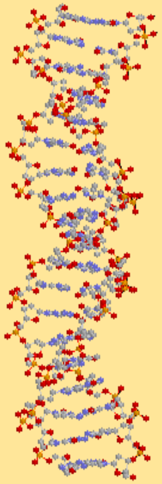
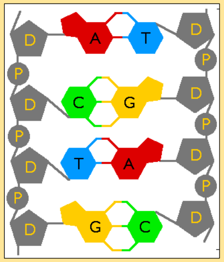
International Conference on Advanced Research in Mathematical Sciences (ICARMS)
24 February 2021



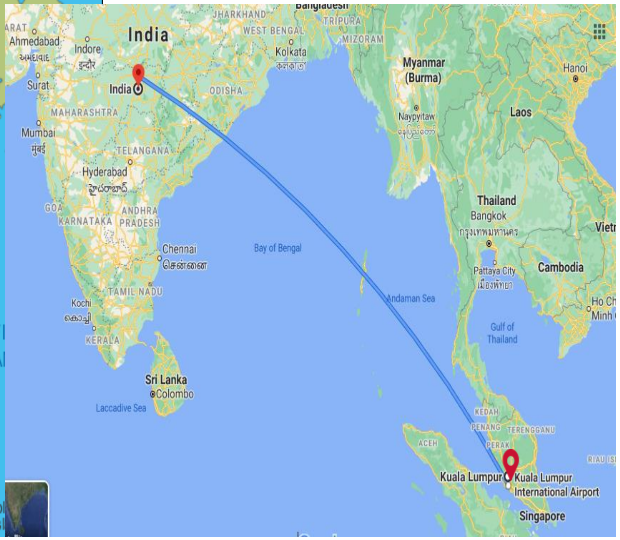
DNA Splicing System - From Birth to Present

Prof. Dr. Nor Haniza Sarmin
Department of Mathematical Sciences, Faculty of Science,
Universiti Teknologi Malaysia, Johor, Malaysia.
Email: nhs@utm.my



INDIA - MALAYSIA

© 2020 Geology.com

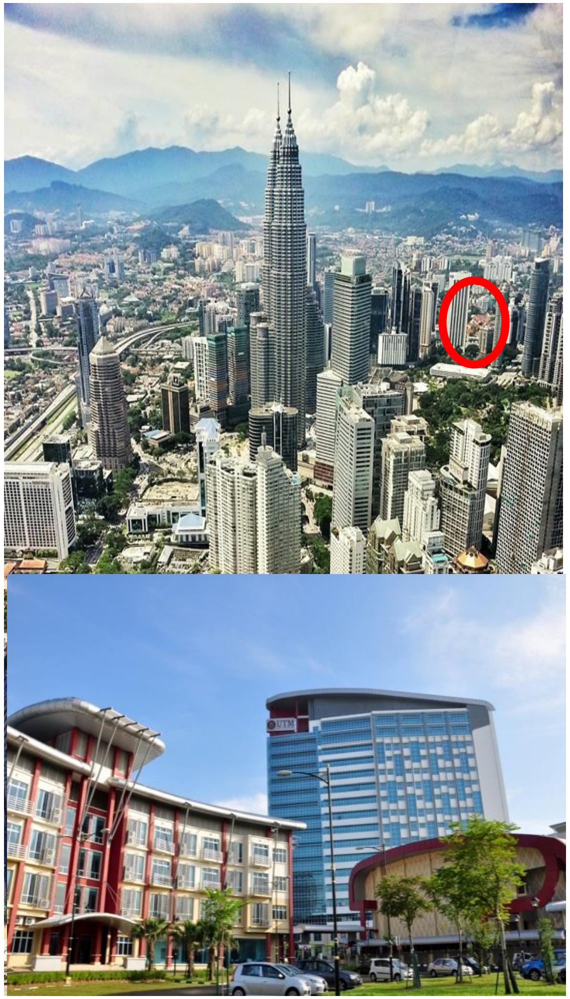


3,081 km

Distance from India to Malaysia

UTM Main Campuses

UTM Johor Bahru
(1145 hectares) – main campus



UTM Kuala Lumpur Campus – branch (19 hectares)

UTM Campus – Pagoh
Innovation Centre in Agritechnology for Advanced Bioprocessing (ICA)

Faculties in UTM

7 Faculties

200 PG programs

55 Bachelor programs

Engineering

- School of Chemical & Energy Eng.
- School of Electrical Eng.
- School of Civil Eng
- School of Mechanical Eng
- School of Biomedical Eng & Health Sciences
- School of Computing

Social Sciences & Humanities

- School of Education
- School of Human Resource Dev. & Psychology
- Islamic Civilization Academy
- Language Academy
- Raja Zarith Sofiah Center for Advanced Studies on Islam, Science and Civilization

Built Environment & Surveying

- Geoinformation
- Real Estate
- Quantity Surveying
- Architecture
- Landscape Architecture
- Urban Regional Planning

Science

- Physics
- Chemistry
- **Mathematical Sciences**
- Biosciences

Azman Hashim International Business School

- Business Administration
- Accounting and Finance
- Information System
- School of Computing

Razak Technology & Informatics

- Engineering & Technology
- Science Management and Design
- Advanced Informatics
- Perdana Center

Malaysia-Japan International Institute of Technology

- Electronic Systems Engineering
- Mechanical Precision Engineering
- Chemical Process Eng. & Sustainable Systems
- Technology & Innovation Management
- Disaster Risk Management
- Sustainability and Environmental Sciences

UTM Quick Facts and Figures

1,416
Academic Staff

7
Faculties

23,476
Students

14,565
UG Students

8,911
PG Students

4,495
International Students

627
Mobility Outbound
2019

2,248
Mobility Inbound
2019

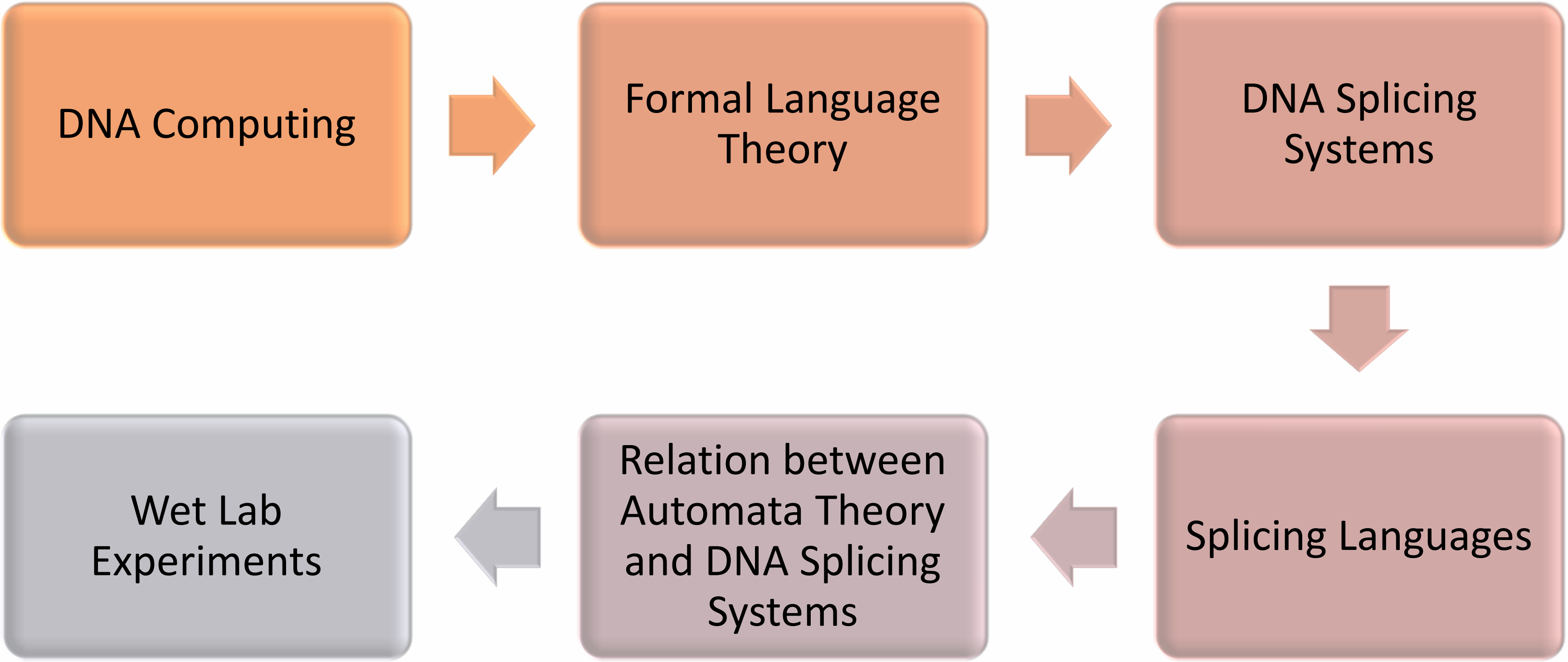
updated June 2020

DNA Splicing System - From Birth to Present

ABSTRACT

The mathematical modelling of **DNA splicing system** has first been introduced by Tom Head in 1987 which is simulated by the technique of **recombinant DNA molecules** that relies on **restriction enzymes** in the study of **formal language theory**. In this talk, **variants of splicing models** that have been developed over the years will be presented. Furthermore, the splicing systems will be shown as **graphically represented** using the concepts of **automata**. Lastly, **molecular works** for splicing systems which are conducted through **wet-lab experiments** will be shared to validate the existence of splicing languages.

Presentation Outlines



DNA Computing

Formal Language Theory

DNA Splicing Systems

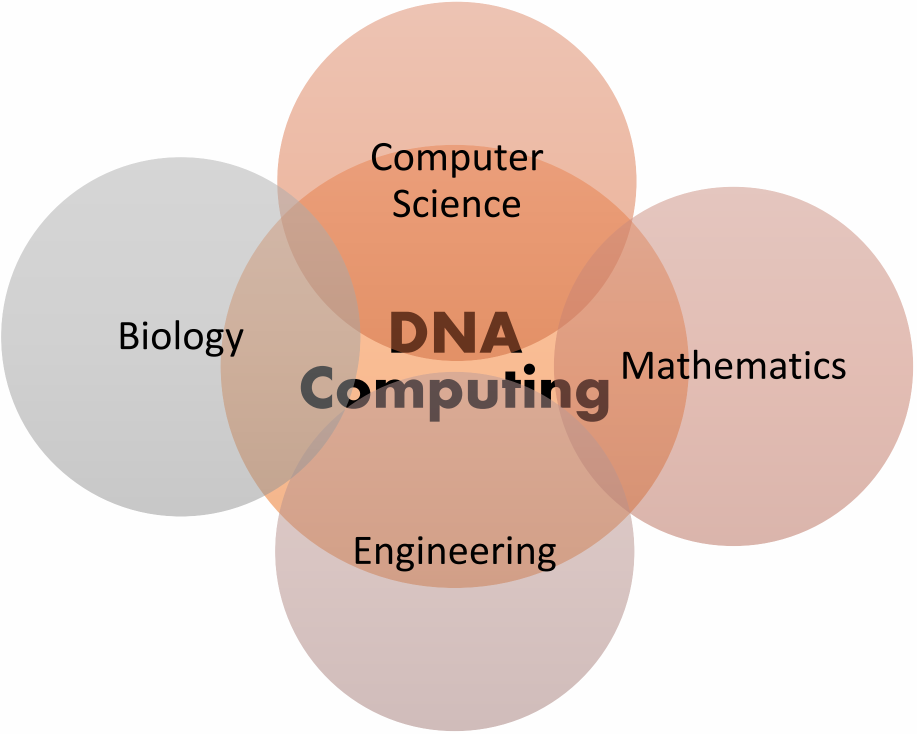
Splicing Languages

Relation between Automata Theory and DNA Splicing Systems

Wet Lab Experiments

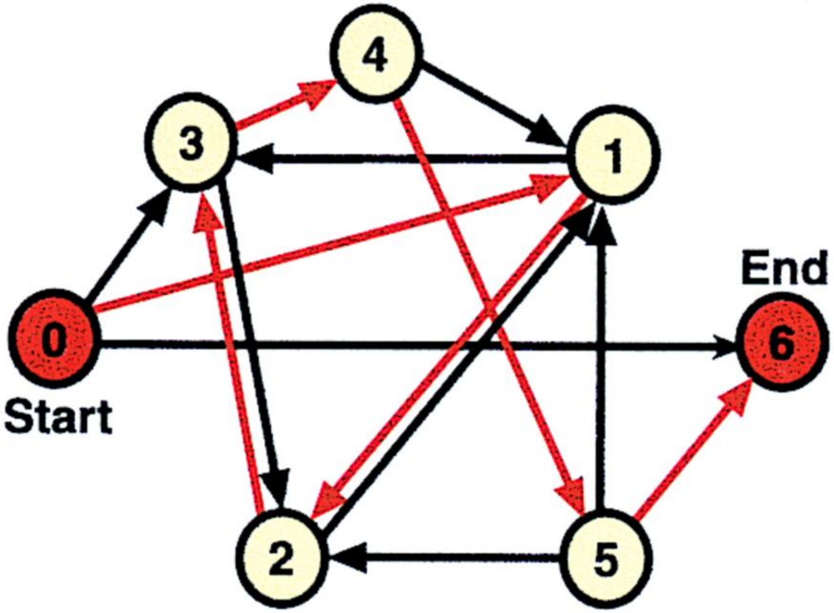
DNA Computing

- DNA computing has emerged in the last twenty years as an exciting new research field at the intersection of Computer Science, Biology, Engineering and Mathematics.



DNA Computing (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.



Adleman, L. M. (1994). Molecular Computation of Solutions to Combinatorial Problems. *Science, 266*(5187), 1021-1024. doi:https://doi.org/10.1126/science.7973651

DNA Computing (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.

Formal Language Theory

Splicing System

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

Formal Language Theory (cont.)

- Formal language theory is a study of formal languages which are defined with respect to an alphabet.
- The alphabet of language is a finite set of symbols or letters.
- The elements of the alphabet can be "ordinary" letters (A to Z, a to z), digits (0, 1, 2, …) or any symbols (eg: #, $, …).
- A finite sequence that concatenate from symbols or letters is called a string.

Example

001 is a string over the alphabet {0, 1}

abba is a string over the alphabet {a, b}

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

Formal Language Theory (cont.)

- A formal language is a set of strings formed over an alphabet.
- Some notations and operations on languages are given in the following:-

Symbols	Description
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
$\cup$	Union
$\cdot$	Concatenation
$*$	Star-closure
{} or ()	Parentheses

Formal Language Theory (cont.)

Example

Let *A* = {a, b, c, …, y, z}. Then *A** is the set of all strings over the letters of Latin alphabet. In particular, the following are formal languages:

- the set of strings consisting of vowels only

  {a, e, i, o, u, aa, ii, oo, …}
- the set of palindromes: strings that read the same from right to left or from left to right

  eg: {level, madam, mom, …}
- the set {i, you, he, she, it, we, they}
- the empty set.

I graduated from State University of New York SUNY) at Binghamton (now known as Binghamton University), New York, USA

BSc (1989), MA (1990), PhD (1995-1998)

DNA Splicing Systems

- The mathematical modelling of splicing system was first defined by Head 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.

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. doi:https://doi.org/10.1007/BF02481771

Tom Head’s visit to UTM, 2004

Visit to SUNY Binghamton, USA, 2007

Visit and Research Seminar at SUNY Binghamton, 2007

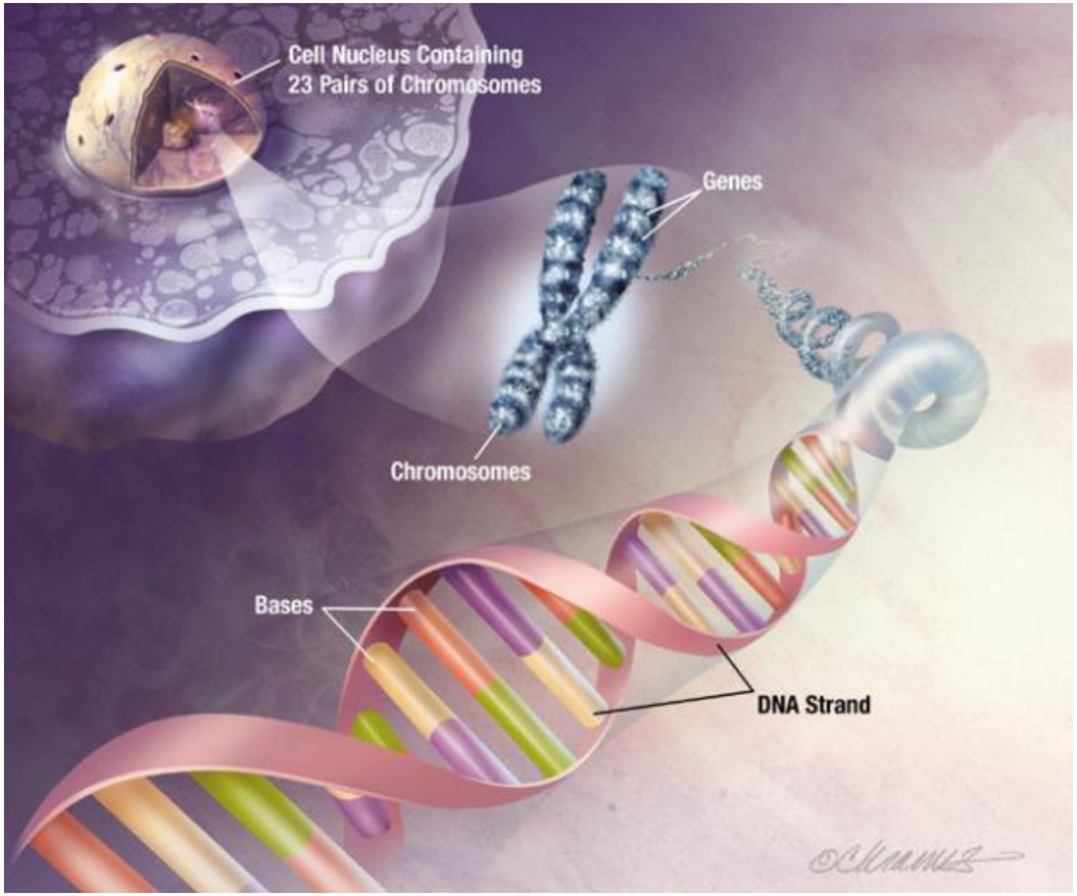
Conference on Biomathematical Computing: Past, Present and Prospects, SUNY Binghamton, USA, 2008

DNA Splicing Systems

- Structure of DNA
- Restriction Enzymes
- DNA Ligase

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

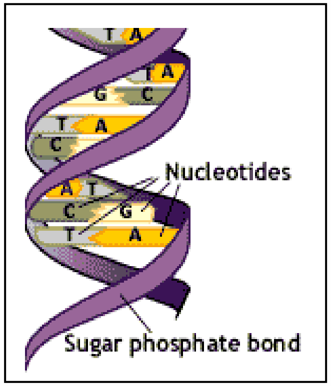
Made up of

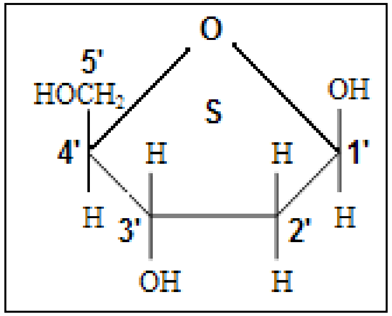
Nucleotides

Phosphate group

Base

Pentose (a 5-carbon sugar)

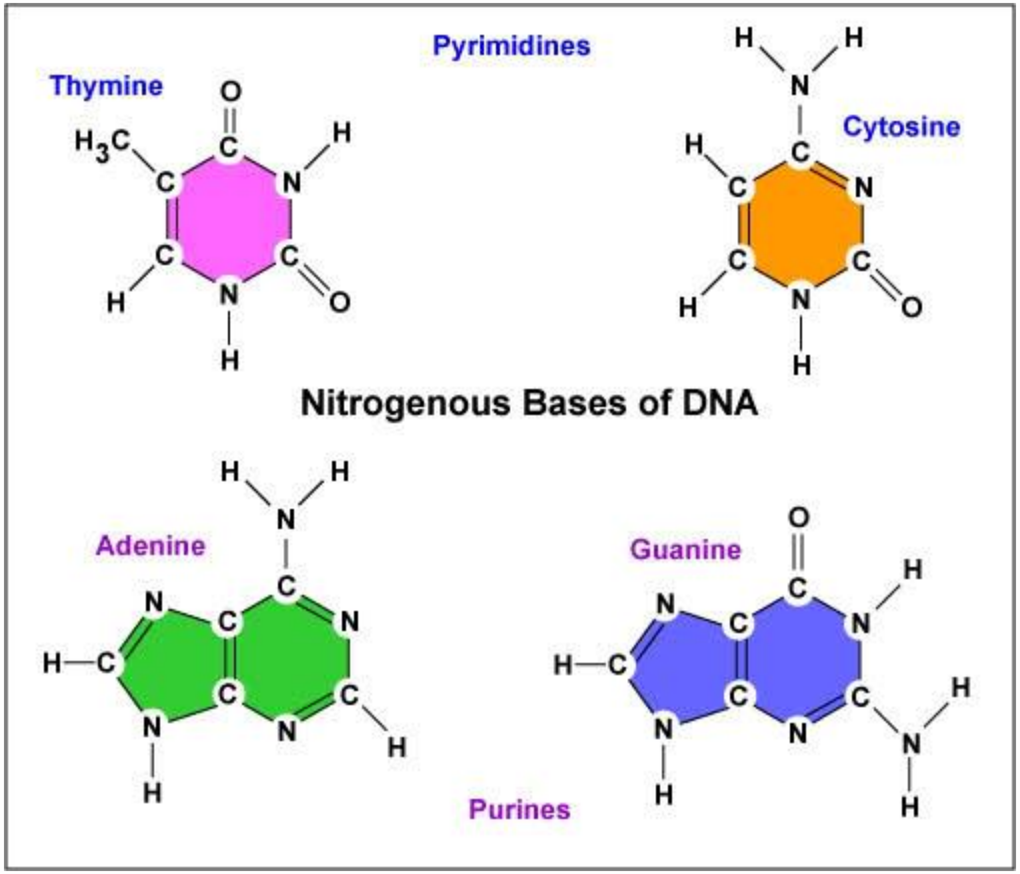




Adenine (A)
Cytosine (C)
Guanine (G)
Thymine (T)

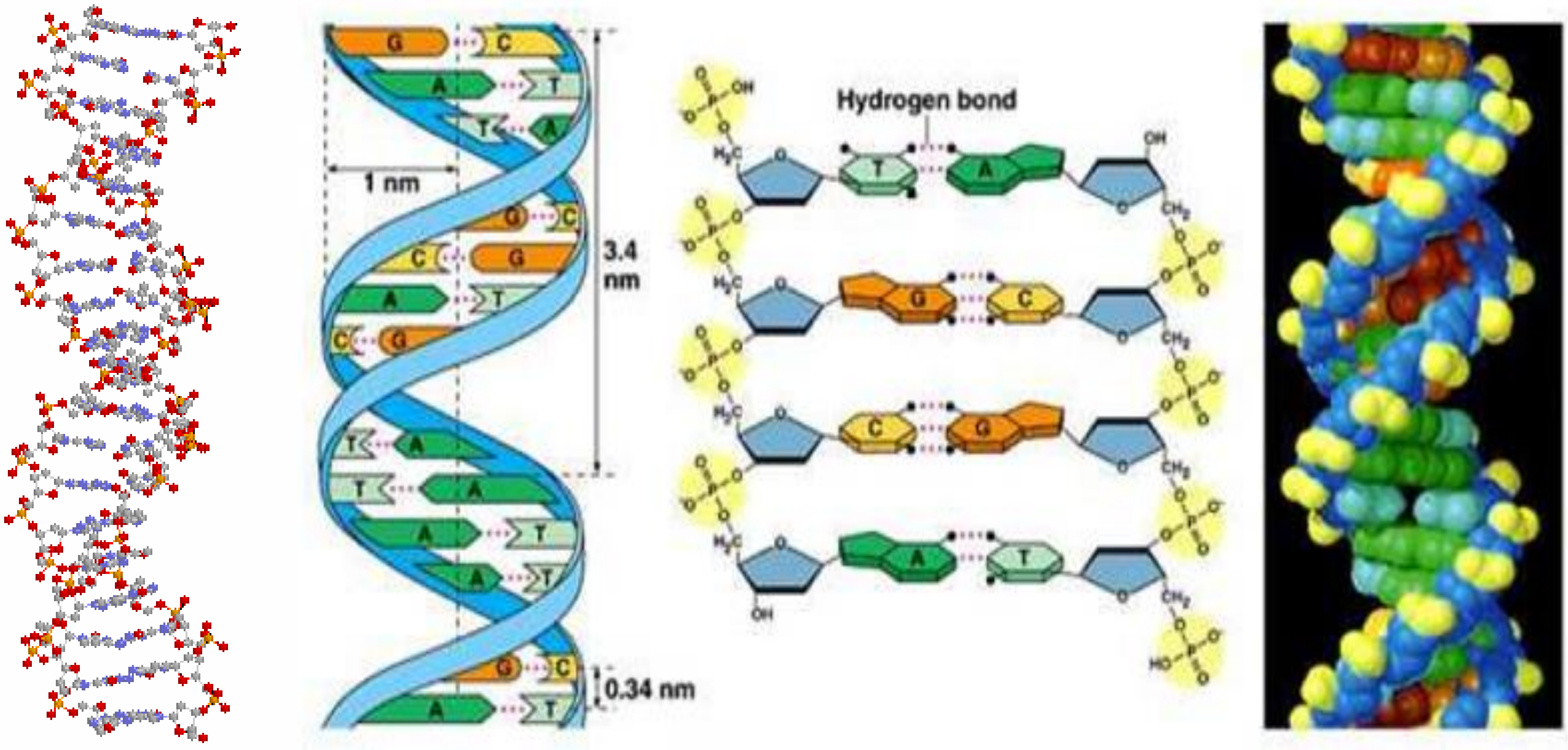
Structure of DNA (cont.)

- 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.



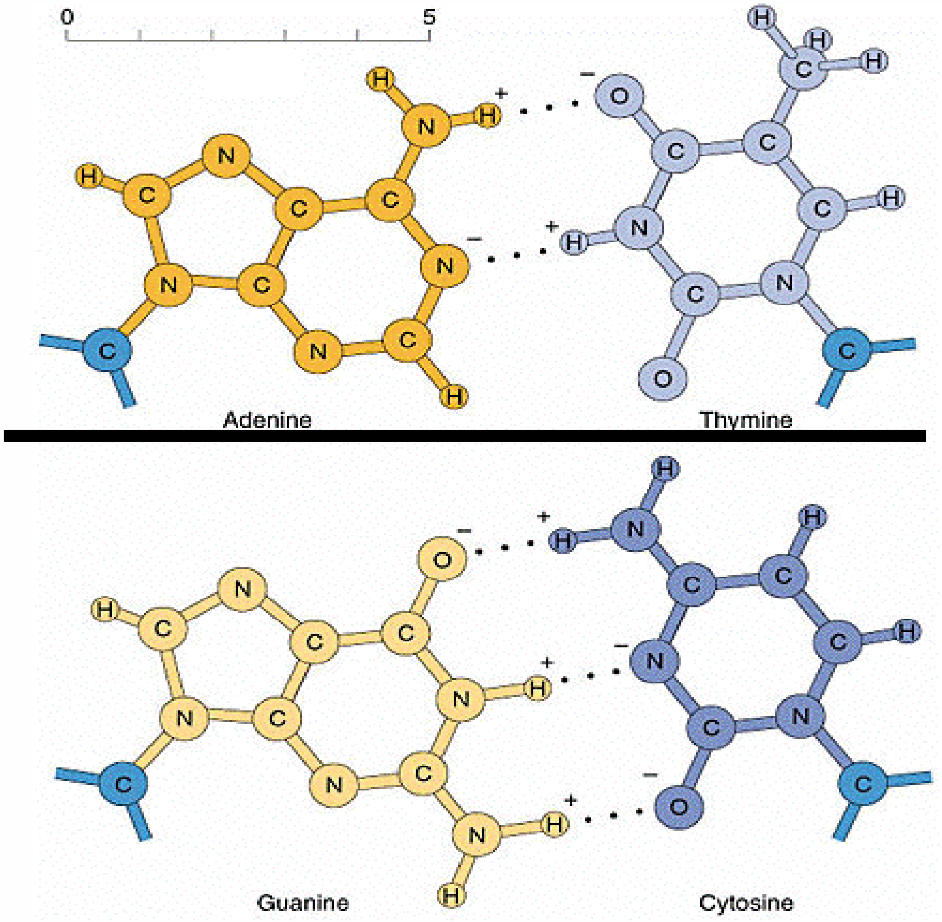
Structure of DNA (cont.)

- 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 Enzymes

- 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▼CGC…3’

3’…GGC▲G…5’

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

5’…AGC▼GCT…3’

3’…TCG▲CGA…5’

Restriction Enzymes (cont.)

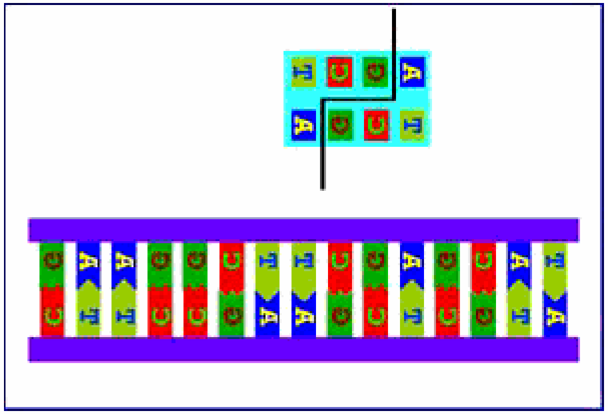
- Every restriction enzyme has a triple known as the cleavage pattern of the enzyme.
- The triple is denoted as a rule for the restriction enzyme which consists of left context, crossing and right context (Head, 1998).
- The restriction enzyme *EcoR*I is isolated from the bacterium Escherichia coli with strain serotype *R*; I indicates the first enzyme discovered from the bacterium.
- The cleavage pattern of restriction enzyme *EcoR*I is

$$\underbrace{(\mathrm{g},}_{\text{left context}}\ \underbrace{\mathrm{aatt},}_{\textbf{crossing}}\ \underbrace{\mathrm{c})}_{\text{right context}}$$

Head, T. (1998). Splicing Representations of Strictly Locally Testable Languages. *Discrete Applied Mathematics, 87*(1), 139-147. doi:https://doi.org/10.1016/S0166-218X(98)00053-5

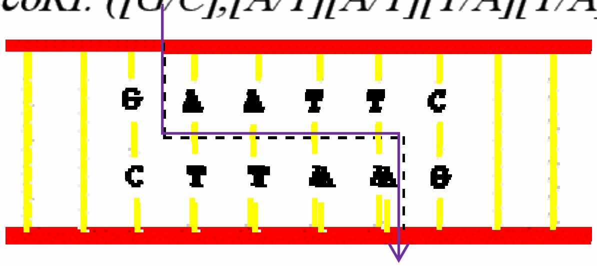
Restriction Enzymes (cont.)

- A restriction enzyme is an enzyme that cuts double-stranded or single stranded DNA at specific recognized nucleotide sequences, known as restriction sites.

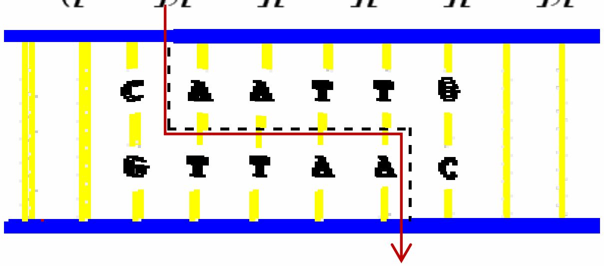


Restriction Enzymes (cont.)

EcoRI: ([G/C],[A/T][A/T][T/A][T/A],[C/G])

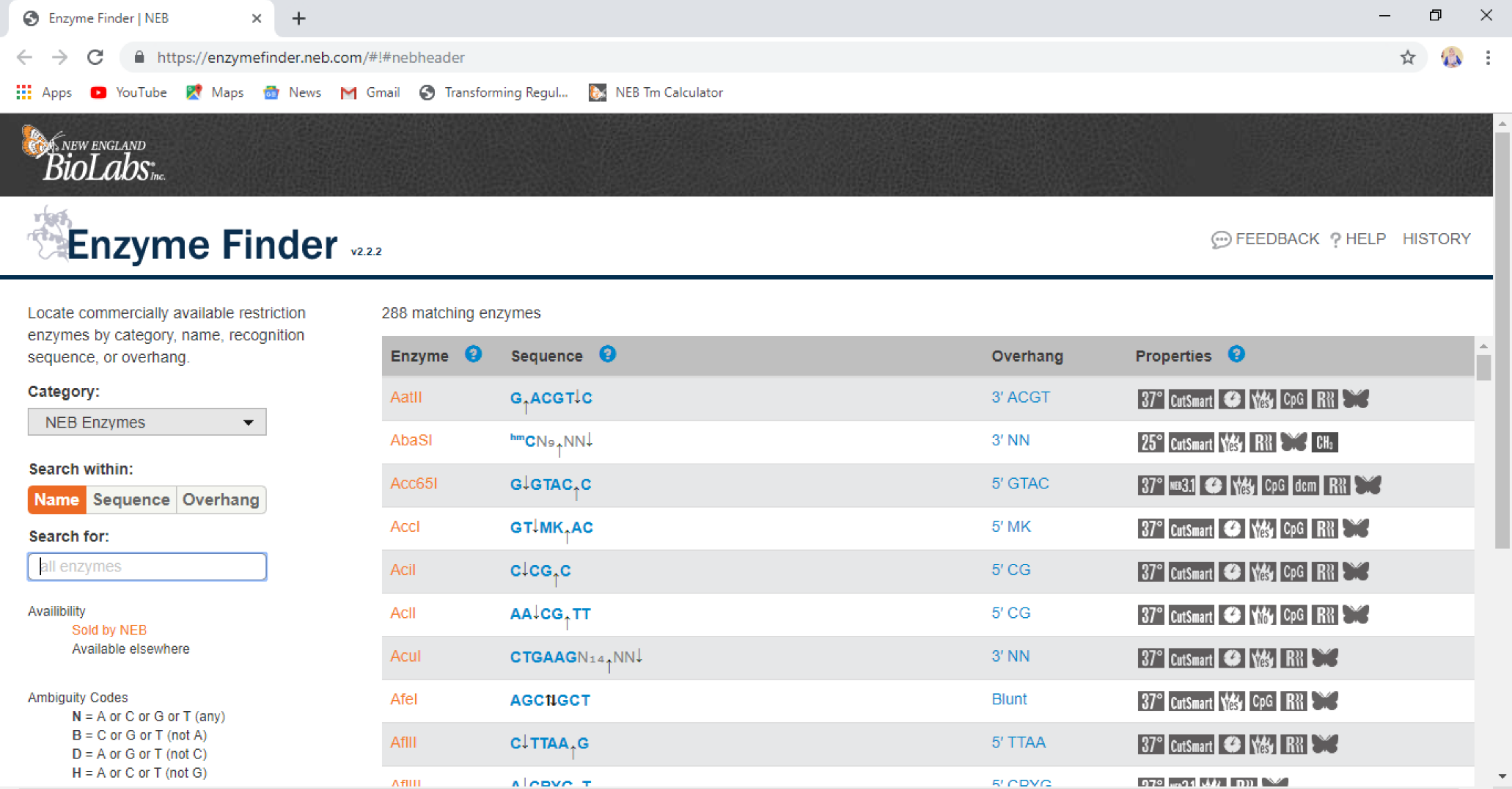


MfeI: ([C/G],[A/T][A/T][T/A][T/A],[G/C])



Restriction Enzyme (cont.)

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



Enzyme Finder v2.2.2

Locate commercially available restriction enzymes by category, name, recognition sequence, or overhang.

Category: NEB Enzymes

Search within: Name | Sequence | Overhang

Search for: all enzymes

Availibility
Sold by NEB
Available elsewhere

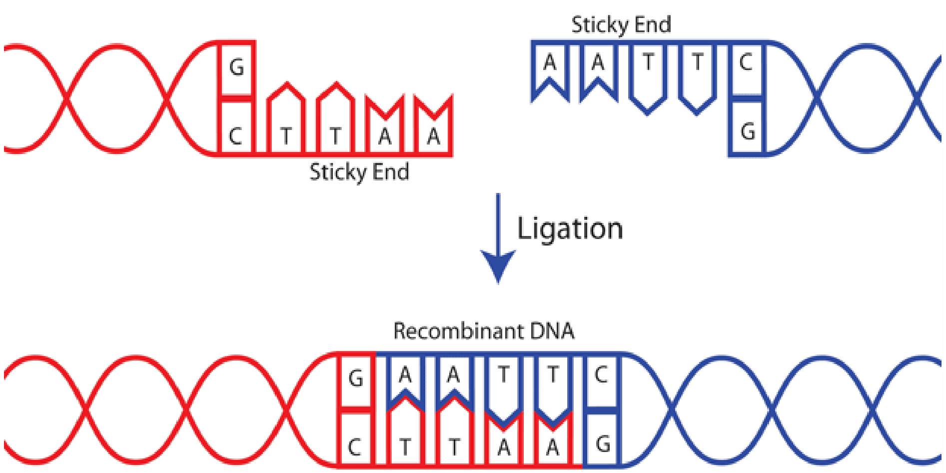
Ambiguity Codes
N = A or C or G or T (any)
B = C or G or T (not A)
D = A or G or T (not C)
H = A or C or T (not G)

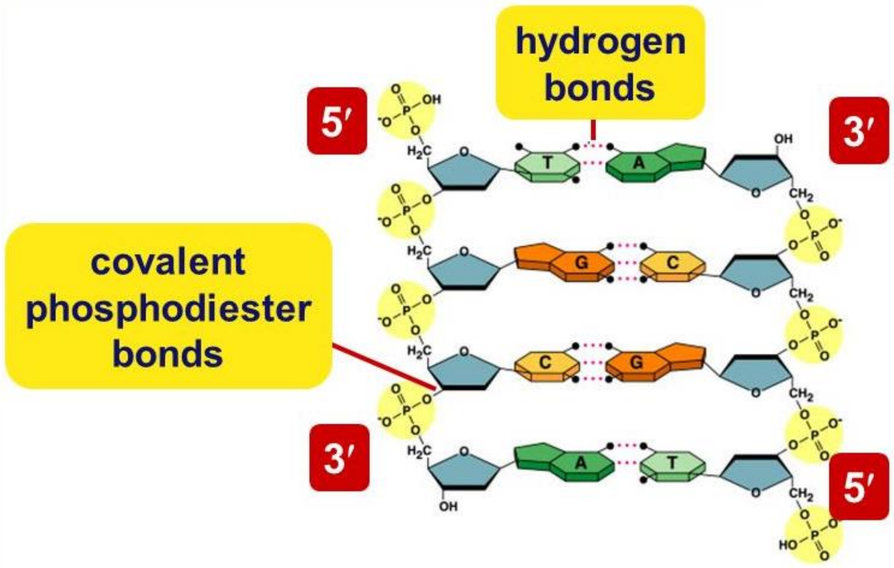
288 matching enzymes

Enzyme	Sequence	Overhang	Properties
AatII	G↑ACGT↓C	3′ ACGT	37° CutSmart 🕘 Yes CpG R 🦋
AbaSI	^{hm}CN$_{9}$↑NN↓	3′ NN	25° CutSmart Yes R 🦋 CH$_3$
Acc65I	G↓GTAC↑C	5′ GTAC	37° NEB3.1 🕘 Yes CpG dcm R 🦋
AccI	GT↓MK↑AC	5′ MK	37° CutSmart 🕘 Yes CpG R 🦋
AciI	C↓CG↑C	5′ CG	37° CutSmart 🕘 Yes CpG R 🦋
AclI	AA↓CG↑TT	5′ CG	37° CutSmart 🕘 No CpG R 🦋
AcuI	CTGAAGN$_{14}$↑NN↓	3′ NN	37° CutSmart 🕘 Yes R 🦋
AfeI	AGC⇅GCT	Blunt	37° CutSmart Yes CpG R 🦋
AflII	C↓TTAA↑G	5′ TTAA	37° CutSmart 🕘 Yes R 🦋

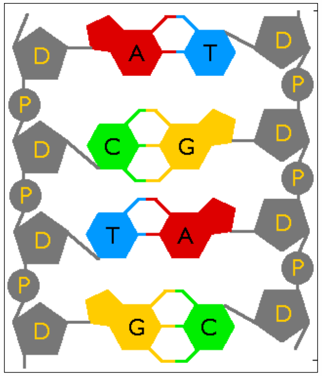
DNA Ligase

- DNA ligase is an enzyme that can catalyze the linking of DNA strands together by forming a phosphodiester bond.





Modelling of DNA Splicing System



Nitrogenous base pairings

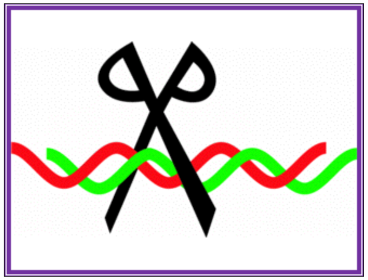
↓

dsDNA

Nucleotide chains

↓

DNA strings

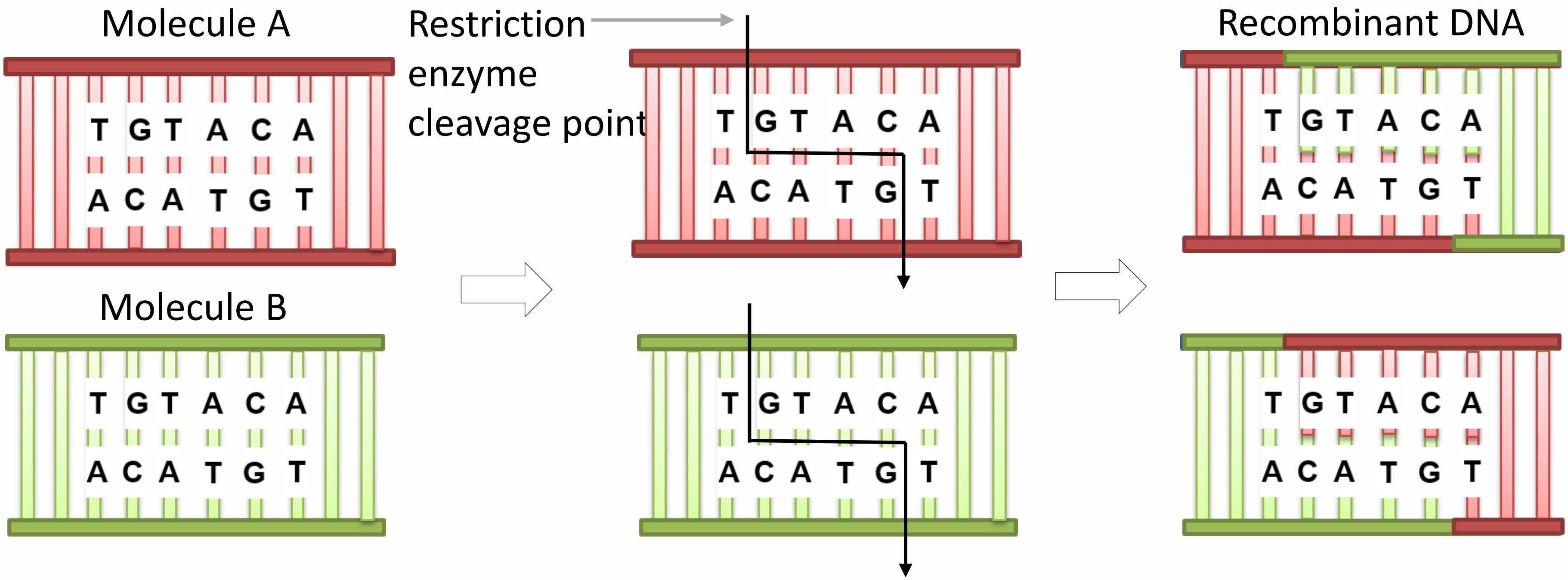


Restriction enzymes

↓

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

DNA Splicing System

Definition 1 (Head, 1987) 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^*
- B and C: finite sets of triples (c, x, d) with c, x and d in A^*
- Triples in B are called left patterns
- Triples in C are called right patterns

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 $L = L(S)$.

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. doi:https://doi.org/10.1007/BF02481771

DNA Splicing System (cont.)

Example

Suppose that *S* = (*A*, *I*, *B*, *C*) is a splicing system in which *A* = $\left\{\begin{matrix}\mathrm{A}, & \mathrm{C}, & \mathrm{G}, & \mathrm{T}\\ \mathrm{T} & \mathrm{G} & \mathrm{C} & \mathrm{A}\end{matrix}\right\}$ is the set of dsDNA symbols, *I* = $\left\{\begin{matrix}\mathrm{GAATTCTCTGTAAT}\\ \mathrm{CTTAAGAGACATTA}\end{matrix}\right\}$ is the set consisting of an initial string of molecules, set *B* = $\left\{\left(\begin{matrix}\mathrm{G} & \mathrm{AATT} & \mathrm{C}\\ \mathrm{C}' & \mathrm{TTAA}' & \mathrm{G}\end{matrix}\right)\right\}$ is the set of cleavage pattern for the enzyme ***Eco*RI** and set *C* is the empty set.

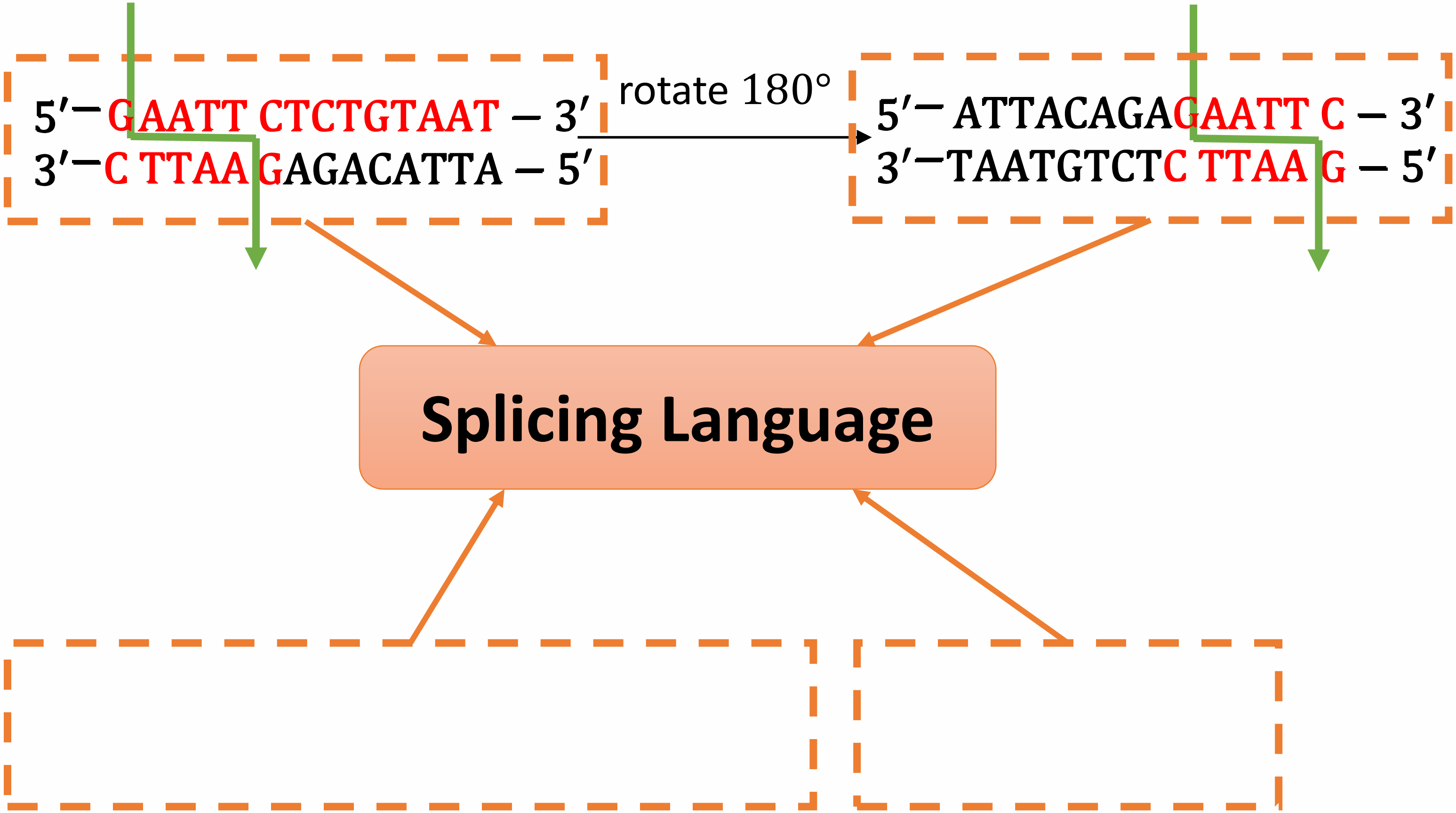
The initial string is shown in the following:

$$\begin{matrix}5'-\mathrm{GAATTCTCTGTAAT}-3'\\ 3'-\mathrm{CTTAAGAGACATTA}-5''\end{matrix}$$

or written 180 degree wise,

$$\begin{matrix}5'-\mathrm{ATTACAGAGAATTC}-3'\\ 3'-\mathrm{TAATGTCTCTTAAG}-5'.\end{matrix}$$

DNA Splicing System (cont.)



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

rotate 180°

5′− ATTACAGAGAATT C − 3′
3′− TAATGTCTC TTAA G − 5′

Splicing Language

Variant of Splicing Models

Head's Splicing System 1987

Paun's Splicing System 1996

Pixton's Splicing System 1996

Goode-Pixton Splicing System 2004

Yusof-Goode Splicing System 2013

Fuzzy Splicing System 2014

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. doi:https://doi.org/10.1007/BF02481771

Păun, G. (1996). On the Splicing Operation. *Discrete Applied Mathematics, 70*(1), 57-79. doi:https://doi.org/10.1016/0166-218X(96)00101-1

Pixton, D. (1996). Regularity of Splicing Languages. *Discrete Applied Mathematics, 69*(1-2), 101-124. doi:https://doi.org/10.1016/0166-218X(95)00079-7

Goode, E., & Pixton, D. (2004). Splicing to the Limit. In N. Jonoska, G. Păun, & G. Rozenberg (Eds.), *Aspects of Molecular Computing, Lecture Notes in Computer Science* (pp. 189-201). Germany: Springer-Verlag.

Yusof, Y., Sarmin, N. H., Fong, W. H., Goode, T. E., & Ahmad, M. A. (2013). *An Analysis of Four Variants of Splicing System.* Paper presented at the 20th National Symposium on Mathematical Sciences - Research in Mathematical Sciences: A Catalyst for Creativity and Innovation (SKSM 2012).

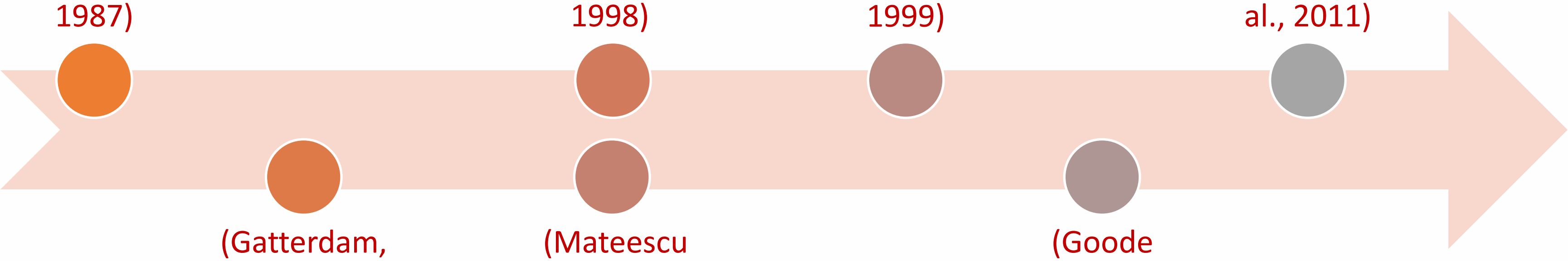
Karimi, F., Turaev, S., Sarmin, N. H., & Fong, W. H. (2014). Fuzzy Splicing Systems. In D. Hwang, J. J. Jung, & N. T. Nguyen (Eds.), *Computational Collective Intelligence. Technologies and Applications, ICCCI 2014, Lecture Notes in Computer Science* (pp. 20-29). Cham, Switzerland: Springer International Publishing.

Notations of Splicing Systems

Splicing System	**Head** $S = (A, I, B, C)$	**Paun** $\sigma = (A, R)$	**Pixton** $\zeta = (R, I)$	**Goode-Pixton** $(w, w') \vdash_r z$	**Yusof-Goode** $S = (A, I, R)$	**Fuzzy** $S = (A, T, I, R, \odot)$
Initial String	$ucxdv$ $pexfq$	uu_1u_2v $u'u_3u_4v'$	$\xi\alpha\eta$ $\xi'\alpha'\eta'$	$w = xuvy$ $w' = x'u'v'y'$	$\alpha uxv\beta$ $\gamma yz\delta$	(ua, x) (bv, y) $x, y \in [0, 1]$
Rule	(c, x, d) (e, x, f)	$u_1\#u_2\$u_3\#u_4$	$(\alpha, \alpha' : \beta)$	$r = (u, v; u'; v')$	$(u, x, v : y, x, z)$	$u\#a\$b\#v$
Splicing Language	$ucxfq$ $pexdv$	uu_1u_4v'	$\xi\beta\eta'$	$z = xuv'y'$	$\alpha uxz\delta$ $\gamma yxv\beta$	$(uv, x \odot y)$ $\odot$ is fuzzy operation

- S: splicing system
- σ and ζ: splicing schemes
- A: finite alphabet
- $T \subseteq A$: terminal alphabet
- I: set of initial strings
- B: set of rules with left pattern
- C: set of rules with right pattern
- $R = B \cup C$: set of rules

Types of Splicing Systems



Null Context, Uniform & Persistent Splicing Systems (Head, 1987)

(Gatterdam, 1992) Permanent and Crossing Disjoint Splicing Systems

S_k Splicing System (Head, 1998)

(Mateescu et al, 1998) Simple Splicing System

Semi-Null Splicing System (Goode, 1999)

(Goode and Pixton, 2001) Semi-Simple Splicing System

Crossing-Preserved and Self-Closed Splicing Systems (Karimi et al., 2011)

Types of Splicing Systems (cont.)

Definition 2 (Head, 1987) Null-Context Splicing System

A null-context splicing system is a splicing system $S = (A, I, B, C)$ for which each cleavage pattern in B and each in C has the form $(1, x, 1)$.

Definition 3 (Head, 1987) Uniform Splicing System

A uniform splicing system is a null context splicing system $S = (A, I, X, X)$ for which there is a positive integer P such that $X = A^P$. A language L is a uniform splicing language if there is a uniform splicing system S for which $L = L(S)$.

Definition 4 (Head, 1987) 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 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*.

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. doi:https://doi.org/10.1007/BF02481771

Types of Splicing Systems (cont.)

Definition 5 (Gatterdam, 1992) 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 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* is an occurrence of the crossing of a site in *ucxfq.*

Definition 6 (Gatterdam, 1992) Crossing Disjoint Splicing System
A splicing system *S = (A, I, B, C*) is crossing disjoint if there do not exist patterns (*a*, *x*, *b*) in *B* and (*c*, *x*, *d*) in *C* with the same crossing *x*.

Gatterdam, R. W. (1992). Algorithms for Splicing Systems. *SIAM Journal on Computing, 21*(3), 507-520.

Types of Splicing Systems (cont.)

Definition 7 (Head, 1998) S_k Splicing System
Let k be an integer ≥ -1 . An S_k splicing system is a null-context splicing system $S = (A, I, B, C)$ for which, for each rule r in B or C, length $r < k$.

Definition 8 (Mateescu et al., 1998) Simple Splicing System
Let $S = (A, I, R)$ be a splicing system in which all rules in R have the form $(a, 1; a, 1)$ where $a \in A$. Then S is called a simple splicing system.

Head, T. (1998). Splicing Representations of Strictly Locally Testable Languages. *Discrete Applied Mathematics, 87*(1), 139-147. doi:https://doi.org/10.1016/S0166-218X(98)00053-5

Mateescu, A., Păun, G., Rozenberg, G., & Salomaa, A. (1998). Simple splicing systems. *Discrete Applied Mathematics, 84*(1-3), 145-163.

Types of Splicing Systems (cont.)

Definition 9 (Goode, 1999) Semi-Null Splicing System
Let (A, I, R) be a splicing system in which I and R are finite and every rule in R has the form $(u, 1; v, 1)$, where u, v are in A^+. Thus σ = (A, R) is called a semi-null splicing scheme, and (A, I, R) a semi-null splicing system.

Definition 10 (Goode and Pixton, 2001) Semi-Simple Splicing System
Let (A, I, R) be a splicing system in which I and R are finite and every rule in R has the form $(a, 1; b, 1)$, where a, b are in A. Thus σ = (A, R) is called a semi-simple splicing scheme and (A, I, R) a semi-simple splicing system.

Laun, T. E. G. (1999). *Constants and Splicing Systems.* (Ph.D. Thesis), State University of New York, Binghamton.

Goode, E., & Pixton, D. (2001). Semi-Simple Splicing Systems. In C. Martín-Vide & V. Mitrana (Eds.), *Where Mathematics, Computer Science, Linguistics and Biology Meet* (pp. 343-352). Dordrecht: Springer.

Types of Splicing Systems (cont.)

Definition 11 (Karimi et al., 2011) Self-Closed Splicing System

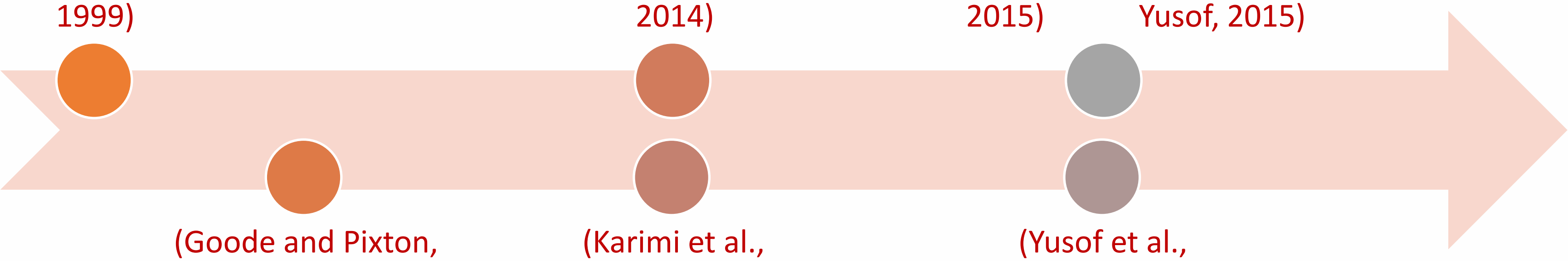
A set of splicing patterns *B* is called self-closed if the set of its sites is closed under its splicing rules. A splicing system is called self-closed if the sets of their patterns are self-closed .

Definition 12 (Karimi et al., 2011) Crossing-Preserved Splicing System

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.

Karimi, F., Sarmin, N. H., & Heng, F. W. (2011). *Crossing-Preserved and Persistent Splicing Systems.* Paper presented at the 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications, Penang, Malaysia.

Types of Splicing Languages



Adult Language
(Laun and Reddy, 1999)

(Goode and Pixton, 2004)
Transient, Limit and n^{th} Order Limit Languages

Second Order Limit Language
(Ahmad et al., 2014)

(Karimi et al., 2014)
Fuzzy and Crisp Languages

Single Stage Splicing Language
(Lim et al., 2015)

(Yusof et al., 2015)
Active Persistent Language

Two Stages Splicing Language
(Mudaber and Yusof, 2015)

Types of Splicing Languages (cont.)

Definition 13 (Laun and Reddy, 1999) Adult Language

Adult language in *L*(*S*) is the set of strings in *L*(*S*) which cannot be used for splicing in a splicing system.

Definition 14 (Goode and Pixton, 2004) Transient Language

A splicing language is called transient if a set of strings is eventually used up and disappears in a given system.

Definition 15 (Goode and Pixton, 2004) Limit language

A limit language is the set of words that are predicted to appear if some amount of each initial molecule is present, and sufficient time has passed for the reaction to reach its equilibrium state, regardless of the balance of the reactants in a particular experimental run of the reaction.

Laun, E., & Reddy, K. J. (1999). *Wet Splicing Systems.* Paper presented at the 3rd DIMACS Workshop on DNA Based Computers, University of Pennsylvania, Philadelphia.

Goode, E., & Pixton, D. (2004). Splicing to the Limit. In N. Jonoska, G. Păun, & G. Rozenberg (Eds.), *Aspects of Molecular Computing, Lecture Notes in Computer Science* (pp. 189-201). Germany: Springer-Verlag.

Types of Splicing Languages (cont.)

Definition 16 (Goode and Pixton, 2004) n^{th} Order Limit language
Let L_{n-1} be the set of second order limit words of L, the set L_n of n^{th} order limit words of L to be the set of first order limit word of L_{n-1}. We obtain L_n from L_{n-1} by deleting the words that are transient in L_{n-1}.

Definition 17 (Ahmad et al., 2014) Second Order Limit Language
Let L_1 be the set of second order limit words of L, the set L_2 of second order limit words of L to be the set of first order limits of L_1. We obtain L_2 from L_1 by deleting words that are transient in L_1.

Goode, E., & Pixton, D. (2004). Splicing to the Limit. In N. Jonoska, G. Păun, & G. Rozenberg (Eds.), *Aspects of Molecular Computing, Lecture Notes in Computer Science* (pp. 189-201). Germany: Springer-Verlag.

Ahmad, M. A., Sarmin, N. H., Heng, F. W., & Yusof, Y. (2014). *An Extension of First Order Limit Language.* Paper presented at the 3rd International Conference on Mathematical Sciences (ICMS3), Kuala Lumpur, Malaysia.

Types of Splicing Languages (cont.)

Definition 18 (Karimi et al., 2014) Fuzzy Language
The fuzzy language generated by a fuzzy splicing system $\gamma = (V, T, A, R, \mu, \odot)$ is defined as $L_f(\gamma) = \{(z, \mu(z)) \in \sigma * f(A) : z \in T^*\}$.

Definition 19 (Karimi et al., 2014) Crisp language
The crisp language generated by a fuzzy splicing system $\gamma = (V, T, A, R, \mu, \odot)$ is defined as $L_c(\gamma) = \{z : (z, \mu(z)) \in L_f(\gamma)\}$.

Karimi, F., Turaev, S., Sarmin, N. H., & Fong, W. H. (2014). Fuzzy Splicing Systems. In D. Hwang, J. J. Jung, & N. T. Nguyen (Eds.), *Computational Collective Intelligence. Technologies and Applications, ICCCI 2014, Lecture Notes in Computer Science* (pp. 20-29). Cham, Switzerland: Springer International Publishing.

Types of Splicing Languages (cont.)

Definition 20 (Lim et al., 2015) Single Stage Splicing Language

Single stage splicing language is defined as

$$[L_1 = L_1(S)] \cong \sum_{r-1}^{n} (Rr + I_{\ r} + l)$$

$$R_r = \text{set of rules, } 1 \leq r \leq n$$

$$I_r = \text{set of initial strings, } 1 \leq r \leq n$$

$$l = \text{ligases}$$

Let $S = (A, I\ I, R)$ be the Y-G splicing system. The set of single stage splicing language, $L_1 = L_1(S)$, models the set of all molecule types which appear when all the restriction enzymes, double stranded deoxyribonucleic acid strings and ligases act simultaneously in a single buffer.

Lim, W. L., Yusof, Y., & Mudaber, M. H. (2015). *Modeling of DNA Single Stage Splicing Language via Yusof-Goode Approach: One String with Two Rules.* Paper presented at the 2nd ISM International Statictical Conference 2014 (ISM-II), Kuantan, Malaysia.

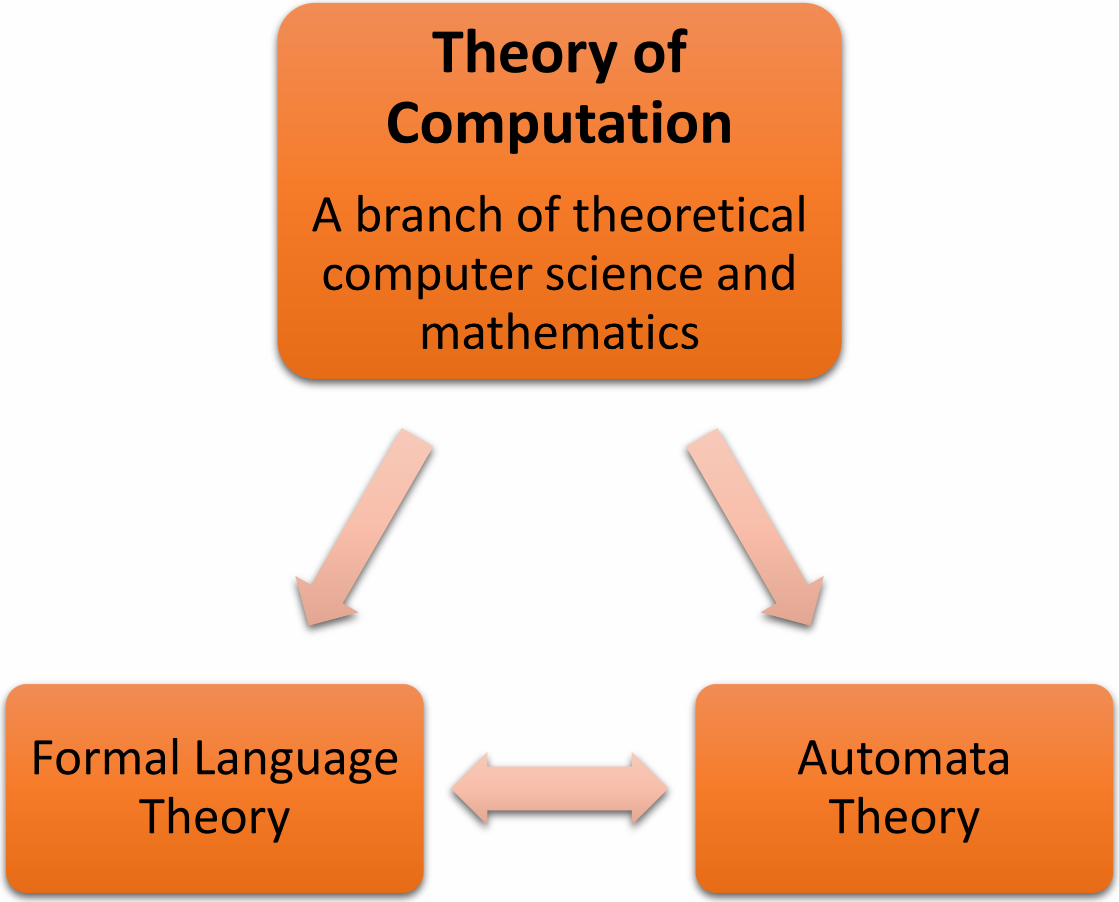
Types of Splicing Languages (cont.)

Definition 21 (Mudaber and Yusof, 2015) Two Stages Splicing Language
Let $S = (A, I, R)$ is a splicing system. Furthermore, let $L = L(S)$ is the set of stage one splicing languages produced by splicing system S and $L = L(S)$ is the 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.

Definition 22 (Yusof et al., 2015) Active Persistent Language
An active persistent language is a set of strings that participate in further splicing and is also contained in the limit language, L_∞.

Mudaber, M. H., & Yusof, Y. (2015). *Two Stages Splicing System.* Paper presented at the International Conference on Mathematics, Engineering & Industrial Applications 2014 (ICoMEIA 2014), Penang, Malaysia.

Yusof, Y., Lim, W. L., Goode, T. E., Sarmin, N. H., Heng, F. W., & Wahab, M. F. A. (2015). *Molecular Aspects of DNA Splicing System.* Paper presented at the International Conference on Mathematics, Engineering and Industrial Applications 2014 (ICoMEIA 2014), Penang, Malaysia.



Theory of Computation

A branch of theoretical computer science and mathematics

Formal Language Theory

Automata Theory

An automaton is a finite representation of a formal language.

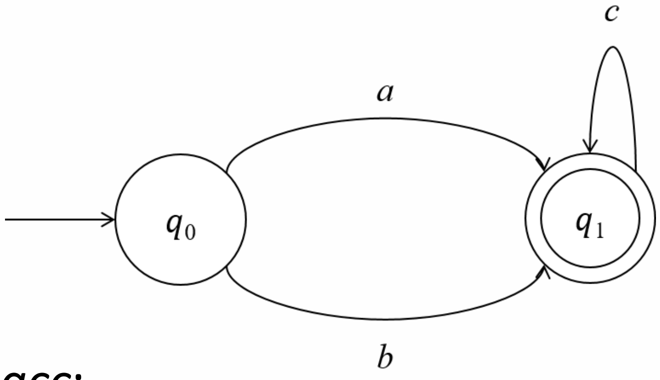
Automata Theory

- The theory of automata is a study of abstract machines or automata where the computational problems can be solved using the machines.
- Automaton is an abstract model of machine that performs computation on inputs by moving through a series of states according to the transition functions and stops moving at the final states.

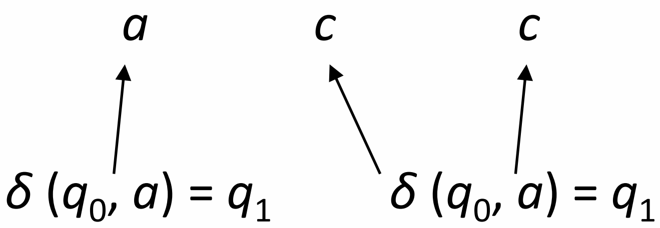
Symbols	**Description**
○	State
◎	Final State
→	Transition

Automata Theory (cont.)

- The automaton reads the symbols of input one after another and transitions from state to state until a string is formed.



- Derivation of a string *acc*:



a *c* *c*

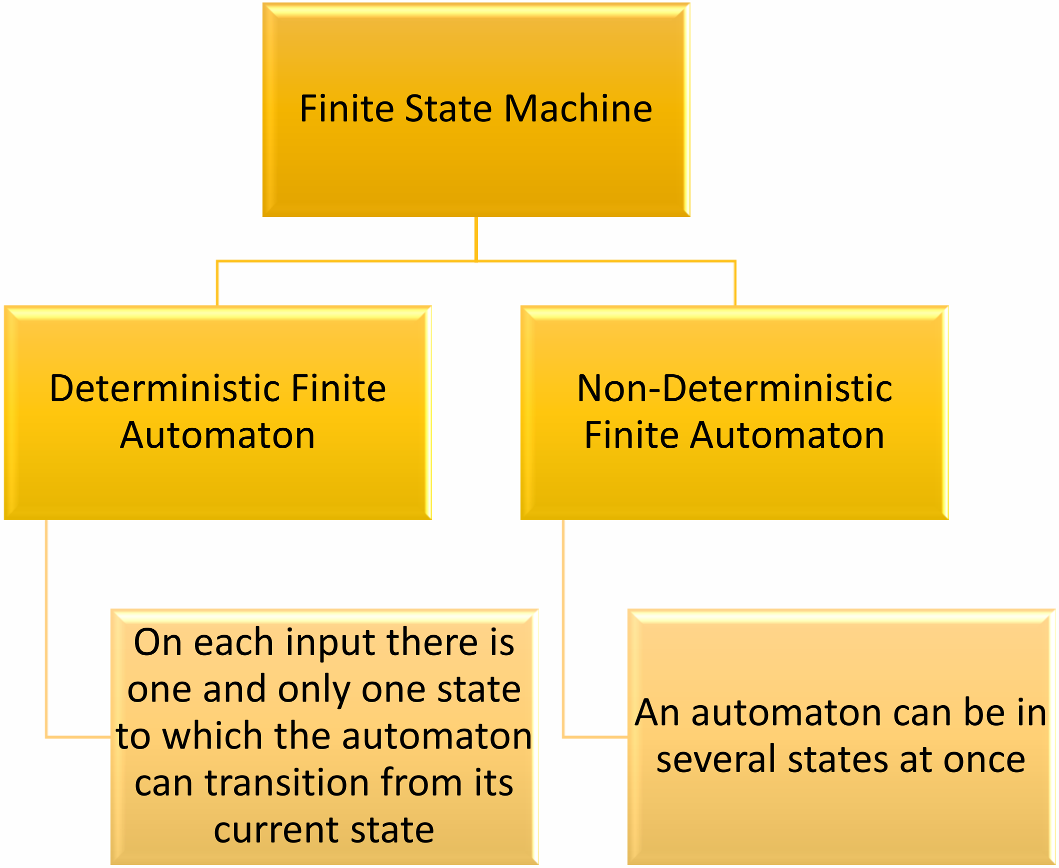
$\delta\ (q_0, a) = q_1$ $\delta\ (q_0, a) = q_1$

- The language *L* accepted by the automaton is the set of all strings formed by the automaton.

$L = \{a, ac, acc, \ldots, b, bc, bcc, \ldots\}$

Automata Theory (cont.)

- An automaton that contains only a finite number of states is called a finite state machine.



Finite State Machine

Deterministic Finite Automaton

Non-Deterministic Finite Automaton

On each input there is one and only one state to which the automaton can transition from its current state

An automaton can be in several states at once

Automata Theory (cont.)

Definition 23 (Linz, 2006) 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 $\delta : Q \times \Sigma \rightarrow Q$, an initial state $q_0 \in Q$ and a set of final states $F \subseteq Q$.

Definition 24 (Linz, 2006) Non-Deterministic Finite Automaton

A non-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 $\delta : Q \times \Sigma \rightarrow P(Q)$ where $P(Q)$ denotes the power set of Q, an initial state $q_0 \in Q$ and a set of final states $F \subseteq Q$.

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

Automata for DNA Splicing Systems

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

Author	Description
Gatterdam 1992	The associated automaton for crossing disjoint, reduced, permanent splicing system.
Bonizzoni et al. 1998	The properties of splicing systems and finite state automata
Fong et al. 2008	Non-deterministic finite automaton diagrams for simple splicing systems
Khairuddin et al. 2019	Deterministic finite Automaton and its application for splicing system
Ismail et al. 2020	The generalisations of splicing languages in deterministic finite automata

Gatterdam, R. W. (1992). Algorithms for Splicing Systems. *SIAM Journal on Computing, 21*(3), 507-520.

Bonizzoni, P., Ferretti, C., Mauri, G., & a di Milano-Bicocca, D.-U. (1998). Splicing Systems with Marked Rules. *Romanian Journal of Information Science and Technology, 1*(4), 295-306.

Fong, W. H., Sarmin, N. H., & Ibrahim, Z. (2008). Recognition of Simple Splicing Systems using SHAutomaton. *Malaysian Journal of Fundamental and Applied Sciences, 4*(2), 337-342.

Khairuddin, S., Ahmad, M., & Adzhar, N. (2019). *Splicing System in Automata Theory: A Review.* Paper presented at the 2nd International Conference on Applied & Industrial Mathematics and Statistics, Kuantan, Pahang, Malaysia.

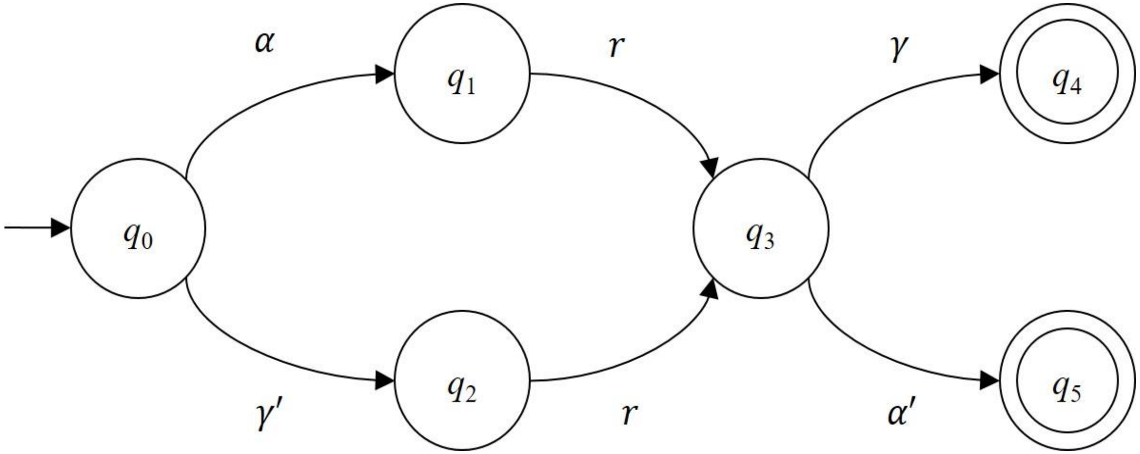
Ismail, N. I., Fong, W. H., & Sarmin, N. H. (2020). Automata for DNA Splicing Languages with Two Restriction Enzymes. *ASM Science Journal, 13*, 1-7.

Automaton for Simple Splicing System with One Cutting Site of a Palindromic Rule

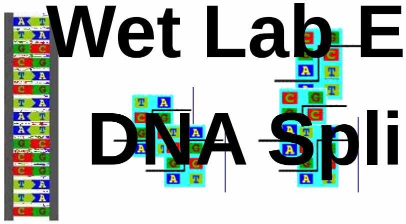
Theorem 1

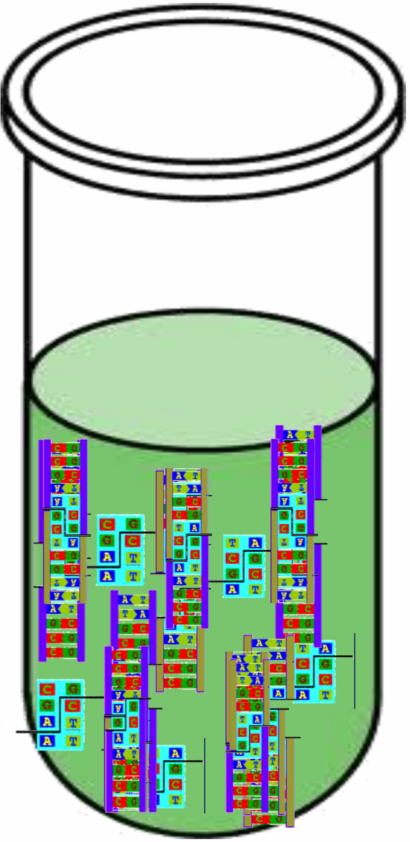
Given $S = (A, \alpha r \gamma, r)$ is a simple splicing system involving a cutting site of a palindromic rule r where $A = \{\alpha, r, \gamma\}$ are the set of variables used to denote any arbitrary dsDNA string, and α', r' and γ' are α, r and γ respectively after 180° rotation, $M = (Q, \Sigma, \delta, q_0, F)$ 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

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



Wet Lab Experiment for DNA Splicing System





$$S=(A,\ I,\ B,\ C)$$

Previous Molecular Works on Splicing Systems

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

Laun, E., & Reddy, K. J. (1999). *Wet Splicing Systems.* Paper presented at the 3rd DIMACS Workshop on DNA Based Computers, University of Pennsylvania, Philadelphia.

Fong, W. H. (2008). *Modelling of Splicing Systems using Formal Language Theory.* (Ph.D. Thesis), Universiti Teknologi Malaysia, Skudai, Malaysia.

Karimi, F. (2013). *Mathematical Modelling of Persistent Splicing Systems in DNA Computing.* (Ph.D. Thesis), Universiti Teknologi Malaysia, Johor, Malaysia.

Yusof, Y., Lim, W. L., Goode, T. E., Sarmin, N. H., Heng, F. W., & Wahab, M. F. A. (2015). *Molecular Aspects of DNA Splicing System.* Paper presented at the International Conference on Mathematics, Engineering and Industrial Applications 2014 (ICoMEIA 2014), Penang, Malaysia.

Ahmad, M. A., Sarmin, N. H., Abdul-Wahab, M. F., Heng, F. W., & Yusof, Y. (2018). Biomolecular Aspects of Second Order Limit Language. *Malaysian Journal of Fundamental and Applied Sciences, 14*(1), 15-19.

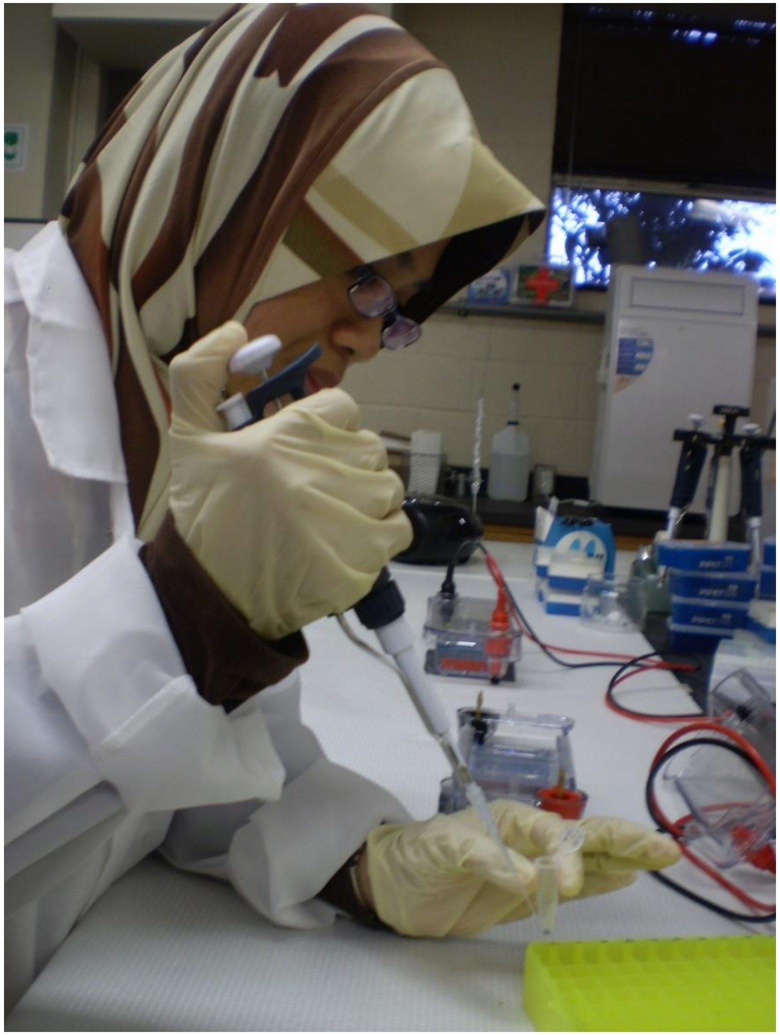
Wet Lab Experiment

Visit to wet lab, SUNY Binghamton, USA, 2007

Wet Lab Experiment (cont.)

Universiti Teknologi Malaysia, 2007

Wet Lab Experiment (cont.)



Towson University, USA, 2010

Wet Lab Experiment (cont.)

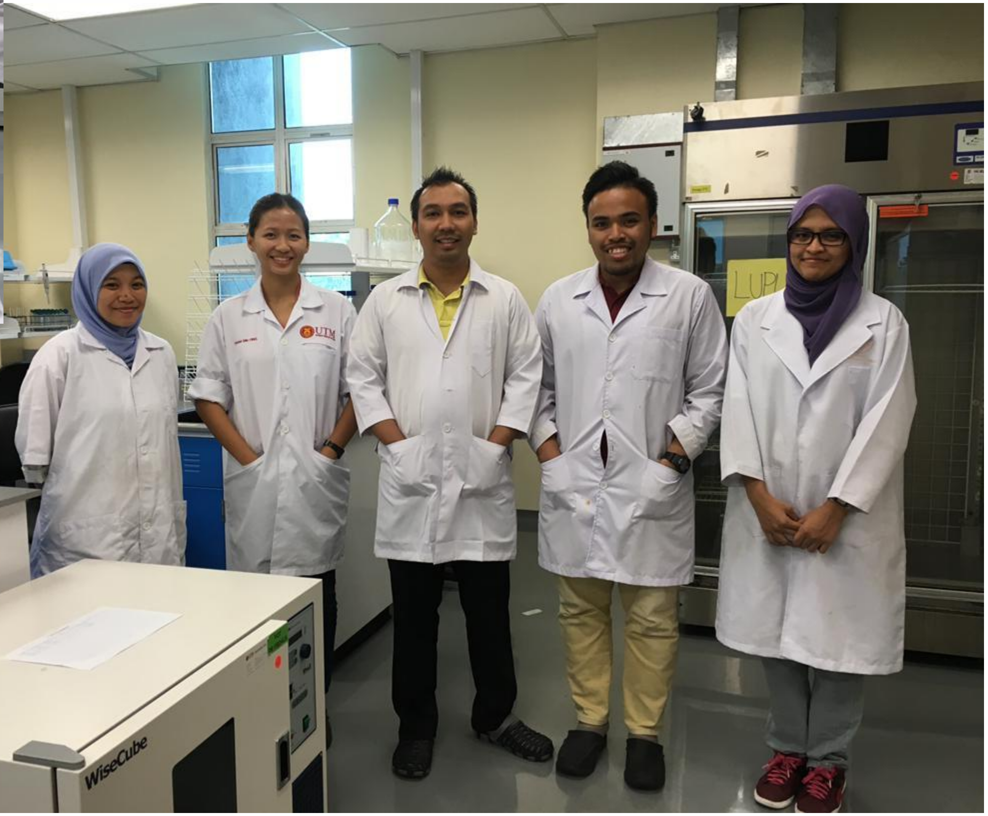
Research collaboration with Towson University, USA, 2010

Wet Lab Experiment (cont.)

Universiti Teknologi Malaysia, 2012

Wet Lab Experiment (cont.)

Universiti Teknologi Malaysia, 2015



Wet Lab Experiment (cont.)

Universiti Teknologi Malaysia, 2020

Laboratory Procedure

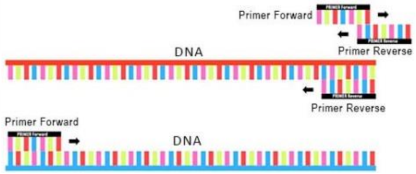
Polymerase chain reaction (PCR)	• To make several copies of a specific DNA segment
Process of Restriction Enzyme Digestion and Ligation	• The restriction enzymes recognize specific restriction sites in DNA molecules • The restriction enzymes and ligase then cut and rejoin the molecules to generate further molecules
Polyacrylamide gel electrophoresis (PAGE)	• to separate proteins based on their molecular weight

Polymerase Chain Reaction (PCR)

PCR machine

PCR tube

Lambda DNA

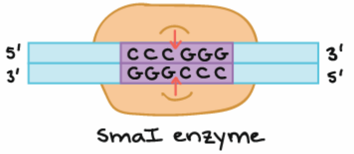


Primers

DNA Polymerase
(PCR reagent)

Process of Restriction Enzyme Digestion and Ligation

PCR Product



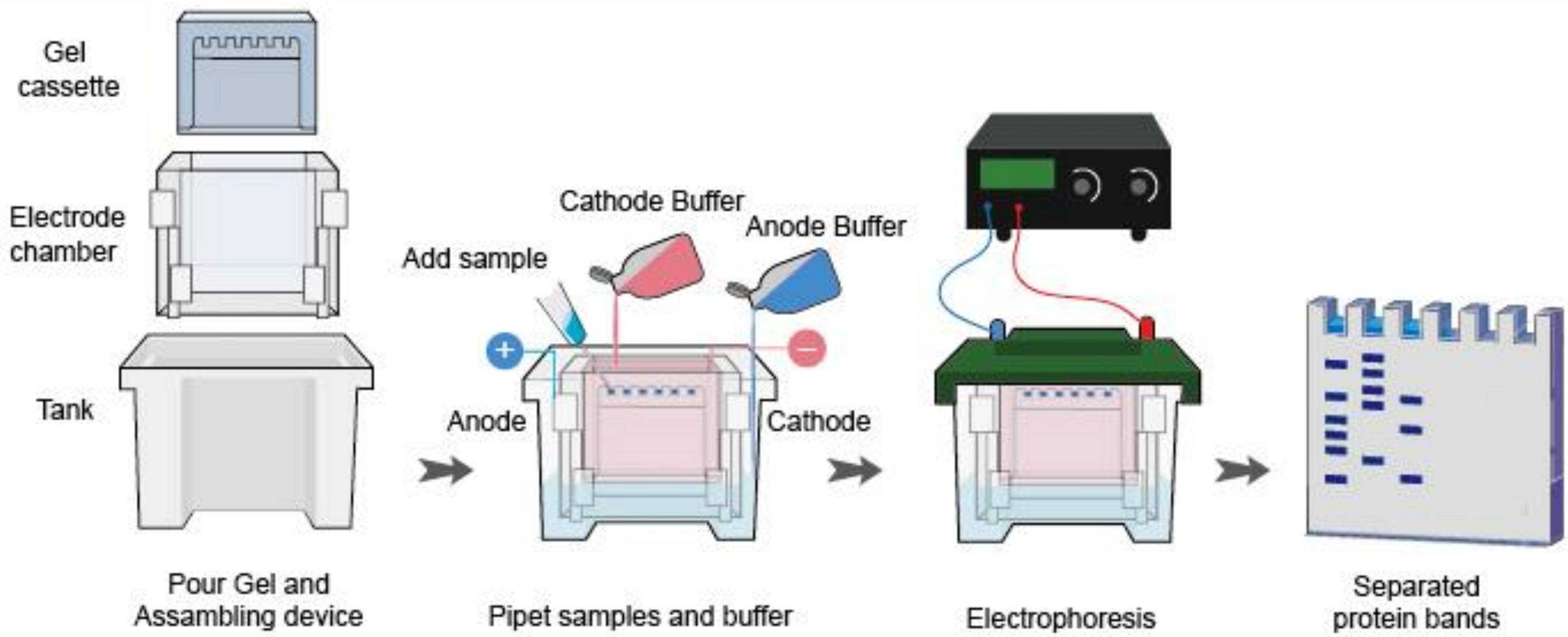
Restriction Enzyme

Ligase and Buffer

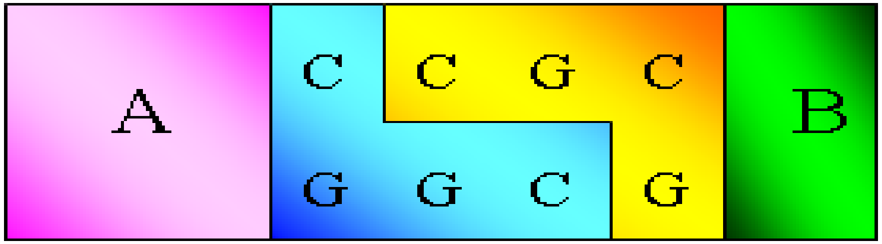
Thermomixer

Microcentrifuge Tube

Polyacrylamide gel electrophoresis (PAGE)

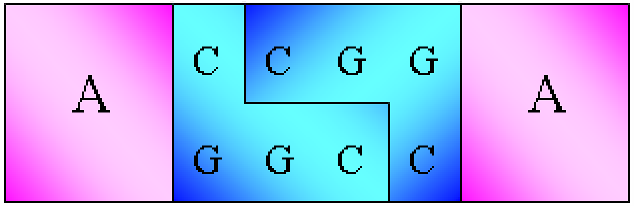


Example of a wet lab experiment



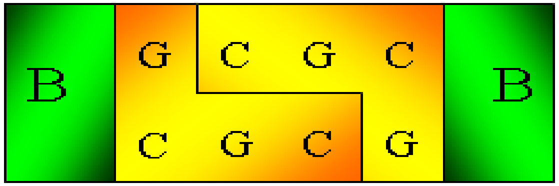
75 bp

Limit Language



90 bp

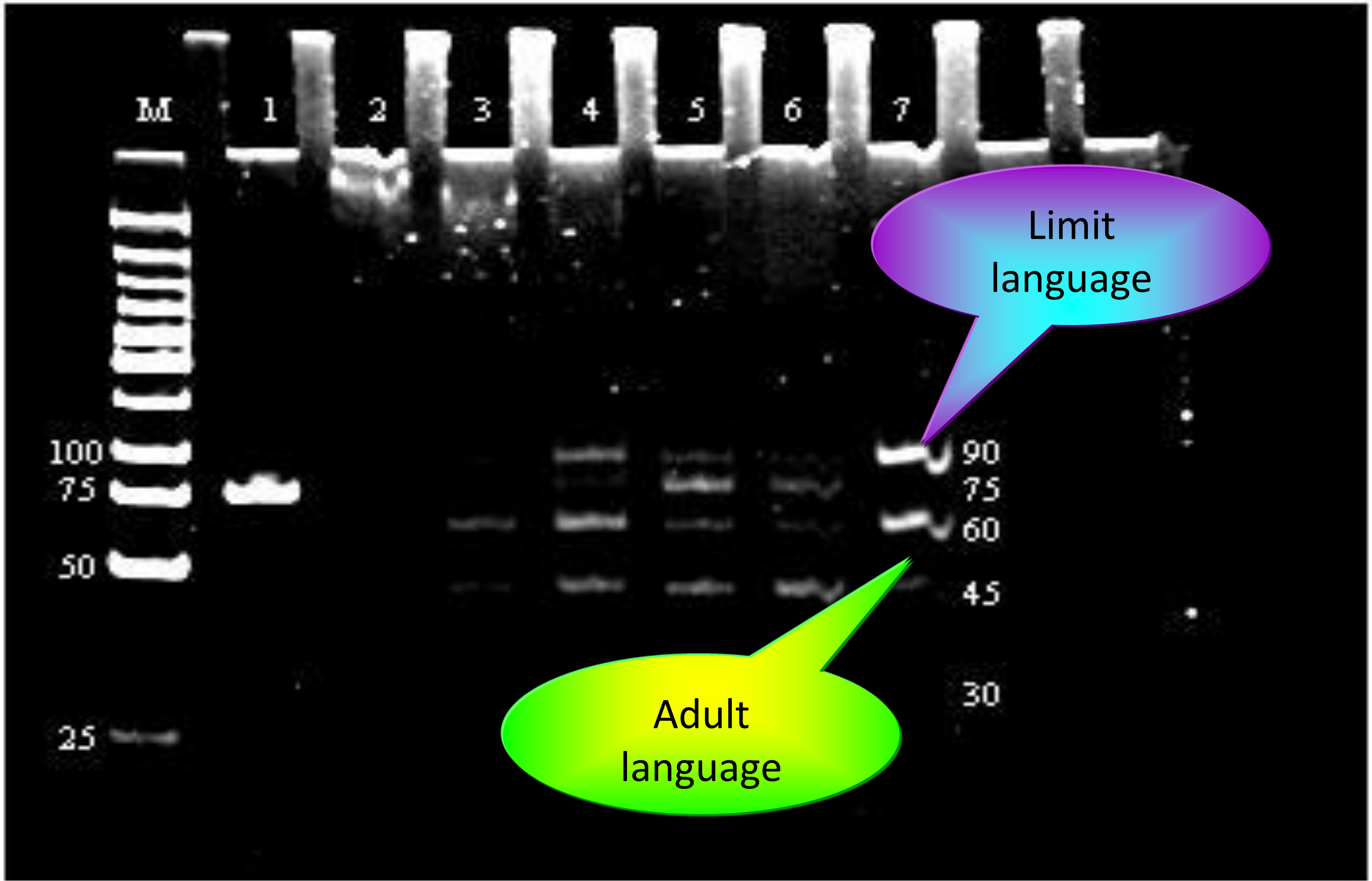
Adult Language



60 bp

Example of a gel photo

LMW Marker	Uncut 75bp	0 min	5 min	10 min	20 min	45 min	overnight



Limit language

Adult language

Experimental Design for DNA Splicing System involving Enzymes *Cvi*QI and *Aci*I

- 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.

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

$|\alpha|$ = 35 bp

$|\beta|$ = 44 bp

$|\gamma|$ = 81 bp

$|\alpha - \beta|$ = 79 bp

$|\beta - \gamma|$ = 125 bp

Experimental Design for DNA Splicing System involving Enzymes *Cvi*QI and *Aci*I (cont.)

- Initial Molecule (42958 and 43117 from bacteriophage lambda)

ggactatcgaagagtgcaaggcgatcaaggcagagtaccaacagaaactcaaagacctgcgaaatagcagaagtgaggccgcatgacgttctcagtaaaaaccattccagacatgctcgttgaaacatacggaaatcagacagaagtagcacgcagactg (160 bp)

- Enzyme *Cvi*QI (palindromic)

5’…G▼TAC…3’

3’…CAT▲G…5’

- Enzyme *Aci*I (non-palindromic)

5’…C▼CGC…3’

3’…GGC▲G…5’

DNA Splicing System involving *Cvi*QI and *Aci*I

The splicing language from this splicing system S 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:

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

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

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

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

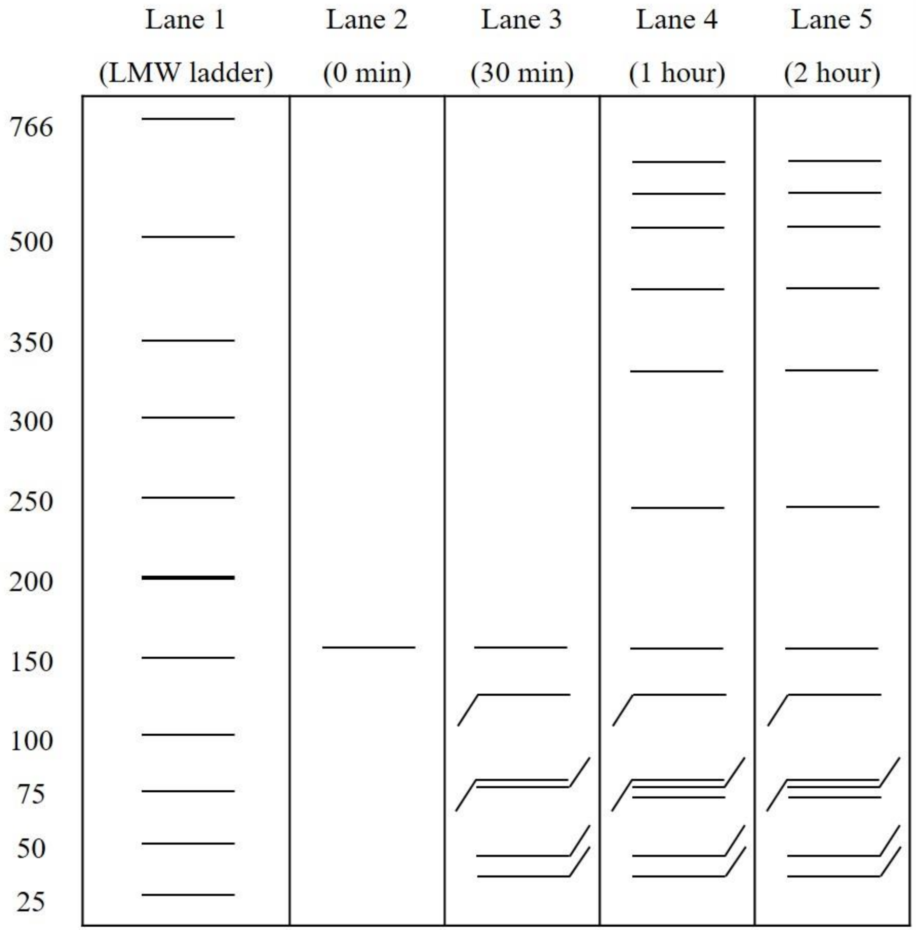
$\gamma'\, \underline{gcg} \left(\underline{g}\, \beta'\, \underline{gtac}\, \beta\, \underline{ccg}\right)^{n-1} \underline{c}\, \gamma$ and

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

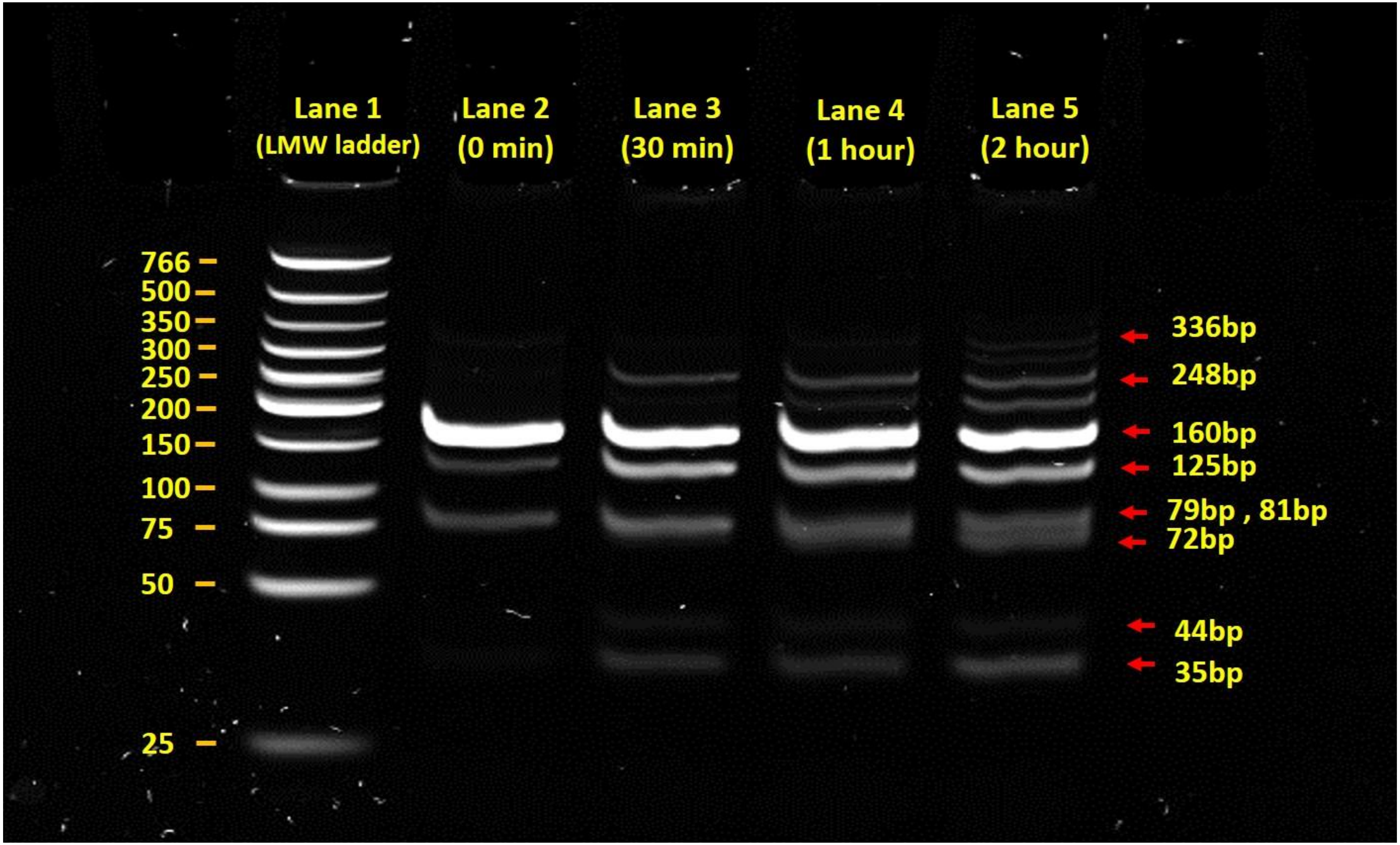
where $n \in \mathbb{Z}^+$ represents multiple copies of the specific strings.

DNA Splicing System involving *Cvi*QI and *Aci*I (cont.)

No.	Molecule	Size (bp)
1.		35
2.		44
3.		81
4.		37
5.		44
6.		79
7.		79
8.		125
9.		160
10.		72, 160, 248, 336,
11.		160, 248, 336, 424,
12.		160, 248, 336, 424,
13.		160, 248, 336, 424,



DNA Splicing System involving *Cvi*QI and *Aci*I (cont.)



List of Publications

Link: https://people.utm.my/nizasarmin/

2010

Nor Haniza Sarmin, Yuhani Yusof and Fong Wan Heng, Some Characterizations in Splicing Systems, *International Conference on Mathematical Sciences (ICMS 2010)*, Abant Ỳzzet Baysal Üniversitesi, Bolu, Turkey, 23 – 27 Nov 2010, American Institute of Physics (AIP) Conference Proceedings, Melville, New York, Vol 1309, pg. 411-418, (ISBN 978-0-7354-0863-0).

2012

Sherzod Turaev, Gan Yee Siang, Mohamed Othman, Nor Haniza Sarmin and Fong Wan Heng, Weighted Splicing Systems, Computational Intelligence and Intelligent Systems, Communications in Computer and Information Science (CCIS), ISBN 978-3-642-34288-2, Volume 316, 2012, pg 416-424, Proceedings in *The 6th International Symposium on Intelligence Computation and Applications (ISICA 2012)*, Wuhan, China, 27-28 October 2012.

2013

Nurhidaya Mohamad Jan, Fong Wan Heng and Nor Haniza Sarmin, Regular Languages, Regular Grammars and Automata in Splicing Systems, *Proceedings of the 20th National Symposium on Mathematical Sciences (SKSM 20)*, AIP Conf. Proc., Vol 1522, 2013, pg. 856-863 (ISSN: 1551-7616).

Yuhani Yusof, Nor Haniza Sarmin, Fong Wan Heng, T. Elizabeth Goode and Muhammad Azrin Ahmad, An Analysis of Four Variants of Splicing System, *Proceedings of the 20th National Symposium on Mathematical Sciences (SKSM 20)*, AIP Conf. Proc., Vol 1522, 2013, pg. 888-895 (ISSN: 1551-7616).

Mathuri Selvarajoo, Fong Wan Heng, Nor Haniza Sarmin and Sherzod Turaev, Some Characteristics of Probabilistic One-Sided Splicing Systems, *Proceedings of the 20th National Symposium on Mathematical Sciences (SKSM 20)*, AIP Conf. Proc., Vol 1522, 2013, pg. 967-975(ISSN: 1551-7616).

Sherzod Turaev, Mathuri Selvarajoo, Mohd Hasan Selamat, Nor Haniza Sarmin and Fong Wan Heng, Probabilistic Splicing Systems, Advanced Methods for Computational Collective Intelligence, Studies in Computational Intelligence, ISBN 978-3-642-34300-1,Volume 457, 2013, pg 259-268, proceedings in *4th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2012)*, Ho Chi Minh city, Vietnam, 28-30 November 2012.

2014

Muhammad Azrin Ahmad, Nor Haniza Sarmin, Fong Wan Heng, Yuhani Yusof, An Extension of First Order Limit Language, *Proceedings of the 3rd International Conference on Mathematical Sciences (ICMS3)*, AIP Conf. Proc., Vol 1602, 2014, pg. 627-631. (ISBN: 978-0-7354-1236-1)

Mathuri Selvarajoo, Fong Wan Heng, Nor Haniza Sarmin and Sherzod Turaev, Probabilistic Simple Splicing Systems, *Proceedings of the 3rd International Conference on Mathematical Sciences (ICMS3)*, AIP Conf. Proc., Vol 1602, 2014, pg. 760-766. (ISBN: 978-0-7354-1236-1)

Yuhani Yusof, Nor Haniza Sarmin, Fong Wan Heng, The Concepts of Persistent and Permanent in Non Semi-Simple DNA Splicing System, *Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM 21)*, AIP Conf. Proc., Vol 1605, 2014, pg. 586-590 (ISBN: 978-0-7354-1241-5).

Fariba Karimi, Sherzod Turaev, Nor Haniza Sarmin and Wan Heng Fong, Fuzzy Splicing Systems, Lecture Notes in Artificial Intelligence, Computational Collective Intelligence: Technologies and Applications, Volume 8733, ISBN 978-3-319-11288-6, pg 20-29, proceedings in 6th *International Conference on Computational Collective Intelligence Technologies and Applications* (ICCCI 2014), Seoul, Korea, September 24-26, 2014.

List of Publications (cont.)

Link: https://people.utm.my/nizasarmin/

2015

Yuhani Yusof, Wen Li Lim, T.Elizabeth Goode, Nor Haniza Sarmin, Fong Wan Heng, Mohd Firdaus Abd Wahab, Molecular Aspects of DNA Splicing Systems, International Conference on Mathematics, Engineering & Industrial Applications 2014 (ICoMEIA 2014), 28-30 May, 2014, The Gurney Resort Hotel & Residences Penang, AIP Conf. Proc., Vol. 1660, 2015, pg. 050045 (ISSN: 1551-7616).

Yee Siang Gan, Wan Heng Fong, Nor Haniza Sarmin and Sherzod Turaev, Some Characteristics on the Generative Power of Weighted One-Sided Splicing Systems, *Proceedings of the 22nd National Symposium on Mathematical Sciences (SKSM 22)*, AIP Conf. Proc., Vol 1682, 2015, pg. 020044 1-7 (ISBN 978-0-7354-1329-0).

Muhammad Azrin Ahmad, Nor Haniza Sarmin, Yuhani Yusof and Wan Heng Fong, Some Restrictions on the Existence of Second Order Limit Language, *Proceedings of the 22nd National Symposium on Mathematical Sciences (SKSM 22)*, AIP Conf. Proc., Vol 1682, 2015, pg. 020048 1-7 (ISBN: 978-0-7354-1329-0).

2016

Wan Heng Fong, Yee Siang Gan, Nor Haniza Sarmin, and Sherzod Turaev, The Generative Capacity of Weighted Simple and Semi-Simple Splicing Systems, *Proceedings of the 23rd National Symposium on Mathematical Sciences (SKSM 23)*, AIP Conf. Proc., **Vol 1750, 2016**, pg. 050013 1-6 (ISBN: 978-0-7354-1407-5).

2017

Mathuri Selvarajoo, Wan Heng Fong, Nor Haniza Sarmin and Sherzod Turaev, Probabilistic Simple Sticker Systems, *Proceedings of the 4th International Conference on Mathematical Sciences* (ICMS4 2016), AIP. Conf. Proc., 1830 (1), 2017, pg. 020057 1-9 (ISBN: 978-0-7354-1498-3)

2018

Nurul Izzaty Ismail, Wan Heng Fong and Nor Haniza Sarmin, Computation of Splicing Languages from DNA Splicing System with One Palindromic Restriction Enzyme, Malaysian Journal of Fundamental and Applied Sciences 2018, 14(2): pg 188-192, (ISSN 2289-5981)

Muhammad Azrin Ahmad, Nor Haniza Sarmin, Mohd Firdaus Abdul-Wahab, Fong Wan Heng, Yuhani Yusof, Biomolecular Aspects of Second Order Limit Language, Malaysian Journal of Fundamental and Applied Sciences 2018, 14(1): pg 15-19, (ISSN 2289-5981).

2019

Wan Heng Fong, Nurul Izzaty Ismail and Nor Haniza Sarmin, Automata for DNA Splicing Languages with Palindromic and Non-Palindromic Restriction Enzymes using Grammars, MATEMATIKA: *Malaysian Journal of Industrial and Applied Mathematics,* December 2019, Special Issue: pg 1-14, (eISSN: 0127-9602)

Nurul Izzaty Ismail, Wan Heng Fong and Nor Haniza Sarmin, DNA Splicing Systems with at Most Two Cutting Sites of a Non-Palindromic Restriction Enzyme, MATEMATIKA: *Malaysian Journal of Industrial and Applied Mathematics,* 2019, 35(2): pg 129-137, (eISSN: 0127-9602)

Thank You

PROF. DR. NOR HANIZA SARMIN

https://people.utm.my/nizasarmin/

nhs@utm.my

Niza Sarmin

Niza Sarmin