Australian Journal of Basic and Applied Sciences, 5(1): 20-24, 2011
ISSN 1991-8178

Some Sufficient Conditions for Persistent Splicing Systems

Fariba Karimi, Nor Haniza Sarmin, Fong Wan Heng

Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia,
81310 UTM Skudai, Johor, Malaysia
Ibnu Sina Institute for Fundamental Science Studies, Universiti Teknologi Malaysia,
81310 UTM Skudai, Johor, Malaysia

Abstract: Splicing system was first defined by Head in 1987 as a mathematical model of the generative capacity of a biological system containing DNA molecules in the presence of appropriate enzymes. The formalism of splicing system is illustrated under the framework of Formal Language Theory which is a branch of applied discrete mathematics and theoretical computer science. In fact, this is a mathematical model for the recombinant behavior of DNA molecules under the influence of restriction enzymes. In this sense, DNA molecules and restriction enzymes are associated with strings and rules, respectively. There are many different types of splicing systems that have been defined by Head and other mathematicians. Some important ones are persistent, permanent and strictly locally testable splicing systems. In this paper, we provide some sufficient conditions for splicing systems to be persistent. Besides, some real examples are provided to support the theorems in the biological sense.

Key words: Splicing Systems; Persistent Splicing Systems; DNA Molecules; Formal Language Theory.

INTRODUCTION

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. The information in DNA is stored as a code made up of four chemical bases: adenine (*A*), guanine (*G*), cytosine (*C*), and thymine (*T*). The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. DNA bases pair up with each other, *A* with *T* and *C* with *G*, to form units called base pairs.

Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder. So DNA can be represented as strings over four alphabets, i.e. $D = \{[A/T],[C/G],[G/C],[T/A]\}$. A restriction enzyme is an enzyme that cuts double-stranded or single stranded DNA at specific recognition nucleotide sequences, known as restriction sites. The recombination behavior of restriction enzymes was modeled in the form of splicing rules by Head. In this research, we investigate some characteristics of persistent splicing systems.

II. Basic Definitions and Notations:

This section gives the main concepts and notations that are used in this paper. Definiton 2.1. (P. Linz, 2001) (Alphabet, String) A finite, nonempty set *A* of symbols is called *alphabet*. Any finite sequence of symbols from alphabet is called a *string.*We use 1to denote the *empty string* which is a string with no symbols at all. If *A* is an alphabet, we use *A** to denote the set of strings obtained by concatenating zero or more symbols from *A.*

Definition 2.2. (T. Head, 1987) (Splicing System, Splicing Language):

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 = (LS)* generated by *S* consists of the strings

Corresponding Author: Fariba Karimi, Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
E-mail: fk.karimi@gmail.com, nhs@utm.my

in *I* and all strings that can be obtained by adjoining the words *ucxfq* and *pexdv* to *L*, 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).

Definition 2.3. (T. Head, 1987) (Persistent):

Let S = (*A,I,B,C*) be a splicing system. Then *S* is *persistent* if for each pair of strings *ucxdv,* and *pexfq,* in *A** with *(c,x,d)* and (*e,x,f*) patterns of the same hand: If *y* is a subsegment of *ucx* (respectively *xfq)* that is the crossing of a site in *ucxdv* pectively *pexfq)* then this same subsegment *y* of *ucxfq* contains an occurrence of the crossing of a site in *ucxfq.*

III. Main Results:

In this section, we provide some sufficient conditions in which splicing systems are persistent as well as real examples that illustrate the theorems.

Theorem 3.1.

Let $S=(A, I, B, \varnothing)$ be a splicing system such that $B=\{(c_i, x_i, d_i): 1\le i\le n\}$ If elements of *B* do not have the same crossing and c_i (respectively d_i) is not a factor of c_j (respectively d_j), $\forall\ 1\le i,j\le n\ (i\ne j)$ then S is persistent.

Proof.

To show that S is persistent, the patterns with the same crossings should be considered. Since the crossings of S are disjoint, each pattern can be considered only with itself. So for every pair of strings $uc_ix_id_iv$ and $pc_ix_id_iq\in A^*$ such that $(c_i, x_i, d_i)\in$ B:

If *Y* is a subsegment of uc_ix_i (respectively)x_id_iq that is the crossing of a site in $uc_ix_id_iv$ (respectively $pc_ix_id_iq$) then according to *B* there exists $1\le sk\le n$ much that Y = X_K and $c_kx_kd_K$ is a factor of $uc_ix_id_iv$ (respectively $pc_ix_id_iq$).

First we prove that $c_kx_kd_K$ is a factor of $u_i\ c_i\ x_i\ d_i\ v_i$ (respectively $c_i\ x_i\ d_i\ q_i$):
In the case, $c_k\ x_k\ d_K$ is not a factor of $u_i\ c_i\ x_i\ d_i$ (respectively $c_i\ x_i\ d_i\ q_i$): since, (respectively $x_k\ d_k$) is a factor of $u_i\ c_i\ x_i$ (respectively $x_i\ d_i\ q_i$) and simultaneously $c_k\ x_k\ d_K$ is also a factor of $uc_ix_id_iv$ (respectively $pc_ix_id_iq$)

Thus, d_k should begin before or from d_i and continue after d_i until reach v and therefore d_k will be a factor of d_i (respectively, c_K should begin before c_i and end after it and therefore c_K will be a factor of c_i). But this contradicts the hypothesis of the theorem, so this case does not happen. Thus, $c_k\ x_k\ d_K$ is also a factor of $U_i\ c_i\ d_i\ x_i\ q_i$ and it means that x_K contains an occurrence of the crossing of a site in $U_i\ c_i\ d_i\ x_i\ q_i$. Thus *S* satisfies the definition of persistent.

Example 3.2.

Let *S* be the splicing system associated with the restriction enzymes {*Age*I, *Mse*I} that is,

$S=(D, I, B, \varnothing)$ such that $D=\{[A/T],[C/G],[G/C],[T/A]\}$ and *I* is an arbitrary subset of *A** and

$B=\{([A/T],[C/G][C/G][G/C][G/C],[T/A]),\quad ([T/A],[T/A][A/T],[A/T])\}$.

Since the crossings of Sare disjoint and $[A/T]$ and $[T/A]$ are not factors of each other, according to Theorem 3.1, S is persistent.

Theorem 3.3:

Let $S=(A, I, B, \varnothing)$ be a splicing system such that

$B=\{(c_i, x_i, d_i): 1\le i\le n\} \cup \{(1, x_i, 1): 1\le i\le n\}$. Then S is persistent.

Proof.

To show that S is persistent, according to the definition, the patterns with the same crossings should be considered. So suppose that uc_ixd_iv and pc_jxd_jq be two arbitrary strings from *A** such that (c_i, x, d_i) and $(c_j, x, d_j)\in B$have the same crossing *x* and by splicing them the string uc_ixd_jq will be obtained. Now, if *Y* is a subsegment of uc_ix (respectively xd_jq) that is the crossing of a site in uc_ixd_iv (respectively pc_jxd_jq) then according to *B* there exists $1\le k\le n$ such that $y=x_k$ and $(c_k, x_k, d_k), (1, x_k, 1)\in B.$ Now X_k is a factor of uc_ixd_jq and it is the crossing of the site $(1, x_k, 1)$

So S is persistent.

Example 3.4.

Given a set consisting of four restriction enzymes: *Bam*HI, *Bgl*II, *Bcl*I and *Dpn*II. Since there are four enzymes in the set, there are 2^4 = 16 subsets for this set of four enzymes. Let *B* be any subset of these sets that contain *Dpn*II. Every splicing system with such set *B* as its right patterns is persistent.Indeed, the cleavage patterns of the enzymes *Bam*HI, *Bgl*II, *Bcl*I and *Dpn*II are

$([A/T],[G/C][A/T][T/A][C/G],[T/A])$,

$([G/C],[G/C][A/T][T/A][C/G],[C/G])$,

$([T/A],[G/C][A/T][T/A][C/G],[A/T])$ and $(1,[G/C][A/T][T/A][C/G],1)$ respectively.

So for every such splicing system the conditions of the previous theorem are satisfied and it is persistent.

Theorem 3.5.

Let $S=(A, I, B, \varnothing)$ be a splicing system such that

$B=\{(c_i, x_1x_2, d_i): 1\le i\le n\} \cup \{(1, x_1x_2, 1), (x_1, 1, x_2)\}$ $c_i, d_i, x_1, x_2\in A^*$ such that x_1 is not a factor of x_2 and vice versa. Then *S* is persistent.

Proof. To show that *S* is persistent, the patterns with the same crossings should be considered. Here two cases can happen.

Case 1:

Suppose that $uc_ix_1x_2d_iv$ and $pc_jx_1x_2d_jq$ be two arbitrary strings from *A** with $(c_i, x_1x_2, d_i)\text{and}(c_j, x_1x_2, d_j)\in B$ with the same crossing x_1 x_2 and with splicing them the string $uc_ix_1x_2d_jq$ will be obtained. Now, if *y* is a subsegment of $uc_ix_1x_2$ (respectively $x_1x_2d_jq$) that is the

crossing of a site in $uc_i x_1 x_2 d_i v$ (respectively $pc_j x_1 x_2 d_j q$) then according to *B*, y = x_1 x_2 or 1. If y = x_1 x_2 , it is the crossing of the site (1,x_1 x_2, 1) in $uc_i x_1 x_2 d_j q$ and so the desired result is satisfied. If y = 1, because *y* is crossing of a site in $uc_i x_1 x_2 d_i v$ it should follow x_1 and precede x_2. Therefore, $uc_i x_1 x_2$ cannot be in the form, $uc_i x_1 x_2 y$ (in other word, although we can consider $uc_i x_1 x_2 = uc_i x_1 x_2 1$but this 1 is not the y that we mean here). From this and the previous hypothesis that y is in $uc_i x_1 x_2$ and simultaneously crossing of a site in $uc_i x_1 x_2 d_i v$, it can be concluded that x_1 y x_2 is a factor of $uc_i x_1 x_2$ and so it is a factor of $uc_i x_1 x_2 d_j q$ Thus, y contains an occurrence of the crossing of a site in $uc_i x_1 x_2 d_j q$ and the desired result is satisfied.

Case 2:

Suppose $ux_1 1 x_2 v$ and $px_1 1 x_2 q$ be two arbitrary strings from A^* with $(x_1, 1, x_2) \in B$ and with splicing them the string $ux_1 1 x_2 q$ will be obtained. Similarly, if *y* is a subsegment of $ux_1 1$ (respectively $1 x_2 v$) that is the crossing of a site in $ux_1 1 x_2 v$ (respectively $px_1 1 x_2 q$) then according to *B*, *y* = x_1 x_2 or 1.

If y = x_1 x_2, it is the crossing of the site (1,x_1 x_2, 1) in $ux_1 1 x_2 q$ and so the desired result is satisfied.

If y = 1, because *y* is crossing of a site in $ux_1 1 x_2 v$ it should follow x_1 and precede x_2. On the other hand, since *y* = 1 should be in $ux_1 1$, x_1 *y* x_2 is a factor of $ux_1 1 x_2$ and it cannot continue until *v*. Indeed, if *y* is located between i^{th} and (i + 1)th element of $ux_1 1$ $i \leq, |ux_1| \Rightarrow i + |x_2| \leq |ux_1 x_2|$ so when x_2 follows *y* it cannot go after $ux_1 x_2$ Therefore, *y* is the crossing of the site x_1 *y* x_2 in $ux_1 1 x_2 q$and the desired result is satisfied.

Example 3.6.

Given a set consisting of four restriction enzymes: *Dpn*I, *Dpn*II, *Bam*HI and *Bcl*I. Since there are four enzymes in the set, there are $2^4 = 16$ subsets for this set of four enzymes. Let *B* be any subset that contains at least {*Dpn*I,*Dpn*II}. Every splicing system with such set *B* as its right patterns is persistent. Indeed, the cleavage patterns of the enzymes, *Dpn*I, *Dpn*II, *Bam*HI and *Bcl*I are :

$([G/C][A/T],1,[T/A][C/G]),(1,[G/C][A/T][T/A][C/G],1),$

$([G/C],[G/C][A/T][T/A][C/G],[C/G])$ and $([T/A],[G/C][A/T][T/A][C/G],[A/T])$

respectively. So for every such splicing system the conditions of the previous theorem are satisfied and it is persistent

***Conclusion*:**

This paper investigates on persistent splicing systems and introduces some sufficient conditions for a splicing system to be persistent. Some real examples are also presented to illustrate this theory.

ACKNOWLEDGMENTS

We would like to thank the Ministry of Higher Education (MOHE) and Research Management Centre (RMC), UTM for the FRGS Vote No. 78482.

REFERENCES

Bonizzoni, P., C. De Felice, and G. Mauri, 2005. "Recombinant DNA, gene splicing as generative devices of Formal Languages", Proceedings "CiE 2005: New Computational Paradigms" (Special Session on Biological Computation), Lecture Notes in Computer Science 3526: 65-67 Springer.

Fong, W.H. 2008. “Modelling of splicing systems using formal language theory.” Ph.D. Thesis. Universiti Teknologi Malaysia

Head, T., 1987. “Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors.” Bulletin of Mathematical Biology, 49: 737-759.

Head, T., 1998. “Splicing representations of strictly locally testable languages.” Discrete Applied Mathematics, 87: 139-147.

Linz, P., 2001.“An introduction to formal languages and automata.” 3rd. ed. USA: Jones and Bartlett Publishers, Inc.