MATHEMATICAL MODELLING OF PERSISTENT SPLICING SYSTEMS IN DNA COMPUTING

FARIBA KARIMI

A thesis submitted in fulfilment of the

requirements for the award of the degree of

Doctor of Philosophy (Mathematics)

Faculty of Science

Universiti Teknologi Malaysia

JUNE 2013

Especially to my beloved, Mother and Father

Who are the meaning of life to me

Also to

My dear brothers,

Abolfaz and Ali

ACKNOWLEDGMENT

A profound depth of gratitude is owed to my dear supervisor, Assoc. Prof. Dr. Nor Haniza Sarmin and my dear co supervisors Dr. Fong Wan Heng. and Dr. Mohd Firdaus Abdul Wahab for their continuous support, help and guidance from the gathering of ideas until the completion of the report writing. Their unwavering belief in me has been a great source of my strength.

Furthermore, I would like to give very special thanks to Assoc. Prof. Dr. Nor Haniza Sarmin who encouraged me to attend in many conferences and I could gain experience. Also, she always read and responded to the drafts of my work and emails quicker than I could have hoped.

Also, I would like to give my special thanks to Dr. Fong Wan Heng who permanently helped me and her comments have been always perceptive, helpful, and appropriate.

Moreover, I should appreciate Dr. Mohd Firdaus Abdul Wahab patience and his kind help during my work in laboratory. Indeed the practical part of my thesis would not have been possible without his expert guidance.

I acknowledge, with gratitude, to the School of Postgraduate Studies (SPS) UTM for the award of UTM International Doctoral Fellowship (IDF) for three semesters.

I would like to express my sincere gratitude to my co-author, Dr Sherzod Turaev from the Department of Department of Computer Science, International Islamic University Malaysia for his assistance in completing some parts of this thesis.

Last but not least, my special thanks to my family, especially my parents, for their continuous support and patience and all my friends for their love and support over the last few years at Universiti Teknologi Malaysia.

ABSTRACT

Splicing system is a bio-inspired computational model that interprets the cutting and pasting behavior of DNA molecules in the presence of restriction enzymes. Splicing system is defined under the framework of formal language theory. In this research the relation between different types of splicing systems and languages such as uniform, null-context, permanent, persistent and strictly locally testable languages are investigated. Then, the characteristics of persistent and permanent splicing systems are explored in detail. The interesting point about these two systems is that if restriction enzymes are chosen from actual biological sense, then the resulting systems are often persistent and permanent. Their main feature is that the property of crossing of a site is preserved and consequently, the enzymes cleavage process persists. Some sufficient conditions are provided for splicing systems to be persistent and permanent. New concepts of self-closed, crossing-preserved and extended crossing-preserved are introduced. These new concepts are closely connected to the notions of persistent and permanent systems. Moreover, fuzzy splicing system is introduced as an extension of splicing systems. In fact, by considering the threshold languages generated by fuzzy splicing systems, their computational power is increased. In other words, there are some fuzzy splicing systems that generate non-regular languages, while splicing systems with finite components can only generate regular languages. At the end of this research, a laboratory experiment has been conducted to biologically validate the behavior of persistent splicing systems.

ABSTRAK

Sistem hiris-cantum adalah satu model pengkomputeran bioinspirasi yang mentafsirkan tingkah laku potong dan tampal molekul-molekul DNA dengan kehadiran enzim-enzim pembatas. Sistem hiris-cantum ditakrifkan menggunakan Teori Bahasa Formal. Dalam penyelidikan ini hubungan antara pelbagai jenis sistem hiris-cantum dan bahasa seperti bahasa seragam, berkonteks-pincang, kekal, berterusan dan bolehuji setempat yang tegas disiasat. Kemudian, ciri sistem hiris-cantum berterusan dan kekal diterokai dengan terperinci. Perkara menarik tentang kedua-dua sistem ini adalah jika enzim pembatas dipilih daripada keadaan biologi sebenar, sistem yang terhasil selalunya adalah berterusan dan kekal. Ciri utama sistem hiris-cantum berterusan dan kekal iaitu sifat sebagai lintasan tapak adalah dipelihara dan akibatnya, proses belahan enzim adalah berterusan. Beberapa syarat cukup bagi sistem hiris-cantum untuk menjadi berterusan dan kekal diberikan. Tambahan pula, konsep baru tertutup-sendiri, lintasan-dipelihara dan peluasan lintasan-dipelihara diperkenalkan yang mana ia berhubung rapat dengan tanggapan sistem berterusan dan kekal. Seterusnya, sistem hiris-cantum kabur diperkenalkan sebagai peluasan sistem hiris-cantum. Dengan mempertimbangkan bahasa ambang yang dijana oleh sistem hiris-cantum kabur, kuasa pengiraan sistem hiris-cantum meningkat. Dengan kata lain, terdapat beberapa sistem hiris-cantum kabur yang menjana bahasa tidak biasa. Manakala sistem hiris-cantum dengan komponen terhingga hanya boleh menjana bahasa biasa. Di akhir penyelidikan ini, satu eksperimen makmal telah dijalankan untuk mengesahkan tingkah laku sistem hiris-cantum berterusan secara biologi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ ABBREVIATIONS/ NOTATIONS

a	-	[A/T]
a'	-	Complement of a
A	-	alphabet
A^*	-	strings obtained by concatenating operation of zero or more symbols from A
A^P	-	set of strings with length P
bp	-	base pair
B	-	set of all patterns associated with enzymes that either produced 5′ overhangs or blunt end/ left pattern rule
c	-	[C/G]
C	-	set of all patterns associated with enzymes that either produced 3′ overhangs / right pattern rule
CF	-	context free
CS	-	context sensitive
DNA	-	deoxyribonucleic acid
dsDNA	-	double-stranded DNA
FIN	-	family of finite languages
g	-	[G/C]
I	-	set of initial strings obtained from A^*
iff	-	if and only if
L	-	language
$L(S)$	-	language generated by the splicing system S
L_{CF}	-	family of context free languages
L_{CS}	-	familiy of context sensitive languages
$L_f(\gamma)$	-	language generated bye the fuzzy splicing system γ
L_{FIN}	-	family of finite languages

L_{RE}	-	familiy of recursively enumerable languages
L_{REG}	-	familiy of regular languages
NEB	-	New England Biolabs
PCR	-	polymerase chain reaction
R	-	set of splicing rules
r/r'	-	sequence of a DNA or its complement reverse
RE	-	recursively enumerable
REG	-	regular
S	-	splicing system
t	-	[T/A]
T	-	set of terminal alphabets
Y-G	-	Yusof-Goode
γ	-	fuzzy splicing system
μ	-	fuzzy membership function
μL	-	micro liter
μM	-	micro Molar
σ	-	splicing scheme
° C	-	degree Celsius
▼... ...▲	-	crossing of restriction site of restriction enzyme
□	-	end of proof
■	-	end of example
$\odot$	-	fuzzy operation over [0,1]
$\mapsto_r$	-	derived by splicing using the rule r
$\prod$	-	product of a sequence
$\in$	-	element of
$\notin$	-	not element of
$\exists$	-	there exist
$\forall$	-	for all
$\neq$	-	not equal to
$\cap$		intersection
$\cup$		union

LIST OF APPENDICES

CHAPTER 1

INRODUCTION

1.1 Introduction

DNA computing is an emerging field that connects the concepts from biology and genetics to computation. Splicing system is a mathematical model in DNA computing which was first introduced by Head in 1987 [1]. This model is defined under the framework of formal language theory which is a subcategory of theoretical computer science and applied discrete mathematics. The behavior of restriction enzymes on DNA molecules can be interpreted by splicing system. Also, splicing systems are important in terms of theoretical computer science and natural computing since they can be regarded as language generating devices. In this thesis, the characteristics of persistent and permanent splicing systems are investigated and some new concepts related to them are introduced. Persistent and permanent splicing systems are two types of splicing system that preserve the splicing operation and consequently, the enzymes cleavage process persists. Also, many restriction enzymes form persistent and permanent splicing systems. Moreover, the new concept of fuzzy splicing systems is introduced as an extension of Paun splicing system whereby its computational power has improved in comparison with Paun splicing systems.

1.2 Background of the Research

The concept of splicing system was first introduced by Head in 1987 to make a new connection between formal language theory and DNA computing. This mathematical model helps to interpret the behavior of restriction enzymes on DNA molecules when they are cut and pasted. Splicing systems were studied from many aspects. Also, several types of splicing systems and languages have been defined by different mathematicians. In some definitions it was considered from biological perspective and in some others from computation theory and its generative power. Two types of initial Head splicing systems are persistent and permanent splicing systems. In this research, the properties of them are investigated and two notions of self-closed and crossing-preserved splicing systems which are related to them. Also, the new concept of fuzzy splicing systems in terms of Paun splicing system is introduced as an enhanced generative power system. In fact, Paun splicing system is more related to computational aspects and its generative power is widely studied by many computer scientists.

1.3 Statement of the Problem

The focus of this research is to answer the following problems satisfactorily.

1. What are the necessary and sufficient conditions for splicing systems to be persistent and permanent?
2. How are the concepts of crossing-preserved, persistent and permanent splicing systems related?
3. How can the concept of persistent splicing system be presented in terms of Paun splicing system?
4. How can the generative power of splicing systems be increased?
5. How can some properties of persistent splicing systems be viewed biologically?

1.4 Research Objectives

This research embarks on the following objectives:

1. To study the characteristics of persistent and permanent splicing systems.
2. To introduce the new concepts of self-closed and crossing-preserved splicing systems and investigate their relation with persistent and permanent splicing systems.
3. To present the notion of persistent splicing systems in terms of Paun splicing systems and to explore some of their properties.
4. To increase the computational power of splicing systems through the introduction of the new concept of fuzzy splicing systems.
5. To perform laboratory experiments in illustrating and validating some properties of persistent splicing systems.

1.5 Scope of the Study

In this research, the characteristics of persistent and permanent splicing systems and languages are focused on. Also, some computational properties of fuzzy splicing systems are investigated.

1.6 Significance of Study

Biomathematics is a fast growing field in the world that has been gaining significant interest among the academic communities of Malaysia. The result of this research will lead to links of research between biologists and mathematicians. It can also establish and strengthen this interdisciplinary area in Malaysian academia. Moreover, the following benefits are expected to be obtained:

i. **Novel Theories/New Findings/Knowledge:**

The new concepts, theorems and propositions that will be developed for splicing systems and languages will contribute to knowledge in DNA computing and formal language theory.

ii. **Research Publications:**

The research has been presented in written reports, followed by publications in indexed national/ international proceedings and journals in theoretical computer science, mathematical biology and mathematics. (Refer to Appendix A)

iii. **Specific or Potential Applications:**

The short-term benefit of this result is that mathematics researchers can verify the validity of the results in wet-lab and apply it to actual biological studies. Also, the theoretical results can provide common research interest for the communities of mathematicians, computer scientists and biologists. The results obtained can be applied to research in bio-inspired computing. Its long-term benefit is to replace the long hours wet-lab work with the results of this mathematical modelling. Moreover, the results provide theoretical model for developing DNA based computers that can do massive parallel computations and solve some computationally hard problems

1.7 Research Methodology

This research has been conducted according to the eight phases as follows:

Phase 1: Literature review

To explore on the different concepts related to splicing systems, particularly, persistent and permanent splicing systems and languages. Also, to study the

biological concepts related to splicing systems and languages and to see how their properties are interpreted biologically.

Phase 2: Providing some necessary and sufficient conditions for splicing systems to be persistent and permanent.

To develop and prove some theorems and propositions which provide some necessary and sufficient conditions for splicing systems to be persistent and permanent.

Phase 3: Introducing the new concepts of self-closed and crossing-preserved

To introduce the new concepts of self-closed and crossing-preserved splicing systems and investigate their relations to persistent and permanent splicing systems in terms of theorems, corollaries, examples and counter-examples.

Phase 4: Presenting the concept of persistent splicing systems in terms of Paun splicing system

To present the notion of persistent extended H systems and investigate their properties and compare with the Head persistent splicing system.

Phase 5: Increasing the computational power of splicing systems with finite components

To introduce the concept of fuzzy splicing systems and to state their properties as lemmas and theorems. Also, to provide some fuzzy splicing examples that generate non regular languages.

Phase 6: Laboratory experiment

To conduct laboratory experiments to explore some properties of persistent splicing systems and investigate their behavior with one DNA strand and two chosen restriction enzymes that are used in three different reactions.

Phase 7: Thesis writing and presentation.

To write up, submit and present PhD thesis.

1.8 Thesis Outlines

This thesis is structurally organized into seven main chapters. The first chapter provides the general introduction to research conducted in this thesis. It includes the background of research, problem statement, objectives, scope, significance of the research and methodology.

In Chapter 2, some biological concepts related to splicing systems are presented. It begins with an overview of deoxyribonucleic acid (DNA) and restriction enzymes structure, follows by an introduction on the concept of splicing systems and their biological interpretation. It also provides a historical review of splicing systems. This chapter ends with preliminaries and basic definitions and concepts that are used in this thesis.

In Chapter 3, the characteristics of different splicing systems and languages are investigated. It is explored on how splicing languages can be generated by different types of splicing systems. Also, some closure properties of persistent and permanent splicing systems are studied. It ends with some theorems that provide necessary and sufficient conditions for splicing systems to be persistent and permanent.

Chapter 4 presents some new concepts that are closely connected to persistent and permanent splicing systems. This includes the concepts of self-closed, crossing-preserved and extended crossing-preserved splicing systems. It is followed by the investigation of their closure properties. Then, the relation between those new notions with persistent and permanent splicing systems is presented in terms of theorems and examples. Also, two special kinds of splicing systems are introduced as common-crossing and blunt splicing systems. The persistency properties of these two kinds of splicing systems, their relation with the concepts of self-closed and crossing-preserved splicing systems are stated in some theorems.

In Chapter 5 the concept of persistent splicing system is presented in terms of Paun splicing system and its properties are studied. Then the new notion of fuzzy splicing systems is introduced with general fuzzy operations. Some properties of multiplication, min and max fuzzy splicing systems are presented in different theorems. This chapter ends with some examples that illustrate how fuzzy threshold languages can increase the computational power of splicing systems and produce some non regular languages.

Chapter 6 focuses on the biological behavior of splicing systems. The results of some conducted experimental work are discussed in this chapter. Here, DNA strand and two restriction enzymes are chosen for three different reactions. The results of these three experiments are studied and compared and it is shown that the laboratory results are the same as the theoretical predictions by splicing system model.

Finally, the last chapter presents the summary and conclusion of this research. Some suggestions for future research on the persistent splicing systems as well as fuzzy splicing system are given in this chapter.

REFERENCES

1 Head, T. Formal language theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors. *Bulletin of Mathematical Biology*. 1987. 49(6): 737-759.

2. Turaev, S., Gan, Y. S., Othman, M., Sarmin, N. H. and Fong, W. H. Weighted Splicing Systems. In: Z. Li, *et al.* eds. *Computational Intelligence and Intelligent Systems*. Springer Berlin Heidelberg. 416-424; 2012.

3. Tamarin, R. H. *Principles of Genetics*: WCB/McGraw-Hill. 1999.

4. Paun, G., Rozenberg, G. and Salomaa, A. *DNA Computing: New Computing Paradigms*: Springer. 2006.

5. Bhd., R. B. S. New England Biolabs 2007-08 Catalog & Technical Reference: USA: Catalogue. 2007.

6. Linz, P. *An Introduction to Formal Languages and Automata*: Jones & Bartlett Learning. 2011.

7. Rozenberg, G. and Salomaa, A. *Handbook of Formal Languages: Volume 1. Word, Language, Grammar*: Springer-Verlag GmbH. 1997.

8. Harrison, M. A. *Introduction to Formal Language Theory*: Addison-Wesley Pub. Co. 1978.

9. Takashi, Y. Learning Local Languages and Their Application to DNA Sequence Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1998. 20(10): 1067-1079.

10. Gatterdam, R. W. Algorithms for Splicing Systems. *SIAM J. Comput.* 1992. 21(3): 507-520.

11. Fong, W. H., Sarmin, N.H. and Yusof, Y. Persistent and Permanent Splicing System. *Proceedings* of *Second International Conference and Workshops on Basic and Applied Sciences & Regional Annual Fundamental Science Seminar (ICORAFSS 2009)*. The ZON Regency Hotel, Johor Bahru, Malaysia. 2009. 17-20.

12. Kim, S. Computational Modeling for Genetic Splicing Systems. *SIAM Journal on Computing*. 1997. 26(5): 1284-1309.
13. Head, T. Splicing representations of strictly locally testable languages. *Discrete Applied Mathematics*. 1998. 87(1–3): 139-147.
14. Sarmin, N. H. and Fong, W. H. Strictly Locally Testable Languages in Splicing Systems. *Proceeding of Simposium Kebangsaan Sains Matematik Ke-15 2007 (SKSM15)*. 4-6th June 2007. Hotel Concorde Shah Alam. 2007. 61-69.
15. Fong, W. H. *Modelling of Splicing Systems Using Formal Language Theory*. Ph.D. Thesis, Universiti Teknologi Malaysia; 2008
16. Sarmin, N. H. and Fong, W. H. Non Strictly Locally Testable Languages and Molecular Considerations on Splicing Systems. *Mathematics Department Technical Report, LT/M Bil*. 7/2006.
17. Goode, E. and Pixton, D. Splicing to the Limit. In: Jonoska, N., Păun, G. and Rozenberg, G. eds. *Aspects of Molecular Computing*. Springer Berlin Heidelberg. 189-201; 2004
18. Goode, E. and Reddy, K. J. Wet Splicing Systems. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. 1999. 48: 73-83.
19. Yusof, Y., Sarmin, N. H., Goode, T. E., Mahmud, M. and Fong, W. H. An Extension of DNA Splicing System. *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on*. 27-29 Sept. 2011. 2011. 246-248.
20. Yusof, Y., Sarmin, N. H., Goode, T. E., Mahmud, M. and Fong, W. H. Hierarchy of Certain Types of DNA Splicing Systems. *International Journal of Modern Physics: Conference Series*. 2012. 09: 271-277.
21. Yusof, Y. *DNA Splicing System Inspired by Bio Molecular Operations*. Ph.D. Thesis, Universiti Teknologi Malaysia; 2012
22. Adleman, L. Molecular Computation Of Solutions To Combinatorial Problems. *Science*. 1994. 266(5187): 1021-1024.
23. Lipton, R. DNA Solution of Hard Computational Problems. *Science*. 1995. 268(5210): 542-545.
24. Boneh, D., Dunworth, C., Lipton, R. J. and Sgall, J. On the Computational Power of DNA. *Discrete Appl. Math.* 1996. 71(1-3): 79-94.

25. Păun, G., Rozenberg, G. and Salomaa, A. Computing by Splicing. *Theoretical Computer Science*. 1996. 168(2): 321-336.
26. Pixton, D. Regularity of Splicing Languages. *Discrete Applied Mathematics*. 1996. 69(1–2): 101-124.
27. Păun, G. On the Splicing Operation. *Discrete Applied Mathematics*. 1996. 70(1): 57-79.
28. Bonizzoni, P., Ferretti, C., Mauri, G. and Zizza, R. Separating Some Splicing Models. *Information Processing Letters*. 2001. 79(6): 255-259.
29. Gatterdam, R. W. Splicing Systems and Regularity. *International Journal of Computer Mathematics*. 1989. 31(1-2): 63-67.
30. Goode, E. *Constants and Splicing Systems*. Ph.D. Thesis, Binghamton University; 1999
31. Bonizzoni, P., De Felice, C., Fici, G. and Zizza, R. On the Regularity of Circular Splicing Languages: A Survey and New Developments. *Natural Computing*. 2010. 9(2): 397-420.
32. De Felice, C., Fici, G. and Zizza, R. A Characterization of Regular Circular Languages Generated by Marked Splicing Systems. *Theoretical Computer Science*. 2009. 410(47–49): 4937-4960.
33. Bonizzoni, P., De Felice, C. and Zizza, R. A Characterization of (Regular) Circular Languages Generated by Monotone Complete Splicing Systems. *Theoretical Computer Science*. 2010. 411(48): 4149-4161.
34. Turaev, S., Selvarajoo, M., Selamat, M., Sarmin, N. H. and Fong, W. H. Probabilistic Splicing Systems. In: N. T. Nguyen, *et al.* eds. *Advanced Methods for Computational Collective Intelligence*. Springer Berlin Heidelberg. 259-268; 2013
35. Mitrana, V., Petre, I. and Rogojin, V. Accepting Splicing Systems. *Theoretical Computer Science*. 2010. 411(25): 2414-2422.