Jurnal
Teknologi

Full Paper

MODELLING OF TWO STAGES DNA SPLICING LANGUAGES ON DE BRUIJN GRAPH

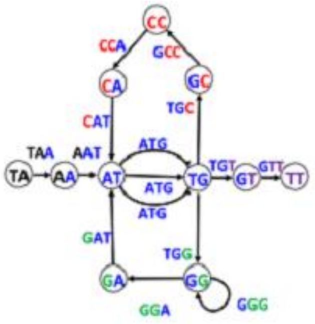
Article history
Received
17 March 2015
Received in revised form
7 July 2015
Accepted
15 December 2015

Mohammad Hassan Mudaber[a*], Yuhani Yusof[b], Mohd Sham Mohamad[b], Aizi Nor Mazila Ramli[b], Wen Li, Lim[b]

[a]Department of Mathematics, Faculty of Natural Sciences, Kabul Education University, Afshar District, Kabul, Afghanistan
[b]Faculty Industrial Sciences and Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan Pahang Darul Makmur, Malaysia

*Corresponding author
hmudaber@yahoo.com

Graphical abstract



Abstract

Finding the sequence of the genome from its compositions as well as a mathematical graph is the most interesting topic in a field of DNA molecular. Since lack of technology is the big obstacle that biologists are facing to read a long sequence of the genome from beginning up to the end, therefore finding the compositions of the genome having very long sequence and also its description via de Bruijn graph is challenging or even impossible. In this paper, Yusof-Goode (Y-G) approach is used to generate the DNA splicing languages based on cutting sites of initial strings (one or two cutting sites) and crossing and contexts factors of restriction enzymes. The two short sequences of DNA (8bp) and two restriction enzymes are considered to create a connection between mathematics and DNA molecular. This relation will be presented as de Bruijn graph so that every edge of the de Bruijn graph gives a k-mer composition of DNA molecule and also each path of the de Bruijn graph gives a DNA sequence and vice-versa. Besides, the persistency and permanency of two stages DNA splicing languages can be predicted using this model.

Keywords: de Bruijn graph, permanent, persistent, two stages splicing languages, Y-G splicing system

Abstrak

Mencari jujukan genom daripada komposisinya serta graf matematik adalah topik yang paling menarik dalam bidang molekul DNA. Oleh kerana kekurangan teknologi adalah halangan besar yang dihadapi oleh ahli biologi iaitu dalam membaca jujukan genom yang panjang dari awal hingga ke akhir, maka, mencari komposisi genom yang mempunyai jujukan panjang serta keterangannya melalui graf de Bruijn adalah satu cabaran atau mungkin mustahil. Dalam kertas kerja ini, pendekatan Yusof-Goode (Y-G) digunakan untuk menghasilkan bahasa hiris-cantum DNA berdasarkan bahagian pemotongan jujukan awal (satu atau dua bahagian pemotongan) dan pindah silang dan faktor konteks bagi enzim pembatas. Dua jujukan pendek DNA (8bp) dan dua enzim pembatas adalah dipertimbangkan untuk mencipta hubungan antara matematik dan molekul DNA. Hubungan ini akan dipersembahkan sebagai graf de Bruijn supaya setiap pinggir graf de Bruijn memberikan komposisi k-mer molekul DNA dan juga setiap laluan daripada graf de Bruijn memberikan jujukan DNA dan sebaliknya. Di samping itu, konsep berterusan dan kekal bagi dua peringkat bahasa hiris-cantum DNA boleh diramal dengan menggunakan model ini.

Kata kunci: graf de Bruijn, kekal, berterusan, bahasa hiris-cantum berperingkat dua, sistem hiris-cantum Y-G

1.0 INTRODUCTION

Deoxyribonucleic acid (DNA) is the genetic material of the living organism constructing from nucleotides. Therefore, each sequence of DNA consists of three parts: nitrogenous bases (adenine, guanine, cytosine and thymine), deoxyribose sugar and phosphate group. The structure of DNA is a double helix; the sugar-phosphate (backbone) is the two sides, and the bases are inside the double helix. Since DNA is a double-stranded molecule, the two strands of DNA are antiparallel and have two opposite directions. The bases (adenine with thymine and guanine with cytosine and vice-versa) are connected to each other by hydrogen bonding. There are two sharing hydrogen bonds between adenine and thymine and three sharing hydrogen bonds between guanine and cytosine [1]. In the other word, adenine is only paired with thymine and guanine is paired by cytosine, and vice-versa. This rule of pairing is symbolized as [*A*/*T*], [*T*/*A*], [*C*/*G*] and [*G*/*C*]. Therefore, based on these rule pairings, the two strands of DNA are mutually complementary to each other as presented below.

$$5'\ldots CGAGCTCG\ldots 3'$$
$$3'\ldots GCTCGAGC\ldots 5'$$

A restriction enzyme cuts the DNA from specific sequences resulting in molecules with staggered or blunt ends. There are different types of restriction enzymes. Each of them recognizes and cut the strand of DNA in a different way. Then the fragments of DNA with their complementariness can be combined by ligase to produce new DNA molecules [2].

Since the double-stranded DNA (dsDNA) consists of millions of base pairs; it is hard to read the nucleotides from beginning up to end with recent technology. Therefore, the scientists and researchers consider the short sequence of DNA mostly between 500 bp-1200 bp long [3]. The sequence of DNA was described on de Bruijn graph with respect to its 3-mer compositions [3-4]. The way of constructing the bi-directed de Bruijn graph for illustrating the genome path, where each *k*-molecule is represented only once was shown [5]. The splicing languages that are produced by the null-context, uniform and simple splicing systems were modelled on the automaton diagram [6-7]. Mudaber [8] introduced the concept of two stages in splicing system and defined the languages, which are produced by splicing system as two stages splicing languages. In this paper, the two stages DNA splicing languages are modelled on a de Bruijn graph. The important point of this research, which is different from previous researchers is that, it describes more than one splicing languages (DNA sequences) on the de Bruijn graph.

2.0 PRELIMINARIES

In this section, some definitions and concepts related to this research are given. De Bruijn graph is an appropriate way of representing the DNA strands according to its reads. The recombinant DNA strands are divided into *k*-mers composition, and by connecting the *k*-mers to each other the desired de Bruijn graph will be constructed. In the resulted de Bruijn graph, each *k*-mer of reads shows an edge of the de Burijn graph, and each *k*-1 prefix of each *k*-mer ((*k*-1)-mer) shows a node of the de Bruijn graph. Therefore, to develop a de Bruijn graph from the resulting DNA strands, Y-G splicing system is preferred among all models of splicing system in generating the two stages DNA splicing languages, since Y-G splicing system is the appropriate approach on studying the biological aspect of DNA splicing system in a transparent way. Hence, Y-G splicing system is defined below.

Definition 1: [9] Yusof-Goode (Y-G) Splicing System
A Y-G splicing system is the form $S=\left(A,I,R\right)$, where A is the set of four alphabets a,g,c and t, I is the set of initial strings of double- stranded DNA and R is the set of splicing rules that indicates the set of enzymatic operation. The rule R in Y-G model is presented as $\left(u;x,v:y;x,z\right)$ and $\left(u,x;v:y,x;z\right)$ which show the left pattern and right pattern, respectively. However, the notation $\left(u,x,v:y,x,z\right)$ indicates that both patterns of rules were applied on DNA string. If $r\in R$, where $r=\left(u,x,v:y,x,z\right)$ and $s_1=\alpha uxv\beta$ and $s_2=\gamma yxz\delta$ are elements of I, then splicing s_1 and s_2 using r produce the initial string I together with $\alpha uxz\delta$ and $\gamma yxv\beta$, presented in either order where $\alpha,\beta,\gamma,\delta,u,v,x$ and $y\in A^*$ are free monoid generated by A with the concatenation operation and 1 as the identity element.□

The most important characteristic of a string and a rule in splicing system is palindromic. Being a string or crossing site of a rule palindromic effects on the number of resulted DNA splicing languages. If the crossing site of the splicing rule be palindromic, the rotation of 180 degrees of string fragments after cutting by the rules can also re-join to produce new DNA splicing languages. Therefore, the definition of palindromic string, which was introduced by Yusof [9] is viewed below.

Definition 2: [9] Palindromic
A string I of double stranded deoxyribonucleic acid (dsDNA) is said to be palindromic if the sequence from the left side of the upper single strand is equal with the sequence from the right side of the lower single strand.□

In the next definition, the two concepts of prefix and suffix, which are used for finding the *k*-mer compositions of the recombinant DNA strand from its sequences, are viewed bellow.

Definition 3: [10] Prefix and Suffix of a String
Any string of consecutive symbols in some word w is said to be a substring of w. If $w=vu$ then the substrings v and u are said to be a prefix and a suffix of w, respectively. For example, if $w=abbab$, then $\left\{\lambda,a,ab,abb,abba,abbab\right\}$ is the set of all prefixes of w, while $\left\{b,ab,bab,bbab\right\}$ is the set all suffixes of w. □

In terms of DNA recombination, when the restriction enzyme is applied on initial DNA strand, at the existence of ligase the fragments of DNA molecules will be fused together with its complementary ends to form the new hybrids DNA molecule. The resulted DNA molecules sometimes contain the cleavage pattern of restriction enzyme and can be cut by the restriction enzyme if the reaction needs to achieve at second stage. Mathematically, for this property of recombinant DNA strands is called persistent. Hence, the definition of persistent is given next.

Definition 4: [11] Persistent Splicing System
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 hands: if y is a sub segment of ucx (respectively xfq) that is crossing of a site in $ucxdv$ (respectively $pexfq$) then this same sub segment y of $ucxfq$ contains an occurrence of a crossing of a site in $ucxfq$.□

In the next definition, a proper case of persistent, which is called permanent, is defined below.

Definition 5: [12] Permanent Splicing System
Let $S=(A,I,B,C)$ be a splicing system. Then S is permanent if for each pair of strings $ucxdv$ and $pexfq$ in A^* with (c,x,d) and (e,x,f) patterns of the same hands: if y is a sub segment of ucx (respectively xfq) that is crossing of a site in $ucxdv$ (respectively $pexfq$) then this same sub segment y of $ucxfq$ is an occurrence of a crossing of a site in $ucxfq$.□

To develop a de Bruijn graph from the resulting DNA strands, the definition of k-mers composition is first given.

Definition 6: [13] *K*-mers Composition
Given a string Text, its *k*-mers composition COMPOSITION$_k$(Text) is the collection of all *k*-mers substrings of Text (including repeated *k*-mers).□
For example, the composition$_3$ of the genome $TATGGGGTGC$ are as follows:
$ATG, GGG, GGG, GGT, GTG, TAT, TGC, TGG.$
The set of *k*-mers composition of a genome is called a *k*-spectrum. In terms of genome assembly, de Bruijn graph was defined by Pevzner *et al*. (2013) in order to represent the genome sequence as a model. The definition of de Brujin graph is stated below.

Definition 7: [13] de Bruijn graph
A de Bruijn graph is a directed graph where a *k*-mer composition of a genome is assigned to each edge and a (*k*-1)- mer represents each node of the graph and also a path in the graph visits every edge exactly once.□

Since in this investigation the two stages splicing languages are modelled on de Bruijn graph, thus the definition of two stages splicing languages that was introduced by Mudaber [8] is stated below.

Definition 8: [8] Two Stages Splicing Languages
Let $S=(A,I,R)$ is a Y-G splicing system. Furthermore, let $L=L(S)$ be set of stage one splicing languages produced by splicing system S and $L'=L'(S)$ be set of stage two splicing languages produced by S that consists of $L=L(S)$ and all splicing languages that can be resulted by splicing L. Then, the union of stage one and stage two splicing languages are called two stages splicing languages.□

The above definition can be represented in the form of diagram as well to show $L(S)$, $L'(S)$ and $L(S)\cup L'(S)$.

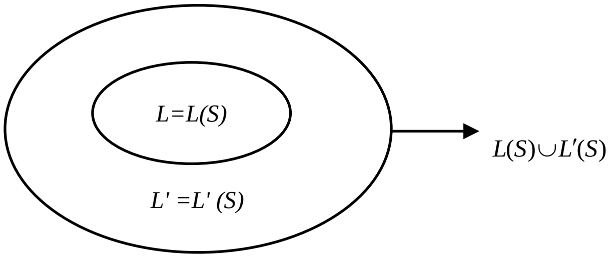


Figure 1 Graphical Representation of Definition of Two Stages Splicing Languages

In terms of DNA splicing system, the splicing languages which are produced by splicing operation are a set of recombinant DNA strands and each of this splicing language represents a distinct path on the graph. Therefore, to illustrate all the resulted DNA splicing languages on de Bruijn graph, the definition of de Bruijn graph is rewritten as de Bruijn multi-graph in terms of splicing system.

A directed graph with respect to Y-G splicing system is called a de Bruijn multi-graph if it has the following properties:

- Each edge of the graph represents a *k*-mer composition of DNA splicing language.
- Each node of the graph represents a (*k*-1)-mer composition of splicing language.
- The graph has n, $n\in\mathbb{Z}^+$ distinct paths, where each path represents a splicing language.
- Each path visits every edge of the graph that are containing on that path exactly once.
- The loop is allowed to each node.

In the next section, the de Bruijn graph of two stages DNA splicing languages is discussed.

3.0 DEBRUIJN GRAPH OF TWO STAGES SPLICING LANGUAGES

In this section, the generating DNA splicing languages of two stages are described on de Bruijn graph. Since the concentration of this research is on the number of cutting sites of DNA strands and crossing sites factor of splicing rules, therefore two different cases are considered to model the recombinant DNA strands on de Bruijn graph. Thus, for each case a de Bruijn graph from the resulting splicing languages will be constructed containing all sets of stage one and stage two splicing languages. In addition, the graph determines whether the two stages DNA splicing languages are persistent and permanent or not.

Case 1: Modelling de Bruijn graph of two stages DNA splicing languages with respect to two initial strings (with one cutting site) and two rules (with palindromic disjoint crossing sites)

In this case, the two stages DNA splicing languages, which are produced by Y-G splicing system consisting of two initial strings and two rules with disjoint crossing sites and palindromic sequences, are represented on de Bruijn graph. Assume there is a test tube contains of two initial DNA strands named s_1 and s_2 and two restriction enzymes namely $Mse\text{I}$ and $Hpa\text{II}$ with NEB CutSmart™ Buffer and appropriate ligase. After the initial DNA strands are cut by the restriction enzymes, then their fragments will be rejoined to form the new DNA strands. This process of DNA recombination is mathematically presented in the following example.

Example 3.1: Let $S=(A,I,R)$ is a Y-G splicing system such that $A=\{a,g,c,t\}$ and the set R consists of two rules namely, $(t;ta,a:t;ta,a)$ and $(c;cg,g:c;cg,g)$. Suppose $s_1=gcttaacg$ and $s_2=atccggta$ are two arbitrary initial strings in I. When splicing takes place the four distinct DNA splicing languages are generated at two stages besides the initial strings as listed below.

$$gcttaagc, cgttaacg, atccggat, taccggta\ .$$

Since the initial strings have different sequences and also the crossing sites of rules are disjoint, thus there is no possibility for recombining the string fragments after splitting by the rules. The only way is to recombine the fragments of each string with themselves. Therefore, the red and blue colours splicing languages show the splicing languages which are generated by $s_1=gcttaacg$ and $s_2=atccggta$, respectively at the existence of rules $(t;ta,a:t;ta,a)$ and $(c;cg,g:c;cg,g)$. Since the lengths of restriction enzymes are four base-pairs (4bp), the compositions₄ of the two stages splicing languages is considered. Therefore, using Mathematica V9.01 (StringCases["splicing language", _~~_~~_~~_~~ , Overlaps -> True]), the compositions₄ of the above splicing languages is considered.

$$gcttaacg = gctt, ctta, ttaa, taac, aacg$$
$$gcttaagc = gctt, ctta, ttaa, taag, aagc$$
$$cgttaacg = cgtt, gtta, ttaa, taac, aagc$$
$$atccggat = atcc, tccg, ccgg, cgga, ggat$$
$$taccggta = tacc, accg, ccgg, cggt, ggta$$
$$atccggta = atcc, tccg, ccgg, cggt, ggta$$

The two concepts of prefix and suffix are used for finding the nodes of the graph that refer to the first (k-1) nucleotides and last (k−1) nucleotides of a k-mer, respectively. Moreover, to glue the nodes together for constructing the de Bruijn graph, the suffix of a k-mer should be equal to the prefix of the following k-mer in the recombinant DNA strands. The prefix$(gctt)=gct$ and suffix$(gctt)=ctt$. Therefore, suffix $(gctt)=$prefix$(ctta)=ctt$, using this approach the nodes of the graph from the 4-mer compositions of the above splicing languages are found as below.

$$gct, ctt, tta, taa, aac, acg$$
$$gct, ctt, tta, taa, aag, agc$$
$$cgt, gtt, tta, taa, aac, agc$$
$$atc, tcc, ccg, cgg, ggt, gta$$
$$atc, tcc, ccg, cgg, gga, gat$$
$$tac, acc, ccg, cgg, ggt, gta$$

Based on the above representation, a de Bruijn graph with thirteen nodes and fifteen edges is constructed in order to model the two stages DNA splicing languages. The Figure 2 shows the de Bruijn of the above two stages splicing languages.

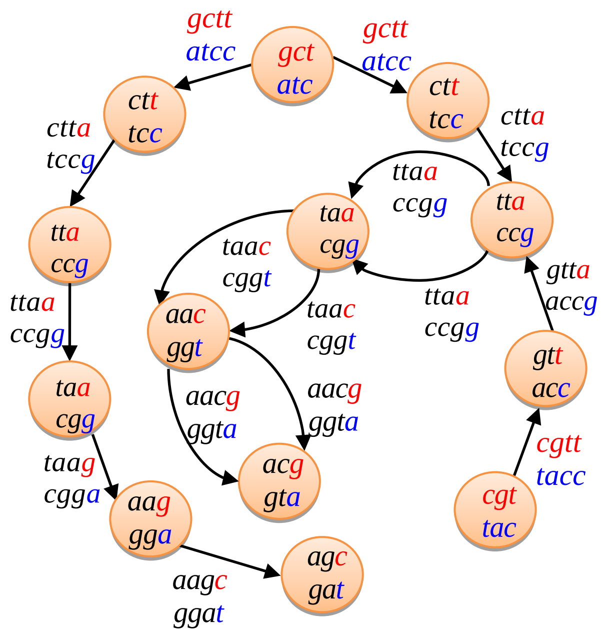


Figure 2 De Bruijn graph of two stages splicing languages with respect to two initial strings and two rules with palindromic disjointed crossing sites.

The above developed de Bruijn graph describes six distinct splicing languages and each of them represents a path of the de Bruijn graph. There are two distinct k-mers assigned to each edge of the graph, where the upper *k*-mers show the composition of the above three DNA splicing languages, which highlighted in red colour, and the lower *k*-mers show the composition of the three DNA splicing languages with blue colour. The initial strings have only one cutting sites and have length eight of base-pairs (8bp), and the resulted splicing languages also have the same length as initial strings. Therefore, in this situation, all the paths in the de Bruijn graph have the same length. Since each path of this graph contains five distinct edges and there is an edge in every path that contains the restriction enzyme sequence, thus the six DNA splicing languages that can be found from the above graph according to Definition 4.1 and 4.2 are persistent and permanent. Thus, as can be clearly seen from the above de Bruijn graph, there is a connectivity between two stages splicing languages which are generated by Y-G approach and field of DNA recombination, so that each path of the graph gives a persistent DNA splicing language (recombinant DNA strand) and reversely each DNA sequence is a path of the graph. These sequences of DNA (persistent and permanent splicing languages) can be resulted by walking along the paths of de Bruijn graph which all have overlapped connections and highlighted by red and blue colours.

In the next, the de Bruijn graph of two stages DNA splicing languages is constructed based on two initial strings and two rules with palindromic identical crossing sites.

Case 2: Modelling de Bruijn graph of two stages DNA splicing languages with respect to two initial strings and two rules with palindromic identical crossing sites

In this case, a biological example is provided with respect to Y-G splicing system consisting of two initial strings and two rules with identical palindromic crossing sites. The purpose of this example is to develop a de Bruijn graph for presenting the two stages DNA splicing languages as well as to illustrate the persistency and permanency of two stages DNA splicing languages.

Example 3.2: Let $S=(A,I,R)$ is a Y-G splicing system such that $A=\{a,g,c,t\}$ and the set of splicing rule R consists two restriction enzymes namely $CviQI$ and $BfaI$ which are represented mathematically in the form $(g;ta,c:c;ta,g)$. Suppose $s_1=cagtactg$ and $s_2=ctctagag$ are two arbitrary initial strings in I. Since the whole sequences of initial strings are palindromic, by applying the restriction enzymes on the initial DNA strands and then recombining the fragments by DNA ligase the following two distinct splicing languages will be generated at two stages besides initial strings of set I namely ,

$$cagtagag,\ \ ctctactg$$

To model the two stages DNA splicing languages on de Bruijn graph, the 4-mer compositions of DNA splicing languages with respect to length of restriction enzymes sequences are considered. Since the composition of the DNA splicing languages have overlapping connections, thus the composition$_4$ of the above two stages DNA splicing languages are presented below.

Composition$_4$$\left(cagtactg\right)$: $cagt, agta, gtac, tact, actg.$

Composition$_4$$\left(ctctagag\right)$: $ctct, tcta, ctag, taga, agag.$

Composition$_4$$\left(cagtagag\right)$: $cagt, agta, gtag, taga, agag.$

Composition$_4$$\left(ctctactg\right)$: $ctct, tcta, ctac, tact, actg.$

Based on the above *k*-mers compositions (edges of de Bruijn graph) the nodes of the de Bruijn graph can be obtained using the methods of prefix and suffix, if the (*k*-1) suffix of a *k*-mer is equal to the (*k*-1) prefix of the followed *k*-mer. Thus, based on this approach, the nodes of the graph for each path are obtained as below.

$$cag, agt, gta, tac, act, ctg.$$
$$ctc, tct, cta, tag, aga, gag.$$
$$cag, agt, gta, tag, aga, gag.$$
$$ctc, tct, cta, tac, act, ctg.$$

Since the edges and nodes of the de Bruijn graph have been found, therefore based on them a de Bruijn graph with seventeen nodes and twenty edges is developed, which consists all two stages DNA splicing languages. The graph is presented in Figure 3 below.

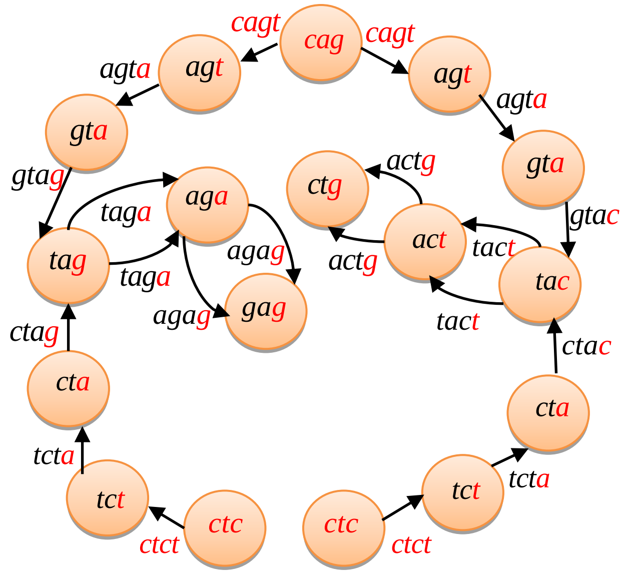


Figure 3 De Bruijn graph of two stages splicing languages with respect to two initial strings and two rules with palindromic identical crossing sites

As it can be clearly seen from the graph, it has four paths that contain four distinct DNA splicing languages. In the other words, this de Bruijn graph gives four different splicing languages that have been highlighted with red colour. This graph shows that the two stages DNA splicing languages are not persistent as well as not permanent, because the sequences of restriction enzymes have not been assigned with any edges in some of the paths in the graph. As a result by walking along each path on the de Bruijn graph four DNA splicing languages will be obtained from this graph. Biologically, these DNA splicing languages are not persistent, since they do not contain the cleavage patterns of restriction enzymes.

4.0 CONCLUSION

This study discusses on modelling of two stages splicing languages. This model which is developed as de Bruijn graph presents the connection between splicing languages which are produced mathematically by Y-G model and field of DNA recombination. The de Bruijn graphs of the generated DNA strands are constructed according to crossing sites of restriction enzymes at two different cases. The difference between these cases is only the crossing sites of restriction enzymes. In the first case, two rules with palindromic disjointed crossing sites is considered, while in Case 2, two rules with palindromic identical crossing sites, respectively are considered. Besides, from developed de Bruijn graphs, the persistency and permanency as well as non-persistency and non-permanency of two stages DNA splicing languages can be determined. If the cleavage pattern of restriction enzyme is assigned to an edge in a path of the graph, then the hidden splicing languages are persistent and permanent otherwise the splicing languages are non-persistent and non-permanent. In reality, there are mutually relations between paths of the de Bruijn graph and DNA sequences, so that every edge of the graph presents a *k*-mer composition of the recombinant DNA molecules as well as every path of the graph gives a unique sequence of the DNA and vice-versa.

Acknowledgement

The authors would like to express their gratitude to Ministry of Education of Malaysia and Research and Innovation Department of Universiti Malaysia Pahang (UMP) for financial funding through RDU Vote: 130354 and RAGS Grant Vote: RDU 131404. The first author also would like to express his especial thanks to Ministry of Higher Education (MOHE) of Afghanistan for his master scholarship and Kabul Education University (KEU) for the study leave.

References

[1] Robin, S., Rondolphe, F and Schbath, S. 2006. *DNA World and Model*. UK. Cambridge University Press.

[2] Walker, J. M. and Rapley, R. 2009. *Molecular Biology and Biotechnology*. London: Royal Society of Chemistry.

[3] Kaptcianos, J. 2008. A Graph Theoretical Aapproach to DNA Fragment Assembly. *American Journal of Undergraduate Research*. 7(1): ISSN 1536-4558.

[4] Pevzner, P. A. and Tang, H. 2001. Fragment Assembly with Double-Barreled Data. *Bioinformatics*. 17: 225-233.

[5] Medvedev, P. and Brudno, M. 2009. Maximum Likehood Genome Assembly. *Journal of Computational Biology*. 16: 1101-1116.

[6] Jinn, L. S., Fong, W. H., Sarmin, N. H. and Karimi, Fariba. 2011. Mathematical Modelling of Some Null-Context and Uniform Splicing System. *Journal of Fundamental and Applied Sciences*. 7(2): 145-149.

[7] Fong, W.H., Sarmin, N.H. and Ibrahim, Z. 2009. Recognition of Simple Splicing System Using SH-Automata. *Journal of Fundamental and Applied Sciences*. 4(2): 337-342.

[8] Mudaber, M.H., Yusof, Y. and Mohamad, S. M. Some Relations between Two Stages DNA Splicing Languages. *AIP Conf. Proc*. 1602: 254-259.

[9] Yusof, Y. 2012. *DNA* Splicing System Inspired by Bio Molecular Operation. Ph.D. Thesis. Universiti Teknilogi Malaysia.

[10] Linz, P. 2006. *An Introduction to Formal Languages Theory and Automata*. USA. Jones and Barlett Publisher.

[11] Head, T. 1987. Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors. *Bulletin of Mathematical Biology*. 49(6): 737-759.

[12] Gatterdam, R. 1992. Algorithm for Splicing System. *SIAM Journal of Computing*. 21(3): 507-520.

[13] Pevzner, P., Conpeau, P. E. C. and Vyahhi, Nikolay. 2013. How to Assemble Genome? *Bioinformatics Algorithms*. (online) https://www.coursera.org/-course/bioinformatics.