MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Biomolecular aspects of second order limit language

Muhammad Azrin Ahmad [a,*], Nor Haniza Sarmin [b], Mohd Firdaus Abdul-Wahab [c], Fong Wan Heng [b], Yuhani Yusof [a]

[a] *Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, 26300 Lebuhraya Tun Razak, UMP Gambang, Pahang, Malaysia*
[b] *Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*
[c] *Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*

* *Corresponding author: azrin@ump.edu.my*

Article history
Submitted 6 August 2017
Revised 4 October 2018
Accepted 2 January 2018
Published Online 8 March 2018

Abstract

The study on the recombinant behavior of double-stranded DNA molecules has led to the mathematical modelling of DNA splicing system. The interdisciplinary study is founded from the knowledge of informational macromolecules and formal language theory. A splicing language is resulted from a splicing system, in which different types of splicing langauges have been researched previous, namely adult/inert, transcient and first order limit language. Recently, second order limit language has been extensively explored from the first order limit language. Therefore, in this paper, a laboratory experiment was conducted to validate the existence of a second order limit language. To accomplish it, an initial strand of double-stranded DNA, amplified from bacteriophage lambda, was generated through polymerase chain reaction to generate thousands of copies of double-stranded DNA molecules. A restriction enzyme and ligase were added to the solution to complete the reaction. The reaction mixture was then subjected to polyacrylamide gel electrophoresis to separate biological macromolecules according to their sizes. A mathematical model derived at the early study was used to predict the approximate length of each string in the splicing language. The results obtained from the experiment are then used to verify the mathematical model of a second order limit language. This study shows that the theory on the second order limit language is biologically proven hence the model has been validated.

Keywords: DNA, mathematical modelling, splicing system, splicing language, limit language

INTRODUCTION

Each living organism is differentiated based on the unique molecule that is deoxyribonucleic acid (DNA). Two important roles of DNA are to generate code for the production of protein and self-replication that allows it to transfer information from parent cells to offspring cells [1]. DNA molecules are composed of nucleotides. A complete nucleotide consists of a phosphate group, a sugar group and a nitrogenous base. There are four types of nitrogenous bases which are adenine (A), guanine (G), cytosine (C) and thymine (T), which later can be grouped as purine (A and G) and pyrimidine (C and T). Watson-Crick complementarity [2] stated that the only possible pairings that could exist between the nucleotides are A with T, C with G and vice versa. Phosphodiester bonds link one nucleotide to another to form a single-stranded DNA. The pairs among the bases then form hydrogen bonds. By combining those two bonds, a double-stranded DNA (dsDNA) is formed [3].

Restriction enzyme is a type of enzyme that cuts DNA at a particular nucleotide sequence [4]. The DNA molecule that is cut by the enzyme can have blunt or staggered ends ($5'$-overhang or $3'$-overhang). Those fragments can be joined together by the presence of an enzyme called ligase. It will then produce the same or new hybrid DNA molecules.

Head, who is a pioneer in the mathematical modelling of a splicing system has created an interdisciplinary study that links both the informational macromolecules and formal language theory [5]. In the model, $S=(A,I,B,C)$ represents a set of finite alphabets A, a set of initial strings I, and a set of finite sets, namely pattern B or C, which is a triple of c, x, d in A^*, where A^* denotes the set of all strings over an alphabet A, which is obtained by concatenating zero or more symbols from A [6]. The complementary bases, [A/T], [C/G], [G/C] and [T/A] can be written as a set of alphabets a, c, g and t respectively. The DNA molecule that has been cut by the restriction enzyme may produce $5'$-overhang or blunt end which is assigned to pattern B and the molecule with $3'$-overhang is assigned to pattern C.

Yusof-Goode (Y-G) splicing system, which is a revised model of Head and Goode-Pixton splicing system was introduced in 2011 [7, 8]. The research on the splicing system has covered two issues: a model based on the generation of language, and a model to preserve the biological characteristics of the splicing process. The Y-G model is used in this research since it is proven to present the transparent behavior of the DNA splicing process.

Different types of splicing language were proven to exist when some experiments were carried out to verify the existence of various types of splicing language. Basically, the experiment is conducted either in one stage or two stages. The experiment that is conducted in one stage involves one or more restriction enzymes at a time. Meanwhile, the experiment in two stages involves only one restriction enzyme at a time in the reaction. Then, another enzyme is added to continue the reaction.

Laun and Reddy [9] conducted a one stage experiment by using *Bgl*I and *Dra*III to investigate the accuracy of the model which predicts the behavior of Head splicing system. Fong in 2008 [10] had conducted a two stages experiment in order to verify the mathematical model of splicing system and to show the difference between adult and limit

language by using two restriction enzymes namely, *Aci*I and *Hpa*II. In 2011, Yusof [11] conducted a one stage experiment that involved *Aci*I and *Acl*I to verify the existence of inert persistent, transient and active languages and also to validate the non-uniqueness of limit language. Karimi in 2013 used *CviQ*I and *Acc*65I in a two stages experiment to validate the behavior of persistent splicing system. Other than that, Colosimo [12] also used the concepts of informational macromolecules and formal language theory but with a different purpose, that is to explore the non-random structures in gene sequences. Biologically, a limit language is the remaining molecules after the splicing system has reached its equilibrium state or is completed [13]. Recently, the study of limit language has been extended to the second order limit language [14].

In this paper, a mathematical model of a second order limit language is proposed for the Y-G splicing system involving a restriction enzyme, namely *Dpn*II at one stage and the expected results are presented. In addition, the procedures of conducting the experiment are discussed. The results obtained from both means are interpreted and discussed.

This paper is organised as follows: the first section is the introduction, followed by the second section in which fundamental definitions are presented. The third section discusses the mathematical modelling of the second order limit language. In the fourth section, the procedures of conducting the experiment are discussed. Finally, a conclusion on the findings is presented in the last section.

PRELIMINARIES

In this section, some fundamental definitions of this research and some types of splicing language used in this paper are given.

The first three basic definitions relate to formal language theory, namely alphabet, string and language.

Definition 1 [6] An alphabet, *A*, is a finite, nonempty set of symbols.

Definition 2 [6] A string is a finite sequence of symbols from the alphabet.

Definition 3 [6] A set of strings all of which are chosen from some A^*, where *A* is a particular alphabet, is called a language.

In the molecular biology, complementary bases are represented as a set of alphabets and initial strands of dsDNA molecules are represented as a set of initial string. Besides, restriction enzymes are represented as a set of rules. Meanwhile, dsDNA molecules obtained from the splicing process are represented by language.

The concatenation between two languages, L_1 and L_2 has been given in [2] where $L_1L_2 = \{xy \mid x \in L_1, y \in L_2\}$.

Furthermore:

$L^0 = \{\lambda\}$,

$L^{i+1} = LL^i,\ i \geq 0$,

$L^* = \bigcup_{i=0}^{\infty} L^i$ (the * - Kleene closure),

$L^+ = \bigcup_{i=1}^{\infty} L^i$ (the + - Kleene closure).

The definition of a Y-G splicing system is given in the following.

Definition 4 [7] A splicing system $S = (A, I, R)$ consists of a set of alphabets *A*, a set of initial strings *I* in A^* and a set of rules, $r \in R$ where *r* = (*u*, *x*, *v*: *y*, *x*, *z*). For $s_1 = \alpha uxv\beta$ and $s_2 = \gamma yxz\delta$ elements of *I*, splicing s_1 and s_2 using *r* produces the initial string *I* together with $\alpha uxz\delta$ and $\gamma yxv\beta$, presented in either order where $\alpha, \beta, \gamma, \delta, u, x, v, y$ and $z \in A^*$ are the free monoids generated by *A* with the concatenation operation and 1 as the identity element.

Two types of splicing languages are discussed in this paper, namely transient and limit languages [14]. Experimentally, a splicing language is called transient if a set of strings is eventually used up and disappear in a given system. On the other hand, a splicing language is a limit language or known as the first order limit language given that it is the set of words that is predicted to appear if some amount of each initial molecule is present, and sufficient time has passed for the reaction to reach its equilibrium state, regardless of the balance of the reactants in a particular experimental run of the reaction.

In the following, the definition of the second order limit language is given.

Definition 5 [14] Let $L(S)$ be a splicing language of a splicing system, *S* and $L_1(S)$ is the first order limit language. A splicing language is called a second order limit language, $L_2(S)$ if the set of strings produced in $L_2(S)$ is distinct from the set of strings of $L(S)$ in which $L_2(S) \cap L(S) = \varnothing$ and $L_1(S) \not\subset L_2(S)$.

In the next section, a mathematical model of the second order limit language is developed.

MATHEMATICAL MODELLING OF SECOND ORDER LIMIT LANGUAGE

In this section, a mathematical model of the second order limit language that has two cutting sites is developed. The splicing language is generated based on the given rule. Next, the second order limit language is also generated in the general form.

Let $S = (A, I, R)$ be a Y-G splicing system consisting of a set of alphabets, $A = \{a, c, g, t\}$, a set of initial strings, $I = \{\alpha gatc\beta gatc\gamma\}$, and a set of rules, $R = \{r\}$ such that $r = (1; gatc, 1:1; gatc, 1)$ where $\alpha, \beta, \gamma \in A^*$. Based on the molecular perspective, a solution contains multiple copies of dsDNA where its 180° degree rotation is also considered which indicates the presence of string α', β', γ' also in A^*. Other than that, the left and right context of the set of rules are represented as 1 which explain their null form. This shows the cutting process occur within the restriction site, *gatc* which starts at *g* and ends after the alphabet *c* on the DNA molecule.

The following splicing language is generated from the splicing system:

$$\{\alpha gatc\beta gatc\gamma\} \xrightarrow{R} I \cup \left\{ \begin{array}{l} \alpha gatc\gamma, \alpha gatc\alpha', \gamma' gatc\gamma, \\ \alpha gatc\beta gatc\alpha', \gamma' gatc\beta gatc\gamma, \\ \alpha gatc\beta' gatc\gamma, \alpha gatc\beta gatc\beta' gatc\gamma, \\ \alpha gatc\beta gatc\beta' gatc\alpha', \\ \alpha gatc\beta' gatc\beta gatc\gamma, \\ \alpha gatc\beta' gatc\beta gatc\alpha', \\ \gamma' gatc\beta gatc\beta' gatc\gamma, \\ \gamma' gatc\beta' gatc\beta gatc\gamma \end{array} \right\}.$$

Based on the definition of the second order limit language, $L_2(S) \cap L(S) = \varnothing$ since $L_1(S) \subseteq L(S)$ thus $L_1(S) \not\subset L_2(S)$. Hence, by further splicing the set of strings of $L(S)$, the second order limit language is presented in the following general form:

$$L_2(S) \xrightarrow{R} I \cup \left\{ \begin{array}{l} \alpha (gatc\beta \cup gatc\beta')^* gatc\alpha', \\ \gamma' (gatc\beta' \cup gatc\beta)^* gatc\gamma, \\ \alpha (gatc\beta' \cup gatc\beta)^* gatc\gamma \end{array} \right\}.$$

In the next section, a restriction enzyme called *diplococcus pneumonia*, *Dpn*II is represented as a set of rules and a set of initial dsDNA molecules is chosen from the amplified bacteriophage lambda that contains exactly two restriction sites, *gatc* of *Dpn*II is represented as a set of initial strings. In additions, the procedures of conducting the experiment are discussed in detail.

THE WET MODEL

In this experiment, an initial strand of dsDNA was chosen from lambda phage DNA (New England Biolabs, USA).

Fragment 1 of this lambda phage DNA is the fragment of interest for the initial string *I* since it contains two times the site for the restriction enzyme *Dpn*II. The length of the fragment is given as follows.

Fragment 1: $A - Dpn\text{II site} - B - Dpn\text{II site} - D$

$$|A| = 68\text{bp}$$
$$|Dpn\text{II}| = 4\text{bp}$$
$$|B| = 38\text{bp}$$
$$|Dpn\text{II}| = 4\text{bp}$$
$$|D| = 40\text{bp}.$$

The genome location for the strand $(A - Dpn\text{II site} - B - Dpn\text{II site} - D)$ is between 5396 and 5549 which gives 154 base pairs (bp) long. This strand has exactly two cutting sites of restriction enzyme *Dpn*II. Therefore, the strand is denoted as $A - B - D$.

Polymerase chain reaction (PCR) can generate thousands of copies of DNA molecules.The strand was generated by PCR using MyCycler™ Thermal Cycler (Bio-Rad). OneTaq® Hot Start 2X Master Mix (New England Biolabs, USA) with standard buffer was used for $A - B - D$ strand. The PCR recipe of $A - B - D$ strand is 4 µl of lambda DNA (10 ng/µl), 1 µl of forward and reverse primer with concentration 10 µM, 25 µl of OneTaq® Hot Start 2X Master Mix and nuclease-free water to obtain 50 µl total volume of the reaction mixture.

The forward and reverse primers were designed using Primer3 Input [15] at Primer3 website version 4.0.0. (http://bioinfo.ut.ee/primer3/). Full Enterobacteria lambda phage DNA sequence (NCBI reference sequence NC_001416.1) was used as template to design the forward and reverse primers. The primer sites were chosen such that they flank the desired restriction site, assisted by the software Primer3. The forward primer site correspond to nucleotide no. 5396-5414 of the lambda phage sequence, and the reverse primer site correspond to nucleotide no. 5532-5549. The two restriction sites chosen are located at nucleotide no. 5464-5467 and 5506-5509. The melting point of each primers and the annealing temperature suitable for both primers were obtained from T_m Calculator version 1.7.2, New England Biolabs website (http://tmcalculator.neb.com).

The A – B – D strand was amplified with the forward and reverse primers as shown in Table 1. Prior to 35 cycles of the program, the DNA was denatured for 30 seconds at 94ºC. The temperature program was as follows: 30 seconds at 94ºC, 40 seconds at 58ºC, 30 seconds at 72ºC. The final elongation step was at 72ºC for 5 minutes before being stored at 4ºC.

Restriction enzyme digestion and ligation with time-sequence sampling is the final step before running the aliquoted reaction mixture on 12% polyacrylamide gel electrophoresis (PAGE). The recipe of the time sequence reaction of restriction enzymes digestion and ligations in the five wells (Lane 3 to Lane 7) of agarose gel is 60 µL of purified lambda DNA, 10 µL of T4 DNA ligase buffer, 4 µL of concentrated T4 DNA ligase, 5 µL of *Dpn*II and nuclease-free water to obtain 100 µL total volume of the reaction mixture.

The time interval for time-sequence sampling was conducted, where aliquots 20 µL each was taken out from the micro centrifuge tube incubated at 37 ºC. Each aliquots for reaction tube at 0, 15, 30 and 60 minutes were stored immediately at -20 ºC to stop the reaction.

RESULTS AND DISCUSSION

In this experiment, a few assumptions were made such that the initial strand *I* and the restriction enzyme were of the same amount. In addition, the probabilities of digestion and ligation efficiency are assumed to be equal. By following all procedures, the solution was then subjected to 12% polyacrylamide gel electrophoresis (PAGE) for 90 minutes. On the other hand, Gangaraj *et al*. [16] used molecular genetic in identifying milk protein genotypes where the method involves polymerase chain reaction and the solution is subjected to the polyacrylamide gel electrophoresis to observe the pattern of kappa-caseine gene in buffaloes. Hence, the predicted gel is presented in Figure 1 below.

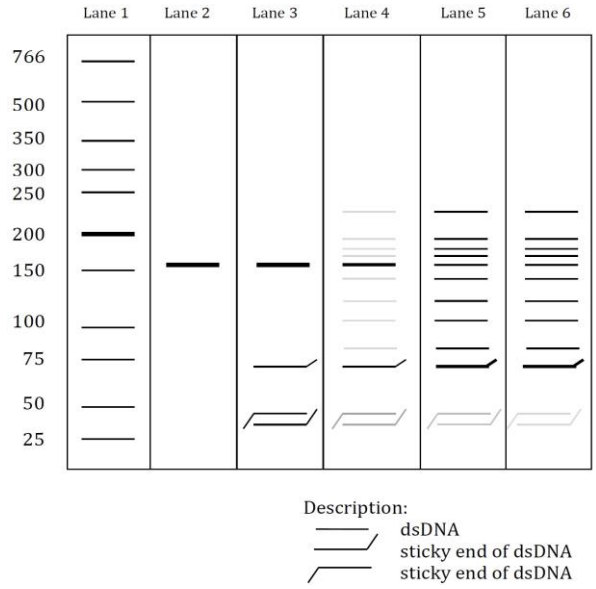

Fig. 1 Predicted gel of *Dpn*II digestion and ligation towards *I*. Lane 1: LMW Ladder, Lane 2: Lambda DNA $\alpha - \beta - \gamma$ (purified), Lane 3: the time sequence reaction mixture at t = 0 minutes, Lane 4: the time sequence reaction mixture at t = 15 minutes, Lane 5: the time sequence reaction mixture at t = 30 minutes and Lane 6: the time sequence reaction mixture at t = 60 minutes.

The first lane shows the existence of eleven bands of LMW DNA ladder that behave as DNA Marker represented by single lines according to their molecular weights 25, 50, 75, 100, 150, 200, 250, 300, 350, 500 and 766bp respectively. In Lane 2, one line exists between 150 and 200bp due to the existence of initial strand of dsDNA, *I* which is the value of $A - B - D$. Note that, $A - B - D$ is equal to 154bp. In Lane 3, the initial strand *I* and the sticky ends of *A*, *B* and *D* exist. In Lane 4 to Lane 6, the lines that appear are summarized in the following table.

Table 1 The size (bp) of predicted molecules.

No.	Molecule	Size (bp)
1.	A	68
2.	B	42
3.	D	44
4.	$A - B - D$	154
5.	$A - D$	112
6.	$A - A'$	140
7.	$D' - D$	84
8.	$A - B - A'$	182
9.	$D' - B - D$	126
10.	$A - B' - D$	154
11.	$A - B - B' - D$	196
12.	$A - B - B' - A'$	224
13.	$A - B' - B - D$	196
14.	$A - B' - B - A'$	224
15.	$D' - B - B' - D$	168
16.	$D' - B' - B - D$	168

In Figure 2, extra bands appear in Lane 4 – Lane 6 as compared to Figure 1. In consequence, those bands represent the second order limit language in Y-G splicing system.

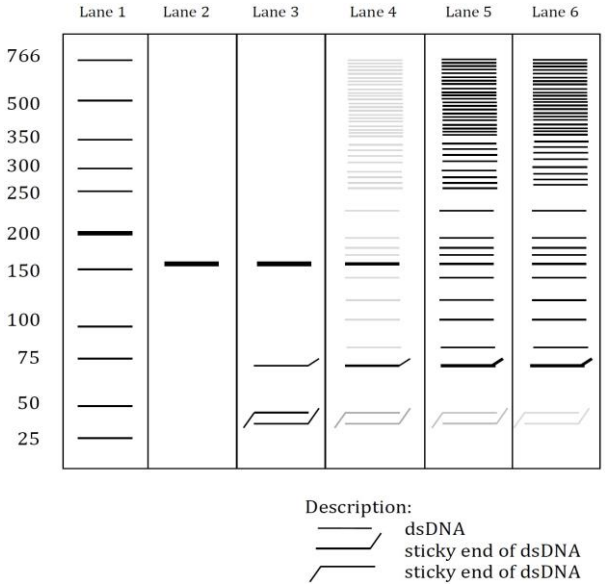

Fig. 2 Predicted gel of the second order limit language.

PAGE gel was stained with EtBr and visualized using a UV transilluminator. Figure 3 shows the second order limit language observed experimentally.

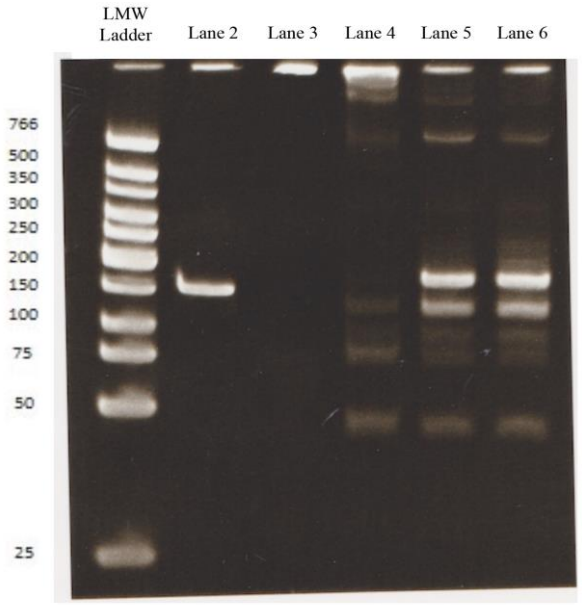

Fig. 3 Gel photo with the splicing pattern of enzyme *Dpn*II. Lane 1: LMW Ladder, Lane 2: Lambda DNA $\alpha - \beta - \gamma$ (purified), Lane 3: the time sequence reaction mixture at t = 0 minutes, Lane 4: the time sequence reaction mixture at t = 15 minutes, Lane 5: the time sequence reaction mixture at t = 30 minutes and Lane 6: the time sequence reaction mixture at t = 60 minutes.

In Lane 2 of Figure 3, a single band appeared between 150 and 200 bp, indicating the initial strand of dsDNA that is 154 bp. Lane 3, there are no distinct bands appearing probably due to the molecules are at their intermediate states, due to the dynamics of *Dpn*II digestion. In Lane 4, several bands appeared ranging from 42 bp to 766 bp (and larger) indicating the successful ligation of 42, 44, 68 bp products (and other larger products) by T4 DNA ligase. The appearances of some bands in Figure 2 in Line 5 and Line 6 in addition to those in Figure 1, which are on the same lanes, show the existence of a second order limit language in the Y-G splicing system. Across the lane, the bands are similar as compared to Lane 4 except that the bands below and above 150 bp became more intense. This is due to the overlapping of other bands where the difference from one band to another is close. The 42 bp, 44 bp and 68 bp molecules with sticky ends are not considered to be in the splicing language since they are not well-formed dsDNA molecules. Therefore, the complete DNA strands produced in this process are the same as we anticipated in Figure 2.

CONCLUSION

In conclusion, it was shown that the second order limit language is proven to exist through an experiment. The action of cutting and pasting was predicted to result in a splicing system which would converge to a particular set of the second order limit language with the presence of enzyme *Dpn*II. It has been correctly predicted from the splicing system in its wet-lab procedure that as time increases, the visibility of the initial strings decreases, while the second order limit language increases. Figure 1 shows the prediction of the bands which represent dsDNA molecules that are produced when the digestion and ligation of *Dpn*II took place. Figure 2 then shows the extra bands other than the bands in Figure 1 that indicate the existence of the second order limit language. It can be concluded that the mathematical model of the second order limit language has been verified experimentally since the dsDNA molecules produced in the experiment are as predicted in the model.

ACKNOWLEDGEMENT

The first author would like to thank Universiti Malaysia Pahang for the financial funding through Vote No. RDU1703278. The second and fourth authors would like to thank Ministry of Higher Education (MOHE) and Research Management Centre (RMC), Universiti Teknologi Malaysia for the financial funding through UTM Research University Grant Vote No. 13H18.

REFERENCES

[1] I. E. Alcamo, *DNA Technology: The Awesome Skill* (*2nd ed.*), USA: Academic Press, 2001.
[2] G. Paun, G. Rozenberg, A. Salomaa, *DNA computing: New computing paradigms*, New York: Springer-Verlag Berlin Heidelberg, 1998.
[3] S. S. Purohit, *Biotechnology: Fundamentals and applications*, Jodhpur, India: Agrobios, 2005.
[4] M. Wink, *An introduction to molecular biotechnology*, Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2006.
[5] T. Head, Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviors. *Bull. Math. Biol.*, vol. 49, pp. 737–759, 1987.
[6] P. Linz, *An introduction to formal languages and automata* (*5th ed.*), USA: Jones and Bartlett, LLC, 2012.
[7] Y. Yusof, N. H. Sarmin, T. E. Goode, M. Mahmud, W. H. Fong, in: Abdullah R., Khader A. T., Venkat I., Wong L. P., Subramaniam K. G. (eds.), An extension of DNA splicing system, IEEE Conference Proceedings: *6th International Conference on Bio-Inspired Computing: Theories and Application (BIC-TA 2011)*, September 27-29, 2011, California, USA: IEEE Computer Society, 2011, p. 246-248.
[8] Y. Yusof, N. H. Sarmin, T. E. Goode, M. Mahmud, W. H. Fong, Hierarchy of certain types of DNA splicing systems. *Int. J. Mod. Phys. Conf. Ser.*, vol. 9, pp. 271-277, 2012.
[9] E. Laun, K. J. Reddy, in: Rubin H., Wood D. H. (eds.), Wet splicing system, DIMAC Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Rhode Island, USA, 1999, p. 73-83.
[10] W. H. Fong, N. H. Sarmin, Mathematical modelling of splicing systems, *Proceedings of the 1st International Conference on Natural Resources Engineering & Technology*, July 24-26, 2006, Putrajaya, Malaysia, 2006, p. 524-527.
[11] Y. Yusof, W. L. Lim, T. E. Goode, N. H. Sarmin, W. H. Fong, M. F. A. Wahab, in: Ramli M. F., Junoh A. K., Roslan N., Masnan M. J., Kharuddin M. H. (eds), Molecular aspects of DNA splicing system, *International Conference on Mathematics, Engineering and Industrial Applications 2014 (ICoMEIA 2014)*, AIP Publishing, vol. 1660, 2015, p. 050045.

[12] A. Colosimo, A. D. Luca, Special factors in biological strings. *J. Theor. Biol.*, vol. 204, pp. 29–46, 2000.
[13] E. Goode, D. Pixton, in: Janoska N, Paun Gh., Rozenberg G. (eds), *Splicing to the limits aspects of molecular computing*, New York: Springer-Verlag Berlin Heidelberg, 2004, p. 189.
[14] M. A. Ahmad, N. H. Sarmin, W. H. Fong, Y. Yusof, in: Zain WZW, Dzul-Kifli S. C., Razak F. A., Ishak A. (eds), An extension of first order limit language, *Proceedings of the 3rd International Conference on Mathematical Sciences*, AIP Conf. Proc., vol. 1602, 2014, p. 627-631.
[15] S. Rozen, H. Skaletsky, Primer3 on the WWW for General Users and for Biologist Programmers. In: Misener S, Krawetz SA (eds) *Bioinformatics Methods and Protocols*, Totowa, New Jersey: Humana Press, 1999, p. 365-386.
[16] D. R. Gangaraj, S. Shetty, M. Govindaiah, C. Nagaraja, S. Byregowda, M. Jayashankar, Molecular characterization of kappa-casein gene in buffaloes, *ScienceAsia*, vol. 34, pp. 435–439, 2008.