GRAMMAR SYSTEMS IN SIMPLE SPLICING

NURUL ‘ALYAA’ BINTI SHUKRI

UNIVERSITI TEKNOLOGI MALAYSIA

GRAMMAR SYSTEMS IN SIMPLE SPLICING

NURUL ‘ALYAA’ BINTI SHUKRI

A dissertation submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

JUNE 2013

To my beloved father, mother and sisters

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest appreciation to my supervisor, Dr. Fong Wan Heng for her guidance and support. I am thankful for her valuable comments, effort and patience in the completion of this dissertation.

Besides, I would like to thank the Department of Mathematical Sciences, UTM, for providing the facilities in completing my study.

Last but not least, deepest thanks to my beloved family and friends for their understanding and endless love as they keep encouraging and supporting me to complete my dissertation. Thank you Allah S.W.T for this wonderful journey. I treasure my experience in completing this dissertation.

ABSTRACT

Splicing systems was first introduced by Head in 1987 as a mathematical model of the generative formalism that initiates the connection between formal language theory and the study of deoxyribonucleic acid (DNA). The mathematical modeling of splicing was developed by involving the activities of restriction enzymes and ligases on a set of DNA molecules. The language resulted from the splicing systems is called the splicing language. The splicing languages are then formalized and investigated by using concepts in formal language theory. Among the different types of splicing systems is the simple splicing system. In this research, some molecular examples on the reduction process of splicing systems into simple splicing systems by using the concept of solid codes are presented. This research focuses on simple splicing systems and its relation with grammar systems where grammar is one of the basic concepts in formal language theory. Among the grammar systems presented in this research are the four types of simple splicing grammar systems (SSGS), pattern grammar systems and pure pattern grammar systems. Then, SSGS is further applied in the test tube system, known as the simple test tube systems. The languages resulted from the various types of SSGS in simple test tube systems are then being analyzed and compared.

ABSTRAK

Sistem hiris-cantum pada asalnya diperkenalkan oleh Head pada 1987 sebagai model matematik untuk formalisme penjanaan yang merupakan asas kepada perhubungan antara teori bahasa formal dan asid deoksiribonukleik (DNA). Model matematik untuk hiris-cantum dibangunkan melalui penglibatan enzim-enzim pembatas dan enzim-enzim penyambung pada set molekul-molekul DNA. Bahasa yang terhasil daripada sistem hiris-cantum dipanggil bahasa hiris-cantum. Bahasa hiris-cantum ini kemudiannya diatur dan diselidik dengan menggunakan konsep-konsep dalam teori bahasa formal. Antara jenis-jenis sistem hiris-cantum ialah sistem hiris-cantum mudah. Dalam kaji selidik ini, contoh-contoh bermolekul proses penurunan sistem hiris-cantum kepada sistem hiris-cantum mudah dengan menggunakan konsep kod padu dipersembahkan. Kaji selidik ini memfokuskan kepada sistem hiris-cantum mudah dan hubungannya dengan sistem tatabahasa yang mana tatabahasa merupakan salah satu konsep asas dalam teori bahasa formal. Antara sistem-sistem tatabahasa yang dibentangkan dalam kaji selidik ini ialah empat jenis sistem tatabahasa hiris-cantum mudah, sistem tatabahasa bercorak dan sistem tatabahasa bercorak tulen. Seterusnya, sistem tatabahasa hiris-cantum mudah ini diteroka dengan lebih lanjut dalam sistem tabung uji yang disebut sebagai sistem tabung uji mudah. Bahasa yang terhasil daripada pelbagai jenis sistem tatabahasa hiris-cantum dalam sistem tabung uji mudah ini kemudiannya dianalisis dan dibandingkan.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	DECLARATION	**ii**
	DEDICATION	**iii**
	ACKNOWLEDGEMENTS	**iv**
	ABSTRACT	**v**
	ABSTRAK	**vi**
	TABLE OF CONTENTS	**vii**
	LIST OF TABLES	**x**
	LIST OF FIGURES	**xi**
	LIST OF ABBREVIATION AND SYMBOLS	**xii**
1	**INTRODUCTION**	**1**
	1.1 Introduction	1
	1.2 Research Background	2
	1.3 Problem Statements	3
	1.4 Research Objectives	4
	1.5 Scope of the Study	4
	1.6 Research Methodology	4
	1.7 Significance of the Research	5
	1.8 Dissertation Organization	5

2	**LITERATURE REVIEW**		**7**
	2.1	Introduction	7
	2.2	Structure of DNA	7
	2.3	Some Basic Concepts in Formal Language Theory	9
		2.3.1 Languages	10
		2.3.2 Grammars	12
	2.4	Splicing Systems and Splicing Languages	13
	2.5	Simple Splicing Systems and Simple Splicing Languages	16
	2.6	Conclusion	17
3	**REDUCTION OF SPLICING SYSTEMS INTO SIMPLE SPLICING SYSTEMS**		**18**
	3.1	Introduction	18
	3.2	Some Concepts of Solid Codes	18
	3.3	Reduction of Molecular Splicing Systems	20
	3.4	Conclusion	23
4	**SIMPLE SPLICING GRAMMAR SYSTEMS, PATTERN GRAMMAR AND PURE PATTERN GRAMMAR**		**24**
	4.1	Introduction	24
	4.2	Languages Generated by the Grammars	25
	4.3	Simple Splicing Rules and Grammar Systems	28
	4.4	Pattern Grammar	37
	4.5	Pure Pattern Grammar	41
	4.6	Application of Simple Splicing Rules in Pattern and Pure Pattern Grammar Systems	43
	4.7	Conclusion	48

5 SIMPLE TEST TUBE SYSTEMS 49

5.1 Introduction 49
5.2 Test Tube Systems 49
5.3 Simple Test Tube Systems 57
5.4 Application of Test Tube System with Simple Splicing Grammars 59
5.5 Conclusion 63

6 SUMMARY AND SUGGESTIONS 64

6.1 Summary of the Research 64
6.2 Suggestions for Future Research 65

REFERENCES 66

LIST OF TABLES

TABLE NO.	TITLE	PAGE
4.1	Splicing rules and their corresponding simple splicing grammar systems (SSGS)	28

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	The double helix structure of DNA	8
2.2	The hydrogen bonding between nitrogenous bases in DNA	9

LIST OF ABBREVIATION AND SYMBOLS

A	-	Alphabet
A	-	Adenine
B	-	Set of all patterns associated with enzymes that either produce 5' overhangs or produce blunt ends
C	-	Set of all patterns associated with enzymes that produce 3' overhangs
C	-	Cytosine
CFG	-	Context-free grammar
CFL	-	Context-free language
DNA	-	Deoxyribonucleic acid
dsDNA	-	Double stranded DNA
G	-	Grammar
G	-	Guanine
H	-	Hydrogen
I	-	Set of initial strings / states
L	-	Language
$L(G)$	-	Language generated by G
$L(S)$	-	Language generated by a splicing system S
mtDNA	-	Mitochondrial DNA
N	-	Nitrogen
O	-	Oxygen
Q	-	Set of states
R	-	Set of splicing rules
SSGS	-	Simple splicing grammar system
S	-	Splicing system
SH	-	Simple splicing
S_kH	-	Families of splicing systems

T	-	Thymine
T	-	Set of terminal states
λ	-	Null string
ϕ	-	Empty set
$\subseteq$	-	Subset
$\in$	-	Element of
$\cup$	-	Union
$\neq$	-	Not equal to
$\leq$	-	Less than or equal to
$\geq$	-	Greater than or equal to
|, #	-	Cutting sites
$	-	Combination site

CHAPTER 1

INTRODUCTION

1.1 Introduction

Deoxyribonucleic acid, which is commonly known as DNA, is a molecular instructions for life that can be found in the cells of almost all living things. DNA is mostly located in the cell nucleus and it is called the nuclear of DNA. However, DNA can also be found in the mitochondria but only in just a small portion. The DNA that was found in the mitochondria is called mitochondrial DNA or mtDNA [1]. DNA molecules can be cut by restriction enzymes and after cutting, the resulting fragments will be pasted by a ligase. After this cut and paste process, new molecules of recombinant DNA will be produced.

Splicing systems that was introduced by Head in 1987 is an important field in DNA computing. Head introduced splicing systems as a mathematical model for a new generative formalism of a biological system containing double-stranded DNA (dsDNA) with the enzymes which allow the cutting and ligating operation on dsDNA [2]. Restriction enzymes are DNA cutting enzymes which can cut DNA sequences in parts; while ligase enzymes are enzymes that catalyze reactions which make bonds join together to form a new DNA sequences. The field of DNA computing which is also known as molecular computing, is a form of computing which uses DNA, molecular biology and biochemistry. Usually, electronic computers and human can only solve an

issue or problem at one time. However by using DNA computing, complex problems can be solved simultaneously.

Splicing system is a system involving a finite set of initial strings over an alphabet with a finite set of rules [3]. On the other hand, formal language theory is a branch of theoretical computer science that is devoted to the study of sets of finite strings (called languages) of symbols chosen from a prescribed finite set (called an alphabet) [2]. In formal language theory, DNA molecules are considered as one single string. This is because formal language theory does not deal directly with linked pair of strings. In splicing systems, the language resulting from the splicing process is called the splicing language. Different languages can result from this recombinant behaviour and these languages are analyzed using some concepts of languages in formal language theory [4]. There are different types of splicing languages including simple splicing language, semi-simple splicing language, persistent splicing language, strictly locally testable language, uniform splicing language and null-context splicing language resulting from the various types of splicing systems [5].

1.2 Research Background

The mathematical modeling of splicing system that was developed by Head initiates the new relationship between formal language theory and the study of macromolecules [2]. This mathematical model (A, I, B, C) consists of [6]:

A - the four bases of A, G, C and T,

I - a finite set of initial strings of DNA molecules,

B - the set of rules consisting 5' overhangs and blunt end of restriction enzymes, and

C - the set of rules consisting 3' overhangs of restriction enzymes.

Splicing systems are categorized into various types such as the simple splicing system and the null-context splicing system. The various types of splicing system depend on the property of cutting site on DNA molecules when a splicing operation is performed with the presence of appropriate enzymes. Some splicing system can be reduced to simple splicing system by using the concept of solid codes [5].

Language and grammar are two concepts in formal language theory and both of them have been extensively investigated in a grammar system. The idea of a grammar system is that several usual grammars are being considered to generate a common language.

The mathematical modeling of splicing system involves several rules and enzymes which can decrease the efficiency of DNA computation. Hence, in order to solve this problem, researches of DNA computing produced a test tube system. In test tube systems which was first mentioned in 1996 [7], the components in every tube will be spliced according to the rules of the tubes respectively. The strings that may arise from the splicing process in every tube are then either inserted to the next tube or remain in the previous tube according to the rules of the next tube.

1.3 Problem Statements

In formal language theory, a language acts as a mathematical statement while a grammar is a formalism that gives a finite representation of a language. Simple splicing grammar, pattern grammar and pure pattern grammar are among the different types of grammar, where each of them works thoroughly by their rules. In addition, simple splicing grammar system can be applied in the test tube systems. Thus, in this research, the following questions will be answered:

1) What are simple splicing grammar system, pattern grammar, pure pattern grammar and their generated languages?

2) How can simple test tube systems be related to the four types of simple splicing grammar system?

1.4 Research Objectives

The objectives of this research are:

i. to study the basic concepts on DNA, splicing systems and splicing languages,
ii. to understand the concept of solid codes in the reduction of splicing systems to simple splicing system,
iii. to recognize the various types of simple splicing grammar systems,
iv. to relate the simple splicing rules in pattern and pure pattern grammar,
v. to present the relation between simple test tube systems and simple splicing grammar.

1.5 Scope of the Study

This research focuses on the simple splicing grammar systems, simple splicing rules and simple test tube systems.

1.6 Research Methodology

This research method is carried out by reading some books and journals which are related to the scope and topic of the research. The languages generated by various

types of grammar systems are then being analyzed by using concepts of formal language theory. Then the relations of different types of grammar systems are investigated by using properties of grammar. These properties are also used in solving the problem of test tube systems. In test tube systems, the components in every tube are spliced according to the rules of simple splicing grammar systems.

1.7 Significance of the Research

The results of this study can be beneficial for both mathematicians and biologists. Meanwhile, DNA computing has become an important area in revolutionizing the entire field of bio-mathematics. Hence, it is proven that the DNA computing is one of the newest exciting areas to be explored by mathematicians and bio-molecular scientists. Besides that, the presence of test tube system can benefit in the progress of lab work as it can overcome the decreasing efficiency of the DNA computing with the normal splicing system.

1.8 Dissertation Organization

As stated in this chapter, the main purpose of the study has been well-mentioned and the whole research has been briefly explained. The DNA concepts are discussed, and also the relation of splicing systems with formal language theory is explained. This chapter also gives introduction of research background, the problem statements and the research objectives. The scope of the study, research methodology and significance of research are also stated.

Chapter 2 elaborates the literature review of this research. The first section of this chapter discusses about the structure of deoxyribonucleic acid (DNA) and followed

by the study of some basic concepts in a formal language theory. The definitions of terms that are mentioned frequently in this research are also included.

This research focuses more on simple splicing systems. Therefore, the main focus in Chapter 3 is the reduction process of splicing systems to the simple splicing systems by using the concept of solid codes. The definitions of the concept of solid codes and some examples of solid and non solid codes are also listed in this chapter. Other than that, at the end of this chapter, two molecular examples are included to illustrate the reduction process.

Chapter 4 discusses thoroughly on the simple splicing grammar systems (SSGS). Various types of SSGS are observed. Besides that, this chapter also discussed on the concept of pattern and pure pattern grammar. Then, at the end of this chapter, those concepts that are discussed earlier in the chapter are combined. For each part in this chapter, examples are included to give the clear view of every concept that are mentioned in the overall study.

In Chapter 5, this research has extended SSGS into the simple test tube systems. As in SSGS, the four types of SSGS are also being considered in simple test tube systems. The languages that are generated by simple test tube systems are discussed in this chapter.

Last but not least, the sixth chapter summarizes the whole research. It also includes some suggestions for further research in this area.

REFERENCES

1. Moran, L. A., Horton, R.A., Scrimgeour, G. and Perry, M. *Principles of Biochemistry*. 5th. ed. Prentice Hall. 2011.
2. Head, T. Formal Language Theory and DNA: An Analysis Of The Generative Capacity Of Specific Recombinant Behaviors. *Bulletin of Mathematical Biology*. 1987. 49 : 737-759.
3. Lim, S. J., Fong, W. H., Sarmin, N. H. and Karimi, F. Mathematical Modelling of Some Null-Context and Uniform Splicing Systems. *Journal of Fundamental Sciences.* 2011. 7(2) : 145 – 149.
4. Sarmin, N.H. and Fong, W. H. Mathematical Modelling of Splicing Systems. *Proceedings of the 1st International Conference on Natural Resources Engineering & Technology*. 24 – 25 July 2006. Putrajaya, Malaysia, 524-527.
5. Fong, W. H. *Modelling of Splicing Systems Using Formal Language Theory*. Ph.D. Thesis. Universiti Teknologi Malaysia; 2008.
6. Yusof, Y., Sarmin, N.H., Fong, W. H. and Karimi, F. Some Relations on Different Types of Splicing Systems. *Journal of Fundamental Sciences*. 2010. 6 : 143-147. 28.
7. Csuhaj-Varju, E., Kari, L. and Paun, E. G. Test Tube Distributed Systems Based On Splicing. *Computers and Artificial Intelligence*, 1996. 2-3: 211-232. 29.
8. Freudenrich, C. *How DNA Works*. Ph.D. Thesis. University of Pittsburgh School of Medicine.2007.
9. Paun, G., Rozernberg, G., and Salomaa, A. *DNA Computing: New Computing Paradigms*. Germany: Springer-Verlag Berlin Heidelberg New York. 1998.

10. Ellsworth, D. L. and Manolio, T. A. The Emerging Importance of Genetics in Epidemiologic Research. I. Basic Concepts in Human Genetics and Laboratory Technology. *Ann Epidemiol.* 1999. 9. 1-16.
11. Linz, P. *An Introduction to Formal Languages and Automata*. 3rd. ed. USA: Jones and Barlett Publishers, Inc. 2001.
12. Shallit, J. *A Second Course In Formal Languages and Automata Theory*. New York: Cambridge University Press. 2009.
13. Kelley, D. *Automata and Formal Languages: An Introduction*. Englewood Cliffs, NJ: Prentice Hall. 1995.
14. Martin, J.C. *Introduction to Languages and the Theory of Computation.* 4th. ed. New York, NY: McGraw-Hill. 2011.
15. Griffiths, J. F., Wessler, S. R., Suzuki, D. T., Miller, J. H. *An Introduction to Genetic Analysis.* 8th. ed. W. H. Freeman. 2004.
16. Kim, S. M. Computational Modelling for Genetic Splicing Systems, *SIAM J. Comput.* 1997. 26(5): 1284-1309.
17. Fong, W. H., Sarmin, N. H. and Ibrahim, Z. Reduction of Splicing System Using Solid Codes. *Prosiding Simposium Kebangsaan Sains Matematik ke -16.* 3-5 June 2008.
18. Bonizzoni, P., Felice, C. D., Mauri, G. and Zizza, R. Decision Problems for Linear and Circular Splicing Systems. In: Ito, M. and Toyama, M. (Eds). *DLT 2002*, LNCS 2450. 78-92; 2003.
19. Head, T. Splicing Representations of Strictly Locally Testable Languages. *Discrete Applied Mathematics.* 1998. 87: 139-147.
20. Mateescu, A., Paun, G. H., Rozernberg, G.and Salomaa, A. Simple Splicing System. 1998. 84. 145 – 163.
21. Laun, E. G. *Constants and Splicing Systems.* Ph.D. Thesis. State University of New York at Binghamton; 1999.
22. New England Biolabs 2011.12 Catalog & Technical Reference.
23. Dassow, J. and Mitrana,V. Splicing Grammar Systems. *Computers and Artificial Intelligence,* 1996, 15: 109-122.

24. Dersanambika, K. S. and Krithivasan, K. and Subramaniam K. G. Simple Splicing Grammar Systems. *Proceedings of Grammar Systems Week.* 170-178. 2004.

25. Dassow, J. Paun, E. G. and Salomaa, A. Grammars Based On Patterns. *International Journal of Foundations of Computer Science*. 1993. 4: 1-14.

26. Sindhu, J. K., Abisha, P. J. and Thomas, D. G. Simple Splicing Pattern and Pure Pattern Grammar Systems. *International Conference on Bio-Inspired Computing: Theories and Applications.* 2011. 6: 220-224.

27. Abisha, P. J., Subramanian, K. G. and Thomas, D. G. Pure Pattern Grammars. *Proceedings of International Workshop On Grammar Systems*. 2000. 253-262.

28. Krithivasan, K., Harsha, P. and Talupur, M. Communicating Distributed H Systems With Simple Splicing Rules. *Proceedings of the 2006 International Conference on Computer Design & Conference on Computing in Nanotechnology*. 2006: Las Vegas, Novada, USA. 107-111.