Proceedings of the 1st International Conference on Natural Resources Engineering & Technology 2006
24-25th July 2006; Putrajaya, Malaysia, 524-527

Mathematical Modelling of Splicing Systems

Nor Haniza Sarmin* and Fong Wan Heng

Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.

Abstract

Every living organism has DNA that makes the organism unique. Since a DNA strand can be viewed as a string over a four letter alphabet (*a*, *c*, *g*, and *t*) which is the four deoxyribonucleotides, thus the modelling can be done within the framework of formal language theory. When restriction enzymes and ligase are added to initial strings of DNA molecules, additional strings of molecules can result. Those resulting molecules are adjoined into a language called a splicing language, which can then be analyzed using concepts in formal language theory. This process is modelled as a splicing system. The development of mathematical representation of the generative capacity of a splicing system was initiated by Tom Head in 1987. This research initiates the connection between formal language theory and the study of informational macromolecules.

Keywords: Mathematical modelling, splicing systems, splicing language, DNA, formal language theory.

1.0 Introduction

This research initiates the connection between formal language theory and the study of informational macromolecules. Previously, these two branches of studies are independent of each other. However, when splicing system is introduced, the generative capacity of system of enzymes acting on a set of DNA molecules is established formally using formal language theory. Different languages can result from this recombinant behaviour and are analyzed using some concepts of languages in formal language theory.

There are more than 200 types of readily available restriction enzymes as listed in the New England Biolabs catalogue. These restriction enzymes can cut strings of DNA molecules at specific places, resulting in molecules with sticky ends. New molecules then arise when molecules previously cut by restriction enzymes are pasted together by a ligase. Splicing system was defined to model the recombinant action of restriction enzyme and a ligase on DNA molecules. The language which results from a splicing system is called a splicing language. This language contains the initial strings of DNA molecules and is closed under the application of splicing rules. This splicing language is further studied using formal language theory, which is a branch of applied discrete mathematics and theoretical computer science. It concerns with sets of strings called languages and different mechanism for generating and recognizing them.

* *Corresponding author: Tel: 07-5534266 Fax: 07-5566162 E-mail: nhs@mel.fs.utm.my, fwheng@yahoo.com*

The potential effect of sets of restriction enzymes and a ligase that allow DNA molecules to be cleaved and reassociated to produce further molecules can be found in [1]. The associated languages are analyzed by means of a new generative formalism called a splicing system. A new relationship between formal language theory and the study of informational macromolecules was thus initiated. Formal language theory is a branch of applied discrete mathematics and theoretical computer science that is devoted to the study of sets of finite strings (called languages) of symbols chosen from a prescribed finite set (called an alphabet). The set of double-stranded DNA molecules that may arise from an initial set of DNA molecules in the presence of specified enzyme activities is represented as a language over the four-symbol alphabet of deoxyribonucleotide pairs.

1.1 Splicing system

A splicing system $S = (A, I, B, C)$ consists of a finite alphabet A, a finite set I of initial strings in A^*, and finite sets B and C of triples (c, x, d) with c, x and d in A^*. Each such triple in B or C is called a pattern. For each such triple the string cxd is called a site and the string x is called a crossing. Patterns in B are called left patterns and patterns in C are called right patterns. The language $L = L(S)$ generated by S consists of the strings in I and all strings that can be obtained by adjoining to L $ucxfq$ and $pexdv$ whenever $ucxdv$ and $pexfq$ are in L and (c, x, d) and (e, x, f) are patterns of the same hand. A language L is a splicing language if there exists a splicing system S for which $L = L(S)$.

A splicing language consists of all the strings, that is, dsDNA molecules without sticky ends, that ever occurred in the generation process. Molecules with remaining sticky ends are not considered to be in the splicing language. Only fully double stranded DNA molecules are considered to form the splicing language. A splicing language can consist of two types of languages which are adult language and limit language.

Adult language consists of the strings that are produced by the system but do not participate in further operations. Strings lying in this language cannot be cut any further. Limit language consists of those strings that will be present after a splicing system has reached equilibrium, without considering whether or not they will participate in further operations. Strings in this language may still continue to be cut and re-ligated.

Adult language and limit language are both considered to be in the splicing language, since they do not consist of molecules with sticky ends. Any string that is in the adult language is automatically in the limit language, but not vice versa. Thus, the adult language is always a subset of the limit language.

In 1999, the viability of Head's "dry model" has been experimented in a wet-lab procedure by Laun and Reddy in [2]. A simple example of a wet-lab procedure is shown to generate in vitro the splicing language predicted by the corresponding "dry model". This example shows the production of an adult language which is identical to the limit language. Later, in 2004, the result of a wet-lab procedure after the reaction has run to its completion is considered by Goode and Pixton in [3].

The objective of this paper is to initiate a wet-lab procedure which will illustrate a simple possible case in which the adult language and limit language are distinct.

2.0 Materials and Methods

2.1 Wet-lab Procedure

In this section, we give an example to illustrate the concepts in splicing system and some related concepts mentioned in section 1.

A simple example of a splicing system is $S = (A, I, B, C)$ where A is the set $\{a, g, c, t\}$, I is the set of initial string $\{\alpha ccgc\beta\}$, with α and β in $\{a, c, g, t\}^* = A^*$. This initial string is the sequence which appear in the genome of the bacteriophage lambda. The string is assumed to be dephosphorylated on its 5′ ends in order to prevent blunt end ligation. The substring *ccgc* appearing in the initial string is the recognition site for the restriction enzyme AciI. The pattern (*c*, *cg*, *c*), which is the element in set *B*, encode the action of the enzyme AciI in the presence of a DNA ligase. In this example, set *C* is empty. The restriction site (*c*, *cg*, *c*), when cut, leave 5′ CG single-stranded overhangs which allow religation. There are no other recognition sites for this enzyme in the initial string, so the splicing language $L(S)$ for this example is the set $I \cup \{\alpha ccgg\alpha, \beta gcgc\beta\}$.

We will verify experimentally that this language is in fact generated in vitro as predicted by the splicing model, and that these two strings {αccggα, βgcgcβ} are both in adult language and limit language. The string αccggα can further be cut by the restriction enzyme HpaII. However after cutting, this string can only combine with itself, so this string remains in the limit language. As for the string βgcgcβ, it cannot be cut again by any other enzyme present. Thus this string lies in the adult language, which is automatically in the limit language.

This experiment is to produce a laboratory verification of Head's "dry model" of the actual wet-lab procedure. This experiment will test the hypothesis that a particular splicing system will converge to a fixed set of strings. The initial set is the sequence of linear dsDNA with dephosphorylated 5′ ends, and having the restriction site of the enzyme AciI. The action of iterated cleavage and religation was predicted to result in a dynamical splicing system which would converge to a particular set of adult strings with the presence of enzyme AciI only; but which would converge to a particular set of adult and limit strings with the presence of both enzymes AciI and HpaII. Figure 1 below shows the initial and adult strings which are the expected products of the wet-lab procedure. The 150bp string will lie in the limit language with the presence of the restriction enzyme HpaII. This experiment is designed to see whether the molecular splicing system behaved as predicted by Head's "dry model".

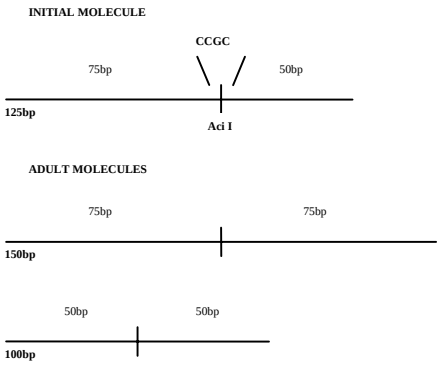

Figure 1 The initial and adult molecules which are the expected products of the wet-lab procedure

It is predicted that an analysis of the dynamical behaviour of this splicing system in its wet-lab procedure would show a time-related decrease in the initial strings, and a corresponding increase in the adult and limit strings. Intermediate fragments having 5′ CG overhangs are expected also.

3.0 Conclusion

This paper initiates the connection between formal language theory and the study of informational macromolecules. Splicing system is defined to model the recombinant action of restriction enzyme and a ligase on DNA molecules. The language which results from a splicing system is called a splicing language. The concept of adult language and limit language is discussed in this paper. A dry mathematical model is used to predict the actual behaviour of the corresponding wet-lab procedure, where the adult language and limit language are distinct.

Acknowledgements

We would like to express our gratitude to Prof Tom Head[1], Assoc. Prof. Dr. Noor Aini Abdul Rashid and Mohd Firdaus Abdul Wahab for their collaborations in this research. We would also like to thank the Ministry of Science, Technology and Environment Malaysia for the financial funding through IRPA Vot 74259.

References

[1] Head, Tom. 1987. Formal Language Theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *Bulletin of Mathematical Biology*. 49: 737-759.
[2] Laun, Elizabeth and K. J. Reddy. 1999. Wet Splicing Systems. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. 48: 73-83.
[3] Goode, Elizabeth and D. Pixton. 2004. *Splicing to the limit. Aspects of Molecular Computing - Essays Dedicated to Tom Head on the Occasion of His 70th Birthday* (N. Jonoska, G. Paun, G. Rozenberg eds.). Springer-Verlag. Lecture Notes in Computer Science, v. 2950.