Crossing-Preserved and Persistent Splicing Systems

Fariba Karimi, Nor Haniza Sarmin
Department of Mathematics, Faculty of Science,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia.
fk.karimi@gmail.com, nhs@utm.my

Fong Wan Heng
Ibnu Sina Institute for Fundamental Science Studies,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia.
fwh@utm.my

Abstract— **By the introduction of the notion of splicing system by Head in 1987, a new approach for bio-inspired problems was made. This mathematical model helps to interpret the behavior of restriction enzymes on DNA molecules when they are cut and pasted. The theoretical skeleton of this model was based on formal language theory. Several types of splicing systems have been defined by different mathematicians. One of those is the persistent splicing system in which the property of being crossing of a site is preserved. In this paper, we introduced two new concepts, namely self-closed and crossing-preserved splicing patterns. The connection of these concepts with the persistent splicing systems is investigated. Some examples are provided to illustrate the relations.**

Keywords- Splicing Systems; Persistent Splicing Systems; DNA Molecules; Formal Language Theory

I. INTRODUCTION

Head defined splicing system to model the following biological process. If there are some finite DNA molecules and restriction enzymes in a test tube, after DNAs are cut and reassociated by restriction enzymes, some new hybrid DNAs will be produced [1]. DNA or deoxyribonucleic acid that is the genetic material is a chain of nucleotides. Each nucleotide consists of three components, phosphate, sugar and a nitrogenous base. The sugar molecule consists of five carbons that are numbered 1' to 5'. By attaching a phosphate and a base to the 5' and 1' carbon of a sugar respectively, a nucleotide is formed. Nucleotides themselves can join together and form a single stranded DNA. Each DNA has four kinds of bases, that are adenine, guanine, cytosine and thymine, which are abbreviated by A, G, C and T respectively. A double stranded DNA is the genetic material of all cellular organisms except some viruses. It is made of two single strands of DNAs that are linked together in a helical shape with the hydrogen bonds between the bases. In 1953, it was shown that the bases can join with a complementarity, A with T and C with G respectively [2].

Restriction enzymes are sequence-specific enzymes that can recognize a particular sequence of bases in a DNA and break it in two fragments with complementary staggered or blunt ends that are cut straightly. The cutting is done by breaking the phosphodiester bond between adjacent nucleotides. For the sticky ends, the single stranded tails of the fragments can be in two types in order for their free 3'-OH group or 5'-phosphate group to be able to join to the phosphate group and OH group of another nucleotide, respectively. The tails that terminate with an open 5' end and open 3' end are called 5' overhangs and 3' overhangs, respectively. If there is another kind of enzyme which is called DNA ligase, the fragments can rejoin together and make new DNAs. In fragments with staggered or sticky ends, the important point is that only the ones that have ends with complementary bases and the same overhangs can join together [2].

According to the structure of DNA, DNA molecule can be considered as a sequence over four alphabets [A/T], [T/A], [C/G] and [G/C]. The patterns of restriction enzymes for cutting DNAs are associated with some rules in splicing systems. So, the collection of DNAs that can be produced after the action of enzymes are associated with a language called a splicing language [1]. After presenting the definition of splicing system by Head on the basis of this biological problem, various related definitions and theorems were developed by him and many other mathematicians from different points of views [3]. Some of the definitions were made in the way that they preserve its biological background and can model and analyze the biological problems [4-5]; while some others paid attention to it from language generating point of view and its computational power [6-7].

In this paper, we mainly focus on persistent splicing systems. The interesting point about these systems is that if restriction enzymes are chosen from actual biological sense, then the resulting systems are often persistent [8]. Moreover, any splicing system that has only one restriction enzyme is persistent [1]. Two new concepts of self-closed and crossing-preserved splicing systems are introduced here for some characterization of the persistent splicing systems.

II. PRELIMINARIES

This section gives some basic concepts and notations that are used in this paper.

Definiton 1. [9] **(Alphabet, String)**
A finite, nonempty set A of symbols is called an *alphabet*. Any finite sequence of symbols from alphabet is called a *string*.

The *empty string* which is a string with no symbol at all is usually denoted by 1.

If A is an alphabet, we use A^* to denote the set of strings obtained by concatenating zero or more symbols from A.

Any subset of A^* is called a *language* over A.

978-0-7695-4514-1/11 $26.00 © 2011 IEEE
DOI 10.1109/BIC-TA.2011.23

Definition 2. [1] **(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 $L = L(S)$ generated by S consists of the strings 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 3. [1] **(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 $ucxdv$ (respectively $pexfq$) then this same subsegment y of $ucxfq$ contains an occurrence of the crossing of a site in $ucxfq$.

III. Main Results

In this section, two new concepts are introduced, namely self-closed and crossing-preserved. The relation between these notions and persistent splicing systems are stated in the main theorem.

Definition 1. (Self-closed)

A set of splicing patterns B is called *self-closed* if the set of its sites is closed under its splicing rules. A splicing system is called *self-closed* if the sets of their patterns are self closed.

Definition 2. (Crossing-preserved)

A set of splicing patterns B is called *crossing-preserved* if for every pattern (c, x, d) that its crossing x contains a substring x_1 that is crossing of another site, then the pattern (cu, x_1, vd) is also in B where $x=ux_1v$. A splicing system is called *crossing-preserved* if the sets of their patterns are crossing-preserved.

Main Theorem

If a splicing system S is self-closed and crossing-preserved and its crossings are substrings of each other then it is persistent.

Proof.

To show that S is persistent, the patterns with the same crossings and same hand should be considered. Suppose that $ucxdv$ and $pexfq$ are two strings from A^* such that (c, x, d) and (e, x, f) are in B. The string $ucxfq$ will be obtained via splicing. If y is a subsegment of ucx that is the crossing of a site in $ucxfq$ then according to the hypothesis that all crossings are substrings of each other, two cases may happen:

Case one: Suppose that y contains x. Since S is self-closed, (c, x, f) also is in B. So y contains x as a crossing of the site cxf in $ucxfq$.

Case two: Suppose that y is a substring of x. Since S is crossing-preserved, B contains the splicing pattern (cx_1, y, x_2f) where $x = x_1yx_2$. Therefore, y is crossing of the site cx_1yx_2f in $ucxfq$. Thus, S is persistent. □

Although the appearance of the three conditions in the main theorem is sufficient for a splicing system to be persistent, they are not necessary conditions. In other words the converse of the theorem is not true.

In the following example it is shown that there exists a persistent splicing system S that is crossing-preserved but not self-closed. Thus, being self-closed is not a necessary condition for being persistent.

Example 1

Let $S = (A, I, B, C)$ be a splicing system with the alphabet $A = \{b, c, x, d, e, f\}$, I is an arbitrary subset of A^*, $B = \{(bc, x, d), (e, x, f), (c, x, f), (e, x, d), (c, x, d)\}$ and C is the empty set.

S is not self-closed. In fact, if we consider the strings $bcxd$ and exf then the string $bcxf$ obtained from those by splicing is not in B.

S is crossing-preserved. Indeed, all the crossings are the same.

However, S is persistent. To show that it is persistent, the patterns with the same hand and the same crossings should be considered. But, in this example there exists only one crossing x. Suppose that $u_1w_1xz_1v_1$ and $u_2w_2xz_2v_2$ are two strings from A^* such that (w_1, x, z_1) and (w_2, x, z_2) are in B. The string $u_1w_1xz_2v_2$ will be obtained via splicing. If y is a substring of u_1w_1x that is crossing of a site in $u_1w_1xz_1v_1$, then y is equal to x (since all crossings are the same) and there exists w_3 and z_3 such that (w_3, y, z_3) is in B and w_3yz_3 is a substring of $u_1w_1xz_1v_1$. Now, we should show that y contains crossing of a site in $u_1w_1xz_2v_2$. The following two cases will be considered.

Case one: If w_3yz_3 is a substring of u_1w_1x, then y is again the crossing of the site w_3yz_3 in $u_1w_1xz_2v_2$.

Case two: Suppose w_3yz_3 is not a substring of u_1w_1x. Since by hypothesis, w_3y is a substring of u_1w_1x and according to the patterns of B, z_3 can only be the two single alphabets d or f, therefore, z_3 should appear after x and y which coincides with x in $u_1w_1xz_1v_1$. Then, w_1yz_2 *is* a substring of $u_1w_1xz_2v_2$. The string w_1 can be ab, b or e. In all cases, y is a crossing of one of the sites $(b, y=x, z_2)$ or $(e, y=x, z_2)$ where $z_2=d$ or f. Thus S is persistent. □

In the following example it is shown that there exists a persistent splicing system S that is neither self-closed nor crossing-preserved. So, being crossing-preserved is not a necessary condition for being persistent either.

Example 2

Let S be the splicing system associated with the enzymes *CviQI* and *Acc*65I with the patterns (g, ta, c) and $(g, gtac, c)$ respectively. In fact, $S = (A, I, B, C)$ where $A = \{a, c, g, t\}$, I (unspecified), $B = \{(g, ta, c), (g, gtac, c)\}$ and C is the empty set.

S is not self-closed because it does not contain the string $gtacc$ as its site.

S is not crossing-preserved because it does not contain the pattern (gg, ta, cc) as its splicing patterns.

However, S is persistent. Indeed, since the two enzymes have 5' overhangs, they are of the same hand. To

show that S is persistent, the patterns with the same crossings should be considered. There are just two different crossings in this system. So, if $u_1w_1xz_1v_1$ and $u_2w_2xz_2v_2$ are two strings from A^* such that (w_1, x, z_1) and (w_2, x, z_2) are in B, then two cases may happen:

Case 1: If $x = ta$ and therefore $(w_1, x, z_1) = (w_2, x, z_2) = (g, ta, c)$, $u_1w_1xz_1v_1 = u_1gtacv_1$ and $u_2w_2xz_2v_2 = u_2gtacv_2$.

Now, we should show that if y is a substring of u_1gta that is crossing of a site in u_1gtacv_1, then y contains crossing of a site in u_1gtacv_2. Since, y is crossing of a site in u_1gtacv_1, according to B, $y = ta$ or $y = gtac$ and in both cases the site gyc is a substring of u_1gtacv_1. On the other hand, gy is a substring of u_1gta. So, if $y = gtac$, then gyc should be in u_1 and the desired result will be obtained. In fact, y again is the crossing of the site gyc in u_1gtacv_2. Otherwise if $y = ta$, then gy is either in u_1 (that obviously the desired result will be obtained) or is the rightmost substring of u_1gta. In the latter case, y is crossing of the site $gtac$ in u_1gtacv_2.

Case 2: If $x = gtac$ and therefore $(w_1, x, z_1) = (w_2, x, z_2) = (g, gtac, c)$, $u_1w_1xz_1v_1 = u_1ggtaccv_1$ and $u_2w_2xz_2v_2 = u_2ggtaccv_2$.

Similarly, we should show that if y is a substring of u_1ggtac that is crossing of a site in $u_1ggtaccv_1$, then y contains crossing of a site in $u_1ggtaccv_2$. Since, y is crossing of a site in $u_1ggtaccv_1$, according to B, $y = ta$ or $y = gtac$ and in both cases the site gyc is a substring of u_1gtacv_1. On the other hand, gy is a substring of u_1ggtac. So in both cases of y the string gy should be in u_1 (that obviously the desired result will be obtained) or in $ggtac$. If $y = ta$ and in $ggtac$ then y is crossing of the site in $u_1ggtaccv_2$. If $y = gtac$ and in $ggtac$ then y is crossing of the site $ggtacc$ in $u_1ggtaccv_2$. Thus S is persistent. □

In the following example it is shown that there exists a splicing system S that is self-closed and its crossings are substrings of each other but not persistent. In fact, self-closed itself is not a sufficient condition for being persistent and a splicing system should meet all the conditions of the main theorem.

Example 3

Let $S = (A, I, B, C)$ be a splicing system with the alphabet $A=\{c, x_1, x_2, d_1, d_2, e, f\}$, I is an arbitrary subset of A^*, $B=\{(c, x_1x_2, d_1d_2), (e, x_1x_2, f), (c, x_1x_2, f), (e, x_1x_2, d_1d_2), (c, x_1, x_2d_1)\}$ and C is the empty set.

Then S is self-closed. In fact, it can be easily seen that the set of its sites $\{cx_1x_2d_1d_2, ex_1x_2f, cx_1x_2f, ex_1x_2d_1d_2, cx_1x_2d_1\}$ is closed via the splicing rules in B.

However S is not persistent. Indeed, if we consider the strings $cx_1x_2d_1d_2$ and ex_1x_2f and the string cx_1x_2f obtained from those by splicing. Then, x_1 is a substring of cx_1x_2 that is crossing of the site $cx_1x_2d_1$. But, x_1 does not contain an occurrence of the crossing of any site in cx_1x_2f. Thus S is not persistent. □

IV. Conclusion

In this paper, two new notions of self-closed and crossing-preserved splicing systems are introduced. The relation between these splicing systems and persistent splicing systems has been investigated in the main theorem. Some examples are provided to clarify the relations, showing that the converse of the main theorem is not true.

Acknowledgment

We would like to thank the Ministry of Higher Education (MOHE) and Research Management Centre (RMC), UTM for the RUG Vote No. Q. J130000.7126.02J65.

References

[1] T. Head, "Formal language theory and DNA: An analysis of the generative capacity specific recombinant behaviors," Bull. Math. Biology, vol. 49, 1987, pp. 737-759.

[2] R. H. Tamarin, Principle of Genetics. USA: The McGraw-Hill Companies, 2001.

[3] P. Bonizzoni, C. Ferretti, G. Mauri, and R. Zizzaet, "Separating some splicing models" Information Processing Letters, vol. 79(6), 2001, pp. 255-259.

[4] R.W. Gatterdam, "Splicing systems and regularity," International Journal of Computer Math., vol. 31, 1989, pp. 63-67.

[5] Goode, E. and D. Pixton, "Splicing to the limit," in Aspects of Molecular Computing, N. Jonoska, G. Paun, and G. Rozenberg, Editors., Springer Berlin / Heidelberg. 2004, pp. 236-254.

[6] C. De Felice, G. Fici, and R. Zizza, "A characterization of regular circular languages generated by marked splicing systems," Theoretical Computer Science, vol. 410(47-49), 2009, pp. 4937-4960.

[7] V. Mitrana, I. Petre, and V. Rogojin, "Accepting splicing systems, " Theoretical Computer Science, vol. 411(25), 2010, pp. 2414-2422.

[8] T. Yokomori, and S. Kobayashi, "Learning local languages and their application to DNA sequence analysis,". IEEE Trans. Pattern Anal. Mach. Intell, vol. 20(10), 1998, pp. 1067-1079.

[9] P. Linz, An introduction to formal languages and automata. Jones and Bartlett Publishers, 2006.