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.

(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 deoksibonukleik 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 deoksibonukleik 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 deoksibonukleik 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 rantaian 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 *Msp*I, *Aci*I, dan *Mse*I untuk Model 1 dan *Age*I, *Eag*I, *BspE*I, dan *Avr*I 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 *Msp*I, *Aci*I, and *Mse*I for Model 1 and *Age*I, *Eag*I, *BspE*I, and *AvrI*I for Model 2.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mu', \gamma', \varepsilon', \phi', \eta', \omega'$	Arbitrary String and a Concatenation of Symbols in A
$\in$	Element of
$\varnothing$	Empty Set
Σ	Input Alphabet
$\cap$	Intersect
L	Language
μL	Micro Litre
$\not\subset$	Not Subset
$\cup$	Union
▼… …▲	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

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

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

Paun, G. (1996). On the Splicing Operation. *Discrete Applied Mathematics*. 70(1996): 57-79.
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

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.