

MODELLING OF n -TH ORDER LIMIT
LANGUAGE USING AUTOMATA THEORY IN
DNA SPLICING SYSTEM

SITI HAJAR BINTI MOHD KHAIRUDDIN

MASTER OF SCIENCE

UNIVERSITI MALAYSIA PAHANG

SUPERVISOR'S DECLARATION

We hereby declare that we have checked this thesis and in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science.



(Supervisor's Signature)

Full Name : MUHAMMAD AZRIN BIN AHMAD

Position : SENIOR LECTURER

Date : 18/01/2022



(Co-supervisor's Signature)

Full Name : NORAZIAH BINTI ADZHAR

Position : SENIOR LECTURER

Date : 18/01/2022



STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

A handwritten signature in black ink, appearing to read 'Siti Hajar', is written above a horizontal line.

(Student's Signature)

Full Name : SITI HAJAR BINTI MOHD KHAIRUDDIN

ID Number : MSE19001

Date : 18/02/2022

MODELLING OF n -TH ORDER LIMIT LANGUAGE USING AUTOMATA
THEORY IN DNA SPLICING SYSTEM

SITI HAJAR BINTI MOHD KHAIRUDDIN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Master of Science

Centre for Mathematical Sciences
UNIVERSITI MALAYSIA PAHANG

JANUARY 2022

ACKNOWLEDGEMENTS

First and foremost, I want to express my gratitude to Allah, the Almighty, for showering His blessings on me during my research, allowing me to complete it smoothly.

I would like to convey my heartfelt appreciation to Dr. Muhammad Azrin bin Ahmad, my research supervisor, for enabling me to conduct research and giving essential advice during this project. He's taught me how to do research and explain my findings as simply as possible. Working and studying under his direction was a wonderful honour and privilege. I also want to thank my co-supervisor, Dr. Noraziah binti Adzhar, for her willingness to improve my writing process throughout my research. I am extremely grateful for what they have done to improve my thesis. Moreover, I would like to thank Assoc. Prof. Dr. Aizi Nor Mazila Ramli, an Associate Professor from Faculty of Industrial Sciences and Technology (FIST), UMP and her team for their willingness to help me during my lab work for almost three months at Enzyme Microb and Proteomic / Food and Feed Research Laboratory.

I am also indebted to the Ministry of Higher Education (MOHE) Malaysia for financial funding through the Fundamental Research Grant Scheme (FRGS), FGRS/1/2018/STG06/UMP/03/1.

I am extremely grateful to my parents and siblings for their love, prayers, caring, and sacrifices for educating and preparing me for my future. Not to forget my friends, I'll forever be grateful for the encouragement they gave me when I needed it.

ABSTRAK

Kajian sistem hiris-cantum berkembang pesat setelah Head mendedahkan dalam penyelidikannya mengenai pemodelan proses biokimia yang melibatkan asid deoksiribonukleik pada tahun 1987. Proses sistem hiris-cantum terdiri daripada potongan dan tampalan asid deoksiribonukleik pilin ganda. Bahasa hiris-cantum yang dihasilkan oleh sistem hiris-cantum dapat diklasifikasikan kepada tiga kategori: bahasa lengai, fana dan bahasa batas. Sebelum ini, Goode juga telah mengemukakan bahasa batas ke- n , tetapi beliau hanya menyatakan bahawa bahasa batas sebelumnya berlainan dengan bahasa batas semasa. Dalam penyelidikan ini, kajian bahasa batas ke- n dikaji dengan menyelidiki bilangan asid deoksiribonukleik pilin ganda pemula dan peraturan yang terlibat dalam sistem hiris-cantum. Penyelidikan ini terbahagi kepada tiga kategori: konsep bahasa batas ke- n , automata dan eksperimen biologi. Definisi bahasa batas ke- n diimprovisasi dan diselidiki dengan menggunakan bilangan asid deoksiribonukleik pilin ganda pemula dan peraturan yang mempunyai panjang yang sama. Kemudian, rajah automata digunakan untuk mengubah bahasa batas ke- n dan dihasilkan dengan menggunakan tatabahasa. Akhir sekali, bahasa batas ke- n dibincangkan dalam aspek biologi. Penyelidikan ini merangkumi dua eksperimen iaitu pencernaan majmuk, ligasi dan tindak balas rangkaian polimerase untuk merangkumi definisi bahasa batas ke- n , iaitu model bahasa batas ketiga dan keempat menggunakan tiga dan empat enzim sekatan. Enzim yang digunakan dalam eksperimen adalah *MspI*, *AciI*, dan *MseI* untuk Model 1 dan *AgeI*, *EagI*, *BspEI*, dan *AvrII* untuk Model 2.

ABSTRACT

The study of splicing systems swiftly grew after Head revealed in his research about modelling the biochemical process involving the deoxyribonucleic acid in 1987. The process of the splicing system consists of a cut and paste of the double-stranded deoxyribonucleic acid. Splicing language produced by the splicing system can be classified into three categories: inert, transient and limit language. Previously, Goode also has defined the n -th order limit language, but she just stated that the previous order of the limit language is distinct from the current order limit language. In this research, the n -th order limit language study is studied by investigating the number of initial strings and rules involved in the splicing system. This research is divided into three categories: the concept of n -th order limit language, automata and biological experiment. The definition of n -th order limit language is improvised and investigated using the number of initial strings and rules involved in the splicing system, where the rule must have the same length. Then, the n -th order limit language is transformed into an automaton diagram by using grammar. Lastly, the n -th order limit language is discussed from the biological aspects. This research provides two experiments that involve several procedures such as multiple digestion, ligation and Polymerase Chain Reaction to generalise the formation of n -th order limit language, which are third and fourth-order limit language models using three and four restriction enzymes, respectively. The enzymes used in the experiments are *MspI*, *AciI*, and *MseI* for Model 1 and *AgeI*, *EagI*, *BspEI*, and *AvrII* for Model 2.

TABLE OF CONTENT

DECLARATION	
TITLE PAGE	
ACKNOWLEDGEMENTS	ii
ABSTRAK	iii
ABSTRACT	iv
TABLE OF CONTENT	v
LIST OF TABLES	viii
LIST OF FIGURES	x
LIST OF SYMBOLS	xii
LIST OF ABBREVIATIONS	xiv
CHAPTER 1 INTRODUCTION	1
1.1 An Overview	1
1.2 Research Background	3
1.3 Research Questions	4
1.4 Research Objectives	4
1.5 Research Scope	4
1.6 Research Significance	5
1.7 Research Workflow	5
1.8 Thesis Outline	6
1.9 Conclusion	7
CHAPTER 2 LITERATURE REVIEW	8
2.1 Introduction	8
2.2 Deoxyribonucleic Acid	8

2.3	Splicing System	10
2.4	Splicing Language	17
	2.4.1 Limit Language	19
2.5	Grammar	21
2.6	Automata Theory	22
	2.6.1 Type of Finite State Machine	23
	2.6.2 Automaton Diagram	25
	2.6.3 Application of Automata Theory in DNA Splicing System	26
2.7	Conclusion	27
CHAPTER 3 PROPERTIES OF n-TH ORDER LIMIT LANGUAGE		29
3.1	Introduction	29
3.2	Preliminaries	29
	3.2.1 New Concept of n -th Order Limit language	29
3.3	Recognition of an n -th Order Limit Language	30
3.4	Conclusion	49
CHAPTER 4 APPLICATION OF AUTOMATA THEORY ON n-TH ORDER LIMIT LANGUAGE		51
4.1	Introduction	51
4.2	The Basic Concept of Automata Theory	52
	4.2.1 Regular Expression and Automaton Diagram	52
4.3	Application of Automata Theory on n -th Order Limit Language	53
4.4	Conclusion	66

CHAPTER 5 <i>n</i>-TH ORDER LIMIT LANGUAGE IN THE BIOLOGICAL ASPECTS	68
5.1 Introduction	68
5.1.1 Procedure and Equipment	68
5.2 A Mathematical Model of Head Splicing System for <i>n</i> -th Order Limit Language	70
5.3 Biological Laboratory Experiment	74
5.3.1 Prediction of Results of the Experimental Design	79
5.4 Results and Discussion	83
5.5 Conclusion	86
CHAPTER 6 CONCLUSION	88
6.1 Summary of Research	88
6.2 Suggestion for Future Research	89
REFERENCES	91
APPENDICES	95
APPENDIX A: LIST OF PUBLICATIONS	96
APPENDIX B: LIST OF CONFERENCES AND SYMPOSIUMS	97

LIST OF TABLES

Table 2.1	Development of Splicing System and the Case Covered by the Splicing System	17
Table 2.2	Notation of Automaton Diagram and Its Functions	25
Table 3.1	Summarisation of Cases	35
Table 3.2	Number of Combination of the String of the Language	42
Table 3.3	The Number of the Combination of the String of the Language and the Pattern of the Iteration Produce Based on the Number of the Combination of the String	42
Table 3.4	The Number of Pattern of the String of the Language	45
Table 3.5	The Number of the Pattern of the String of the Language and the Pattern of the Iteration Produce Based on the Number of the Pattern of the String	46
Table 4.1	Number of Looping in the Automaton Diagram	61
Table 4.2	The Number of the End State of the Automaton Diagram	65
Table 5.1	All Possible Generated String for Model 1	71
Table 5.2	Resulted Strings of DNA Splicing System for Model 1	72
Table 5.3	All Possible Generated String for Model 2	73
Table 5.4	Resulted Strings of DNA Splicing System for Model 2	73
Table 5.5	The PCR Recipe of $A-B-C-D$ Strand	75
Table 5.6	Forward and Reverse Primer of Strand $A-B-C-D$	75
Table 5.7	Thermocycling Conditions for a Routine PCR for Model 1	76
Table 5.8	Digestion and Ligation Recipe for Model 1	76
Table 5.9	The PCR Recipe of $F-G-H-I-J$ Strand	78
Table 5.10	Forward and Reverse Primer of $F-G-H-I-J$ Strand	78
Table 5.11	Thermocycling Conditions for a Routine PCR for Model 2	78
Table 5.12	Digestion and Ligation Recipe for Model 2	79
Table 5.13	The Fragment Size (bp) of Predicted Molecules for Model 1	80

Table 5.14	The Fragment Size (bp) of Predicted Molecules for Model 2	82
Table 5.15	Aliquots for Each Well for Model 1	84
Table 5.16	Aliquots for Each Well for Model 2	85

LIST OF FIGURES

Figure 2.1	DNA Structure	9
Figure 2.2	Structure of dsDNA	10
Figure 2.3	The Initial String of dsDNA contains Restriction Sites of <i>AbsI</i>	12
Figure 2.4	The Cutting Site of dsDNA producing 5'–overhang	12
Figure 2.5	The Initial String of dsDNA contains Restriction Sites of <i>AbsI</i> after the Cutting Process	12
Figure 2.6	The Initial String of dsDNA contains Restriction Sites of <i>AatII</i>	12
Figure 2.7	The Cutting Site of dsDNA producing 3'–overhang	13
Figure 2.8	The Initial String of dsDNA contains Restriction Sites of <i>AatII</i> after the Cutting Process	13
Figure 2.9	The Initial String of dsDNA contains Restriction Sites of <i>AanI</i>	13
Figure 2.10	The Cutting Site of dsDNA producing a Blunt End	13
Figure 2.11	The Initial String The Initial String of dsDNA contains Restriction Sites of <i>AanI</i> after the Cutting Process	13
Figure 2.12	Types of Finite State Machine	23
Figure 2.13	Summary of Literature Review	27
Figure 4.1	Framework for Application of Automata Theory on n -th Order Limit Language	51
Figure 4.2	Automaton Diagram for M_1	52
Figure 4.3	Automaton Diagram for First Order Limit Language: One Initial String and One Rule	54
Figure 4.4	Automaton Diagram for Second Order Limit Language: One Initial String and Two Rules.	55
Figure 4.5	Automaton Diagram for n -th Order Limit Language: One Initial String and p Rules.	56
Figure 4.6	Automaton Diagram for First Order Limit Language: Two Initial Strings and One Rule	58
Figure 4.7	Automaton Diagram for First Order Limit Language: Two Initial Strings and Two Rules	60

Figure 4.8	Automaton Diagram for Case 1 in Chapter 3 Theorem 3.1	62
Figure 4.9	Automaton Diagram for Case 2 in Chapter 3 Theorem 3.1	62
Figure 4.10	Automaton Diagram for Case 7 in Chapter 3 Lemma 3.1	63
Figure 4.11	Automaton Diagram for Case 8 in Chapter 3 Lemma 3.1	64
Figure 5.1	MyCycler Thermal Cycler	69
Figure 5.2	Electrophoresis Equipment	70
Figure 5.3	Blue LED Transilluminator	70
Figure 5.4	Predicted Image on Gel for PCR Products Lane 1: 1kb DNA Ladder, Lane 2: PCR Product for Model 1, Lane 3 PCR Product for Model 2	80
Figure 5.5	Predicted Image on Gel for Model 1 Lane 1: 1kb ladder; Lane 2: DNA after PCR; Lane 3: DNA Before Digestion and Ligation; Lane 4: DNA after Digestion and Ligation; Lane 5: DNA Incubated Overnight at 4°C	81
Figure 5.6	Predicted Image on Gel for Model 2. Lane 1: 1kb ladder; Lane 2: DNA after PCR; Lane 3: DNA Before Digestion and Ligation; Lane 4: DNA after Digestion and Ligation; Lane 5: DNA Incubated Overnight at 4°C	83
Figure 5.7	Gel Electrophoresis for PCR Product	84
Figure 5.8	Gel Electrophoresis for Model 1	85
Figure 5.9	Gel Electrophoresis for Model 2	86

LIST OF SYMBOLS

$\mu', \gamma', \varepsilon', \phi', \eta', \omega'$	Arbitrary String and a Concatenation of Symbols in A
\in	Element of
\emptyset	Empty Set
Σ	Input Alphabet
\cap	Intersect
L	Language
μL	Micro Litre
$\not\subset$	Not Subset
\cup	Union
$\blacktriangledown \cdots \cdots \blacktriangle$	Crossing of Restriction Site of Restriction Enzyme
\geq	Greater or Equal to
\leq	Less or Equal to
■	End of Theorems and Lemmas
a	Pairing Between Adenine and Thymine (A-T)
A	Adenine
bp	Base Pairs
c	Pairing Between Guanine and Cytosine (G-C)
C	Cytosine
g	Pairing Between Cytosine and Thymine (C-G)

G	Guanine
q_0	Start State
t	Pairing Between Thymine and Adenine (T-A)
T	Thymine

LIST OF ABBREVIATIONS

DFA	Deterministic Finite Automata
DNA	Deoxyribonucleic Acid
dsDNA	Double-Stranded DNA
FSA	Finite State Automaton
FSM	Finite State Machine
G-P	Goode-Pixton
NFA	Non-deterministic Finite Automata
Y-G	Yusof-Goode

REFERENCES

- Ahmad, M. A. (2016). *Second Order Limit Language and its Properties in Yusof-Goode Splicing System*. Ph.D. Thesis. Universiti Teknologi Malaysia.
- Ahmad, M. A., Sarmin, N. H., Abdul Wahab, M. F., Fong, W.H. and Yusof, Y. (2018). Biomolecular aspects of second order limit language. *Malaysian Journal of Fundamental and Applied Sciences*. 14(1): 15-19.
<http://dx.doi.org/10.11113/mjfas.v14n1.727>
- Ahmad, M. A., Sarmin, N. H., Fong, W. H., and Yusof, Y. (2014). A New Relation of Second Order Limit Language in Simple and Semi-Simple Splicing System. *Jurnal Teknologi*. 71(5): 13-15.
<https://doi.org/10.11113/jt.v71.3845>
- Ahmad, M. A., Sarmin, N. H., Fong, W. H. and Yusof, Y. (2015.) A Comparison of Second Order and Non-Second Order Limit Language Generated by Yusof-Goode Splicing System. *Jurnal Teknologi*. 72(1): 27-31.
<https://doi.org/10.11113/jt.v72.3062>
- Ahmad, M. A., Sarmin, N. H., Heng, F. W., and Yusof, Y. (2014). *Second Order Limit Language in Variants of Splicing System*. AIP Conference Proceedings, pp. 639-643.
<https://doi.org/10.1063/1.4887664>
- Alberts, B., Johnson, A., Lewis, J., Raff, M. and Roberts, W. P. (2017). *Molecular Biology of the Cell*. Garland: WW Norton and Company.
<https://doi.org/10.1201/9781315735368>
- Berg, J., Tymoczko, J. and L, S. (2002). *Biochemistry*. New York: W.H. Freeman and Company.
- Chomsky, N. (1959). On Certain Formal Properties of Grammars. *Information and Control*. 2(2): 137-167.
[https://doi.org/10.1016/s0019-9958\(59\)90362-6](https://doi.org/10.1016/s0019-9958(59)90362-6)
- Dwyer, C. L. and Lebeck, A.R. (2008). *An Introduction to DNA Self Assembled Computer Design*. Boston, London: Artech House, Inc.
- Fong, W. H. (2009). Recognition of simple splicing systems using SH automaton. *Malaysian Journal of Fundamental and Applied Sciences*. 2(4): 337–342.
<https://doi.org/10.11113/mjfas.v4n2.41>
- Freund, R., Kari, L. and Paun, G. (1999). DNA Computing Based on Splicing: The Existence of Universal Computers. *Theory of Computing Systems*. 32(1): 69–112.
<https://doi.org/10.1007/s002240000112>
- Gan, Y. S. (2015). *Finite Automata in DNA Splicing and Sticker Systems with Weights*.

- Ph.D. Thesis. Universiti Teknologi Malaysia.
- Goode, E., and Pixton, D. (2003). Splicing to the Limit: In Aspects of Molecular Computing Heidelberg, Berlin: Springer.
https://doi.org/10.1007/978-3-540-24635-0_13
- Goode, T. E. and DeLorbe W. (2007). *DNA Splicing Systems: An Ordinary Differential Equations Model and Simulation*. Proceedings of the 13th International Conference on DNA Computing, pp. 1-12.
http://dx.doi.org/10.1007/978-3-540-77962-9_25
- Guyton, A. (2016). *Textbook of Medical Physiology*. Philadelphia, PA: Elsevier Inc.
https://dx.doi.org/10.4103%2Fsni.sni_327_17
- Head, T. (1987). Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors. *Bulletin of Mathematical Biology*. 49(6): 737–759.
[https://doi.org/10.1016/S0092-8240\(87\)90018-8](https://doi.org/10.1016/S0092-8240(87)90018-8)
- Head, T. (1997). Splicing Languages Generated With one Sided Context. Working Paper. Singapore: Springer-Verlag.
- Hopcroft, J. E., Motwani, R. and Ullman, J.D. (2013). *Introduction to Automata Theory, Languages and Computation: Pearson New International Edition*. London: Pearson Education.
<https://doi.org/10.1145/568438.568455>
- Karimi, F., Turaev, S., Sarmin, N. H. and Fong, W. H. (2014). *Fuzzy Splicing Systems*. International Conference on Computational Collective Intelligence, pp. 20-29.
http://dx.doi.org/10.1007/978-3-319-11289-3_3
- Kumar, K. S., and Malathi, D. (2015). A Novel Method to Construct Deterministic Finite Automata From a Given Regular Grammar. *Journal of Science, Engineering and Research*. 6(3): 106-111.
- Lim, S. J. (2011). *Mathematical Modelling of Some Uniform and Null Context Splicing Systems*. Master of Science Thesis. Universiti Teknologi Malaysia.
- M. Khairuddin, S. H., Ahmad, M. A. and Adzhar, N. (2020). Effect of m Number of Initial String on the n -th Order Limit Language. *Technology Reports of Kansai University*. 62(07): 3681–3689.
- Mealy, G. H. (1955). A Method for Synthesizing Sequential Circuits. *Bell System Technical Journal*. 34(5):1045–1079.
<https://doi.org/10.2307/2963669>
- Mohamad Jan, N., Fong, W. H. and Sarmin, N. H. (2013). *Regular Languages, Regular Grammars and Automata in Splicing Systems*. 20th National Symposium on

- Mathematical Sciences: Research in Mathematical Sciences: A Catalyst for Creativity and Innovation, pp. 856–863.
<https://doi.org/10.1063/1.4801217>
- Moore, E. F. (1956.) *Automata Studies: Gedenken- Experiment on Sequention Mathematics*. Princeton: Princeton University Press.
<https://doi.org/10.1515/9781400882618-006>
- Mudaber, M. H., Yusof, Y. and Mohamad, M. S. (2017). Differentiating the Persistency and Permanency of Some Two Stages DNA Splicing Language via Yusof-Goode (Y-G) approach. *Journal of Physics: Conference Series*. 890(1): 1-10.
<http://dx.doi.org/10.1088/1742-6596/890/1/012081>
- New England BioLab. Enzyme Finder (online). <https://enzymefinder.neb.com> (22 April 2019).
- Paun, G. and Rozenberg, A. S. (1998). *DNA computing: New Computing Paradigms*. Berlin Heidelberg: Springer-Verlag.
[https://doi.org/10.1016/S0303-2647\(99\)00031-3](https://doi.org/10.1016/S0303-2647(99)00031-3)
- Paun, G. (1996). On the Splicing Operation. *Discrete Applied Mathematics*. 70(1996): 57-79.
[https://doi.org/10.1016/0166-218X\(96\)00101-1](https://doi.org/10.1016/0166-218X(96)00101-1)
- Peter, L. (2017). *An Introduction to Formal Languages and Automata (6th ed.)*. USA: Jones and Bartlett, LLC.
https://doi.org/10.1142/9789812817495_0002
- Pixton, D. (1996). Regularity of Splicing Languages. *Discrete Applied Mathematics*. 69(1-2): 101-124.
[https://doi.org/10.1016/0166-218X\(95\)00079-7](https://doi.org/10.1016/0166-218X(95)00079-7)
- Rich, E. A. (2008). *Automata, Computability and Complexity: Theory and Applications*. London: Pearson Education.
- Russell, P. (2001). *iGenetic*. New York: Benjamin Cummings.
- Santhanam, V. (2017). Pattern Matching using Computational and Automata Theory. *International Research Journal of Engineering and Technology*. 10(4): 1295–1299.
- Watson, J. D. and Crick, F. H. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. 171(4356): 737-738.
<https://doi.org/10.1038/171737a0>
- Wright, D. R. (2005). *Finite State Machines*. Ph.D. Thesis. Carolina State University, USA.

Yan, S. Y. (1998). *An Introduction to Formal Languages and Machine Computation*. Singapore: World Scientific Publishing Co. Pte. Ltd.
<https://doi.org/10.1142/2659>

Yusof, Y. (2012). *DNA Splicing System Inspired by Bio Molecular Operation*. Ph.D. Thesis Universiti Teknologi Malaysia.