School of Computing Science

# Investigating routes for *in vitro* and *in vivo* data storage

*Annunziata Lopiccolo, MSc*

Submitted for the degree of Doctor of Philosophy in the School of Computing Science, Newcastle University

September 2016

# Declaration of Authorship

I, Annunziata Lopiccolo, declare that this thesis titled, "Investigating routes for *in vitro* and *in vivo* data storage" are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

*"The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to the sky, would not be enough for the investigation of so vast a subject... And so this knowledge will be unfolded only through long successive ages. There will come a time when our descendants will be amazed that we did not know things that are so plain to them... Many discoveries are reserved for ages still to come, when memory of us will have been effaced"*

Seneca

*To my brother Luca,*

*living in every ray of light.*

# Abstract

Computing Science, Synthetic Biology and Nanotechnology are converging. Synthetic Biology and Nanotechnology compose the “hardware” platform, whilst Computing Science formulates the logic, data storage and processing pipelines in order to create complex yet controlled behaviour at the nanoscale.
Although much work has been done on information processing at the nanoscale via *in vivo* constructs, *e.g.* logic gates in various organisms, relatively little has been done on implementing data structure, a fundamental building block for computation.
This dissertation proposes and investigates methods to implement data structures by employing biological molecules via both a Synthetic Biology and a Nanotechnological approach. A data structure implemented at the nanoscale could help to substantially increase the complexity of behaviours that could be programmed and embedded in living cells or at the interface between living cells and other nano-substrates, with potential applications in intelligent drug factories and delivery nanosystems, biosensors, and environmental cleaning bionanotechnologies.

This work explores the possibility of implementing via DNA constructs, both *in vitro* and *in vivo*, "list-like" data structure that can potentially hold an unlimited number of items. This has not been achieved before. Thus, the text describes designs and test prototypes.

Firstly, this thesis focuses on an *in vitro* approach. This is achieved through a DNA-based machinery implementing a signal recorder based on DNA strand displacement reactions. Such DNA architecture can in principle implement a stack machine, capable of *storing data* providing a dynamic temporary memory capable of pushing and popping data-items encoded in DNA nanostructures (called DNA "bricks”). The "list-like" data is thus represented by a growing (or shrinking) chain of DNA bricks.

Secondly, I introduce a potential design and initial experiments for an *in vivo* approach presenting, a synthetic genetic circuit designed to record and accumulate extracellular signals digitally within a "*tape*" DNA molecule inside a living cell. The core is based on the engineering of the self-splicing group II retrotransposon *Ll*.LtrB of *Lactococcus lactis*.

Together, these two *in vitro* and *in vivo* routes expand our knowledge in the context of molecular memory devices and the biological operations we can compute.

# Acknowledgements

This thesis could not have been written without substantial assistance from many people. Here I am pleased to thank my supervisor, Prof. Natalio Krasnogor, I am deeply grateful and owe many thanks to him for his excellent guidance, scientific approach, sincere encouragement through all the hard work, and most importantly for believing in me.

I am also very grateful to Dr. Harold Fellermann for his expertise, knowledge, consistent motivation, and generous guidance that helped me during all my research at Newcastle University.

I wish to thank Dr. Stephan Heeb, my second supervisor whilst I was at the University of Nottingham, for his research directions, discussions and advices for the *in vivo* prototype presented in this thesis. My sincere thanks goes also to Dr.Birgit Koch, for her immense kindness and Dr. Pawel Widera, Dr. Omer Markovitch and Dr. Ben Shirt-Ediss for their conversations and practical advices.

A very special thank goes to my colleague and dear friend Jerzy Kozyra, providing support and friendship through all these years of interdisciplinary research. Appreciation also goes out to Dr. Daven Sanassy, to his help and our long conversation and fraternal words of encouragement. Thanks to my closest friends Dr. Maria Franco for her friendship without borders, Nicola Lazzarini, Jonathan Naylor, Charles Winterhalter and Alessandro Ceccarelli for their support and encouragement through the hard work.

Many friends helped me to stay strong through these important years. Their support and care helped me to stay focused and enjoy my life in England. I greatly value their friendship and their support during my weakest moments.

Special thank to my family, my heart-felt gratitude goes to them, my constant source of love, none of this would have been possible without their sacrifices and patience.

Finally, I wish to thank the financial support from the EPSRC funded Synthetic Biology

project: “AUDACIOUS: Towards a Universal Biological-Cell Operating System” (EP/J004111/1 & EP/J004111/2) that funded the research discussed in this thesis.

# Contents

# List of Figures

# List of Tables

# Acronyms

**5-FOA** 5-Fluoroorotic Acid

**AB** Agrobacterium Minimal Media

**AS** Anti-Start

**att** Attachment Site

**Au** Gold

**BHQ** Black Hole Quencher

**BLAST** Basic Local Alignment Search Tool

**bp** Base Pair

**CCR** Carbon Catabolite Repression

**Cro** Causes Recombination Protein

**CRNs** Chemical Reaction Networks

**Cy3** Cyanine 3

**DDS** Digital Data Storage

**DNA** Deoxyribonucleic Acid

**dsDNA** Double Stranded DNA

**EBS** Exon Binding Site

**EDTA** Ethylene Diamino Tetraacetic Acid

**FLP** Flippase

**FRT** Flippase Recognition target

**IBS** Intron Binding Site

**IEP** Intron Encoded Protein

**int** Integrase

**IPTG** Isopropyl Beta D Thiogalactopyranoside

**JCVI** The J.Craig Venter Institute

**LB** Luria Bertani media

**Lox** Locus Of X-over

**M9** Minimal Media for E.coli

**NCBI** National Centre for Biotechnology Information

**nM** Nano Molar Concentration

**nt** Nucleotides

**OD** Optical Density

**PAGE** Denaturing Polyacrylamide Gel Electrophoresis

**PBS** Phosphate Buffered Saline

**PCR** Polymerase Chain Reaction

**PLy** Push-Write Ly

**PX** Push-Write X

**PY** Push-Write Y

**PZ** Push-Write X or Y **QP** Pop-Push

**RBH** Reciprocal Best Hits

**RMCE** Recombinase-mediated Cassette Exchange

**RPMI** Roswell Park Memorial Institute Media

**RNA** Ribonucleic Acid

**RNP** Ribonucleoprotein Complex

**RPM** Ribonucleoprotein Complex

**RLy** Read-Write Ly

**RX** Read-Write X

**RY** Read-Write Y

**RZ** Read-Write X or Y

**SB** Synthetic Biology

**SP** Start-Push

**SPLy** Start-Push-Write Ly

**SPX** Start-Push-Write X

**SPY** Start-Push-Write Y

**SPZ** Start-Push-Write X or Y

**ssDNA** Single Stranded DNA

**SSR** Specific Site Recombinase

**TAE** Tris-Acetate-EDTA buffer

**TEG** Tetra-Ethyleneglycol

**TEM** Transmission Electron Microscope

# Part I

# Chapter 1

# Introduction

This chapter includes an introduction on the importance of my research, the motivation behind, objectives and contributions to the topic. My research questions have been addressed at the interface between Computing Science, Nanotechnology and Synthetic Biology, to investigate routes for *in vitro* and *in vivo* data storage. Besides, I will include a brief description of the thesis structure and outline, providing a focus on the way that the research tasks have been carried out.

## 1.1 Background and Motivation

When we think about data storage, our minds naturally envision disks, electronic microchips and other technologies that have only existed for less than a century. Over the last few decades, there have been great leaps in data storage media, as we have moved from floppy disks to CDs (which could store many floppy discs worth of data), to DVD, Blu-ray disks and beyond. The technology sector is always endeavouring to improve the next generation of data storage.

Nature has employed biochemical data storage for billions of years to encode the complexity of life, with enough dynamism to support the process of biological evolution. This data storage is realised by the famous deoxyribonucleic acid (*DNA*) molecule and has proven to be a medium that is highly stable, resilient, capable of replication and long-term information storage [1]. Furthermore, with regard to data density, DNA that encodes the entire human

genome fits within a single biological cell.

Digital data storage (DDS) [2] represents a method to store data in a steady and stable way on a storage device, providing a location and a memory for data content. Both computational and genetic data storage are analogous in nature; DNA is a digital storage medium that is quaternary instead of binary. The exponential growth of digital data means researchers are continuously searching for a stable and durable archiving solution. DNA enables a theoretical storage capacity maximum of two bits per nucleotide or 455 exabytes per gram of single-stranded DNA [3]. Another reason to use the "double helix" as storage media is its inherent stability governed by specific hydrogen bonding patterns as well as a natural "backup" of genetic information from DNA strand complementarity. In the scientific state-of-the-art, there is evidence that DNA is an optimal substrate to support both technological applications *in vitro* and *in vivo*.

Going forward, macromolecules represent an important opportunity for advances in data storage capability, allowing complex molecular architectures to act as an information device with the ability to change configuration and switch between states in response to a signal [4, 5].

Furthermore, controllable devices can be crafted using molecular units, with addressable input and readable output [6]. Achieving these goals would give one the resources to build efficient and reliable molecular memory storage in the vision of a DNA "tape". This DNA media would be far more compact and robust whilst having significantly greater storage capability than its antiquated electromagnetic predecessor. Despite all these advantages, data access speed still remains a disadvantage. A magnetic tape takes just few seconds to read the stored informations, whereas storing data using DNA has a high cost in terms of retrieving the information stored (*e.g., sequencing)*. But still DNA is very remarkable in terms of capability. But crucially, the ultimate aim of achieving reliable data structures with DNA (and RNA) is to be able to implement complex computations *in vivo* or at the interface between living and non-living matter that could process information based on previously stored data states (something that has not yet been achieved to any meaningful level).

Biological implementation of primitive data structure models demands biological self-organization, and required controllable and predictable behaviours, achieved either *in vivo* or *in vitro* via manipulating macromolecules.

## 1.2 Aims and Scope

My overarching goal in this dissertation is to advance the state of the art in molecular data structures. This goal is translated into four objectives:

I Designing a "list-like" prototype for implementing an *in vitro* data structure. As a demonstrator, I will use a stack data structure.

II Designing a "list-like" prototype for implementing a structure for *in vivo* operations. As a demonstrator, I will use a tape data structure.

III Engineering both prototypes in the laboratory.

IV Testing and assessing limitations of the design in objective I and objective II

This thesis presents different methodologies to investigate two different routes for DNA data storage structures. In computer science, a *list* data structure is an ordered sequence of items, that allows to store,access and change elements. To work towards these goals, I present a *list-like* DNA data storage structures (allows to store but not to change and access any element), where one is in the form of a *stack*, for the *in vitro* approach and the second one is in the form of a *tape*, for an *in vivo* application.

The objective of this thesis is to introduce an original and feasible *in vitro* & *in vivo* prototype for DNA digital data storage. The first avenue explored in this work investigates the *in vitro* implementation of a DNA architecture designed to implement a stack machine. This is a design where kinetics and thermodynamics (tuned by the usage of a genetic algorithm) play a key role in the performance of a dynamic DNA memory, implemented via strand displacement [7]. The second route describes an innovative *in vivo* approach where the central mechanism of the genetic network is based on TargeTron technology [8][9], that can in principle hold an unlimited "list-like" data structure within a DNA molecule.

## 1.3 Main Contribution

The work presented in this thesis contributes to the EPSRC funded synthetic biology project: "AUDACIOUS: Towards a Universal Biological-Cell Operating System" (EP/J004111/1).

As part of this research work, this dissertation contributes to the area of Computing Science, Synthetic Biology, DNA computing and Nanotechnology, pointing to a novel thinking and routes towards *in vitro* and *in vivo* data storage. The interdisciplinary nature of this study contributes to expand knowledge and learning how to critically compare and combine approaches that can require different standards, identify problems, propose solutions, integrate literature, theoretical analysis, data analysis and drawing possible further research directions.

Specific contributions:

1. Design, implementation, testing and functional demonstration of an *in vitro* stack data structure.
2. Design, implementation and testing of an *in vivo* prototype *tape* data structure
3. New biological and nanotechnological materials and methods for others to replicate and expand this work.

This is a highly interdisciplinary work where I have conducted all the experimental work in its entirety. The specification of the stack data structure was done collaboratively with J. Kozyra who also contributed the computational analysis of the *in vitro* DNA data structure. Both directions presented in this scientific study can be seen as a critical and deep analysis into design methodologies and solutions directly addressing the rigorous optimization of flexible bio-constructs used as the basis for a versatile data storage system. This has resulted in the following papers being submitted:

- Annunziata Lopiccolo, Harold Fellermann, Jerzy Kozyra, Natalio Krasnogor "*In Vitro implementation of stack data structure based on DNA strand displacement*". International Conference on Unconventional Computation and Natural Computation (UCNC), Manchester, UK, July 11-15 2016.

- Ben Shirt Ediss, Harold Fellermann, Annunziata Lopiccolo, Natalio Krasnogor "*DNA Stack Recorder: In-Silico Insights into an In-Vitro Implementation*". Submitted to the 22nd International Conference on DNA Computing and Molecular Programming, Munich, Germany, September 4-8, at Ludwig-Maximilians-Universität (LMU).

- Harold Fellermann, Annunziata Lopiccolo, Jerzy Kozyra, Ben Shirt-Ediss, Natalio Krasnogor "A DNA-Based Signal Recorded Studied In Vitro and Simulation".2016 Conference on Systems Chemistry SYSCHEM'16, Valtice, Check Republic.

- Jerzy Kozyra, Alessandro Ceccarelli, Annunziata Lopiccolo, Jing-Ying Gu, Harold Fellermann, Ulrich Stimming, and Natalio Krasnogor. "Designing uniquely addressable bio-orthogonal synthetic scaffolds for DNA and RNA origami". (submitted)

## 1.4 Thesis Structure

The thesis is divided into five parts. Part I includes this and Chapter 2- "Background and State of the Art". It provides an overview of DNA Nanotechnology, where molecular programming achieved via DNA hybridization and DNA strand displacement are described. Moreover, this chapter presents an extended literature of Synthetic Biology and its usage of recombinases and a retrotransposon systems to implement new classes of synthetic circuits.
Part II includes:
Chapter 3 -"Specification and Design of an In Vitro DNA Stack Data Structure": firstly introduces the concept of a stack data structure, and, secondly reviews the *in vitro* approach, with a detailed description of the design, specifications and rationale behind them. Furthermore, it presents a full illustration and explanation of all the operations that the prototype can compute.
Chapter 4 - "Specification and Design of an In Vivo DNA Tape Data Structure": presents the *in vivo* approach, with a high-level description of the prototype circuit implementing the tape data structure, a more detailed explanation of the system components, design and specification of each bio-part and limitations of the approach and technologies involved. It concludes Part II.

Part III includes:

Chapter 5 -"Engineering of an In Vitro DNA Stack Data Structure" demonstrates the implementation of an *in vitro* DNA stack data structure via Hybridization and DNA Strand Displacement Cascades.

Chapter 6 -"Engineering of an In Vivo DNA Tape Data Structure" describes the *in vivo*, *in vitro* and *in silico* methods used the implementation of a tape data structure via a site-specific recombination approach. It concludes Part II.

Part IV includes:

Chapter 7 -" Characterization and Result of an In Vitro DNA Stack Data Structure" : presents the results obtained toward the final implementation of an *in vitro* DNA stack data structure. All the results have been obtained via the methods described in Chapter 5.

Chapter 8 -"Characterization and Result of an In Vivo DNA Tape Data Structure": presents the results obtained toward the preliminary implementation of an *in vivo* DNA tape data structure. All the results have been obtained via the methods described in Chapter 6 and conclude Part III.

Part V includes:

Chapter 9 -"Discussion and Conclusions": offers a discussion of the results obtained via the *in vitro* and *in vivo* approaches. It integrates and synthesizes the conclusions of this work.

Chapter 10 -" Future Research Directions ": offers a glimpse of possible future avenues of research derived from this work.

# Chapter 2

# Background and State of the Art

This chapter offers a detailed review of the literature, in both state of the art Nanotechnology and Synthetic Biology. It contains an overview of the main technologies and their recent applications. Additionally, it will fully describe the technologies that will be used to implement the stack and the tape data structures, presented in Part II.

## 2.1 Introduction

Creation of novel biological systems requires a deep understanding and insight into the operating principles that govern living organisms [10]. Innovation in DNA nanotechnology and synthetic biology aims to understand and expanding those principles. These disciplines are important because they also impinge on other areas such as production of biofuels, biomaterials, bioremediation and biomedicine [11]. Both are a platform for wholly new biological scenarios to emerge. They provide an *in vitro* and *in vivo* design space to build new biological functions and systems otherwise not found in nature.

## 2.2 DNA: Structure, Functionality and Operations

Biological systems are described as biochemical "machines", processing a genetic instruction encoded in the deoxyribonucleic acid (DNA) code, to the ultimate aim of replicate

themselves [12]. This molecule is a member of the nucleic acid family, which appears under the form of a polymer, where the "deoxyribose" is the principal sugar. This polymer contains two strands running in opposite directions and made of building blocks called nucleotides, which will intertwine to form a double helix. If we imagine unwinding the double helix, the strands chemical structure will appear as a repeating unit of nucleotides, where each nucleotide is made up of three components: a five carbon sugar, a phosphate group and one of four possible nitrogenous bases (adenine [A], guanine [B], thymine [T] and cytosine [C]) (Fig. 2.1) [13][14]

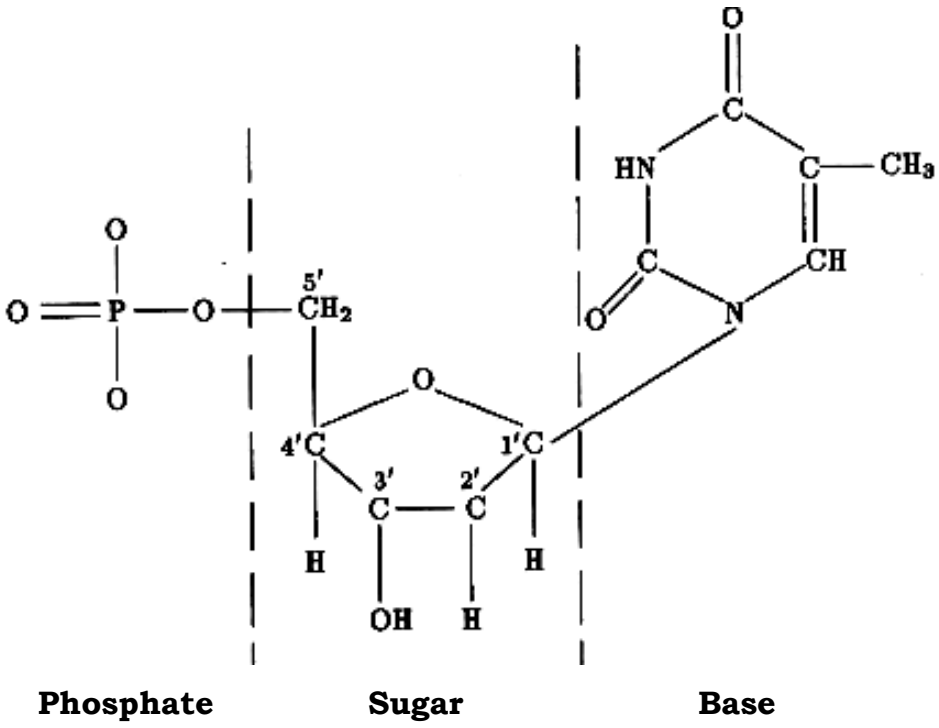



Figure 2.1: **The nucleotide chemical structure.** Image taken from [13] showing the three components of a nucleotide with thymine base.

The most common double helix conformation is known as B-DNA form, where one base pairs measures 0.34 nm and one turn of the helix, corresponds to 3.4 nm [15] as schematic represented in Fig 2.2.

Although nucleotides come together through covalent bonds in the backbone, the two DNA strands interact through non-covalent hydrogen bonds between the bases. Each base form

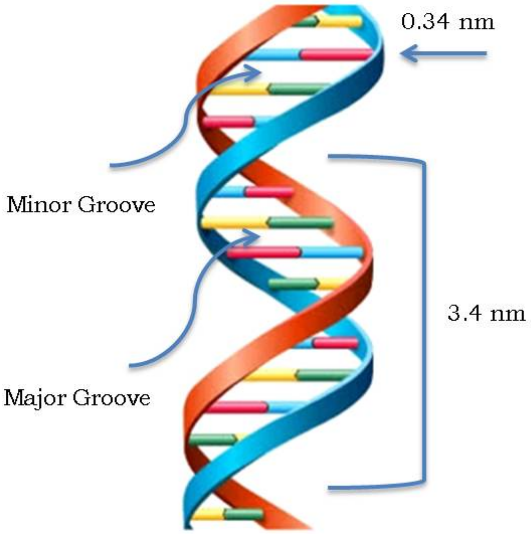



Figure 2.2: **DNA double helix B-Form**.(Image taken from Power Point and annotated). In the B-Helical structure there are two repeating and alternating spaces, the major and minor grooves. One base pairs measures 0.34 nm, whereas one turn of helix, corresponds to 3.4 nm (10 base pairs).

multiple hydrogen bonds with its complementary one on the opposite strand and with high specificity of base pairing (A-T, two hydrogen bonds, and G-C, three hydrogen bonds). Nucleotides attach to each other in the DNA strand by phosphodiester bonds. The phosphate group of one nucleotide binds to the 3'oxygen of the neighbouring nucleotide. Thus, the sugar and phosphate group make up the DNA backbone. The carbon number of the sugar is the key to describe the directionality of the DNA strand 5'$\rightarrow$ 3'. Looking within the sugar there is an intrinsic orientation difference between the two strands. On the top strand the 5' carbon of each sugar is on the left and 3' carbon is on the right. The opposite is true for the bottom strand. Reading left to right, makes the top strand orientation 5'$\rightarrow$ 3', and the bottom strand orientation 3'$\rightarrow$ 5'. The double helix structure is highly regular. Each turn of the helix measures approximately 10 base pairs. In addition to the hydrogen bonding between the bases, a phenomena known as DNA stacking, contributes to stabilizes the double helix structure [16]. The pi-clouds electrons of the aromatic rings of the base pairs, have an hydrophobic nature and interact via hydrophobic bounding, thus creating a pile, a "stack" of nucleobases [16, 17]. The geometry regularity of the helical structure forms two repeating and alternating spaces, called the major and minor grooves (Fig 2.2). These grooves act as a base pairs recognition and binding sites for proteins. They are not equal in terms of size and functionality. In the minor groove the sugar protrudes out from the base pairs with an

angle of 120°, whereas in the main grooves it does with an angle of 220°. The difference in the pattern and narrow angles, is crucial for proteins to recognise the DNA to bind. The minor grooves are less accessible to accommodate protein amino acidic chains, and a minority of non-sequence specific proteins can unwind a small tract of the double helix and bind to the minor groove. On the other side, the the major groove is more accessible to the majority of the proteins, that are instead sequence specific. The weak nature of DNA hydrogen bounds make possible to separate the two strands, and this process is known as "denaturation". When the hydrogen bonds break, the double helix is able to dissociate. The dissociation can be induced by heating a DNA solution and this will break the double helix structure. The breaking of the hydrogen bounds and helices dissociation is a process known as "DNA melting". The DNA melting occurs at specific temperature, called "melting temperature" (Tm), defined as the temperature at which half of the dissociation will happen. This process can be monitored by the amount of UV light absorbed by a DNA molecule, which increases with the helices dissociation. On the other side, decreasing the temperature of a DNA solution containing two separated strands, will re-associate or "hybridize" the strands, this process is known as "annealing". Another important reaction, is the "DNA ligation". A *ligase* protein is responsible for the formation of covalent bonds between adjacent fragments of DNA. Ligation, in molecular cloning represent the technique that allows to incorporate a DNA fragment (insert) into a plasmid (vector). The reaction catalyses the formation of a phosphodiester bond between the 5' phosphate (P) and 3' hydroxil (OH) adjacent groups. Going forward, the elongation of the DNA, is instead, performed by the DNA *polymerase*, an enzyme that is able to "add" nucleotides. This protein is responsible for the DNA replication. It requires a single stranded DNA (ssDNA) as template, nucleotides as building blocks and a short existing sequence called *primer*, that is bound to the template and will initiate the extension. To remove nucleotides, the DNA *nucleases*, are instead, able to cleave one nucleotide at the time and degrade in an orientated specific manner, the DNA strands. Restriction enzymes of class II, have instead the capability of cutting DNA sequences in specific sites, breaking phosphodiester bond and generating overhangs called sticky ends or blunt ends, if the molecule ends with paired or unpaired nucleotides.

### *2.2.1 DNA vs RNA*

The RNA molecule presents the ribose as sugar backbone, thus, it is a ribonucleic acid. This molecule express the information contained in the DNA, playing the important messenger rule. The step that converts the DNA in messanger RNA (mRNA) is well known as transcription that will lead to the translation of the information into an actual protein. The thymine base, is substituted in RNA with uracil (U), meaning that every adenine base will be read as a uracil.

## 2.3 DNA Nanotechnology Overview

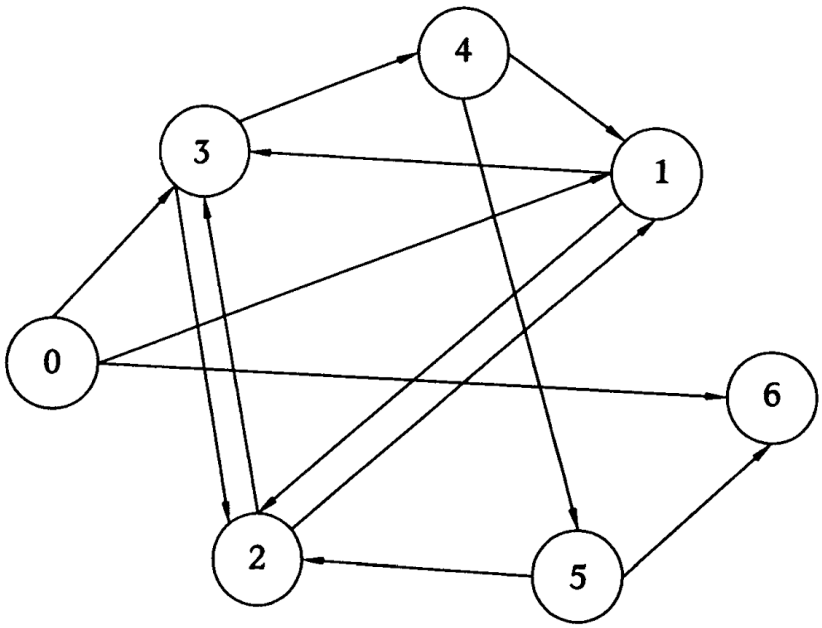
Watson and Crick nucleic acid double helix predictability, has been completely revolutionary for DNA nanotechnology, in the past 20 years, leading the scenario for a broad range of nanomachines such as switches, circuits, 2D-3D DNA origami folding [18] [19], novel DNA geometrical objects [20] and ultimately, towards DNA programmability [21]. Behind the DNA complementarity (A-T and G-C), exists an intrinsically algorithmic nature that makes DNA a valuable molecular kit to assemble molecular circuits at a nanoscale level, with DNA nanostructures as a way to design complex (possibly dynamical) structures. Structural DNA nanotechnology, takes advantage of DNA complementarity, relying on the hydrogen bounding of the DNA hybridization [22] as a fundamental mechanism operating at the molecular level. Sequence design marks the crucial first step to engineer DNA nanocircuits [23]

The idea started to became popular in 1982 when [24] *Seeman et al,* introduced the concept that nucleic acids can associate to form junctions capable to covalently engineer 3D nucleic acids structures [25] [26] [27].

But roots of this novel approach are back in 1959, when an American physicist, Richard Feynman had the intuition that building "sub-microscopic" computers was possible and that 50 atoms were capable to store one bit of information [28]. But it was only in 1994 that an American Computer Scientist, Leonard Adleman [29] used DNA double helix predictability to compute and perform the first experiment in the subject.

Adelman's experiment was a proof-of-principle study, consisting in designing a given $n$-vertex graph, and solving the Hamiltonian Path problem [30] (Fig 2.3 Panel A and B). The

problem was testing all the possible paths from vertex "0" to vertex "6" according to an established set of rules, to find the only combination visiting each vertex just one time. The experiment, marked a new wave of doing computation and an initial excitement for nanoscale assembling and computation.




$O\text{-}_2$ TATCGGATATCGGTATATCCGA
$O\text{-}_3$ GCTATTCGAGCTTAAAGCTA
$O\text{-}_{2\text{ to }3}$ **GTATATCCGAGCTATTCGAG**

$O\text{-}_2$ TATCGGATATCGGTATATCCGAGCTATTCGAGCTTAAAGCTA $O\text{-}_3$
**GTATATCCGAGCTATTCGAG**

Figure 2.3: **Hamiltonian Path problem**. (Top image taken from [29]) showing the direct graph and the path that goes from vertex "0" to vertex "6", and stops only once in each vertex. Bottom: example of vertex and distances between vertexes encoded in DNA sequences with direction 5'→3'

The nanotechnology community is widely attracted by the DNA in the form of a single-strand, which is a crucial intermediate in many cellular processes (*e.g.,* replication, tran-

scription and recombination of DNA) [31]. A single stranded DNA molecule, has the physical property of folding into a secondary structure, creating loops which provides a wide range of functionalities *e.g.,* protecting the ssDNA from the degrading nucleases activity [32]. A ssDNA is considered as very flexible and dynamic structure free from hydrogen bounds [33], although molecular stacking has to be taken into account, since it has been observed in ssDNA non containing secondary structures, at temperatures below 10°C [34]. In Table 2.1 is listed a range of software currently available to predict ssDNA and RNA secondary structures.

| Software | Specifications |
|---|---|
| Vienna Fold [35] | Predicts a secondary structure from a single sequence |
| M-fold [36] | web server for nucleic acid folding and hybridization prediction |
| NUPACK [37] | Analysis and design of nucleic acid systems. |
| oxDNA [38, 39] | Coarse-grained DNA model. |

Table 2.1: Popular DNA/RNA secondary structure prediction software

### *2.3.1 DNA Nanotechnologies used in this study*

DNA nanotechnology is successfully becoming a practical platform for implementing *in vitro* nanoscale structures exploiting the capabilities of kinetic and binding rates, where DNA sequences are decisive for a successful hybridization.
Hybridization is followed by component assembly and double stranded "sticky ends" (overhanging ends) can be used to link the sugar phosphate residue of a nearby DNA duplex molecule. This mechanism will tightly lock the structure, with covalent bonds and high affinity, thus, sticking the individual DNA pieces in a stable and dynamic structure.

Along the side of DNA hybridization, DNA strand displacement (also known as "Toehold mediated strand displacement") [7] has made possible the implementation of kinetically controlled DNA systems (*e.g,* DNA gate architecture [40] [41], DNA transport device inspired by microtubulis [42] and a molecular engine [43]), and DNA-fuelled molecular machines made of DNA, where DNA is used as "fuel" and structural material [44]. Reactions involved in both hybridization and displacement have the advantage of being enzyme-free and the molecular species are represented by DNA strands [45]. Limitations of the technology are represented by the possible interference between molecules in the system. Sequences have

to be designed carefully to prevent undesired interactions.

DNA strand displacement reaction rate is controlled *via* a single strand overhang (henceforth "toehold") exchange that has been proved to be much slower than DNA hybridization [46–48], thus, representing another limitation. Toehold stability, represents another bottleneck for an efficient displacement where mismatches, composition and length of the sequences involved can significantly affect the reaction rate constant [49].

#### 2.3.1.1 DNA hybridization

DNA hybridization is a well known mechanism that leads to the association of two random-coil single-stranded DNA covalently bound in a DNA duplex (Fig.2.4). The phenomena was introduced in 1963 and described as system where a phase called "nucleation" is followed by a second stage called "zippering". [50].

DNA nucleation is the initial process where two random-coil single-stranded DNA (ssDNA), are recruited in an order of milliseconds, and few base pairs will bind together [51], while searching for the correct complementary one. The base pairs will not immediately find the complementary ones, and there is a phase of attaching and detaching before achieving an equilibrium and the correct assembly. Models described by a coarse-grained DNA model, called "oxDNA" [38] demonstrated that high temperature can decrease the efficiency of the process and decrease the rate constant [39] .

The "zippering", is that phase that will follow the nucleation, and has a very fast progression. Hybridization reaction has been widely used in the scenario of nanotechnologies, because of its strong non covalent bound, extremely crucial in important molecular events (*e.g.*, DNA repair).

DNA hybridization, is a thermodynamically well known strand association, and the rate constant strongly increase with temperature, salt concentration [52], [16] [53] and depends on pH, length and composition of the sequences[53] [16]. The kinetics of duplex formation was explored using the temperature jump technique [52], and classified as a second order. Sequences with a high percentage of GC content can slower the helix dissociation from six to eight order of magnitude and enthalpies of activation are about 25 to 50 kcal/mole [52].

Duplex stability is much better understood than duplex reactivity. Two main ingredients

control the stability of the DNA double helix: i) base pairing, and, ii) base stacking. Looking closely at the thermodynamic of the DNA duplex, stability does not directly depend on base pairing. As a matter of fact, the aromatic rings of the base pairs show a coplanar alignment of the $\pi$ orbital [16] which is sequence dependent [54]. Kinetics of DNA hybridization is still not completely resolved, and thus remains particularly arduous to control.

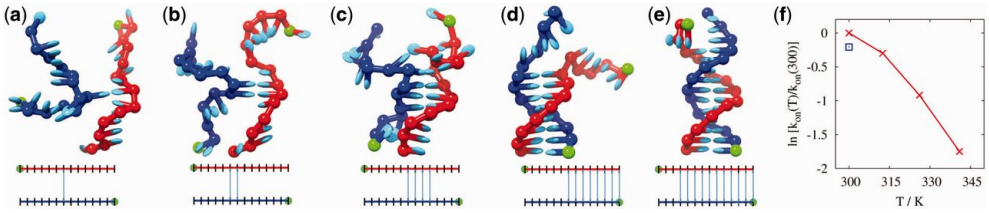



Figure 2.4: **Stages of DNA hybridization**. (Image taken from [39]). Schematic of the hybridization process and hybridization rates, (a-e) represents the stages of the hybridization, from the nucleation to the zippering. (f) Hybridization rate depending on Temperature.

#### 2.3.1.2 DNA Strand displacement

A Toehold mediated strand displacement or DNA strand displacement reaction (SDR), is a mechanism mediated by the toehold (short overhang domain) that will allow a DNA duplex to receive a new single strand DNA, called "invader" strand. The toehold can hybridize with the "invader" strand, thus, progressively releasing and replacing the originally bound strand through a process known as "three-way branch migration" [55], leading to a toehold-mediated DNA strand displacement [46] where the original hybridized "incumbent" strand will be pushed apart. The toehold mediated reaction is an elegant example of how an orthogonal reactions can support complex behaviours without the usage of enzyme and at room temperature. The hybridization is determined by the length of the toehold [7]. The crucial steps of the process are (i) the Toehold binding, (ii) the branch migration and finally (iii) the strand displacement. A Schematic of the species involved is represented in Fig.2.5.

(i) Toehold binding: thermodynamically downhill, either reversible or irreversible.
The toehold is activated via hybridization. The sequence is typically 4-20 nucleotides. The effect of the toehold length has been reported by *Yurke* at al., [7] as a limiting factor for the kinetic of the displacement. The longer the toehold, the higher the biomolecular rate constant, which can differ over six orders of magnitude. The toehold binding with the

"invader strand" can be reversible, depending on sequence composition and length it can be more prone to dissociate or to continue the reaction toward the displacement. The incumbent strand is a region that is complementary to the target strand, but also to the "invader". Thus, the "incumbent" strand and the "invader" strand can hybridize (Fig.2.5). It is important that the reaction is set in a chronological order, to provide control and programmability. Studies also demonstrate that there is a rate at which the incumbent can dissociate from the target strand [7].

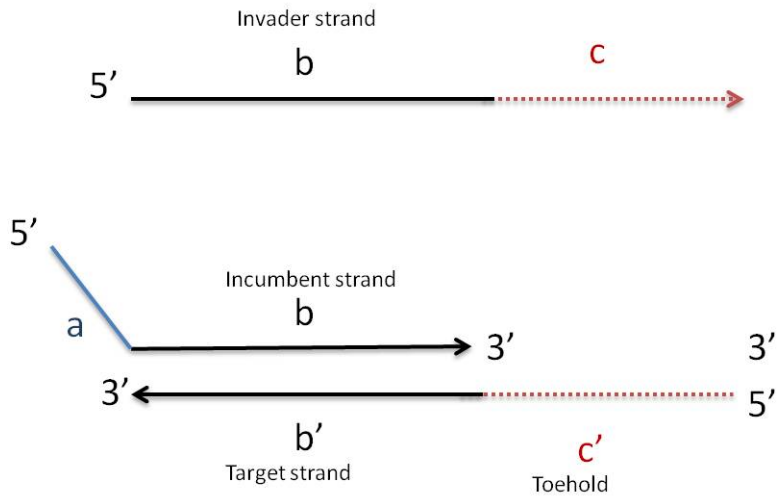



Figure 2.5: **DNA strand displacement molecular species**. A brief description of the molecular species involved during the hybridization, branch migration and displacement reactions. Sequences indicated with an apex are complementary to the correspondent ones. Black lines indicate the branch migration domains (b, b'). Red dotted lines indicates toeholds domains (c, c'). Blue line overhang sequence does not participate in the reactions (a).

(ii) Branch migration: thermodynamically neutral, random walk, reversible.

Both "invader" and "incumbent" strand are provided with a branch migration domain that will ultimately promote the strands exchange. At this point of the reaction, the "incumbent" and the "invader" will compete to bind the target, exchanging base pairs with it. The branch point will slowly proceed back and forth (the movement will occur at the same rate), and this action has been described as a random walk [56].

The reaction is thermodynamically favourite by the toehold binding, thus promoting the reaction forward. The presence of mismatches at different positions along the branch migration domain can affect the displacement reaction, and this could not be completed. Despite researchers demonstrating how length and mismatches can be fundamental parameters and kinetic rate can exponentially depend on them [47], it remains still very challenging to quan-

titatively predict the kinetic constant from toehold sequence. Rate constant for 1 nucleotide toehold and a 20 nucleotides branch is on the order of $10^6$ $M^{-1}$ $s^{-1}$ [47].

(iii) Strand displacement: thermodynamically downhill, irreversible.
Eventually, after the branch migration is complete, the incumbent strand will be totally displaced, whereas the invader will bind the target strand with a fully Watson-Crick hydrogen bonding. An overview of the strand displacement process is described in Fig. 2.6.

Along this line, the method represent a powerful tool for high level molecular programming and computer based simulation.

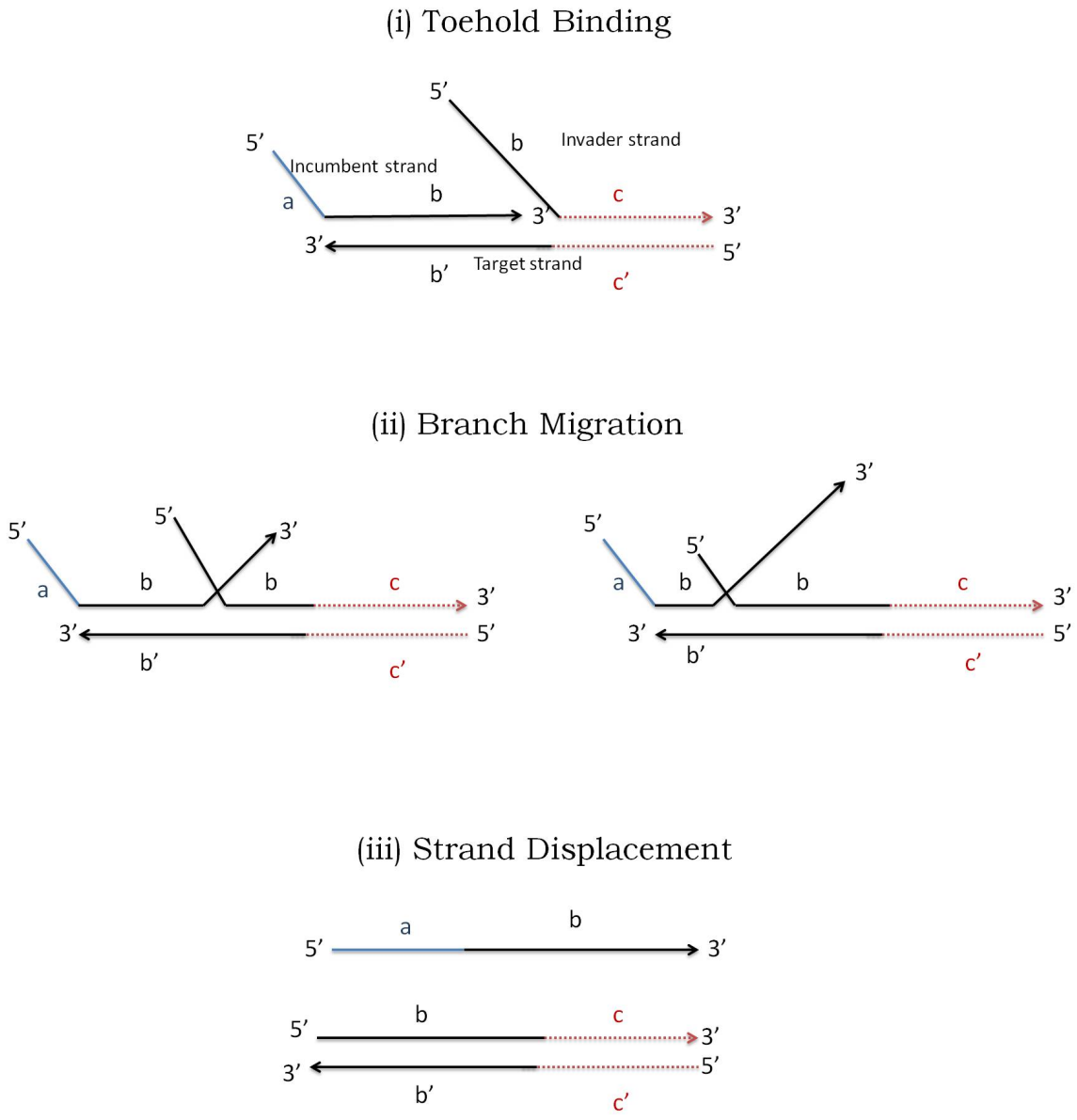



Figure 2.6: **DNA strand displacement cascade**. A brief description of DNA strand displacement cascade. The first step of the process is represented by the (i) Toehold Binding, followed by the (ii) Branch Migration, and the completion of the (iii) displacement.

#### 2.3.1.3 Applications and State of the Art

Recent applications of laboratory-scale DNA computing (Table 2.2) includes examples of finite automaton [57], autonomous DNA programmable modules that can control the levels of messanger RNA *in vitro* [58], DNA Boolean networks based on catalytic nucleic acids (DNAzymes) systems [59] [60], and the implementation of a finite state machine where DNA sequences represents state and transition rules triggered by a clock signal which allows parallel operation [61]. These works demonstrate that DNA not only represents a promising material to solve computational problems, as demonstrated by Adelman *et al*, [29], but also a novel platform to built DNA compilers [62] and implement Boolean logic gates [63] [64]. Moreover, it provides a novel computational answer to chemical reaction networks (CRNs), but also a way to expand the complexity of CRNs molecular components to a DNA polymer level [45] and built smart DNA XOR gates networks with autonomous behaviours [40]. In the CRNs design proposed by Qian *et al* in 2011, they employ reversible interactions to directly implement a reversible-stack data structure and switch the device from one configuration into another via strand displacement cascades, using a combination of three single strands. The majority of the work done so far has been about information processing and thus there is a lack of data structures and hence the relevance, importance and novelty of this dissertation.

| MACHINE | OPERATION SYSTEM AND REACTIONS |
|---|---|
| Programmable finite automaton [57] | Software and inputs encoded by dsDNA. Usage of restriction enzymes, DNA hybridization, ligation |
| *In vitro* modular DNA automaton [58] | Usage of ssDNA and DNA Hybridization to control mRNA levels |
| Libraries of DNAzyme subunits [59] | DNA based computational platform. For assembly of logic gates and half adder . |
| "MAYA" molecular automaton [59],[60] | A computational approach to build deoxyribozyme Boolean network of 23 molecular logic gates. |
| Efficient Turing-universal computation with DNA polymers [45] | Design of a strand displacement cascades for addition and removal of monomers from the end of a DNA polymer. |
| A clocked finite state machine [61] | States, transition rules and input represented by ssDNA. |
| This work: A DNA stack machine [65] | Based on DNA hybridization, DNA strand displacement and ssDNA data and operations. |

Table 2.2: Recent nanotechnological approaches implemented via DNA Hybridization and/or DNA Strand Displacement.

## 2.4 Synthetic Biology Overview

The term “synthetic biology” (SB) was used in 1979 by a biologist from the University of Freiburg, Barbara Hobom, in an article for a german newspaper, *Frankfurter Allgemeine Zeitung* [66]. Hobom used the expression as synonym of bioengineering, referring to the concept of bacteria genetically engineered via recombinant DNA technology. Later on, synthetic biology became associated with the idea of "redesigning" life. This would be realised through the synthesis of organic molecules that mimic natural ones (such as enzymes) [67], engineering-driven novel applications, *de novo* DNA synthesis for artificial gene networks [68], and computational redesign to reprogram signalling pathways [69]. The field started to become an increasingly inclusive concept [70, 71], with a different approach from classical genetic engineering, embracing both system design and system fabrication. In [72], "Ten Grand Challenges" were presented for the Synthetic Biology community, which are still relevant today and likely to remain so for many years to come, namely:

1. Reaching a consensus on synthetic and streamlined genomes

2. Cooking from scratch (bottom-up)
3. Learning from nature: naturally evolved reduced minimal genomes
4. Refine and make reality the notion of biological chassis
5. Manufacturing engineered biosystems
6. Overcoming physical and chemical constraints
7. From models to cells and back
8. Replication and reproduction
9. Towards an integrated design strategy of synthetic organisms
10. Coupling scientific development and public opinion information

There has also been an effort to guarantee "plug and play" adaptability via implementation and use of standard biological parts [73].
The area lies at the interface of a variety of disciplines ranging from biology through chemistry, physics, computer science, mathematics and engineering [67], thus representing a multidisciplinary area of biological research, employing the expertise of a range of fields. It is a rapidly growing platform for the design and manufacture of non-natural biological systems with the grand aim of allowing cells to carry out functions designed by scientists, where, a deep understanding of transcription networks, networks motifs and signal processing represents the key to synthesize novel biological networks to implement novel functionalities (*e.g.,"genetic memory"* and circuits acting as cellular event counters).

The discipline builds a deep understanding of biological systems with conceptual *in silico* tools using the descriptive language of Computer Science (*e.g.* Models, Simulation, Evolutionary Design & Optimization) to address old questions and challenges with approaches inspired by traditional engineering disciplines (Fig. 2.7).

These novel bio-systems could potentially be programmable computing devices, encoding genetic circuits that function like their electronic counterparts (*e.g.*, logic gates, oscillators and pulse generators) and genetic programs to be executed by introducing data into live cells to be processed.

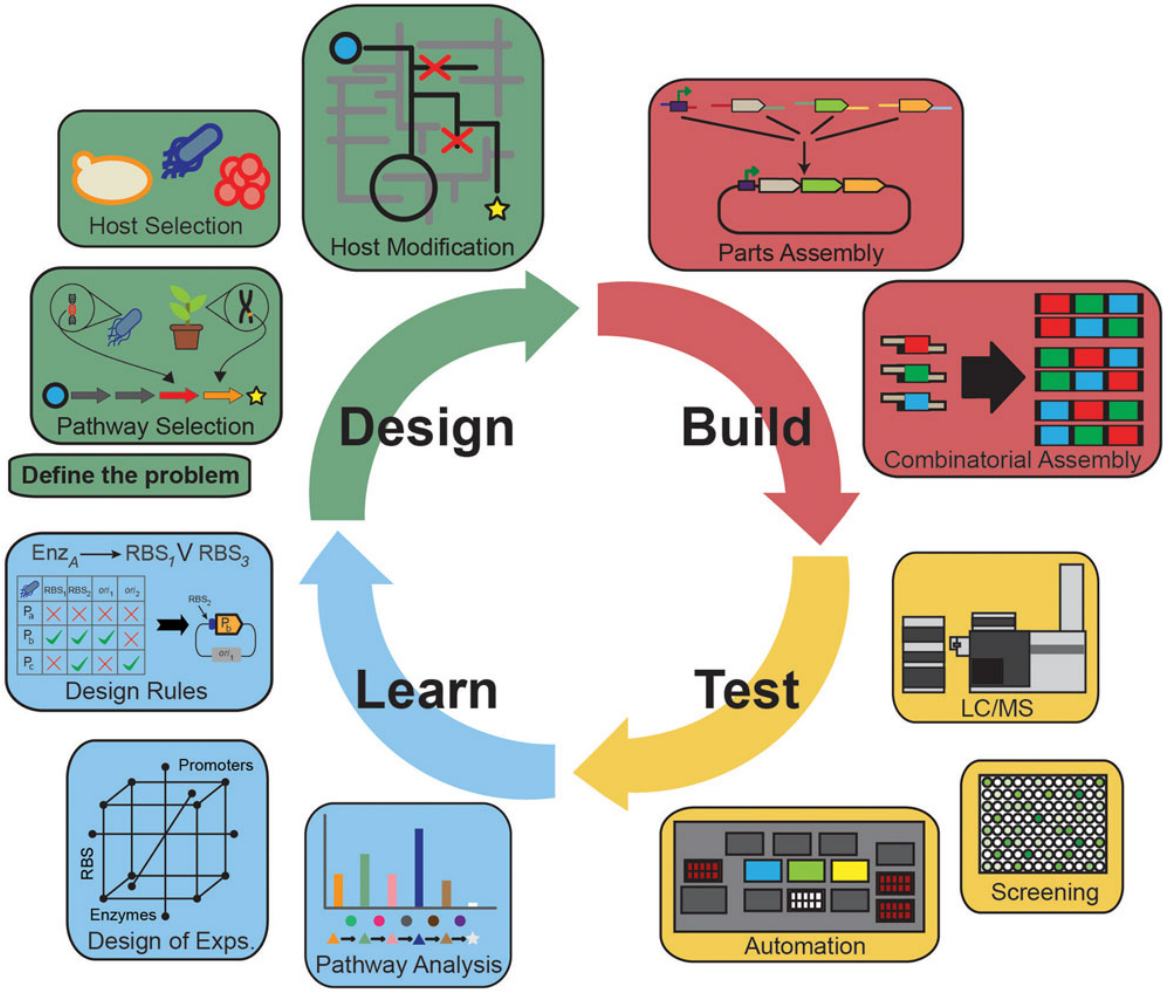



Figure 2.7: **The design–build–test–learn cycle of Synthetic Biology.**. Source: Royal Academy of Engineer, 2009 Report: "Synthetic Biology, scope applications and implications.

### *2.4.1 DNA recombinant technologies used in synthetic biology*

To work towards Synthetic Biology goals and ultimately engineering bacteria, recombinant DNA technologies, offer a wide range of molecular tools. In the early 1970s, the advent of genetic engineering using recombinant DNA technology revolutionized molecular biology [74]. DNA cloning technology made use of restriction enzymes, to recognize and cleave specific sequences on a DNA strand breaking phosphodiester bonds [75], and DNA ligase allowing the ligation of DNA fragments into specific vectors. Soon after, in 1983 the advent of another technology, the polymerase chain reaction (PCR), [76] marked new extended possibilities, capable to amplify a specific region of a DNA strand, with the usage of two primers that are complementary to the 3' ends of each of the sense and anti-sense strand of the DNA target. In 2003, 20 years after the revolutionary PCR, The J. Craig Venter Institute (JCVI) assembled in a unprecedented amount of time (14 days) a synthetic genome. The assembling was used to complete the 5386 bp genome of a $\phi$X174 Bacteriophage from synthetic oligonucleotides [77]. Since then, a variety of new assembly and genome editing methods were released and nowadays routinely used in molecular biology laboratories [78]. The most promising is represented by the CRISPR/CAS9 genome editing system [79]. The system introduces a break in dsDNA targeting a specific chromosome site recognised via a guide RNA (gRNA) that matches a 20nt sequence on the chromosome, flanked by an adjacent domain (PAM sequence). The DNA cleavage produces a blunt cut and is than repaired via homologous recombination. The system is a promising and attractive genome editing platform.

Therefore, usage of recombinant DNA technologies represents a powerful platform to advance in the synthetic biology research. Recombination systems from yeast and bacteriophage have been applied to experiments targeting suitable chromosomal positions, promoting expression cassettes exchange, inversion and translocation. *Saccaromyces Cerevisiae* FLP/FRT and bacteripohage P1 CRE/lox systems have been used for synthetic biology purposes [80] [81]. They are widely used because of their ability to mediate both integration and excision at exact sites of the genome. These systems also have site-specific recombination (SSR) which avoids the random insertion of genes into the chromosome.

#### 2.4.1.1 Cell-based genetic memory

More sophisticated designs have aimed to build logic systems with a temporary memory, through the usage of recombinant DNA technologies. Synthetic Biologist researches have tried to implement genetic memories circuits, but few advances were achieved in the last 5 years with various degrees of success. First steps were taken in 2008 when a “double inversion recombination switch” memory system encoding its state in DNA using an invertase-based system was capable of ‘"flipping" a DNA region to turn on and off specific genes [82]. In 2009, Friedland *et. al.,* [83] took significant steps to build genetic counters that lead to the expression of desired proteins after processing input signals, including the "Riboregulated Transcriptional Cascade" (RTC) counter and the“Single Invertase Memory Module” (SIMM). But it is only in 2010, that Bioengineers of Stanford University have implemented a rewritable digital data storage using DNA: called the "RAD" module [84]. This rewritable DNA-based memory module can reliably store digital information in living cells and could be used to track cell division and differentiation events in studies of ageing and cancer. The module is a recombinase addressable device, that can flip the DNA fragments. It uses a serine integrase and excisionase recombinase and chemical inputs to set or reset operations, and activate the transcription of specific DNA fragments respectively corresponding to “1” or “0” of a binary alphabet for 1 bit of memory. But expanding the capacity of the system represent a big limitation in terms of controllability, to get 8-bit of memory, for instance, the counter would require 256 input pulses and 16 recombinases, thus making the system very unstable and difficult to control. In 2014, a platform called *SCRIBE (Synthetic Cellular Recorders Integrating Biological Events)* was released for analogue, rewritable, cellular recording by generating single-stranded DNA (ssDNA) [85]. The circuit is able to convert genomic DNA into a living tape that can store information into a cell population in the presence of a specific signal (*e.g.,* a small molecule), an mRNA containing a specific target with a point mutation sequence is produced, a reverse transcriptase, recognise the sequence and partially reverse transcribe the mRNA, creating an hybrid cDNA/mRNA moelcule (msDNA). Specific proteins will bind the msDNA and target a sequence on the genome and catalyse a site specific recombination. After cell division, one of the daughter cells contains the msDNA and will pass the information to the next generation. The disadvantage is that the frequency of recombination is very low so just a small number of cells will undergo the

recombination.

In these works the memory capacity and scalability represents a crucial aspect for the storage of information in living cells. The usage of a combination of promoters could in theory make these systems more controllable but still with unlimited capacity. This is the aim of the *in vivo* tape prototype, presented chapter 4. The system core, uses a novel retrotransposon technology, called Targetron, which could potentially implement an unlimited capacity. This could represents a new advance in the storage of digital information in living cells providing an opportunity to make "biology easier to engineer". Recombination systems and usage of the described circuits are listed in Table 2.3

| CIRCUIT | RECOMBINATION SYSTEM AND USAGE |
|---|---|
| Double inversion recombination switch [82] | *fim* and *hin* invertase systems. Permits the inversion of specific chromosomal segments. |
| RTC counter [83] | T7 RNA polymerase. Riboregulators with a stem-loop structure. Count brief arabinose pulses. |
| SIMM [83] | Recombinase *Cre* and *flp*. Synthetic circuits to maintain genetic memory of low-frequency events. |
| RAD [84] | Bacteriophage integrase and excisionase system. Rewritable memory module to track cell division. |
| SCRIBE [85] | Expression of ssDNAs from engineered retrons that use a reverse transcriptase protein to produce hybrid RNA-ssDNA molecules. |
| This work: An *in vivo* Tape Data Structure | Based on intron II technology and integrase system. Stores and accumulates digital information in living cell. |

Table 2.3: Main features of systems representing advances in molecular circuits

### *2.4.2 Molecular Biology technologies used in this study*

The *in vivo* tape prototype, presented chapter 4, is based on an Intron II technology and integrase system with the aim of storing and accumulating digital information in living cell. The following subsections provide a detailed overview and applications of both technologies.

#### 2.4.2.1 TargetTron Technology

The TargeTron gene knockout system from Sigma Aldrich® allows for the knocking out of bacterial genes by insertion of group II introns. The system is based on a specific ribozyme, comprising by the mobile group II intron from *Ll*.LtrB from *Lactococcus lactis*, and it

operates via a site-specific retroelement that can invade a cognate sequence [8][9]. Two key features of group II introns are the intron RNA and the LtrA intron encoding protein (IEP). To direct the group II intron to disrupt a specific gene, the RNA portion of the intron must be mutated in three specific regions, denoted as exon binding sites EBS2, EBS1, and $\sigma$ (complementary DNA target sequences are the intron binding sites IBS1, IBS2, and $\sigma$') involved in the base pairing and located in two different RNA stem loops, denoted as d1 and d3, showed in Fig 2.8 [86].

The intron RNA and the LtrA protein are transcribed into a single transcript, the intron RNA is encoded at the 5' end, while the IEP is at the 3' end. The IEP is a protein which provides for four functions, a maturation function for intron RNA splicing, a DNA binding activity for target site, an endonuclease activity and reverse transcriptase activity [8]. Following translation of the LtrA protein, the next step is the maturation activity of the LtrA protein. After translation, this protein binds the unspliced intron, and promotes the formation of the active ribozyme, which is essential for the intron splicing.

Additionally the IEP, reverse transcriptase activity will reverse transcribe the mature intron and insert it as cDNA into the target site. Upon intron splicing, the LtrA will form a ribonucleoprotein complex (RNP) with the excised *Ll*.LtrB intron lariat, thus mediating the mobility of the intron . The intron RNA and the intron-encoded protein (LtrA) recognise the DNA target and perform the reverse splicing into the intron insertion site (IS) in the target DNA, following in ligation of the 3' end of the intron RNA to the 5' end of the 3' exon DNA. The IEP cleave the bottom strand, and reverse transcribes the intron RNA. After, the intron cDNA is connected to the 5' exon. This mechanism is known as "retrohoming of intron II" is followed by degradation of the intron and second-strand synthesis (Fig.2.9) A stable intron insertion is born to disrupt the target; this disruption is permanent, specific and stable. The targetron gene knockout system from Sigma Aldrich provides an efficient, specific and permanent method to generate gene knockouts [87] [88]

The mutation occurs into a 350 bp DNA fragment by using oligonucleotides that target the intron to the gene of interest. These oligonucleotides are generated using the TargeTron algorithm [89][90]Frazier et al. [8][87]. After multiple round of PCR, the mutated intron PCR template is purified, cut and ligated into the TargeTron expression vector (Protocols and Specifications are provided by Sigma Aldrich®) and inserted into the host cell ei-

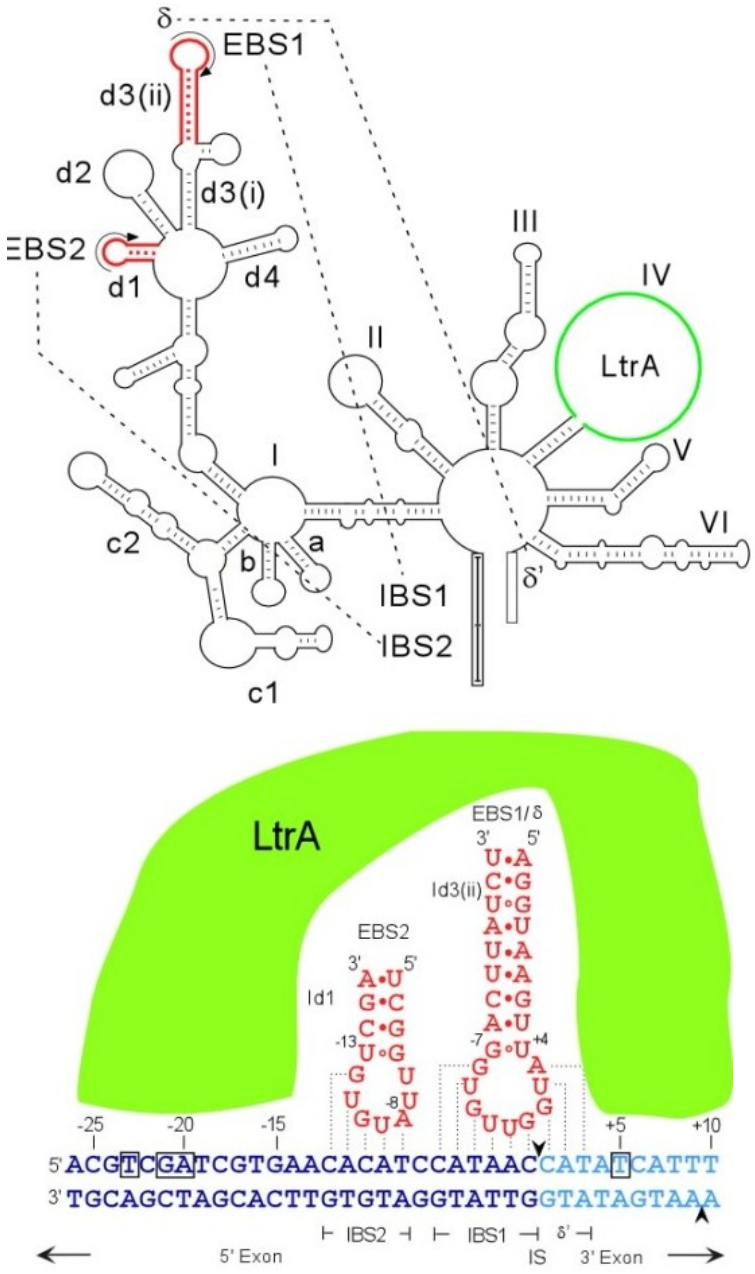



Figure 2.8: **TargetTron Technology.** Top: Tertiary structure of the ribozyme and RNA loops involved in the targeting process (EBS1, EBS2 and $\sigma$). The green loop represents the portion transcribing the LtrA protein. Bottom: LtrA protein (green) and EBS sites targeting the correspondent IBS sequences. Image adapted from the Targetron website (http://www.targetrons.com)

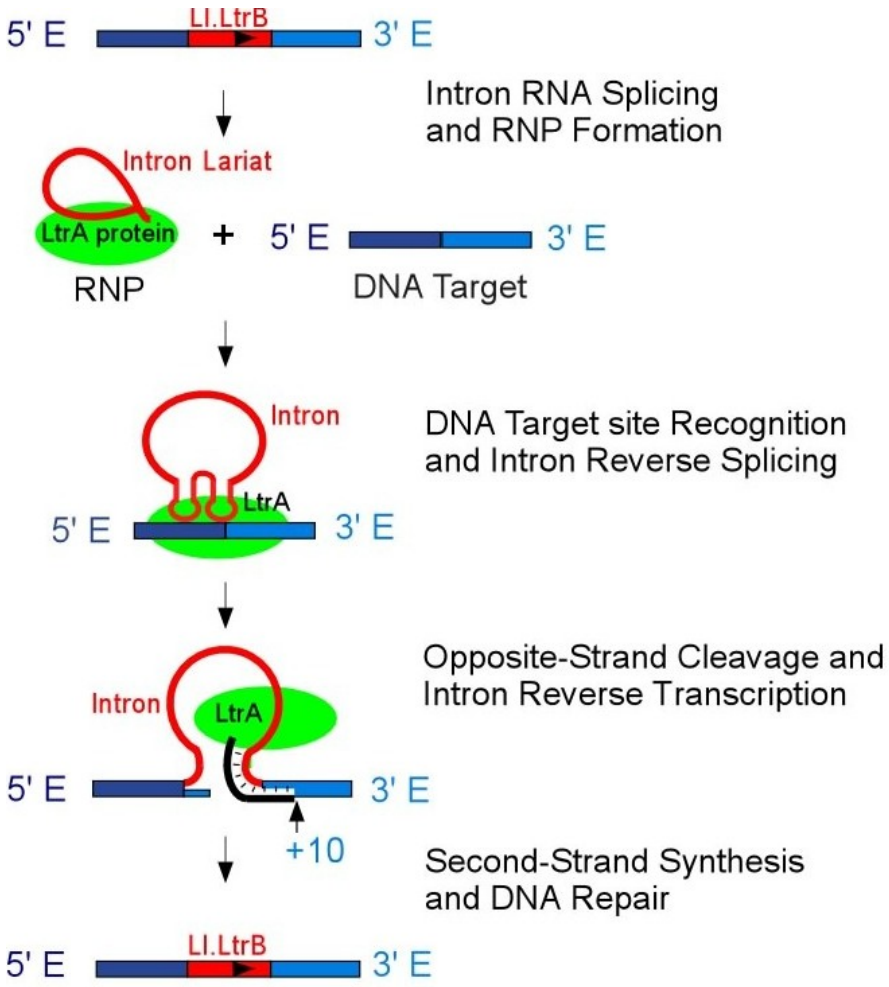



Figure 2.9: **Retrohoming of intron II**. Adapted from the Targetron website (http://www.targetrons.com)

ther chemically or by electroporation. Utility of this technology has been demonstrated for prokaryotic genetic engineering, systems biology and functional genomics approaches, *e.g* ClosTron system, which allows the directed construction of stable mutants in *Clostridium* species using a bacterial group II intron [91] [92] [88] TargeTron technology has been confirmed for use in prokaryotic genomes, to establish mutants (*e.g.* ClosTron system, which allows the directed construction of stable mutants in *Clostridium* species) [91].

#### 2.4.2.2 $\phi$C31 Integrase and related recombination systems

Specific site recombinase (SSR) based approaches are becoming attractive methods since they enables a specific integration in defined sites. Systems like *Cre* from Bacteriophage P1, and *FLP/FRT* system adapted from yeast [93], catalyze recombination reaction as excision, inversion and DNA translocation. Both systems can catalyze a trans event related to the insertion of circular DNA into the genome with the consequent creation of two cis-positioned sites which are substrate for the excisionase [94]. Both systems do not show an efficient selection-free identification of insertion [93] [94] and recently $\phi$C31 from *Streptomyces* seems to be ideal for the insertion in mammalian genome [94]. The *Streptomyces* $\phi$C31 derived integrase, has been widely studied and it belongs to the to serine recombinase family. This family of recombinase catalyze the recombination event, recognizing specific sequences known as *attB* and *attP* [94], chosen as candidate in the Tape System presented in this study. The recombination event produce hybrid sites (*attL* and *attR*). These new sequences cannot be recognized by the integrase. The *FLP/FRP* and *Cre-loxP* systems can be considered as a first generation of specific site recombinase (SSR) systems used as recombinase based technology, and based on them, many genetic tools have been designed in the last 20 years [95]. A second generation, can be considered the RMCE technique, which permits a fast modification of a defined chromosomal locus through targeted integration of transgenes [93].

## 2.5 Conclusions

This chapter highlighted the advances in Nanotechnology via DNA Hybridization and DNA strand Displacement, and in Synthetic Biology via recombination systems. The next part (Part II) of this dissertation, will focus on the design and specifications of the *in vitro* and *in vivo* prototypes. Firstly, chapter 3, will describe the hybridization and displacement reaction cascades, implementing the *in vitro* stack data structure. Secondly, chapter 4 will describe the *in vivo* tape design and specification using the SSR-based approaches via TargetTron and $\phi$C31 Integrase systems.

# Part II

# Chapter 3

# Specification and Design of an In Vitro DNA Stack Data Structure

This chapter introduces the design and specifications of a DNA stack data structure, operating via building blocks called "bricks". The bricks will interact via DNA hybridization and DNA strand displacement, methods introduced in the previous chapter (chapter 2). A coarse-grained DNA model was used to model the tertiary structure of the stack prototype.

## 3.1 Introduction

I propose a prototype of a DNA stack data structure that implements a stack machine and represents a dynamic memory. A stack is an abstract data structure that serves as a linear collection of elements, with two principal operations: *push* adds an element to the stack, and *pop* removes the most recently added element that was not yet removed. Formally, this is achieved through the interface

$$push : \text{stack} \times \text{element} \longrightarrow \text{stack}$$

$$pop : \text{stack} \longrightarrow \text{stack} \times \text{element}$$

with the invariant

$$pop(push(\text{stack}, \text{element})) = \text{stack}, \text{element}$$

*e.g.*,

$$push(\text{d,abc}) = \text{abcd}$$

$$pop(\text{abcd}) = \text{d,abc}$$

with the invariant

$$pop(push(\text{d, abc}) = \text{abc,d}$$

to guarantee *last-in-first-out* operation. Further common but non-essential operations such as *peek* (return the last element without removal) and *empty* (return true if the stack experienced at least as many *pop* as *push* operations) are not provided in this prototype implementation.

Fully implementing this data type in DNA requires molecular realizations of the assembled stack, all potential elements, as well as the push and pop operations. This is achieved by associating each data element and each operation with a single-stranded DNA (ssDNA) with partial secondary structures. Those strands are called "bricks", schematic representation is shown in Fig.3.1.

Table 3.1 lists the nucleotide sequences of all domains designed by my collaborator Jerzy Kozyra with a genetic algorithm [65].

| **domain** | **sequence** | **length** | **domain** | **sequence** | **length** |
|---|---|---|---|---|---|
| $a$ | TCTCCC | 6 | $h_L y$ | GCACGCTCGAGCTCGTATCGCAGTA | 25 |
| $b$ | GCCA | 4 | $k_x, k_y$ | CTCTAATCAC | 10 |
| $c$ | GCACACACTTC | 11 | $k_L y$ | CATCCCTATA | 10 |
| $d$ | ACACCACTTC | 10 | $l_x, l_y$ | AGACAAAAAA | 10 |
| $e$ | GGGAGACCAA | 10 | $l_L y$ | ATTTTTTTCC | 10 |
| $f$ | CGGCGG | 6 | $x/y$ | AGACCGCTAAA/ATACTGCTTTA | 11 |
| $g$ | CTGCC | 5 | $m$ | TATGACTGCAA | 11 |
| $h_x, h_y$ | ATTAGTAGGT | 10 | $Ly$ | ATACTGCTTTA | 11 |

Table 3.1: Sequence specification of domains in the design. The ssDNA strand sequences has been divided into domains. Sequences are indicated in 5'→3' direction.

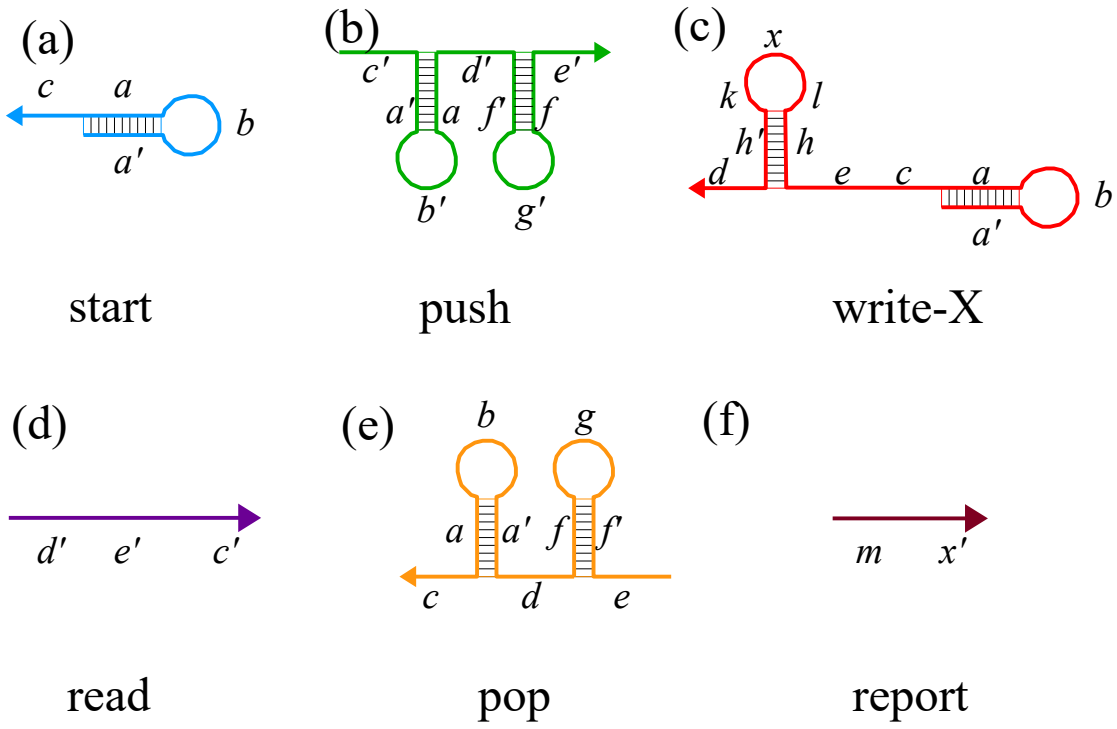



Figure 3.1: Schematic of the different bricks involved in the DNA stack.(a-c) bricks involved in the data storage cycle. (d-e) bricks involved in the reading cycle. (f) non essential brick in the stack operations. Domains are specified by letters, where the apex implies complementary domains, *e.g.*: a' is complementary to a. Different letters *e.g.* a and b are implied to be non-complementary. Arrows indicate 5' $\rightarrow$ 3' direction.

The stack data structure is built from bricks via hybridization of complementary DNA domains. More precisely, the stack forms a double-stranded DNA (dsDNA) assembly with essentially no single-stranded regions but one active toehold domain, that offers an entry for operation. Data bricks form the top strand and push bricks form the bottom strand of this dsDNA assembly. To prevent runaway processes (uncontrolled elongation of the stack) that might occur when adding bricks in realistic concentrations, the device to toggle between two states in all modes of operation referred as *data state* and *operator state*. When the stack is in data state, it will accept a single data brick. Upon binding this data element, the device toggles into the operator state in which it cannot further interact with data bricks, but instead awaits a new operator brick such as *push*. Again, only a single operator brick is accepted, and by interacting with it, the stack toggles back into the data state.

The design differs from the one proposed by Qian *et al.* in several important aspects:

1. all data and operations were implemented as single DNA strands, whereas Qian *et al.* employ bricks of up to three DNA strands.

2. The assembled DNA stack is entirely double-stranded and does not feature any dangling single-stranded overhangs, which is used by Qian *et al*. in 2010, to store the actual data elements.

3. Instead, in this design data is encoded in internal secondary structure motifs in the double strand, namely in hairpin loops that form holiday junctions.

4. The stack modes of operation are based on DNA interactions that are effectively irreversible at the operating temperature. Qian *et al*.'s design, in contrast, employs only reversible interactions and relies on detailed balance to drive the device from one configuration into another.

The stack prototype take these design decisions, in order to minimize the amount of required distinct DNA sequences and to obtain maximally robust modes of operation, especially when envisioning ultimate in vivo applications.

### *3.1.1 Data and Operator Brick Design*

The DNA data structure device operates with six distinct DNA bricks (Fig.3.1) and is able to store combinations of two different signals, encoded by two types of data elements. Two further bricks are added for experimental analysis.

- *Start*: data brick designating the beginning of the stack data structure. It features a toehold domain for interaction with *push* and a hairpin motif at the 5' end. This hairpin undergoes branch migration with a complementary hairpin in *push* but is otherwise not functional in the current design. The 3' end can be biotynilated for further functionalization.

- *Push*: operator brick to initiate subsequent signal data storage. The brick contains the complementary toehold for interaction with *start*, a hairpin motif complementary to the one in *start*, a second hairpin that does not participate in branch migration and exists for structural reasons, and two toehold domains, one on each side of the structural hairpin, to bind *write* bricks.

- *Write*: data bricks that can be stored in the stack data structure. These bricks contain two toehold domains complementary to the push toeholds, a structural hairpin that does not undergo branch migration, plus the same toehold domain and 5' hairpin that form the *start* brick. Toehold domains and branch migration hairpins are identical for all types of *write* bricks. Thus, they can only differ in their structural hairpin motif. Since these hairpins do not participate in hybridization or branch migration, they can be functionalized to host any desired functionality such as recognition sites for DNA binding proteins. Spacing and spatial arrangement of these data loops can be controlled by altering the length of toehold domains.

  Two different types of *write* bricks, denoted as *write-X/Y* and *write-Ly*, were employed. *Write-Ly* features a longer hairpin stem than *write-X/Y* (twenty-five base pairs against ten base pairs) and has a different sequence in its stem loop. Whereas all other hairpins in the design employ shortest possible stable loops, the structural loop of *write* bricks contains a significantly longer single-stranded loop domain.

- *Read*: operator brick that removes the rightmost *write* operation. The brick is the complement of all toehold domains used in *write*'s. Notably, it does not contain any domains that interact with the structural hairpin of *write* bricks.

- *Pop*: data brick that undoes the rightmost *push* operation. This brick is the exact complement of *push*

- *Report*: non-essential bricks for experimental analysis. Report bricks do not participate directly in the operations of the stack. Instead, they interact with the data domains of structural hairpins in the *write* bricks. *Reporters* can be added to the device in any configuration since their binding sites in the data hairpins are always accessible and since they do not interfere with the operating modes of the device.

  Domain sizes have been chosen with the following objectives: toeholds are long enough to span a single helical turn when hybridized with their complements (10 nt) which should promote irreversible hybridization. Hairpin loops that participate in branch migration are long enough to promote stable stems (6 base pair stems with 4-5 nt loops) but short enough to obtain quick branch migration times. The single-stranded loop domains of *write* bricks is long enough (31 nt) to prevent steric restrictions to hybridization with complementary *reports*. The structural hairpin loop of *write* bricks together with the unpaired domain of *report* are long enough to accommodate 5 nm and 10 nm diameter nanoparticles in close vicinity to the device.

### *3.1.2 Modes of Operation*

DNA hybridization, branch migration and strand displacement are the three processes governing all involved DNA interactions. All reactions are energetically downhill, driven by the binding energy of the closing toehold domains.

#### 3.1.2.1 Data Storage

A schematic of the Data Storage process is shown in Fig.3.2. Starting from an empty stack (Fig.3.2 (a)), which is represented by the *start* brick, the device is toggled into its data state

by providing a *push* operator (Fig.3.2 (b)). The *start-push* interaction begins by irreversibly binding toehold *c* and continues via branch migration among the two complementary *aba'* domains (Fig.3.2 (c,d)). The stack (Fig.3.2 (d)) is now in its data state, where a single open toehold region (*d'e'*) can recruit a *write* brick (Fig.3.2 (e)). The *write* will partially hybridise with the *d'e' push* toeholds, thus toggling the stack back into its operator state (Fig.3.2 (f)). In this state, the stack exposes the same toehold-hairpin interface that characterises the *start* brick, which allows the device to undergo subsequent rounds of data storage .

Note that the assembled stack is essentially double-stranded with a single exposed toehold domain. Because the structural hairpins of neither the *push* nor the *write* participate in branch migration, the stack will form holiday junctions for each stored data element. As data specific domains are encoded in the loop regions of this holiday junction, the data storage cycle is independent on the actual written data.

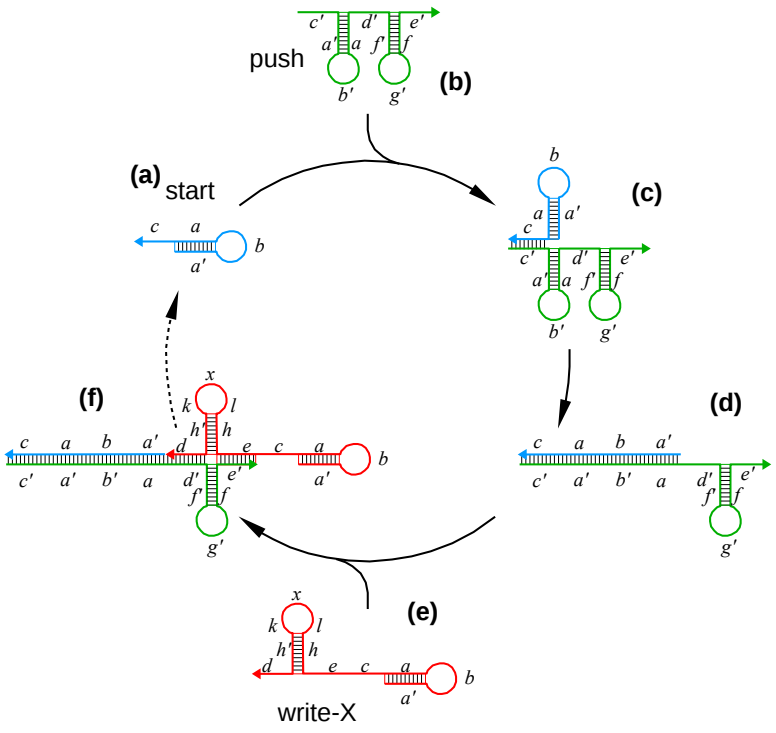



Figure 3.2: Schematic of the data storage process.

### 3.1.2.2 Reading

While data storage proceeds from left to right in the schematic Fig. 3.2, reading will proceed from right to left, thereby undoing any data stored in the last-in first-out manner required by

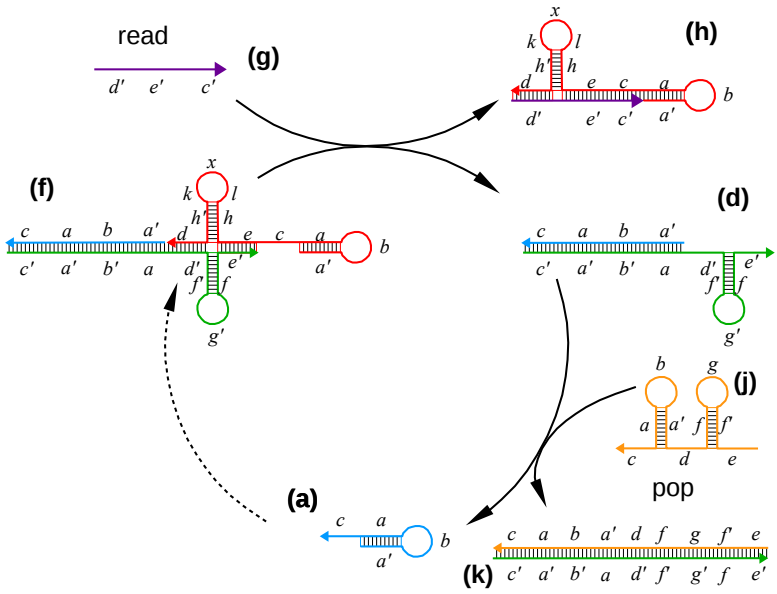



Figure 3.3: Schematic of the reading process.

the stack specification. A schematic of the data storage is shown in Fig. 3.3.

Assuming the stack in operator state **(f)**, providing a *read* brick **(g)** will peel off the last stored *write* brick **(h)**, thereby toggling the device back into the data state **(h)**. This reaction proceeds in two steps: first, the *read* brick hybridizes to the stack at its unique exposed $c$ domain. Secondly, the dangling $d'e'$ domains of the *read* brick initiates a three-way branch migration with the $d'e'$ domains of the adjacent *push* brick against the domains of the *write* brick, until the *push* strand is completely displaced. Note that the data hairpin of the *write* brick does not participate in the branch migration. This ensures that a unique *read* brick can interact with any *write* brick, ensuring that data elements can be read from the stack without a need to know which information has been stored. The resulting *read-write* complex **(h)** is completely double stranded and will not participate in further DNA interactions. In its data state **(h)**, the stack can either be extended again with another data element by switching to the data storage operation, or reading can be completed by toggling the stack back into its operator state. The latter is done by providing a *pop* brick **(j)** that will interact with and peel off the exposed *push* brick. Analogue to the previous reaction, *pop-push* interactions are composed of their initial irreversible toehold hybridization, subsequent branch migration and eventual strand displacement. Again, the resulting *push-pop* complex **(k)** is completely double-stranded and will not participate in further DNA interactions. Strictly speaking, this implementation of the reading cycle violates the stack invariant: while data storage and

subsequent reading of a single data element leaves the stack invariant, the data element itself has been altered as it is now hybridized with a read strand. This could be easily amended by extending the *read* brick with an additional domain that does not interact with any domain present in this design. This would introduce an active toehold into the currently inactive *read-write* complex. An additional DNA brick, which is completely complementary to this extended *read*, could then be introduced in order to displace the *read* brick from the *write* brick, bringing the latter back into its single-stranded, active configuration. As the emphasis of this work is on signal data storage, for which reactivation of the *read-write* complex is not required, I have not followed this extension of the design.

| **brick** | **domains** |
|---|---|
| start | $a'bac$ |
| push | $c'a'b'ad'f'g'fe'$ |
| write$_i$ | $a'bacehl\ x_i\ kh'd$ |
| read | $d'e'c'$ |
| pop | $ef'gfda'bac$ |
| report$_i$ | $mx_i$ or $x_i'm$ |

| **domain** | **length [nt]** | **domain** | **length [nt]** |
|---|---|---|---|
| $a, a'$ | 6 | $g, g'$ | 5 |
| $b, b'$ | 4 | $h, h'$ | 10 |
| $c, c'$ | 11 | $k$ | 10 |
| $d, d'$ | 10 | $l$ | 10 |
| $e, e'$ | 10 | $x_i, x_i'$ | 11 |
| $f, f'$ | 6 | $m$ | 11 |

Table 3.2: **Specification of domains in the design**. The ssDNA strand sequences has been divided into domains. Domains are specified by letters, where the apex implies complementary domains, *e.g.*: a is complementary to a'.

## 3.2 Reporter Strands and Biotin Functionalization

Data bricks can be functionalized to host many desired functionality. In this study, *report* strands were 5'biotinylated via a 2.6 nm tetra-ethyleneglycol (TEG) spacer. These *report* bricks were functionalized with streptavidin coated gold nanoparticles of different diameters (10 nm for *report-X* and 5 nm for *report-Y*), which allows for easy recognition using transmission electron microscopy (TEM). The estimation of the tertiary structure of the chain has been provided by my collaborators Dr. Harold Fellermann and Jerzy Kozyra from some simple considerations: Assuming A-DNA conformation (raise 0.24 nm/bp, rotation 33.6°/bp) signal hairpins are separated by about 11 nm and lie in a 139° degree turn. The signal hairpin is orthogonal and 2.4 nm long. The biotin is separated from the signal binding site by a to 2.6 nm spacer-TEG (tetra-ethyleneglycol,15 atom). It has been added to the 3'-ends of the start brick, in order to attach it to streptavidin coated surfaces, if necessary. Assuming B-DNA conformation (raise 0.34 nm/bp, rotation 35.9°/bp) signal hairpins are separated by

about 15 nm and lie in a 247° degree turn (-112.7$^0$). The signal hairpin is orthogonal and 2.4 nm long. Fig.3.4 (left) shows a schematic of this estimated structure. However, oxDNA simulations [96] of a single *start-push-write* complex indicate that the assembled chain does not necessarily extend straight forward but might instead turn in an angle at the holiday junctions (Fig. 3.4 right).

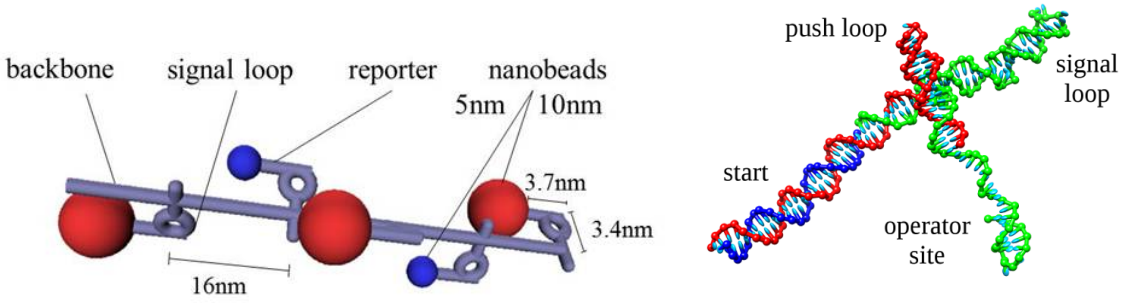



Figure 3.4: **3D B-Conformation assembled structure**.Representative predicted image of X-Y-X-Y-X-Y Where *reporter X* has been attached to a 10 nm gold nanoparticle and *reporter Y* to a 5 nm gold nanopartcle.

## 3.3 Conclusion

This design contributes to the art of designing a DNA data structure using DNA hybridization and DNA strand displacement cascades. The architecture implements a stack data structure with push and pop operations and allows for DNA data storage of multiple signal types. Currently the lack of laboratory scale implementation of data structures are lacking, hence the relevance, importance and novelty of this dissertation. Part III, chapter 5, will introduce the experimental methods used for the implementation of the DNA stack data structure in the laboratory.

# Chapter 4

# Specification and Design of an In Vivo DNA Tape Data Structure

This chapter describes the design and the specifications of a prototype for a tape data structure *in vivo* which could enable adding and storing inputs in a plasmid that can then be analysed afterwards. Sequencing of the *tape* plasmid after the recorder bacteria is exposed to the signals could reveal the information stored in a temporal order. The circuit is designed in order to perform tasks sequentially under the control of two recombinase systems (TargetTron and $\phi$31 Integrase) which were presented in chapter 2.

## 4.1 Introduction

The implementation of the tape requires the assembling of a synthetic gene network and the usage of a "*blank*" plasmid, recognition sequences (*attR*) in principle capable to accept and store one event. The gene network is induced by an external signal (as to avoid run-off events), and upon storing one event, the plasmid switches into a locked state in which it cannot further accept and store a second signal, unless is re-setted, thus, another bit of information can be accepted and the tape can be extended. The tape device presents a three-module design, where each module accomplishes a specific task: signal detection (Module I) data storage (Module II) and, system reset (Module III):

1. **Module I**: includes a repressor gene, which will maintain the circuit in a standby

state. This module represent an interface element that allows an external signal (a chemical inducer) to switch the state of the system ON and OFF. It is detects the external inducer and initiates the transcription of Module II (circuit ON).

2. **Module II**: includes a recombinase gene (henceforth, retrotransposon) responsible for the translation of a multifunctional protein that promotes the tape extension *i.e.* implementing the actual data storage.

3. **Module III**: includes a second recombinase (henceforth, integrase) and hold its standby state until a second external input activates the transcription of the integrase, responsible for the tape reset.
   This is schematically shown in Fig.4.1

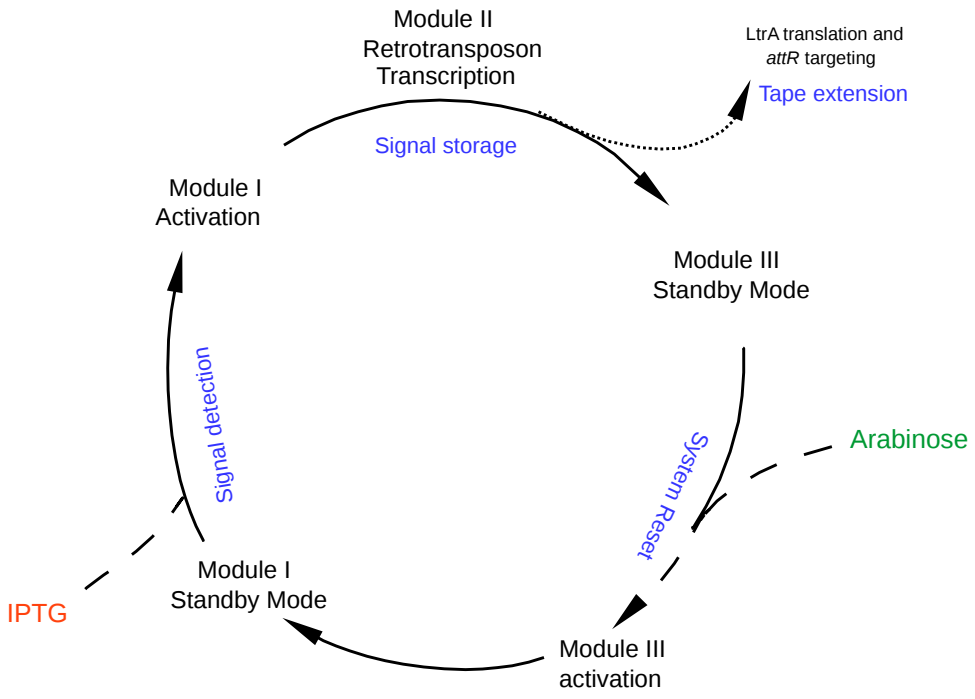



Figure 4.1: **Schematic representation of the tape recorder modules.** Module I detects an external inducer represented by the IPTG signal. Module II represents the core for the signal storage transcribing the a self-splicing retrotrasposon, that targets and extends the tape plasmid. Module III resets the system.

## 4.2 Tape Implementation

The recorder is a synthetic system which, upon detection of a signal (operated by Module I), for example isopropyl beta D thiogalactopyranoside (IPTG) can store the event in a DNA plasmid tape. Unlike other designs for *in vivo* data storage (*e.g*, [84, 85]) the presented tape is not limited to a small fixed number of bits of information and can grow arbitrarily large. In the current prototype I am only considering the storage of one event *type*, representing a unary alphabet, as the modification of the system to binary or even larger order alphabets (*e.g.* representing different events types) is trivial. The circuit is designed in order to perform sequentially different tasks and the control is mediated via the induction of specific genes transcribing either a self-splicing retrotransposon to record the signal of interest or a gene encoding a site-specific recombinase to reset the system. As briefly mentioned in the previous section, the whole system is constituted of three components. Each step is executed by a functional module to accomplish a specific function and carrying specific genes and promoters, specifically:

1. **Module I**: comprises the *lacI* gene and the pTac promoter. Its task is to detect the extracellular signal, IPTG (Fig.4.2). The module is flanked by SacI and NotI restriction enzymes cutting sites.

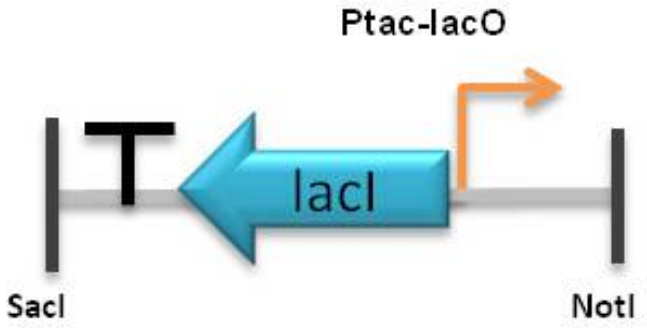



Figure 4.2: **Schematic representation of Module I.** The Module detects an external inducer represented by the IPTG signal.

2. **Module II**: carries the trimethoprim (Tp) antibiotic *dhfR* gene and a non functional *pyrF* (henceforth indicated as *pyrF**) genes, as well as the engineered *Ll*.LtrB retrotransposon based on the TargetTron system, [89][8][87] (see chapter 2, section 2.4.2) which will target and disrupt a recognition sequences, referred as *attR*, on the "blank"

plasmid. This retro-transposition is the central mechanism of the circuit and is essential for the system, because it permits to run once and then become locked in a stationary but reversible state, making the Module II the core of the recording mechanism. (Fig.4.3).

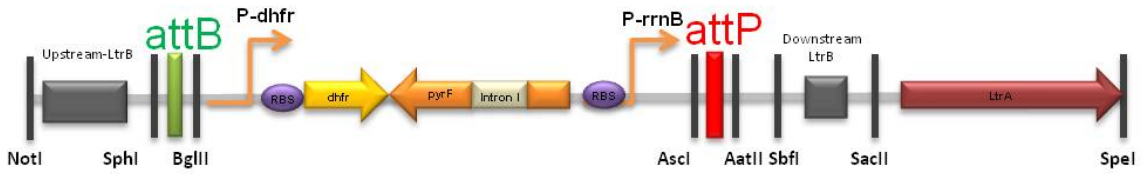



Figure 4.3: **Schematic representation of Module II.** The module represents the core for the signal storage transcribing a self-splicing retrotrasposon, that targets and extends the tape plasmid. The engineered targeTron is flanked by the Upstream and Downstream sequences of the LtrB intron II (grey rectangles) and NotI and SpeI restriction enzymes flank the entire module, comprising the LtrA protein.

3. **Module III**: carries the AraC protein and int-$\phi$C31, and represent the necessary element to reset and reconstruct the disrupted *attR* sequence. In this way the system returns to standby and a new signal can be recorded *attR* target site, on the target plasmid. (Fig.4.4).

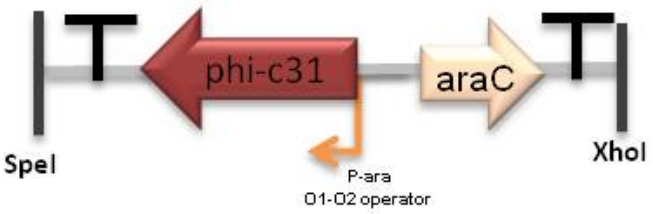



Figure 4.4: **Schematic representation of Module III.** The Module resets the system and is flanked by SpeI and XhoI restriction enzymes sites.

A schematic representation of the entire circuit assembled is shown in Fig.4.5.

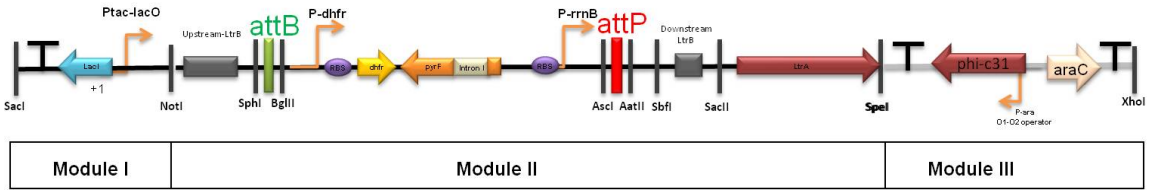



Figure 4.5: **Schematic representation of the entire tape circuit assembled.** Full map of tape including the restrictions sites used for its construction. The entire tape system is 9 kilobases (kb)

### *4.2.1 Intron re-tareting*

The exon recognition sequences (ESB) on the *Ll*.LtrB intron II (Chapter 2, section 2.4.2.1) were re-targeted introducing a suitable and cognate intron recognition sequences (*attR*) on the target plasmid from the bacteriophage $\phi$C31 [97]. The intron recognition sequences was designed using the *attB* and *attP* sequences (Chapter 2, section 2.4.2.2) and an online tool implementing the Perutka Algorithm [90], the same algorithm used for the ClosTron intron design for gene knock-out in *Clostridium* [98]. Once the *attR* is recognised by the intron II, the Module II (excluding the *LtrA* gene) is inserted in the target plasmid disrupting the *attR* (Fig.4.6). At this point, the plasmid will not be accessible to store another signal and will have instead the *attB* and *attP* sequences.

$$\textit{retrotransposon} : \text{attR} \longrightarrow \text{attB} \times \text{attP}$$

The $\phi$C31 site specific recombinase recognises the specific sequence *attB* and *attP* on the target plasmid and irreversibly recombines them, to reset the plasmid and rebuild the *attR*, as shown in Fig.4.7:

$$\textit{Integrase} : \text{attB} \times \text{attP} \longrightarrow \text{attR}$$

A schematic of the accumulation of the signal is shown in Fig.4.8:

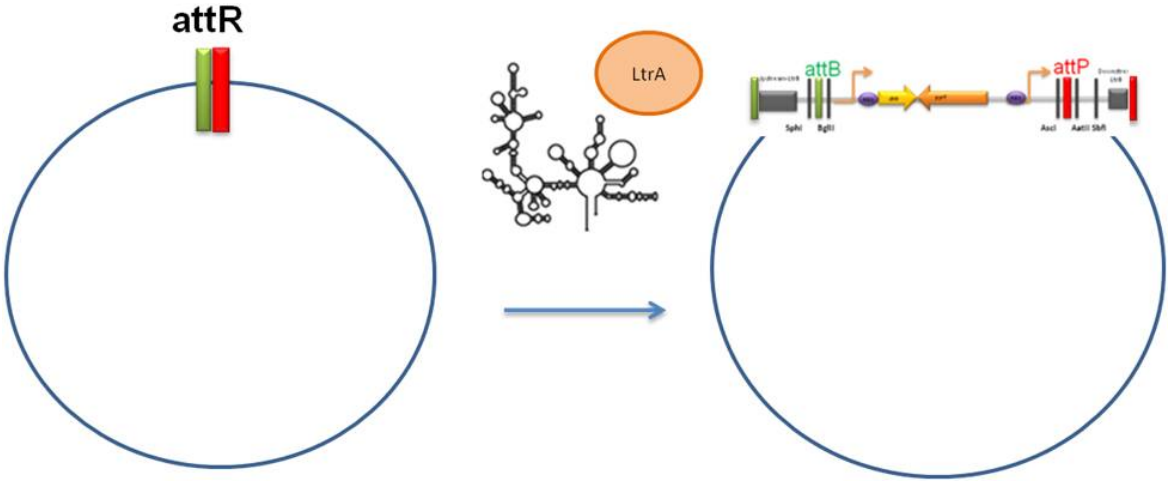



Figure 4.6: **Retrotransposon attR targeting. Left**: *attR* sequence on the target plasmid. **Right**: The retrotransposase is produced and the remaining RNA folds into a retrotransposon which target the *attR* site on the target plasmid and inserts the Module II carrying *attB* (green rectangle) and *attP* (red rectangle). The system is now locked.

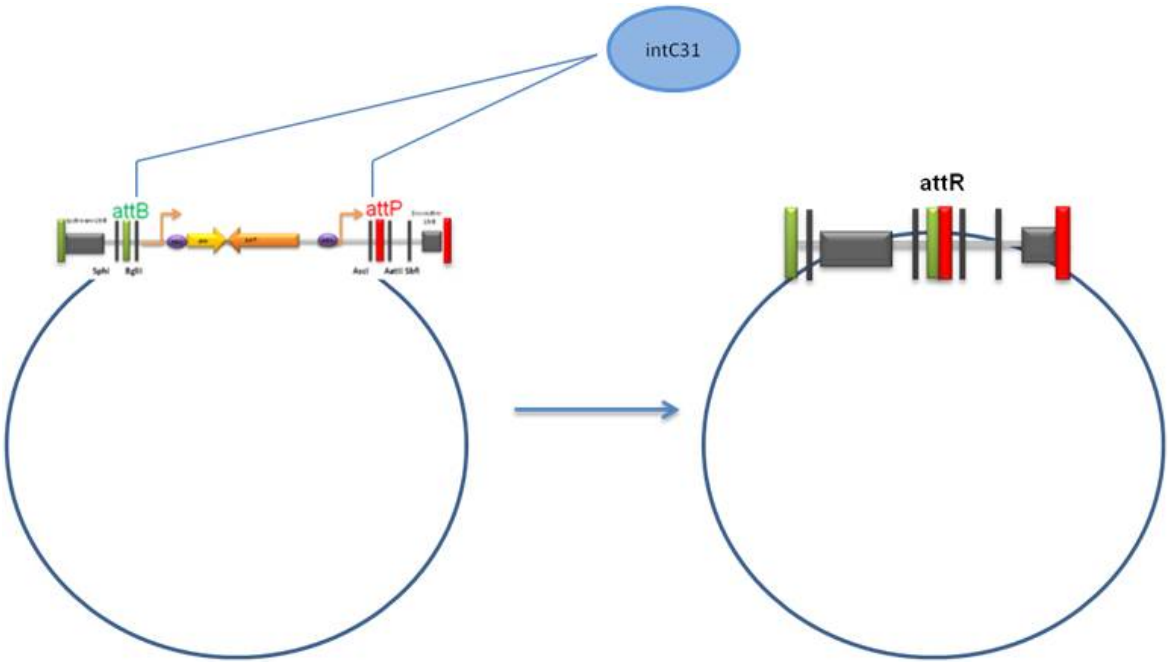



Figure 4.7: **Integrase attB and attP targeting.** The int-$\phi$C31 recombinase controlled by the arabinose promoter (Para) induce the integrase to recombine *attB* and *attP* to reconstruct *attR* site.

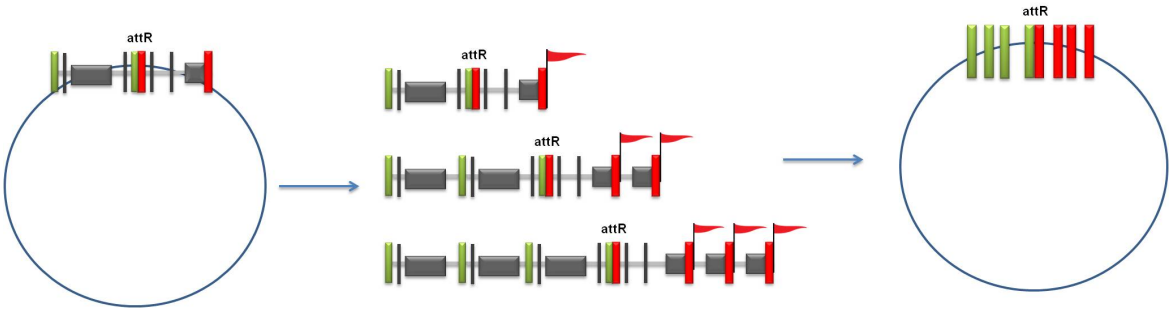



Figure 4.8: **Accumulation of the signal. Left**: the int-$\phi$C31 reconstruct the *attR* site. The system is now unlocked and can accept a second signal. **Centre**: representation of the accumulation of the signal (red flag) after 3 cycles of recording. **Right:**a schematic of the target plasmid with signal accumulation.

## 4.3 Tape Operations

The tape system is designed to function in growing *E. coli* Δ*pyrF* cells undergoing cycles formed by three consecutive steps: (i) standby mode, (ii) signal detection and recording and (iii) system reset to standby mode.

### *4.3.1 Standby mode*

The Tape data structure system incorporates the tac promoter which is a hybrid between the strong trp and the inducible lac promoters [99]. In standby mode, Ptac is repressed by the LacI protein and transcription of the engineered tape Ll.*LtrB* retrotransposon directly attached to it does not occur. Approximately 0.9-kb downstream of the transcription initiation point of the Module II Ll.*LtrB* retrotransposon [100], a dihydrofolate reductase gene (*dhfR*) constitutively transcribed from its own promoter confers resistance to the cells to trimethoprim (TpR). Trimethoprim acts by interfering with the action of the dihydrofolate reductase gene naturally encoded by the *E. coli* chromosome, inhibiting synthesis of tetrahydrofolic acid, which is an essential precursor in the de novo synthesis of the intermediate thymidine monophosphate (dTMP), precursor of DNA metabolite thymidine triphosphate. The dihydrofolate reductase gene *dhfR* present in the tape is however insensitive to trimethoprim, allowing for the selection of cells carrying it. Transcription of the remaining part of the tape Ll.*LtrB* from the TpR promoter does not generate a functional retrotransposon as it requires the upstream 0.9-kb to be fully functional. Immediately downstream of the

*dhfR* gene lies in the opposite orientation the *pyrF** gene, constitutively transcribed from its own promoter. The *pyrF** gene does not code for a functional protein as it is interrupted by a self-splicing group I intron that has been oriented so that it only excises itself when transcribed in the opposite orientation to this gene. A functional *pyrF* gene codes for orotidine-5'-phosphate decarboxylase, an essential enzyme for pyrimidine biosynthesis and prototrophy. As the host carrying the tape system is by design a Δ*pyrF* mutant, cells carrying the tape Ll.*LtrB* retrotransposon in standby mode are not only uracil auxotrophs, but also resistant to the metabolic poison 5-fluoroorotic acid (5-fluorouracil-6-carboxylic acid monohydrate; 5-FOA) which would be otherwise converted by functional PyrF enzymes into toxic analogues of 5-fluorouracil. Thus, *E. coli* Δ*pyrF* cells carrying the tape system in standby mode have been designed to be selectable on minimal medium in the presence of 5-FOA and trimethoprim.

### *4.3.2 Signal detection and recording*

Addition of the inducer IPTG inactivates the LacI repressor. IPTG binds LacI, inducing a conformational change that prevents it from binding DNA and allowing transcription from the tac promoter to proceed. Thus, upon detection of this signal the tape *Ll*.LtrB retrotransposon is transcribed into RNA, including the regions encompassing the *dhfR* and *pyrF** genes. Transcription of the *pyrF** genes in this orientation causes its group I self-splicing intron to excise, thus generating an RNA sequence corresponding to a functional *pyrF* gene but on the wrong strand to be coding for a protein. Further downstream of *pyrF* the IPTG-induced transcript then encodes the LtrA retrotransposase which is produced. This multi-function protein will then perform several tasks with the tape *Ll*.LtrB RNA transcript which will itself fold into a ribozyme cofactor. Notably, by a process named retrohoming these two components will (1) splice the part corresponding to *ltrA* out of the transcript, (2) bind to a specific DNA target sequence (*attR*), (3) convert the processed transcript into cDNA by reverse transcription and (4) integrate the cDNA into the target sequence, irreversibly disrupting it (4.6). The part that integrates into the target sequence as double-stranded DNA includes the *dhfR* and the now functional *pyrF* genes flanked by remains of the tape *Ll*.LtrB RNA transcript. Hence, cells having successfully responded to the IPTG and recorded that signal in such a way can be selected on minimal medium in the presence of trimethoprim and

in the abscence of uracil. As an initial setup and to facilitate analysis and troubleshooting of the system, it was decided that the target would be carried by pACYC184, a low-copy number plasmid, moderately unstable cloning vector that confers resistance to chloramphenicol [101].

### *4.3.3 System reset to standby mode*

The *attR* target sequence has been disrupted and the system is now locked, thus implying that no additional signals can be recorded. To reset it, a mechanism to (1) excise the *dhfR* and *pyrF* genes flanked by remains of the retrotransposon and (2) reconstitute a new target sequence was incorportated. This is accomplished in a single step by (1) flanking the *dhfR* and *pyrF* genes by site-specific recombination sites (SSRS) and (2) making the SSRS that results after recombination the target of the tape Ll.*LtrB* retrotransposon [86] (Fig.4.7). After searching for SSRS that could be amenable to targeting by *Ll*.LtrB retrohoming [97], *attR* from bacteriophage $\phi$C31 was found to be the best if not the only suitable candidate. Therefore, to allow the system that was locked after the signal detection and recording to be reset, the *dhfR* and *pyrF* genes were made to be flanked by *attB*$\phi$C31 and *attP*$\phi$C31 respectively, and attR$\phi$C31 was cloned in pACYC184 [101] to constitute the recording plasmid, with the necessary modifications for adequate *Ll*.LtrB function. Finally, the gene encoding the $\phi$C31 recombinase was then placed under an arabinose-inducible pAra promoter accompanied by the *araC* [102] repressor gene. To reset the system, cells are therefore exposed to arabinose which will induce the $\phi$C31 recombinase and excise the *dhfR* and *pyrF* genes from the recording plasmid, rendering the cells again auxotroph for uracil and resistant to 5-FOA. Flanking the *dhfR* and *pyrF** genes in the tape system by *attB*$\phi$C31 and *attP*$\phi$C31 and inducing the $\phi$C31 recombinase will also cause unwanted recombinations, *e.g.* between the *attB/attP* sites present in the chromosomal part of the system, or between the sites present in the chromosome and the plasmid. However, none of these unwanted recombinations will end up causing uracil auxotrophy and resistance to 5-FOA, allowing to select only the cells having performed the desired operation. Selection by the trimethoprime resistance gene *dhfR* in the chromosome further guarantees the maintenance of the chromosomal element of the system.

## 4.4 Conclusions

Currently there are no computational modelling platforms available to simulate complex next generation of synthetic circuits based on the heavy use of the rewriting of the genetic structure of the synthetic device via, *e.g.* retrotransposons, such as the one I specified and designed here. The available platforms can model transcription and translation processes but not complex recombination events. Thus, this represents an issue in modelling the re-writing by the core of the tape and represents a future opportunity for computational Synthetic Biology research. The engineering and characterization of this system will be discussed in Part III, chapter 6 and Part IV, chapter 8 of this dissertation. This work represents the primary steps on the specification and design for a potential *in vivo* tape data structure.

# Part III

# Chapter 5

# Engineering of an In Vitro DNA Stack Data Structure

This chapter offers a complete overview of all the methodologies used to engineer and implement the *in vitro* DNA data stack prototype. The usage of a novel electrophoresis method will be introduced and the limitations and advantages discussed. Furthermore, Transmission Electron Microscopy was used to image the final product and the application of the Molecular Beacons was begun to be explored.

## 5.1 Implementation of the DNA "Bricks"

The DNA Stack implementation is achieved through ssDNA bricks encoding data and operations:

- *Start*: data brick designating the beginning of the stack data structure.
- *Push*: operator brick to initiate subsequent signal data storage within the stack.
- *Write*: data bricks that can be stored within the stack.
- *Pop*: data brick that undoes the rightmost *push* operation.
- *Read*: operator brick that removes the rightmost *write* operation.

All the bricks are provided with toeholds domains which should promote irreversible hybridization and hairpin loops long enough to promote stable stems but also short enough to be easily accessible and obtain a fast branch migration.

The bricks nucleotide sequences for the data and operations were provided by my collaborators Dr. Harold Fellermann and Jerzy Kozyra (Table 5.1). Nucleotide sequences were obtained with a custom-made genetic algorithm implemented with the open-source genetic programming framework *inspyred* [103–106]. In the algorithm, genotypes consist of candidate primary sequences for all domains of the design. The fitness of an individual is evaluated based on two factors: desired secondary structure and binding energies. Bricks sequences folding, design and structure were tested using the Vienna RNA secondary structure program RNAcofold [107, 108] with DNA interaction parameters. Energy landscapes were calculated with the oxDNA tool [46].

| bricks | Sequence (5'→3') | Purification |
|---|---|---|
| start | GGGAGAGCCATCTCCCGCACACACTTC - [Bio] | PAGE |
| SmaI-start | GGGAGAGCCATCTCCCGCACACACTTC | |
| | GCGAACCCGGGGCTC - [Bio] | PAGE |
| SmaI | [Bio] - GAGCCCCGGGTTCGC | PAGE |
| SwaI-start | GGGAGAGCCATCTCCCGCACACACTTCCCGACATTTAAATCAGC - [Bio] | PAGE |
| SwaI | [Bio] - GCTGATTTAAATGTCGG | PAGE |
| push | GAAGTGTGTGCGGGAGATGGCTCTCCCGAAGTGGTGTCCGCC | PAGE |
| | GGGCAGCGGCGGTTGGTCTCCC | |
| write_x | GGGAGAGCCATCTCCCGCACACACTTCGGGAGACCAAATTAGTAGGTAGAC | PAGE |
| | AAAAAAAGACCGCTAAACTCTAATCACACCTACTAATACACCACTTC | |
| write_y | GGGAGAGCCATCTCCCGCACACACTTCGGGAGACCAAATTAGTAGGTAGAC | PAGE |
| | AAAAAAATACTGCTTTACTCTAATCACACCTACTAATACACCACTTC | |
| Long write $\gamma$ | GGGAGAGCCATCTCCCGCACACACTTCGGGAGACCAAGCACGCTCGAGCTC | PAGE |
| | GTATCGCAGTAATTTTTTTCCATACTGCTTTACATCCCTATATACTGCGAT | |
| | ACGAGCTCGAGCGTGCACACCACTTC | |
| read | GAAGTGGTGTTTGGTCTCCCGAAGTGTGTGC | PAGE |
| pop | GGGAGACCAACCGCCGCTGCCCGGCGGACACCACTTCGGGAGAGCCATCTC | PAGE |
| | CCGCACACACTTC | |
| report_x | [Bio] - TATGACTGCAATTTAGCGGTCT | PAGE |
| report_y | [Bio] - TATGACTGCAATAAAGCAGTAT | PAGE |
| start | 5' GGGAGAGCCATCTCCCGCACACACTTC - Cy3 - 3' | RP-HPLC |
| push | 5' [BHQ-2] - GAAGTGTGTGCGGGAGATGGCTCTCCCGAAGTGGTGTCCGC- | RP-HPLC |
| | CGGGCAGCGGCGGTTGGTCTCCC 3' | |
| Anti-Start | 5' GGGAGATGGCTCTCCC -[BHQ-2] - 3' | RP-HPLC |
| Start | 5' - [Cy3] - GGGAGAGCCATCTCCCGCACACACTTC 3' | RP-HPLC |

Table 5.1: Bricks sequences set used in this study. For domains specifications see Table 3.1 Chapter 3.

### 5.1.1 Experimental Manipulation of DNA

DNA bricks were requested on a 100$\mu$M synthesis scale, and provided by Eurogentec (Belgium), with a standard desalting procedure or a required denaturing polyacrylamide gel electrophoresis (PAGE) purification for bricks longer than 50 nucleotides and/or any 3', 5' modification. Streptavidin coated gold nanoparticles of 10 and 5 nm diameter were supplied by Life Technologies (Alexa Fluor 488 streptavidin), while 20 nm colloidal gold conjugate was provided by Sigma Aldrich. Samples and stock solutions were stored at -20°C. Standard molecular biology procedures were used [109]. Reactions have been set in a chronological order to limit events of unspecific hybridization. (initially, *start* and *push* brick are mixed together and successively *signal* is added and left to react with the hybridized *start-push* complex).

The DNA stack was prepared using aliquots of 200 nM of each brick. DNA samples were dissolved in a total volume of 20$\mu$L of nuclease free water and 50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate, pH 7.9 buffer at room temperature (25 °C) and incubated for 10 minutes if not otherwise specified. The mixture was shaken at 300 revolutions per minute in an Eppendorf Thermomixer Comfort set at 25°C. Agarose gel electrophoresis was carried out in a 2% agarose gel in 1x Tris-Acetate-EDTA buffer (TAE).

### 5.1.2 On-chip electrophoresis

In order to obtain quantitative and qualitative measurements of the DNA stack data structure the usage of a capillary electrophoresis has been instrumental. The Agilent 2100 Bioanalyzer system provides size, quantification and quality control of nucleic acids on a single platform. Macromolecules are separated in a linear polymer gel according to their charge and molecular weight by applying a high voltage to sample solutions. The kit used was the DNA "High Sensitivity Chip" provided by Agilent Technology and adhered manufacturer protocols. The kit provides a dye that fluoresces upon intercalation with dsDNA and to a lesser extent with ssDNA. The system can load up to 11 samples which can be run concurrently on a disposable chip within 30 minutes. The aliquot required is just 1 $\mu$l, added at the very last step of the protocol. A volume of 9$\mu$l Gel-dye mixture is used to fill the wells before the sample is added. Pressure is applied with a disposable 1-mL syringe to let the

gel-dye matrix enter in the chip microchannels (Fig.5.1).

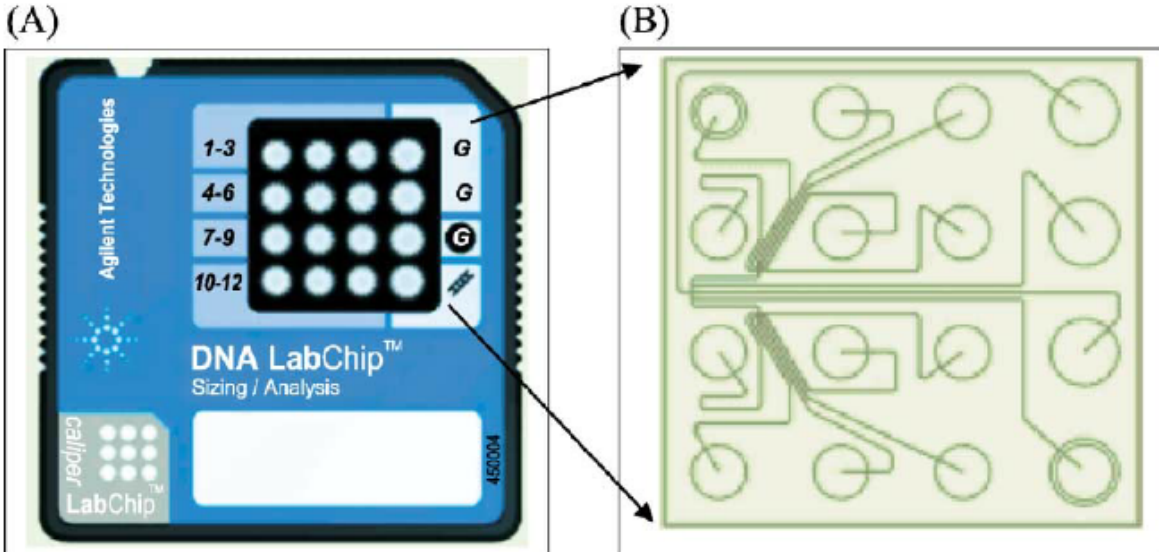



Figure 5.1: **Agilent DNA chip**.Panel A shows a standard DNA High Sensitivity chip. Panel B shows the chip microchannels. Taken from Agilent user Manual

The kit provides lower and higher markers (35bp and 10.000bp) which the software uses to align sample solutions with a DNA ladder of known composition that is run in a separate lane. The ladder range is 50-7000 bp. The software displays samples gel-like images and electropherograms. The electropherogram plots the raw data as arbitrary fluorescents units displayed against either time (seconds) or migration of the fragments (base pairs). Background fluorescence will also be included in the plot that will show the samples as peaks in between the markers (Fig. 5.2).

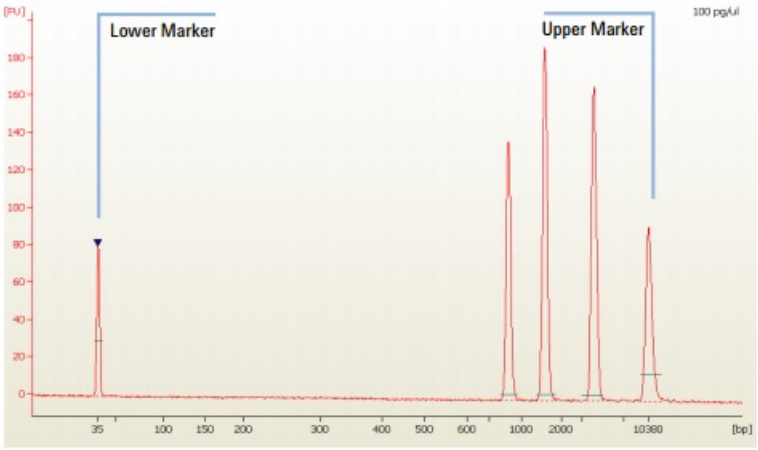



Figure 5.2: **Agilent Electropherogram plot**. Typical Electropherogram resulting from Agilent 2100 Bioanalyzer software. Taken from Agilent user Manual

The raw data is measured and displayed in the form of an electropherogram that plots the arbitrary fluorescents units displayed against either time (seconds) or migration of the fragments (base pairs). Background fluorescence will also be included in the plot. The data will include measurements about the following features:

- **Aligned Migration Time:** This is measured in seconds. The software detects the DNA fragments' migration time according to their size. The migration of small samples is faster than bigger ones, as it happens in a standard agarose electrophoresis gel.

  Differences in sample structure can affect the pattern of time migration, and the same samples can exhibit a different time migration.

- **Fragment Size:** the kit provides a size quantification of the DNA samples in base pairs (bp) units. High sensitivity kit DNA size range is 50 – 7000 bp. Quantification is restricted when working with ssDNA samples, ssDNA-dsDNA hybrids structures, and the presence of secondary structures (*e.g.*, hairpin stem and loops).

- **Area under the peak:** The area for each recognised peak is only calculated if this is aligned with the ladder.

- **Mass Concentration:** nucleic acid concentration is measured in ng/$\mu$L. The software initially calculates the concentration of the lower and upper marker in order to set internal standards and align them with the ladder. The curve under the peak is calculated and the relationship between area, concentration and markers will permit the estimation of each sample concentration. The first peak found will correspond to the lower marker, whilst the last peak found will be assigned to the upper marker. After this, the other peak concentrations can be estimated.

  If unexpected peaks are shown in the electropherogram, and the marks have been set in an incorrect way, one can manually choose a peak to use as a reference marker.

- **Molar Concentration:**

$$Molarity = \frac{Concentration \times 10^6}{660 \times Size}$$

It is measured in nanomoles per liter (nmol/L) and it is calculated by combining size and concentration measurements. The software assumes that the molecular weight of each singular base pair is 660 ng/nmol.

Agilent Bioanalyzer 2100 technology was chosen in order to achieve a good quality gel-like image and to generate a detailed plot (electropherogram) with peaks correspondent to the sample involved.

It is possible to go from the peaks shown in the electropherogram to the actual quantification (expressed as molarity) involved in the reactions.

#### 5.1.2.1 Technology limitations

In this study the Agilent High sensitivity DNA kit was used to assay the samples. Due to the complex secondary structures and ssDNA-dsDNA ratio of our bricks set, the kit does limit a quantitative analysis, but offers a significant approach towards an advanced qualitative analysis. The measurements successfully discriminate the migration times of almost all strands (disregarding *report* strands) with significant differences. Only *start* and *read* cannot be reliably differentiated. As expected, shorter oligomers register with shorter migration times. Striking discrepancies between the known brick sizes and the sizes derived by the software from comparison to the ladder might be attributed to two reasons: firstly, short oligomers such as *start*, *read* and *report* are well below the detection limit of the high sensitivity kit, which can resolve dsDNA fragments between 50 - 7000 base pairs in length, with an accuracy of $\pm 10\%$ from 50 to 600 bp and $\pm$ 20% from 600 to 7000 bp (according to manufacturer specifications). Secondly, the reported deviations might lie in the fact that our bricks contain extensive secondary structures that might affect their motility in the gel matrix. A similar discrepancy is observed in the derived molarity values. This is partly due to the fact that molarity calculation is based on the base pair estimation and will thus suffer from the issues described before, partly because our bricks contain extensive ssDNA regions which interact differently with the fluorescent dye than dsDNA (for which the analysis kit has been designed). Going further, could be possible to explore RNA chips and reagents. However, this is not something that Agilent has validated thus is not possible to know how accurate the quantification would be.

### 5.1.3 Imaging

Gold nanoparticles have been used in this study to label the *reporters* ssDNA (carrying biotin). They bind the signals internally forming Au-DNA conjugates, which can be imaged (as black dots) using TEM. These widespread particles represent an advantage to fluorophores, for their combination of optical, chemical and electrical properties. They are capable of both absorbing and scattering visible and near infrared light [110].

Samples were examined using a Philips CM 100 Compustage (FEI) Transmission Electron Microscope and digital images were collected using an AMT CCD camera (Deben) provided by the Electron Microscopy Research Services at Newcastle University. A volume of 5$\mu$L sample was applied on glow discharge grids preliminary washed with 0.5 mM Magnesium chloride to change the hydrophilic surface charge orientation. Moreover, sample were tested with and without the usage of a staning agent (2% aqueous Uranyl Acetate) or magnesium. In order to detect more than one signal/report at the same time and the stack starting point, biotin-TEG labeled strands were attached to colloidal AuNPs streptavidin coated of 5, 10 and 20 nm size provided by Life Technologies and Sigma Aldrich, at room temperature and then stored at 4°C protected from light. Gold nanoparticles were chosen to be reporters for their strong electron density, a characteristic that makes them favourable for TEM microscopes imaging. They are detected black spots on the grid surface [111].

### 5.1.4 Molecular Beacons design and experiment

Molecular Beacons Cy3/BHQ-2 (Table 5.2) were used in this study to further investigate the hybridization of the *start* and *push* complex using an additional *anti-start* brick. The modified bricks were provided by Eurogentec where Cy3 in 5' is incorporated directly during the synthesis process, as Amidite whereas Cy3 in 3' is coupled manually through a C6 amino linker.

The hybridization reaction was monitored over time by directly labelling the *start*, *anti-start* (Fig.5.3, left) and the *push* brick with a Cy3 fluorophore or a BHQ-2 quencher, as showed in Fig.5.3, right. The fluorescence data were used to determine the rate constants for the hybridization process. The 96-Well Microplate Greiner F-Bottom uses black microplates with a volume of 50 $\mu$l to reduce background autofluorescence. Unused wells were filled

with the reaction buffer sample in triplicate Tecan Sapphire2 Spectrophotometer (Tecan), at 485 nm excitation and 510 nm emission, 5 nm slits, gain = 90 V. The experiment has been designed such that the quencher will absorb the excitation energy from the fluorophore. This means it will not show any fluorescence unless the complex is displaced.

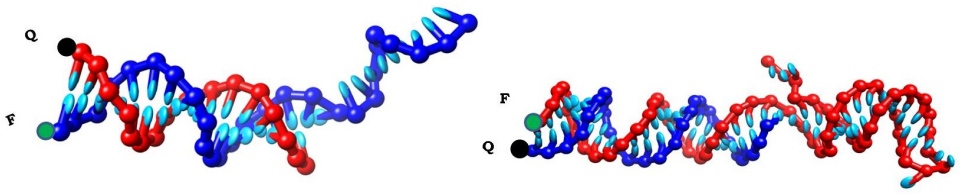



Figure 5.3: **Molecular beacons oxDNA simulation**. Left: In blue is shown the *start*, with the Fluorophore (F) at the 3'end, and in red the *anti-start*, with the Quencher (Q) at the 5' end. Right: In blue is shown the *start*, with the Quencher (Q) at the 3'end, and in red the *push*, with the Fluorophore (F) at the 5' end .

| Fluoropher/Quencher | Excitation - Emission [nm] | Quenching range [nm] |
|---|---|---|
| Cy3/BH2-Q | 550-570 | 550-650 |

Table 5.2: Molecular Beacons used in this study.

## 5.2 Conclusions

All the methods discussed above will be used for the final characterisation and quantification of the stack prototype, and is discussed in the dedicated Chapter in Part IV of this thesis (chapter 7). The next chapter of this thesis will introduce the experimental methods used for the *in vivo* tape data structure implementation.

# Chapter 6

# Engineering of an In Vivo DNA Tape Data Structure

This chapter presents experimental procedures performed to build and implement the *in vivo* DNA tape data recorder. It describes the standard molecular biology and microbiology procedures (*e.g,* agarose gel electrophoresis, PCR, enzymatic digestion, bacterial transformation) that were used to characterise and analyse the correct assembling and implementation of the recorder.

## 6.1 DNA Implementation of the *in Vivo* Tape Data Structure

The tape system was designed to be hosted and replicate within a bacterial platform. During the testing of the system, a variety of *Escherichia coli* bacterial same line strains, for plasmids maintenance and tape replication, have been used, as well as, primers to perform PCR screening or troubleshoot defects in the tape sequence. The following sections present a description of growth conditions, assays and parameters used for the strains, the plasmids, oligonucleotides sequences and usage.

### 6.1.1 *Strains, plasmids and oligonucleotides used in this study*

Bacterial strains are shown in table 6.1. They were grown in Luria Broth (LB media) or minimal media (M9 media) at 37°C overnight with shaking at 250 rpm or left static on agar plates. The Δ*pyrF* strains were grown with uracil supplementation in LB media at 20$\mu$g/mL. The *Escherichia coli* strains with pKD46-RecA$_p$*a* recombineering vectors (**recomb**inogenic eng**ineering**) [112] were grown at 30°C to prevent loss of vector [113]. All of the strains carrying the recorder tape system were grown in the presence of glucose to cause catabolite repression [114]. This reinforces the repressive action of LacI, in order to avoid spurious induction of the system. The following antibiotics were added to the media when required: carbenicillin (100$\mu$g/mL); chloramphenicol (20$\mu$g/mL); gentamicin (20$\mu$g/mL); kanamycin (25$\mu$g/mL); tetracycline (25$\mu$g/mL); nalidixic acid (15 $\mu$g/mL) and trimethoprim (100 $\mu$g/mL). The bacterial glycerol stocks were prepared with 0.5 mL of overnight bacterial culture in a sterile tube with 0.5 mL of sterile glycerol added, and subsequently stored at - 80°C. The plasmids used in this study are listed in Table 6.2. Maps and sequences of each vector are shown in Appendix A. The primers sequences used in this study are listed in Table 6.3. Primers Tn7L and Tn7R were designed by *Choi et. al.*,[115]; primers attTn7-1 and attTn7-2 were designed by *Schweizer et.al.,* [116]. Plasmids and primers were always stored at -20°C.

| Strains | Description |
|---|---|
| *Escherichia Coli* DH5$\alpha$ | F- $\phi$80lacZΔ15 Δ(lacZYA-argF)U169 deoR ΔrecA1 endA1 hsdR17(rk-, mk +) phoA supE44 thi-1 gyrA96 relA1 $\lambda$- [117]. |
| TLoST01 | DH5$\alpha$ ΔlacI ΔsdiA ΔpyrF, grown at 30°C. Origin of replication: R101. TcR or GmR . |
| TLoST02 | TLoST01:: tape . |
| Tape 2.0 | JW1273:: Tape (with Module I repaired). The strain implements the tape recorder (chapter 4, Fig. 4.5) |
| CC118($\lambda$pir) | $\lambda$pir strain for cloning miniTn7::Tape [118]. |
| S17-1($\lambda$pir) | TpR, SmR, recA, thi, pro, hsdR-M+RP4: 2-Tc:Mu: Km Tn7 $\lambda$pir.Strain for maintaining and conjugating R6K replicons [119] |
| BW25113 | F-, Δ(araD-araB)567, ΔlacZ4787::rrnB-3, $\lambda$-, rph-1, Δ(rhaD-rhaB)568, hsdR51 [120] [121]. |
| JW1273 | *E. coli* BW25113 derivative ΔpyrF789::KmR [121] |

Table 6.1: **Strains used in this study**. Strains indicated as $\lambda$pir contains the *pir* gene necessary for the cloning and propagation of plasmids with R6K origin of replication. Knocked out genes are indicted as Δ.

| Plasmid | Description |
|---|---|
| pKD46-RecA$_p a$ | Thermosensitive plasmid, provides transiently the RecA recombinase. |
| pTNS2 | helper plasmid pTNS2 for the transposition of miniTn7 elements, provides Tn7-specific transposase genes. |
| pACYC184 | Cloning vector, CmR, TcR, p15A origin of replication, low copy number. |
| pACYC184::attR | CmR, this plasmid is the target "blank" recorder. AttR is formed by annealing oligonucleotides AttR-F and AttR-R and ligating into pACYC184 cut with BamHI and SphI (disrupts TcR) |
| pBluescript II KS | Cloning vector, ColE1 replicon, CbR, ApR, lacZ, allowing selection via blue-white screening. |
| pBSTape | Tape system (SacI-XhoI) cloned in pBluescript, ApR |
| pUC18R6K-mini-Tn7 | Cloning vector, R6K replicon, TcR, propagated only in $\lambda$pir strain contains the *pir* gene. |
| pTape-01 | Tape system (SacI-XhoI) in miniTn7 delivery plasmid (pUC18R6K-mini-Tn7),TcR, ApR, R6K origin of replication |

Table 6.2: **Plasmids used in this study**

| Oligos | Sequence (5'→3') |
|---|---|
| PyrFcF | TTCTTCCCGCGCTGTTAC |
| PyrFcR | CGCTGTAAAGAGGCGTTGA |
| Tn7L | ATTAGCTTACGACGCTACACCC |
| Tn7R | CACAGCATAACTGGACTGATTTC |
| attTn7-1 | GATGCTGGTGGCGAAGCTGTC |
| attTn7-2 | GATGACGGTTTGTCACATGGAG |
| AttR F | GATCCTCGAGTGAGGTGTAGAA CGCGCCCGGGGAGCCCAAAGG TTACCCCAGTTGGGGCACGGCATG |
| AttR R | CCGTGCCCCAACTGGGGTAACCTT TGGGCTCCCCGGGCGCGTTCTACA CCTCACTCGAG |
| ModI R | GCGGCCGCTGTGGAATTGTGAGCGC TCACAATTCCACACATTATACGAGCCG |
| ModI F | AACGAAAGGCTCAGTCGAAA |

Table 6.3: **Primers used in this study**

### 6.1.2 Media and growth conditions

It is known that *E. coli DH5α* exhibits inferior growth in minimal media (M9) compared to other *E. coli* strains [122] resulting from unknown accumulated mutations of the *purB* gene, involved in the *de novo* purine nucleotide pathway. For this reason the M9 minimal medium used was improved by adding adenine. M9 basic salts were autoclaved and used with the following working concentrations: potassium phosphate monobasic ($KH_2PO_4$) 3g/L, sodium phosphate dibasic ($Na_2HPO_4$) 6g/L, ammonium chloride ($NH_4Cl$) 1g/L and sodium chloride (NaCl) 0.5g/L. Glucose was used as the carbon source and the media was supplemented with 100 mg/L D-Ca panthenate, 100 mg/L Choline c-hloride, 200 mg/L i-Inositol, 100 mg/L Pyridoxine-HCl, 10 mg/L Riboflavin, 100 mg/L Thiamine-HCl,2 mM Magnesium sulfate ($MgSO_4$) and casaminoacids 1%. Bacteria were grown in M9 media optimized at 37°C overnight with shaking at 250 rpm or left static on agar plates.

### 6.1.3 5-FOA Medium

This assay is used as a screening method to identify strains that carry a deleted *pyrF* gene. This gene encodes for an essential enzyme in the pyrimidine biosynthetic pathway. Cells that lack this activity can grow by uracil supplementation on minimal medium, which can be converted to uridine 5'-monophosphate (UMP) through a salvage pathway. 5-FOA is converted to 5-fluorouracil (5-FU), by orotidylate decarboxylase. 5-FU is uracyl highly toxic analogue as shown in Fig.6.1. Cells that lack the enzyme encoded by *pyrF* are no longer sensitive to 5'-fluorootic acid (FOA) [123], which provides a powerful counter-selection for the loss of the gene. For the preparation of 5-FOA medium, 20 g of agar was autoclaved in 750 ml of water for molecular biology usage. The following reagents were added to a flask containing 250 ml of water: ammonium sulphate ($(NH_4)2SO_4$) 5.0 g; Yeast Nitrogen without amino acid 1.7 g; Dextrose 20 g; 5-FOA 1.0 g. The mix was filter sterilized with a sterile 0.22 $\mu$m filter. Both solutions were put in a water bath at 50°C for 1h. When the solutions are both completely melted the reagent mix containing 5-FOA was added to the agar and poured into sterile plastic petri dishes. These were subsequently dried in a laminar flow hood for 30 minutes and eventually stored at 4° (if not used directly afterwards).

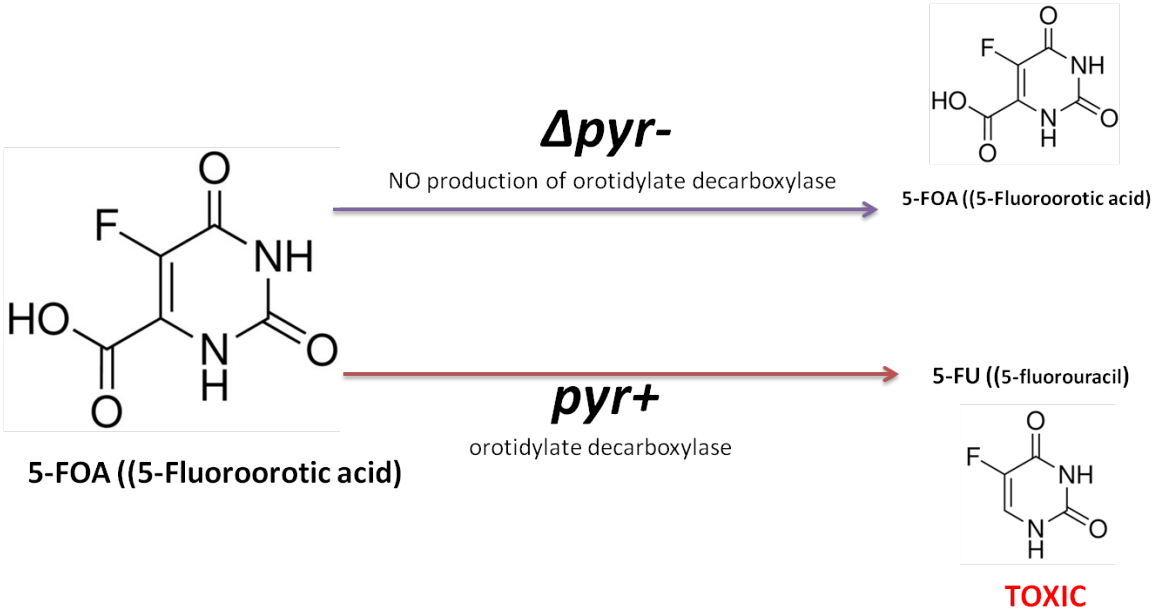



Figure 6.1: Schematic representation of the 5-FOA conversion pathway.

### *6.1.4 Growth assays*

Agrobacterium (AB) minimal medium, RPMI-1640 medium supplied by Sigma Aldrich and M9 minimal medium were used in this study for comparing growth response of *E. coli* uracil auxotroph strains. The media was supplemented with a glucose and vitamins mixture. Three colonies for each strain used as samples were inoculated in a 5 mL culture of LB media and incubated at 37°C in a falcon tube overnight. Cells were collected by centrifugation at 4000 rpm for 5 minutes and washed with phosphate buffered saline (PBS). The washing step was repeated twice and the culture was then resuspended in 5 mL of PBS. Afterwards, the samples were transferred into 5 mL of each minimal media to a final concentration to a final optical density of 0.01 (O.D at 600 nm) and incubated at 37°C. The optical densities were measured at 24h, 48h and 72h. All the experiments were conducted in triplicate and each essay was repeated at least three times.

### *6.1.5 Chromosomal insertion of mini-Tn7 constructs in E. coli*

The pTape-01 suicide delivery vector was co-transformed with pTNS2 helper plasmid encoding for the transposases necessary for the chromosomial insertion. Cells were made electrocompetent using glycerol and used promptly or kept at -80°C until usage in aliquots

of 50$\mu$l. Cells were electroporated with 50 ng of mini-Tn7 and 50 ng of pTNS2 helper plasmid and then transferred into a 0.2 cm electroporation cuvette. BIO-RAD MicroPulser program "EC 2" was used for administering the electric shocks. Cells were recovered in 950 ml of LB medium and incubated at 37°C with shaking 1h 30'. 100 $\mu$l of culture was plated in LB with tetracycline (25$\mu$g/mL) and incubated overnight at 37°C. Colonies were counted and screened the next day.

### *6.1.6 Population dynamics*

A population dynamics test has been used in this study to determine the recorder spontaneous reversion frequency to trimethoprim sensitive (TmS) or *pyrF*+. The recorder strain was grown for 24h to stationary phase in 5 ml of selective M9 liquid media at 37°C with shaking at 200 rpm. At day 1, the previous culture was inoculated in a falcon tube with a fresh 5 ml dose of non selective M9 liquid media at 37°C with shaking to obtain a one million fold dilution (dilution step). 200 $\mu$l of the dilution was plated onto M9 non-selective agar plates and grown for 24h at 37°C with the aim of generating 100 individual colonies (plating step). At day 2, the one million fold dilution step was repeated starting from the previous culture and the same aliquot was inoculated again in 5 ml of non selective M9 liquid media at 37°C with shaking. The plating step was then repeated. All the colonies obtained were picked and screened onto selective and non-selective media in order to determine the percentages of a reverting subpopulation within a population. The entire process was repeated for 5 days (Fig.6.2). This method was used in all population dynamics tests described in this study in order to measure the recorder strain population dynamics under different conditions.

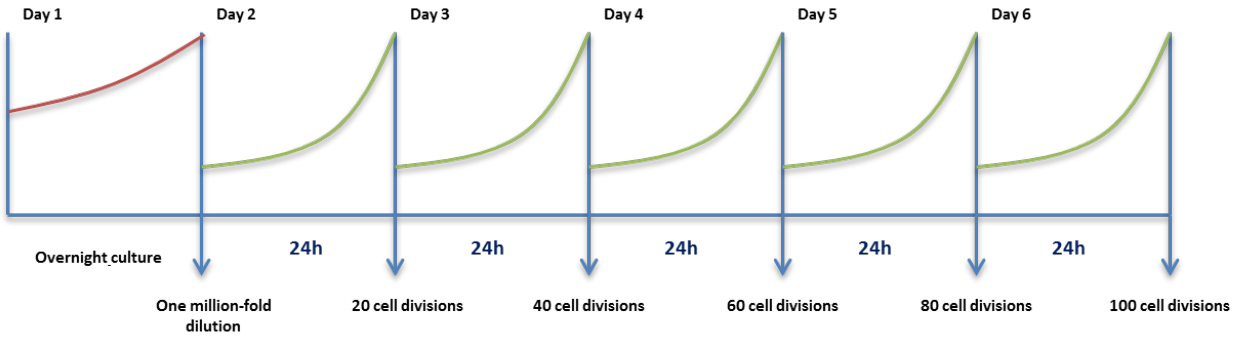



Figure 6.2: **Basic model of dynamic population experiment** The red curve represents the overnight culture to stationary phase. The culture was diluted to one million fold.

### *6.1.7 Molecular Manipulation of DNA*

Standard molecular biology procedures were used for the implementation of the tape. Low copy number plasmids were prepared using the Plasmid Midi Kit (QIAGEN) and Plasmid Midi Kits For Large Scale Purification of High quality Plasmid DNA (Thermo Scientific GeneJET$^{TM}$). High copy number plasmids were prepared with QIAprep Mini Kit (QIAGEN) and Plasmid Mini kit (Thermo Scientific GeneJET$^{TM}$). Chromosome extractions were performed with Wizard® Genomic DNA Purification Kit (Promega). Gel purifications were performed using the DNA gel extraction kit (Zymogen Zymoclean Gel DNA Recovery Kit). We followed the supplier's protocol for every kit.

For restriction enzyme digestions DNA samples were dissolved in total volumes of 20$\mu$L, using the manufacturer's recommended 10x digestion buffer (Promega). Samples were incubated with restriction enzymes at 37°C for 1.5h using 10-20$\mu$L of DNA solution.

Ligation reactions were performed by estimating the vector and insert concentrations using 2$\mu$L of T4 DNA ligase Promega and 2$\mu$L of ligase buffer 10X up to 20$\mu$l of nuclease free water. In order to check the ligation, controls were included.

Polymerase chain reactions (PCRs) of bacterial chromosomal DNA or plasmids were performed with GoTaq polymerase (Promega) according to the manufacturer's manual. Primers were designed manually or using Primer3 software [124] and were synthesized by Sigma Aldrich or Eurofins. For colony PCR bacterial colonies were first diluted in 80$\mu$l of water and heated for 10 minutes at 96°C. Afterwards, 1$\mu$l of the sample was added to the PCR mixture. The PCR programme started with an initial denaturation of 5 minutes at 95°C.

Annealing was performed at 55°C for 1 minute and elongation at 72°C for 1 minute. The whole process was carried out for 33 cycles, ending with a final elongation step at 72°C for 5 minutes. Agarose gel electrophoresis was carried out in a 0.7% agarose gel in 1x Tris-Acetate-EDTA buffer (TAE). Electrocompetent cells were prepared using glycerol or sucrose [109]. The competent cells were used directly; otherwise these were kept at -80°C until usage. The transformation was made by electroporation using BIO-RAD MicroPulser. Competent cells of *E.coli DH5α* [125], *CC118(λpir)*[118], *S17-1(λpir)* [119], *BW25113* [120], [121] and *JW1273* [121] were routinely used for cloning. *DH5α* and *S17-1(λpir)* were used to maintain the construct pBS::Tape.

### *6.1.8 Bioinformatic analysis*

To further validate any problem in the tape sequence, I used the Reciprocal Best Hits (RBH) method. Alignments of the entire tape sequence were run against all the biological parts used to build the system.

### *6.1.9 IPTG induction*

A colony of Tape-2.0, was picked from a fresh plate and growth overnight in 5 ml LB and in presence of two antibiotics: trimethoprim and kanamycin. The strain was induced with IPTG at a concentration of 0 mM (negative control), 0.5 mM, 1 mM, 2 mM and 4 mM and incubated at 37°C for 1.5, 3, 6, 8h and overnight ( 12h). For this assay, all cultures were grown at 37°C in a 50 mL tube with shaking at 250 rpm. In all experiments, the overnight culture was diluted 1:1000 (to 0.01 O.D) in M9 and grown for about 6-8 hours until an O.D of 0.6.

## 6.2 Conclusions

All the methods outlined above will be used for the final characterisation of the tape prototype. This is discussed in the dedicated Chapter in Part IV of this thesis (chapter 8).

# Part IV

# Chapter 7

# Characterization and Results of an In Vitro DNA Stack Data Structure

In this chapter, the experimental results of the operation performed by the *in vitro* DNA stack data structure are presented. Experiments were performed using both standard molecular biology procedures (agarose gel electrophoresis) and a sensitive DNA quantification method (capillary electrophoresis). The structure was visualized with a Transmission Electron Microscope. Finally, the usage of molecular beacons was explored as a method to potentially investigate further. The presented results from the laboratory scale implementation confirm that the predicted DNA stack data structure prototype forms the final expected product.

## 7.1 Agarose Gel Electrophoresis

This section presents experiments conducted with agarose gel electrophoresis, aimed at the characterization of the ssDNA concentration for the stack assembly, validation of data storage and reading out cycles

### *7.1.1 Characterization of the bricks concentration*

Agarose gel electrophoresis was used to determine the minimal concentration of DNA bricks necessary to carry on the assembling of the DNA stack data structure. The experiment has

been run on agarose gel, starting from a concentration of 50 nM for each brick and gradually increasing up to 250 nM. The minimal concentration detected was 200 nM and was used as reference in all the experiments described in this chapter, if not otherwise specified. A representative result is shown in Fig. 7.1.

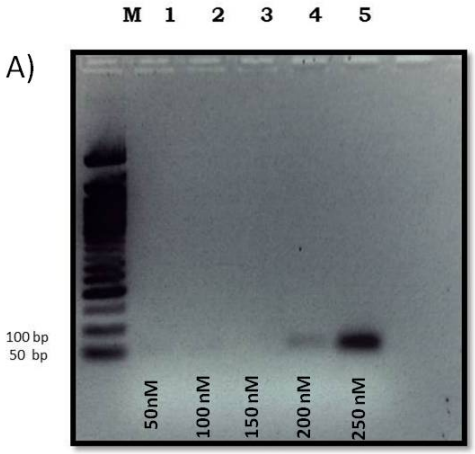



Figure 7.1: **Agarose gel Electrophoresis**. Electrophoresis gel showing results for a different range of concentration. In lane M is showed the NEB Low Molecular Weight DNA Ladder. Lane 1 to 5 show the *push* brick at a concentration of 50 nmol, 100 nmol, 150 nmol, 200 nmol, 250 nmol; (panel B) In lane M, NEB Low Molecular Weight DNA Ladder.

## *7.1.2 Data storage experiments*

### 7.1.2.1 DNA stack data storage of One Signal: Start-Push-Write X

Agarose gel was also critical to first prove the progressive elongation and shrinkage of the DNA stack data structure during the pushing and popping operations. The electric field moves the charged bricks through the agarose matrix according to their size, and, when they hybridize correctly, is clearly visible a delayed electrophoresis mobility of the assembled components during the pushing step, occurring via progressive elongation (Fig:7.2).

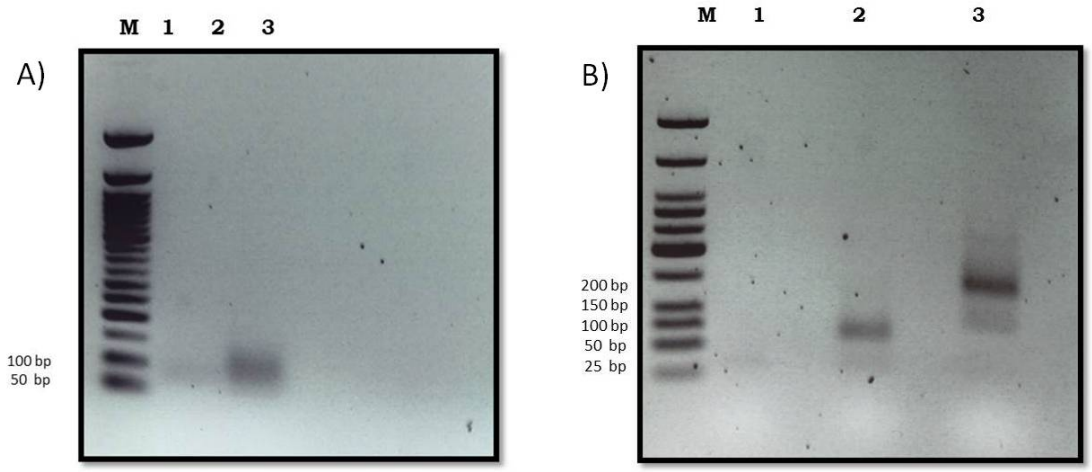



Figure 7.2: **Agarose gel Elechtrophoresis**. (panel A) In lane M is showed the NEB Low Molecular Weight DNA Ladder. Lane 1 shows the *start* (27nt) brick at a concentration of 200 nmol; In Lane 2 *push* (64 nt) brick was added and reacted with the start, forming a complex of roughly 91 bp. (panel B) In lane M, NEB Low Molecular Weight DNA Ladder. Lane 1: lane 1: *start* (27nt); lane 2 *start+push* (91 nt); lane 3 *start-push-write x* roughly (189 nt). Agarose gel 2%.

### 7.1.2.2 DNA stack data storage of eight write X signals

In Fig 7.3, the experiment consisted in the recording of eight write X. Although the gel presents smears background, it is possible to observe the elongation of the stack, and a slower mobility of the assembled ssDNA.

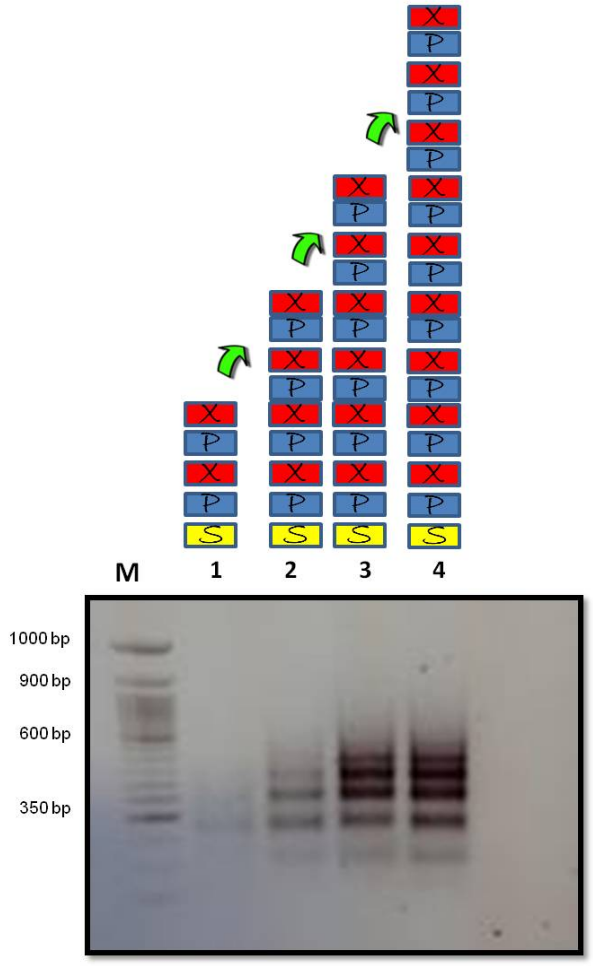



Figure 7.3: **Agarose Gel Electrophoresis of the data storage process** Lane M = NEB 1Kb ladder. Lane 1=SPXPX (2 Signals, 351 bp); Lane2=SPXPXPXPX (4 Signals, 675 bp); Lane 3=SPXPXPXPXPXPX (6 Signals, 999 bp) ; Lane 4=SPXPXPXPXPXPXPXPX (8 signals, 1.323 bp) ; Data obtained from four parallel experiments. On top of the Gel image, a schematic representation of the bricks involved in the 8 signals stack data storage.

### *7.1.3 Reading Out Experiments*

In Fig 7.4, the experiment consisted in the data storage of three write signals and the reading out of two signals. It is easy to see that after a progressive elongation, follows a faster bands migration, when the DNA stack data structure is subject to shrinkage (reading). Smears in the gel suggests the presence of intermediates species and/or a possible unspecific bricks hybridization.

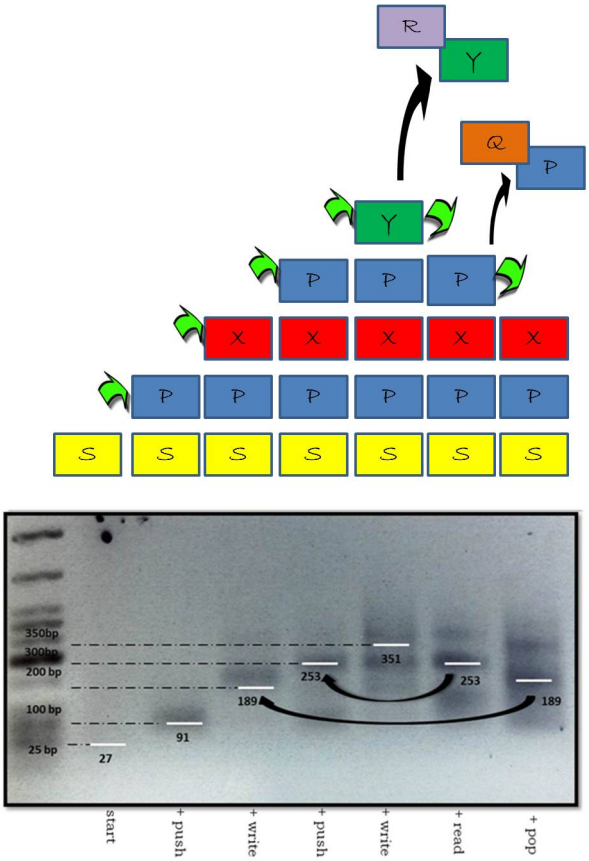



Figure 7.4: **Electrophoresis gel showing the pushing and popping products**. In lane M is showed the NEB Low Molecular Weight DNA Ladder. Lane 1 to 5 show the data structure and the implementation of the *pushing* operation: lane 1: start (27nt); lane 2 start+push (91 nt); lane 3 start+push+signal (189 nt); lane 4 start+push+signal+push (253 nt); lane 5 start+push+signal+push+signal (351 nt). Lane 6 and 7 show the *popping* operation:lane 6 goes back to start+push+signal+push (253 nt), peeling off the last signal exposed and in lane 7 goes back to start+push+signal (189 nt), peeling off the last push exposed. Strand concentration 200 nM, agarose gel 2%. On top of the Gel image, a schematic explanatory representation of the bricks involved.

## 7.2 Capillary Electrophoresis

This section presents experiments conducted with capillary electrophoresis, aimed at the characterization of the ssDNA concentration, calibration of the bricks and validation of data storage and reading out cycles

### *7.2.1 Characterization of the bricks concentration*

An additional investigation for the characterization of the concentration was run by capillary electrophoresis (with the commercial Agilent 2100 BioAnalyzer) used as a more precise and sensitive system detection than the standard agarose gel. The instrument was able to detect samples down to 25 nM. A representative result conducted on the *start* brick is shown in Fig. 7.5.

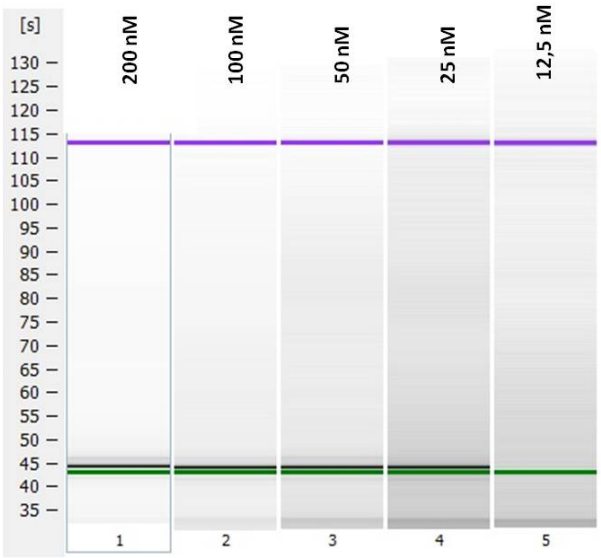



Figure 7.5: **On-chip electrophoresis start concentrations.** Agilent 2100 Bioanalyzer comparison tool, has been used to facilitate the analysis of the bands detected from each sample concentrations provided to the instrument. Below 25 nM the machine, seems to not detect the ssDNA bricks samples.

### *7.2.2 Single Brick Calibration*

Capillary electrophoresis measurements were performed for all individual bricks in order to determine the response of the Agilent 2100 Bioanalyzer High Sensitivity DNA Assay for

our non-standard DNA. All bricks where provided in 200 nM concentrations. Electropherograms always detected a single clear peak per brick as showed in Fig 7.6 and 7.7. Table 7.1 summarizes for each brick its known size, the measured migration time and fluorescence area under the peak, as well as the calculated size and molarity derived by the instrument software from comparison to the reference ladder. Averages and standard deviations were calculated from at least three independent measurements.

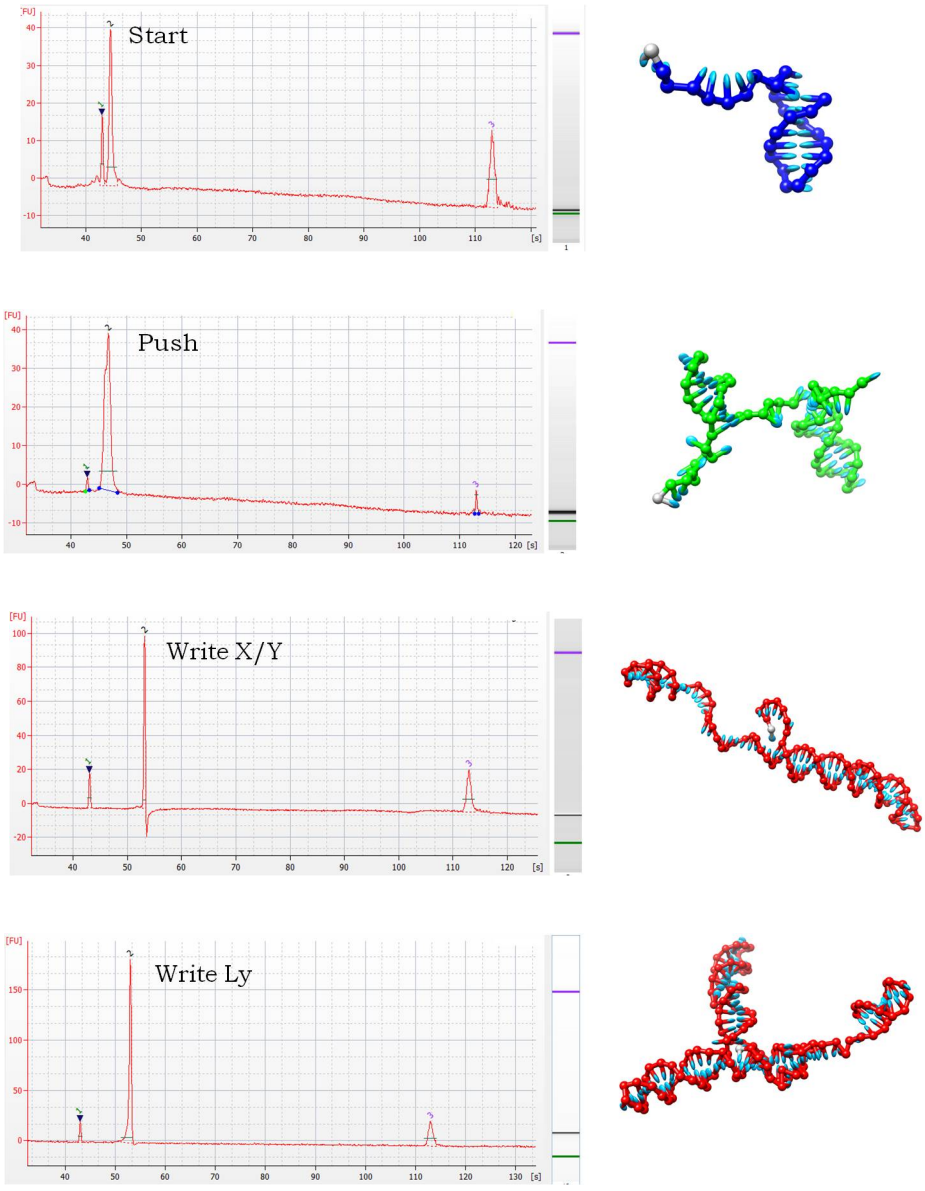



Figure 7.6: **Data Storage cycle Single Brick Calibration Profile**. Right:Electropherograms for each of the bricks involved in the data storage cycle. Each one shows a unique profile. Left: brick model by oxDNA

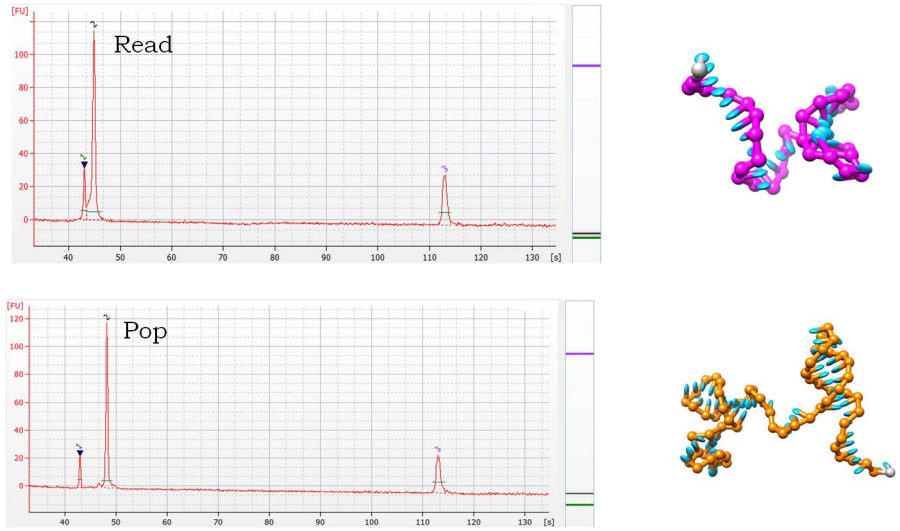



Figure 7.7: **Reading cycle: Single Brick Calibration Profile**. Right: Electropherograms for each of the bricks involved in the reading cycle. Each one shows a unique profile. Left: brick model by oxDNA

| | | measured | | derived | |
|---|---|---|---|---|---|
| **brick** | **size[nt]** | **time [s]** | **area [FU]** | **size [bp]** | **molarity [nM]]** |
| *start* (S) | 27 | 45.22±0.92 | 94.6±61.23 | 51±7.6 | 34.80±15.92 |
| *push* (P) | 64 | 46.81±0.76 | 74.4±39.2 | 64±6.9 | 8.08±0.174 |
| *write-X* (X) | 98 | 53.27±0.34 | 55.93±39.65 | 128±3.78 | 5.961±0.473 |
| *write-Y* (Y) | 98 | 53.45±0.32 | 55.8±36.65 | 129±3.04 | 6,296±0.629 |
| *write-Ly* (Ly) | 128 | 55.35±0.06 | 5.27±1.15 | 147±0.8 | 0.845±0.221 |
| *report-X* (Rx) | 22 | 44.81±0.81 | 248.5±60.57 | 47±6.4 | 78.25±16.81 |
| *report-Y* (Ry) | 22 | 45.18±1.02 | 241.3±84.49 | 47±11.3 | 86.44±12.77 |
| *read* (R) | 31 | 44.61±0.35 | 73.85±15.76 | 46±2.82 | 31.67±1.21 |
| *pop* (Q) | 64 | 47.89±0.28 | 28.13±25.4 | 74±3.4 | 6.602±6.78 |

Table 7.1: Calibration results (given as averages and standard deviation) for all individual strands provided in 200 nM concentrations.

## 7.3 Reaction Time Estimation

To determine kinetic rates for the different strand interactions, the effect of varying reaction times was investigated . This has been done by running on-chip electrophoresis measurements of *start-push* solutions as well as sequentially prepared *start-push-write* solutions with 5 to 240 minutes reaction time. From this section onwards, the quantitative analysis of each experiment is summarised in tables. Each column header indicates the single brick or

the bricks involved in the reaction. If the column header contains a "+ *brick*" *e.g.,* + SP", this indicates that the mentioned bricks were added to the same test tube of the previous column/reaction. The first column always contains "start 100%", meaning that the start is alone and not reacting with any brick in solution, until the next one is added. The row header, on the other hand identifies all the species recognised in the electropheragram (peaks), for each reaction (column). Each cell represents the intersection point between a column and row header and contains the molarity or the area registered for the specie indicated in the row header, when the total start (100%) reacts with the next brick provided, specified in the column header.

### *7.3.1 start-push*

In Fig.7.8, for *start-push*, were not obtained significant differences, even for a reaction time of only 5 minutes: the electropherograms show a single dominant peak that was associate with *start-push*. According to molarity estimates of the Bioanalyzer software, the reaction runs into 80-90% completion as showed in Table 7.2, indicates the experiment reaction time in minutes and the bricks reacting with the *start*. However, limitations of the instrument with regard to single stranded DNA with a partial secondary structure apply, hence, were took in consideration the percentage of the total Area measured by the software (Table 7.3)

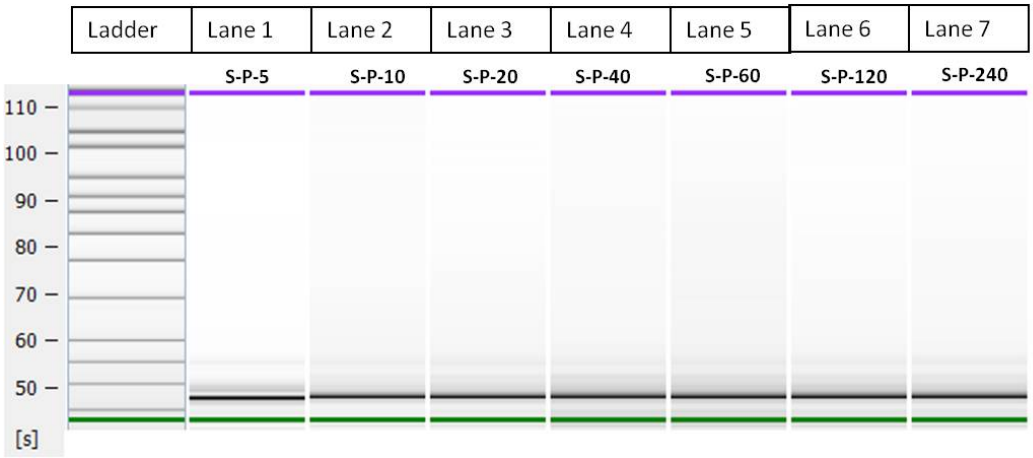



Figure 7.8: Capillary electrophoresis of the *start-push* reaction time. Data obtained from five parallel experiments starting from a reaction time of 240 to 5 minutes from Lane 1 to 5.

-

| | | 5 [min] | 10 [min] | 20 [min] | 40 [min] | 60 [min] | 120 [min] | 240 [min] |
|---|---|---|---|---|---|---|---|---|
| | **S** | **P** | **P** | **P** | **P** | **P** | **P** | **P** |
| **S** | 100% | 6.51% | 7.94 | 9.75% | 10.10% | 4.89% | - % | 8.88% |
| **P** | | 6.93% | 8.03% | 7.87% | 8.15% | 8.10% | 8.48% | 10.94% |
| **SP** | | 84% | 84.03% | 82.38% | 81.75% | 84.69% | 92% | 77.5% |
| **Undefined** | | 2.18% | - | - | - | 2.31% | - | 1.96% |
| | | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.2: **Percentage of the total Molarity**. Molarity, of each species recognized for each individual experiment, derived from the Molarity values measurements of *start-push* by the Agilent Software.

| | | 5 [min] | 10 [min] | 20 [min] | 40 [min] | 60 [min] | 120 [min] | 240 [min] |
|---|---|---|---|---|---|---|---|---|
| | S | P | P | P | P | P | P | P |
| **S** | 100% | 4.2% | 5.1% | 6.4% | 6.7% | 3.1% | - | 5.6% |
| **P** | | 6.1% | 6.9% | 7% | 7.4% | 7% | 7.4% | 9.2% |
| **SP** | | 85.3% | 88% | 86.6% | 85.9% | 85.4% | 92.6% | 80.4% |
| **Undefined** | | 4.4% | - | - | - | 4.5% | - | 4.8% |
| | | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.3: **Percentage of the total Area.** The table shows the percentage of the total area, of each species recognized for each individual experiment, as calculated for *start-push* by the Agilent software.

### *7.3.2 start-push-write X*

For *start-push-write*, in Fig.7.9, it is possible to see that the branch migration reaction of the *write* association does not equilibrate within 5 minutes and it takes at least 10 minutes for the reaction to equilibrate. According to the molarity detected by the Agilent Software, was found that after 5 minutes only between 30-40% of *start-push* complexes were extended by the *write* bricks. (Table 7.4). Molarities value has to be taken with consideration, due to an instrument underestimation of molarities values. The percentage of the total Area (Table7.5), shows instead a more accurate value, since it is calculated using arbitrary fluorescent units (FU) and the area under the curve

Henceforth, for all the experiments presented in this chapter were taken in consideration the percentage of the total Area. A comparison of the Area, for measurements of *start-push* solutions and *start-push-write* solutions with 5 to 240 minutes reaction time is showed in the histograms in Fig.7.10.

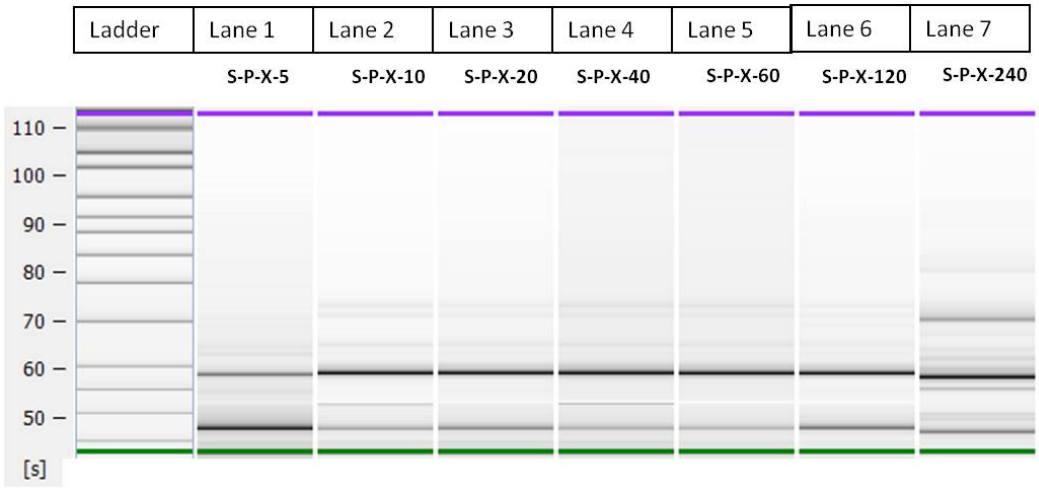



Figure 7.9: Capillary electrophoresis of the *start-push-write* reaction time. Data obtained from five parallel experiments starting from a reaction time of 5 to 240 minutes from Lane 1 to 5.

| | | 5 [min] | 10 [min] | 20 [min] | 40 [min] | 60 [min] | 120 [min] | 240 [min] |
|---|---|---|---|---|---|---|---|---|
| | **S** | **PX** | **PX** | **PX** | **PX** | **PX** | **PX** | **PX** |
| **S** | 100% | 7.15% | 15.44% | 12.66% | 16.60% | 12.43% | 6.90% | 4.45% |
| **P** | | 6.33% | 5.22% | - | 4.95% | - | 6.21% | - |
| **X** | | - | 5.48% | - | 4.35% | - | - | 11.35% |
| **SP** | | 77.18% | 38.02% | 52.40% | 36.97% | 43.42% | 58.82% | 43.17% |
| **SPX** | | 9.34% | 31.48% | 33.96% | 34.33% | 41.92% | 27.47% | 27.35% |
| **SPXP** | | - | 2.61% | 0.97% | 1.30% | 1.40% | 0.60% | 3.23% |
| **SPXPX** | | - | 1.70% | - | 1.49% | 0.84% | - | 5.33% |
| **Undefined** | | - | - | - | - | - | - | 5.12% |
| | | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.4: **Percentage of the total Molarity**. The table shows the percentage of the total Molarity, of each species recognized for each individual experiment, derived from the Molarity values measurements of *start-push-write* by the Agilent Software .

| | | 5 [min] | 10 [min] | 20 [min] | 40 [min] | 60 [min] | 120 [min] | 240 [min] |
|---|---|---|---|---|---|---|---|---|
| | **S** | **PX** | **PX** | **PX** | **PX** | **PX** | **PX** | **PX** |
| **S** | 100% | 4.1% | 6% | 5.2% | 6.4% | 4.5% | 3% | 1.8% |
| **P** | | 4.9% | 2.6% | - | 2.5% | - | 3.7% | - |
| **X** | | - | 5.3% | - | 4.2% | - | - | 7.2% |
| **SP** | | 67.7% | 22.1% | 32.5% | 21.7% | 23.9% | 40% | 20.8% |
| **SPX** | | 23.3% | 52% | 59.8% | 56.7% | 65.5% | 51.7% | 41.1% |
| **SPXP** | | - | 6% | 2.5% | 3.1% | 3.2% | 1.6% | 6.4% |
| **SPXPX** | | - | 6% | - | 5.4% | 2.9% | - | 15.9% |
| **Undefined** | | - | - | - | - | - | - | 6.2% |
| | | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.5: **Percentage of the total Area.** The table shows the percentage of the total area, of each species recognized for each individual experiment, as calculated for *start-push-write* by the Agilent software.

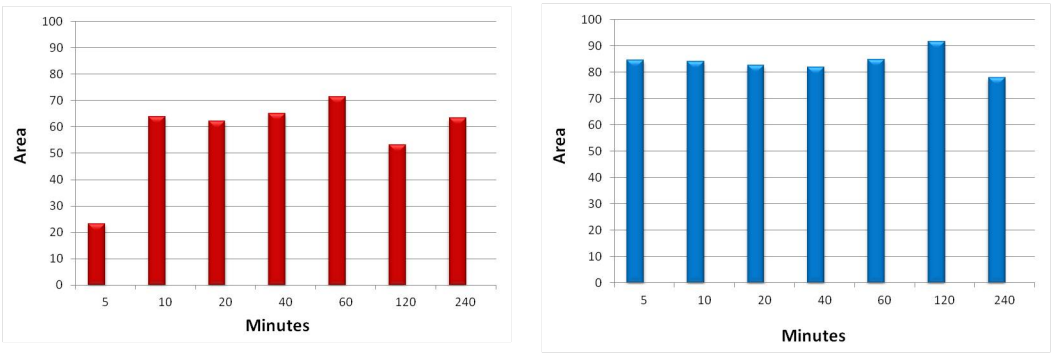



Figure 7.10: **A comparison of the Areas, for measurements with 5 to 240 minutes reaction time**. **Left:**Bar graph of the *start-push-write* data obtained from the measured Area. **Right:** Same bars representation for the extended *start-push* over time.

## 7.4 DNA Stack Data Storage Experiment

### *7.4.1 Data storage differences between write X - write Y and write Ly*

In Fig 7.11 was performed a data storage with write X and write Ly signals and analysed the effects on the chain length and time migration. Experiments were run in six parallel eppendorf tubes. In the first two lanes, addition of *write-X* and *write-Y* brick respectively is followed by the appearance of clear peaks correspondent to 1 signal data storage in the electropherogram (43.7% and 54.3% of the total fluorescence and a migration time of roughly 59 seconds). In the third lane, it was stored one signal using the *write Ly* which accounts for the 37% of total fluorescence and migrates at 63 seconds, slower than an SPX/SPY stack. In Lane 4 (SPXPY) the gel corresponds to the storage of two signals where only about the 8.3% of the total fluorescence, accounts as the second signal stored and a 10% still correspond to the first signal (SPX). In this kind of situation it was not possible to distinguish between X and Y. The substitution of a *write Y* with a *write Ly* as a second signal, in lane 5 leads to the appearance of a 24.4% that was interpreted as a SPLy. It was noticed that the same peak does appear as 14.9% in lane 3. It was identified the peak in lane 3 as a PXP complex which migrates at the same time of SPLy, thus the 24.4% of the total fluorescence appearing in lane might more likely be due for the two species. Only the 6.6% of the total accounts as SPXPLy or an PXPXP. In Lane 6 finally two signals Ly (SPLyPLy) were stored. A well defined peak at about 81s migration time might correspond to the desired product, although corresponding to only the 8.8% of the total fluorescence. Lane peaks indicate clearly

a difference between *write Y* and *write Ly* and the fluorescence of undefined complexes do not exceed the 7% of the total.

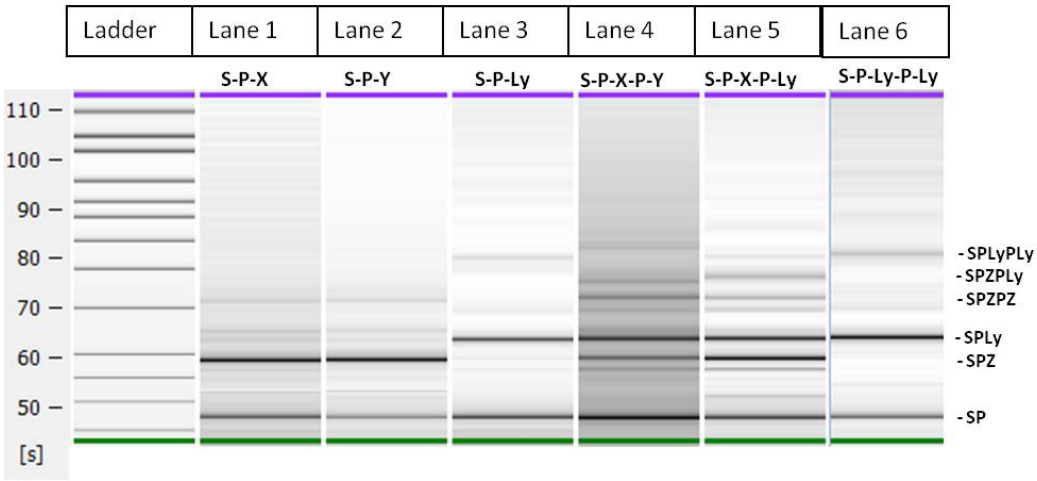



Figure 7.11: **Capillary electrophoresis for one and two signals using *write X*, *write Y* and *write Ly***. Data obtained from six parallel experiments. Lane 1= SPX and Lane 2 = SPY; In Lane 3, SPLy is visible as a higher strong band. Lane 4 shows a two signal stored using *write X*and *write Y*. In Lane 5, the two signals are represented by X and Ly, whereas in Lane 6 the signals are given by Ly and Ly. Letter "Z" indicates either write X or Y.

| | | Lane1 | Lane 2 | Lane3 | Lane 4 | Lane 5 | Lane 6 |
|---|---|---|---|---|---|---|---|
| | S | PX | PY | PLy | PXPY | PXPLy | PLyPLy |
| **S** | 100% | 4.6% | 4.4% | - | 2.6% | - | - |
| **P** | | 3.8% | 2.4% | - | - | - | - |
| **Z** | | 2.2% | - | - | 5.7% | 2.2% | - |
| **Ly** | | - | - | - | 2.3% | | - |
| **SP** | | 30.4% | 25.7% | 59.5% | 22.4% | 27.9% | 36.8% |
| **PZ** | | - | - | - | 4.9% | 4.5% | - |
| **SPZ** | | 43.7% | 54.3% | - | 10.8% | 27.3% | - |
| **PZP/SPLy** | | 4.8% | 3.5% | 37.1% | 14.9% | 24.4% | 54.5 |
| **PZPZ/PLyP** | | - | - | 0.2% | 4.8% | 2.3% | - |
| **SPZP** | | 5.7% | 4.6% | - | 6.7% | - | - |
| **SPZPZ/SPZPY** | | 4.8% | 5.1% | - | 8.3% | 4.8% | - |
| **SPLyPLy** | | - | - | 3.2% | - | - | 8.8% |
| **PZPZP/SPZPLY** | | - | - | - | 6.9% | 6.6% | - |
| **SPZPZPZ** | | - | - | - | 2.7% | - | - |
| **Undefined** | | - | - | - | 7% | - | - |
| | | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.6: **Percentage of the total Area.** The table shows the percentage of the total area, of each species recognized in the capillary electrophoresis in Fig.7.11 for each individual experiment, as calculated by the Agilent software. Letter "Z" indicates either write X or Y.

### 7.4.2 *DNA stack data storage of five signals "X-X-X-LY-X"*

To probe the performance of the data storage (push) cycle, it was performed experiments in which were sequentially stored five signals $(X,X,X,Y,X)$ onto the growing stack. Experiments were run in five parallel eppendorf tubes and stopped at different steps in the protocol. Gel-like images of the Bioanalyzer output is shown in Figure 7.12. Table 7.7 quantifies the relative size of each peak by calculating the relative area under the fluorescence curve.

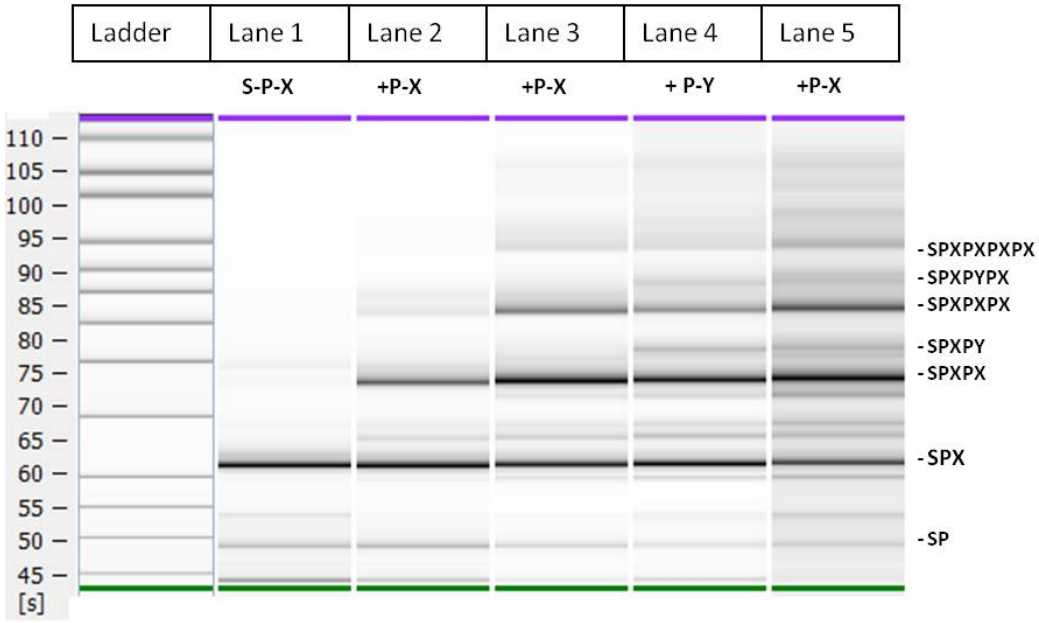



Figure 7.12: Capillary electrophoresis of the data storage process. Lane 1=SPX; Lane2=SPXPX; Lane 3=SPXPXPX; Lane 4=SPXPXPXPY; Lane 5=SPXPXPXPYPX. Data obtained from five parallel experiments.

For the first three stored signals, addition of each *write-X* brick is accompanied by the appearance of a new clear peak in the spectrum: after addition of the first *write-X* brick this peak (*start-push-write-X* complex, or SPX) accounts for more than 58% of the total fluorescence. Lane 2 shows the appearance of a second peak (SPXPX) that corresponds to the two signals. However, this second peak accounts for only about 22% of the total fluorescence, whereas almost 40% still correspond to the first signal (SPX). The situation repeats in the third lane, where the correct complex (SPXPXPX) accounts for slightly more than 17% of the fluorescence, the second signal peak (SPXPX) for about 30% and the first peak still for about 23%.

The addition of *write-Y* in lane 4 leads to the appearance of several new peaks, which was

| | | Lane1 | Lane 2 | Lane3 | Lane 4 | Lane 5 |
|---|---|---|---|---|---|---|
| | **S** | **+PX** | **+PX** | **+PX** | **+PY** | **+PX** |
| **S** | 100% | 17.5% | 7.1% | 1.8% | 3.9% | 2.2% |
| **P** | – | 1.5% | 1.4% | – | – | 0.5% |
| **X** | – | 3.6% | – | 1.9% | – | 0.6% |
| **Y** | – | – | – | – | 1.3% | 3.6% |
| **PX** | – | – | – | – | 2.3% | 3.4% |
| **SP** | – | 13.9% | 10.8% | 5% | 3.2% | 4.8% |
| **SPX** | – | 58.2% | 39.5% | 23.3% | 24.3% | 13.1% |
| **SPXP** | – | 1.5% | 6.8% | 7% | 8% | 9.2% |
| **SPXPX** | – | 1.2% | 22.3% | 29.8% | 27.3% | 22.4% |
| **SPXPY** | – | – | – | – | 7.8% | 6.2% |
| **SPXPXP** | – | 2.6% | 4.5% | 4.5% | – | – |
| **SPXPXPX** | – | – | 5.6% | 17.6% | 11.6% | 11.5% |
| **SPXPYPX** | – | – | – | – | 4.6% | 5.5% |
| **SPXPXPXPX** | - | - | - | 3.7% | 2.6% | 4.8% |
| **Undefined** | – | – | 2% | 5.4% | 3.1% | 12.2 |
| **Total** | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.7: Percentage of the total Area. The table shows the percentage of the total area, of each species recognized for each individual experiment, as calculated by the Agilent software.

identified as SPY, SPXPY, and SPXPXPY. A very faint peak at about 98 s migration time might correspond to the desired SPXPXPXPY, but the signal is too weak to be properly identified by the analysis software. Lane 5 essentially shows the same peaks as lane 4, with peak sizes changing as expected: peaks from complexes ending in a *write-Y* brick become smaller, whereas the corresponding complexes with added *write-X* become proportionally larger.

In all lanes faint higher peaks indicate that there is a very small potential for run-away processes to create complexes with more signals than the provided ones. Yet, in all cases, the fluorescence of all these longer bands combined does not exceed 10% of the total.

### 7.4.3 *Reading out of five signals "X, Ly, Ly, Ly, X"*

Next, were performed experiments to test the reading/popping cycle of the DNA stack. In this experiment, five signals ($X, Ly, Ly, Ly, X$) where pushed onto the stack and subsequently removed by adding *read* ($R$) and *pop* ($Q$) bricks in molarities equal to the *start*, *push* and *write* bricks (200 nM) and waiting time 240 minutes between each $R$ and $Q$. Figure 7.13 shows the gel-like images and table 7.8 the correspondent percentage of the fluorescence for each experiment.

Lanes 1 through 5 reconfirm the working of the data storage cycle with the same observations than for the experiment of the last section: each added *write* brick generates a new peak in the spectrum with very little evidence for run-away processes and persistence of peaks that indicate intermediate complexes.

Lane 6 to 9 shows the response of the device after provision of 200 nM *read* (R) and *pop* (Q), to trigger four readout cycle, starting from the last signal X stored in lane 5 ($X, Ly, Ly, Ly, X$). Lane 10 and 11 show the response after the provision of 200 nM *read* (R) and *pop* (Q), to trigger one readout cycle starting from a sample taken at a stage where only 4 signals were stored, corresponding to lane 4 ($X, Ly, Ly, Ly$). Lane 12 and 13 shows the response after the provision of 200 nM *read* (R) and *pop* (Q), to trigger one readout cycle starting from a sample taken at a stage where only 3 signals were stored, corresponding to lane 3 ($X, Ly, Ly$). From lane 6 to 13, newly created *push-pop* as well as *read-write* complexes result in the appearance of three new peaks at around 47.42(QP), 52.22(RX), and 57.39 (RY) seconds. The *push-pop* (QP) complexes account for more values from the 33.6% to the 68% of the fluorescence, whereas *start-write-X* (RX) and *start-write-Ly*(RLy) account for 2.8 to 16.8% and 3.7 to 20.5% respectively. Peaks associated with the different stack states SPXPLyPLy SPXPLyPXPLyPX, ... SPLyPLy, SPXPLy, and SPLy decrease accordingly. For the second readout cycles there is an increase of *read-write*(RX or RLy), and the *push-pop* (QP) peaks simultaneously reduces intensities of the corresponding stack complexes. Noteworthily, after reading out the five stored signals, 11.5% of the fluorescence results from the *start-push* complex whereas peaks of stacks that still contain stored information only register with 17.2, 9.9, 27, 13.1, 9.3 and 5.4%.

| | Data Storage | | | | | Reading Lane 3 | | | | Reading Lane 4 | | Reading Lane 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Lane 1 | Lane 2 | Lane 3 | Lane 4 | Lane 5 | Lane 6 | Lane 7 | Lane 8 | Lane 9 | Lane 10 | Lane 11 | Lane 12 | Lane 13 |
| | S | +X | +Ly | +Ly | +X | +RQ [5(e) to 4(e)] | +RQ [4(e) to 3(e)] | +RQ [3(e) to 2(e)] | +RQ [2(e) to 1(e)] | +RQ [4(d) to 3(d)] | +RQ [2(d) to 1(d)] | +RQ [3(c) to 2(c)] | +RQ [2(c) to 1(c)] |
| **S/R** | 100% | 21.5% | 9.8% | 19.2% | - | 1.5% | 1.5% - | - | - | - | 7.7% | 8.8% | 4.4% |
| **P** | - | - | - | 1.9 | - | - | - | - | - | - | - | - | - |
| **X/RX** | | 5% | 4.1% | 1.5% | - | 2.2% | 17.2% | 16.8% | 13.8% | 5.7% | 6.5% | 2.8% | 1.9% |
| **Ly/RLy** | | - | 1.8% | 4.8% | 2.3% | 6.6% | 12.5% | 4.5% | 3.9% | 3.7% | 12.9% | 12.7% | 12% | 20.5% |
| **QP** | | - | - | - | - | - | 68% | 54.1% | 59.9% | 59.1% | 49% | 33.6% | 38% | 38.5% |
| **SP** | | 20.1% | 14.1% | 10.4% | 13.7% | 11.5% | - | 12.8% | 11.9% | 15.2% | 15.9% | 14.7% | 9.5% | 14.1% |
| **PX** | | - | - | 5.6% | 1.7% | - | - | - | - | - | - | - | - | - |
| **SPX** | | 53.4% | 28% | 10.5% | 13.4% | 17.2% | 6.3% | 7.3% | 5.9% | 6.7% | 8.5% | 8.6% | 10.6% | 8.3% |
| **PXP/SPLy** | | - | 18.2% | 11.9% | 20.9% | 9.9% | - | 2.6% | 1.6% | - | 4.6% | 5.5% | 8.2% | 4.2% |
| **SPXP** | | - | 1.8% | - | - | - | 1.2% | - | | 1.5% | - | 0.9%% | - | 4.8% |
| **SPXPX** | | - | 4.4% | 6.7% | 4.1% | 27% | 1% | - | - | - | - | 1.9% | 3.5% | 3.3% |
| **SPXPLy** | | - | 15.3% | 7.8% | 25% | 13.1% | 7.3% | - | - | - | - | - | - | - |
| **SPLyPLy** | | - | - | 12.8% | - | - | - | - | - | - | - | - | - | - |
| **SPXPXPX** | | - | - | - | 14.2% | 9.3% | - | - | - | - | - | - | - | - |
| **SPXPLyPLy** | | - | - | 6.9% | - | 5.4% | - | - | - | - | - | - | 6.6% | - |
| **SPXPXPXPX** | | - | - | - | 2.2% | - | - | - | - | - | - | - | - | - |
| **Undefined** | | - | 2.5% | - | 2.5% | - | - | - | - | 3.4% | 7.9% | | | |
| | 100% | | 100% | | 100% | 100% | 100% | 100% | 100% | 100% | 100% | | | |

Table 7.8: Percentage of the total area. The table shows the percentage of the total area, of each species recognised for each individual experiment, as calculated by the Agilent software.

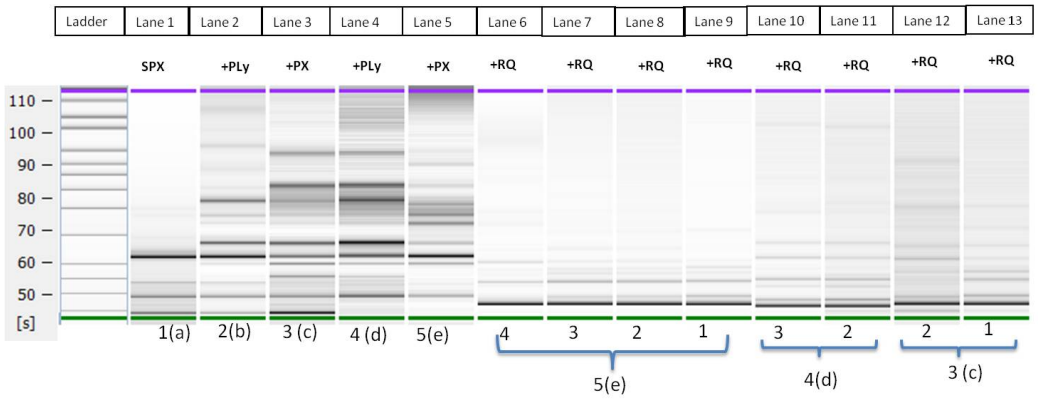



Figure 7.13: Capillary electrophoresis of the data storage and reading of three signals. Data storage: Lane 1=SPX [1(a)]; Lane2=SPXPLy [2(b)]; Lane 3=SPXPLyPX [3(c)] ; Lane 4=SPXPLyPXPLy [4(d)]; ; Lane 5=SPXPLyPXPLyPX [5(e)]. Reading: Lanes 6 to 9, popping of 4 signals from sample 5(e) = SPXPLyPXPLyPX - RQRQRQRQ; Lanes 10 to 11, popping of 4 signals from sample 4(d) =SPXPLyPXPLy+RQRQ; Lanes 12 to 13, popping of 2 signal from sample 3(c) =SPXPLyPX + RQRQ.

### *7.4.4 Optimization of the reading out*

Lanes 1 through 5 reconfirm the working of the data storage cycle with the same observations than for the experiment of the data storage section. Here it was showed the response of the device after provision of *push* at each cycle. The popping was optimized using half of the concentration (100 nM for each R and Q)

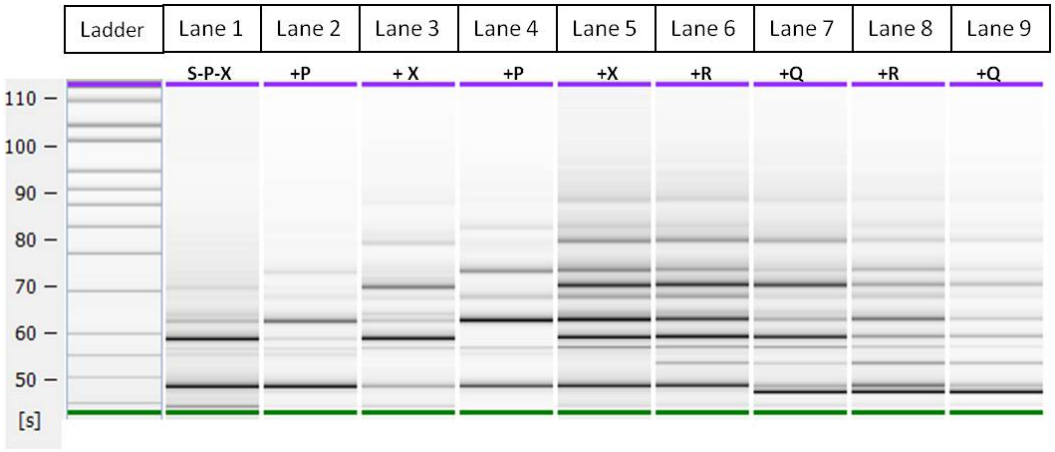



Figure 7.14: Capillary electrophoresis of the data storage and reading of three signals. data storage: Lane 1=SPX; Lane2=SPXPY; Lane 3=SPXPYPY. Reading: Lane 4=SPX-PYPY+RQ; Lane 5=SPXPYPY+RQRQ.

| | | **Lane1** | **Lane 2** | **Lane3** | **Lane 4** | **Lane 5** | **Lane 6** | **Lane 7** | **Lane 8** | **Lane 9** |
|---|---|---|---|---|---|---|---|---|---|---|
| | **S** | **+PX** | **+P** | **+X** | **+P** | **+X** | **+R** | **+Q** | **+R** | **+Q** |
| **S** | 100% | 11.2% | 2.2% | 4.9% | - | 2% | 1.7% | 1.9% | 1.5% | 2.6% |
| **P** | | - | - | - | - | 0% | - | - | - | - |
| **X** | | 4.3% | 6.5% | 3.8% | 6% | 4.2% | 4.6% | 4.2% | 4% | 1.7% |
| **SP** | | 40.8% | 52.5% | 13.6% | 33.9% | 19.9% | 18.2% | 7.9% | 16% | 12.4% |
| **SPX** | | 31.1% | 5% | 33.7% | 0% | 16.1% | 16.2% | 19.1% | 8.7% | 9.1% |
| **SPXP** | | 10.6% | 25.8% | 11.8% | 36% | 19.3% | 15.8% | 9% | 14.9% | 8% |
| **SPXPX** | | 2% | 4% | 22.6% | 7.5% | 19.8% | 20.7% | 20.8% | 12.8% | 11.7% |
| **SPXPXP** | | | 4% | 3.9% | 14.1% | 9% | 8% | 4.8% | 7.5% | 4% |
| **SPXPXPX** | | | | 5.7% | 0% | 5.7% | 6.1% | 6.8% | 3.5% | 3.8% |
| **SPXPXPXP** | | | | | 2.5 | 2.3% | 2.2% | 1.1% | 2.1% | 1.4% |
| **SPXPXPXPX** | | | | | | 1.7% | 2.6% | 1.9% | 1.5% | 1.2% |
| **RX** | | | | | | | 3.9% | 2% | 7.4% | 7.5% |
| **QP** | | | | | | | | 20.5% | 20.1% | 36.6% |
| | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Table 7.9: The table shows the percentage of the total area, of each species recognised for each individual experiment, as calculated by the Agilent software, when half of the Q and R are provided (100 nM)

## 7.5 Molecular Beacons

In this study I propose a preliminary example of the usage of the indodicarbocyanine, Cy5 and Black Hole Quencher-2, BHQ-2 to monitor the *start-push* hybridization reaction over time by directly labelling the bricks with the molecular beacons. It was achieved with a 16 nt long sequence ( henceforth *anti-start* or AS),labelled at the 3' with BHQ-2, designed in order to be partially complementary to the *start* brick. The experiment were performed using a The 96-Well Microplate, as described in chapter 5, section 5.1.5.

#### 7.5.0.1 Hybridization

In Fig.7.15 results show the analysis conducted using a 5' [BHQ-2]-*push*-3' and 5'*start*-[Cy3]-3'. Measurements show a very low FU signal for all the reaction set. The experiment confirms the *start* and *push* hybridization although the estimation of the reaction time still remains very unclear (Fig.7.16).

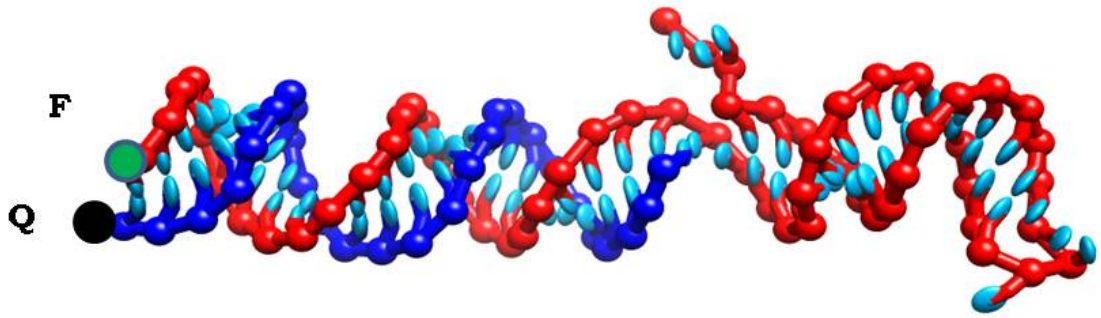



Figure 7.15: **Start-Push Molecular beacons oxDNA simulation**. In blue is shown the *start*, with the Quencher (Q) at the 3'end, and in red the *push*, with the Fluorophore (F) at the 5' end.

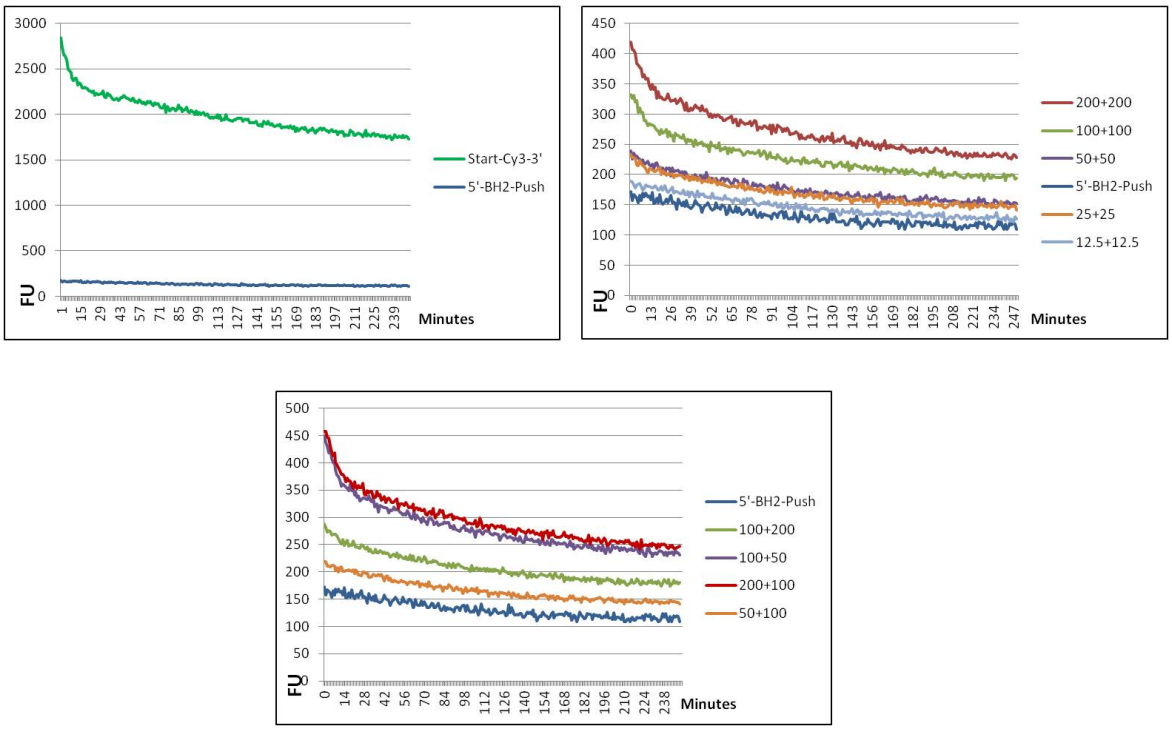



Figure 7.16: **Molecular Beacons measurements for 5'-BHQ-2-push-3' and 5'-Cy3-start-3'**.The assay was carried out the including control measurements for 5'-BHQ-2-push-3' (blue line) and 5'-Cy3-start - 3'(green line), as shown in the first plot. Second plot on the right displays results until 500 FU: 200 nM S + 200 nM P (dark red); 100 nM S + 100 nM P (light green); 50 nM S + 50 nM P (purple); 25 nM S + 25 nM P (orange); 12,5 nM S + 12,5 nM P(light blue). In the last graph, results are displayed up to 500 FU and loaded as follow: 100 nM S + 200 nM P (light green); 100 nM S + 50 nM P (purple); 200 nM S + 100 nM P (red); 50 nM S + 100 nM P (orange). Samples were run straight after their preparation.

#### 7.5.0.2 Strand Displacement

The experiment in Fig 7.17. Samples were samples with a variety of molar concentrations to improve our knowledges of concentration *vs* reaction time. Plate reader FU measurements indicates that no fluorescence is detected when the *anti-start* is used alone. The 5'*start*-[Cy3]-3', on the other side is detected with high FU signal, as expected. The assay was carried out mixing in a eppendorf tube *anti-start* and *start*, and the *push* brick afterwards. Results are shown in Fig.7.18. Measurements clearly show a highest FU signal for 200 nM concentration respect to the usage of 50 nM. The expectation is that the AS hybridize with the start *c* domain and no fluorescence is detected, due to the action of the quenching molecule BHQ-2. Addition of the *push* should in principle separate and displace the AS, so as soon as the Cy3 is not quenched by fluorescence is detected.

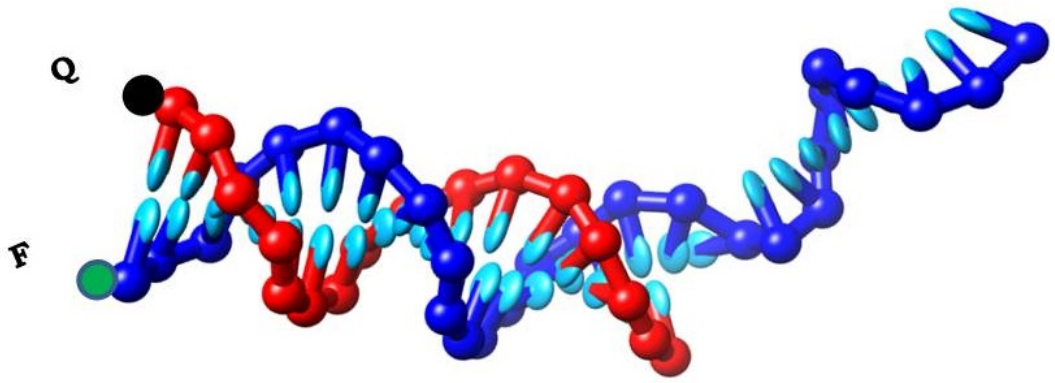



Figure 7.17: **Antistart-Start Molecular beacons oxDNA simulation**. In blue is shown the *start*, with the Fluorophore (F) at the 3'end, and in red the *anti-start*, with the Quencher (Q) at the 5' end.

The experiments are just a preliminary example of the usage of molecular beacons. It is important to better understand the time of degradation of the Cy3, since studies suggests that the degradation occurs nearly three times more rapidly when oligomers are externally labelled. Thus, in further works the design could be improved with an internal labelling of

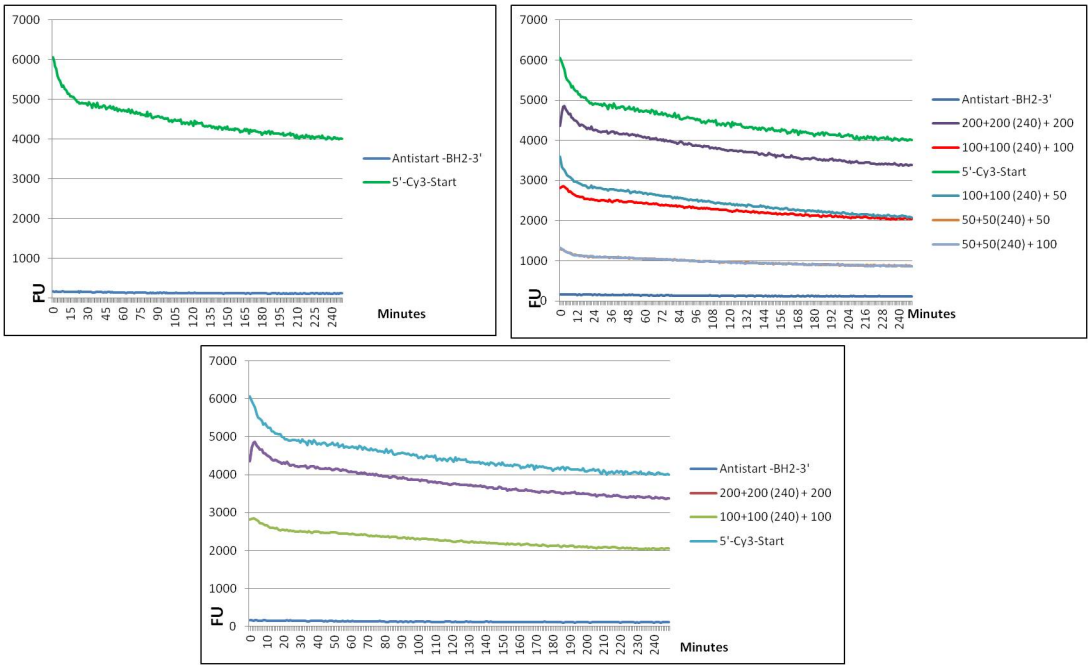



Figure 7.18: **Molecular Beacons measurements for 5'-Anti-Start BHQ-2-3' and 5'-Cy3-start-3'**. The graphs reports on the y axis the Fluorescent unit and on the x axis the time in minutes. The assay was carried out the including control measurements for 5'-Anti-Start BHQ-2-3' (blue line) and 5'-Cy3-start-3'(green line), as shown in the first plot. In the second plot on the right, measurements were carried out as follow: 200 nM AS + 200 nM S (240 minutes reaction) and 200 nM unlabelled push (purple);100 nM AS+ 100 nM S (240 minutes reaction) and 100 nM unlabelled push (red); 100 nM AS+ 100 nM S (240 minutes reaction) and 50 nM unlabelled push (light blue); 50 nM AS + 50 nM S (240 minutes reaction) and 50 nM unlabelled push (orange); 50 nM AS + 50 nM S (240 minutes reaction) and 100 nM unlabelled push (light purple). In the last graph experiments are represented by 200 nM AS + 200 nM S (120 minutes reaction) and 200 nM unlabelled push (purple); 200 nM AS+ 200 nM S (60 minutes reaction) and 200 nM unlabelled push (light purple); 200 nM AS+ 200 nM S (60 minutes reaction) and 200 nM unlabelled push (orange).

the samples, to better track the branch migration walk.

## 7.6 Imaging

For additional confirmation of the data storage within the stack, the assembled nanodevice was imaged using transmission electron microscopy (TEM). For this purpose, assembled chains were mixed with *report* strands that, in turn, are decorated with 10 and 20 nm gold nanoparticles. *Reporters* associate with their respective *write* bricks at any position in the assembled stack. Nanoparticles appear in TEM images as black dots that can be easily distinguished and classified.

Samples were examined using a Philips CM 100 Compustage (FEI) TEM and digital images were collected using an AMT CCD camera (Deben) provided by the Electron Microscopy Research Services at Newcastle University. A volume of 5$\mu$L sample was applied on glow discharge grids preliminary washed with 0.5 mM magnesium chloride to change the hydrophilic surface charge orientation.

Samples were prepared as described in the dedicated section in chapter 5, with and without the usage of a staining agent. Results are not significantly different under the presence of Uranyl Acetate as showed in Fig. 7.19.

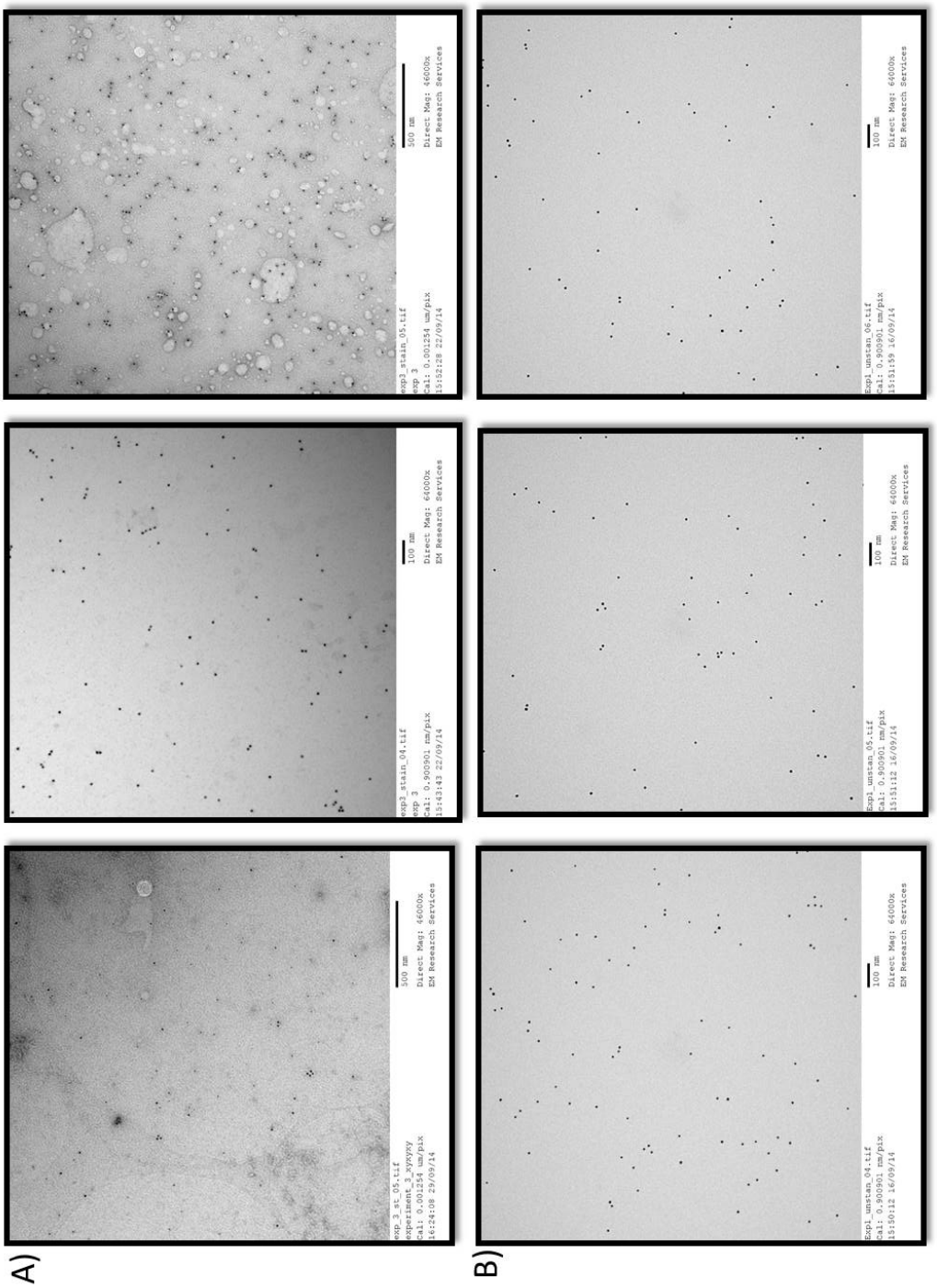



Figure 7.19: **TEM sample with and without Uranyl Acetate**. In Panel A) are shown TEM images produced using staning agent, whereas in Panel B) gold particles are perfectly visible without the need of staining. Scale 100 nm and 500 nm

.

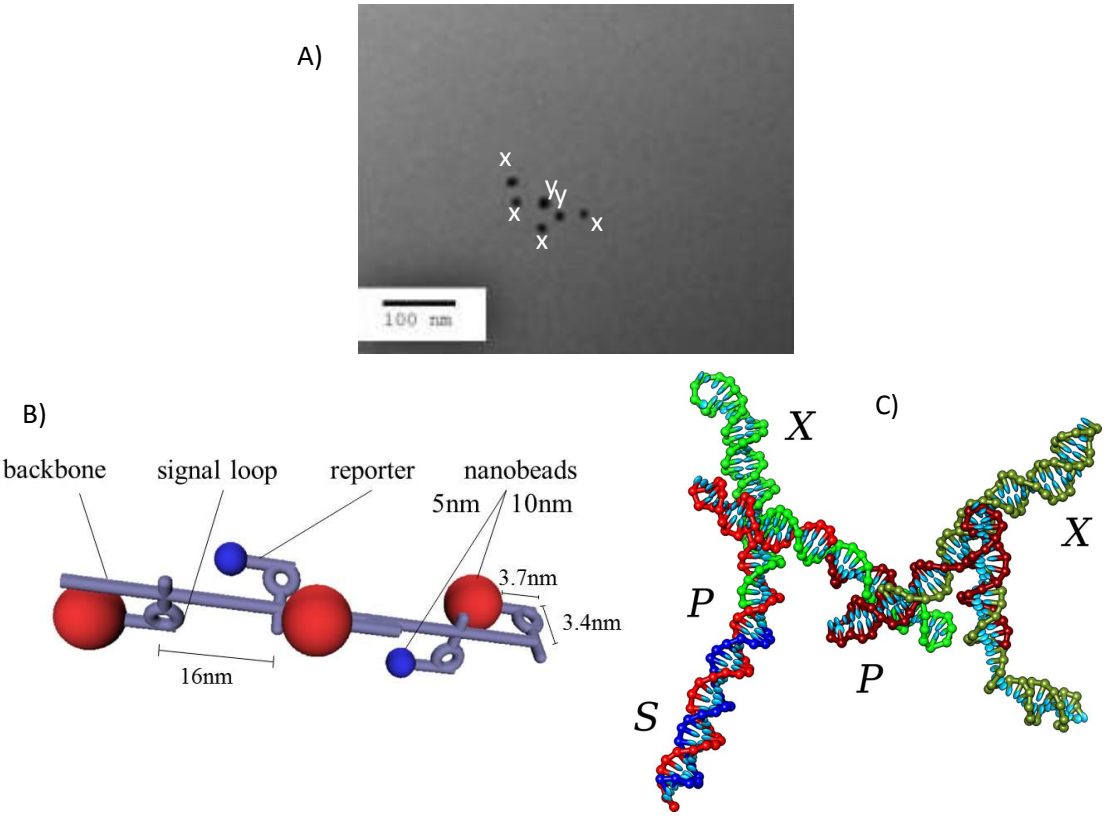



Figure 7.20: **Top:** TEM image, scale 100 nm. High Magnification. **left:** 3D B-Conformation assembled structure. Representative image of x-x-y-x-x Where *report-x* has been attached to a 10 nm gold nanoparticle and *report-y* to a 5 nm gold nanopartcle (A-B). **Right:** oxDNA simulations of a single *start-push-write* complex (C).

In Fig.7.20, the original experiment consists of the data storage of five signals "X-Y-X-X-X". OxDNA simulations [96] (Fig. 7.20, right) indicate that the assembled stack does not necessarily extend straight forward but might instead contain a kink at each *signal-push* holiday junction. Fig. 7.20 (top) shows TEM results from an experiment where five signals $(X,Y,X,X,X)$ have been recorded. The image show a stack with just one extra *write-X* on the left side of the recorder, resulting in a stack with six signals $(X,X,Y,X,X,X)$. The image shows a separation of 15-20 nm between the nanoparticles with a zig-zag configuration predicted by the simulations.

The image shows a separation of 15-20 nm between the nanoparticles as estimated in the 3D conformation (Fig 7.20 left). Moreover the zig-zag alignment confirms the geometric prediction of our oxDNA simulations (Fig 7.20, right).

In Fig.7.21 a representative TEM images of Au-conjugated bio-bricks assembled with alter-

nate X and Y signals respectively attached to a 10 nm and 5 nm gold nanobeads.

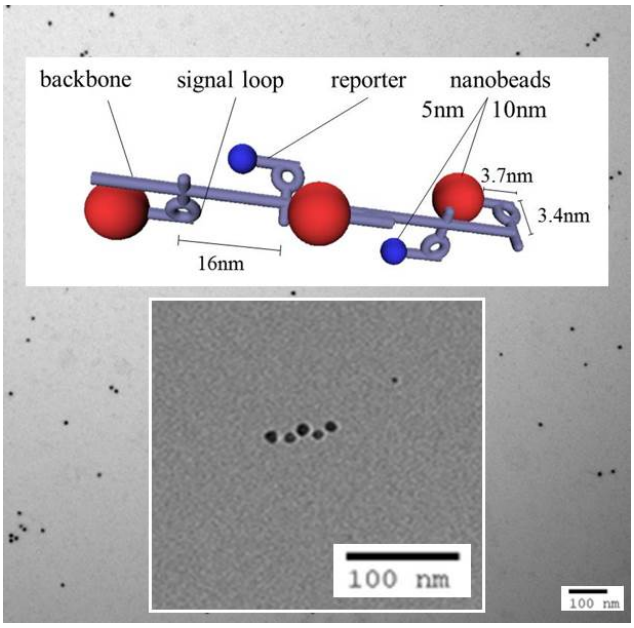



Figure 7.21: **High-magnification Au-conjugated TEM image**. The image shows Au-conjugated bio-bricks assembled with alternate x and y signals (x-y-x-y-x), five signals stored). Simultaneous use of 5 nm and 10 nm Au gold nanoparticles has been used respectively to image the x and the y signal. Reporter bricks has been used to bind the signal. The image also shows the predicted 3D structure pattern that has been confirmed with the imaging. Scale 100 nm.

In Fig.7.22, the original experiment consists of the data storage of ten signals "20 Au- X-Y-Y-Y-X-Y-X-X-Y-X" with the start attached to a 20 nm gold nanoparticle, X and Y respectively to a 5 and a 10 nm gold nanoparticle. It was possible to recognize different stack, and especially in Fig.7.23 (bottom) a stack starting exactly with a 20 Au attached to a 20-Au-X-Y-Y-Y.

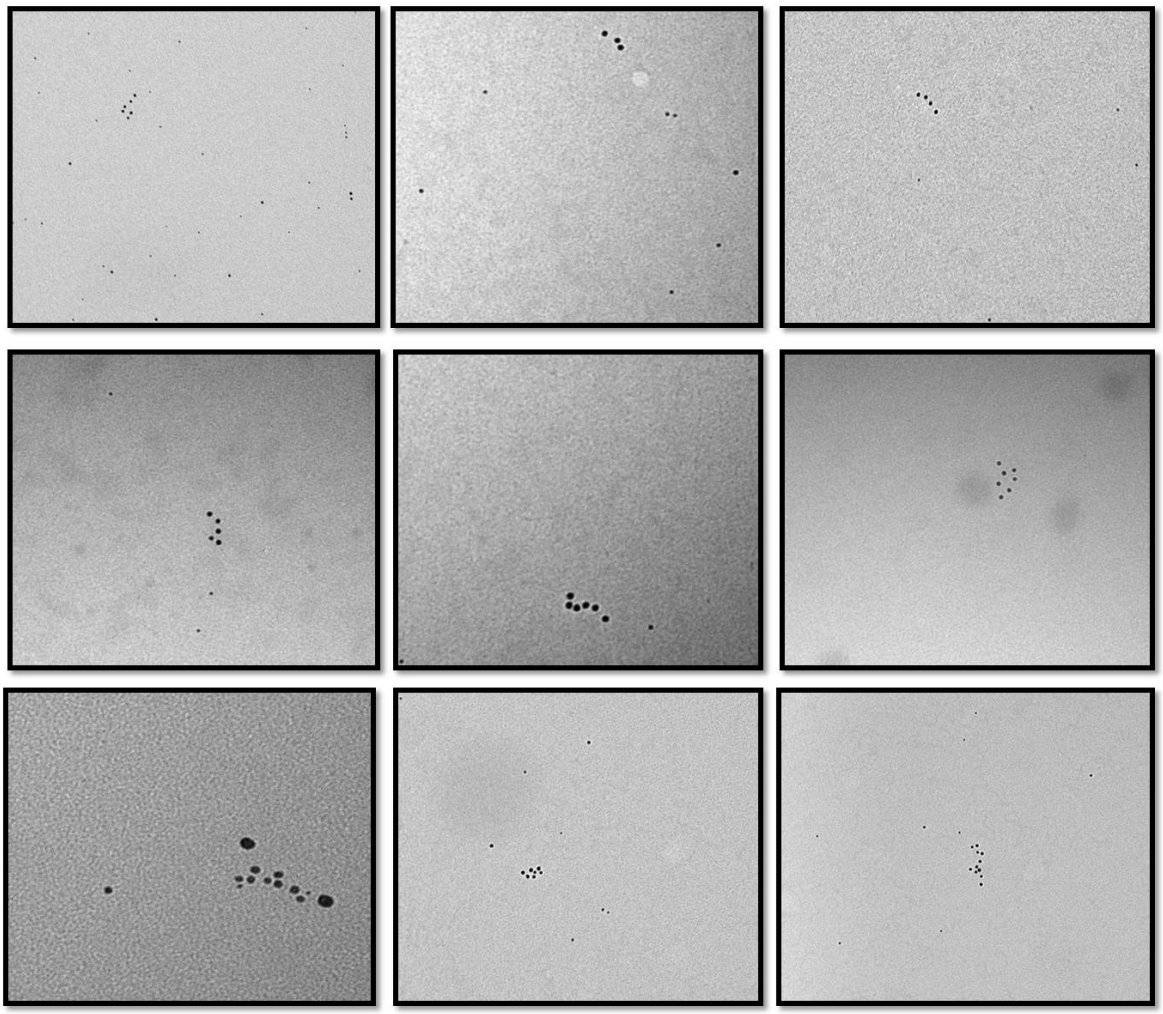


Figure 7.22: **Imaging DNA Data Stack Storage**. Representative TEM images of Au-conjugated. The original experiments consists of data storage of 10 signals 20-Au-XYYYXYXXYX. Was possible to find representative images of data storage, but not in the X- Y order that the experiment was set. The images give an idea of how the DNA stack data structure is visualized at the TEM

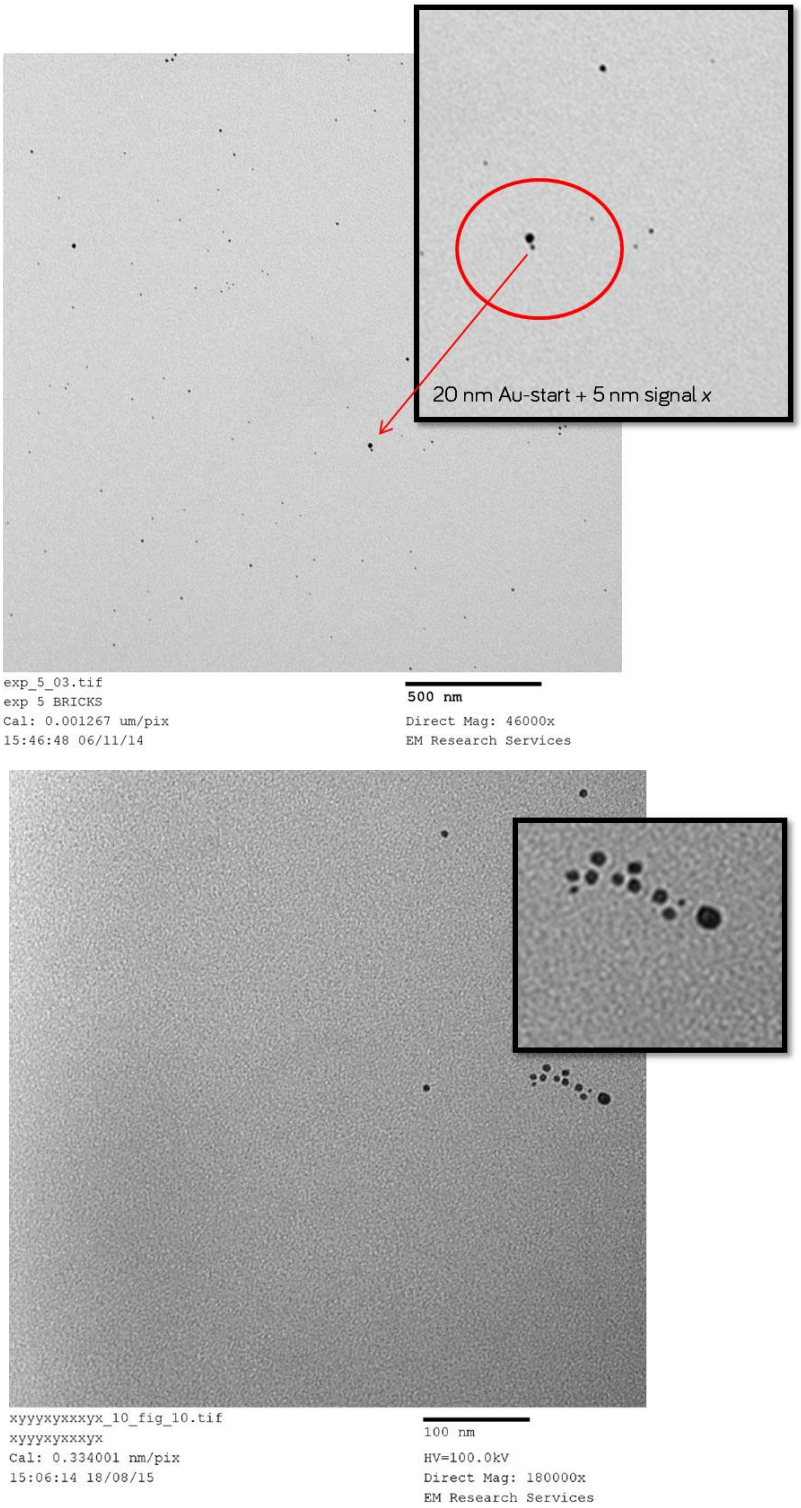



Figure 7.23: **TEM image of a 20 nm Au particle binding the start bio-brick**. The start brick has been attached to a 20 nm gold nanopartcle to facilitate the visualization of the beginning of the chain. Scale 500 nm.**TEM image of 10 signals stored**. 10 signals (x-y-y-y-x-y-x-x-x-y-x). The beginning of the chain is marked by a 20 nm Au gold nanoparticle. The assembling does not show exactly the one proposed in the experiment, but it does suggest a very close structure. Scale 100 nm.

## 7.7 Preventing Runaway Process

In order to prevent an uncontrolled elongation of the stack through previously unattached bricks present in the solution (henceforth, runaway process), I explored the possibility of introducing a washing step in the protocol. The washing was performed with (i) Hydrophilic Streptavidin 2 $\mu m$ diameter Magnetic beads provided by New England Biolab and (ii) a size-selection filtering via Amicon-Ultra 0.5 Centrifugal Filter Unit devices with 100 KDa cut off. Both were used as instructed by the provided supplier protocols. The magnetic nanobeads are covalently conjugate to streptavidin, in order to strongly bind the biotinylated ligand. The *start* brick was incubated with magnetic beads as instructed in the provided supplier protocol. A magnetic rack was used for the separations of the unattached biotynilated bricks. To release the stack, I explored the usage of two different versions of the *start* brick, modified by introduction of a *SmaI* and a *SwaI* restriction enzymes site, sequences are listed in table 5.1. These enzymes have been used for their ability to work at 25°C (approximately room temperature). The digestion reaction was carried out in the test tube containing the magnetic beads and the samples. The provided enzymes buffer was used as for the reaction, and left the mixture reacting for 2h. Amicon Centrifugal Filters for DNA were also used to attempt the filtration of the unattached bricks. The average weight of one base pair is 650 Daltons [126]. Conversion of our bricks and complexes into Daltons have been listed in Tables 7.10. I used either a 10 or a 20 nm gold nanoparticles to grant the filtration.

| **Brick** | **Size [KDa]** |
|---|---|
| Start | 17.5 |
| Push | 41.6 |
| Start-push | 59.1 |
| Write X/Y | 63.7 |
| Write LY | 83.2 |
| Start-Push-Write X/Y | 122.8 |
| Start-Push-Write LY | 142.3 |

Table 7.10: Bricks nucleotide size converted in Dalton units.

#### 7.7.0.1 Limitations of the washing step

Preliminary results showed that the proposed methods did not produce enough yield to be detected by the agarose gel and on chip electrophoresis, suggesting that these washing meth-

ods are not applicable. A potential explanation may be that both enzymes do not cut within the restriction sites, thus the DNA stack is not released. This is possibly due to lack of space between the beads and the restriction site. On the other side, the recovery rate using Amicon filters was also not detectable, possibly due to the fact that the majority of *start* bricks bound to a 10 or 20 nm gold nanoparticle presumably get stuck in the filter. It is possible that gold is progressively blocking the membrane until eventually it is not functioning at its MWCO (100 KDa). The washing step represents a desired but not necessary condition for the prototype implementation, thus although was not possible to prevent the uncontrolled elongation of the stack, it has been still successfully possible to implement and observe the operations a laboratory scale level.

## 7.8 Conclusions

The stack data structure characterization shows the capacity of the molecular device to store a sequence of events encoded via the use of DNA single strands. Although the protocols need to be improved for a better yield, the prototype is a viable mechanism to read and write information from the device and ensures robust operations over multiple rounds.

# Chapter 8

# Characterization and Results of an In Vivo DNA Tape Data Structure

This chapter presents experimental results of the phenotypic characterization of the strains used as platforms for the tape data structure, using standard microbiology and molecular biology techniques. The tape was initially cloned in a low copy plasmid and successfully inserted into the *E.coli* chromosome. Due to the nature of the recombinase system, it is difficult to control, and thus the final tape data structure remains very difficult to achieve in the laboratory setting.

## 8.1 Phenotypic Characterization of Δ*PyrF* Strains

The laboratory workhorse strain *E. coli DH5* was modified to become a strain platform to host the tape system. It was achieved by deleting the *pyrF* gene, to render the strain auxotroph for pyrimidines, required as a selection and counter-selection marked in this system. The next subsections, introduce the phenotipic characterization of Δ*pyrF* strains (TLoST01 and TLoST02, chapter 5 section Table 5.5)

#### 8.1.0.1 Antibiotic sensivity

In previous work while me and my supervisor team were at Nottingham (where this dissertation work started) EPSRC grant EP/HO24905/1 The Logistic of Small Things - A Cross

Disciplinary Feasibility Account, (TLoST), unpublished, *DH5α* was genetically modified to knock out the *lacI* gene encoding the *lac* operon repressor [127] as well as the *pyrF* gene encoding for orotidine -5'-phosphate decarboxylase [128] and the *sdiA* gene encoding a transcriptional regulator of the LuxR family. The strains *TLoST01*, *TLoST02* did not grow in the presence of tetracycline or gentamicine antibiotics, therefore not exhibiting the phenotypes that they should. The strain *TLoST02*, was also unable to grow under the presence of trimethoprim resulting sensitive to the such antibiotic, suggesting that it probably do not includes the tape system.

#### 8.1.0.2 Uracil auxotrophy

The *pyrF* gene encodes for orotidine-5'-phosphate decarboxylase, an essential enzyme in the pyrimidine biosynthetic pathway which deletion confers uracil auxotrophy [123] . *E. coli* strains with this gene knocked-out should not grow in the absence of uracil in minimal medium. In order to verify the deletion of *pyrF* in the strains *TLoST01/pKD46RecAPa*, and *TLoST02*, were grown on solid minimal media (M9) in the presence and absence of uracil. Results after 24 hours and 48 hours incubaton, at 37°C are shown in the 8.1 and 8.2.

| | **U -** | **U** |
|---|---|---|
| DH5α(*control*) | - | - |
| TLoST01/pKD46RecAPa | - | - |
| TLoST02 | - | - |

Table 8.1: Results after 24 hours shown that *TLoST01/pKD46RecAPa*, *TLoST02* did not grow under the presence and absence of uracil. *DH5α* was used as control

| | **U -** | **U** |
|---|---|---|
| DH5α(*control*) | + | + |
| TLoST01/pKD46RecAPa | + | + |
| TLoST02 | + | + |

Table 8.2: Results after 48 hours shown that TLoST01/ pKD46RecAPa, TLoST02 grew under the presence and absence of uracil. DH5α was used as control.

*E. coli* DH5α exhibits inferior growth phenotype, especially in minimal media (M9) compared to other *E. coli* strains. Strains which present a mutated *purB* gene need a supplementation of adenine [122].Results after 24 hours and 48 hours incubaton with and without

adenine, at 37°C are shown in the 8.3 and 8.4. Both strains were grown again on solid agar under the presence and absence of uracil and adenine. Results after 24 hours of incubation, at 37°C are shown in table 8.5.

| | **A -** | **A** |
|---|---|---|
| DH5$\alpha$(*control*) | - | - |
| TLoST01/pKD46RecAPa | - | - |
| TLoST02 | - | - |

Table 8.3: Results after 24 hours shown that *TLoST01/pKD46RecAPa*, *TLoST02* did not grow under the presence and absence of adenine. *DH5*$\alpha$ was used as control

| | **A -** | **A** |
|---|---|---|
| DH5$\alpha$(*control*) | + | + |
| TLoST01/pKD46RecAPa | + | + |
| TLoST02 | + | + |

Table 8.4: Results after 48 hours shown that TLoST01/ pKD46RecAPa, TLoST02 grew under the presence and absence of adenine. DH5$\alpha$ was used as control.

| | **U- + A -** | **U - + A** | **U + A** |
|---|---|---|---|
| DH5$\alpha$(*control*) | + | + | + |
| TLoST01/pKD46RecAPa | + | + | + |
| TLoST02 | + | + | + |

Table 8.5: Results after adenine supplementation shown that the strains *TLoST01/pKD46RecAPa*, *DH5*$\alpha$, *TLoST02* grown in the presence and absence of both uracil and adenine. *DH5*$\alpha$ was used as control.

#### 8.1.0.3 5-Fluoroorotic Acid (5-FOA) selection for Δ*pyrF* strains

5-Fluoroorotic Acid (5-FOA) has been used as selective agent. A functional *pyrF* gene encodes for orotidylate decarboxylase, (involved in uracyl biosynthesis). 5-FOA is converted to 5-fluorouracil (5-FU), by orotidylate decarboxylase. 5-FU is uracyl highly toxic analogue. *E. coli DH5*$\alpha$ does not grow in presence of 5-FOA. Test conducted on *TLoST01/pKD46RecAPa*, and *TLoST02* show that both strains are sensitive to 5-FOA, confirming that the strains do not carry the tape system. Results are shown in table 8.6

| | **5-FOA$^-$** | **5-FOA** |
|---|---|---|
| DH5$\alpha(control)$ | + | - |
| TLoST01/pKD46RecAPa | + | - |
| TLoST02 | + | - |

Table 8.6: Results after 24 hours shown that TLoST01/ pKD46RecAPa, TLoST02 do not grow under the presence of 5-FOA. DH5$\alpha$ was used as control.

#### 8.1.0.4 *pyrF* gene amplification

Primers pyrFcF and pyrFcR were used to amplify the *pyrF* gene in order to verify their presence in the strains *TLoST01/pKD46RecAPa, TLoST02*. The experiment was performed in order to compare the PCR results with the ones performed in M9 media. As shown in Fig. 8.1 a PCR product for *pyrF* could be amplified, confirming that the strain was not correctly knocked out.

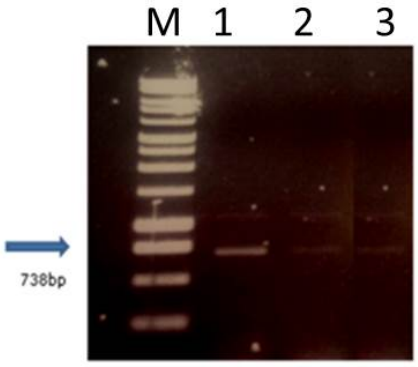



Figure 8.1: **Agarose gel electrophoresis showing PCR product of pyrF gene amplification.** In lane 1 is the DNA 1kb ladder from Promega, used to determine the band size of the amplified fragments. In lane 2 is a clear PCR product, amplified from DH5$\alpha$. Lanes 2 and lane 3 show respectively the products amplified from the strains *TLoST01/pKD46RecAPa* and *TLoST02*

## 8.2 Characterization of PBluescript::Tape

In previous work, the tape system has been cloned in the pBluescript vector to produce pBluescript::Tape. Further analysis of characterization has been performed, to check the

presence of the tape system in the plasmid, by antibiotic sensitivity assay,restriction analysis and sequencing.

#### 8.2.0.1 Antibiotic sensivity

The plasmid contains the *bla* gene, encoding for beta-lactamase that confers resistance to ampicillin. The DH5$\alpha$ strain has been transformed by electroporation with the plasmid and grown on solid agar with carbenicillin 100 $\mu$g/mL.

#### 8.2.0.2 Restriction analysis

pBluescript::Tape DNA was extracted and digested with *XhoI* and *SacI* to release and verify the 9-kb Tape fragment that should be present and the 2.9-kb pBluescript backbone. Electrophoresis gel of the digestion showed, in several occasions, abnormal clones as shown in Fig 8.2.

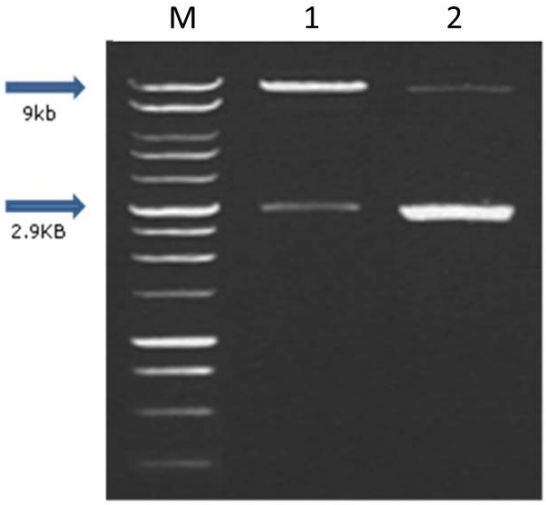



Figure 8.2: **Electrophoresis gel showing the digested pBluescript :: Tape**. 500 ng of DNA were loaded in lanes 1 and 2. In Lane M is showed the Promega 1Kb ladder. Lane 2: Digestion with *SacI/XhoI*, normal pattern Lane 3 Digestion with *SacI/XhoI*, abnormal pattern.

#### 8.2.0.3 Sequencing

Extremities of the insert present in pBluescript::Tape showing the expected digestion pattern was sequenced with universal primers T3 and T7 and were analyzed using SeqMan Pro sequence assembly programme (Fig 8.3).

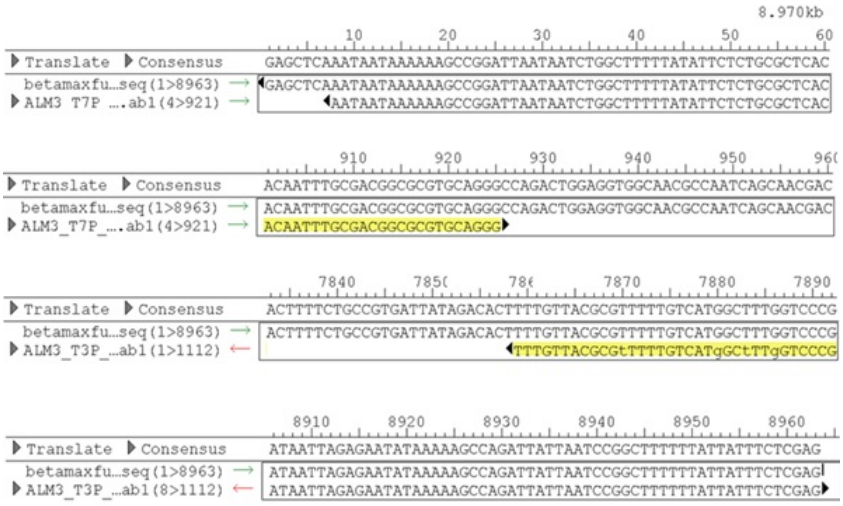



Figure 8.3: **Sequencing analysis.** Results shows that Tape ends were free of error

## 8.3 Bioinformatic Analysis

The original Tape sequence was synthesised by GeneArt ™ Gene Synthesis. The sequence annotation was provided in the Artemis table format [129]. To further validate any defect in the system sequence, I identified all the system components using NCBI's Blast. The analysis revealed the absence of the lacO (AATTGTGAGCGGATAACAATT ) in the whole sequence, a short DNA fragment up-stream of the promoter for LacI repressor which binding is very crucial, the system will not work without this operator in the sequence. The correct Tape full sequence annotation is provided in the Appendix A.

### *8.3.1 Module I repair*

In this work I propose a strategy to repair the module I. The original Tape system design (EPSRC project EP/J004111/1 "The Logistic of small things, unpublished) has been found lacking the lactose operator (lacO) site.The system will not work without these operators in sequences. LacO wild type has been improved in a tight bind version, named "symmetric lacO", consisting of 15 bais pair segment from the half of the natural operator sequence [130]. Primers Mod1R Containing pTac – lacOsym -NotI, and Mod1F were used to amplify the Module I from a previous version of minitn7::bmx received at the beginning of the project. PCR product was cloned into the Thermo Scientific CloneJET ™ PCR Cloning Kit and sequencing and enzymatic digestion confirmed that the module was successfully repaired as shown in Fig 8.4.

## 8.4 Construction and Characterization of PTape01

To insert the tape system in the *E.coli* chromosome was used a delivery plasmid based on the Tn7 transposition. Firstly, I performed the cloning of the tape in the miniT7 delivery plasmid. Growth conditions for the strain carrying the miniTn7::Tape (pTape01), was optimized. The last step, consists of the insertion of the tape in the chromosome of *E.coli* and an analysis of the cell population dynamic. Finally, preliminary results are shown using IPTG to induce the system.

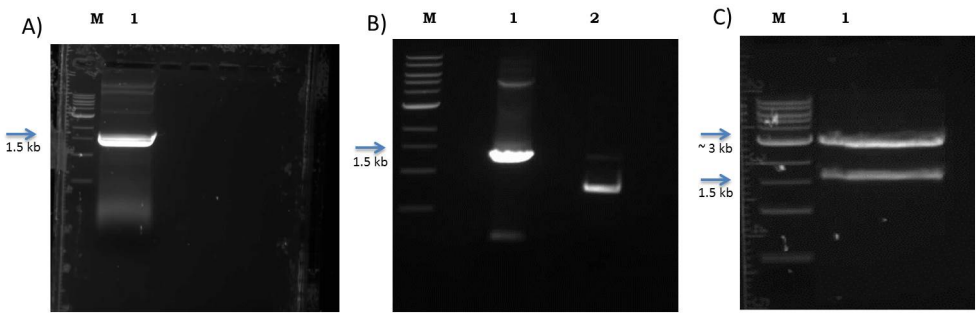



Figure 8.4: **Module I Containing pTac – lacOsym -NotI**.Panel A) M: 1kb NEB ladder;Lane 1: Module I + lacO PCR product 1.5 Kb. Panel B) M: 1kb NEB ladder;Lane 1: pJet vector :: Module I + lacO PCR product 1.5 Kb Lane 2: pJET1.2/blunt Cloning Vector vector without insert used as control. Panel C) M: 1kb NEB ladder; Lane 1: pJET1.2::lacO digested with *SacI/NotI*

### *8.4.1 Cloning and transformation*

The Tape system lacking of the lacOsym was cloned in the mini-Tn7 delivery plasmid (pTape01) and then repaired with the right module I. Both insert and vector were digested with *SacI/XhoI*. *E. coli CC118(λpir)* was used as bacterial host. Transformants were selected in LB agar plates with trimethoprim 160$\mu$g/mL and tetracycline 25$\mu$g/mL The plasmid was extracted and digested with *SacI/XhoI* in order to verify the ligation (Fig 8.5).

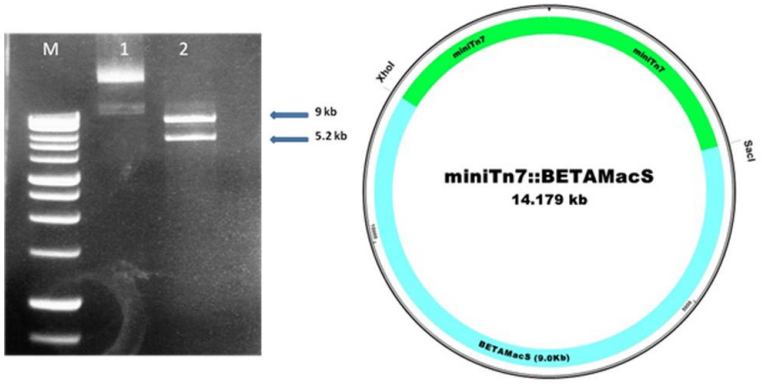



Figure 8.5: **Electrophoresis gel showing the digested pTape01**. 500 ng of DNA were loaded in lanes 1 and 2. In Lane M is shown the Promega 1Kb ladder. Lane 1: PUC18R6K::Tape not digested. Lane 2: PUC18R6K:: Tape digested with SacI/XhoI, with the corresponding bands of Tape 9kb and PUC18R6K::TcR of 5.2 Kb; map of PUC18R6K::TcR miniTn7 delivery plasmid with the Tape .

#### 8.4.1.1 Antibiotic concentration

Antibiotic concentration essay was performed to define the amount of trimetroprim antibiotic which could inhibit the visible growth of the host cell after overnight incubation. Results shows that Tape keep on growing with concentration until 160 $\mu$g/mL.

### *8.4.2 Optimization of media*

*E. coli DH5*$\alpha$ and *E. coli BW25113* were used as positive control, whilst *E. coli JW1273* was used as the negative control. Optimization of the media was performed for growing the *E. coli* $\Delta$*pyrF* cells carrying the Tape system selectable in the presence of 5-FOA and trimethoprim. Tape strains were grown in minimal media (M9 media) supplemented with MgSO4 (2mM), casaminoacids and vitamins. Tape carries a *pyrF* gene, knocked-out to render the strain auxotroph for pyrimidines, required as a selection and counter-selection marker in this system. *E. coli* strains with this gene knocked-out should not grow in the absence of uracil in minimal medium.

In Fig 8.7 results after 24 hours at 37°C and in Fig 8.8 with and without uracil supplementation.

LB, RPMI, AB and M9 media were chosen for minimal media optimization experiments due to their common use in molecular biology.

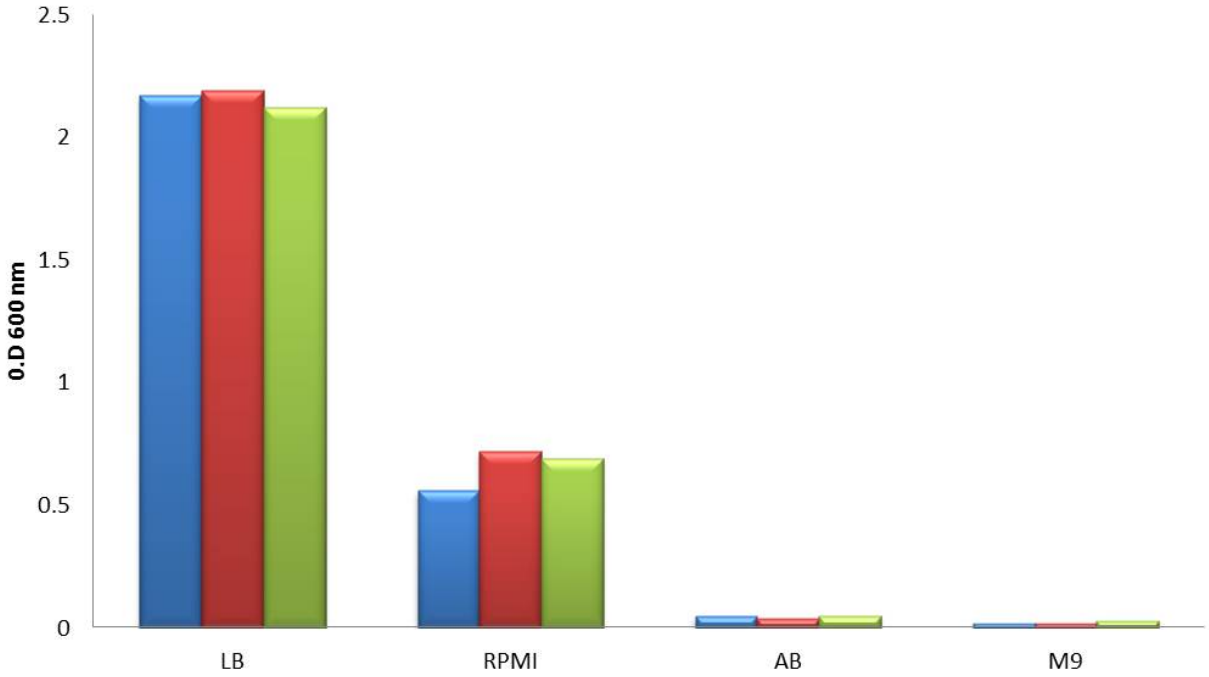



Figure 8.6: **Growth comparison in different defined media at 37°C**.Growth comparisons between strains *DH5α* (blue bar), *JW1273* (*pyrF*-, red bar), and *BW25113* (wt, green bar). In LB growth is visible after 24h with a Max O.D of 2.2. In RPMI growth is visible after 24h with a Max O.D of 0.7. In M9 media supplemented with vitamins mixture growth is visible after 24h with a Max O.D of 1,7

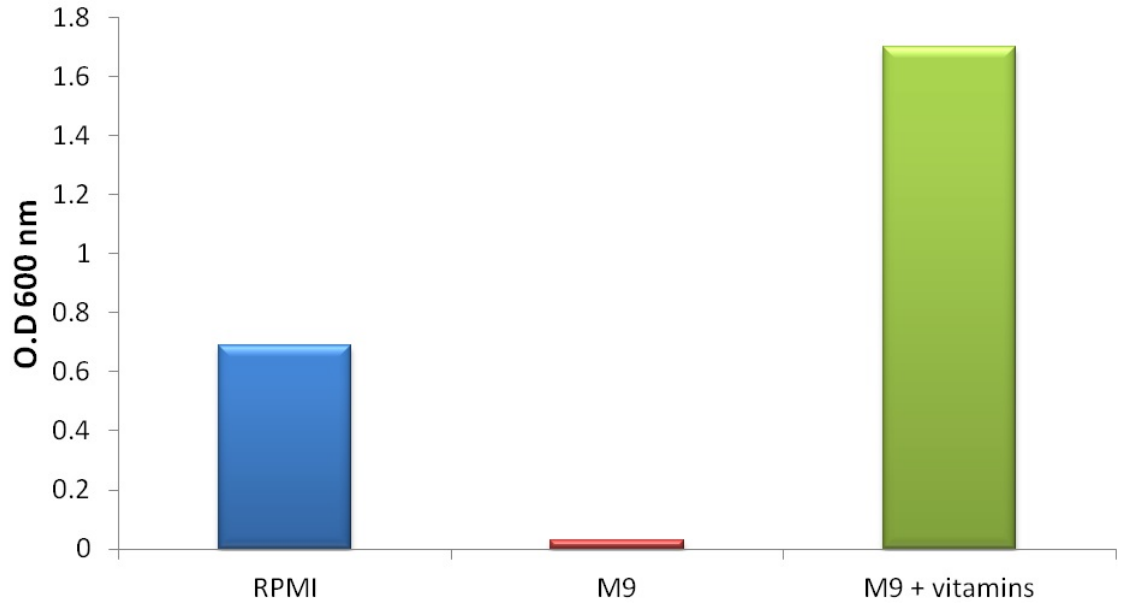



Figure 8.7: **Growth comparison between RPMI, M9 and M9 +vitamins.** *E.coli BW25113* growth comparison between RPMI media M9 media and M9 supplemented with vitamins mixture, at 37°C. In RPMI growth is visible after 24h with a Max O.D of 0.7. In M9 media not supplemented with vitamins mixture growth is visible after 24h with a Max O.D of 0.03. In M9 media supplemented with vitamins mixture growth is visible after 24h with a Max O.D of 1,7

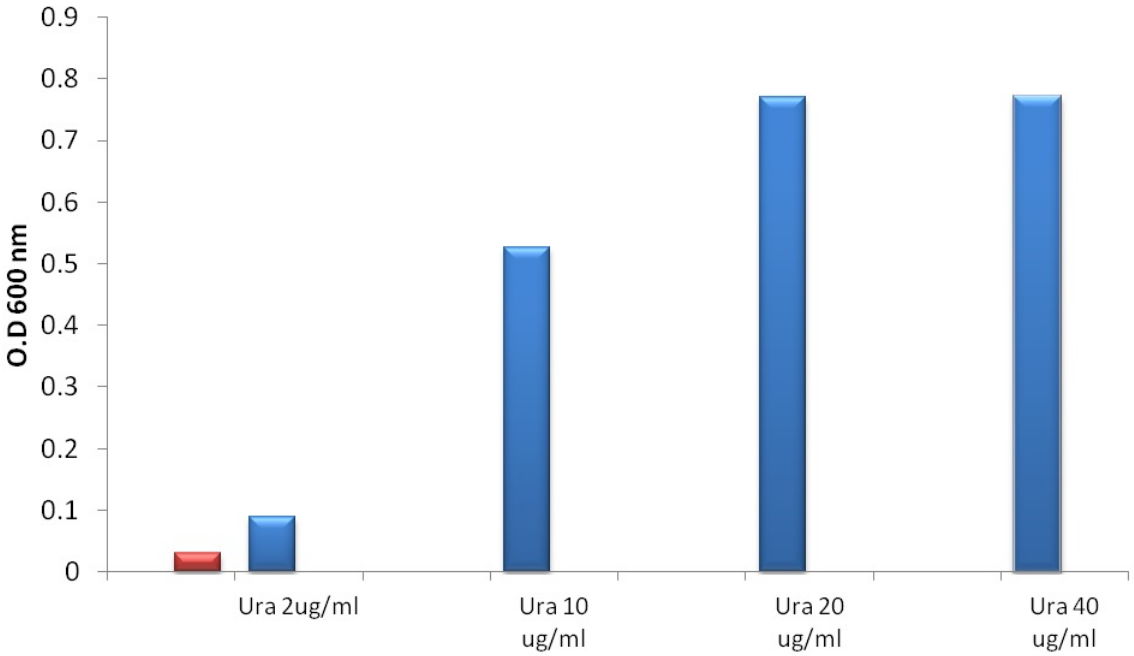



Figure 8.8: **Growth comparisons between *JW1273* (*pyrF*-) in M9 + vitamins medium, with and without uracil supplementation.** Strain *pyrF*- are not able to grow without uracil supplementation (red bar). Different range of uracil were used to detect the suitable concentration (blue bars) necessary for strain lacking of pyrF activity. Results show that growth is visible with a concentration of at least 10$\mu$g/ml of uracil.

### *8.4.3 Chromosome insertion of pTape01 in E. coli*

Colony PCR was performed to confirm the insertion of mini-Tn7 in the chromosome of *E. coli* (Fig 8.9 ) .The mini-Tn7 element should be inserted downstream the *glmS* gene and upstream the *pstS* gene, with a specific orientation [115].

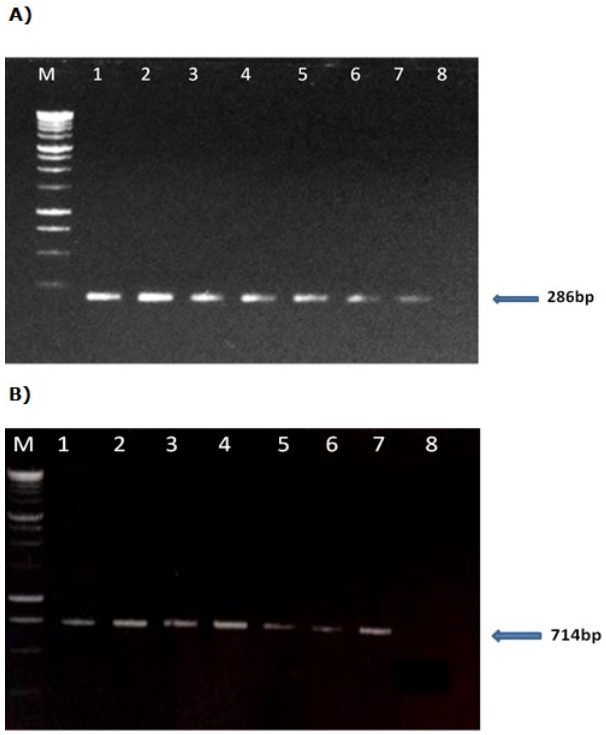



Figure 8.9: **Electrophoresis gel showing the PCR products using primers Tn7R/attTn7-1.** .(panel A) and primers Tn7L/attTn7-2 (B) In Lane M is showed the Promega 1Kb ladder. Lane 1 to 6: JW1273::Tape colonies were used as samples. Lane 7: DH5$\alpha$::Tn7 as positive control. Lane 8: JW1273 as negative control.

### *8.4.4 Tape01 Population dynamics*

Instability of the system can be due: a) accidental excision between attB and attP as consequence of a spurious expression of the integrase which provoke loss of trimetoprim resistance; b) spurious transcription from the tac promoter, followed by spurious retrotranscription into non-specific target which provoke loss uracil auxotrophicity. Therefore it is necessary estimate the percentage of colonies spontaneously reverse to trimetoprim sensitive (TmS) or pyrF+, in M9 minimal medium in standby conditions (Fig 8.10 panel A and C) and by adding 1mM arabinose as inducer (Fig 8.10 panel B and D). M9 media with trimetoprim 100 $\mu$g/mL was used for the estimation of antibiotic loss while M9 media with uracil 20 $\mu$g/mL was used for the estimation of auxotrophicity loss.

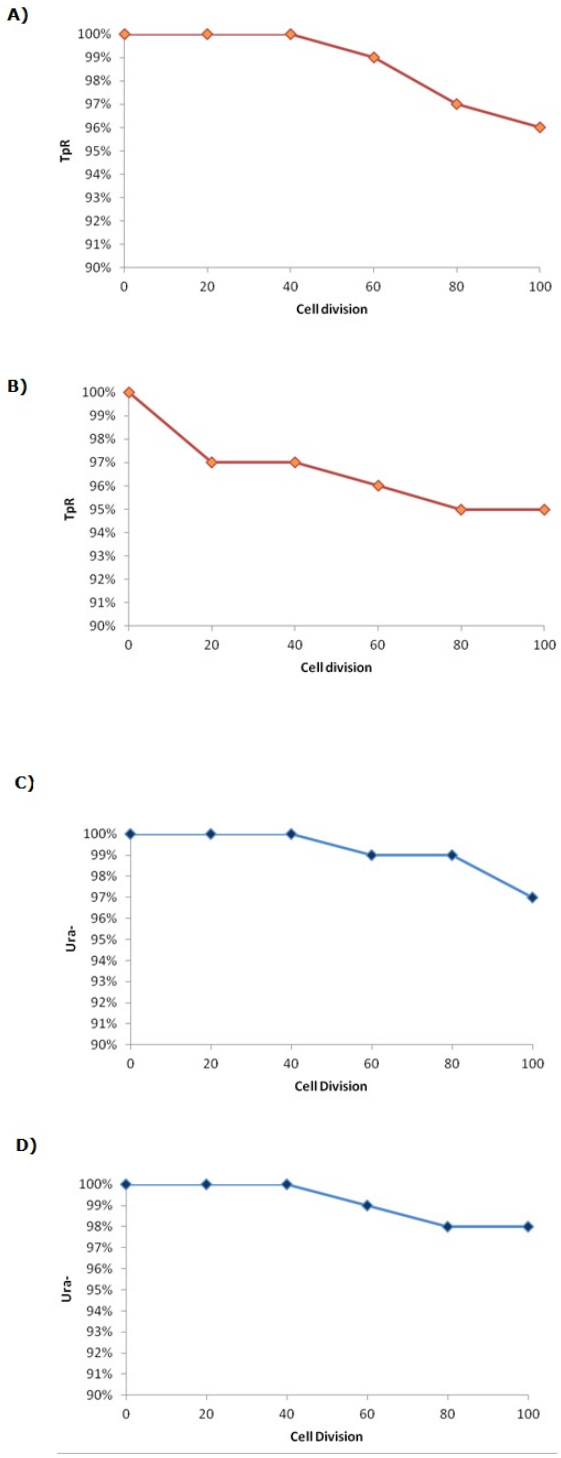



Figure 8.10: **(A) Trimethoprim resistance loss in M9 media:** accidental excision between *attB* and attP can be detected by the loss of Trimethoprim resistance. The curves shows the percentage of colonies trimetoprim sensitive (TmS) after 100 cell division in standby. **(B)** Trimethoprim resistance loss in M9 media + 1mM arabinose. **(C) auxotrophicity loss in M9 media:** retrotranscription into non-specific target can be detected by the loss of Uracil auxotrophicity. The curve shows the percentage of colonies pyrF+ after 100 cell division **(D)** Uracil auxotrophicity loss in M9 media + 1mM arabinose .

### *8.4.5 Recording of one event*

During this study a strategy has been designed in order to verify the recording of one event (1bit).

Tape 2.0 strain has been induced with IPTG 0 mM 0.5 mM, 1 mM, 2 mM and 4 mM and incubated for 1.5 h, 3 h, 6 h, 8h and overnight ( 12h), has showed in Table 8.7.

| | **0 mM** | **0.5 Mm** | **1 mM** | **2 mM** | **4 mM** |
|---|---|---|---|---|---|
| 1.5 h | - | 2 | - | - | - |
| 3 h | 10 | - | - | - | - |
| 6 h | 5 | 9 | - | - | 1 |
| 8 h | - | - | - | 2 | - |
| 12 h | 3 | 3 | - | - | - |

Table 8.7: Results after IPTG induction.

## 8.5 Conclusions

The tape system characterization results show that it is very unstable and not easy to control, therefore making the full implementation very challenging. I have however stably introduced into the chromosome the tape data structure and future optimisation of the various devices components might render the device more controllable. This chapter concludes Part IV of this thesis.

# Part V

# Chapter 9

# Discussion and Conclusions

The objective of the approaches presented in this dissertation, is to produce a DNA data structure, that can be in principle programmed and embedded in living cells, offering multiple applications in bio and nanotechnologies. My target is to advance the state of the art of designing biological embedded data structures implemented at the nanoscale level. Toward my goals I presented:

1. an *in vitro* "list-like" represented by a DNA stack data structure. The work present the design, engineering and characterization of the prototype of a stack machine that offers control via DNA hybridization and DNA strand displacement

2. an *in vivo* "list-like" represented by a DNA tape data structure. The work presents the design, engineering and preliminary characterization of the prototype moving the first steps towards an *in vivo* implementation of a tape data structure that exhibits control via site specific recombination.

3. identification of limitations and solutions of DNA data structure designs.

The next sections provides a discussion and conclusion of the contents introduced in the previous chapters, for the two routes presented in this dissertation.

## 9.1 Evaluation of the DNA Stack Data Structure

This dissertation presents a design and experimental evidence for the working of an *in-vitro* signal data storage based on DNA strand assembly and displacement. The prototype implements a stack data structure with push and pop operations and allows for storing two signal types with the usage of DNA oligonucleotides, indicated as "bricks". The implementation requires the usage of standard simple reagents, but the characterization requires the usage of sophisticated equipments.

### *9.1.1 Limitation of the quantitative analysis*

Experiments revealed a decreasing of the mobility when the stack is elongated during the data storage cycle, and a faster band migration when the opposite is true during the reading cycle, in both agarose and capillary electrophoresis. Because we employ single stranded DNA with a partial secondary structure, the electrophoresis analysis software does not correctly detect molecular concentrations, which prevents us from gaining a precise quantitative picture of the involved processes. The DNA High sensivity kit, can quantify with high affinity dsDNA. Discrepancies between the known brick sizes and the sizes derived by the software from comparison to the ladder might be attributed to two reasons: firstly, short bricks such as *start*, *read* and *report* are well below the detection limit of the High sensitivity kit, which can resolve ssDNA fragments between 50 -7000 base pairs in length, with an accuracy of $\pm$10% from 50 to 600 bp and $\pm$ 20% from 600 to 7000 bp (according to manufacturer specifications). Secondly, the reported deviations might lie in the fact that our bricks contain extensive secondary structures that might affect their motility in the gel matrix. Despite some limitations, it is possible to observe that the data storage cycle works with relatively high fidelity. However it is difficult to quantify the molar yield due to the changing secondary structures of the DNA bricks. Nonetheless, capillary electrophoresis and TEM imaging indicate that the nanodevice is able to store at least three consecutive signals and does not suffer from problematic runaway processes. However, after storing several signals, electrophoresis analysis indicates that the device is not only present in the desired final state, but also in several intermediate recording states. Because of the limits of experimental quantification, we can currently not offer a satisfying explanation for these

intermediate peaks. This currently impacts the readout cycle, as the pop operation interacts with all present stacks and thus returns a superposition of recorded signals. While this is contrary to the intended working, such a superposition might also have advantages, as it might allow one to reverse engineer the composition and order of recorded information from a single electrophoresis read out. Other issues that could quite explain the difference in fluorescence could be: pipetting differences and eventually sample storage and temperature that could lead to a difference in signal strength and a consequent over or under estimation when the software tries to normalize the values.

### *9.1.2 Imaging limitation of the DNA stack data structure*

The ssDNA bricks have been conjugated with gold for the TEM measurements. Note that the particle can have geometric variations, that I confirmed with some measurements taken from the image analysis software, to verify the particle size and distances. Some images are clearly distinguishable and other look like gold agglomerates, limiting the capacity of distinctly recognise the stack data structure. The work contributes to nanotechnology by proposing a novel design of a nanoscale architecture, assembled via DNA hybridization and displacement cascades where irreversible reactions make possible to write and read information from the nanodevice.

## 9.2 Evaluation of a Tape Data Structure

In the *in vivo* approach, the goal requires the usage of a novel system based on the usage of an an Intron II and $\phi 31$ integrase. The implementation required the usage of standard reagents, and standard microbiology and molecular biology techniques.

### *9.2.1 Strains phenotype characterization*

The strains TLoST01/pKD46RecAPa and TLoST02 were grown in the presence of tetracycline and gentamycine antibiotics, separately. The experiment was performed in order to determine which miniTn7 vector was used to make this construct, because it was not

specified. The strain resulted sensitive to both antibiotics. TLoST02 resulted sensitive to trimethroprim and these results indicate that these strains could not be used. *pyrF*, is a gene encoding for orotidine-5'-phosphate decarboxylase (EC 4.1.1.23). This enzyme carries out an essential step in the pyrimidine biosynthetic pathway. Cells that lack this activity can grow by uracil supplementation on minimal medium, which can be converted to uridine 5'-monophosphate (UMP) through a salvage pathway. Cells that lack the enzyme encoded by pyrF are no longer sensitive to 5'-fluorootic acid (FOA) [123], which provides a powerful counter-selection for the loss of the gene.

### *9.2.2 Minimal Media Optimization*

It is known that *E. coli DH5α* exhibits inferior growth in minimal media (M9) compared to other *E. coli* strains [122]. Inferior growth rates in this strain have resulted from unknown accumulated mutations during its development process. Strains which present a mutated *purB* gene need a supplementation of adenine [122] This gene catalyzes two reactions in de novo purine nucleotide pathway. In *E. coli* there are two pathways for conversion of adenine into guanine nucleotides, both involving the intermediary formation of inosine monophosphate (IMP). The major pathway involves conversion of adenine into hypoxanthine in three steps via adenosine and inosine, with subsequent phosphoribosylation of hypoxanthine to IMP. The minor pathway involves formation of ATP, which is converted via the histidine pathway to the purine intermediate 5-amino-4-imidazolecarboxamide ribonucleotide and subsequently to IMP. For this reason the M9 minimal medium used was improved by adding adenine and experiments showed after the 24 hours a restoration of growth. Comparison between optimized and traditional minimal media demonstrates that vitamins supplementation permit to increase the O.D after 24h growth, whilst a poor growth is observed after 48h using the traditional media composition.

### *9.2.3 pyrF strains characterization*

Strains previously developed were grown on solid minimal media (M9) in the presence and absence of uracil to verify the deletion of *pyrF* gene within the chromosome. Results after 24 hours and 48 hours were dissimilar. Results show that the *pyrF* gene was probably not

deleted because strains did not stop to grow when uracil was omitted. PCR was performed to verify the presence of *pyrF* gene and DH5$\alpha$ was used as control. The reaction produced a band of 738 bp product in both *pyrF* mutants. These results indicate that the strains were not lacking a *pyrF* gene. An *E. coli* strain carrying the tape system cloned in the pBluescript vector containing the ampicillin resistance was sucessfully grown on carbenicillin.

### *9.2.4 Instability of the system*

Restriction analysis of pBluescript::tape plasmids with *SacI* and *XhoI* was performed in order to release the tape 9kb insert from the vector to verify it. In more than a few occasions, abnormal clones were detected. Electrophoresis analysis showed a more intense fluorescence in the 2.9 kb band corresponing to the pBluescript backbone than the 9kb band of the tape, probably due to a recombination event. Such results strongly suggest that the system is unstable, and has a tendency to recombine in undesired ways. In the tape system the int-$\phi$C31 recombinase is controlled by the arabinose promoter (Para). In absence of arabinose, the AraC repressor protein is produced and binds as a dimer to the Para promoter, forming a loop that prevents the RNA polymerase from binding [102]. Expression of AraC can potentially direct transcription from the Para promoter, leading to the expression of int-$\phi$C31 recombinase. This can promote abnormal recombination events to occur in the system, causing unwanted alterations. This effect is more significant especially when the tape is cloned into pBluescript, since it is a high copy number plasmid and makes the probabilities of the recombinase being produced higher. In order to avoid undesirable recombination events, the cells containing the tape system were grown in medium supplemented with glucose. This results in carbon source catabolite repression (CCR) [131]. The most efficient carbon source for *E. coli* is glucose, therefore the cells will use this sugar rather than other less efficient carbohydrates. Utilization of glucose causes reduction in the intracellular levels of cAMP, and this in turns reduces the fortuitous induction of the Ptac and Para promoters by the cAMP receptor protein (CAP). In this case, the presence of glucose is expected to tighten the expression from Para. A clone showing the expected pattern in the electrophoresis was purified, and the extremities of the insert in pBluescript sequenced. This confirmed the autheticity of the pBluescript::tape plasmid of this particular clone, which was then stored as a stock at -80°C

### 9.2.5 *Insertion of the tape system into the Escherichia coli chromosome*

*E. coli CC118* ($\lambda$pir) was used as intermediate host for the plasmid pTape01 with the tetracycline cassette and trimethoprim resistance conferred by the synthetic insert, because pUC18R6KT-miniTn7T is R6K a replicon [115], and thus they can only replicate in $\lambda$pir strains. An important goal achieved is represented by the transposition of the tape element into *E. coli* which represents an essential step. Experiments have been performed to promote the transposition of the mini-Tn7 element containing the tape system in the *E. coli* chromosome, co-transforming the strain with the pTNS2 plasmid, encoding the Tn7 transposase genes. PCR was performed to verify the presence of tape insertion in the strain *E. coli* JW1273 where *pyrF* gene is already knocked out. Each pair of primers produced a specific amplification of a fragment of the predicted size respectively of 286bp from the couple Tn7R/attTn7-1 and 714bp obtained by using Tn7L/attTn7-2. *E. coli* BW25113 was used as negative control, and it clearly does not show the amplification of those fragments.

### 9.2.6 *Tape dynamic population*

The tape dynamic population was checked by calculating the colony forming units (cfu) spontaneously reverting to *pyrF*+, after spread in M9 minimal medium lacking uracil in standby condition and adding arabinose. The gene coding integrase $\phi$31 is actively transcribed once the arabinose is added. Instability of the system can be due: a) accidental excision between *attB* and *attP* as consequence of a spurious expression of the integrase; b) spurious transcription from the Tac promoter, followed by spurious retrotranscription into non-specific target. Arabinose increase the frequency of undesired recombination. The Perutka alghoritm has been used to re-target the transposon and generate the *attR* target sequence. The *attR* site is present in the target plasmid but it has three substitutions with respect to a true *attR* site. The third substitution is unexpected, likely due to a PCR error, and there are no evidence if these substitutions can affect the efficiency of the retrotransposition.

### *9.2.7 Intron re-targeting*

An enzymatic digestion of all the clones containing the plasmid was performed after the recording experiment to verify the 900 bp insertion in the disrupted *attR* site. The selection is necessary to identify the required phenotype capable to store information. It was not possible to identify any clone carrying the insertion. One of the possibilities is that the intron is not targeting the *attR* sequence, due to two reasons:

1. The *Ll*.LtrB promotes the RNA splicing of the catalitical RNA structure. To do so it needs the presence of magnesium. Although *E.coli* cellular environment should provide its presence, the exon binding sites (EBS) and intron binding sites (IBS) sequences can interact creating cross links [132], disturbing the retrotransposon stability. The mutate EBS, in this way cannot recognise the *attR* sequence on the tape plasmid. In order to obtain the targeting of the *attR* sequence on plasmid, the intron has to correctly fold into its tertiary structure.

2. Although there is no need for a perfect attR site in the target plasmid to assess the first retrotransposition, to troubleshoot the system is more likely needed a design of a completely optimal target for the existing retrotransposon. To do this, it is important a full understanding of the constraints that define a good "Perutka score" (after the paper by Perutka *et al*. 2004) and reverse-engineer the perfect target sequence.

# Chapter 10

# Future Research Directions

## 10.1 A DNA Stack Data Structure

### *10.1.1 Optimization of the Design*

The work carried out in this thesis has revealed many promising results from a very first design of the bricks set, but also let me suggest some optimizations of the design for the next results:

1. optimize the *start* brick, using a longer sequence. The brick is a 27 nucleotide string that is detected as a roughly 50 base pairs fragment via capillary electrphoresis, due to the ladder limitations. The usage of a longer *start* could lead towards a more precise estimation and detection of the brick, with the advantage of improving the interpretation of the electrophoresis output, without changing the functionality.

2. signal bricks contain a hairpin loop that does not participate in the strand displacement or hybridization. These hairpins can host any desired functionality for experimental detection, such as reporter binding motifs. Vienna Fold simulation [108] shows the stem structure of the *write* with the hairpin loop binding the *reporter*, as expected. On the other hand OxDNA simulations [96] revealed that *report* bricks have difficulties to associate with their respective *write* (Fig.10.1), due to the double helix formation of the loop domain despite sequence non-complementary. The model shows that those

domains can hybridize, although they are not exactly complementary. This will just leave the *x* domain in the form of a single strand, which does represent just a small subset of the hairpin. The *report* complementary domain *x'* will bind but the limited hairpin space could promote a reversible binding and not the expected irreversible one. Having a longer *h* and *l* hairpin domains could reduce this unexpected hybridization reaction between the two domains and obtain the desired irreversible *write-reporter* binding. If oxDNA prediction, is actually the most accurate one, the scarce *report* binding could quite well explain the difficulty to find presence of structures when looking at the TEM.

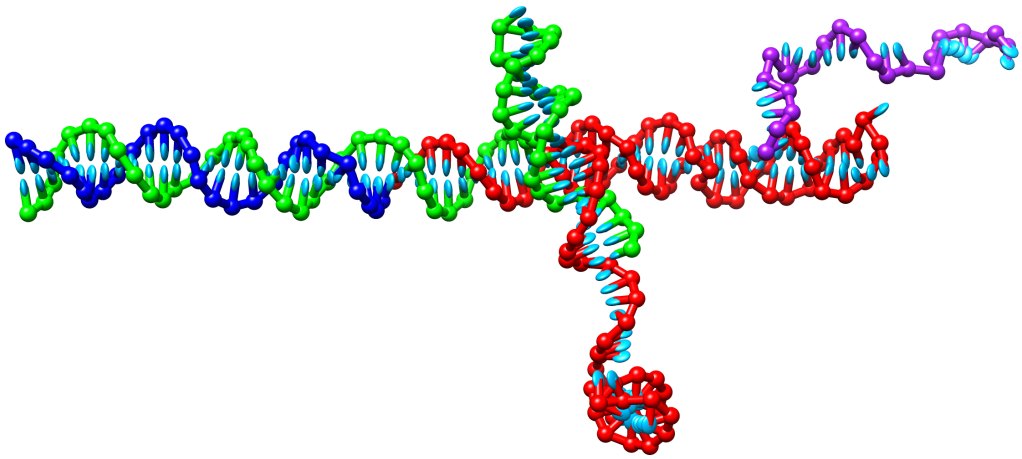


Figure 10.1: oxDNA reporter binding prediction of double helix formation of the loop domain between non-complementary sequences. Start(blue),Push(green),Write(red), Reporter (purple).

### *10.1.2 Optimization of the reading*

The reading cycle concentrations need to be optimized, to prevent runaway processes. Using a concentration of 200 nM (as for the Data storage cycle), seems to dramatically pop elements from the stack. When using half of the concentration, is possible to follow the reading cycle, looking at new species or intermediates, under the form of capillary electrophoresis peaks. To choose the ideal concentration for read and pop, it is required either a quantitative characterization of the results or alternatively a robust prevention of runaway processes.

### *10.1.3 Optimization of the quantitative analysis*

Also, the molecular beacons could represent a more powerful method to follow the hybridization and the displacement. In ref [133] has been found that Cy3 photo-degradation occurs nearly three times more rapidly for externally labelled oligonucleotides. Thus, one of the possibilities is represented by the internal labelling of the bricks hairpin, thus monitoring the branch migration, *e.g,* labelling the *start* b domain (hairpin) and the *push* b' domain (hairpin), respectively with a quencher and a fluoropher. Thus, the fluorescence should decrease whilst the branch migration proceeds. Additionally the plan is to improve experimental analysis methods using different capillary electrophoresis analysis kits (such as RNA assay kits). Better experimental quantification will allow us to calibrate computational models that will in turn help us increase our understanding of the fidelity of the device. Alternatively it could be possible to introduce a washing step, to reduce the surplus bio-bricks, using a streptavidin surface to attach the stack and release it via strand displacement (*e.g.,* designing a start releaser).

### *10.1.4 In vivo implementation*

Ultimately, the goal is to move forward into an *in vivo* implementation. For a Universal construction and explore the stack backbone as programmable scaffold for large-scale assembly of DNA origamis, and ideally achieving Universal computing with the designs of multiple simultaneous stacks and signals that can be written and read multiple times. Alternatively, the device could be used to programmatically and reversibly arrange matter such as liposomes [134, 135] or DNA origami on the nanoscale. The design is based on ssDNA bricks, and the entire data structure could, in principle, be expressed *in vivo* by a living cell as an RNA data structure and post-transcriptionally controlled. All strands (except of start and record) have already been designed to start with a sequence that encodes a promoter. As the data storage happens in a double-stranded fashion rather than in dangling single strands, an *in vivo* realization is likely to suffer less from enzymatic attack. Also, bricks have been designed with washing through RNAse in mind. A dangling strand would thus be susceptible to 3'$\rightarrow$ 5' RNAse digestion since this design meets the requirements.

## 10.2 A Tape Data Structure

### *10.2.1 Testing the system limits*

The results presented in this thesis indicate that it is necessary to test the limits of the system, control of levels and timings of recombinase expression, recording capabilities and number of inputs able to be recorded. It could be also very useful to test the tape without the reset module carrying the araC used just for the resetting. Fortunately, the system was designed since the beginning with plenty of unique restriction sites all over the sequence, so that replacing or removing a module should be fairly easy to do. As modelling in synthetic biology represent an approach to optimize the synthetic systems and make their functions and usage more predictable, a version of the tape modelling could be developed, but there is a lack of tools that permit to model retrotransposons.

### *10.2.2 Gene fusion plasmid*

The usage of a repressor with the *attR*, could lead to engineer an *attR* reporter. The idea is to create a hybrid gene formed from the *attR* sequence and a repressor (*e.g Cro* from *E.coli* [136]). The gene is controlled by a Promoter (Px) and the transcription starts from the AUG start codon, proceeding until the TGA stop codon. In this condition the repressor is transcribed and translated and will bind a sequence upstream the chosen reporter *e.g,* antibiotic, light reporter or a fluorescence protein. When the *attR* is disrupted by the insertion of the stored bit of information, as a result, it can occur the transcription and translation of the reporter, as the repressor will not be transcribed (Fig.10.2)

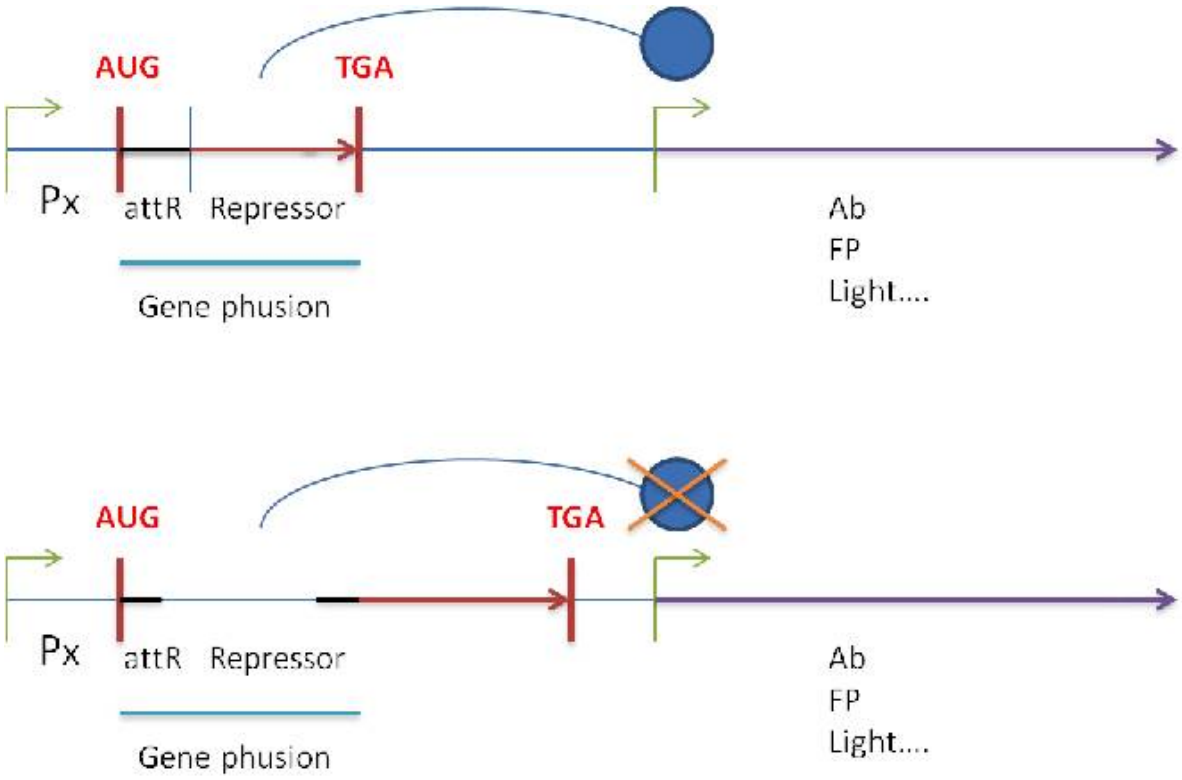



Figure 10.2: Gene fusion reporter idea for a new tape plasmid.

### *10.2.3 Methods to verify the data storage*

Tape strains able to record 1 bit can be selected with 5-FOA and uracil, or alternatively with a plasmid PCR to verify the 900 bp insertion in the disrupted *attR* site. The selection is necessary to identify the required phenotype capable to store information. A similar experiment needs to be run for the resetting of the system using arabinose.

### *10.2.4 Tape platform optimization*

*E. coli JW1273* has been chosen as potential synthetic biology platform. In order to avoid any potential interference with the tape system, the *lacI* gene encoding for the lac repressor in Lac Operon could be knocked out, although this is not an essential requirement. On the contrary, it may end up being beneficial to have additional copies of this repressor gene. Other strains or even organisms are not excluded to be tested as alternative platforms. Once the recording experiment will be successful will be important to understand after how many generations the long-term memory is maintained, expecting a performance at least of 100

cells division [137].

# Appendix A

# Appendix

## A.1 Kit, Reagents, Equipments and Software

All Kit, reagents, equipments and software used for the the engineering and characterization of the presented prototypes are listed in Tables A.1, A.2, A.3, A.4.

| Kit | Supplier |
|---|---|
| Zymoclean Gel DNA Recovery Kit | Zymo Research |
| QIAquick Gel Extraction Kit | Quiagen |
| *Escherichia coli* 5-alpha Competent *E. coli* (High Efficiency) | NEB |
| Plasmid Miniprep Kit | Quiagen |
| Quiagen Plasmid Purification | Quiagen |
| CloneJET PCR Cloning | Thermo Fisher Scientific |
| DNA High Sensivity Kit | Agilent Technology |
| Amicon Ultra-0.5 mL Centrifugal Filters for DNA | Amicon |

Table A.1: **Kits used in this study**

| Reagents | Supplier |
|---|---|
| potassium phosphate monobasic ($KH_2PO_4$) | Sigma Aldrich |
| ammonium chloride ($NH_4Cl$) | Sigma Aldrich |
| ammonium sulphate ($(NH_4)2SO_4$) | Sigma Aldrich |
| Yeast Nitrogen without amino acid | Sigma Aldrich |
| 5-FOA | Sigma Aldrich |
| RPMI-1640 medium | Sigma Aldrich |
| sodium chloride (NaCl) | Sigma Aldrich |
| Glycerol | Sigma Aldrich |
| Sucrose | Sigma Aldrich |
| Dextrose | Sigma Aldrich |
| D-Ca panthenate | Sigma Aldrich |
| Choline c-hloride | Sigma Aldrich |
| i-Inositol | Sigma Aldrich |
| Pyridoxine-HCl | Sigma Aldrich |
| Riboflavin | Sigma Aldrich |
| Thiamine-HCl | Sigma Aldrich |
| Magnesium sulfate ($MgSO_4$) | Sigma Aldrich |
| casaminoacids | Sigma Aldrich |
| T4 DNA ligase | Promega |
| Nuclease free water | Sigma Aldrich |
| GoTaq polymerase | Promega |
| Deoxynucleotide (dNTP) Solution Mix | NEB |
| UltraPure TAE Buffer, 10X | Invitrogen |

Table A.2: **Reagents used in this study**

| Equipments | Supplier |
|---|---|
| Thermomixer Comfort | Eppendorf |
| T3000 Thermocycler | Peltier technology |
| Thermocycler 48 | Sensoquest |
| Horizontal Electrophoresis Systems | BIO-RAD |
| 2100 Bioanalyzer | Agilent Technology |
| MicroPulser | BIO-RAD |
| CM-100 Compustage (FEI) Transmission Electron Microscope | Philips |
| AMT CCD camera | Deben |

Table A.3: **Equipments used in this study**

| Software | Description |
|---|---|
| Primer3 | web interface for primer design |
| Agilent Technologies 2100 Bioanalyser - 2100 Expert v.B02.08.SI1648 (SR3) | web interface for Agilent Bioanalyser data. |

Table A.4: **Software used in this study**

## A.2 DNA Stack Data Structure Assembly

Protocol Example Using 6 write (X-Y-X-Y-X-Y) in a 20 $\mu$L reaction, and 200 nM bricks from a 10 $\mu$M stock

**Eppendorf tube:** X-Y-X-Y-X-Y

1. nuclease free water 17.6 $\mu$L
2. Add START 0.4 $\mu$L
3. shaking 10 minute 300 rpm
4. Add PUSH 0.4 $\mu$L
5. shaking 10 minute 300 rpm
6. Add WRITE X 0.4 $\mu$L
7. shaking 10 minute 300 rpm
8. Add PUSH 0.4 $\mu$L
9. shaking 10 minute 300 rpm
10. Add WRITE Y 0.4 $\mu$L
11. shaking 10 minute 300 rpm
12. Add PUSH 0.4 $\mu$L
13. shaking 10 minute 300 rpm
14. Add WRITE X 0.4 $\mu$L
15. shaking 10 minute 300 rpm
16. Add PUSH 0.4 $\mu$L
17. shaking 10 minute 300 rpm
18. Add WRITE Y 0.4 $\mu$L

19. shaking 10 minute 300 rpm
20. Add PUSH 0.4 $\mu$L
21. shaking 10 minute 300 rpm
22. Add WRITE X 0.4 $\mu$L item shaking 10 minute 300 rpm
23. Add WRITE Y 0.4 $\mu$L
24. shaking 10 minute 300 rpm

## A.3 Plasmids Maps

All plasmid maps used for the the engineering and characterization of the *in vivo* prototype are presented in Figures A.1, A.2, A.3, A.4, A.5.

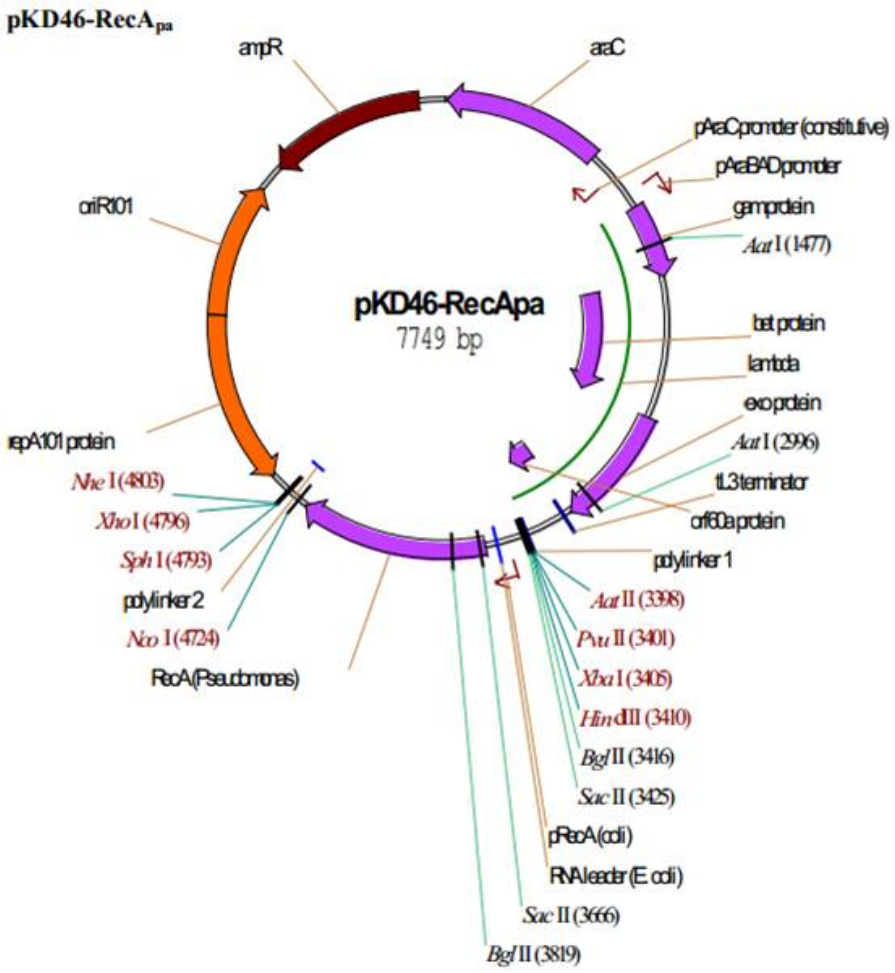



Figure A.1: Thermosensitive plasmid, provides transiently the RecA recombinase necessary for some transposition events, like the insertion of mini-Tn7 constructs in the chromosome. Image taken from [138].

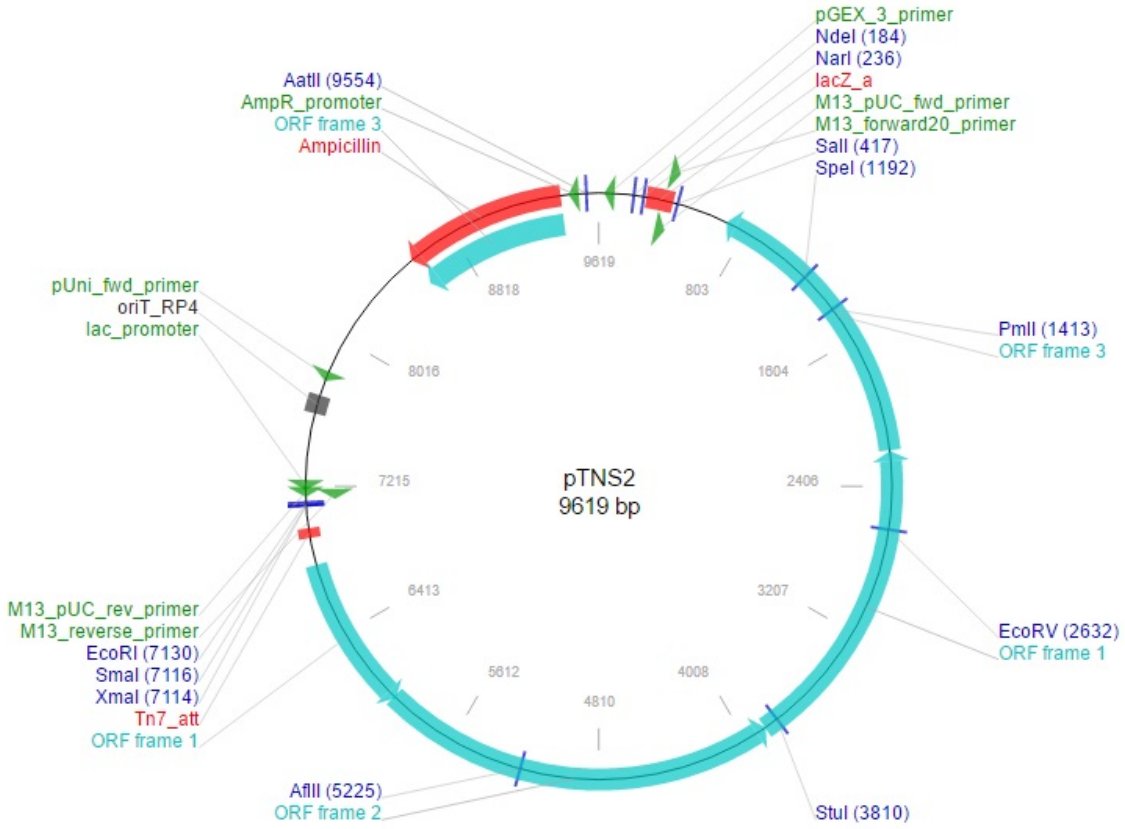



Figure A.2: Helper plasmid pTNS2 for the transposition of miniTn7 elements, provides Tn7-specific transposase genes. Image taken from www.addgene.org/vector-database/.

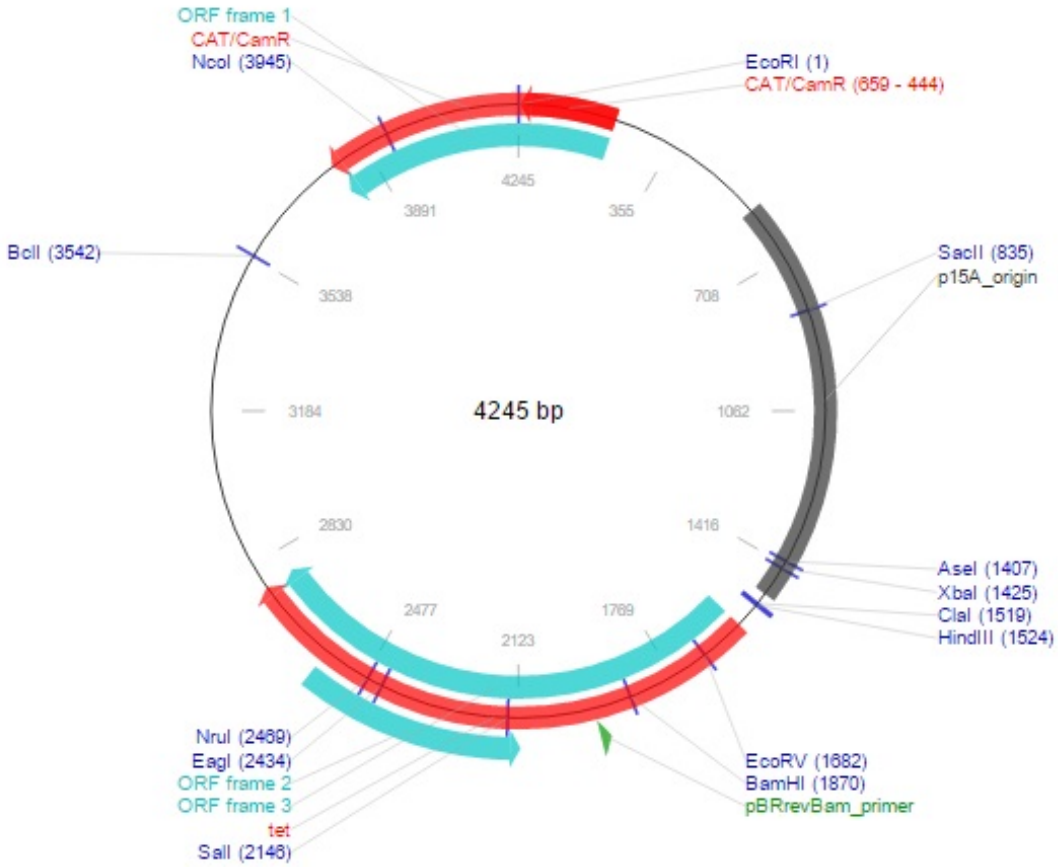



Figure A.3: pACYC184 cloning vector, CmR, TcR, p15A origin of replication, low copy number. Image taken from www.addgene.org/vector-database/.

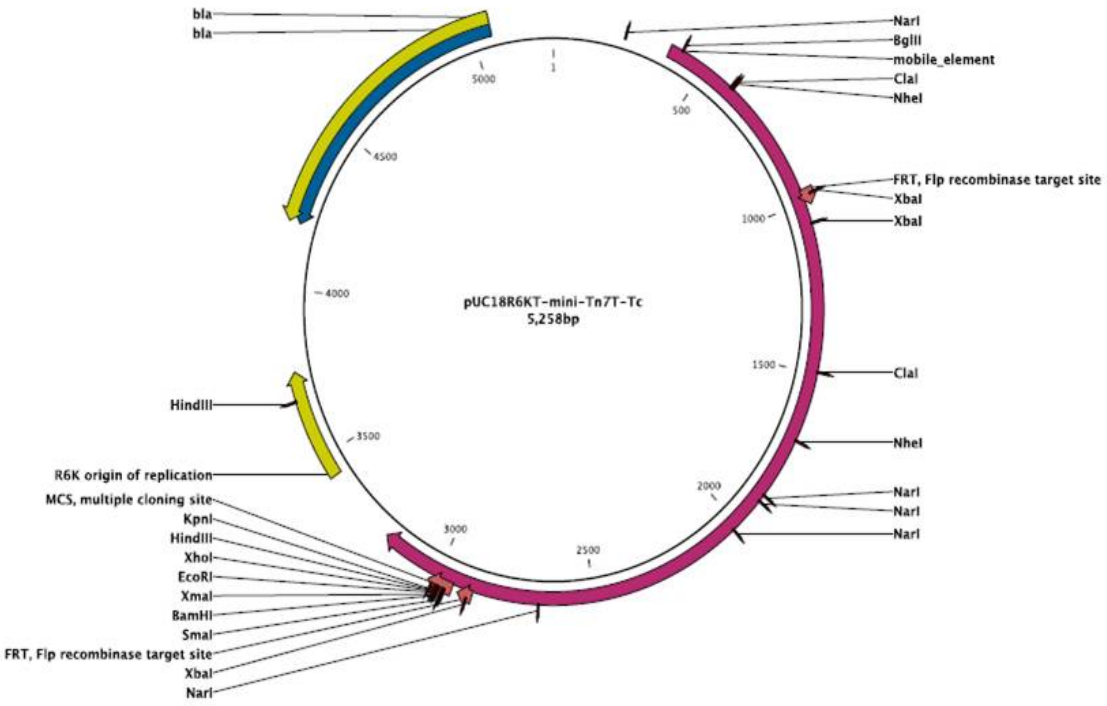



Figure A.4: miniTn7 delivery plasmid (pUC18R6K-mini-Tn7),TcR, ApR, R6K origin of replication.

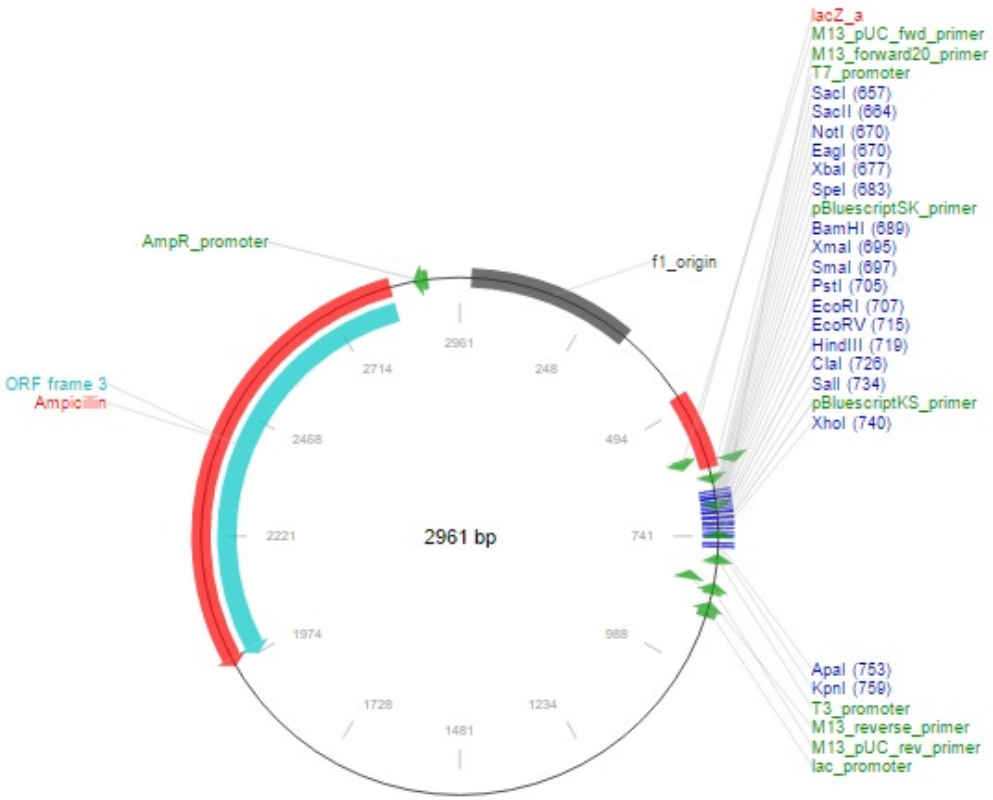



Figure A.5: pBluescript II KS cloning vector, ColE1 replicon, CbR, ApR, lacZ, allowing selection via blue-white screening. Image taken from www.addgene.org/vector-database/.

## A.4 Tape Complete Sequence Annotation

| Position | Functional part | Sequence |
|---|---|---|
| 1-6 | misc_binding (SacI restriction site) | GAGCTC |
| 8-52 | transcriptional terminator | AATAATAAAAAAGCCGGATTAATAATCTGGCTTTTTATATTCTCT |
| 57-1139 | CDS (Lac Repressor) | TCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCA<br>TTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATT<br>GGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAAC<br>AGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCA<br>AGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTT<br>GATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTA<br>TCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCC<br>CGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATC<br>GTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGC<br>ATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGC<br>CTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATA<br>TTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTT<br>AATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGA<br>CCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAA<br>AATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAAT<br>AACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCAT<br>CCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCG<br>TTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACG<br>CCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTT<br>GATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGC<br>GTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGAC<br>TGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAAT<br>TCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGC<br>AGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGA<br>TAAGAGACACCGGCATACTCTGCGACATCGTATAACGTTACTG<br>GTTTCAT |
| 1451-1479 | pTac promoter | TTGACAATTAATCATCGGCTCGTATAATG |
| 1483 | start of transcription from pTac promoter | G |

| | | |
|---|---|---|
| 1481 – 1491 | *NotI* restriction site | GCGGCCGC |
| 1492-2099 | upstream ltrB intron retargeted for *attR* | AAAAAAGCTTATAATTATCCTTACGGGGCGCCCAAGTGCGC<br>CCAGATAGGGTGTTAAGTCAAGTAGTTTAAGGTACTACTCT<br>GTAAGATAACACAGAAAACAGCCAACCTAACCGAAAAGCGA<br>AAGCTGATACGGGAACAGAGCACGGTTGGAAAGCGATGAGT<br>TACCTAAAGACAATCGGGTACGACTGAGTCGCAATGTTAAT<br>CAGATATAAGGTATAAGTTGTGTTTACTGAACGCAAGTTTC<br>TAATTTCGGTTCCCCGTCGATAGAGGAAAGTGTCTGAAACC<br>TCTAGTACAAAGAAAGGTAAGTTACTTTGGGCGACTTATCT<br>GTTATCACCACATTTGTACAATCTGTAGGAGAACCTATGGG<br>AACGAAACGAAAGCGATGCCGAGAATCTGAATTTACCAAGA<br>CTTAACACTAACTGGGGATACCCTAAACAAGAATGCCTAAT<br>AGAAAGGAGGAAAAAGGCTATAGCACTAGAGCTTGAAAATC<br>TTGCAAGGGTACGGAGTACTCGTAGTAGTCTGAGAAGGGTA<br>ACGCCCTTTACATGGCAAAGGGGTACAGTTATTGTGTACTA<br>AAATTAAAAATTGATTAGGGAGGAAAACCTCAAA |
| 2218-2223 | *SphI* restriction site | GCATGC |
| 2225-2280 | *attB* site | TCGAGTGAGGTGGAGTACGCGCCCGGGGAGCCCAAGGGCA<br>CGCCCTGGCACCCGCA |
| 2283-2288 | *BglII* restriction site | AGATCT |
| 2304-2331 | dhfr promoter | TAACCCTGATAAATGCTTCAATAATATT |
| 2337-2342 | RBS | AGGAGG |
| 2349-2585 | CDS (trimethoprim gene) | ATGGAACGAAGTAGCAATGAAGTCAGTAATCCAGTTGCTG<br>GCAATTTTGTATTCCCATCGAACGCCACGTTTGGTATGGG<br>AGATCGCGTGCGCAAGAAATCCGGCGCCGCCTGGCAAGGT<br>CAGATTGTCGGGTGGTACTGCACAAATTTGACCCCCGAAG<br>GCTACGCCGTCGAGTCTGAGGCTCACCCAGGCTCAGTACA<br>GATTTATCCTGTTGCGGCGCTTGAACGCATCAACTGA |
| 2590-3218 | CDS (downstream pyrF exon) | TCATGCACTCCGCTGTAAAGAGGCGTTGATCGCTTTCAG<br>CGTCTGCGCTGGATCTACCGATTGCGTTACCGGGCGACC<br>AATCACCATATAATCAACACCAGCCGACAACGCCTGTTC |

| | | |
|---|---|---|
| | | TGGCGTCATAATGCGGCGCTGGTCACCAGCTTCACTCCC<br>CTGCGGACGAATGCCCGGCGTAACCAGTTTGAACTCCTG<br>ACCGAATACCTGTTTAAAGCGCACAGCTTCCTGAGCAGA<br>ACACACCACACCATCAAGGCCACATTTTTGCGTCAGTGC<br>CGCCAGACGTTCTGCATAATCTGCAGGTGACAGTGTCAT<br>GCCAAGATCGACCAGGTCGCTGGCTTCCATGCTGGTCAA<br>CACTGTCACAGCAATCAAAAGCGGTGCATCTTTGCCAAA<br>CGGAACCAGTGCCTCACGCGCTGCGGTCATCATACGCGC<br>CCCACCAGAGGCATGAACATTCACCATCCACACGCCTAA<br>GTCAGCTGCAGCAGCGACAGCGTGCGCTGCAGTGTTGGG<br>GATATCGTGGAATTTCAGGTCAAGAAAGATATCAAAACC<br>ACGCTGTTGAAGTTCGCGCACAAACTGTGGCCCAAACAA<br>TGTAAACATCTCTTTGCCGACCTTCAGACGACAATCTCT<br>TGGGT |
| 3219-3611 | td group I intron | TAATTGAGGCCTGAGTATAAGGTGACTTATACTTGTAAT<br>CTATCTAAACGGGGAACCTCTCTAGTAGACAATCCCGTG<br>CTAAATTGTAGGACTGCCCTTTAATAAATACTTCTATAT<br>TTAAAGAGGTATTTATGAAAAGCGGAATTTATCAGATTA<br>AAAATACTTTCTCTAGAGAAAATTTCGTCTGGATTAGTT<br>ACTTATCGTGTAAAATCTGATAAATGGAATTGGTTCTAC<br>ATAAATGCCTAACGACTATCCCTTTGGGGAGTAGGGTCA<br>AGTGACTCGAAACGATAGACAACTTGCTTTAACAAGTTG<br>GAGATATAGTCTGCTCTGCATGGTGACATGCAGCTGGAT<br>ATAATTCCGGGGTAAGATTAACGACCTTATCTGAACATA<br>ATG |
| 3612-3720 | CDS (upstream pyrF exon) | CTATCTTGTCGACAAAGGCCAGCGCGTCATCACGATTAT<br>GATAATCAAGGGCAACAACCACAGGAGAATTCGTAACAG<br>CGCGGGAAGAAGATGAAGCAGTTAACGTCAT |
| 3727-3732 | RBS | CCTCCT |
| 3745-3776 | promoter - rrnB p2 promoter (core promoter) | ATTATAGGGAGTTATTCCGGCCTGACAAGAGA |
| 3782-3789 | *AscI* restriction site | GGCGCGCC |
| 3790-3844 | *attP* site | CTAGACCCTACGCCCCCAACTGAGAGAACTCA |

| | | |
|---|---|---|
| | | AAGGTTACCCCAGTTGGGGCACG |
| 3845-3850 | *AatII* restriction site | GACGTC |
| 3851-3971 | PCR tag site 1 | GATTACACATTACAGATTACA |
| 3872-3879 | *SbfI* restriction site | CCTGCAGG |
| 3880-3998 | intron (downstream ltrB intron) | ATTTTTACGAACGAACAATAACAGAGCCGTATA CTCCGAGAGGGGTACGTACGGTTCCCGAAGAGG GTGGTGCAAACCAGTCACAGTAATGTGAACAAG GCGGTACCTCCCTACTTCA |
| 3999-4008 | exon (3' exon site) | CATATCATTT |
| 4012-4017 | *SacII* restriction site | CCGCGG |
| 4018-4023 | RBS | AGGAGG |
| 4030-5829 | CDS (ltrA) | ATGAAACCAACAATGGCAATTTTAGAAAGAATC AGTAAAAATTCACAAGAAAATATAGACGAAGTT TTTACAAGACTTTATCGTTATCTTTTACGTCCA GATATTTATTACGTGGCGTATCAAAATTTATAT TCCAATAAAGGAGCTTCCACAAAAGGAATATTA GATGATACAGCGGATGGCTTTAGTGAAGAAAAA ATAAAAAAGATTATTCAATCTTTAAAAGACGGA ACTTACTATCCTCAACCTGTACGAAGAATGTAT ATTGCAAAAAAGAATTCTAAAAAGATGAGACCT TTAGGAATTCCAACTTTCACAGATAAATTGATC CAAGAAGCTGTGAGAATAATTCTTGAATCTATC TATGAACCGGTATTCGAAGATGTGTCTCACGGT TTTAGACCTCAACGAAGCTGTCACACAGCTTTG AAAACAATCAAAAGAGAGTTTGGCGGCGCAAGA TGGTTTGTGGAGGGAGATATAAAAGGCTGCTTC GATAATATAGACCACGTTACACTCATTGGACTC ATCAATCTTAAAATCAAAGATATGAAAATGAGC CAATTGATTTATAAATTTCTAAAAGCAGGTTAT CTGGAAAACTGGCAGTATCACAAAACTTACAGC GGAACACCTCAAGGTGGAATTCTATCTCCTCTT TTGGCCAACATCTATCTTCATGAATTGGATAAG TTTGTTTTACAACTCAAAATGAAGTTTGACCGA GAAAGTCCAGAAAGAATAACACCTGAATATCGG GAACTTCACAATGAGATAAAAAGAATTTCTCAC |

| | | |
|---|---|---|
| | | CGTCTCAAGAAGTTGGAGGGTGAAGAAAAAGCT<br>AAAGTTCTTTTAGAATATCAAGAAAAACGTAAA<br>AGATTACCCACACTCCCCTGTACCTCACAGACA<br>AATAAAGTATTGAAATACGTCCGGTATGCGGAC<br>GACTTCATTATCTCTGTTAAAGGAAGCAAAGAG<br>GACTGTCAATGGATAAAAGAACAATTAAAACTT<br>TTTATTCATAACAAGCTAAAAATGGAATTGAGT<br>GAAGAAAAAACACTCATCACACATAGCAGTCAA<br>CCCGCTCGTTTTCTGGGATATGATATACGAGTA<br>AGGAGAAGTGGAACGATAAAACGATCTGGTAAA<br>GTCAAAAAGAGAACACTCAATGGGAGTGTAGAA<br>CTCCTTATTCCTCTTCAAGACAAAATTCGTCAA<br>TTTATTTTTGACAAGAAAATAGCTATCCAAAAG<br>AAAGATAGCTCATGGTTTCCAGTTCACAGGAAA<br>TATCTTATTCGTTCAACAGACTTAGAAATCATC<br>ACAATTTATAATTCTGAATTAAGAGGGATTTGT<br>AATTACTACGGTCTAGCAAGTAATTTTAACCAG<br>CTCAATTATTTTGCTTATCTTATGGAATACAGC<br>TGTCTAAAAACGATAGCCTCCAAACATAAGGGA<br>ACACTTTCAAAAACCATTTCCATGTTTAAAGAT<br>GGAAGTGGTTCGTGGGGCATCCCGTATGAGATA<br>AAGCAAGGTAAGCAGCGCCGTTATTTTGCAAAT<br>TTTAGTGAATGTAAATCCCCTTATCAATTTACG<br>GATGAGATAAGTCAAGCTCCTGTATTGTATGGC<br>TATGCCCGGAATACTCTTGAAAACAGGTTAAAA<br>GCTAAATGTTGTGAATTATGTGGAACATCTGAT<br>GAAAATACTTCCTATGAAATTCACCATGTCAAT<br>AAGGTCAAAAATCTTAAAGGCAAAGAAAAATGG<br>GAAATGGCAATGATAGCGAAACAACGTAAAACT<br>CTTGTTGTATGCTTTCATTGTCATCGTCACGTG<br>ATTCATAAACACAAGTGA |
| 5830-5835 | *SpeI* restriction site | ACTAGT |
| 5838-5869 | terminator (Bidirectional Rho independent transcriptional terminator) | AGTCAAAAGCCTCCGGTCGGAGGCTTTTGACT |
| 5872-7989 | CDS (phiC31 recombinase) | CTACGCCGCTACGTCTTCCGTGCCGTCCTGGGC |

| | | |
|---|---|---|
| | | GTCGTCTTCGTCGTCGTCGGTCGGCGGCTTCGC |
| | | CCACGTGATCGAAGCGCGCTTCTCGATGGGCGT |
| | | TCCCTGCCCCCTGCCCGTAGTCGACTTCGTGAC |
| | | AACGATCTTGTCTACGAAGAGCCCGACGAACAC |
| | | GCGCTTGTCGTCTACTGACGCGCGCCCCCACCA |
| | | CGACTTAGGGCCGGTCGGGTCAGCGTCGGCGTC |
| | | TTCGGGGAACCATTGGTCAAGGGGAAGCTTCGG |
| | | GGCTTCGGCGGCTTCAAGTTCGGCAAGCCGCTC |
| | | TTCCGCCCCTTGCTGCCGGAGCGTCAGCGCTGC |
| | | CTGTTGCTTCCGGAAGTGCTTCCTGCCAACGGG |
| | | TCCGTCGTACGCGCCTGCCGCGCGGTCTTCGTA |
| | | CAGCTCTTCAAGGGCGTTCAGGGCGTCGGCGCG |
| | | CTCCGCAACAAGGTTCGCCCGTTCGCCGCTCTT |
| | | CTCAGGCGCCTCAGTGAGCTTGCCGAAGCGTCG |
| | | GGCGGCTTCCCACAGAAGCGCCAACGTCTCTTC |
| | | GTCGCCTTCGGCGTGCCTGATCTTGTTGAAGAT |
| | | GCGTTCCGCAACGAACTTGTCGAGTGCCGCCAT |
| | | GCTGACGTTGCACGTGCCTTCGTGCTGCCCAGG |
| | | TGCGGACGGGTCGACCACCTTCCGGCGACGGCA |
| | | GCGGTAAGAGTCCTTGATCGATTCTTCCCCGCG |
| | | CTTCGAAGTCATGACGGCGCCACACTCGCAGTA |
| | | CAGCTTGTCCATGGCGGACAGAATGGCTTGCCC |
| | | CCGGGAAAGCCCCTTGCCGCGCCCCCTGCCGTC |
| | | CAACCACGCCTGAAGCTCATACCACTCAGCGGG |
| | | CTCGATGATCGGTCCGCAATCAAGCTCGACCGG |
| | | CCGGAGCGTGATCGGGTCGCGCTGAATGCGGTA |
| | | ACCCTCAATCTTCGTGGTCGGCGTGCCGTCCGG |
| | | CTTCTTCTTGTAGATCACCTCAGCGGCGAAGCC |
| | | CGCAATACGCGGGTCCCGAAGGATTCGCATAAC |
| | | GGTTGCCGGGTCCCAGGCGCTTGAAGCGGTCTT |
| | | CTTCCCAATCGTCTCGCCCCGGGTCGGCACGGC |
| | | GTCAGCGTCCATGCGCTTACAAAGCCCCGTGAT |
| | | GCTGCCCGGGTGAATGGCGGCTTGACTGCCCGG |
| | | CTTGAAGGGAAGGTGTTTGTGCGTCTTGATCTC |
| | | ACGCCACCACCACCGGATTACGTCGGGCTCGAA |
| | | CTCGAAGGGTCCGGTAAGGGGAGTGGTCGAGTG |
| | | CGCAAGCTTGTTGATGACGACATTGACCATTCG |
| | | GCCGTTGCGCGTGATCTCCTTCGTCTCCGAAAC |

| | | |
|---|---|---|
| | | AAGCTCGAAGCCGTAAGGCGCCTTCCCGCCGAC<br>GTACCCGCCCAATTCGCGCTGAAGGTTCTTCGT<br>GTCGAGAATCTTCGCCGACTTCAGCGAAGATTC<br>TTTGTGCGACGCGTCGAGCCGCATAATCAGGTG<br>AATCAGGTCCATGACGTTTCCCTGCCGGAAGAC<br>GCCTTCCTGAGTGGAAACAATCGTCACGCCCAG<br>GGCGAGCAATTCCGAGACAATCGGAATCGCGTC<br>CATGACCTTCAGGCGCGAGAAGCGCGACACGTC<br>ATAGACAATGATCATGTTGAGCCGCCCGGCGCG<br>GCATTCGTTCAGGATGCGTTCGAACTCCGGGCG<br>CTCCGCCGTCCCGAACGCCGACGTGCCCGGCGC<br>TTCGCTGAAATGCCCGACGAACCTGAACCGGCC<br>CCCGTCGCGCTCGACTTCGCGCTGAAGGTCGGC<br>CGCCTTGTCTTCGTTGGCGCTACGCTGTGTCGC<br>TGGGCTTGCTGCGCTCGAATTCTCGCGCTCGCG<br>CGACTGACGGTCGTAAGCACCCGCGTACGTGTC<br>CAC |
| 7696-7701 | RBS | CCTCCT |
| 7728-7755 | araBAD promoter | ACAGTAGAGAGTTGCGATAAAAAGCGTC |
| 7754-7792 | operator I2 + I1 sites | TCAGGTAGGATCCGCTAATCTTATGGATAAAAATGCTAT |
| 7788-7801 | CAP site | GCTATGGCATAGCA |
| 7823 – 7844 | operator 1(o1) site | ATCAATGTGGACTTTTCTGCCG |
| 7852 – 7880 | promoter | AGACACTTTTGTTACGCGTTTTTGTCATG |
| 7985 – 8002 | operator 2(o2) site | ATATGGACAATTGGTTTC |
| 8031-8909 | CDS (arabinose repressor araC) | ATGGCTGAAGCGCAAAATGATCCCCTGCTGCCG<br>GGATACTCGTTTAATGCCCATCTGGTGGCGGGT<br>TTAACGCCGATTGAGGCCAACGGTTATCTCGAT<br>TTTTTTATCGACCGACCGCTGGGAATGAAAGGT<br>TATATTCTCAATCTCACCATTCGCGGTCAGGGG<br>GTGGTGAAAAATCAGGGACGAGAATTTGTTTGC<br>CGACCGGGTGATATTTTGCTGTTCCCGCCAGGA<br>GAGATTCATCACTACGGTCGTCATCCGGAGGCT<br>CGCGAATGGTATCACCAGTGGGTTTACTTTCGT |

| | | |
|---|---|---|
| | | CCGCGCGCCTACTGGCATGAATGGCTTAACTGG<br>CCGTCAATATTTGCCAATACGGGGTTCTTTCGC<br>CCGGATGAAGCGCACCAGCCGCATTTCAGCGAC<br>CTGTTTGGGCAAATCATTAACGCCGGGCAAGGG<br>GAAGGGCGCTATTCGGAGCTGCTGGCGATAAAT<br>CTGCTTGAGCAATTGTTACTGCGGCGCATGGAA<br>GCGATTAACGAGTCGCTCCATCCACCGATGGAT<br>AATCGGGTACGCGAGGCTTGTCAGTACATCAGC<br>GATCACCTGGCAGACAGCAATTTTGATATCGCC<br>AGCGTCGCACAGCATGTTTGCTTGTCGCCGTCG<br>CGTCTGTCACATCTTTTCCGCCAGCAGTTAGGG<br>ATTAGCGTCTTAAGCTGGCGCGAGGACCAACG<br>TATCAGCCAGGCGAAGCTGCTTTTGAGCACCAC<br>CCGGATGCCTATCGCCACCGTCGGTCGCAATGT<br>TGGTTTTGACGATCAACTCTATTTCTCGCGGGT<br>ATTTAAAAAATGCACCGGGGCCAGCCCGAGCGA<br>GTTCCGTGCCGGTTGTGAAGAAAAAGTGAATGA<br>TGTAGCCGTCAAGTTGTCATAA |
| 8912-8957 | transcriptional terminator | AGAGAATATAAAAAGCCAGATTATTAATCCGGC<br>TTTTTTATTATTT |
| 8958-8963 | *Xho* restriction site | CTCGAG |

Table A.5: Tape sequence full annotation.

# Bibliography


[1] Carter Bancroft, Timothy Bowler, Brian Bloom, and Catherine Taylor Clelland. Long-term storage of information in dna. *Science (New York, NY)*, 293(5536):1763–1765, 2001.

[2] S Gladwin and Matthew England. Digital data storage system, September 30 2005. US Patent App. 11/241,555.

[3] George M Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012.

[4] George Pasparakis, Natalio Krasnogor, Leroy Cronin, Benjamin G Davis, and Cameron Alexander. Controlled polymer synthesis—from biomimicry towards synthetic biology. *Chemical Society Reviews*, 39(1):286–300, 2010.

[5] George Pasparakis, Maria Vamvakaki, Natalio Krasnogor, and Cameron Alexander. Diol–boronic acid complexes integrated by responsive polymers—a route to chemical sensing and logic operations. *Soft Matter*, 5(20):3839–3841, 2009.

[6] Andrew N Shipway, Eugenii Katz, and Itamar Willner. Molecular memory and processing devices in solution and on surfaces. In *Molecular Machines and Motors*, pages 237–281. Springer, 2001.

[7] Bernard Yurke and Allen P Mills Jr. Using dna to power nanostructures. *Genetic Programming and Evolvable Machines*, 4(2):111–122, 2003.

[8] Courtney L Frazier, Joseph San Filippo, Alan M Lambowitz, and David A Mills. Genetic manipulation of lactococcus lactis by using targeted group ii introns: generation of stable insertions without selection. *Applied and Environmental microbiology*, 69 (2):1121–1128, 2003.

[9] Yi Wang, Xiangzhen Li, Caroline B Milne, Holger Janssen, Weiyin Lin, Gloria Phan, Huiying Hu, Yong-Su Jin, Nathan D Price, and Hans P Blaschek. Development of a gene knockout system using mobile group ii introns (targetron) and genetic disruption of acid production pathways in clostridium beijerinckii. *Applied and environmental microbiology*, 79(19):5853–5863, 2013.

[10] Ron Weiss, Subhayu Basu, Sara Hooshangi, Abigail Kalmbach, David Karig, Rishabh Mehreja, and Ilka Netravali. Genetic circuit building blocks for cellular computation, communications, and signal processing. *Natural Computing*, 2(1):47–84, 2003.

[11] Yaakov Benenson. Biocomputers: from test tubes to live cells. *Molecular BioSystems*, 5(7):675–685, 2009.

[12] Richard Dawkins. *The selfish gene*. Number 199. Oxford university press, 2006.

[13] Gheorghe Paun, Grzegorz Rozenberg, and Arto Salomaa. *DNA computing: new computing paradigms*. Springer Science & Business Media, 2005.

[14] Francis HC Crick and James D Watson. The complementary structure of deoxyribonucleic acid. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 223, pages 80–96. The Royal Society, 1954.

[15] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D Watson, and AV Grimstone. Molecular biology of the cell (3rd edn). *Trends in Biochemical Sciences*, 20(5):210–210, 1995.

[16] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic acids research*, 34(2):564–574, 2006.

[17] Julius Marmur and Paul Doty. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *Journal of molecular biology*, 5(1):109–118, 1962.

[18] Paul WK Rothemund. Folding dna to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.

[19] Shawn M Douglas, Adam H Marblestone, Surat Teerapittayanon, Alejandro Vazquez, George M Church, and William M Shih. Rapid prototyping of 3d dna-origami shapes with cadnano. *Nucleic Acids Research*, page gkp436, 2009.

[20] Nadrian C Seeman. Dna nanotechnology: novel dna constructions. *Annual review of biophysics and biomolecular structure*, 27(1):225–248, 1998.

[21] Nadrian C Seeman. Nanomaterials based on dna. *Annual review of biochemistry*, 79: 65, 2010.

[22] Donald Voet and Alexander Rich. The crystal structures of purines, pyrimidines and their intermolecular complexes. *Progress in nucleic acid research and molecular biology*, 10:183, 1970.

[23] Anne Condon. Designed dna molecules: principles and applications of molecular nanotechnology. *Nature Reviews Genetics*, 7(7):565–575, 2006.

[24] Nadrian C Seeman. Nucleic acid junctions and lattices. *Journal of theoretical biology*, 99(2):237–247, 1982.

[25] Nadrian C Seeman. Dna in a material world. *Nature*, 421(6921):427–431, 2003.

[26] Nadrian C Seeman. An overview of structural dna nanotechnology. *Molecular biotechnology*, 37(3):246–257, 2007.

[27] Nadrian C Seeman. Nanotechnology and the double helix. *Scientific American*, 17: 30–39, 2007.

[28] RP Feynman. Miniaturization. *Reinhold, New York*, pages 282–296, 1961.

[29] Leonard M Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(5187):1021–1024, 1994.

[30] Frank Rubin. A search procedure for hamilton paths and circuits. *Journal of the ACM (JACM)*, 21(4):576–580, 1974.

[31] Lee Zou and Stephen J Elledge. Sensing dna damage through atrip recognition of rpa-ssdna complexes. *Science*, 300(5625):1542–1548, 2003.

[32] Ronald R Breaker. Dna aptamers and dna enzymes. *Current opinion in chemical biology*, 1(1):26–31, 1997.

[33] Steven B Smith, Yujia Cui, and Carlos Bustamante. Overstretching b-dna: the elastic response of individual double-stranded and single-stranded dna molecules. *Science*, 271(5250):795–799, 1996.

[34] Donald M Crothers, Victor A Bloomfield, and Ignacio Tinoco. *Nucleic acids: structures, properties, and functions*. University Science Books Sausalito, 2000.

[35] Ivo L Hofacker. Vienna rna secondary structure server. *Nucleic acids research*, 31 (13):3429–3431, 2003.

[36] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415, 2003.

[37] Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. Nupack: analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1):170–173, 2011.

[38] Thomas E Ouldridge, Ard A Louis, and Jonathan PK Doye. Dna nanotweezers studied with a coarse-grained model of dna. *Physical Review Letters*, 104(17):178101, 2010.

[39] Thomas E Ouldridge, Petr Šulc, Flavio Romano, Jonathan PK Doye, and Ard A Louis. Dna hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic acids research*, 41(19):8886–8895, 2013.

[40] Lulu Qian, Erik Winfree, and Jehoshua Bruck. Neural network computation with dna strand displacement cascades. *Nature*, 475(7356):368–372, 2011.

[41] Lulu Qian and Erik Winfree. Scaling up digital circuit computation with dna strand displacement cascades. *Science*, 332(6034):1196–1201, 2011.

[42] Jong-Shik Shin and Niles A Pierce. A synthetic dna walker for molecular transport. *Journal of the American Chemical Society*, 126(35):10834–10835, 2004.

[43] Tosan Omabegho, Ruojie Sha, and Nadrian C Seeman. A bipedal dna brownian motor with coordinated legs. *Science*, 324(5923):67–71, 2009.

[44] Bernard Yurke, Andrew J Turberfield, Allen P Mills, Friedrich C Simmel, and Jennifer L Neumann. A dna-fuelled molecular machine made of dna. *Nature*, 406(6796): 605–608, 2000.

[45] Lulu Qian, David Soloveichik, and Erik Winfree. Efficient turing-universal computation with dna polymers. In *DNA computing and molecular programming*, pages 123–140. Springer, 2011.

[46] Niranjan Srinivas, Thomas E Ouldridge, Petr Šulc, Joseph M Schaeffer, Bernard Yurke, Ard A Louis, Jonathan PK Doye, and Erik Winfree. On the biophysics and kinetics of toehold-mediated dna strand displacement. *Nucleic Acids Research*, 41 (22):10641–10658, 2013.

[47] David Yu Zhang and Erik Winfree. Control of dna strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47):17303–17314, 2009.

[48] Anthony J Genot, David Yu Zhang, Jonathan Bath, and Andrew J Turberfield. Remote toehold: a mechanism for flexible control of dna hybridization kinetics. *Journal of the American Chemical Society*, 133(7):2177–2182, 2011.

[49] Robert RF Machinek, Thomas E Ouldridge, Natalie EC Haley, Jonathan Bath, and Andrew J Turberfield. Programmable energy landscapes for kinetic control of dna strand displacement. *Nature communications*, 5, 2014.

[50] Jon Applequist and Vinayak Damle. Theory of the effects of concentration and chain length on helix—coil equilibria in two-stranded nucleic acids. *The Journal of chemical physics*, 39(10):2719–2721, 1963.

[51] Yandong Yin and Xin Sheng Zhao. Kinetics and dynamics of dna hybridization. *Accounts of chemical research*, 44(11):1172–1181, 2011.

[52] Dietmar Pörschke and Manfred Eigen. Co-operative non-enzymatic base recognition iii. kinetics of the helix—coil transition of the oligoribouridylic· oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic ph. *Journal of molecular biology*, 62(2):361IN11365–364IN15381, 1971.

[53] D Pörschke, OC Uhlenbeck, and FH Martin. Thermodynamics and kinetics of the helix-coil transition of oligomers containing gc base pairs. *Biopolymers*, 12(6):1313–1335, 1973.

[54] Petr Šulc, Flavio Romano, Thomas E Ouldridge, Lorenzo Rovigatti, Jonathan PK Doye, and Ard A Louis. Sequence-dependent thermodynamics of a coarse-grained dna model. *The Journal of chemical physics*, 137(13):135101, 2012.

[55] Charles M Radding, Kenneth L Beattie, William K Holloman, and Roger C Wiegand. Uptake of homologous single-stranded fragments by superhelical dna: Iv. branch migration. *Journal of Molecular Biology*, 116(4):825–839, 1977.

[56] Cal Green and Clark Tibbetts. Reassociation rate limited displacement of dna strands by branch migration. *Nucleic Acids Research*, 9(8):1905–1918, 1981.

[57] Yaakov Benenson, Tamar Paz-Elizur, Rivka Adar, Ehud Keinan, Zvi Livneh, and Ehud Shapiro. Programmable and autonomous computing machine made of biomolecules. *Nature*, 414(6862):430–434, 2001.

[58] Yaakov Benenson, Binyamin Gil, Uri Ben-Dor, Rivka Adar, and Ehud Shapiro. An autonomous molecular computer for logical control of gene expression. *Nature*, 429 (6990):423–429, 2004.

[59] Johann Elbaz, Oleg Lioubashevski, Fuan Wang, Françoise Remacle, Raphael D Levine, and Itamar Willner. Dna computing circuits using libraries of dnazyme subunits. *Nature Nanotechnology*, 5(6):417–422, 2010.

[60] Milan N Stojanovic and Darko Stefanovic. A deoxyribozyme-based molecular automaton. *Nature Biotechnology*, 21(9):1069–1074, 2003.

[61] CristinaáCosta Santini et al. A clocked finite state machine built from dna. *Chemical Communications*, 49(3):237–239, 2013.

[62] Paul WK Rothemund. A dna and restriction enzyme implementation of turing machines. *DNA based computers*, 27:75–119, 1996.

[63] Luca Cardelli. Strand algebras for dna computing. In *DNA computing and molecular programming*, pages 12–24. Springer, 2009.

[64] Luca Cardelli. Two-domain dna strand displacement. *Mathematical Structures in Computer Science*, 23(02):247–271, 2013.

[65] Harold Fellermann, Annunziata Lopiccolo, Jerzy Kozyra, and Natalio Krasnogor. In vitro implementation of a stack data structure based on dna strand displacement. In *International Conference on Unconventional Computation and Natural Computation*, pages 87–98. Springer, 2016.

[66] Barbara Hobom. Surgery of genes-at the doorstep of synthetic biology. *Medizinische Klinik*, 75(24):14–21, 1980.

[67] Steven A Benner and A Michael Sismour. Synthetic biology. *Nature Reviews Genetics*, 6(7):533–543, 2005.

[68] Jingdong Tian, Hui Gong, Nijing Sheng, Xiaochuan Zhou, Erdogan Gulari, Xiaolian Gao, and George Church. Accurate multiplex gene synthesis from programmable dna microchips. *Nature*, 432(7020):1050–1054, 2004.

[69] John E Dueber, Brian J Yeh, Kayam Chak, and Wendell A Lim. Reprogramming control of an allosteric signaling switch through modular recombination. *Science*, 301(5641):1904–1908, 2003.

[70] Ernesto Andrianantoandro, Subhayu Basu, David K Karig, and Ron Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology*, 2(1), 2006.

[71] Khashayar Khoshmanesh, Abbas Z Kouzani, Saeid Nahavandi, Sara Baratchi, and JR Kanwar. At a glance: Cellular biology for engineers. *Computational Biology and Chemistry*, 32(5):315–331, 2008.

[72] Manuel Porcar, Antoine Danchin, Victor de Lorenzo, Vitor A Dos Santos, Natalio Krasnogor, Steen Rasmussen, and Andrés Moya. The ten grand challenges of synthetic life. *Systems and synthetic biology*, 5(1-2):1–9, 2011.

[73] Jacqueline Y Quinn, Robert Sidney Cox III, Aaron Adler, Jacob Beal, Swapnil Bhatia, Yizhi Cai, Joanna Chen, Kevin Clancy, Michal Galdzicki, Nathan J Hillson, et al. Sbol visual: A graphical language for genetic designs. *PLoS Biol*, 13(12):e1002310, 2015.

[74] AlanE H Emery. Recombinant dna technology. *The Lancet*, 318(8260):1406–1409, 1981.

[75] N Trun and J Trempy. Fundamental bacterial genetics, chapter gene expression and regulation, 191–212, 2003.

[76] John MS Bartlett and David Stirling. A short history of the polymerase chain reaction. In *PCR protocols*, pages 3–6. Springer, 2003.

[77] Hamilton O Smith, Clyde A Hutchison, Cynthia Pfannkoch, and J Craig Venter. Generating a synthetic genome by whole genome assembly: $\varphi$x174 bacteriophage from synthetic oligonucleotides. *Proceedings of the National Academy of Sciences*, 100 (26):15440–15445, 2003.

[78] Arturo Casini, Marko Storch, Geoffrey S Baldwin, and Tom Ellis. Bricks and blueprints: methods and standards for dna assembly. *Nature Reviews Molecular Cell Biology*, 16(9):568–576, 2015.

[79] Jennifer A Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with crispr-cas9. *Science*, 346(6213):1258096, 2014.

[80] Thomas Schlake and Juergen Bode. Use of mutated flp recognition target (frt) sites for the exchange of expression cassettes at defined chromosomal loci. *Biochemistry*, 33(43):12746–12751, 1994.

[81] Xu-Dong Zhu and Paul D Sadowski. Cleavage-dependent ligation by the flp recombinase characterization of a mutant flp protein with an alteration in a catalytic amino acid. *Journal of Biological Chemistry*, 270(39):23044–23054, 1995.

[82] Timothy S Ham, Sung K Lee, Jay D Keasling, and Adam P Arkin. Design and construction of a double inversion recombination switch for heritable sequential genetic memory. *PLoS One*, 3(7):e2815, 2008.

[83] Ari E Friedland, Timothy K Lu, Xiao Wang, David Shi, George Church, and James J Collins. Synthetic gene networks that count. *science*, 324(5931):1199–1202, 2009.

[84] Jerome Bonnet, Pakpoom Subsoontorn, and Drew Endy. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proceedings of the National Academy of Sciences*, 109(23):8884–8889, 2012.

[85] Fahim Farzadfard and Timothy K Lu. Genomically encoded analog memory with precise in vivo dna writing in living cell populations. *science*, 346(6211):1256272, 2014.

[86] Yue Chen, Bruce A McClane, Derek J Fisher, Julian I Rood, and Phalguni Gupta. Construction of an alpha toxin gene knockout mutant of clostridium perfringens type a by use of a mobile group ii intron. *Applied and Environmental microbiology*, 71 (11):7542–7547, 2005.

[87] Michael Karberg, Huatao Guo, Jin Zhong, Robert Coon, Jiri Perutka, and Alan M Lambowitz. Group ii introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nature Biotechnology*, 19(12):1162–1167, 2001.

[88] Lori H Conlan, Matthew J Stanger, Kenji Ichiyanagi, and Marlene Belfort. Localization, mobility and fidelity of retrotransposed group ii introns in rrna genes. *Nucleic Acids Research*, 33(16):5262–5270, 2005.

[89] Yue Chen, Lori Caruso, Bruce McClane, Derek Fisher, and Phalguni Gupta. Disruption of a toxin gene by introduction of a foreign gene into the chromosome of clostridium perfringens using targetron-induced mutagenesis. *Plasmid*, 58(2):182–189, 2007.

[90] Jiri Perutka, Wenjun Wang, David Goerlitz, and Alan M Lambowitz. Use of computer-designed group ii introns to disrupt escherichia coli dexh/d-box protein and dna helicase genes. *Journal of Molecular Biology*, 336(2):421–439, 2004.

[91] John T Heap, Oliver J Pennington, Stephen T Cartman, Glen P Carter, and Nigel P Minton. The clostron: a universal gene knock-out system for the genus clostridium. *Journal of Microbiological Methods*, 70(3):452–464, 2007.

[92] Gary M Dunny and Larry L McKay. Group ii introns and expression of conjugative transfer functions in lactic acid bacteria. In *Lactic Acid Bacteria: Genetics, Metabolism and Applications*, pages 77–88. Springer, 1999.

[93] Dagmar Wirth, Leonor Gama-Norton, Pamela Riemer, Upneet Sandhu, Roland Schucht, and Hansjörg Hauser. Road to precision: recombinase-based targeting technologies for genome engineering. *Current opinion in Biotechnology*, 18(5):411–419, 2007.

[94] Gusztav Belteki, Marina Gertsenstein, David W Ow, and Andras Nagy. Site-specific cassette exchange and germline transmission with mouse es cells expressing &phgr; c31 integrase. *Nature Biotechnology*, 21(3):321–324, 2003.

[95] Tolga Çağatay, Marc Turcotte, Michael B Elowitz, Jordi Garcia-Ojalvo, and Gürol M Süel. Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139(3):512–522, 2009.

[96] Jonathan P. K. Doye, Thomas E. Ouldridge, Ard A. Louis, Flavio Romano, Petr Šulc, Christian Matek, Benedict E. K. Snodin, Lorenzo Rovigatti, John S. Schreck, Ryan M. Harrison, and William P. J. Smith. Coarse-graining DNA for simulations of

DNA nanotechnology. *Phys. Chem. Chem. Phys.*, 15(47):20395, 2013. ISSN 1463-9076, 1463-9084. doi: 10.1039/c3cp53545b. URL http://xlink.rsc.org/ÿDOI=c3cp53545b.

[97] Annahita Keravala and Michele P Calos. Site-specific chromosomal integration mediated by $\phi$c31 integrase. In *Chromosomal Mutagenesis*, pages 165–173. Springer, 2008.

[98] John T Heap, Sarah A Kuehne, Muhammad Ehsaan, Stephen T Cartman, Clare M Cooksley, Jamie C Scott, and Nigel P Minton. The clostron: mutagenesis in clostridium refined and streamlined. *Journal of microbiological methods*, 80(1):49–55, 2010.

[99] Herman A De Boer, Lisa J Comstock, and Mark Vasser. The tac promoter: a functional hybrid derived from the trp and lac promoters. *Proceedings of the National Academy of Sciences*, 80(1):21–25, 1983.

[100] David A Mills, Larry L McKay, and Gary M Dunny. Splicing of a group ii intron involved in the conjugative transfer of prs01 in lactococci. *Journal of Bacteriology*, 178(12):3531–3538, 1996.

[101] Annie C Chang and Stanley N Cohen. Construction and characterization of amplifiable multicopy dna cloning vehicles derived from the p15a cryptic miniplasmid. *Journal of bacteriology*, 134(3):1141–1156, 1978.

[102] Sung Kuk Lee, Howard H Chou, Brian F Pfleger, Jack D Newman, Yasuo Yoshikuni, and Jay D Keasling. Directed evolution of arac for improved compatibility of arabinose-and lactose-inducible promoters. *Applied and Environmental microbiology*, 73(18):5711–5715, 2007.

[103] Aaron Lee Garrett. inspyred: Bio-inspired Algorithms in Python, 2016. URL https://pypi.python.org/pypi/inspyred. version 1.0, accessed March 9, 2016.

[104] G. Terrazas, M. Gheorghe, G. Kendall, and N. Krasnogor. Evolving tiles for automated self-assembly design. In *IEEE Congress on Evolutionary Computation, 2007. CEC 2007*, pages 2001–2008, 2007. doi: 10.1109/CEC.2007.4424719.

[105] Peter Siepmann, Christopher P. Martin, Ioan Vancea, Philip J. Moriarty, and Natalio Krasnogor. A Genetic Algorithm Approach to Probing the Evolution of Self-Organized Nanostructured Systems. *Nano Letters*, 7(7):1985–1990, 2007. ISSN 1530-6984, 1530-6992. doi: 10.1021/nl070773m. URL http://pubs.acs.org/doi/abs/10.1021/nl070773m.

[106] Richard A. J. Woolley, Julian Stirling, Adrian Radocea, Natalio Krasnogor, and Philip Moriarty. Automated probe microscopy via evolutionary optimization at the atomic scale. *Applied Physics Letters*, 98(25):253104, 2011. ISSN 00036951. doi: 10.1063/1.3600662. URL http://scitation.aip.org/content/aip/journal/apl/98/25/10.1063/1.3600662.

[107] Andreas R Gruber, Ronny Lorenz, Stephan H Bernhart, Richard Neuböck, and Ivo L Hofacker. The vienna rna websuite. *Nucleic acids research*, 36(suppl 2):W70–W74, 2008.

[108] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, 6(1):26, 2011. ISSN 1748-7188. doi: 10.1186/1748-7188-6-26. URL http://www.almob.org/content/6/1/26.

[109] Joseph Sambrook, Edward F Fritsch, Tom Maniatis, et al. *Molecular cloning*, volume 2. Cold spring harbor laboratory press New York, 1989.

[110] Juan Yguerabide and Evangelina E Yguerabide. Light-scattering submicroscopic particles as highly fluorescent analogs and their use as tracer labels in clinical and biological applications: I. theory. *Analytical biochemistry*, 262(2):137–156, 1998.

[111] B Alberts, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. Molecular biology of the cell 5th ed.(garland science, taylor & francis group). 2008.

[112] New England Biolab. Recombineering Vector pkd46-reca and pkd46-recapa recombineering vectors instruction manual. URL http://www.natx.com/assets/ntc-recombineering-plasmid-brochure-2-0.pdf.

[113] N Kleckner, J Way, M Davis, R Simons, and S Halling. Transposon tn10: genetic organization, regulation, and insertion specificity. In *Federation Proceedings*, volume 41, pages 2649–2652, 1982.

[114] VICKY Chan, LISA F Dreolini, KERRY A Flintoff, SONJA J Lloyd, and ANDREA A Mattenley. The effects of glycerol, glucose, galactose, lactose and glucose with galactose on the induction of $\beta$-galactosidase in escherichia coli. *Journal of Experimental Microbiology and Immunology (JEMI) Vol*, 2:130–137, 2002.

[115] Kyoung-Hee Choi, David DeShazer, and Herbert P Schweizer. mini-tn7 insertion in bacteria with multiple glms-linked atttn7 sites: example burkholderia mallei atcc 23344. *Nature Protocols*, 1(1):162–169, 2006.

[116] Kyoung-Hee Choi and Herbert P Schweizer. mini-tn7 insertion in bacteria with single atttn7 sites: example pseudomonas aeruginosa. *Nature Protocols*, 1(1):153–161, 2006.

[117] RUPAP Simon, U Priefer, and Alfred Pühler. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Nature Biotechnology*, 1(9):784–791, 1983.

[118] Marta Herrero, Victor de Lorenzo, and Kenneth N Timmis. Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. *Journal of Bacteriology*, 172(11):6557–6567, 1990.

[119] Victor de Lorenzo, Lindsay Eltis, Birgit Kessler, and Kenneth N Timmis. Analysis of pseudomonas gene products using laciq/ptrp-lac plasmids and transposons that confer conditional phenotypes. *Gene*, 123(1):17–24, 1993.

[120] Kirill A Datsenko and Barry L Wanner. One-step inactivation of chromosomal genes in escherichia coli k-12 using pcr products. *Proceedings of the National Academy of Sciences*, 97(12):6640–6645, 2000.

[121] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Molecular Systems Biology*, 2(1), 2006.

[122] Suk-Chae Jung, Chris L Smith, Ki-Sung Lee, Min-Eui Hong, Dae-Hyuk Kweon, Gregory Stephanopoulos, and Yong-Su Jin. Restoration of growth phenotypes of escherichia coli dh5$\alpha$ in minimal media through reversal of a point mutation in purb. *Applied and environmental microbiology*, 76(18):6307–6309, 2010.

[123] Shital A Tripathi, Daniel G Olson, D Aaron Argyros, Bethany B Miller, Trisha F Barrett, Daniel M Murphy, Jesse D McCool, Anne K Warner, Vineet B Rajgarhia, Lee R Lynd, et al. Development of pyrf-based genetic system for targeted gene deletion in clostridium thermocellum and creation of a pta mutant. *Applied and environmental microbiology*, 76(19):6591–6599, 2010.

[124] Steve Rozen and Helen Skaletsky. Primer3 on the www for general users and for biologist programmers. In *Bioinformatics methods and protocols*, pages 365–386. Springer, 1999.

[125] Seth G Grant, Joel Jessee, Fredric R Bloom, and Douglas Hanahan. Differential plasmid rescue from transgenic mouse dnas into escherichia coli methylation-restriction mutants. *Proceedings of the National Academy of Sciences*, 87(12):4645–4649, 1990.

[126] ITD Integrated DNA Technologies. Molecular facts and figures. *World Wide Web: http:// www.idtdna.com/pages/docs/educational-resources/molecular-facts-and-figures.pdf*.

[127] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

[128] William P Donovan and Sidney R Kushner. Cloning and physical analysis of the pyrf gene (coding for orotidine-5-phosphate decarboxylase) from escherichia coli k-12. *Gene*, 25(1):39–48, 1983.

[129] Barrell B MA. Artemis: sequence visualization and annotation. *Bioinformatics*, 16 (10):944945Schlueter, 2000.

[130] John R Sadler, Henri Sasmor, and Joan L Betz. A perfectly symmetric lac operator binds the lac repressor very tightly. *Proceedings of the National Academy of Sciences*, 80(22):6785–6789, 1983.

[131] Boris Görke and Jörg Stülke. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology*, 6(8):613–624, 2008.

[132] James W Noah and Alan M Lambowitz. Effects of maturase binding and mg2+ concentration on group ii intron rna folding investigated by uv cross-linking. *Biochemistry*, 42(43):12466–12480, 2003.

[133] Wonbae Lee, Peter H von Hippel, and Andrew H Marcus. Internally labeled cy3/cy5 dna constructs show greatly enhanced photo-stability in single-molecule fret experiments. *Nucleic Acids Research*, 42(9):5967–5977, 2014.

[134] M. Hadorn, E. Bonzli, H. Fellermann, P. Eggenberger Hotz, and M. Hanczyc. Specific and reversible DNA-directed self-assembly of emulsion droplets. *Proc. Nat. Acad. Sci. USA*, 109(47), 2012.

[135] H. Fellermann and L. Cardelli. Programmable chemistry in DNA addressable bioreactors. *R. Soc. Interface*, 2014.

[136] Douglas H Ohlendorf, Dale E Tronrud, and Brian W Matthews. Refined structure of cro repressor protein from bacteriophage $\lambda$ suggests both flexibility and plasticity. *Journal of Molecular Biology*, 280(1):129–136, 1998.

[137] Piro Siuti, John Yazbek, and Timothy K Lu. Synthetic circuits integrating logic and memory in living cells. *Nature Biotechnology*, 31(5):448–452, 2013.

[138] James A Williams. Processes for improved strain engineering, May 22 2008. US Patent App. 12/601,505.