of recombinant DNA technologies. It plays a pivotal role in attempts to recombine sets of double-stranded DNA molecules when acted on by restriction enzymes and a ligase, which is currently estimated to cost around USD$300 per restriction endonuclease if to conduct laboratory experiments (NEB, 2014). By focusing on this problem, the objective of this research is to develop a mathematical and computational model in predicting splicing language, which is the language produced by splicing system, based upon the context and crossing of rules with at most two rules and two non-palindromic initial strings with one recognition site in a non semi-null splicing system. In particular, this study is to seek methods that are computationally feasible even for limit language, which are the molecules that will be present on the system after the reaction is complete. Moreover, the following questions will be addressed and answered:

(i) What is non semi-null splicing system?
(ii) What is the definition of single-stage splicing and limit language?
(iii) What are the conditions and methods to predict the number types of single-stage splicing and limit language?
(iv) What is the characteristics in terms of transient, active persistent, inert persistent of single-stage splicing language?
(v) How to develop a mathematical and computational model that can validate the existence of single-stage limit language based on crossing and contexts of restriction enzymes factors in order to optimize time and money?

CHAPTER 3

SINGLE-STAGE SPLICING LANGUAGE

3.1 INTRODUCTION

First and foremost, a non semi-null splicing system that will be used throughout this study is introduced. Besides that, some concepts of single-stage splicing language applying on non semi-null splicing system are given throughout this chapter since when all the restriction enzymes, dsDNA strings and ligases take place simultaneously in a single buffer can optimize the time and money during an experiment with minimum difference in the resulted strings generated. The productions of Y-G splicing system, namely splicing language involving at most two non-palindromic strings and two palindromic rules are predicted based on the theorems formulated.

3.2 NON SEMI-NULL SPLICING SYSTEM

The scope of research for this thesis is restricted to non semi-null splicing system since in semi-null splicing system where a rule has left or right null context can produce splicing language consisting of two recognition sites. Besides, blunt ended rules that involved in a semi-null splicing system are not within the research scope. Furthermore, a rule having a single letter as a crossing site rarely exists in reality. Hence, non semi-null splicing system is investigated in this study. Since Y-G splicing system is used throughout the research, semi-null splicing system from Definition 2.4 is redefined in Y-G notation as follows:

Definition 3.1: (Yusof, 2012) Semi-Null Splicing System

Let $S = (A, I, R)$ be a Y-G splicing system in which I and R are finite and every rule in R has the form $(u,1,1:v,1,1)$, $(1,w,u:1,w,v)$ or $(u,w,1:v,w,1)$ where u, v and w are in A^+. Thus, $S = (A, I, R)$ is a semi-null splicing system. □

Consequently, from Definition 3.1, if a Y-G splicing system $S = (A, I, R)$ is not in the form of semi-null splicing system, that splicing system is called a non semi-null splicing system.

Recombinant dsDNA molecules are a sequence of molecules that does not exist in nature which has been formed by laboratory methods. Many laboratory experiments have been conducted to verify the model of splicing system in either one stage or two stages. In a splicing system with one cutting site, there is minimal difference in between resulted molecules generated from one stage and two stages. Thus, in the next section, single-stage splicing language is introduced to save time and cost.

3.3 THE CONCEPT OF SINGLE-STAGE SPLICING LANGUAGE

The behaviour of recombinant DNA suggests that suppose there is a finite set of DNA molecules and a finite set of restriction enzymes where DNA molecules and restriction enzymes that are given or produced are always obtainable. Instead of running the laboratory experiment stage by stage, what are the molecules that can potentially appear by letting all DNA molecules, restriction enzymes and ligases act simultaneously in a test-tube environment in order to save cost and money? Hence, in this section, the existing definition of splicing language introduced by Head (1987) is further defined specifically as single-stage splicing language to model the set of all molecule types that arise with restriction enzymes, dsDNAstrings and ligases all act in a single buffer.

Definition 3.2: Single-stage Splicing Language

Let $S=(A,I,R)$ be a Y-G splicing system, where A is a set of alphabets a, g, c and t, I is a set of initial strings and R is a set of rules, then a single-stage splicing language $L_1 = L_1(S)$ can be expressed as follows:

$$L_1 = L_1(S) = \{R_p + I_q \mid 1 \le p \le n,\ 1 \le q \le n,\ \forall p,\ q = 1,2,3,...,n\} \tag{3.1}$$

where $R_p = \{r_p\}$ represents a set of rules and $I_q = \{s_q\}$ represents a set of initial strings. □

From Eq. (3.1), if $p=1$ or $q=1$, then $R_1 = \{r_1\}$ and $I_1 = \{I_1\}$. Meanwhile, if $p=2$ or $q=2$, then $R_2 = \{r_1, r_2\}$ and $I_2 = \{I_1, I_2\}$. The same goes to $p=n$ or $q=n$, where we have $R_n = \{r_1, r_2, ..., r_n\}$ and $I_n = \{I_1, I_2, ..., I_n\}$. Despite the number of p and q, single-stage splicing language is the output of splicing system consisting of all DNA molecules with the action of all restriction enzymes in one single buffer.

Fong (2008) has conducted a wet-lab experiment by using *Aci*I and *Hpa*II in two stages. In the experiment, PCR generates thousand copies of $\alpha - \beta$ strands, *I*. OneTaq *GC* Reaction Buffer and OneTaq Standard Reaction Buffer are chosen to be used as the reaction buffer for DNA strands based on the percentage of *G* and *C* bases in the forward and reverse primer. The molecules for this experiment resulting from both stages areas follows:

$$\alpha\cdots CCGC\cdots\beta \qquad \alpha\cdots CCGG\cdots\alpha' \qquad \beta'\cdots GCGC\cdots\beta$$
$$\alpha'\cdots GGCG\cdots\beta', \qquad \alpha'\cdots GGCC\cdots\alpha, \qquad \beta\cdots CGCG\cdots\beta'$$

Next, an example is presented on this wet splicing system in single-stage Y-G splicing system to show that single-stage splicing system can produce the same splicing language as two stages splicing system.