NEW VARIANTS OF INSERTION AND DELETION SYSTEMS IN FORMAL LANGUAGES

AHMAD FIRDAUS YOSMAN

UNIVERSITI TEKNOLOGI MALAYSIA

NEW VARIANTS OF INSERTION AND DELETION SYSTEMS IN FORMAL LANGUAGES

AHMAD FIRDAUS YOSMAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Philosophy

Faculty of Science
Universiti Teknologi Malaysia

AUGUST 2017

To those who continued to believe in me

ACKNOWLEDGEMENT

First and foremost, I would like to give praise and thanks to Allah S.W.T; for without His will and power, none of this would have been possible.

Next, I would like to extend my utmost gratitude to my esteemed supervisor, Dr. Fong Wan Heng for believing in me and in my research. Her guidance and support has pulled me through these grueling times. Not forgetting my co-supervisor, Dr. Sherzod Turaev, who was always prepared to discuss ideas and avenues for the progress of the research.

Other than that, I would like to acknowledge Dr. Bianca Truthe for providing a warm welcome and great hospitality during my time at University of Giessen, Germany. She and Professor Markus Holzer were instrumental in helping me produce wonderful results for the research.

I would like to thank my friends and fellow members of the Applied Analysis and Algebra Group (AAAG) for the support and companionship throughout my research. Every moment we had spent together has been a learning experience. Furthermore, I am very grateful to the participants of the Fourth Biennal International Group Theory Conference (4BIGTC2017) for their invaluable feedback and suggestions.

Lastly, my deepest appreciation goes to my loving parents and brothers.

ABSTRACT

In formal language theory, the operations of insertion and deletion are generalizations of the operations of concatenation and left/right quotients. The insertion operation interpolates one word in an arbitrary place within the other while the deletion operation extracts the word from an arbitrary position of another word. Previously, insertion and deletion have been applied to model the recombinance of DNA and RNA molecules in DNA computing, where contexts were used to mimic the site of enzymatic activity. However, in this research, new systems are introduced by taking motivation from the atomic behaviour of chemical compounds during chemical bonding, in which the concept of a balanced arrangement is required for a successful bonding. Besides that, the relation between insertion and deletion systems and group theory are also shown. Here, insertion and deletion systems are constructed with bonds and also interactions; hence new variants of insertion and deletion systems are introduced. The first is bonded systems, which are introduced by defining systems with restrictions that work on the bonding alphabet. The other variant is systems with interactions, which are introduced by utilizing the binary operations of certain groups as the systems' interactions. From this research, the generative power and closure properties of the newly introduced bonded systems are determined, and a language hierarchy is constructed. In addition, group generating insertion systems are introduced and illustrated using Cayley graphs. Therefore, this research introduced new variants of insertion and deletion systems that contribute to the advancement of DNA computing and also showcased their application in group theory.

ABSTRAK

Dalam teori bahasa formal, operasi penyisipan dan pengguguran ialah pengitlakan kepada operasi penjeraitan dan hasil bahagi kiri/kanan. Operasi penyisipan menginterpolasi suatu kata di sebarangan posisi dalam suatu kata yang lain manakala operasi pengguguran pula mengeluarkan suatu kata daripada sebarangan posisi dalam suatu kata yang lain. Sebelum ini, penyisipan dan pengguguran telah digunakan untuk mengilustrasikan penggabungan semula molekul DNA dan RNA dalam bidang pengkomputeran DNA, di mana konteks telah digunakan untuk mengajuk laman aktiviti enzim. Namun, dalam penyelidikan ini, sistem-sistem baharu diperkenalkan dengan mengambil motivasi daripada tingkahlaku atomik sebatian kimia semasa pengikatan kimia, di mana penyusunan hendaklah terimbang untuk memastikan ikatan yang berjaya. Selain itu, kaitan sistem penyisipan dan pengguguran dengan teori kumpulan juga ditunjukkan. Di sini, sistem penyisipan dan pengguguran dibina dengan ikatan dan juga interaksi; justeru varian-varian baharu sistem-sistem penyisipan dan pengguguran diperkenalkan. Pertama, sistem-sistem terikat diperkenalkan dengan mewujudkan sistem-sistem yang mempunyai batasan yang bertindak atas abjad mengikat. Varian selainnya ialah sistem-sistem dengan interaksi yang diperkenalkan dengan menggunakan operasi biner kumpulan tertentu sebagai interaksi kepada sistem-sistem tersebut. Daripada penyelidikan ini, kuasa penjanaan dan sifat tertutup sistem-sistem terikat yang baharu diperkenal telah dikenalpasti dan hierarki bahasa dibina. Tambahan lagi, sistem-sistem menjana kumpulan telah diperkenalkan dan digambarkan dengan menggunakan graf Cayley. Oleh itu, penyelidikan ini telah memperkenalkan varian sistem-sistem penyisipan dan pengguguran baharu yang menyumbang kepada kemajuan pengkomputeran DNA dan menunjukkan aplikasi sistem-sistem tersebut dalam teori kumpulan.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ipINS-system	-	Bonded Indian parallel insertion system
bPDEL-system	-	Bonded parallel deletion system
bPINS-system	-	Bonded parallel insertion system
bSDEL-system	-	Bonded sequential deletion system
bSINS-system	-	Bonded sequential insertion system
upINS-system	-	Bonded uniformly parallel insertion system
D0L-system	-	Deterministically interactionless Lindenmayer system
DT0L-system	-	Deterministically tabled interactionless Lindenmayer system
DFA	-	Deterministic finite automaton
ED0L-system	-	Extended deterministically interactionless Lindenmayer system
EDT0L-system	-	Extended deterministically tabled interactionless Lindenmayer system
E0L-system	-	Extended interactionless Lindenmayer system
ET0L-system	-	Extended tabled interactionless Lindenmayer system
insdel-system	-	Insertion-deletion system
0L-system	-	Interactionless Lindenmayer system
NFA	-	Nondeterministic finite automaton
∗SINS-system	-	Sequential insertion system with interaction
T0L-system	-	Tabled interactionless Lindenmayer system

LIST OF SYMBOLS

A_n	-	Alternating group of order n
α	-	Axiom word
$*$	-	Binary operation
η	-	Bonded deletion system
γ	-	Bonded insertion system
$\mathcal{B}_\Sigma$	-	Bonding alphabet
$\lvert S\rvert$	-	Cardinality of the set S
$\mathbb{Z}_n$	-	Cyclic group of order n
$\Rightarrow$	-	Derivation relation
β	-	Derived word
D_n	-	Dihedral group of order $2n$
$G \times H$	-	Direct product of G and H
$\in$	-	Element of
$\emptyset$	-	Empty set
λ	-	Empty word
$\mathcal{L}(X)$	-	Family of languages generated by X
F	-	Fluorine atom
$>$	-	Greater than
$\geq$	-	Greater than or equal to
δ_i	-	Insertion word
a^*	-	Kleene closure on the symbol a
$\lvert w\rvert$	-	Length of the word w
$<$	-	Less than
$\leq$	-	Less than or equal to
Li	-	Lithium atom

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Formal language theory is the study of the syntax of formal languages, primarily used as a basis for defining the grammars of programming languages. By breaking down programming languages into their general characteristics, a formal language can be formed [1]. The formation of languages depends on certain production rules or grammars, which are determined by operations on symbols or letters in an alphabet. So far, numerous operations have been studied, which can roughly be classified into three different classes [2]: the class of set operations, comprising union, intersection, and complementation; the class of algebraic operations, which include morphism and substitution; and lastly the class of purely language theoretical operations, namely concatenation, quotient, and Kleene closure.

Kari developed generalizations for the operations of concatenation and quotient, which are insertion and deletion, respectively [3]. Since the operation of concatenation only allows for addition of symbols or words at the rightmost extremity of a given word, Kari had introduced the insertion operation, whereby symbols or words may be added in any place in an initial word. Similarly, there had only been two ways of removing symbols or words from an initial word, which is either left quotient or right quotient, where, the symbol or word can only be removed from the left or right extremities. Hence, Kari had introduced the deletion operation, which enabled removal of symbols or words from any place in a given word. Kari further enhanced her findings by introducing variants of the operations of insertion and deletion [3]. From there, many new findings had been obtained, which contributed to the rapidly blossoming field of DNA computing. Even so, it will be shown in this research that not only do insertion and deletion systems contribute to the advancement in DNA computing, but insertion systems can also be utilized to generate algebraic structures, specifically groups.

1.2 Research Background

Computation models of DNA recombination utilize formal languages due to the similarity of DNA bases with symbols in an alphabet. Over the years, numerous computation models have been introduced, mainly variants of splicing systems, sticker systems or insertion-deletion systems. By starting from the set of axioms, a myriad of languages that possess different powers can be generated by a set of rules as defined in the systems.

The field of DNA computing has been experiencing substantial advancement due to the keen interest of many researchers. This is due to the interdisciplanary nature of the field and its potential real-world applications. Not only that, DNA computation models have also shown to be receptive of mathematical inputs. For instance, groups and probabilities have been used as weights which are implemented onto systems, subsequently increasing the generative power of the original systems [4, 5].

With that being said, Kari and Thierrin in [6] introduced contextual insertion and deletion, where triples are defined so that any insertion or deletion occurs in between two consecutive symbols called contexts. These operations were used in concert to form insertion-deletion systems, which are a close model of the insertion and deletions that occur at restriction sites on DNA strands. These systems have been shown to generate recursively enumerable languages, which are of the highest power in the Chomsky hierarchy. However, they do not take into account the occurrences at the atomic level of DNA recombination. In this research, the modelling of DNA recombination is taken a step deeper, where new variants of insertion and deletion systems are introduced to model the atomic behavior of chemical compounds (such as DNA molecules) during chemical bonding in the process of DNA recombination.

On the other hand, the notion of generating languages transfers seamlessly into group theory, where the generation of groups can be done by insertion systems, as well. It is well-known that groups can be generated by either a set of generators or by the repeated operations among its elements with respect to the binary operation. The closure property of groups mean that no matter how many times the elements act upon each other, the output will still be contained within the group, hence, a group can be considered as a language. From there, insertion systems are introduced to generate languages that equal to groups.

1.3 Problem Statement

The problem statements of this research are to impose some restrictions to the rules of the derivations of insertion and deletion systems in order to develop bonded variants and variants that are able to generate groups. Here, the generative power of the bonded insertion and deletion systems is determined according to the Chomsky hierarchy. Not only that, visual representations of the group generating variant of insertion systems are also presented by using Cayley graphs. Hence, the findings of this research can answer the following questions:

1. How to construct bonded insertion and deletion systems?
2. What is the generative power of bonded insertion and deletion systems?
3. How to construct insertion systems that can generate groups?
4. What are the Cayley graphs of insertion systems that generate groups?

1.4 Objectives of the Research

The objectives of this research are:

1. To investigate the operations of insertion and deletion in formal languags.
2. To introduce bonded insertion and deletion systems by imposing restrictions on the rules of the derivations.
3. To determine the generative power of bonded insertion and deletion systems according to the Chomsky hierarchy and Lindenmayer systems.
4. To introduce insertion systems that generate groups.
5. To construct the Cayley graph of the generation of groups using insertion systems.

1.5 Scope of the Research

In this research, two variants of insertion and deletion are considered, which are sequential insertion and deletion, and parallel insertion and deletion, such that bonded systems of each variant are introduced. These bonded systems utilize the concepts of atomic behavior of chemcial compunds (such as DNA molecules) during chemical bonding in the process of DNA recombination. Furthermore, the families of languages in the Chomsky hierarchy and Lindenmayer systems are used to compare the languages generated by bonded insertion and deletion systems to determine their generative power.

In addition, fundamental concepts in group theory are used to introduce a variant of insertion systems that can generate groups. Moreover, the concepts of finitely generated groups and Cayley graphs are used to visually represent the generation of some groups using insertion systems.

1.6 Significance of the Research

Insertion and deletion systems have gained increasing interest in recent years due to the advancement of computer technology. Given their high significance in serving as mathematical models of biomolecular activities, insertion and deletion systems are pivotal in the field of DNA computing. Therefore, introducing insertion and deletion systems that relate to the atomic behavior during bio-processes will give real-world computation models. Besides that, this research also contributes to the field of group theory, in which insertion systems are introduced to generate languages that equal to finite groups. From there, the Cayley graphs depicting the derivation of elements in a group are presented. Lastly, this research will further strengthen the relationship between researchers from the field of mathematics, computer science and biochemsitry, thus enabling many more interdisciplinary discoveries in science and technology.

1.7 Research Methodology

This research begins with the study on some fundamental concepts in biochemistry, formal languages, and group theory. Firstly, the concept of chemical bonds between biomolecular structures, such as DNA molecules is studied. The atomic behavior of DNA molecules during chemical bonding in the process of DNA recombination provides the motivation to introduce bonded systems. Then, bonded insertion and deletion systems are introduced by relating the aforementioned atomic behavior of DNA molecules to the operations of insertion and deletion. This is reflected in the construction of the systems that work on symbols with bonds attached to them. A derivation is successful if and only if the output words are balanced, similar to a stable octet arrangement of electrons in chemical compounds. The generative power of the bonded systems are then determined according to the Chomsky hierarchy and hierarchy of Lindenmayer systems. Next, sequential insertion systems with interactions are introduced to generate groups by imposing an additional restriction onto the derivations. This is done by ensuring that the inserted symbol interacts with the axiom word using the binary operation of the group of interest. From there, the Cayley graphs of the systems are presented to visually represent the generations of the groups using the sequential insertion systems with interactions. Figure 1.1 depicts the flow of the research.

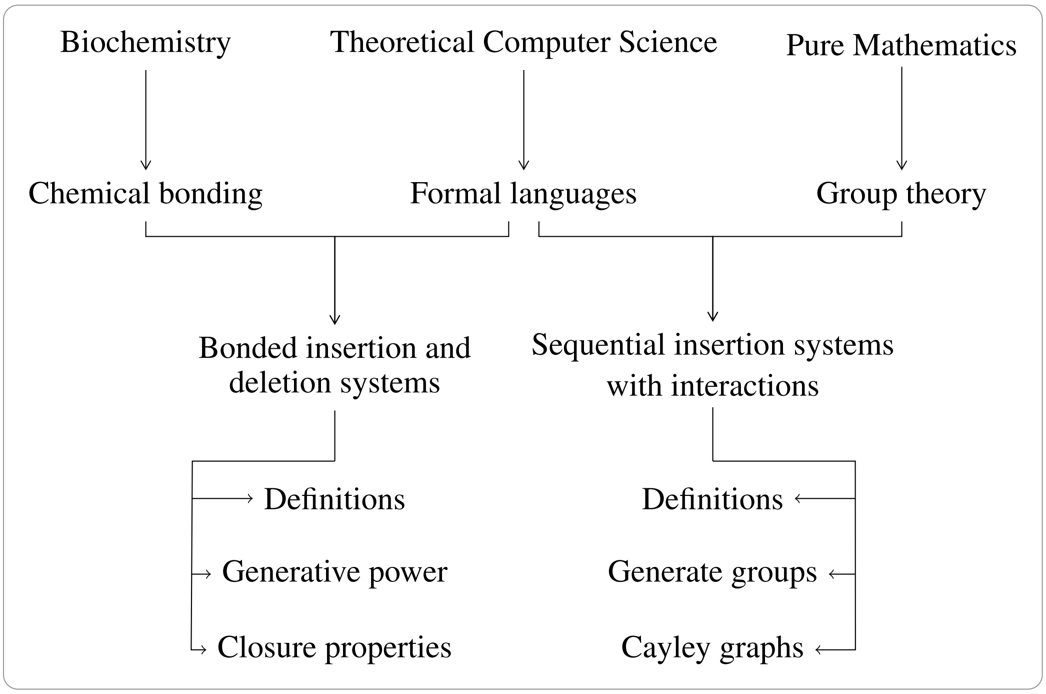

Figure 1.1 Flow chart of the research

1.8 Organization of the Thesis

The introduction to the thesis is provided in Chapter 1. Here, the research background and problem statement are explained. In this chapter, the objectives, scope, significance, and methodology are also provided. Lastly, the organization of the thesis is presented.

The literature review of the research is presented in Chapter 2. Firstly, the historical background of insertion and deletion systems is given, where the progress of the systems is told from its early conception to current works. After that, some concepts in formal languages are explained, which include notations, constructions, definitions, and basic hierarchies of language families. Next, some concepts in group theory are presented, which include definitions of groups and Cayley graphs.

In Chapter 3, results on bonded insertion systems are presented. Variants of bonded insertion systems are introduced and their generative powers are determined. The variants include bonded sequential, parallel, Indian parallel, and uniformly parallel insertion systems. A hierarchy of families of languages generated by bonded insertion systems with respect to the Chomsky hierarchy and Lindenmayer systems hierarchy is construced in the end.

Similarly, results on bonded deletion systems are presented in Chapter 4. Variants of bonded deletion systems are introduced and their generative powers are determined along with some closure properties. The variants include bonded sequential and parallel deletion systems. A hierarchy of families of languages generated by bonded deletion systems with respect to the Chomsky hierarchy and Lindenmayer systems is also construced in the end.

In Chapter 5, the concept of sequential insertion systems with interactions is introduced. Here, the generation of finite cyclic groups, dihedral groups, the quaternion group, and symmetric groups is shown. The Cayley graphs corresponding to the systems are constructed, followed by the introduction of simple sequential insertion systems.

Lastly, the conclusion of the research along with suggestions for future work are provided in Chapter 6.

The organization of the thesis is shown in Figure 1.2.

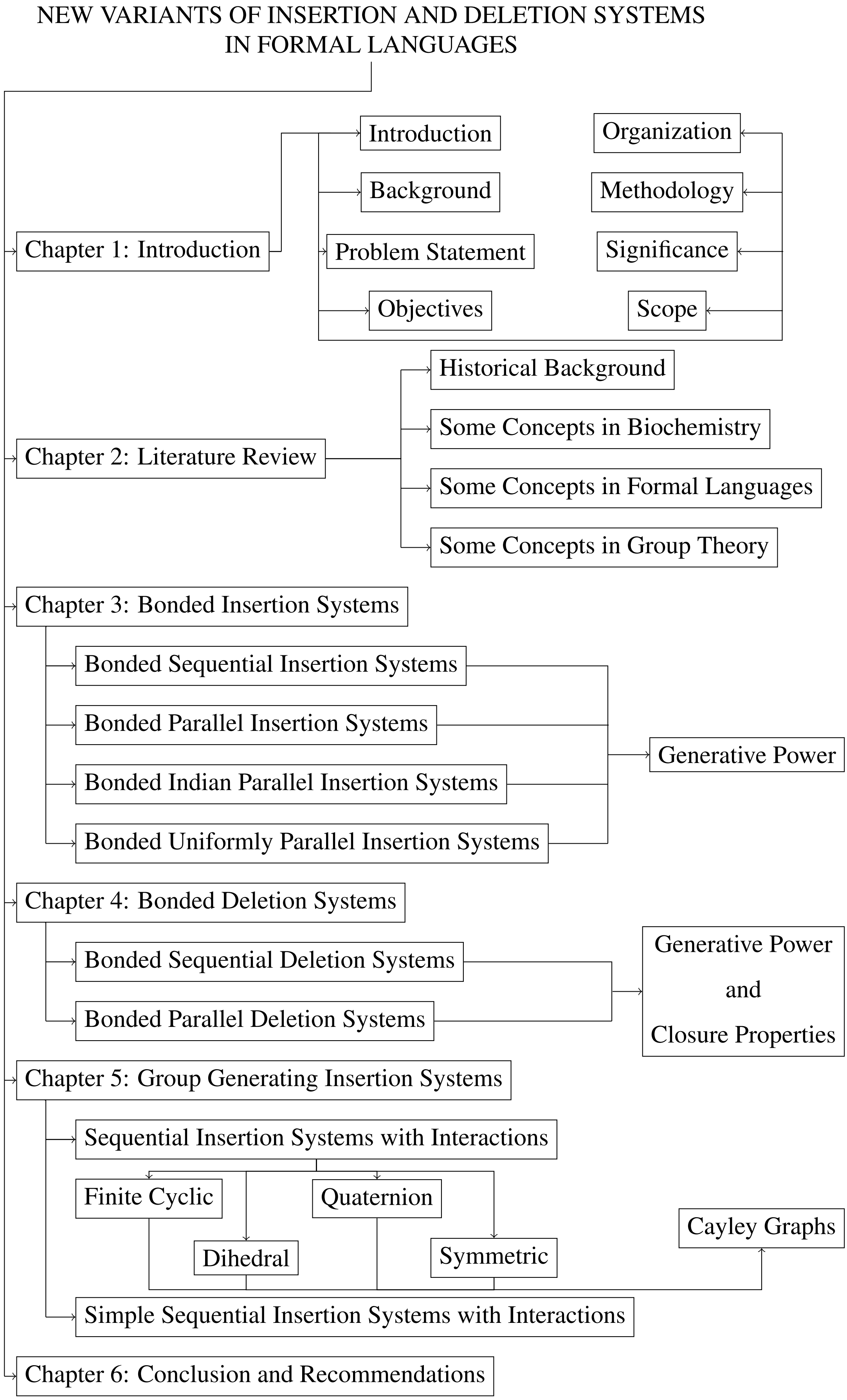

Figure 1.2 Organization of the thesis

1.9 Conclusion

DNA computing is a new advent for advancement in science and technology, where one of the main drivers is the study of insertion and deletion systems. From the formulation of insertion and deletion comes many studies on the various ways alphabets in a language can be acted upon, which has found its way to applications in DNA computing. The coherence of formal language theory and informational macromolecules brought forth new paradigms, and with it a deeper understanding of the recombinant behavior of DNA molecules. Firstly, this research introduces bonded insertion and deletion systems as a reflection on the natural occurrence in DNA recombination, wherein some form of computation is needed to ensure a successful insertion (or deletion) of a DNA molecule into (or out of) a DNA strand. Since the formalization is done within the environment of formal language theory, the generative power of the computed variants of insertion and deletion will be determined. Next, insertion systems that enable the generation of groups are introduced. These systems utilize the concept of binary operations between elements in a group as a way for derivations from one word to another. The generation of languages that equal to groups is visually represented by Cayley graphs.

REFERENCES

1. Linz, P. *An Introduction to Formal Languages and Automata*. 4th ed. Sudbury, MA: Jones and Bartlett Publishers. 2006.
2. Santean, L. Six Arithmetic-Like Operations on Languages. *Revue Roumaine de Linguistique*. 1988. 65-73.
3. Kari, L. *On Insertion and Deletion in Formal Languages*. Ph.D. Thesis. University of Turku; 1991.
4. Gan, Y. S. *Finite Automata in DNA Splicing and Sticker Systems with Weights*. Ph.D. Thesis. Universiti Teknologi Malaysia; 2015.
5. Selvarajoo, M. *Probabilistic Splicing and Sticker Systems in DNA Computing*. Ph.D. Thesis. Universiti Teknologi Malaysia; 2015.
6. Kari, L. and Thierrin, G. Contextual Insertions/Deletions and Computability. *Information and Computation*. 1996. 131(1): 47-61.
7. 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.
8. Adleman, L. Molecular Computations of Solutions to Combinatorial Problems. *Science*. 1994. 226(5187): 1021-1024.
9. Kari, L. Insertion and Deletion of Words: Determinism and Reversibility. *Mathematical Foundations of Computer Science*. 1992. 629: 315-327.
10. Kari, L. Generalized Derivatives. *Fundamenta Informaticae*. 1993. 18 (1): 61-77.
11. Kari, L., Mateescu, A., Păun, G., and Salomaa, A. Deletion Sets. *Fundamenta Informaticae*. 1993. 19: 355-370.
12. Kari, L. Insertion Operation: Closure Properties. *European Association of Theoretical Computer Science (EATCS) Bulletin*. 1993. 51: 181-191.

13. Kari, L. Deletion Operations: Closure Properties. *International Journal of Computer Mathematics*. 1994. 52: 23-42.

14. Kari, L., Mateescu, A., Păun, G., and Salomaa, A. On Parallel Deletions Applied to a Word. *RAIRO – Theoretical Informatics and Applications*. 1995. 29(2): 129-144.

15. Ito, M., Kari, L., and Thierrin, G. Insertion and Deletion Closure of Languages. *Theoretical Computer Science*. 1997. 183: 3-19.

16. Păun, G., Rozenberg, G., and Salomaa, A. *DNA Computing: New Computing Paradigms*. Berlin: Springer-Verlag Berlin Heidelberg. 1998.

17. Daley, M., Kari, L., Gloor, G., and Siromoney, R. Circular Contextual Insertions/Deletions with Applications to Biomolecular Computation. *Proceedings of International Symposium on String Processing and Information Retrieval (SPIRE1999)*. September 21-24, 1999. Los Alamitos, California: IEEE Computer Society Press. 1999. 47–54.

18. Kari, L. and Thierrin, G. At the Crossroads of DNA Computing and Formal Languages: Characterizing RE using Insertion-Deletion Systems. *Proceedings of DNA Based Computers III*. June 16-18, 1997. Providence, Rhode Island: American Mathematical Society. 1999. 329-347.

19. Krassovitskiy, A. *Complexity and Modeling Power of Insertion-Deletion Systems*. Ph.D. Thesis. Universitat Rovira i Virgili; 2011.

20. Fernau, H., Kuppusamy, L., and Raman, I. Descriptional Complexity of Graph-Controlled Insertion-Deletion Systems. *Descriptional Complexity of Formal Systems*. 2016. 111-125.

21. Freund, R., Kogler, M., Rogozhin, Y., and Verlan, S. Graph-Controlled Insertion-Deletion Systems. *Proceedings of the Twelfth Annual Workshop on Descriptional Complexity of Formal Systems*. August 8-10, 2010. Waterloo, New South Wales: Open Publishing Association. 2010. 31: 88-98.

22. Cui, B., Kari, L., and Seki, S. Block Insertion and Deletion on Trajectories. *Theoretical Computer Science*. 2011. 412: 714-728.

23. Petre, I. and Verlan, S. Matrix Insertion-Deletion Systems. *Theoretical Computer Science*. 2012. 456: 80-88.

24. Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. *Campbell Biology*. 10th ed. Amsterdam: North-Holland/America Elsevier. 1975.

25. Rozenberg, G. and Salomaa, A. *Handbook of Formal Languages*. Berlin: Springer-Verlag Berlin Heidelberg. 1997.

26. Rozenberg, G. and Salomaa., A. *The Mathematical Theory of L-systems*. Cambridge, MA: Academic Press. 1980.

27. Duchon, P. On the Enumeration and Generation of Generalized Dyck Words. *Discrete Mathematics*. 2000. 225: 121-135.

28. Fraleigh, J. B. *A First Course in Abstract Algebra*. 7th Edition. Boston, MA: Addison Wesley. 2003.

29. Meier, J. *Groups, Graphs and Trees: An Introduction to the Geometry of Infinite Groups*. Cambridge: Cambridge University Press. 2008.

30. Ito, M., Kari, L., and Thierrin, G. Shuffle and Scattered Deletion Closure of Languages. *Theoretical Computer Science*. 2000. 245(1): 115-133.

31. Hermann, G. T. and Rozenberg, R. *Developmental Systems and Languages*. Upper Saddle River, NJ: Pearson Education, Inc. 2011.

32. Bellare, M., Goldreich, O., and Goldwasser, S. Incremental Cryptography and Application to Virus Protection. *Proceedings of the 27th ACM Symposium on the Theory of Computing*. May 29-June 1, 1995. New York, New York: Association for Computing Machinery. 1995. 45-56.

33. Takahara, A and Yokomori, T. On the Computational Power of Insertion-Deletion Systems. *Natural Computing*. 2003. 2: 321-336.

34. Wang, Y. *The Goldbach Conjecture*. Singapore: World Scientific. 2002.

35. Gan, Y. S., Fong, W. H., Sarmin, N. H., and Turaev, S. Geometrical Representation of Automata over Some Abelian Groups. *Malaysian Journal of Fundamental and Applied Sciences*. 2012. 8(1): 24-30.

36. Gan, Y. S., Fong, W. H., Sarmin, N. H., and Turaev, S. Permutation Groups in Automata Diagrams. *Malaysian Journal of Fundamental and Applied Sciences*. 2013. 9(1): 35-40.

37. Fong, W. H., Gan, Y. S., Sarmin, N. H., and Turaev, S. Automata Representation for Abelian Groups. *AIP Conference Proceedings*. 2013. 1522(1): 875-882.

38. Fong, W. H., Gan, Y. S., Sarmin, N. H., and Turaev, S. Automata for Subgroups. *AIP Conference Proceedings*. 2014. 1602(1): 632-639.

39. Conrad, K. *Generating Sets*. Expository paper; University of Connecticut.