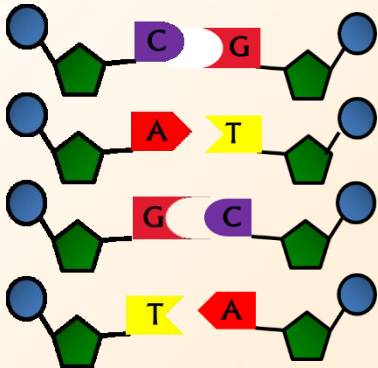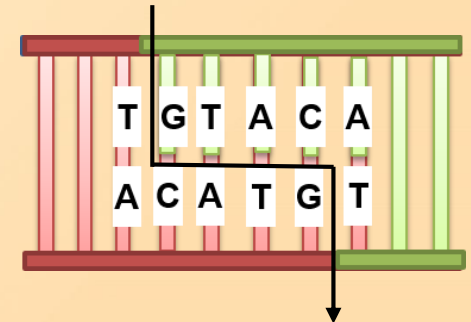# Wet Lab Experiments of DNA Splicing System

Dr. Fong Wan Heng

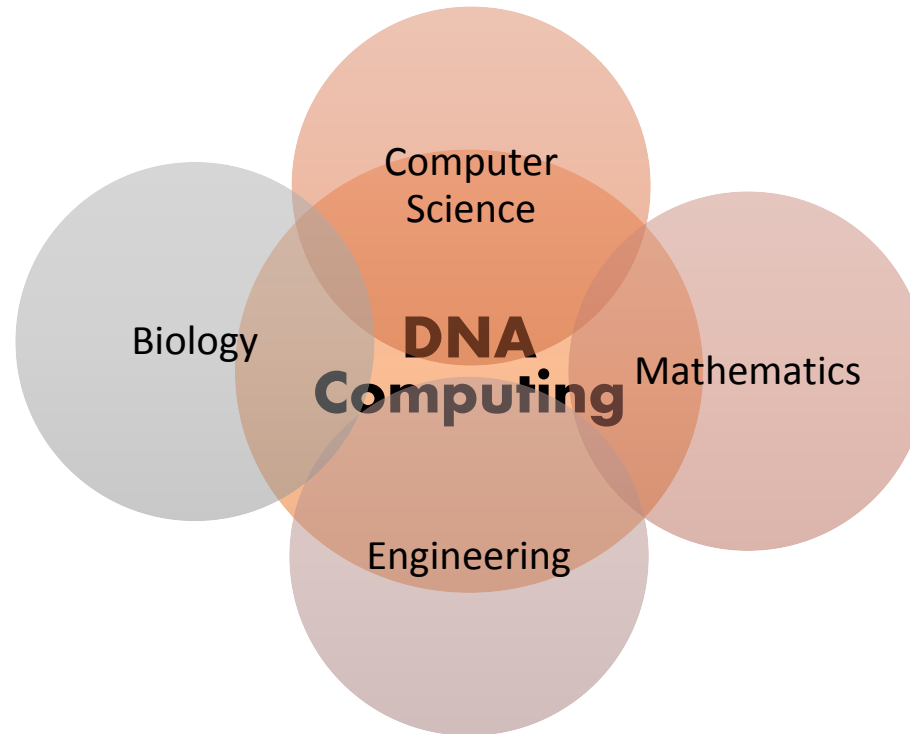**7th Bi-Weekly**

Applied Algebra and Analysis Group (AAAG) Seminar
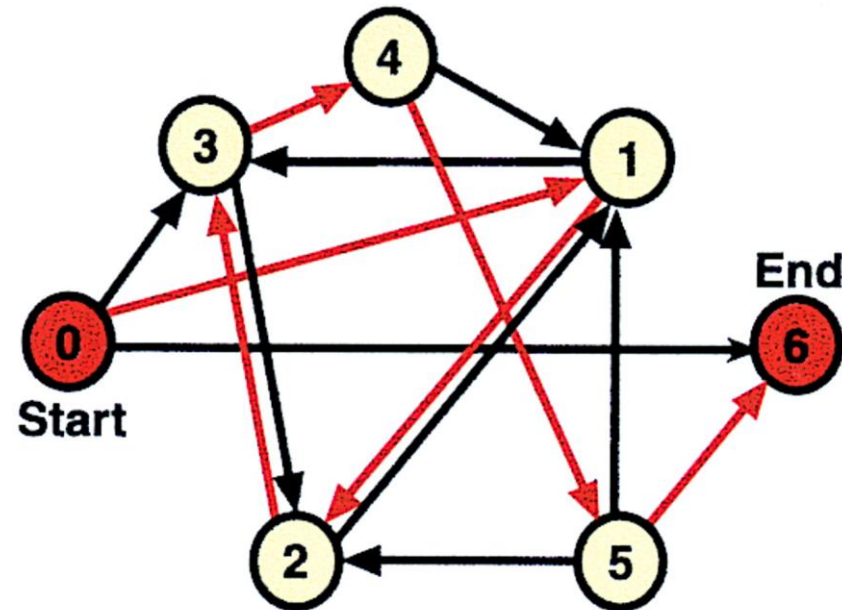
1 February 2021

innovative ● entrepreneurial ● global

DNA Computing

- DNA computating has emerged in the last twenty years as an exciting new research field at the intersection of computer science, biology, engineering and mathematics.

DNA Computation (cont.)

- Although anticipated by Feynman from the 1950s, the notion of performing computations at the molecular level was only realized in 1994, with Adleman's experiment on solving the Hamiltonian Path Problem using DNA [1].

1. L. M. Adleman, "Molecular Computation of Solutions to Combinatorial Problems," *Science*, vol. 266, no. 5187, pp. 1021-1024, 1994.

innovative ● entrepreneurial ● global

www.utm.my

DNA Computation (cont.)

- Since then the field has blossomed rapidly, with development of significant theoretical and experimental results by researchers from interdisciplinary areas.

- Different models of molecular computation have been proposed in scientific society including Splicing Models and Sticker Models.

Splicing Systems

- The mathematical modelling of splicing system was first defined by Head [2] in 1987.

- It was introduced as a mathematical model of the generative capacity of a biological system containing DNA molecules in the presence of appropriate enzymes.

2. T. Head, "Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors," *B. Math. Biol*, vol. 49, no. 6, pp. 737-759, 1987.

innovative ● entrepreneurial ● global

www.utm.my

Tom Head's visit to UTM, 2004

At Tom Head's house in Binghamton, USA, 2007

Research discussion at State University of New York (SUNY) Binghamton, USA, 2007

innovative ● entrepreneurial ● global

www.utm.my

Research seminar at SUNY Binghamton, 2007

Dinner at Prof Kappe's house, 2007

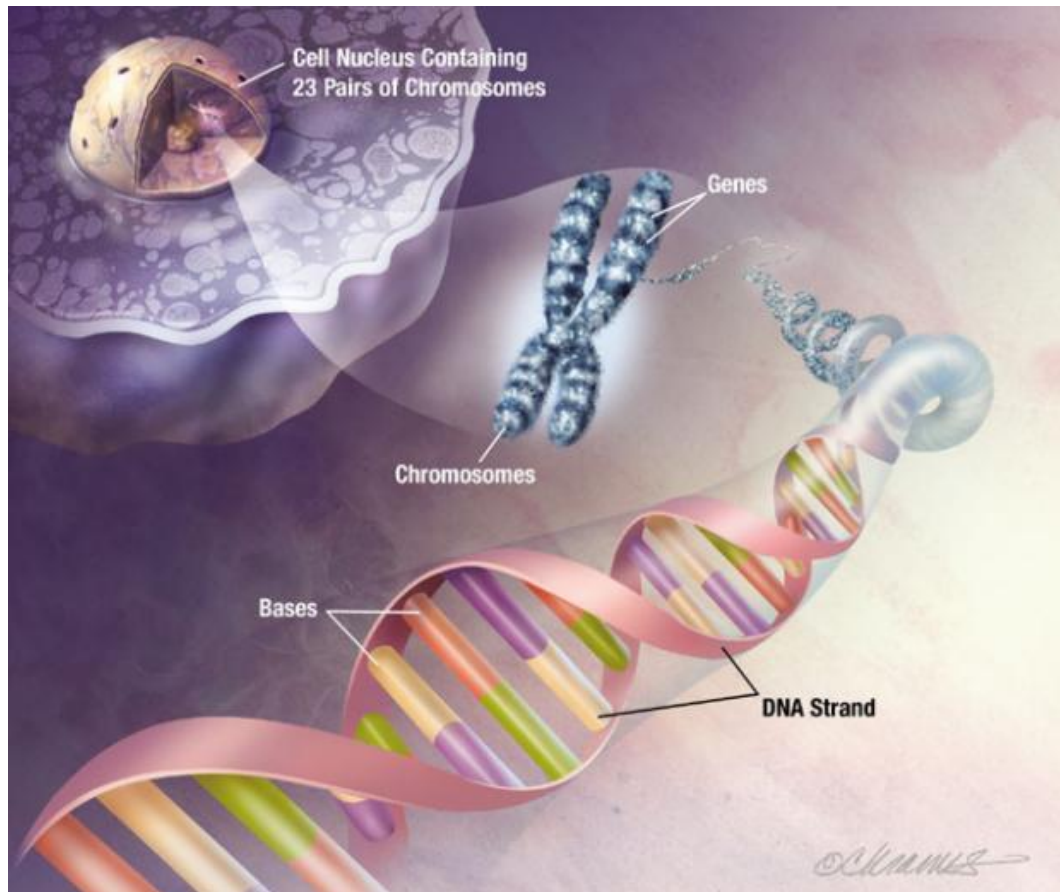**Three generations, 2007**

Lunch with Tom Head, 2007

Conference on Biomathematical Computing: Past, Present and Prospects, SUNY Binghamton, USA, 2008

Conference dinner, SUNY Binghamton, USA, 2008

Structure of DNA

- DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms.

Structure of DNA (cont.)
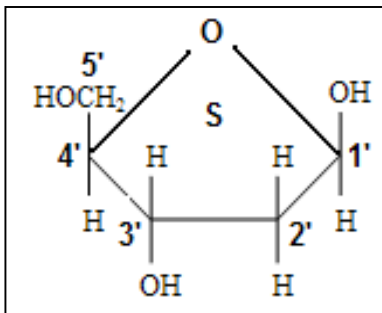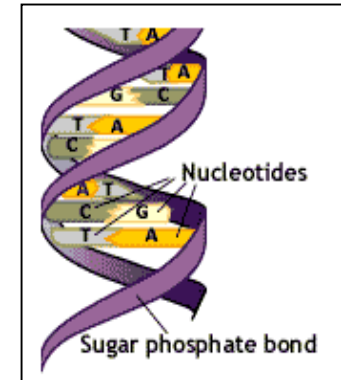
DNA – DeoxyriboNucleic Acid

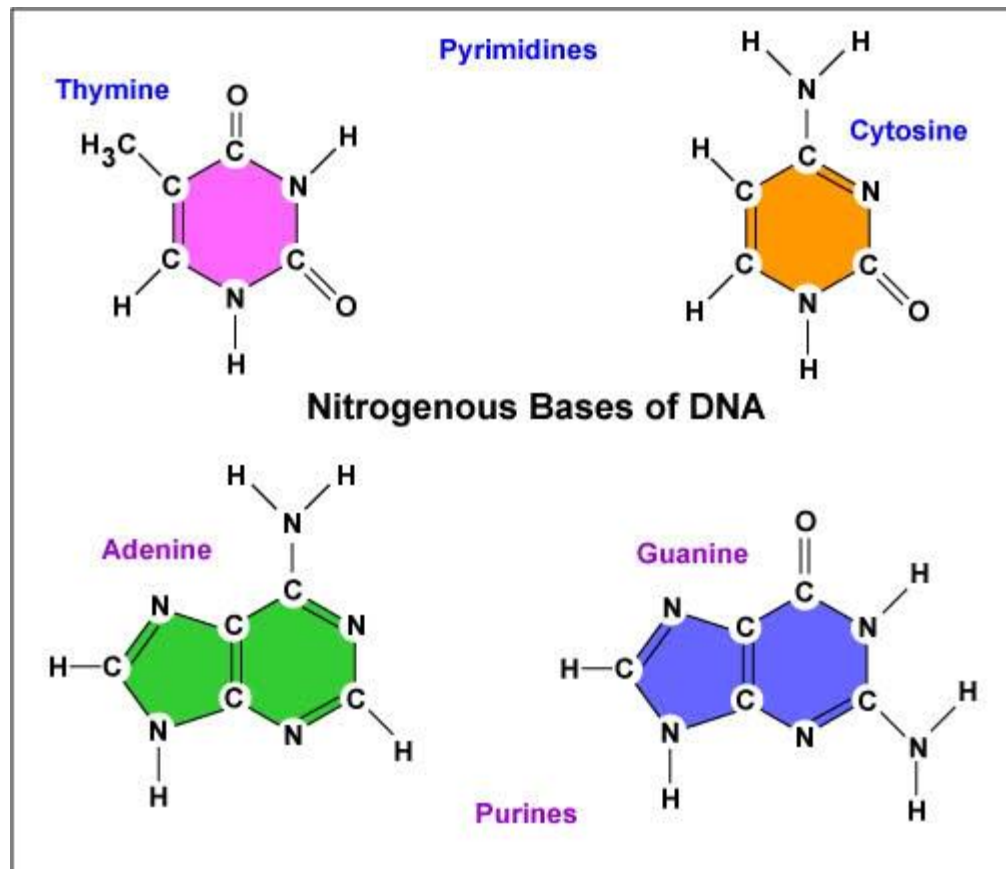Made up of

Nucleotides

Phosphate group

Base

Pentose (a 5-carbon sugar)

Adenine (A)
Cytosine (C)
Guanine (G)
Thymine (T)

- The name of a nucleotide is taken from its base. Each DNA has four kinds of bases, that are adenine, guanine, cytosine and thymine, which are usually abbreviated by A, G, C and T.

Nitrogenous Bases of DNA

Structure of DNA (cont.)

- Two single strands of DNAs can be linked together with the hydrogen bonds between their bases and hence form a helical shape called double stranded DNA (dsDNA).
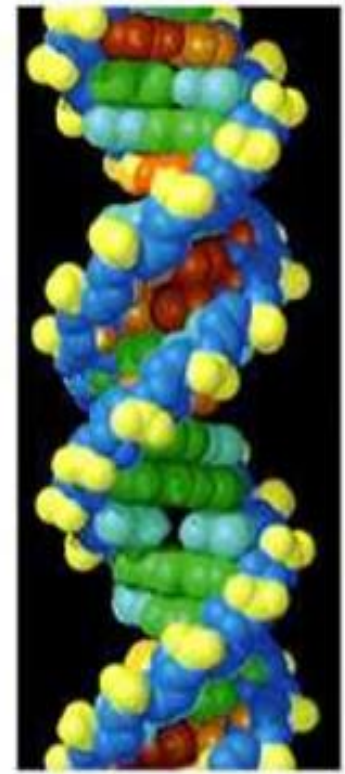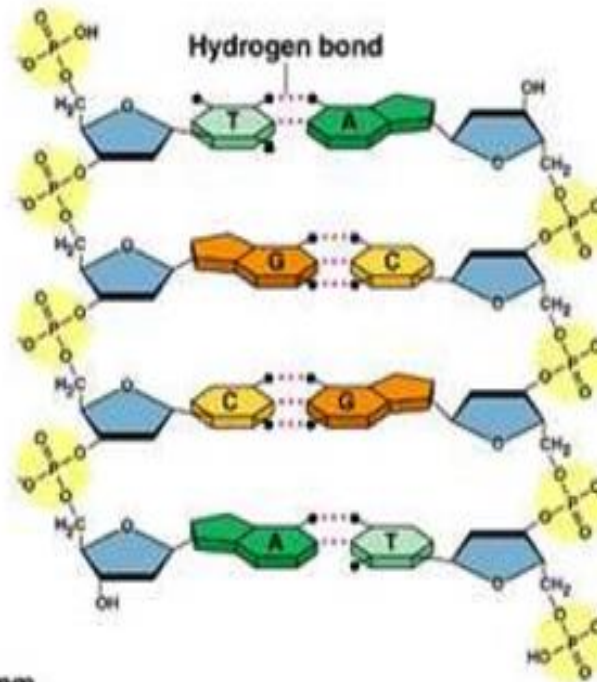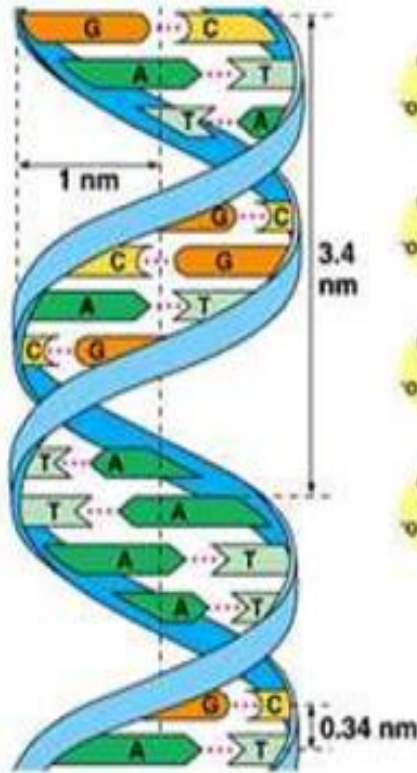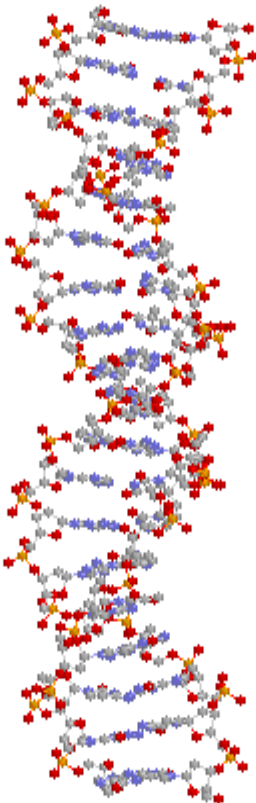
Structure of DNA (cont.)

- In 1953, it was shown that the bases can join only complimentarily, A with T and G with C respectively.
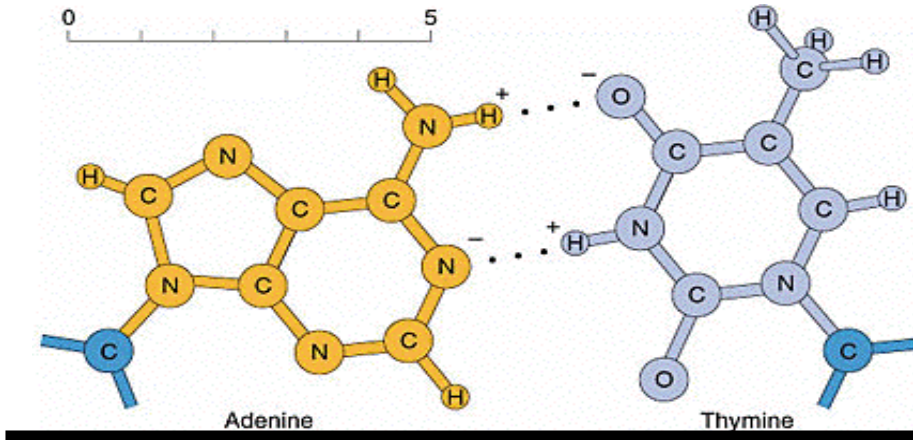
A – T
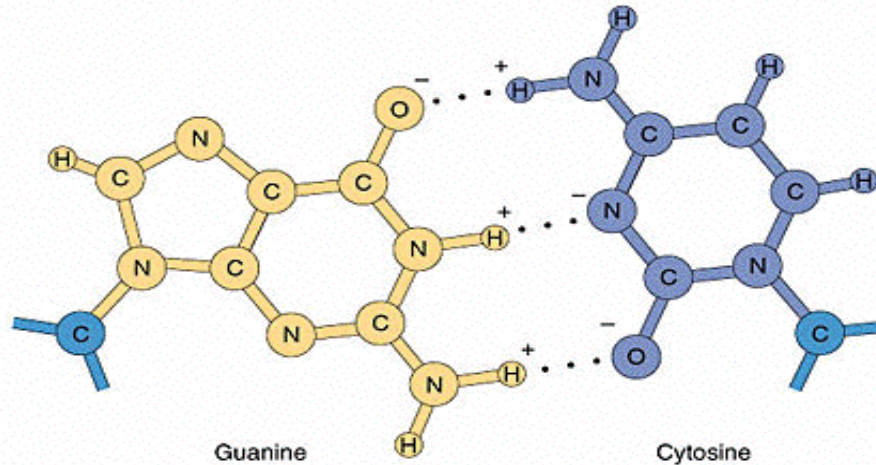
G – C

Structure of DNA (cont.)
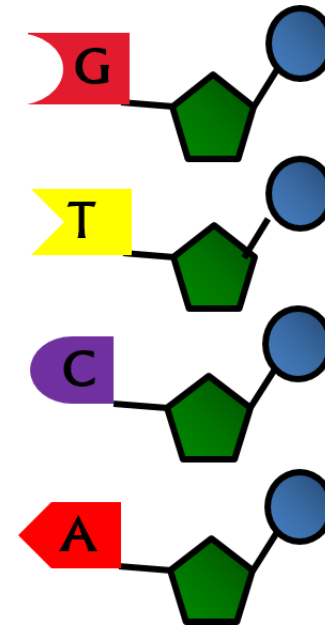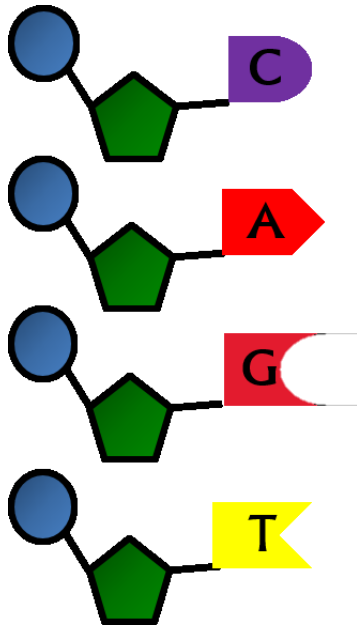
Restriction Enzyme

- DNA molecules can be cut by restriction enzymes at specific places based on the cutting sites of the restriction enzymes.

Sticky end (e.g. *Aci*I)

$$5'...C{\blacktriangledown}CGC...3'$$

$$3'...GGC{\blacktriangle}G...5'$$

Blunt end (e.g. *Afe*I)

$$5'...AGC{\blacktriangledown}GCT...3'$$

$$3'...TCG{\blacktriangle}CGA...5'$$

Restriction Enzyme (cont.)

- Every restriction enzyme has a triple known as the cleavage pattern of the enzyme.

- The triple is denoted as a rule for the restriction enzyme which consists of left context, crossing and right context [3].

- The restriction enzyme *EcoR*I is isolated from the bacterium Escherichia coli with strain serotype *R*; I indicates the first enzyme discovered from the bacterium.

- The cleavage pattern of restriction enzyme *EcoR*I is

$$(g, \, aatt, \, c)$$

left context   **crossing**   right context

3. T. Head, "Splicing Representations of Strictly Locally Testable Languages," *Discrete. Appl. Math.*, vol. 87, no. 1, pp. 139-147, 1998.

Restriction Enzyme (cont.)

- A restriction enzyme is an enzyme that cuts double-stranded or single stranded DNA at specific recognized nucleotide sequences, known as restriction sites.
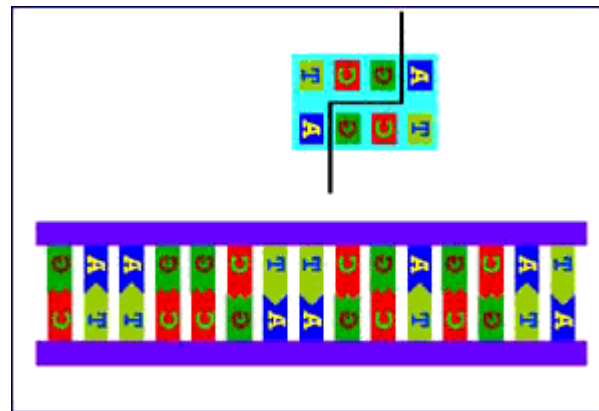
EcoRI: ([G/C],[A/T][A/T][T/A][T/A],[C/G])
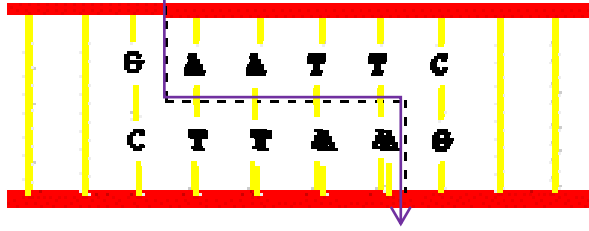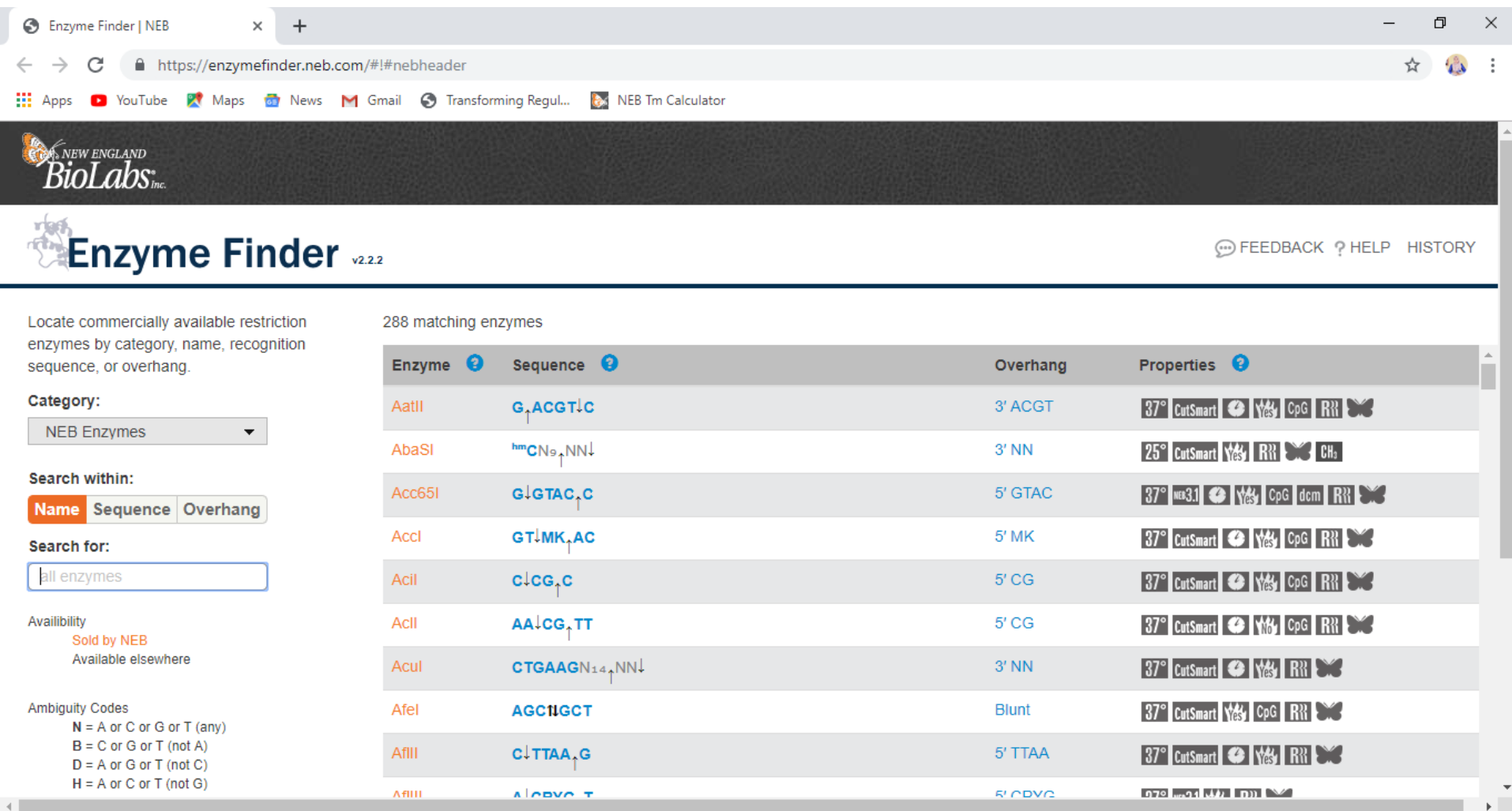
MfeI: ([C/G],[A/T][A/T][T/A][T/A],[G/C])
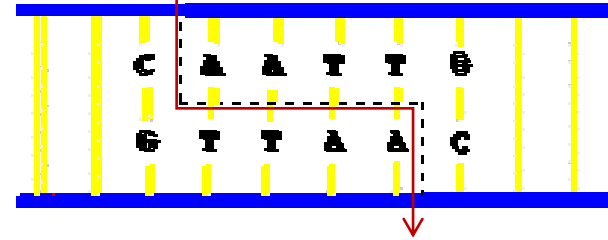
Restriction Enzyme (cont.)

Link: https://enzymefinder.neb.com/#!#nebheader

innovative • entrepreneurial • global

www.utm.my

Definition 1 [4] Palindromic String

A string *l* of a dsDNA is said to be palindromic if the sequence from the left to the right side of the upper single strand is equal to the sequence from the right to the left side of the lower single strand.

Example of a Palindrome:

$$\textbf{H A I N I A H}$$

forwards ⟶ ⟵ backwards

4. Y. Yusof, *DNA Splicing System Inspired by Bio Molecular Operations*. Ph.D. Thesis, Universiti Teknologi Malaysia, 2012.

innovative ● entrepreneurial ● global

www.utm.my

Palindromic and Non-Palindromic Restriction Enzymes

The enzyme *EcoR*I $\begin{array}{c} 5' - \text{GAATTC} - 3' \\ 3' - \text{CTTAAG} - 5' \end{array}$, is a palindromic restriction enzyme:

forwards $\longrightarrow$

$$5' - \textbf{G A A T T C} - 3'$$

$$| \quad | \quad | \quad | \quad | \quad |$$

$$3' - \textbf{C T T A A G} - 5' \quad \longleftarrow \quad \text{backwards}$$

The enzyme *Aci*I $\begin{array}{c} 5' - \text{CCGC} - 3' \\ 3' - \text{GGCG} - 5' \end{array}$ is a non-palindromic restriction enzyme:

forwards $\longrightarrow$

$$5' - \textbf{C C G C} - 3'$$

$$| \quad | \quad | \quad |$$

$$3' - \textbf{G G C G} - 5' \quad \longleftarrow \quad \text{backwards}$$
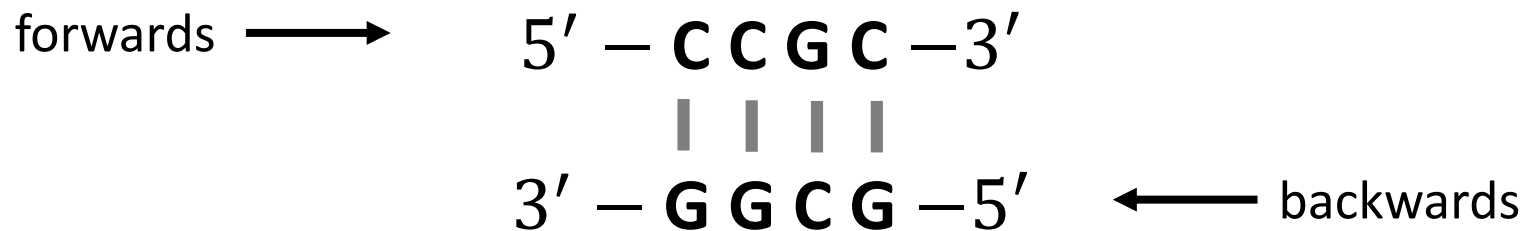
Splicing System

- **Formal Language Theory**
- Applied Discrete Mathematics
- Theoretical Computer Science

Formal Language Theory (cont.)

- Formal language theory is a study on a set of finite strings of symbols or language over an alphabet in which some formation rules are applied in DNA splicing system [5]:-

Symbols	Explanation
$A*$	A set of strings of symbols from an alphabet A
A^+	A set of strings of symbols from an alphabet A without the empty string
λ	Empty string
+	Union
·	Concatenation
*	Star-closure
{} or ()	Parentheses

5. P. Linz, *An Introduction to Formal Languages and Automata*, 4th ed. USA: Jones and Bartlett Publisher, 2006.

innovative ● entrepreneurial ● global

www.utm.my

Modelling of DNA Splicing System
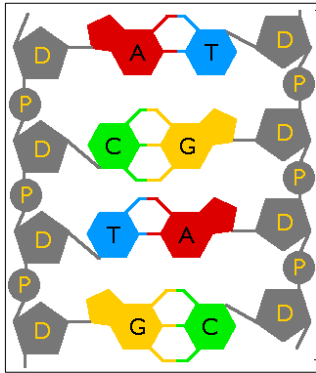
Nitrogenous base pairings
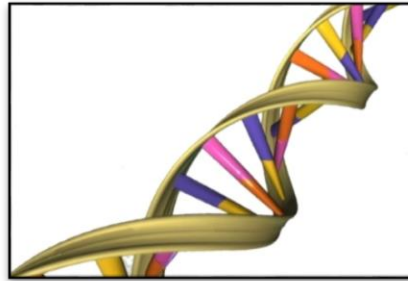
Nucleotide chains

Restriction enzymes

dsDNA

DNA strings

rules

Modelling of DNA Splicing System (cont.)

Molecule A

Molecule B

Restriction enzyme cleavage point

Recombinant DNA
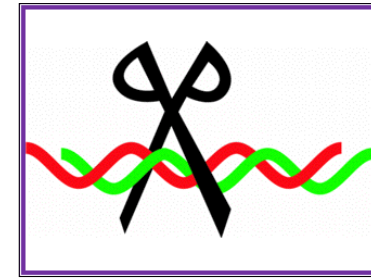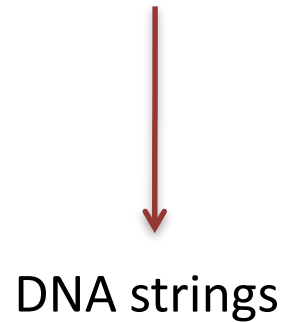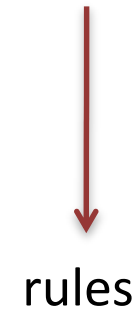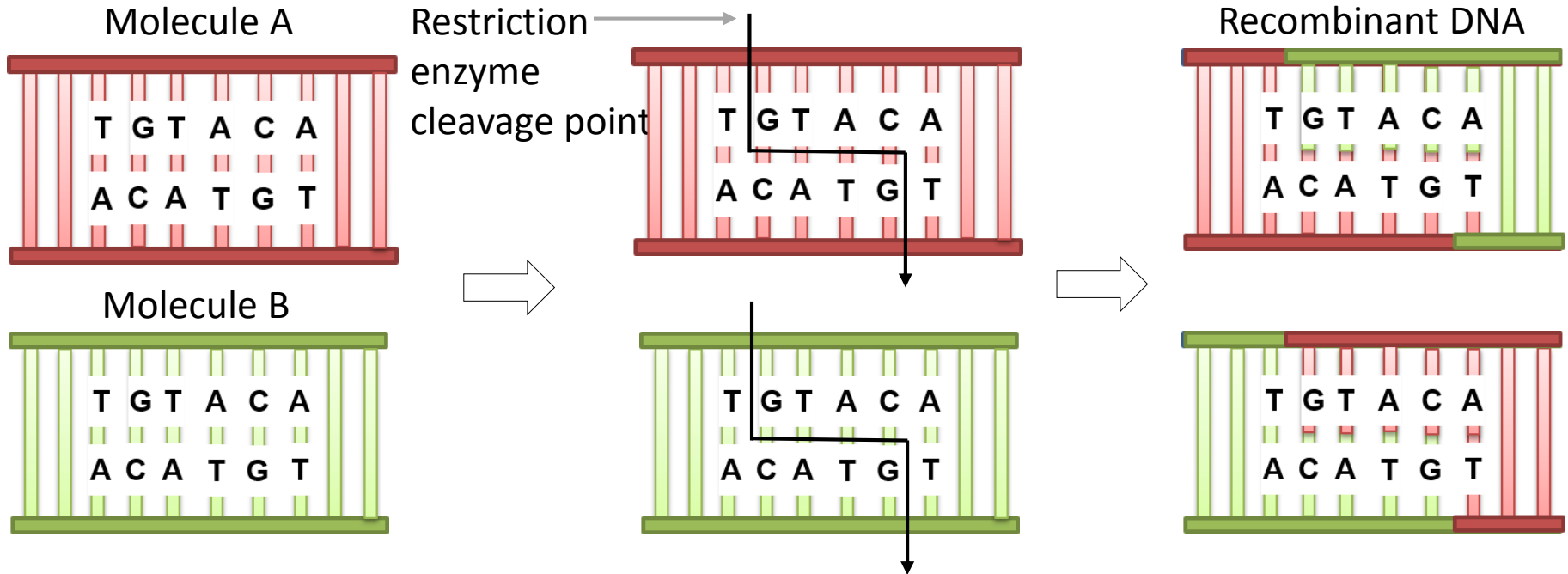
G: Guanine   A: Adenine   C: Cytosine   T: Thymine

Definition 2 [6] Splicing System and Splicing Language

A splicing system, *S = (A, I, B, C)* consists of

- *A*: finite alphabet

- *I*: a finite set of initial strings in *A**

- Patterns *B* and *C*: finite sets of triples *(c, x, d)* with *c , x* and *d* in *A**

For each such triple the string *cxd* is called a site and the string *x* is called a crossing.

A language, *L* is a splicing language if there exists a splicing system *S* for which *L = L(S)*.

6. T. Head, "Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors," *B. Math. Biol*, vol. 49, no. 6, pp. 737-759, 1987.

innovative ● entrepreneurial ● global

www.utm.my

Example 1

Suppose that $S = (A, I, B, C)$ is a splicing system in which $A = \left\{ \begin{matrix} A, & C, & G, & T \\ T & G & C & A \end{matrix} \right\}$ is the set of dsDNA symbols, $I = \left\{ \begin{matrix} GAATTCTCTGTAAT \\ CTTAAGAGACATTA \end{matrix} \right\}$ is the set consisting of an initial string of molecules, set $B = \left\{ \left(\begin{matrix} G & AATT & C \\ C, & TTAA, & G \end{matrix} \right) \right\}$ is the set of cleavage pattern for the enzyme **EcoR**I and set C is the empty set.

The initial string is shown in the following:

$$5'-GAATTCTCTGTAAT-3'$$
$$3'-CTTAAGAGACATTA-5''$$
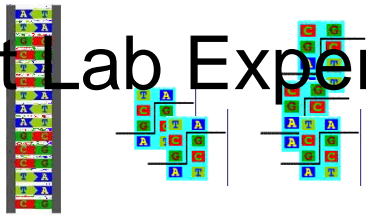
or written 180 degree wise,

$$5'-ATTACAGAGAATTC-3'$$
$$3'-TAATGTCTCTTAAG-5'$$

DNA Splicing System (cont.)

5′– GAATT CTCTGTAAT – 3′
3′– C TTAA GAGACATTA – 5′

rotate 180°

5′– ATTACAGAG AATT C – 3′
3′–TAATGTCTC TTAA G – 5′
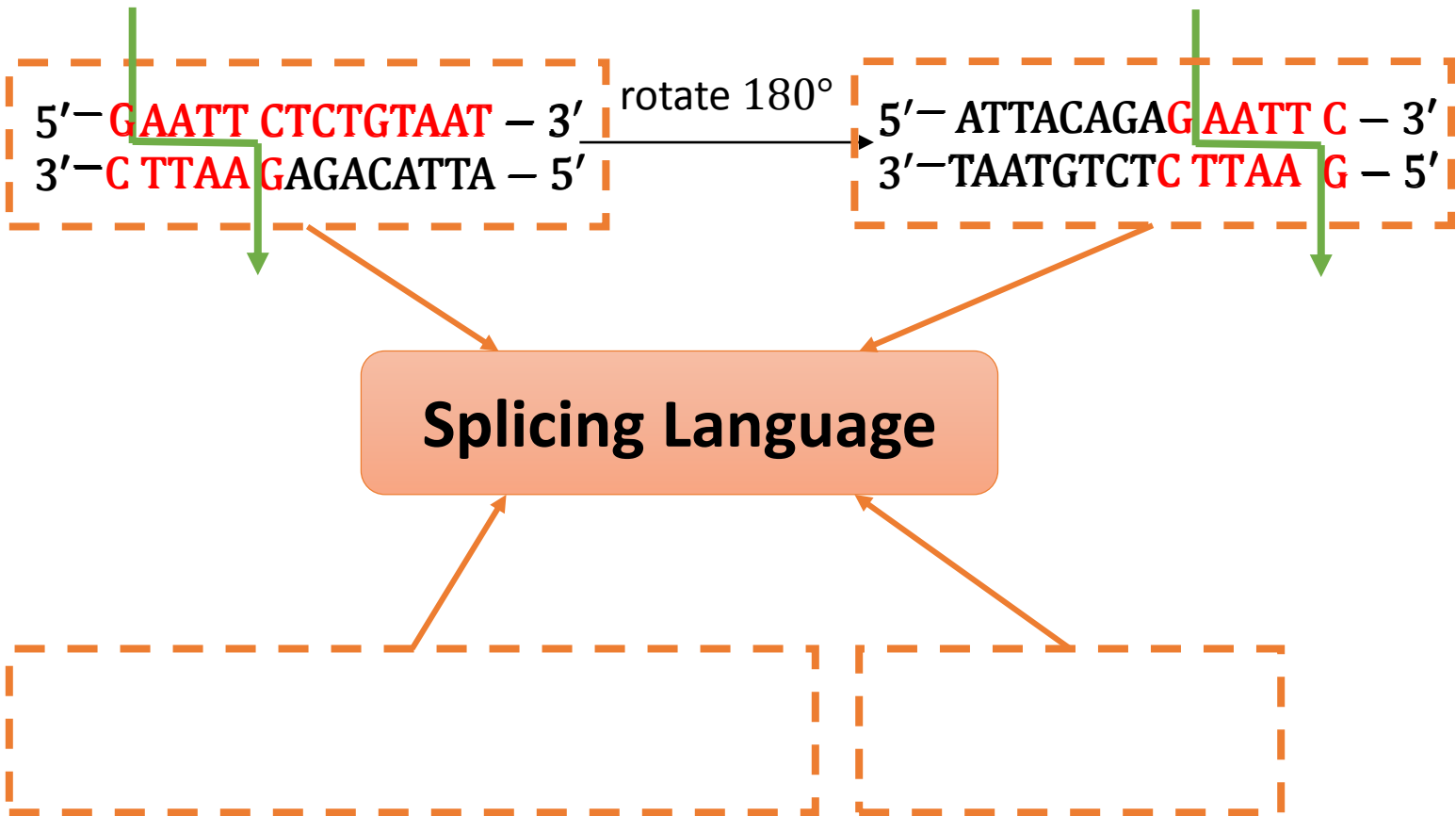
Splicing Language

Wet Lab Experiment

$$S = (A, I, B, C)$$
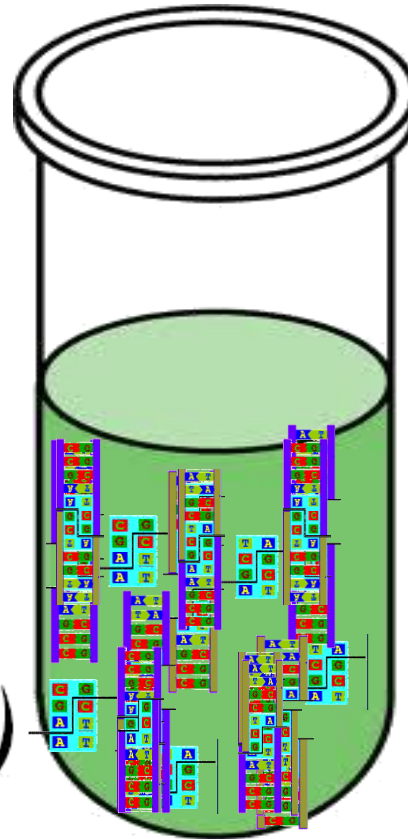
Historical Background of DNA Splicing Systems
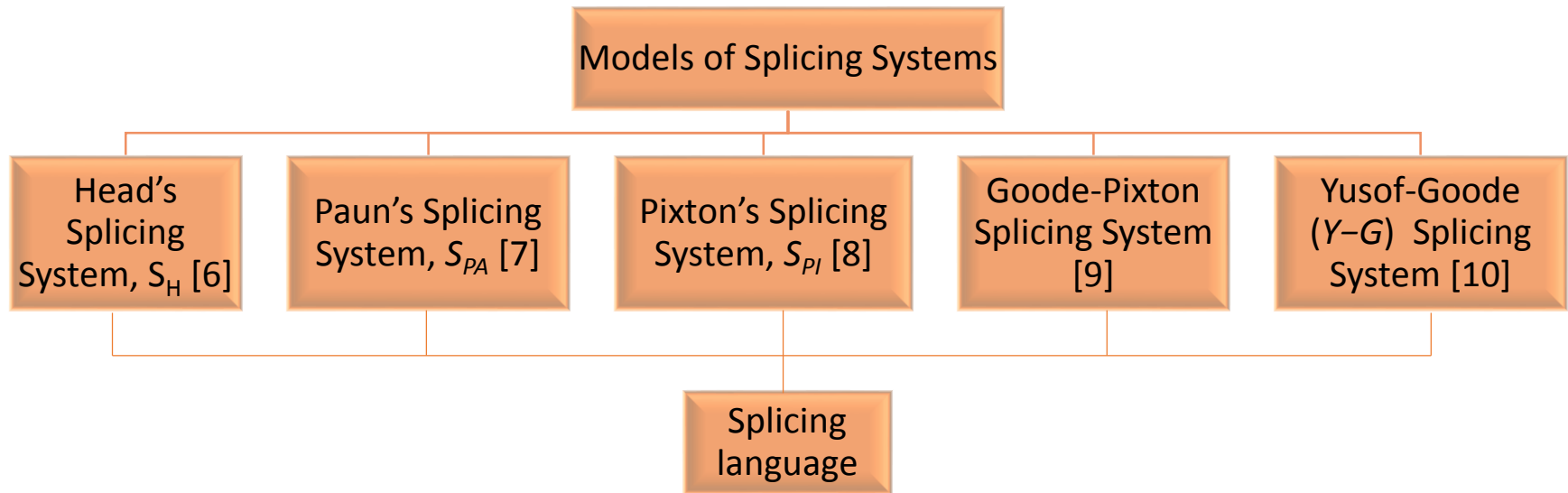
```
                    ┌─────────────────────────┐
                    │ Models of Splicing Systems │
                    └─────────────────────────┘
```

| Head's Splicing System, S_H [6] | Paun's Splicing System, S_{PA} [7] | Pixton's Splicing System, S_{PI} [8] | Goode-Pixton Splicing System [9] | Yusof-Goode ($Y-G$) Splicing System [10] |

Splicing language

6. T. Head, "Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors," *B. Math. Biol*, vol. 49, no. 6, pp. 737-759, 1987.
7. G. Păun, "On the Splicing Operation," *Discrete. Appl. Math.*, vol. 70, no. 1, pp. 57-79, 1996.
8. D. Pixton, "Regularity of Splicing Languages," *Discrete. Appl. Math.*, vol. 69, no. 1-2, pp. 101-124, 1996.
9. E. Goode, D. Pixton, Splicing to the Limit, In: N. Jonoska, G. Păun, G. Rozenberg ed, *Aspects of Molecular Computing, Lecture Notes in Computer Science*. Germany: Springer-Verlag, pp. 189-201, 2004.
10. Y. Yusof, N. H. Sarmin, W. H. Fong, T. E. Goode, M. A. Ahmad, "An Analysis of Four Variants of Splicing System," *Proceedings of the 20th National Symposium on Mathematical Sciences - Research in Mathematical Sciences: A Catalyst for Creativity and Innovation (SKSM 2012)*, Melville, NY, 2013, pp. 888-895.

innovative ● entrepreneurial ● global

www.utm.my

Previous Molecular Works on Splicing Systems

Author	Description
Laun and Reddy [11] 1999	The first experiment on the splicing system using restriction enzymes *Bgl*I and *Dra*III
Fong [12] 2008	The adult and limit languages from Head's splicing model using restriction enzymes *Hpa*II and *Aci*I
Karimi [13] 2013	Verification of the persistency properties of splicing systems involving restriction enzymes *CvaQ*I and *Acc65*I
Yusof et al. [14] 2015	Yusof-Goode splicing system with restriction enzymes *Acl*I and *Aci*I using limit graph approach
Ahmad et al.[15] 2018	Experiment on second order limit language from Yusof-Goode splicing system using restriction enzyme *Dpn*II

12. E. Laun, K. J. Reddy, "Wet Splicing Systems," *Proceedings of the 3rd DIMACS Workshop on DNA Based Computers*, Rhode Island, USA, 1999, pp. 73-84.
13. W. H. Fong, *Modelling of Splicing Systems using Formal Language Theory*. Ph.D. Thesis, Universiti Teknologi Malaysia, 2008.
14. F. Karimi, *Mathematical Modelling of Persistent Splicing Systems in DNA Computing*. Ph.D. Thesis, Universiti Teknologi Malaysia, 2013.
15. Y. Yusof, W. L. Lim, T. E. Goode, N. H. Sarmin, F. W. Heng, M. F. A. Wahab, "Molecular Aspects of DNA Splicing System," *Proceedings of the AIP Conference Proceedings*, 2015, pp. 050045 1-8.
16. M. A. Ahmad, N. H. Sarmin, M. F. Abdul-Wahab, F. W. Heng, Y. Yusof, "Biomolecular Aspects of Second Order Limit Language," vol. 14, no. 1, pp. 15-19, 2018.

innovative ● entrepreneurial ● global

www.utm.my

Visit to SUNY Binghamton, USA, 2007

innovative ● entrepreneurial ● global

www.utm.my

Visit to wet lab, SUNY Binghamton, USA, 2007

Wet lab experiment, UTM, 2007

innovative ● entrepreneurial ● global

Wet lab experiment (Assoc Prof Dr Yuhani Yusof), Towson University, USA, 2010

innovative ● entrepreneurial ● global

Research collaboration with Towson University, USA, 2010
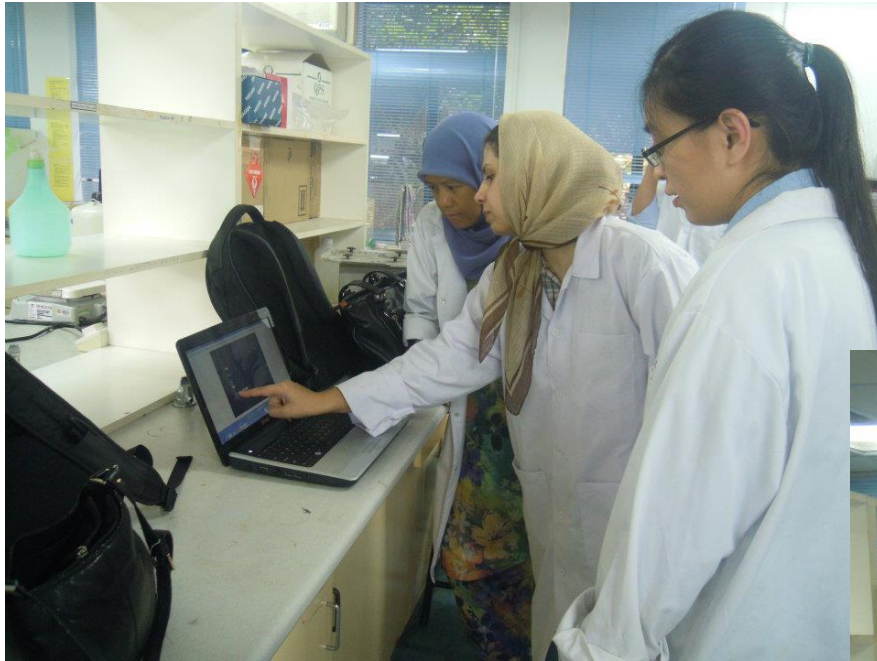
Wet lab experiment (Dr Fariba Karimi), UTM, 2012

**Wet lab experiment
(Dr Muhammad Azrin Ahmad),
UTM, 2015**

Wet lab experiment (Nurul Izzaty Ismail), UTM, 2020

Laboratory Procedure

Polymerase chain reaction (PCR)

- To make several copies of a specific DNA segment

Process of Restriction Enzyme Digestion and Ligation

- The restriction enzymes recognize specific restriction sites in DNA molecules
- The restriction enzymes and ligase then cut and rejoin the molecules to generate further molecules

Polyacrylamide gel electrophoresis (PAGE)

- to separate proteins based on their molecular weight

Polymerase Chain Reaction (PCR)

PCR machine

PCR tube

Lambda DNA

Primers

DNA Polymerase
(PCR reagent)

Process of Restriction Enzyme Digestion and Ligation
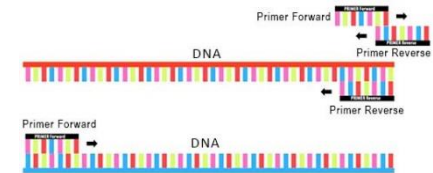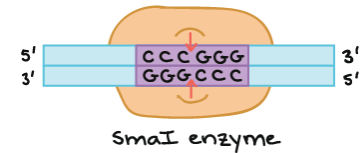

Thermomixer

Microcentrifuge Tube
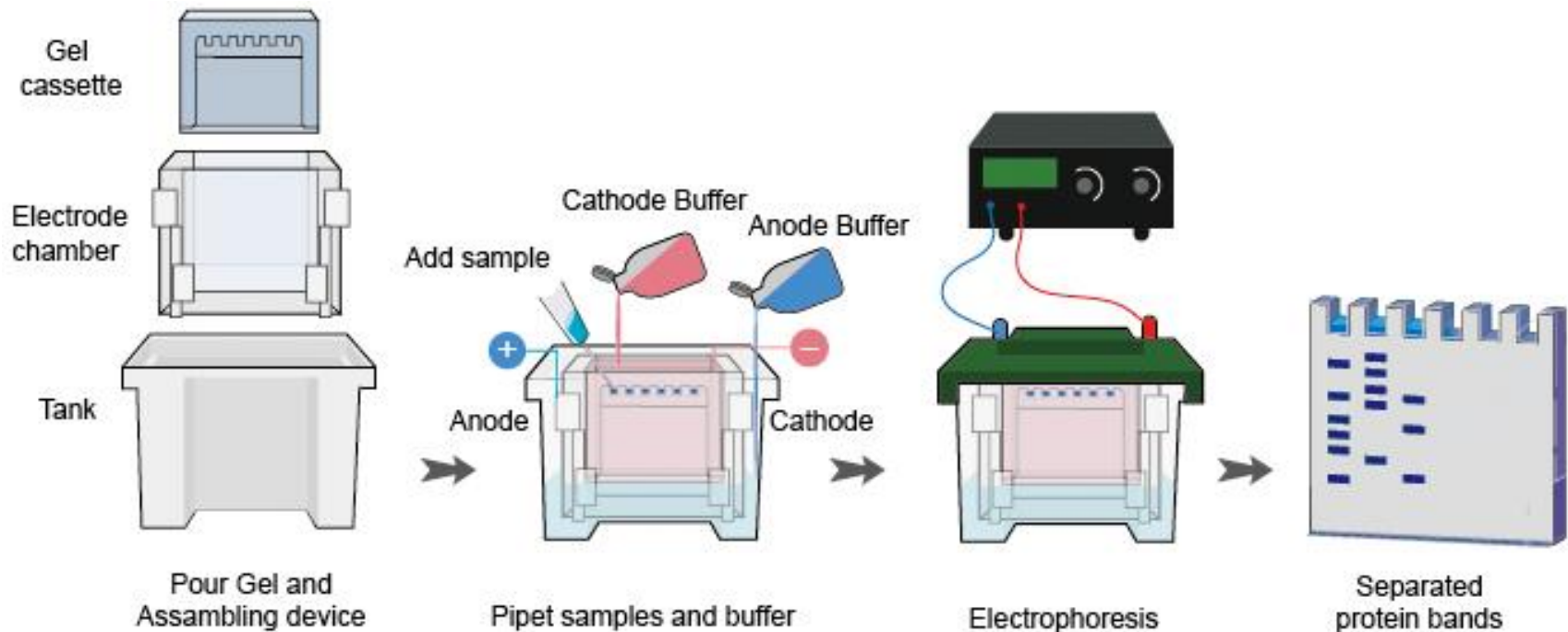
PCR Product

SmaI enzyme

Restriction Enzyme

Ligase and Buffer

Polyacrylamide gel electrophoresis (PAGE)

innovative • entrepreneurial • global

Example of a wet lab experiment
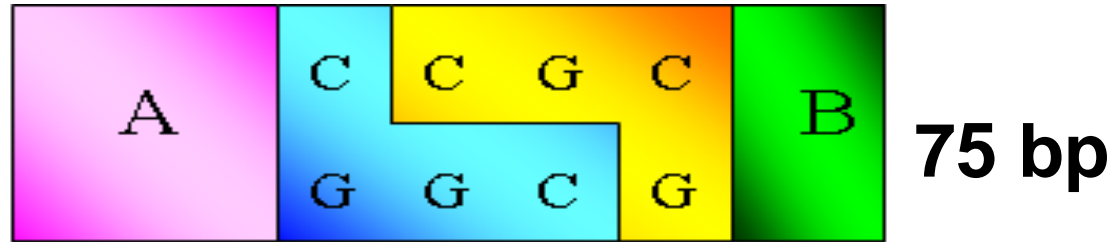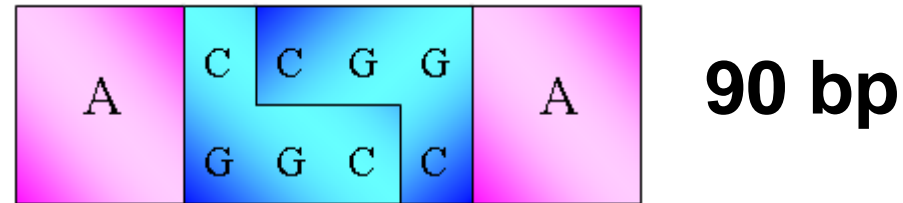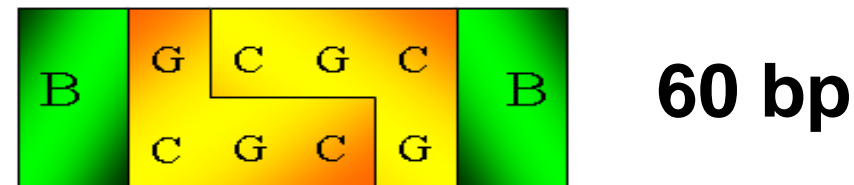
75 bp

Limit Language   90 bp

Adult Language   60 bp

Example of a gel photo
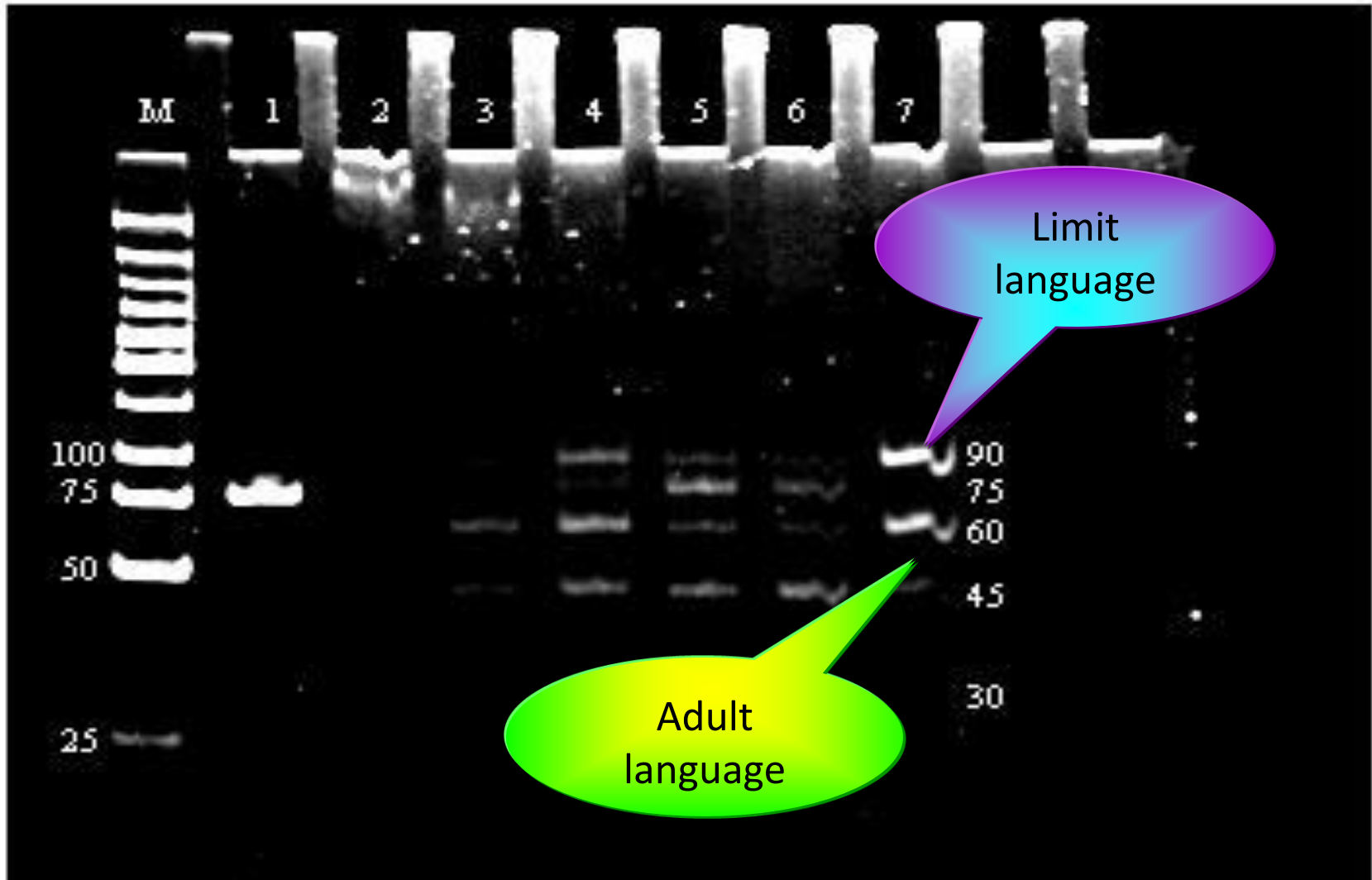
Case 1: Experimental Design for a DNA Splicing System

- An initial DNA molecule *I* used in this splicing model is a small segment taken from bacteriophage lambda between 42958 and 43117 with the length of 160 base pairs (bp).

- The initial molecule contains one cutting site each of the restriction enzymes *Cvi*QI and *Aci*I where the genome locations for the cutting sites are found at 42992-42995 and 43036-43039 respectively.

- Five sticky ends of molecules α, β, γ, $\alpha - \beta$ and $\beta - \gamma$ are produced by the restriction enzymes when cutting the initial molecule. The lengths of fragments for the sticky ends are given in the following.

$$\text{Fragment:} \quad \alpha \xrightarrow{\textit{Cvi}\text{QI site}} \beta \xrightarrow{\textit{Aci}\text{I site}} \gamma$$

$$|\alpha| = 35 \text{ bp}$$

$$|\beta| = 44 \text{ bp}$$

$$|\gamma| = 81 \text{ bp}$$

$$|\alpha - \beta| = 79 \text{ bp}$$

$$|\beta - \gamma| = 125 \text{ bp}$$

- Initial Molecule (42958 and 43117 from bacteriophage lambda)

ggactatcgaagagtgcaaggcgatcaaggcaga<u>gtac</u>caacagaaactcaaagacctgcgaaatagcaga agtgagg<u>ccgc</u>atgacgttctcagtaaaaaccattccagacatgctcgttgaaacatacggaaatcagacagaa gtagcacgcagactg (160 bp)

- Enzyme *Cvi*QI (palindromic)

$$5'...G^{\blacktriangledown}TAC...3'$$

$$3'...CAT_{\blacktriangle}G...5'$$

- Enzyme *Aci*I (non-palindromic)

$$5'...C^{\blacktriangledown}CGC...3'$$

$$3'...GGC_{\blacktriangle}G...5'$$

Corollary 1

Let $S = (A, I, B, C)$ be a DNA splicing system in which $A = \{\alpha, x_1, y, x_2, \beta, w_1, z, w_2, \gamma\}$ is the set of variables used to denote any arbitrary dsDNA, $I = \{\alpha x_1 y x_2 \beta w_1 z w_2 \gamma\}$ is the set consisting of an initial string with one cutting site each of palindromic and non-palindromic rules $x_1 y x_2$ and $w_1 z w_2$ where α, x_1, y, x_2, β, w_1, z, w_2 and γ can be rotated 180°, represented as α', x_1', y', x_2', β, w_1', z', w_2' and γ' respectively, set $B = \{(x_1, y, x_2), (w_1, z, w_2)\}$ is the set of cleavage pattern for the rules where y and z are the crossings and set C is the empty set, then the resulting splicing language consists of strings of the form

$$\{\alpha x_1 y x_2 (\beta w_1 z w_1' \beta' x_1 y x_2)^{n-1} (\alpha' + \beta w_1 z w_2 \gamma)\}$$
$$+\{\gamma' w_2' z (w_1' \beta' x_1 y x_2 \beta w_1 z)^{n-1} (w_2 \gamma + w_1' \beta' x_1 y x_2 \alpha')\}$$

where $n \in \mathbb{Z}^+$ and $\{x_1 y x_2, w_1 z w_2, w_2' z w_1'\} \notin \{\alpha, \beta, \gamma\}$.

From the generalisation of splicing languages in Corollary 1, the splicing language from this splicing system S_1 invoving one cutting site each of palindromic restriction enzyme *Cvi*QI (*g*, *ta*, *c*) and non-palindromic restriction enzyme *Aci*I (*c*, *cg*, *c*) with different palindromic crossings is shown in the following:

$$L(S_1) = \{\alpha gtac(\beta ccgg\beta' gtac)^{n-1}(\alpha' + \beta ccgc\gamma)\}$$
$$+\{\gamma' gcg(g\beta' gtac\beta ccg)^{n-1}(c\gamma + g\beta' gtac\alpha')\}.$$

where $n \in \mathbb{Z}^+$. The fragments of DNA strings in the splicing language $L(S_1)$ are stated as follow:

$$\alpha \underline{gtac} \left(\beta \underline{ccgg} \beta' \underline{gtac}\right)^{n-1} \alpha',$$

$$\alpha \underline{gtac} \left(\beta \underline{ccgg} \beta' \underline{gtac}\right)^{n-1} \beta \underline{ccgc} \gamma,$$

$$\gamma' \underline{gcg} \left(\underline{g} \beta' \underline{gtac} \beta \underline{ccg}\right)^{n-1} \underline{c} \gamma \text{ and}$$

$$\gamma' \underline{gcg} \left(\underline{g} \beta' \underline{gtac} \beta \underline{ccg}\right)^{n-1} \underline{g} \beta' \underline{gtac} \alpha'$$
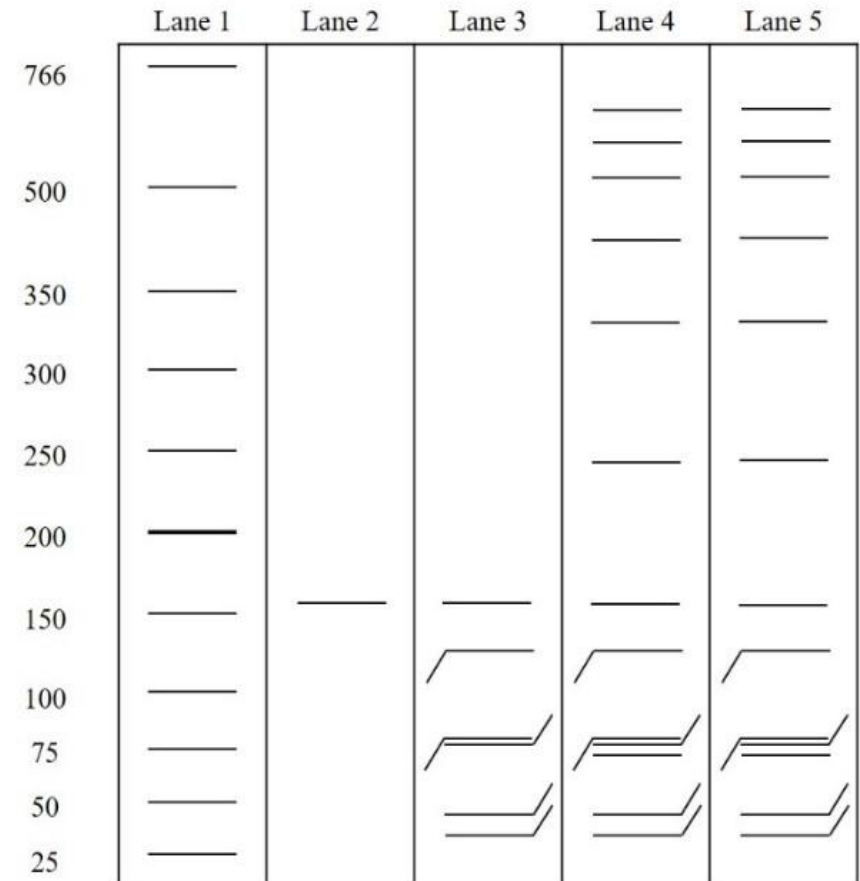
where $n \in \mathbb{Z}^+$ represents multiple copies of the specific strings.

No.	Molecule	Size (bp)
1.	α	35
2.	β	44
3.	γ	81
4.	α'	37
5.	β'	44
6.	γ'	79
7.	$\alpha - \beta$	79
8.	$\beta - \gamma$	125
9.	$\alpha - \beta - \gamma$	160
10.	$\alpha - (\beta - \beta' -)^{n-1} - \alpha'$	72, 160, 248, 336, $\cdots$
11.	$\alpha - (\beta - \beta' -)^{n-1} - \beta - \gamma$	160, 248, 336, 424, $\cdots$
12.	$\gamma' - (-\beta' - \beta -)^{n-1} - \gamma$	160, 248, 336, 424, $\cdots$
13.	$\gamma' - (-\beta' - \beta -)^{n-1} - \beta' - \alpha'$	160, 248, 336, 424, $\cdots$

innovative ● entrepreneurial ● global
www.utm.my

DNA Splicing System involving *Cvi*QI and *Aci*I (cont.)
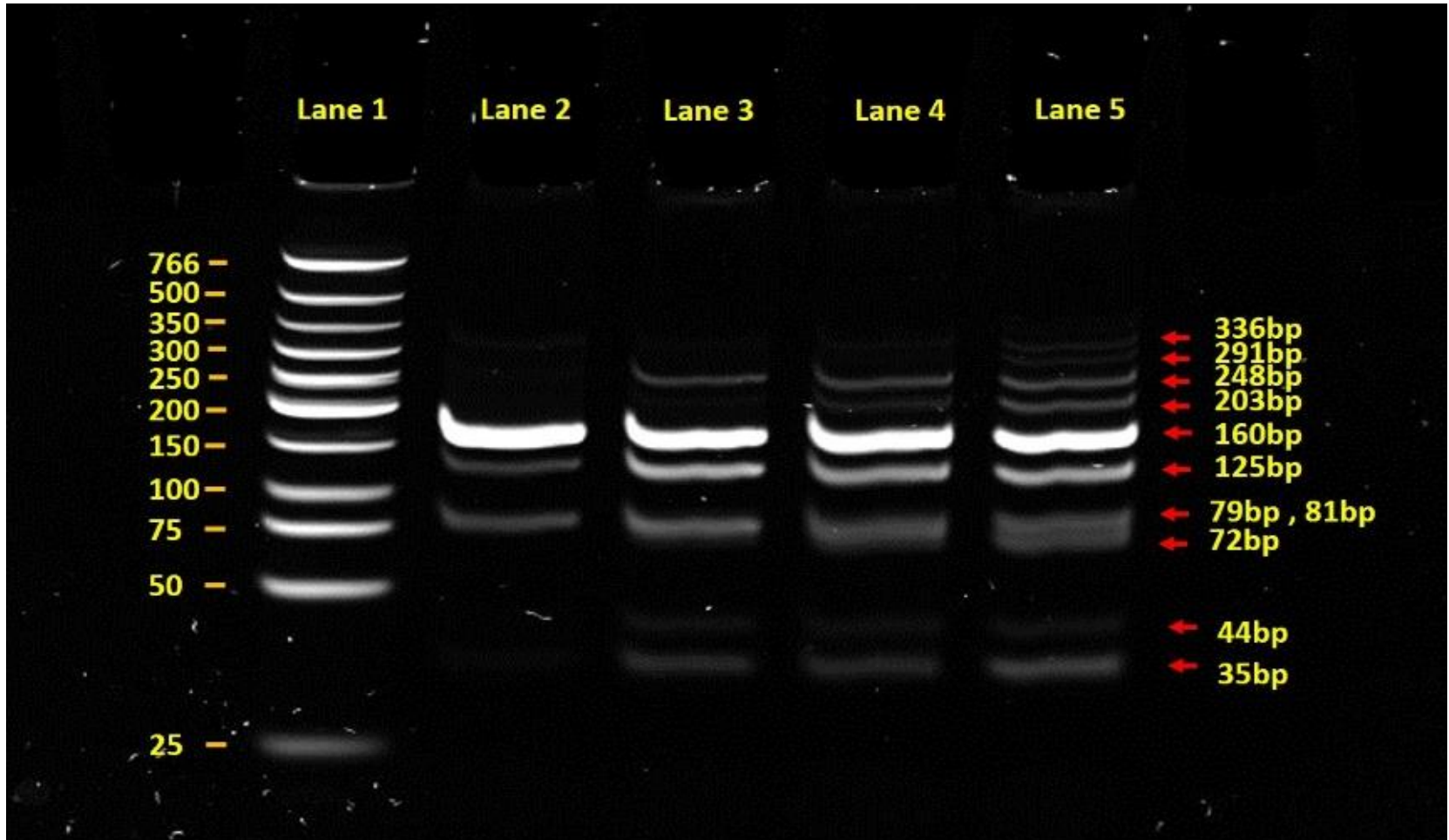
Case 2: Experimental Design for DNA Splicing System involving Enzyme *Cvi*QI

- Initial Molecule (42958 and 43117 from bacteriophage lambda)

ggactatcgaagagtgcaaggcgatcaaggcagagtaccaacagaaactcaaagacctgcgaaatagcaga agtgaggccgcatgacgttctcagtaaaaaccattccagacatgctcgttgaaacatacggaaatcagacagaa gtagcacgcagactg (160 bp)

- Enzyme *Cvi*QI (palindromic)

$$5'...G^{\blacktriangledown}TAC...3'$$

$$3'...CAT_{\blacktriangle}G...5'$$

Corollary 2

Let S = (A, I, B, C) be a DNA splicing system in which A= $\{\alpha, x_1, y, x_2, \delta\}$ is the set of variables used to denote any arbitrary dsDNA, I = $\{\alpha x_1 y x_2 \delta\}$ is the set consisting of an initial string with one cutting site of a palindromic rule $x_1 y x_2$ where α, x_1, y, x_2 and δ can be rotated 180°, represented as α', x_1', y', x_2' and δ' respectively, set B = $\{(x_1, y, x_2)\}$ is the set of cleavage pattern for the rule where y is the crossing and set C is the empty set, then the resulting splicing language consists of strings of the form

$$(\alpha + \delta') x_1 y x_2 (\delta + \alpha')$$

where $x_1 y x_2 \notin \{\alpha, \delta\}$.

From the generalisation of splicing languages in Corollary 2, the splicing language from this splicing system S_2 involving one cutting site of a palindromic restriction enzyme *Cvi*QI (g, ta, c) is shown in the following:

$$L(S_2) = (\alpha + \delta')gtac\big((\beta + \beta')gtac\big)^{n-1}(\delta + \alpha')$$

where δ is a sticky end of $\beta - \gamma$. The fragments of DNA strings in the splicing language $L(S_2)$ are stated as follow:

$$\alpha \xrightarrow{gtac} \delta,$$

$$\alpha \xrightarrow{gtac} \alpha',$$

$$\delta' \xrightarrow{gtac} \delta \text{ and}$$

$$\delta' \xrightarrow{gtac} \alpha'.$$

No.	Molecule	Size (bp)
1.	α	35
2.	δ	125
3.	α'	37
4.	δ'	123
5.	$\alpha - \delta$	160
6.	$\alpha - \alpha'$	72
7.	$\delta' - \delta$	248
8.	$\delta' - \alpha'$	160

innovative ● entrepreneurial ● global

- Initial Molecule (42958 and 43117 from bacteriophage lambda)

*ggactatcgaagagtgcaaggcgatcaaggcagagtaccaacagaaactcaaagacctgcgaaatagcaga
agtgagg<u>ccgc</u>atgacgttctcagtaaaaaccattccagacatgctcgttgaaacatacggaaatcagacagaa
gtagcacgcagactg* (160 bp)

- Enzyme *Aci*I

$$5'...C^{\blacktriangledown}CGC...3'$$

$$3'...GGC_{\blacktriangle}G...5'$$

Corollary 3

Let $S = (A, I, B, C)$ be a DNA splicing system in which $A = \{\eta, x_1, y, x_2, \gamma\}$ is the set of variables used to denote any arbitrary dsDNA, $I = \{\eta x_1 y x_2 \gamma\}$ is the set consisting of an initial string with one cutting site of a non-palindromic rule $x_1 y x_2$ where η, x_1, y, x_2 and γ can be rotated 180º, represented as η', x_1', y', x_2' and γ' respectively, set $B = \{(x_1, y, x_2)\}$ is the set of cleavage pattern for the rule where y is the palindromic crossing and set C is the empty set, then the resulting splicing language consists of strings of the form

$$(\eta x_1 + \gamma' x_2') \, y \, (x_2 \gamma + x_1' \eta')$$

where $x_1 y x_2$, $x_2' y x_1' \notin \{\eta, \gamma\}$.

From the generalisation of splicing languages in Corollary 3, the splicing language from this splicing system S_3 involving one cutting site of a non-palindromic restriction enzyme *Aci*l (c, cg, c) is shown in the following:
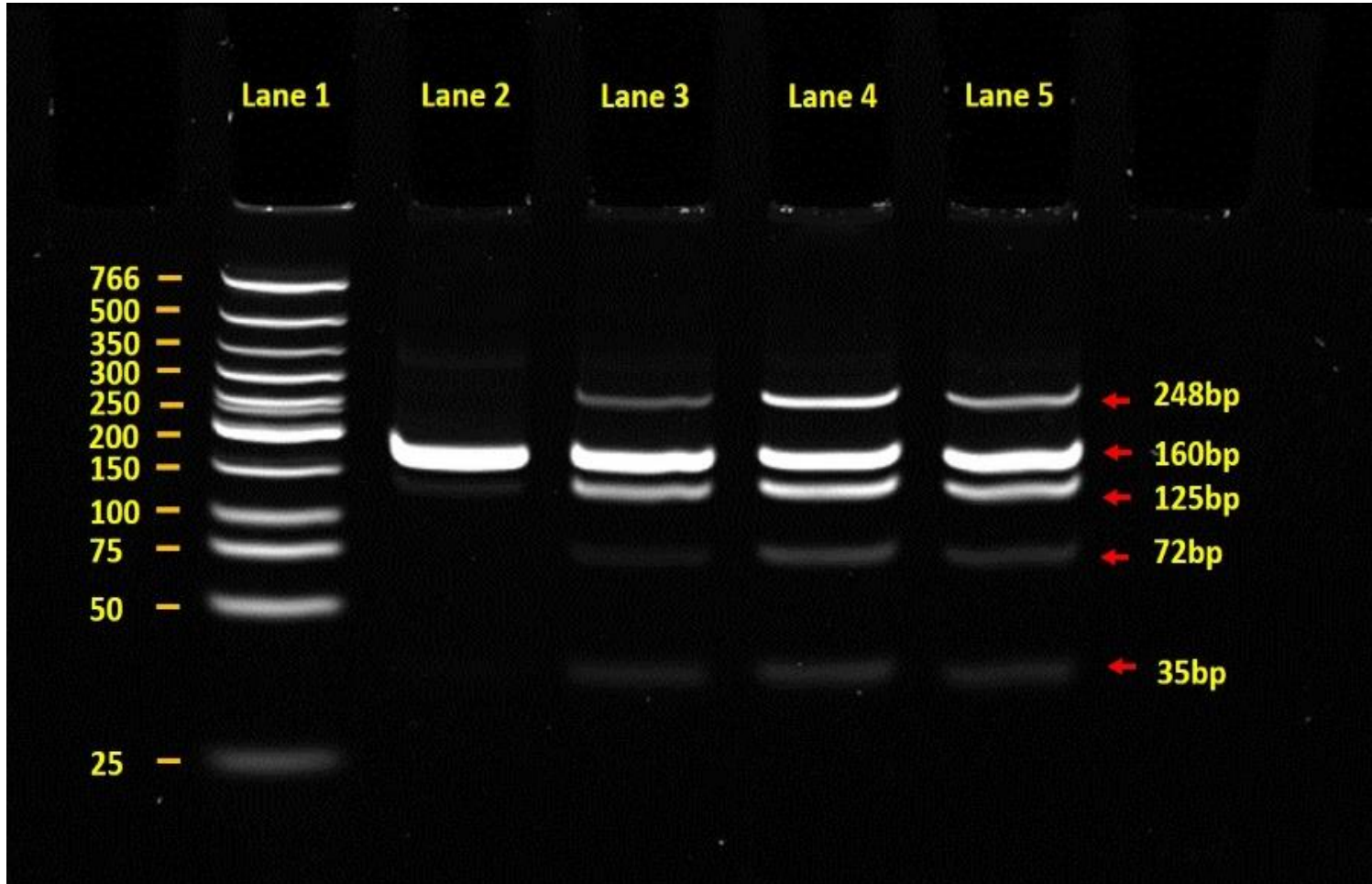
$$L(S_3) = (\eta c + \gamma' g)cg(c\gamma + g\eta')$$

where η is a sticky end of $\alpha - \beta$. The fragments of DNA strings in the splicing language $L(S_3)$ are stated as follow:

$$\eta \frac{ccgc}{} \gamma,$$

$$\eta \frac{ccgg}{} \eta',$$

$$\gamma' \frac{gcgc}{} \gamma \text{ and}$$

$$\gamma' \frac{gcgg}{} \eta'.$$

No.	Molecule	Size (bp)
1.	η	79
2.	γ	81
3.	η'	81
4.	γ'	79
5.	$\eta - \gamma$	160
6.	$\eta - \eta'$	160
7.	$\gamma' - \gamma$	160
8.	$\gamma' - \eta'$	160
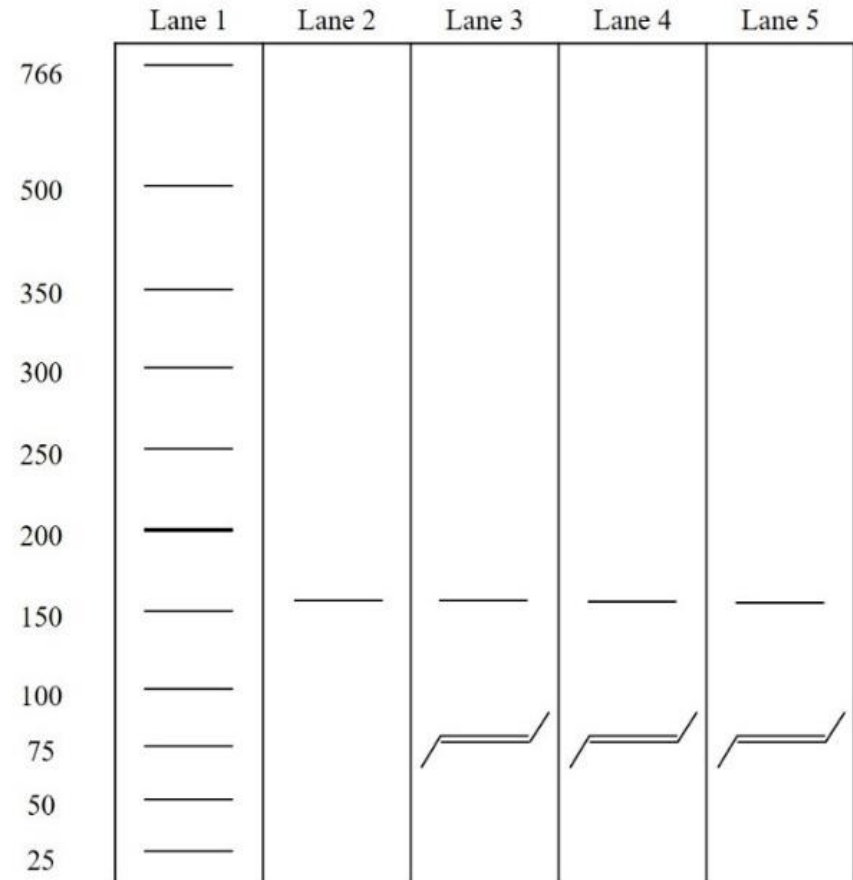
DNA Splicing System involving *Aci*I (cont.)
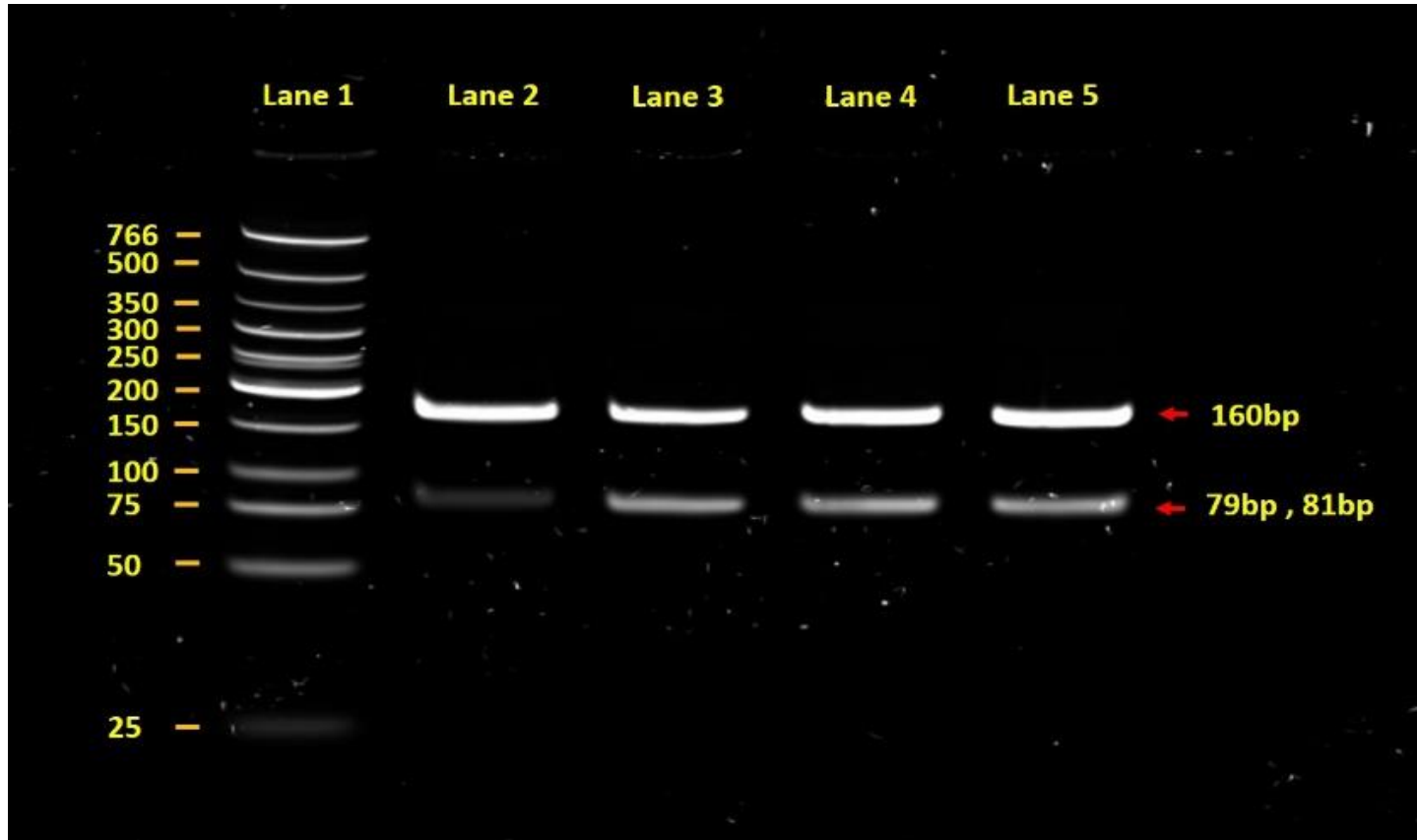
Generalisations of Splicing Languages

Case	Number of Palindromic Rule	Number of Non-Palindromic Rule		Total Number of Cutting Site	Crossing	
		Palindromic Crossing	Non-Palindromic Crossing		Same	Different
1	1			1		
2	1			2		
3	2			2	✓	
4	2			2		✓
5		1		1		
6			1	1		
7		1		2		
8			1	2		
9		2		2	✓	
10			2	2	✓	
11		2		2		✓
12			2	2		✓
13	1	1		2	✓	
14	1		1	2		✓
15	1		1	2		✓

Simple Splicing System

Definition 3 [16] Simple Splicing System

Let $S = (A, I, R)$ be a splicing system in which all rules in R have the form $(r, \lambda; r, \lambda)$, where $r \in R$. Then S is called a simple splicing system.

A splicing language, L is said to be a simple splicing language if L can be generated by a simple splicing system

16. A. Mateescu, G. Păun, G. Rozenberg, A. Salomaa, "Simple splicing systems," *Discrete. Appl. Math.*, vol. 84, no. 1-3, pp. 145-163, 1998.

innovative ● entrepreneurial ● global

www.utm.my

Proposition 1

Let $S = (A, I, R)$ be a simple splicing system with a cutting site of a palindromic rule r where $A = \{\alpha, r, \gamma\}$, $I = \{\alpha r \gamma\}$ and $R = \{r\}$, then the splicing language is $(\alpha + \gamma')r(\gamma + \alpha')$ where $r \notin \{\alpha, \gamma\}$ and α' and γ' are rotation of α and γ respectively through 180°.

Example:

Given a splicing system $S = (A, I, B, C)$ where $I = \{aggactagtct\}$ is the set of initial string, set $B = \{(c, ta, g)\}$ is the set of cleavage pattern for the enzyme BfaI, and set C is the empty set.

The enzyme BfaI, $ctag$ is a palindromic rule since the base sequence of enzyme BfaI reads the same forwards and backwards. The initial string $aggactagtct$ has one cutting site of the enzyme BfaI. Thus, by using Proposition 1, the resulting splicing language is

$$(agga + aga)ctag(tct + tcct)$$

where strings $agga$, aga, $ctag$, tct and $tcct$ indicate strings α, γ', r, γ and α' respectively.

Grammar

- A grammar G is a set of production rules for strings in formal language [5].

- The grammar generates strings by arranging the production rules in sequential order, known as a language generated by the grammar.

Definition 4 [4] Grammar

A grammar *G* is defined as a quadruple *G* = (*V*, *T*, *S*, *P*), where *V* is a finite set of objects called variables, *T* is a finite set of objects called terminal symbols, *S* ∈ *V* is a special symbol called the start variable and *P* is a finite set of productions.

The set $L(G) = \{w \in T^* : S \overset{*}{\Rightarrow} w\}$ is the language generated by *G*, where $\overset{*}{\Rightarrow}$ denotes zero or more steps of sequence of productions.

---

5.   P. Linz, *An Introduction to Formal Languages and Automata*, 4th ed. USA: Jones and Bartlett Publisher, 2006.

innovative ● entrepreneurial ● global

www.utm.my

Grammar (cont.)

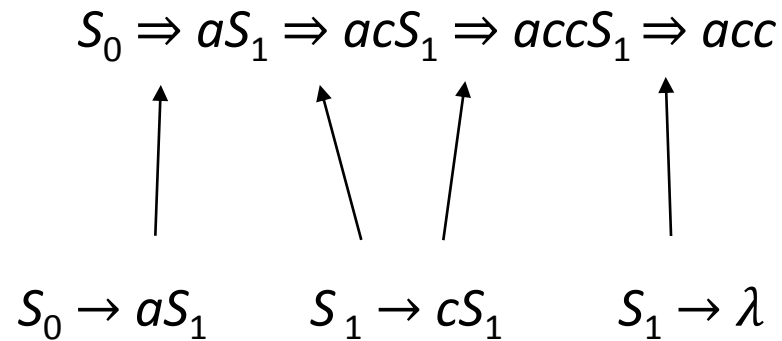**Example:**

- Grammar:
$$S_0 \to aS_1 \mid bS_1,$$
$$S_1 \to cS_1 \mid \lambda$$

- Derivation of sentence *acc*:

$$S_0 \Rightarrow aS_1 \Rightarrow acS_1 \Rightarrow accS_1 \Rightarrow acc$$

$$S_0 \to aS_1 \qquad S_1 \to cS_1 \qquad S_1 \to \lambda$$

Automata

- The concept of automata can be applied in DNA splicing systems.

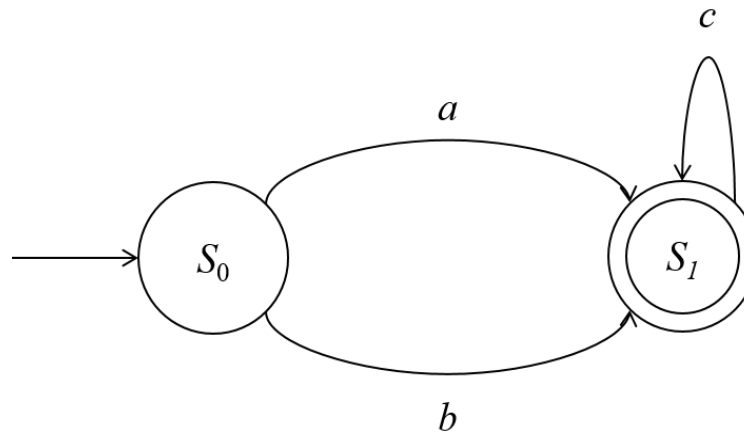- The language generated by the automaton depicts the splicing language from the splicing system.

Definition 5 [5] Deterministic Finite Automaton

A deterministic finite automaton M is a 5-tuple, $(Q, \Sigma, \delta, q_0, F)$ consisting of a finite set of states Q, a finite set of input symbols called the alphabet Σ, a transition function $\delta : Q \times \Sigma \to Q$, an initial state $q_0 \in Q$ and a set of final states $F \subseteq Q$.

---

5. Linz, P. *An Introduction to Formal Languages and Automata*. 4th. ed. USA: Jones and Bartlett Publisher. 2006.

innovative ● entrepreneurial ● global

www.utm.my

Automata (cont.)

Symbols	Description
◯	State
◎	Final State
⟶	Transition

Example:

An example of a deterministic finite automaton

Example:

The figure here shows an example of a deterministic finite automaton that accepts the language $L((a+b) \cdot c^*)$ generated by the grammar with P consisting of the productions
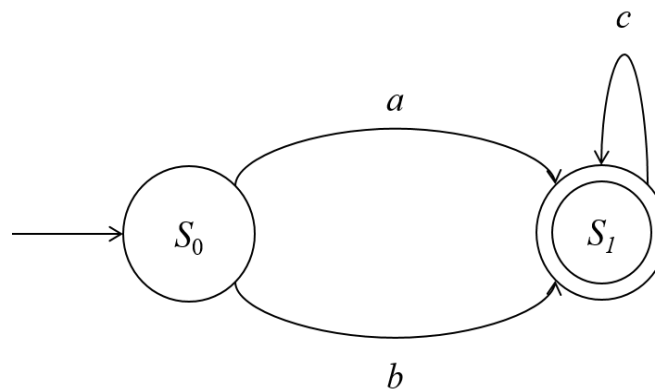
$$S_0 \rightarrow aS_1 \mid bS_1,$$

$$S_1 \rightarrow cS_1 \mid \lambda,$$

where $Q = \{S_0, S_1\}$, $\Sigma = \{a, b, c\}$, S_0 is the initial state, $F = \{S_1\}$ and δ is given by

$\delta(S_0, a) = S_1,$

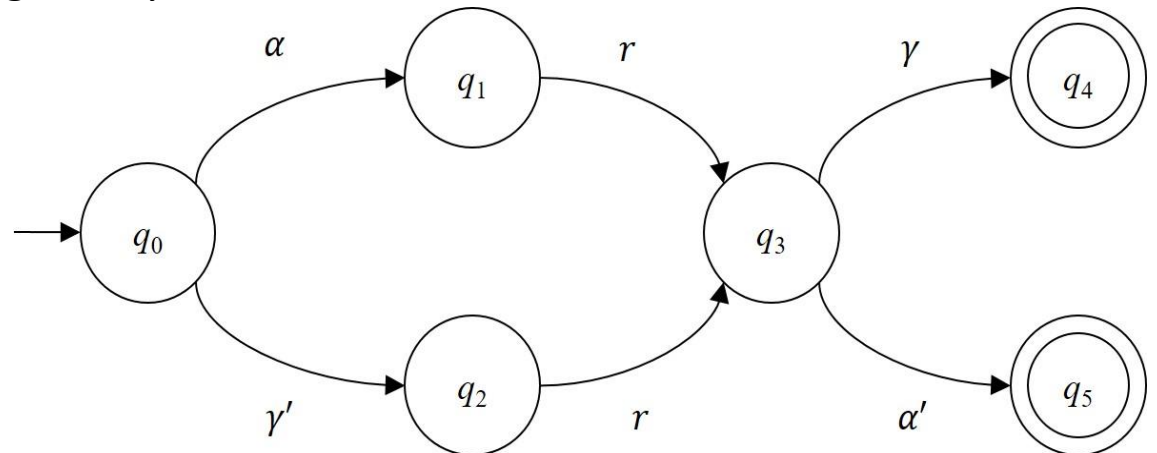$\delta(S_0, b) = S_1,$

$\delta(S_1, c) = S_1.$

Theorem 1

Given $S = (A, \alpha r\gamma, r)$ is a simple splicing system involving a cutting site of a palindromic rule r where $A = \{\alpha, r, \gamma\}$ are the set of variables used to denote any arbitrary dsDNA string, and α', r' and γ' are α, r and γ respectively after 180º rotation, $M = (Q, \Sigma, \delta, q_0, F)$ is a deterministic finite automaton for the splicing system that accepts the language $L(S)$, in which $Q = \{q_0, q_1, q_2, q_3, q_4, q_5\}$ is the set of states where q_0 is the initial state and $F = \{q_4, q_5\}$ is the set of final states, $\Sigma = \{\alpha, \alpha', r, \gamma, \gamma'\}$ is the set of inputs and δ is given by

$\delta\left(q_0, \alpha\right) = q_1$,
$\delta\left(q_0, \gamma'\right) = q_2$,
$\delta\left(q_1, r\right) = q_3$,
$\delta\left(q_2, r\right) = q_3$,
$\delta\left(q_3, \gamma\right) = q_4$ and
$\delta\left(q_3, \alpha'\right) = q_5$.

DNA Splicing Research Group (DSRG)

Group Leader

Prof. Dr. Nor Haniza Sarmin

innovative • entrepreneurial • global

www.utm.my

Group Members

Dr. Fong Wan Heng

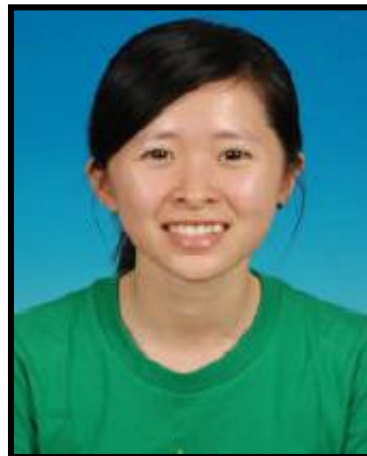**Assoc. Prof. Dr. Yuhani Yusof**

Dr. Fariba Karimi

Deborah Lim Shin Fei

Lim Shun Jinn

Nurul Afidah Mohd Sebry

Group Members (cont.)

Dr. Nur Zatul Akmar Hamzah

Dr. Gan Yee Siang

Dr. Mathuri Selvarajoo

Dr. Muhammad Azrin Ahmad

Dr. Nurhidaya Mohamad Jan

Nurul Izzaty Ismail

Group Members (cont.)

Aqilahfarhana Abdul Rahman

Ahmad Firdaus Yosman

Muhammad Nur Syiham Abdul Razak

Collaborators

Prof. Dr. Susannah Gal

Assoc. Prof. Dr. Elizabeth Goode

Prof. Dr. KG Subramanian

Dr. Sherzod Turaev

Dr. Mohd Firdaus Abdul Wahab

DR. FONG WAN HENG

https://people.utm.my/fwh/

fwh@utm.my

+6075534325 / +60128801875

Thank You

innovative ● entrepreneurial ● global