FORMAL LANGUAGE THEORY AND DNA

NOR MUHAINIAH BINTI MOHD ALI

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

OCTOBER 2004

To my beloved father, mother, husband and family

ACKNOWLEDGEMENT

In preparing this thesis, I wish to express my sincere appreciation to my dissertation supervisor, Associate Professor Dr. Nor Haniza Sarmin, for encouragement, guidance, criticism and friendship. I am also very thankful to Professor Tom Head from Department of Mathematical Sciences, Binghamton University, New York USA for his guidance and lecture series on 26th of July 2004 until 4th of August 2004. Without their continued support and interest, this dissertation would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my M.Sc. study. Librarians also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate students should also be recognized for their support. My sincere appreciation also extends to my beloved husband, Johari bin Mohd Yani for his understanding, love, support and providence assistance at various occasions. Unfortunately, it is not possible to list all of them in this limited space. Lastly, I am very grateful to all my family members.

ABSTRACT

Formal language theory is a branch of applied group theory that is denoted to the study of finite strings called language over some symbols chosen from a prescribed finite set called alphabet. A new manner of relating formal language theory to the study of informational macromolecules is initiated. A language is associated with each pair of sets where the first set consists of double-stranded DNA molecules and the second set consists of the recombinational behaviors allowed by specified classes of enzymatic activities. The scope of this research is on the potential effect of sets of restriction enzymes and ligase that allow DNA molecules to be cleaved and reassociated to produce further molecules. The associated languages are analysed by means of a new generative formalism called a splicing system. Splicing systems were originally developed as a mathematical or dry model of the generative of DNA molecules in the presence of appropriate restriction enzymes and a ligase. A significant subclass of these languages, which we call the persistent splicing languages, is shown to coincide with a class of regular languages which have been previously study in other contexts: the strictly locally testable languages. The relationship between the family *SH* of simple splicing language and the family of strictly locally testable languages is clarified. This study initiates the formal analysis of the generative power of recombinational behaviors in general. The splicing system formalism allows observations to be made concerning the generative power of general recombination and also of sets of enzymes that include general recombination.

ABSTRAK

Teori bahasa formal ialah salah satu cabang kepada aplikasi teori kumpulan dan merupakan satu kajian mengenai bahasa yang mengandungi simbol terhingga yang dinamakan sebagai abjad. Kini terdapat cara baru yang mengaitkan teori bahasa formal dengan kajian maklumat berkaitan makro molekul. Suatu bahasa mengaitkan setiap pasangan set di mana set pertama mengandungi molekul asid deoksiribonukleik (ds-DNA) dan set kedua mengandungi sifat penggabungan yang melibatkan aktiviti enzim bagi kelas yang tertentu. Skop utama penyelidikan ini adalah mengenai kesan potensi set pembatasan enzim dan *ligase* yang membenarkan molekul DNA dipotong untuk bergabung bagi menghasilkan molekul baru. Bahasa yang berkaitan dianalisis menggunakan pendekatan baru yang dinamakan sistem penyambatan (*splicing*). Sistem penyambatan pada asalnya dibangunkan sebagai suatu model matematik atau “model kering” molekul DNA dengan kehadiran enzim pembatasan dan *ligase* yang sesuai. Suatu sub kelas yang penting bagi bahasa ini dinamakan bahasa penyambatan berulang menyamai kelas bahasa biasa (*regular language*) yang sebelum ini dikaji dalam konteks yang berbeza iaitu bahasa teruji setempat terhad (*strictly locally testable language*). Hubungan antara keluarga bagi bahasa penyambatan mudah dan keluarga bagi bahasa teruji setempat terhad (*strictly locally testable*) dijelaskan. Sistem penyambatan yang formal membenarkan pemerhatian dilakukan berkaitan dengan penggabungan umum dan juga yang berkaitan dengan set-set enzim yang terlibat.

CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Σ	-	Alphabet
λ	-	Empty string
$\mid\mid$	-	Length
Σ^k	-	The set of strings of length k
$\in$	-	Element of
$\aleph$	-	Natural numbers
w^R	-	Reversal of w
Σ^*	-	Star closure, Kleene closure or Universal language
Σ^+	-	Positive closure
$\geq$	-	Greater or equal
$\leq$	-	Less or equal
$\subseteq$	-	Subset
$\cup$	-	Union
$\cap$	-	Intersection
DFA	-	Deterministic finite automata
NFA	-	Nondeterministic finite automata
DNA	-	Deoxyribonucleotide acid
ss-DNA	-	Single stranded DNA
ds-DNA	-	Double stranded DNA
SLT	-	Strictly locally testable languages
S_kH	-	S_k Splicing
$LR(S)$	-	Left radius
$RR(S)$	-	Right radius

CHAPTER 1

INTRODUCTION

1.1 Introduction

Formal language theory was developed earliest in the year 1950. This theory was quite related to the non-natural languages that had initiated in computer science. Since that day, this theory has been expanded extensively and has obvious growths, which include applications to the syntactic analysis of programming languages, program schemes, models of biological systems, and relationships with natural languages. In addition, basic research into the abstract properties of families of languages has continued [1].

The theory of formal language is a branch of applied group theory that concerns itself with sets of strings called languages and different mechanisms for generating and recognizing them. Certain finitely process for generating these sets are called grammar. A mathematical theory of such objects was proposed in the latter 1950's and has been extensively developed since that time for application to natural language.

A formal language consists of a set of symbols and some rules of formation by which these symbols can be combined into entities called sentences. A formal language is the set of all strings permitted by the rules of formation [2]. The theory of formal languages has been developed extensively, and has several noticeable trends, which include applications models of biological systems, and relationships with natural languages. This dissertation will discuss applications of formal language theory to the models of biological systems. The application is to establish a new relationship between formal language theory and the study of informational macromolecules.

For several decades it has been realized that molecules of DNA, RNA, and proteins may be idealized as strings of symbols over finite alphabets with the symbols chosen to denote deoxyribonucleotides, ribonucleotides and amino acids, respectively. In considering the primary structure of a protein molecule, an RNA molecule, or a double-stranded DNA molecule or simply abbreviate by ds-DNA it is natural to think of the first as a string over an alphabet of twenty symbols each of which represents an amino acid, the second as a string over an alphabet of four symbols each of which represents a ribonucleotide, and the third as a string over an alphabet of four symbols each of which represents a hydrogen-bonded deoxyribonucleotide pair [3].

In 1987, T. Head [3] introduced a new formalism called splicing systems for the generation of languages and has since received extensive theoretical development. The splicing system concept that introduced in [3] is a formal device for the generation of languages and as a formal model of specific forms of DNA recombination.

Splicing systems were originally developed as a mathematical or dry model of the generative of DNA molecules in the presence of appropriate restriction enzymes and a ligase. According to the original model, a splicing system consists of finite initial set of strings over an alphabet, and a finite set of rules by which the strings can be spliced together to form new strings in addition to the initial strings. The closure of the initial set under the splicing operation generates a splicing language [4].

In the original work of Head [3], he posed the problem of characterizing the splicing languages. In the same work, Head showed the equivalence of the class of strictly locally testable languages to the class of languages, which can be generated by splicing using a restricted, uniform set of rules.

1.2 Objectives

The objectives of this research are:

- To review some important related concepts in formal language theory.
- To study and prove some properties or theorems in formal language theory.
- To study the relationship that exists between formal language theory and the study of informational macromolecules.

1.3 Scope

- This research focuses on formal language theory and its applications to the models of biological systems.

1.4 Introduction to Each Chapter

Chapter 1 is the introduction chapter that covers the background of the research, objectives and scope of the research, and a little introduction to each chapter. In Chapter 2, the basic concepts of formal language theory will be discussed. The basic concepts include alphabet, string, language and the operation on string and languages. In this chapter, finite automata that consists of deterministic finite automata and nondeterministic finite automata, and the concept of regular language will be covered.

The concepts in formal language theory will then be applied to the model of biological systems, which will be discussed in Chapter 3. These include types of languages that are used to model a set of DNA molecules in the absence of enzymes. The action of sets restriction enzymes on sets of DNA molecules is then given with a formal representation and analysis.

The splicing systems concept, which is a formal device for the generation of languages and as a formal model of specific form of DNA recombination will be presented and discussed in Chapter 4. It includes a discussion on restriction enzymes and their roles in the splicing system. The language that is generated by splicing system is called splicing language, and its regularity will also be discussed.

Splicing system can generate the class of languages that is equivalence to the class of strictly locally testable language. This strictly locally testable language will be presented in Chapter 5. The discussion of the problem of generating languages with a null context splicing systems will also be included.

Lastly, Chapter 6 includes the conclusion and suggestion for this dissertation.