FORMAL MODELS AND ALGORITHMS FOR DNA DATA ANALYSIS USING WATSON-CRICK GRAMMARS

BY

NURUL LIYANA BINTI MOHAMAD ZULKUFLI

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

NOVEMBER 2017

ABSTRACT

Genetic information encoding in deoxyribonucleic acid (DNA) and DNA manipulation techniques contributes to the advancement of computing technology; namely DNA computing. However, the research in the counterpart of automata in formal language theory which is grammars, of Watson-Crick automata that is one of DNA computing models, has yet to be completed. Thus, more extended computing models for both computationally and algorithmically efficient methods for general and specific purposes can be further exhibited through the investigation of the grammars. The purpose of this research is to develop novel theoretical models based on DNA computing, Watson-Crick automata specifically and it also aims to design efficient algorithms using these models for the specific purposes especially membership and parsing problems. We have established Watson-Crick grammars; particularly modified version of Watson-Crick regular grammars, novel Watson-Crick linear grammars, and Watson-Crick context-free grammars. The expansion of the grammars starting from Watson-Crick regular grammars to Watson-Crick context-free grammars was based on Chomsky's families of languages hierarchy. The generative power and closure properties of each type of the grammars were investigated. Watson-Crick regular grammars are able to generate non-context-free languages, thus it is more powerful than their counterpart which is regular grammars. The distinction of each type of Watson-Crick grammars is on their ability to generate different levels of balanced parentheses in Dyck language. It is proven that Watson-Crick regular languages are properly included in Watson-Crick linear languages, and the latter are properly included in Watson-Crick context-free languages. The upper bound of the family of Watson-Crick context-free languages is the family of matrix languages. Although they are more powerful, it has been discovered that their closure properties are similar with the properties of the Chomsky grammars' counterparts. The simplification processes of Watson-Crick (WK) context-free grammars were studied. This resulted in a normal form based on Chomsky normal form, but it came with two types of terminal symbols instead of one. Using this WK-Chomsky normal form, we designed a membership algorithm based on Cocke-Younger-Kasami algorithm, which can also be utilized as a parsing method for data in the form of double-stranded string with Watson-Crick complementarity. The expansion of Watson-Crick context-free grammars also led to the development of their automata counterparts, Watson-Crick pushdown automata; with one stack or complementarity stacks. Unlike Watson-Crick automata, complementarity-stack Watson-Crick pushdown automaton includes double stacks connected by Watson-Crick complementarity. We also modified two basic parsing algorithms: top-down parsing and bottom-up parsing to suit double-stranded strings based on this automaton. Further, we introduced Watson-Crick matrix grammars where the rules in Watson-Crick context-free grammars are arranged in matrices, allowing simultaneous control on the production rules. Although it is not more powerful than parallel communicating Watson-Crick automata systems, the result suggests that parallelism features can be employed in the variants of Watson-Crick grammars. These models can be well implemented in string matching problems such as problems occur in DNA analysis and natural language processing.

خلاصة البحث

إن ترميز المعلومات الوراثية في الحمض النووي الريبي منقوص الأوكسجين (DNA) وتقنيات التلاعب في الحمض النووي يسهم في النهوض بتكنولوجيا الحوسبة التي يطلق عليها حوسبة الـ: (دي أن إي). ومع ذلك، فإن الأبحاث في نظرية واتسون –كريك الآلي في اللغة الرسمية وهو قواعد النحو، ونظرية واتسون–كريك الآلية التي هي واحدة من نماذج الحوسبة في الحمض النووي، وهو عمل لم يكتمل بعد. وهكذا، فإن نماذج الحوسبة كانت أكثر سعة وفعالية لكلتا الطريقتين حسابياً ولأغراض عامة ومحددة، ويمكن زيادة عرضها خلال توظيفها في قواعد النحو. والغرض من هذا البحث هو تطوير نماذج نظرية جديدة تقوم على حوسبة الـ: (دي أن إي)، و نظرية واتسون –كريك الآلية على وجه التحديد، ويهدف أيضاً إلى تصميم الخوارزميات الفعالة لاستخدام هذه النماذج لأغراض محددة، وخاصةً العضوية وتحليل المشاكل. لقد وضعنا قواعد النحو عند نظرية واتسون–كريك، ولا سيما النسخة المعدلة من واتسون–كريك في قواعد النحو العادية، و نظرية قواعد النحو الخطية لواتسون–كريك، وقواعد النحو الخالية من السياق لواتسون–كريك، واستند ذلك إلى توسيع قواعد النحو، ومن قواعد النحو العادية لواتسون–كريك، إلى قواعد النحو الخالية من السياق لواتسون–كريك اعتماداً على التسلسل الهرمي لأسر اللغات عند تشومسكي. وقد تم التحقق من القوة التوليدية للخصائص المغلقة من كل نوع في قواعد النحو. إن قواعد النحو لواتسون–كريك العادية قادرة على توليد لغات خالية من السياق، ومن ثمّ فهي أقوى من نظيرتها؛ إذ تميز كل نوع من قواعد النحو عند واتسون–كريك بقدرتها على توليد مستويات مختلفة من الأقواس المتوازنة في لغة العالم اللغوي دايك. وثبت أن اللغات العادية عند نظرية واتسون–كريك مدرجة بشكل صحيح في اللغات الخطية لديها، وتدرج هذه الأخيرة بشكل صحيح في اللغات الخالية من السياق لواتسون–كريك. والحد الأعلى لأسرة واتسون–كريك في اللغات الخالية من السياق هو عائلة اللغات المصفوفة، وعلى الرغم من أنها أكثر قوة، فقد تم اكتشاف أن خصائص الإغلاق تتشابه مع خصائص نظرائها في قواعد النحو لدى تشومسكي. وتمت دراسة عمليات تبسيط واتسون–كريك في قواعد النحو الخالية من السياق، وعُمل ذلك في شكل طبيعي بالاعتماد على النموذج العادي لتشومسكي؛ لكنها جاءت مع نوعين من الرموز النهائية بدلاً من واحدة. ويمكن استخدام هذا النموذج العادي (–WK) عند تشومسكي؛ إذ قمنا بتصميم خوارزمية عضوية على أساس خوارزمية (CYK)، والتي يمكنها أن تستخدم بوصفها وسيلة من وسائل توزيع البيانات على شكل سلسلة مزدوجة متقطعة مع واتسون–كريك المتكاملة. وأدى توسيع قواعد النحو الخالية من السياق لواتسون–كريك إلى تطوير نظرائها الآلية الخاصة بها، وصممت نظرية واتسون–كريك الآلية والموسعة والسفلية، مع مكدسات متكاملة أو مكدس واحد، على عكس تصميم واتسون–كريك الآلي، ويشمل

التكامل لواتسون–كريك الآلي الموسع والسفلي إنسانا آليا للمداخن المزدوجة المتصلة بواسطة نظرية واتسون–كريك المتكاملة، وقمنا بتعديل اثنين من خوارزميات التحليل الأساسية، هي: التحليل من أعلى إلى أسفل، ومن أسفل إلى أعلى وتوزيعها لتتناسب مع السلاسل المزدوجة المتقطعة على أساس أن هذا إنسان. وعلاوة على ذلك، قدمنا قواعد النحو لواتسون–كريك مصفوفة؛ حيث تمّ ترتيب المصفوفات والقواعد في قواعد النحو لواتسون–كريك الخالية من السياق؛ ما يسمح لنا بالتحكم في الوقت نفسه بقواعد الإنتاج. وعلى الرغم من أنها ليست أقوى من موازية التواصل للأنظمة الآلية لواتسون–كريك، تقترح النتيجة أنه يمكن استخدام ملامح التوازي في أنواع من قواعد النحو لواتسون–كريك. وهذه النماذج يمكن تنفيذها بشكل جيد في مشاكل مطابقة السلسلة، مثل: المشاكل التي تحدث في تحليل الحمض النووي ومعالجة اللغة الطبيعية.

APPROVAL PAGE

The thesis of Nurul Liyana binti Mohamad Zulkufli has been approved by the following:

_____
Sherzod Turaev
Supervisor

_____
Mohd. Izzuddin Mohd. Tamrin
Co-Supervisor

_____
Mohamed Ridza Wahiddin
Internal Examiner

_____
Haniza Sarmin
External Examiner

_____
K G Subramanian
External Examiner

_____
Adlina Ariffin
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Nurul Liyana binti Mohamad Zulkufli

Signature..... Date.....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

FORMAL MODELS AND ALGORITHMS FOR DNA DATA ANALYSIS USING WATSON-CRICK GRAMMARS

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Nurul Liyana binti Mohamad Zulkufli

……..…….. ……..

Signature Date

ACKNOWLEDGEMENTS

All praise be to Allah for giving me the chance to glorify Allah more by realising how hard it is to develop only a simple model based on a small part of DNA, which Allah has miraculously created with ease, and how insignificant myself and the human power are, which makes me admiring Allah more and more.

I wish to express my gratitude to my mom, brothers and sister, family, friends, and housemates, who always support me, shining my life, and always giving me motivating and supportive advice especially during hard times. Many thanks to my family, for their unconditional love; and also to my late father – may Allah bless him.

I also wish to express my appreciation and many thanks to my supervisor, Dr. Sherzod, for his dedicated supervision, brilliant discussions, motivating advices and kindness, thus shine my way brightly especially in research. Also, thanks to Dr. Izzuddin for the support in being my co-supervisor and bright smiles. Likewise, thanks to Dr. Messikh for being the chairman of my supervising committee.

Thank you to the important figures and personnel in the department, kulliyyah, and the university, in which through them, I met this fascinating topic, and their cooperation and support are indispensable. Special thanks to the examiners for their invaluable time and energy in examining the thesis, coming to the viva and discussing this research with me.

Finally but importantly, special thanks to my husband, for his patience and support throughout my journey in this research. Thank you for your love, sacrifice and kindness, which I am not able to repay.

May Allah reward all of you with more blessings and happiness.

TABLE OF CONTENTS

Abstract ii
Approval Page v
Declaration vi
Copyright Page vii
Acknowledgements viii
Table of Contents ix
List of Tables xii
List of Figures xiii
List of Abbreviations xiv
List of Symbols xvi

CHAPTER ONE: INTRODUCTION 1
1.1 Motivation of the Study 1
1.2 Statement of The Problem 2
1.3 Research Questions 4
1.4 Reasons, Hypothesis, and Contribution 4
1.5 Purpose of the Study 8
1.6 Research Objectives 8
1.7 Scope of the Study 9
1.8 Research Methodology 9
1.9 Thesis Organisation 13

CHAPTER TWO: LITERATURE REVIEW 14
2.1 Food authentication 14
2.1.1 PCR-Based Authentication 15
2.2 How DNA Works 17
2.2.1 Miraculous Structure of DNA 17
2.2.2 Gel Electrophoresis 20
2.2.3 How DNA is Manipulated 21
2.3 DNA Computing Models 25
2.3.1 Splicing Systems 26
2.3.2 Sticker Systems 27
2.3.3 Watson-Crick Automata 28
2.3.4 Grammars for DNA and RNA 30
2.4 Summary 32

CHAPTER THREE: PRELIMINARIES 33
3.1 Introduction 33
3.2 General notations 33
3.3 String and Languages 34
3.4 Operations with Strings and Languages 34
3.5 Grammars, Chomsky Hierarchy, and Derivation Tree 35
3.6 Finite Automata and Pushdown Automata 39
3.7 Watson-Crick Complementarity 40
3.8 Sticker Systems 41

3.9 Watson-Crick Finite automata 43
3.10 Watson-Crick Regular Grammars 44
3.11 Multi-Head Watson-Crick Finite Automata 45
3.12 Matrix Grammars 46
3.13 Cocke-Younger-Kasami (CYK) Algorithm 46

CHAPTER FOUR: WATSON-CRICK GRAMMARS 48
4.1 Introduction 48
4.2 Definitions 49
4.3 Generative Power 51
4.3.1 Non-context-free Languages 51
4.3.2 Balanced Parentheses 57
4.3.3 Relation with Arbitrary Sticker Systems 61
4.3.4 The Upper Bound for Watson-Crick Context-free Grammars 65
4.3.5 Hierarchy of Watson-Crick Languages 67
4.4 Closure Properties 68
4.4.1 Watson-Crick Regular Grammars 68
4.4.2 Watson-Crick Linear Grammars 71
4.4.3 Watson-Crick Context-free Grammars 72
4.5 Applications of Watson-Crick Grammars 73
4.5.1 DNA Structure Analysis 74
4.5.2 Programming Language Structure Analysis 76
4.5.3 Natural Language Processing 77
4.6 Summary 79

CHAPTER FIVE: GRAMMAR SIMPLIFICATIONS AND NORMAL FORMS 82
5.1 Introduction 82
5.2 Shuffle and Terminal Normal Form 83
5.3 Derivation Tree 85
5.4 Grammar Simplifications 90
5.5 WK-Chomsky Normal Form 95
5.6 Summary 98

CHAPTER SIX: WATSON-CRICK PUSHDOWN AUTOMATA 100
6.1 Introduction 100
6.2 One-Stack Watson-Crick Pushdown Automata 101
6.2.1 Definitions 101
6.2.2 Generative Power 103
6.3 Complementarity-Stack Watson-Crick Pushdown Automata 106
6.3.1 Definitions 107
6.3.2 Generative Power 109
6.4 Summary 111

CHAPTER SEVEN: PARSING 112
7.1 Introduction 112
7.2 Stack-based Parsing 112
7.2.1 Depth-First Top-down Parsing 113
7.2.2 Depth-First Bottom-up Parsing 117

7.3 Modification of CYK Algorithm 122
7.4 Summary 135

CHAPTER EIGHT: WATSON-CRICK MATRIX GRAMMARS 136
8.1 Introduction 136
8.2 Definitions 137
8.3 Generative power 140
8.3.1 Generating Non-regular Unary Language 140
8.3.2 Relation with Multi-Head Watson-Crick Automata 144
8.4 Application in Natural Language Processing 148
8.5 Summary 150

CHAPTER NINE: CONCLUSION AND FUTURE WORK 152
9.1 Open Problems 157
9.2 Future Work 158

REFERENCES 160
List of Publications 166

LIST OF TABLES

Table No.		Page No.
3.1	The closure properties of the families of languages in Chomsky hierarchy	37
4.1	The closure properties of the families of languages in Chomsky hierarchy and Watson-Crick languages	80

LIST OF FIGURES

Figure No.		Page No.
1.1	Synthesis process in bacterial DNA replication	6
1.2	The simulation of a synthesis process with a grammar derivation	6
1.3	The flow chart of research approach	12
2.1	A DNA structure in its double-helix form	18
2.2	A simple representation of DNA strands	18
2.3	The schematic representation of a nucleotide	19
2.4	Gel electrophoresis	21
2.5	DNA lengthening and shortening	22
2.6	DNA ligation	23
2.7	Denaturation, priming, and extension processes on PCR	24
2.8	Splicing operation	26
2.9	Sticking operation	27
2.10	Watson-Crick finite automaton	29
3.1	Possible shapes of 'dominoes' or 'bricks' for sticking	41
4.3.1	The hierarchy of Watson-Crick, sticker, matrix, Chomsky languages	68
5.3.1	A derivation tree for [abc/abc]	87
5.3.2	Different types of derivations in WK context-free grammars	89
6.2.1	One-stack WK pushdown automaton	101
6.3.1	Complementarity-stack WK pushdown automaton	107
7.3.5	The derivation tree using WK-CYK algorithm	134
9.1	The summary of the thesis	156

LIST OF ABBREVIATIONS

WK	Watson-Crick
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
PCR	Polymerase chain reaction
RFLP	Restriction Fragment Length Polymorphism
PCWKS	Parallel communicating WK automata system
CYK	Cocke-Younger-Kasami
FL	Family of languages
RE	Recursive enumerable languages
CS	Context-sensitive languages
CF	Context-free languages
REG	Regular languages
LIN	Linear languages
NFA	Nondeterministic finite automaton
DFA	Deterministic finite automaton
ASL	Arbitrary sticker languages
OSL	One-sided sticker languages
RSL	Regular sticker languages
SSL	Simple sticker languages
NWK	Stateless WK automaton languages
FWK	All-final WK automaton languages
SWK	Simple WK automaton languages

1WK	1-limited WK automaton languages
AWK	Arbitrary WK automaton languages
MAT^{λ}	The family of languages of matrix grammars (without appearance checking, with erasing rules)
WKREG	WK regular languages
WKLIN	WK linear languages
WKCF	WK context-free languages
WKPDA1	One-stack WK pushdown automaton languages
WKPDAc	Complementarity-stack WK pushdown automaton
WK-CYK	CYK algorithm for WK context-free grammars
WKMAT	WK matrix languages

LIST OF SYMBOLS

$\in$	Membership of an element to a set
$\notin$	Negation of $\in$
$\subseteq$	Inclusion
$\subset$	Proper inclusion
$\cup$	Union
$\cap$	Intersection
$\times$	Cross product
-	Difference
$\|X\|$	Cardinality of set X
2^X	Power set of X
$\emptyset$	Empty set
$\{A\}$	Element in a set
$x_{\{i,j\}}$	Index of x
$x_{i,j}$	Index of x
V	Alphabet
V^*	The set of all finite strings over V
V^+	The set of all non-empty finite strings over V
λ	Empty string
$\|w\|$	Length of string w
w	Shuffle
w^R	Mirror image of string w
G	Grammar

N	Set of nonterminal symbols
T	Set of terminal symbols
S	Start symbol
P	Set of production rules
Q	Set of states
δ	Transition function
q_0	Initial state
F	Set of final states
δ^*	Extended transition function
Γ	Stack alphabet
ρ	Symmetric relation
$WK_\rho(V)$	Watson-Crick domain, set of all (well-formed) double-stranded string
$\langle u/v \rangle$	Double-stranded string
$[u/v]$	Complete double-stranded string

CHAPTER ONE

INTRODUCTION

1.1 MOTIVATION OF THE STUDY

-*When DNA meets food authentication, computing, and formal language theory*-

Following the discovery of deoxyribonucleic acids (DNA) in 1950's, thanks to the unique features of DNA: complementarity relation and massive parallelism, technology has been improved in many biology-related fields such as biotechnology and food authentication.

Food authenticity is very important for the Muslim dietary and it is also very important for ensuring health and preventing fraud in trades regardless of beliefs. Prior to this discovery, to differentiate between different kinds of meat products, researchers use methods such as microscopic imaging, analysing thermal compounds and protein. Analysing DNA is found to be much more useful due to its existence in almost all cells, stability against heat and processing, and contains much more information than protein (Ali et al., 2012). With the existence of data in the form of DNA sequences, bioinformatics and computer science field also play important roles in analysing these data especially by the means of, not limited to, string matching algorithms (Jones & Pevzner, 2004).

The discovery also brings novel computation ideas formed into a special field named DNA computing, which began when Adleman solved the Hamiltonian path problem by using DNA in (Adleman, 1994). Heavily related to the theory of formal language and automata field, several computation models have been created to formalise the biological processes performed on DNA, such as sticker system from sticking operation and splicing system form splicing operation (Paun, Rozenberg, &

Salomaa, 1998). Another fascinating model is Watson-Crick automaton (Freund, Paun, Rozenberg, & Salomaa, 1999), which can also be thought as a simple but brilliant extension to the well-known finite automaton, and its grammar counterpart called Watson-Crick regular grammar.

Grammars are also found to be profoundly useful in finding the language structures of DNA, and its sister molecules ribonucleic acids (Damaševičius, 2010). Not only that, it is revealed that digital information can also be successfully stored and rewritten in a rewritable DNA storage (Tabatabaei Yazdi, Yuan, Ma, Zhao, & Milenkovic, 2015; Yazdi, Kiah, Garcia-ruiz, Ma, & Zhao, 2016), which was once only available in archival version (Bornholt et al., 2016), thus may bring a new era of storing digital information.

1.2 STATEMENT OF THE PROBLEM

There is a variety of DNA-based technologies, for example DNA-based food authentication methods and DNA storage. However, these DNA-based technologies have not yet been fully described in formal language theory and automata, despite using biological processes on DNA which can also be categorised as computing. Some, but not all, of the bio-operations such as sticking operation and splicing have been formalised, in the form of sticker systems (Kari, Păun, Rozenberg, Salomaa, & Yu, 1998) and splicing systems (Head, 1987; Păun, 1996), respectively, which are parts of DNA computing.

Although most of DNA-based food authentications are conducted by detecting DNA fragments' length of different species with the use of gel electrophoresis, on the core, authentication can also be considered as identification or recognition of double-stranded sequences of nucleotides. This may be done with a model called Watson-Crick

finite automata (Freund, et. al., 1999), but it is not confirmed if the automata is powerful enough to describe fully the language of DNA, as more researches are still ongoing in finding the grammatical structure of DNA.

Despite having the automata version, their grammars' counterparts are yet to be studied completely. The regular grammars version for double-stranded string with complementarity relation have been introduced in (Subramanian & Venkat, 2012), but not the linear and context-free ones. Moreover, how powerful the grammars are, and what will happen when string operations performed on the languages generated by the grammars are also yet to be known.

There are also many extended models of Watson-Crick finite automata, but since the research on their grammars' counterparts is not yet completed, further implementation to related fields such as natural language processing and programming languages had to be withhold. The necessity to parse double-stranded strings with complementarity relation will arise with the advent of DNA-based computers and technologies, for example storing information in DNA, which has higher storage capacity electronic media and long-term durability, but costly in writing and extracting information (Tabatabaei Yazdi et al., 2015).

Moreover, many of the parsers used in developing programming languages and in analysing natural languages are based on context-free grammars. However, context-free grammars, although useful and easy to use, are not omnipotent: they cannot cover all aspects of natural languages (Dassow, 2004). On the other hand, grammars using double-stranded strings with complementarity relation may yield more powerful computation exceeding the performance of context-free grammars, which are important in parsing thus bringing improvement in other related fields.

1.3 RESEARCH QUESTIONS

1. What are the existing DNA-based technologies, especially food authentication methods and DNA computing models? What are the similarities and differences between them?
2. Which DNA-based processes that are suitable for proposing DNA computing-based models?
3. Which formal language models that are more suitable to adapt for developing the DNA computing-based models?
4. What are the computational power, complexity and closure properties of the proposed models?
5. What types of algorithms for DNA data analysis that could be designed using the proposed models?

1.4 REASONS, HYPOTHESIS, AND CONTRIBUTION

In this section, the questions stated above are answered.

In the attempt to formalize the processes in DNA-based food authentication process, one can successfully use *DNA computing*, which is one of the most exciting new developments of computer science from both theoretical and practical points of view. DNA computing models use two fundamental features, namely, *Watson-Crick complementarity* and *massive parallelism* of DNA molecules. The use of these features has already illustrated that *DNA-based computers* can solve many computationally intractable problems for large instances. Since the main aim of DNA computing paradigms is to develop computationally efficient methods and techniques, models based on DNA computing can be easily implemented at hardware and software levels.

As authentication process at its core intuitively can be regarded as recognition of double-stranded sequences, the DNA computing model is one of ideal model for that purpose, specifically Watson-Crick automata. By using these automata, the symbols (nucleotides) in the input string are examined one by one, which differ from sticker systems and splicing systems where the operations are performed on fragments of substrings.

Another reason to analyse the symbols (nucleotides) one by one, lies in the DNA's principal function consisting of self-replication which allows precise copy to be passed to the offspring cells. As a motivation, how the synthesis process in a bacterial DNA replication (Campbell & Reece, 2005) resembles a grammar derivation is mentioned below. In the beginning of the replication process (see Figure 1.1), the double-stranded string is being unzipped into two strings, but still connected to each other at one or more points, producing shape(s) resembling bubble(s). In the bubble, there are points where the replication will begin, called origin of replications.

Using these parental strands as template, the synthesis process starts from the origin, proceeding in 5’ to 3’ directions of the new strands. Thus, this leads to two types of new strands in the process, called the leading strand and the lagging strand. Enzymes called primase produce one RNA primer to each leading strand, and several RNA primers for lagging strands. Another enzymes called DNA polymerase III (DNA pol III) elongate the new strands by adding nucleotides complement to the parental strands one by one after the primers. The RNA primers are then replaced with DNA by the enzymes DNA polymerase I (DNA pol I). The gaps in between the new strands resulted from this process are filled by the enzymes DNA ligase, thus completing the replication process.

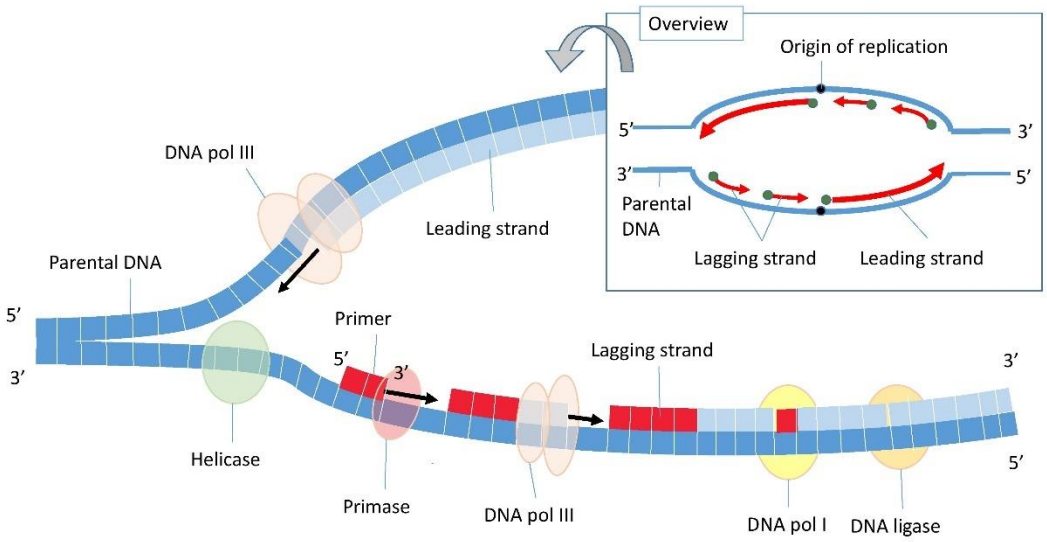

Figure 1.1 Synthesis process in bacterial DNA replication.

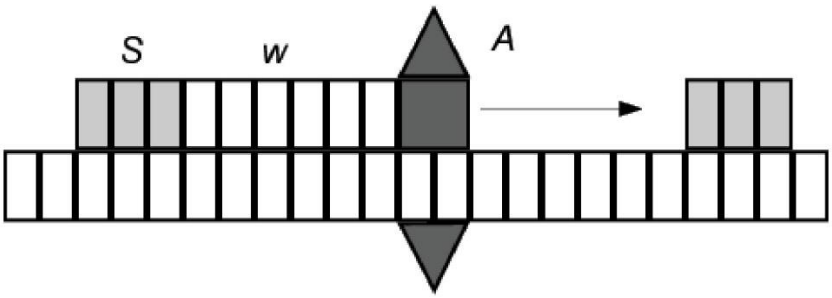

Figure 1.2 The simulation of a synthesis process with a grammar derivation.

Coming from the theory of formal languages, this synthesis process can be seen as a string derivation process in grammars (see Figure 1.2). The primase works as the initial symbol *S*, while the DNA polymerase (I and III) enzymes as the nonterminal symbols (*A*) in the production rules. The symbol *w* in the figure denotes the string generated by the nonterminal symbol *A*. Finally, the enzyme DNA ligase works as the nonterminal symbol which produces only terminal symbols.

Since the grammars' counterparts for Watson-Crick automata have not been completely studied yet, the particular subject has been chosen to be investigated, as the grammars of Watson-Crick finite automata can be extended to linear and context-free versions based on Chomsky hierarchy in formal language theory. From this, equivalent automata can be developed for the particular grammars. By establishing the relations between the automata and grammars, it will be more flexible to switch to one another for different occasions or purposes.

Additionally, it is also fascinating if to generate desired DNA fully within our control, instead of only recognising and generating DNA sequences without full control. In the future, more discoveries about which part in the strands describes the human traits may be found, thus it may be possible to control the genetics according to desired preferences. To do this, grammars approach might be the ideal way, and suitable parsing algorithms specially designed for double-stranded strands with complementarity relation might be necessary in verifying the correctness of the grammar to desired trait (language) and vice versa. As mentioned before, parsing this type of strands might also improve the process of reading and rewriting information in DNA storage.

Insya Allah, the proposed Watson-Crick grammars and automata are computationally powerful, thus they can be considered as new models for DNA computation. Further, these models can provide novel parsing techniques for data in the form of DNA-based sequences. Moreover, a matrix variant is more suitable to manifest the massive parallelism feature of DNA in formal languages.