# Blurry Landscapes in Molecular Evolution A high-throughput directed evolution platform

Vincent Balerdi

▶ **To cite this version:**

Vincent Balerdi. Blurry Landscapes in Molecular Evolution A high-throughput directed evolution platform. Biochemistry, Molecular Biology. Université Paris sciences et lettres, 2021. English. NNT : 2021UPSLS101 . tel-04338583

**HAL Id: tel-04338583**
**https://pastel.hal.science/tel-04338583**

Submitted on 12 Dec 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à ESPCI Paris – PSL : UMR-7083 GULLIVER

# Explorer l'influence du bruit phénotypique sur les paysages adaptatifs lors des processus évolutifs

-

# Blurry Landscapes in Molecular Evolution

*A high-throughput directed evolution platform*


Soutenue par
**Vincent BALERDI**
Le 07/12/2021


Composition du jury :

| | |
|---|---|
| Agathe URVOAS-CISSE<br>Maître de conférences, Université Paris-Sud | *Rapportrice* |
| Jean-Christophe GALAS<br>CR, Sorbonne Université | *Rapporteur* |
| Valérie TALY<br>DR, Université de Paris | *Présidente du jury* |
| Yannick RONDELEZ<br>DR, ESPCI | *Directeur de thèse* |

Ecole doctorale n° 563
**Médicaments Toxicologie Chimie Imagerie (MTCI)**

Spécialité
**Biochimie et Biologie Moléculaire**

ESPCI PARIS | PSL

# Table of Contents

# Abstract

Evolution takes place on many different levels, from the macroscopic scale of populations and species, to the microscopic events acting on proteins and molecules in living organisms. Whichever the system, natural evolution always needs diversity to perform selection, in order to promote the survival of the fittest organisms. Moreover, one common characteristic of these multiscale mechanisms is the consistent presence of noise. Indeed, its effects can be seen through molecular stochasticity to population dynamics. Even though background noise was thought to be detrimental to biological systems, an increasing number of studies hypothesised that it had in fact many positive impacts. Notably, regarding the smoothing of fitness landscapes, and the effects on robustness and evolvability depending on the populations' features.

In this PhD project, we developed an experimental platform in order to quantitatively characterize the influence of artificial noise on protein populations. Based on Holliger's Compartmentalised Self Replication, the objects of the study are DNA polymerases, and in particular the KlenTaq polymerase, fundamental molecular biology tools, notoriously challenging to evolve. In this endeavour, libraries of variants of these proteins are generated, and then put through a fully *in vitro* directed evolution process, consisting of cycles of diversification, compartmentalisation of these variants in droplets using microfluidic devices, and selection for fitness using their ability to replicate and amplify their own genetic material. The resulting library could then be sequenced using NGS, and the data interpreted through Statistical Physics inspired analysis or Machine Learning methods. With the aim of adding an additional source of noise in the system, we also studied the effects of aminoglycoside antibiotics on the *in vitro* translation of the proteins, these being known to interfere with ribosomal accuracy. Being able to efficiently tune the extent of error-prone behaviour during protein synthesis would help in our understanding of the effects of background noise on biological processes such as protein translation and evolution.

# Introduction

This manuscript will detail my work towards the development of a high-throughput, fully *in vitro* experimental platform for the directed evolution of proteins - most notably DNA Polymerases - in the presence of translational noise.

Before diving into the experiments that constituted the bulk of my thesis, we will first present the context in which we hope to set this research effort. Starting from the beginning, we will present general knowledge and information on proteins, primordial constituents of living systems, and the central targets of our endeavours. Going through elementary concepts of biochemistry, kinetics and structural biology, an overview of the many scientific fields that study such biomolecules will be detailed.

Next, we will try to establish the argument that, in biology, the complex and remarkable process of evolution is the be-all and end-all of every inquiry, an indispensable framework without which life as we know it and the infinity of its intricate, constitutive mechanisms lack meaning and purpose. By peering into the fascinating history of the thought of evolution through time and civilisations, the resounding importance of these numerous theories will be addressed, notably towards our current global understanding of living systems.

Coming back to one of the main topics of this thesis, we will present the many techniques that were recently developed in order to further study enzymes and proteins, especially through extensive and/or precise control over their structures and functions. Most of these methods rely on the theoretical knowledge amassed through prior centuries, along with the ground-breaking methodological and technological advancements of the 20$^{th}$ century.

Finally, we will circle back to the major and instrumental role of random perturbations - *i.e.* noise - in living systems and the numerous mechanisms used to counter or to make the best use of these stochastic fluctuations. Especially in the context of evolutive processes, we will show that recent works would hint at several advantages and benefits of noise, contrarily to most belief until now.

## A. Proteins are fundamental components of biological systems

In order to fully understand the purpose of this work, we will first need to remind ourselves what proteins are, and why their ever-expanding study is crucial in nowadays challenges, whether they are scientific, medical, or food-industry related.

Proteins are macromolecules synthesised through several steps by living systems, assembled from individual and relatively small molecules - amino acids - into large chains that exhibit very complex tridimensional structures. Some of them have a structural role, allowing living organisms to develop a very wide panel of biological architectures: well described examples are keratin in hair, feathers and the like; collagen in cartilages; or soluble proteins that polymerise to densify otherwise fluid biological materials. Some possess inherent catalytic properties, allowing them to drastically accelerate the rates of specific biochemical reactions. These proteins are called enzymes, and are necessary for the well-being of most organism's metabolisms. They can perform reactions in seconds or minutes, when it would take days or even years without any catalyst[1]. Enzymes are not consumed during the chemical reaction they assist, nor do they count in the equilibrium of compounds of the reaction. Besides being very efficient, they also are highly specialised workers, whether they work alone or in complexes. Indeed, the vast majority of them will mainly recognise one substrate, used in one specific reaction. However, enzymes sometimes also exhibit a promiscuous activity, a fortuitous side reaction which is much slower and less efficient than their native activity. More often than not, this promiscuity is negligible in the context of successions of biochemical reactions.

These very specialised constructs are expressed from the genetic information encoded in the cells' nuclei. Two slightly different forms of nucleic acids are present inside living cells, the DNA, (DeoxyriboNucleic Acid), and the RNA (RiboNucleic Acid). The only chemical difference between the monomers is the presence of an -OH moiety on the 2'-carbon of the ribose motive for the RNA, contrarily to the DNA (Fig. A.1). The corresponding polymers form a backbone of these (deoxy)ribose molecules, linked to one another via phosphodiester bonds, each motive carrying one of the 4 nucleotides, cytosine (C), guanine (G), adenine (A) and thymine (T), or uracil (U) for the RNA. Theses bases are then coupled to each other via hydrogen bonds: A with T (U for RNA) and C with G. The well-known double-helix structure of the DNA was inferred by Francis Crick and James Watson at the University of Cambridge in 1953, using an X-ray diffraction image obtained by Raymond Gosling and Rosalind Franklin at King's College London a year before. They proposed that two complementary anti-parallel

strands were coiling around each other, defining an asymmetrical pattern with two grooves, one wider than the other (22Å and 12Å respectively).

The supposedly small difference in structure between DNA and RNA actually heavily differentiates the properties of these two biomolecules: while DNA is a very stable polymer that gets replicated during each cell cycle and transmitted between generations, RNAs are easily degraded and mainly act as temporary intermediaries between the precious information in the nucleus and the rest of the organism. Most notably, messenger RNAs (mRNAs), transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) are used for the biosynthesis of proteins in the cytoplasm, while other types (miRNAs, non-coding RNAs, etc) are deeply necessary for proper gene expression and regulation.

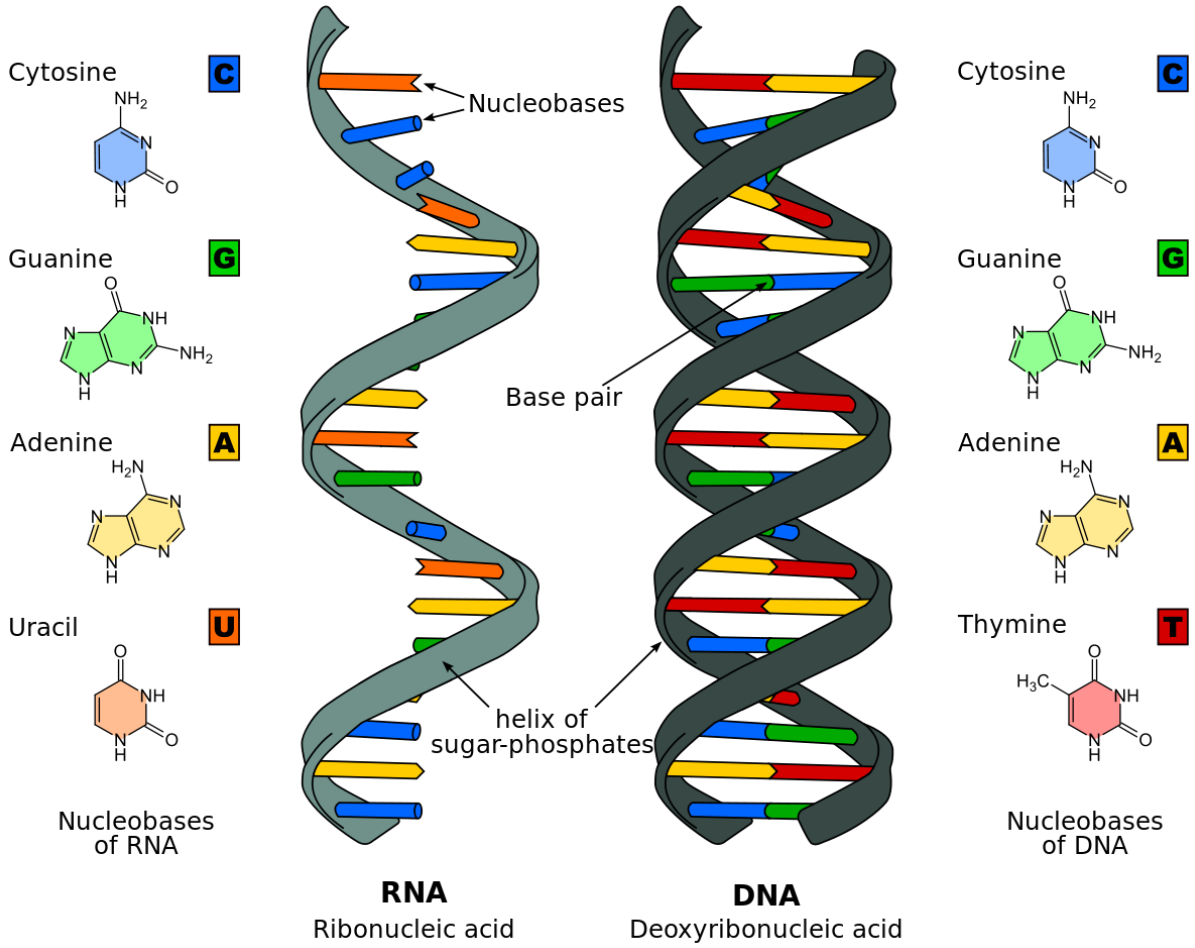



Figure A.1: Differences between DNA and RNA molecules.
The two main differences between those molecules are the building blocks (ribose for RNA, deoxyribose for DNA), and the change of the Thymine (T) nucleobase in Uracil (U) by loss of a -$CH_3$ moiety. While DNA is composed of two polynucleotidic chains coiling around each other, RNA is almost always found as a single strand in nature, and can adopt very intricate and complex 3D structures by folding onto itself. Adapted from Wikipedia.

A specific sequence of nucleotides encoding information (either to be transcripted into RNA, or translated into proteins) is called a gene, and genes are what makes up a large portion of the individual print of every organism. Notable examples in humans include eye, skin and hair colour, along less visible characters such as blood types or higher-risks for specific diseases. However, even though all humans share overall the same phenotypic features, there is still lots of room for diversity. Variants in these common genes (eye or hair colour for example) are called alleles, that reflect the small differences that still exist between every individual of a same population, and even species. As we will see later on, this diversity is paramour for the survival of a species in nature.

## 1. Protein Biosynthesis

As a simple example, we will consider the expression of a gene coding for a protein. The whole process of protein synthesis can be divided into two main phases: the transcription (DNA to mRNA) and translation (mRNA to protein). The transcription takes places in the nucleus, where an enzyme known as helicase unwinds the double-stranded DNA by disrupting the hydrogen bonds between nucleotide pairs (A-T / C-G). Another protein, the RNA Polymerase, binds to one of the two exposed single-strands to read this template from the 3' end to the 5' end. It then synthesises a complementary RNA version of it with RNA nucleotides present in the nucleus, according to base pairing (the thymine being replaced by uracil). The single-stranded mRNA polymer, growing more and more, will eventually break free from the DNA template, being too large to efficiently bind to the DNA via hydrogen bonds. It will then be processed in the nucleus for maturation (post-transcriptional modifications, splicing, etc), before exiting to the cytoplasm for the translation process.

The mRNAs, upon arriving in the extra nuclei environment, get bound to a protein called the Ribosome. This construct will recognise the 5' end of the mRNA, and will begin to move along the strand until it finds a start codon, an AUG sequence, coding for a Methionine amino-acid. The Large Ribosome Subunit will then bind to the mRNA, in order to translate its sequence into a polypeptide, according to the genetic code (Fig. A.1.1), each triplet of nucleobases corresponding to one of the 20 amino-acids or a special codon (Start/Stop). We will now present how the ribosomes are actually synthesising the polypeptides.

The protein is basically obtained through the successive addition of one amino-acid to the other, based on the sequence of the mRNA. These amino-acids are not free-floating in the media, but first bound to another type of RNAs, the transfer RNAs (tRNAs), which serve as the intermediate between the mRNA and the forming protein. These oligonucleotides present a very peculiar structure:

- The primary structure of the RNA is the sequence of ribonucleotides of the mRNA, carrying the genetic information of the transcripted DNA gene.

- The secondary structure reflects the pattern of hydrogen bonds between the bases, determining the three-dimensional form of the biomolecule. The tRNAs are characterised by their cloverleaf shape, exhibiting a crucial unit of three nucleotides known as the anticodon. Each tRNA contains a specific anticodon, coding for an amino-acid. This sequence binds its complementary on the mRNA, allowing a very specific and strong molecular recognition between the two counterparts in the ribosome.
  Looking at the genetic code, it is to be noted that it is redundant, meaning that several amino-acids are coded by several anticodons, such as Leucine, that can be loaded by the ribosomes with tRNAs associated to the codons UUA, UUG, CUU, CUC, CUA and CUG.

- The tertiary structure is the final form of the tRNA's maturation, its spatial shape heavily determined by its primary sequence, and secondary structure to a lesser extent. The "cloverleafs" are actually three-dimensional loops that are recognised by the large subunit of the ribosome.

On the 3' end of the tRNA, depending on its anticodon sequence, the corresponding amino-acid is covalently bound by an amino-acyl tRNA synthetase. Each amino-acid presents its own enzyme to catalyse its covalent binding to one of its cognate tRNA.

Now, onto the actual mechanism of protein synthesis. As mentioned previously, the ribosome is the macromolecular machine that will translate the mRNAs into functional polypeptides. They are composed of two subunits, one larger than the other, in both Prokaryotes and Eukaryotes. These subunits are themselves the reunion of several smaller units, ribosomal RNAs (rRNAs) and other proteins. This makes the full ribosome a very complex and organised worker, necessary for all living organisms.

Considering our work with Prokaryotes, and most notably *E. coli*, we will focus on bacterial ribosomes. As shown in many structural and mechanistic studies, these are the assembly of the 30S and 50S subunits[2,3]. Around 20nm in diameter, they are mostly composed of rRNAs, using ribosomal proteins for their property to scaffold the structure of the whole machinery. The "S" of these constructs corresponds to the Svedberg unit, indicator of their rate of sedimentation when centrifuged.

| First letter | U | C | A | G | Third letter |
|---|---|---|---|---|---|
| U | UUU Phe<br>UUC<br>UUA Leu<br>UUG | UCU<br>UCC<br>UCA<br>UCG Ser | UAU Tyr<br>UAC<br>UAA STOP<br>UAG STOP | UGU Cys<br>UGC<br>UGA STOP<br>UGG Trp | U<br>C<br>A<br>G |
| C | CUU<br>CUC<br>CUA<br>CUG Leu | CCU<br>CCC<br>CCA<br>CCG Pro | CAU His<br>CAC<br>CAA<br>CAG Gln | CGU<br>CGC<br>CGA<br>CGG Arg | U<br>C<br>A<br>G |
| A | AUU<br>AUC<br>AUA Ile<br>AUG Met | ACU<br>ACC<br>ACA<br>ACG Thr | AAU Asn<br>AAC<br>AAA Lys<br>AAG | AGU Ser<br>AGC<br>AGA Arg<br>AGG | U<br>C<br>A<br>G |
| G | GUU<br>GUC<br>GUA<br>GUG Val | GCU<br>GCC<br>GCA<br>GCG Ala | GAU Asp<br>GAC<br>GAA Glu<br>GAG | GGU<br>GGC<br>GGA<br>GGG Gly | U<br>C<br>A<br>G |

Figure A.1.1: Genetic code table.
This set of rules is used by living organisms in order to translate genetic information encoded within their DNA (or mRNA) into proteins. Three-nucleotide codons define which amino acid will be inserted during protein synthesis by the ribosomes. Each of these proteinogenic residues are loaded onto specific tRNAs, recognised by the ribosomes.

Once assembled, the ribosome has three sites for tRNA binding. These are known as the aminoacyl site (A), the peptidyl site (P) and the exit site (E) (Fig. A.1.2). Let us consider the very first tRNA that binds to the mRNA once the ribosome has formed around it.

This newly amino-acylated tRNA (Met) will begin by binding in the A site, if its anticodon is complementary to the exposed AUG codon of the mRNA. It will then switch to the P site, where the ribosome will first link the Methionine amino-acid to a newly entering tRNA (Leu) in the A site. Upon moving the first tRNA to the E site and the second to the P site, the Methionine will be transferred to the second tRNA, forming the start of the polypeptide. Once it's done, the "naked" tRNA will be processed to the exit site, and finally released in the media. Additional amino-acids will get successively added to this growing chain, until one of the stop codons (UAA, UAG and UGA) is read by the ribosome. At this point, no tRNA can recognise and bind to these, but proteins called release factors can. They trigger the hydrolysis of the peptidyl-tRNA, freeing the newly formed protein from the ribosome. The machinery is subsequently recycled: the subunits are separated, releasing the mRNA in the media.

One could think that once the protein is synthesised, it is ready to perform its role in the organism, either catalysis of specific chemical reactions or something else. But it is the start of a tremendously complex process for the polypeptide: its folding into its final three-dimensional structure. This event actually begins during the translation of the protein, as the strand of amino-acids is created, which then folds onto itself step by step. Quite similarly to the DNA and RNA, three degrees of structure can be described for proteins:

- The primary structure: an elongated sequence of covalently bound amino-acids.
- The secondary structure: first folds and coils of the polypeptide through the hydrogen bonds between atoms of the backbone. The most common and repeated structures in protein structures are known as alpha helixes and beta sheets[4].
- The tertiary structure: final and most stable form for most proteins.

Some proteins also operate in multimeric complexes, either with other copies of themselves or different proteins entirely, adding an additional level of structural complexity, known as the quaternary structure of the protein[5,6].

However, in a similar fashion as the mRNAs that need to mature in order to be functional, properly folded proteins also often undergo several stages of post-translational modifications. Their function can vary widely, from enhancements of interaction with substrates or proteins, to alterations of their structures and catalytic activity.

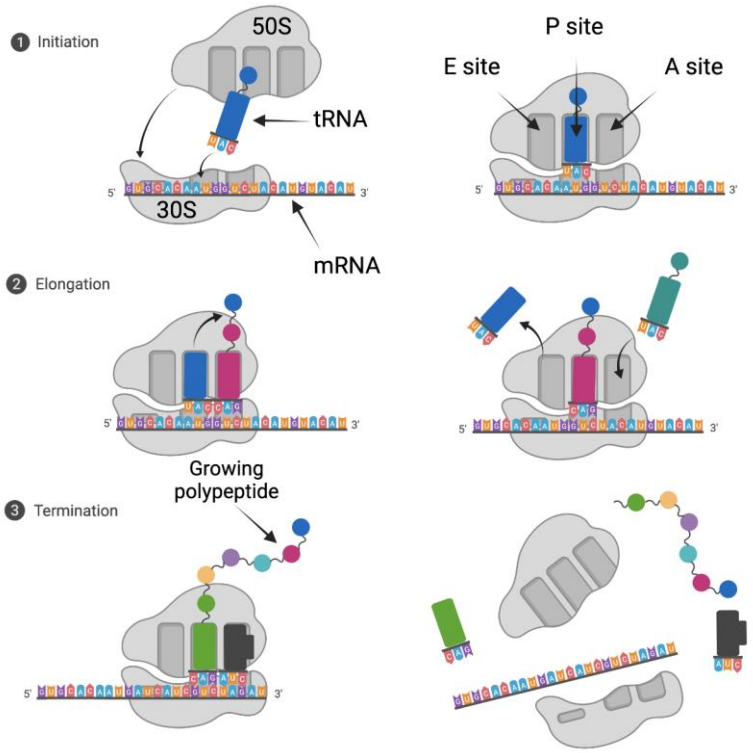



Figure A.1.2: Schematic of ribosome structure and translation process.
First, the two ribosomal subunits (50S and 30S) are assembled with the mRNA to be translated. The first tRNA is always charged with a *N*-Formylmethionine amino acid, on a UAC codon, also known as start codon, and enters at the P site in the ribosome. During the elongation step, successive tRNAs present in the media are randomly inserted in the A site, and if their codon matches the mRNA sequence, the previous amino acid is linked to the one present on the new tRNA. The uncharged tRNA is then discarded through the E site. One by one, amino acids are loaded onto a growing polypeptide chain, until a stop codon is reached on the mRNA (UAA, UGA or UAG). The machinery is then disassembled, and the translated protein released in the media along its cognate mRNA.

## 2. Enzyme Kinetics

As we mentioned before, enzyme are special proteins that exhibit powerful catalytic properties towards a wide panel of chemical reactions. These molecular machines manipulate substrates (often one, but sometimes more) that bind to their active site and, through several steps, get eventually transformed into the chemical products of this reaction. Depending on the enzyme, its substrate(s) and its mechanism of action, kinetic studies of the process can be assessed in order to get a better understanding of how the protein actually works. Historically, the first kinetic models for catalysis revolved around two specific properties of enzymes:

- The saturation parameter. Chemical reactions catalysed by enzymes, contrarily to uncatalysed reactions, display saturation kinetics. A given enzyme always becomes saturated with its particular substrate, if this substrate concentration is too high in the media.

- The intrinsic rate constant. Some enzymes are naturally faster than others, either when catalysing the same reaction, or in absolute terms of turn-over rate, which corresponds to the number of substrate molecules turned into products per catalytic site and per second.

In the beginning of the 20$^{th}$ century, the German biochemist Leonor Michaelis and the Canadian physician Maud Menten proposed a mathematical model to illustrate the kinetics of single-substrate reactions[7] (Fig. A.2.1). This model relies on having an enzyme E with its substrate S forming a complex ES (transition state), that can either dissociate into the enzyme and the product of the reaction P, or revert to the initial composition. Kinetic rates are associated with every step of the process, respectively $k_f$, $k_{cat}$ and $k_r$. The aim of the model is to calculate the rate of reaction $v_0$, defined as the rate of product formation of the enzyme.

$$\mathrm{E} + \mathrm{S} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_{\mathrm{cat}}} \mathrm{E} + \mathrm{P}$$

Figure A.2.1: Chemical equation for the Michaelis-Menten kinetic model. E represents the enzyme and S its unique substrate. Depending on the rates $k_f$, $k_{cat}$ and $k_r$, the complex ES is formed, which leads to the eventual release of the product P and the enzyme.

Under specific assumptions, a system of differential equations can be determined from this chemical equation (Fig. A.2.2). This model considers that [ES] changes very slowly compared to the concentrations of the other species, and that the total concentration of enzymes does not change over time. These conditions are very often true in biological context, and several key parameters of the kinetic model can thus be determined, along with the actual formula for the rate of reaction ν. Namely, $V_{max}$, which is the maximal rate of the enzyme (in $M.s^{-1}$); $K_M$, the Michaelis-Menten constant, concentration (in M) at which the rate of reaction is half of $V_{max}$.

$$v = \frac{\mathrm{d[P]}}{\mathrm{d}t} = V_{\max}\frac{[\mathrm{S}]}{K_\mathrm{M} + [\mathrm{S}]} = k_{\mathrm{cat}}[\mathrm{E}]_0\frac{[\mathrm{S}]}{K_\mathrm{M} + [\mathrm{S}]}$$

$$v_0 \approx \frac{k_{cat}}{K_M}[\mathrm{E}][\mathrm{S}] \qquad K_\mathrm{M} = \frac{k_r + k_{\mathrm{cat}}}{k_f} \qquad V_{\max} = k_{\mathrm{cat}}[\mathrm{E}]_0$$

Figure A.2.2: Resolution of the differential equations given by the Michaelis-Menten kinetic model. ν represents the rate of product P formation, while $K_M$ (in M) is the Michaelis-Menten constant, and $V_{max}$ (in $M.s^{-1}$) the maximal rate of the enzyme.

In order to link this to the previously mentioned foremost important parameters for enzymes, we can see that on one hand, $K_M$ is an indicator of the saturation parameter. The bigger the $K_M$, the more substrate the enzyme needs to attain its maximal/saturating rate. Another way to see $K_M$ would be as the affinity of the enzyme for its substrate: indeed, in most cases, $k_{cat} << k_r, k_f$ (formation/dissociation of ES much faster than product formation). $K_M$ is then the simple ratio of $k_r$ on $k_f$, otherwise known as the dissociation constant of the enzyme. On the other hand, $k_{cat}$ is obviously the turn-over rate of the enzyme. The bigger $k_{cat}$ is, the faster the product P is produced.

Under specific conditions, *i.e.* [S] << $K_M$ (which would be the case for reactions at the initial state), the rate of reaction $v_0$, the initial rate of product formation, can be simplified. This shows a linear relation between $v_0$ and [S], indicating a first-order kinetics situation. However, when [S] >> $K_M$, the reaction is independent of [S] and asymptotically tends to its $V_{max}$ value. All taken together, these equations allow us to plot the v vs [S] graph, as well as the evolution of each of the species' concentrations over time (Fig. A.2.3).

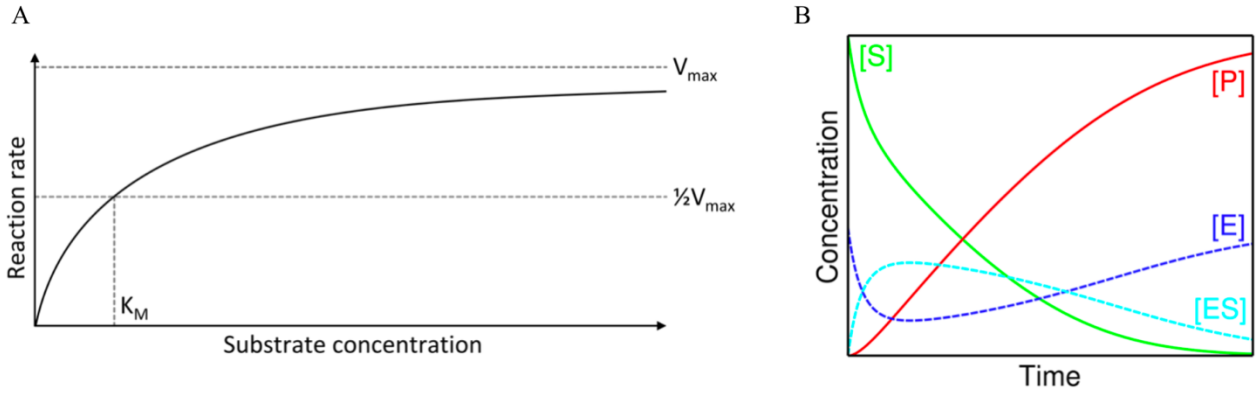



Figure A.2.3: Michaelis-Menten model. A) Graph Plot of the v reaction rate of the enzyme E versus the substrate [S] concentration. While [S] << $K_M$, a linear dependency links the two coordinates. When [S] >> $K_M$, v asymptotically tends to the $V_{max}$ value. B) Graph of the different species' concentrations versus time.

Although the Michaelis-Menten model only works for one-substrate enzyme kinetics, it is a common first approximation for modelling most biochemical reactions, and it can be further complexified to take into account additional steps between the initial state and the product formation, or more complex situations.

## 3. Structure-Activity Relationship

The first conceptualisations of the relationship between a molecule's structure and its biological activity date back to 1865 with the theory of Crum-Brown and Fraser[8], while Emile Fischer proposed his "lock and key" model in 1894 to explain the high specificity observed in enzymatic reactions[9]. But it will not be until the first structural studies of proteins in the mid-1950s using X-Ray crystallography[10] that it became increasingly reasonable to assume that the three-dimensional structure of a protein was linked to its catalytic activity.

Indeed, a solubilised protein can actually shift through several preferred conformations, contradicting a more rigid model of the molecular machines. Since enzymes can be surprisingly quite flexible, local reconfigurations upon substrate binding are possible, in order to further increase enzymatic specificity, as proposed Koshland in 1958 with his "induced fit" model[11]. Upon binding of its substrate(s), the enzyme would slightly change the conformation of its active site, until both are close enough for the biochemical reaction to happen. It is worth to note that this effectively explains the stabilisation of the transition state observed in substrate-enzyme complexes, contrarily to the "lock and key" model.

Therefore, the three-dimensional structure of the catalytic pocket of the enzyme is decisive towards the binding and transformation of the substrate. As we mentioned before, the tertiary structure of a protein is highly predetermined by its primary sequence, *i.e.* the chain of amino-acids, that folds onto itself after translation, leading to its complex spatial shape[12]. While the nature of some residues may not strongly impact the properties, those located around the active site in the 3D structure are in many cases crucial to secure the catalytic activity of the enzyme. Mutations in the corresponding gene, or errors inserted during the transcription/translation processes, may therefore have direct consequences on the enzyme, but also indirect consequences on its biochemical environment. For example, mutations on the surface residues can drastically change the interactions of the protein with its surroundings.

This also explains that the physical, unfolding of proteins often removes catalytic activity from enzymes. Indeed, when exposed to heat or denaturing chemicals, the relatively weak non-covalent interactions that bind the structure together (electrostatic interactions, van der Waals forces, etc) are disrupted, leading to a mainly unfolded protein, which loses (either partially or in totality) its biological activity.

On the other hand, some organisms are known to inhabit quite harsh environments, such as volcanic hot springs[13] or deep-sea trenches[14]. Proteins found in those usually simple beings actually show fascinating properties, from thermophilia to high-pressure resistance. Solving their structures allowed biologists to better understand the molecular determinants of these phenotypes. A very good example is the *Taq* Polymerase, extracted from the thermophilic microorganism *Thermus aquaticus* by Chien et al. in 1976[15]. This heat-resistant DNA polymerase is nowadays a fundamental molecular biology tool.

## 4. On the methods and importance of studying proteins

### 1) Structural Biology

The first works on enzymes date back to the 19th century, but it will not be until the 20th century that the field of enzymology really took off with the numerous methods that were developed by biochemists. These are extensively presented in the *Methods in Enzymology* Volume 1, published in 1955 by Elsevier, but such practices most notably include : the extraction of enzymes from animal tissues[16], plants[17] or microorganisms[18]; purification of functional enzymes from the surrounding biological medium[19–21] and the enrichment of bacterial cultures in specific enzymes[22].

As previously mentioned, structural biology rose in the middle of the 1950's as a cornerstone way to study proteins. Being able to visualise the proteins' structures with or without their substrates really made possible a leap forward in our understanding of the relationship between biological activity and chemical structure of these macromolecules. The most widely used technique to this end was X-ray crystallography, analysing the pattern of diffraction of the electromagnetic radiation on the 3D crystal of the purified protein to determine its structure (Fig. A.4.1). Numerous structures were solved using this technique, with sometimes sub-angstrometric resolution[23]. Yet, this approach necessitates a lot of time, trial and error to yield good results. Moreover, the requirements to purify and crystallise can be quite a burden, especially for flexible targets such as membrane proteins.

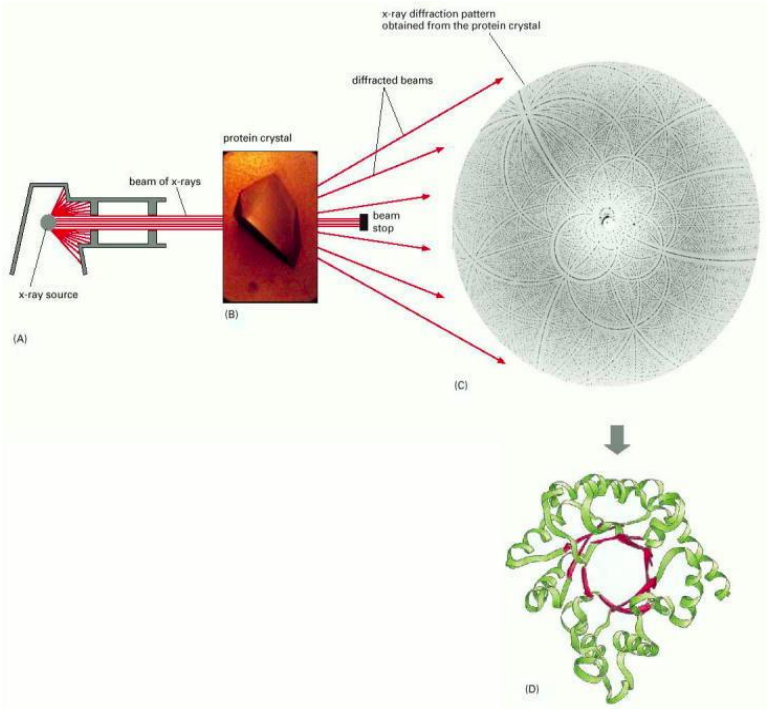



Figure A.4.1: Principle of the X-ray crystallography method.
(A) A beam of x-rays is sent towards a well-ordered protein crystal (B). A fraction of the beam is scattered through the protein crystal by the atoms of the lattice. Diffracted beams appear as a pattern of spots. Here is a figure obtained from diffraction from a RuBisCO (ribulose bisphosphate carboxylase) crystal. (C). The diffraction pattern, through complex calculation, can be used to infer the protein structure. The protein sequence is a necessary information towards this endeavour. (D). Although the exact and complete 3D structure of the protein is hard to recompose, general structural features such as alpha helixes and beta sheets can be associated with the structure through its diffraction pattern.
From *Molecular Biology of the Cell. 4th edition*, Alberts B, Johnson A, Lewis J, et al. New York: Garland Science; 2002.

However, the technological advances of the early 21$^{th}$ century allowed scientists to dive deeper into the structures. Another technique then became increasingly popular: Nuclear Magnetic Resonance (NMR) spectroscopy. While it has commonly been used until the end of the 19$^{th}$ century for the structural analysis of small molecules, several teams of researchers developed new powerful methods in order to solve the structure of small proteins[24].

Given the nanometric accuracy of the NMR spectrometers, probing different nuclei's interactions between themselves in the protein allowed them to map very precisely the structure of the sample, with the help of the powerful computational tools that were emerging at the time. Thus, bigger and bigger protein structures were successfully solved[25]. Because the technique is performed in solution, monitoring changes in the structure is also possible, giving a more dynamic picture than in x-ray crystallography. Indeed, by measuring accurately the relaxation times of nuclei after pulse-sequences of magnetisation, it is possible to visualise the motions of the proteins through time (time-scale ranging from about 10 ps to 10 ns).

The last method used for the gathering of structural information we will present is the Cryogenic Electron Microscopy (CryoEM). Similarly to the regular Electron Microscopy, the images are obtained with an electron beam that is bent and scattered when passing through the samples. The diffraction pattern is then computed to reconstruct the observed 3D structure (Fig. A.4.2). However, due to the damages generated by high-energy radiation beams on the targets, the idea was then to set them in a cryogenised water matrix in order to preserve the sample. Even though the technique dates back to the 1970s[26], recent technological advances for detectors and processing softwares improved the resolution of the images to near-atomic resolution[27]. These promising developments for the imaging of biomolecules even rewarded the creators of the CryoEM with a Nobel Prize in Chemistry in 2017. Just like NMR, this technique is favoured as an alternative to x-ray crystallography, considering the sample is simply vitrified in water and does not need to be crystallised.

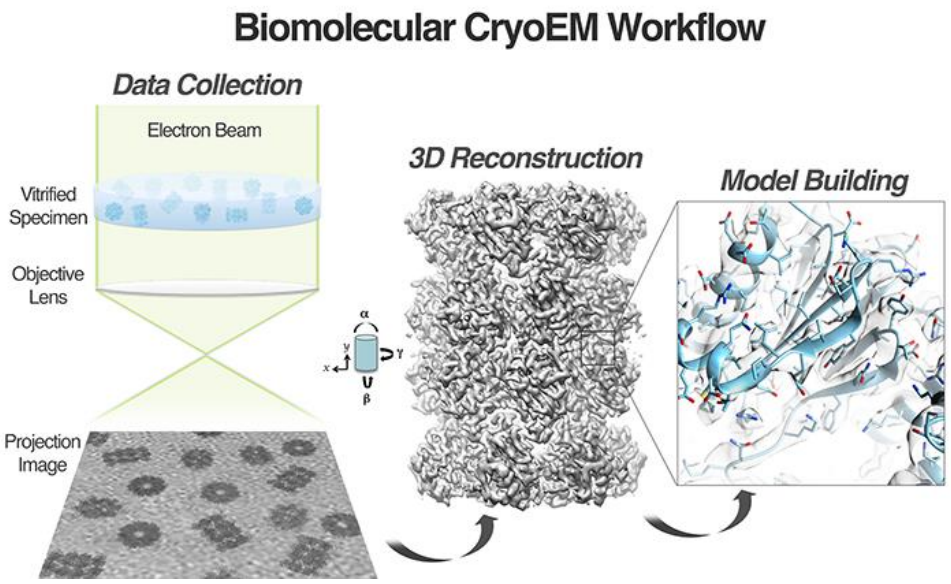



Figure A.4.2: Cryogenic Electron Microscopy Workflow.

First, the biological samples are cryogenised in a pure, thin layer of water. Once the diffraction pattern of the Electron beam is gathered, the 3D structure of the sample is reconstructed using computational softwares, and the structural model is generated from it.
Adapted from Reichow Lab Website, Portland State University.

## 2) Computational Biology

The many improvements in modelling tools that were developed at the start of the 21$^{th}$ century led the way to the increasing usage of computational methods for the prediction of protein and biomolecules structures. With the help of new, more accurate and more affordable sequencing techniques, the number of protein structures solved experimentally quickly paled before to the number of available peptide sequences on either of the PDB or UniProt platforms[28] (Fig. A.4.3). Indeed, the experimental structural tools such as x-ray crystallography or NMR spectroscopy heavily depend on very time-consuming and empiric processes, lagging behind the easier and easier sequencing of protein's encoding genes. Unfortunately, the procedures that allow to infer protein structures from their sequences need a lot of computing power to be efficient: the aim being to model the possible spatial conformations of the macromolecule, calculation of the target free energy and finding the global minimum of this energy are the two most resource-consuming steps of the process.

In order to get a better understanding of what this means, let us consider an unknown, relatively small 50-amino acids long protein sequence. With 49 peptide bonds, the number of possible spatial conformations is ~ $3^{100}$ due to the two radial angles between each peptide. The objective would then be to compute this almost infinite number of 3D structures for this sequence, in order to fully understand how it influences the spatial configuration and biological activity of the protein. Of course, even using already incredibly powerful computers, such a task is not feasible. Two methods thus rose to the challenge, looking for ways to simplify this computation problem[29] : homology and *de novo* modelling, which will be presented in further detail later in the introduction.

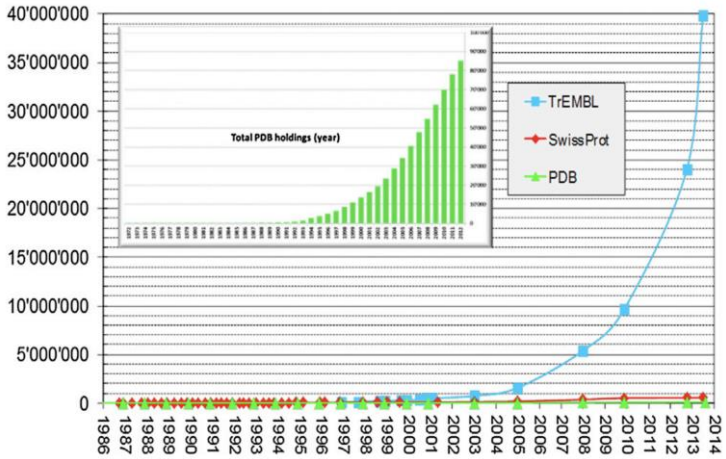



Figure A.4.3: Sequence/Structure Gap.
Number of entries in sequences in TrEMBL, and in structures in SwissProt or the PDB. TrEMBL and SwissProt are part of the UniProt Consortium. We can see the number of protein sequences exponentially increasing from the start of the 2000s, whereas the number of protein structures is still very low more than 10 years later. Inset: growth of PDB holdings from 1972 to 2013. Figure is reproduced from the review of T. Schwede[28].

## B. "Nothing in biology makes sense except in the light of evolution"

We will first present a brief history of the several theories that were developed in order to explain why and how living organisms manage to change their observable characteristics through time and space.

Even though some ideas were conceptualised during ancient times by Greek philosophers such as Plato and his student Aristotle, these were intrinsically linked to some divine purpose that would explain the perfection of nature's diverse creations[30]. The Middle-Ages and the prevalence of Christianity in Europe did not help in nurturing these concepts, contrarily to the philosophic and scientific expansion of knowledge that took place in the Middle-East between the 8th and the 14th century[31]. However, starting from the 17th century, the Renaissance would bring back all this forgotten knowledge to the Western world, and with the beginnings of modern science as we know it today, the theological approaches to evolutionary concepts gradually lose weight against more grounded theories[32,33].

But it will not be until the early 19th century that the truly first evolutionary theory will be proposed by Jean-Baptiste Lamarck in his Philosophie Zoologique (1809): *transformism* refers to the idea that living organisms can change and complexify through time by using their innate life essence, depending on the natural conditions they are exposed to. Use or disuse of organs or limbs would then determine if it is transmitted or not to the offspring, allowing a slow but steady adaptation to the environment. Although his theory mistakenly does not consider that a common ancestor of all living things is plausible, it is the first rational approach that tries to explain how species are linked to one another, and how they can change between generations (Fig. B.1).

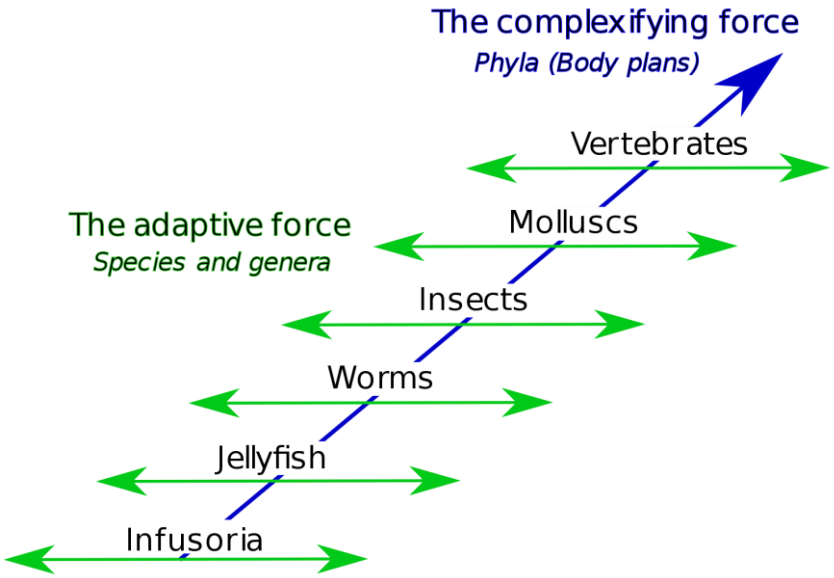



Figure B.1: Lamarck's transformism.
An "innate life force" would drive species to complexify, here from microorganisms and invertebrates to greater life forms such as vertebrates. These species would then slowly adapt to their environment based on the use or disuse of their organs or characteristics.

## 1. Theory of Natural Selection

Nearly 50 years later, Charles Darwin and Alfred Russel Wallace presented in 1858 their separately developed but very similar theories towards evolution[34]: they proposed that every living organism was continuously guided through evolutionary paths depending on the conditions imposed by its environment. This process would be coined as "natural selection", forfeiting the role of an intrinsic life essence in the development of species in nature. To illustrate his theory, Darwin famously described in his works the adaptation of finches' populations among the many islands of the Galapagos Archipelago (Fig. B.1.1).

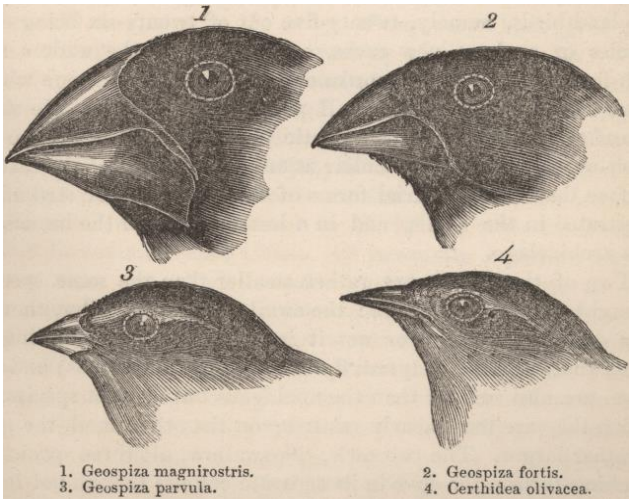



Figure B.1.1: Darwin's finches or Galapagos finches.
Darwin documented the shape and size of the finches' beaks, in accordance to their different food regimes.
From Darwin's Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N.
Voyage of the Beagle, 1845, 2d edition.

These are somewhat similar species, but every island presented a different take on how to best survive in this harsh environment. Most notably, the size of the birds' beak was indicative of a specific food regime, a longer beak in order to tear holes through cacti to eat the pulp, or a shorter one to rip the base of the cacti, and eat insects from the ground. The wide spectrum of species exhibiting strategies between those two phenotypes was a sign of the slow but steady change happening in those finches' populations among the Archipelago.

The reproductive success of a plant or animal indeed relies on several universal parameters, whether that would be access to food, potential mates, or safety from eventual predators. Among the near infinite variations occurring seemingly at random in living organisms, beneficial changes in individuals would mean a better chance for survival, and thus for reproduction. By considering that such changes would then be perpetuated to the offspring (which is the only strong assumption of the theory), this would lead to the establishment of the numerous overall well-adapted populations of species encountered in nature, and the death and extinction of less adapted individuals and/or populations. Darwin would often compare this concept of natural selection to the artificial breeding techniques ancestrally used by humans: just as dog breeders would select or remove traits from a population by breeding the most adapted individuals (colour of the pelt, size, etc), nature would perform the very same process on every living species on Earth, selecting over time the biosphere as we know it.

However, these ideas were still very disruptive at the time, especially towards the deep ethnocentric vision of the world in Western thought, and their popularity in the scientific community stagnated until the rediscovery of Gregory's Mendel works at the beginning of the 20th century[35,36]. Indeed, the most common critics of Darwin and Russel's theory were not about the concept itself, but about the actual mechanism: it was unable to properly explain how nature could achieve such near-infinite variation in life, and how heritable characteristics were passed through generations in populations.

The studies of the Moravian monk were crucial towards our global understanding of heredity and genetics. During almost a decade, tens of thousands of pea plants (*Pisum sativum*) were carefully hybridised in order to determine how individual characters (colours and forms of seeds, flowers, etc) would be transmitted to their offspring. These experiments gave rise to a set of three laws known as the "Mendelian Principles", that dictates how phenotypic trait inheritance takes place (Fig. B.1.2):

- Law of dominance and uniformity: The allele which masks the other is referred to as dominant, while the allele that is masked is referred to as recessive.
- Law of segregation: Organisms pass a randomly selected allele for a trait to their offspring, such that the offspring receives one allele from each parent.
- Law of independent assortment: Genes of different traits can segregate independently during the formation of gametes.

This tremendous work eventually rediscovered, it would jumpstart the development of modern genetics in the early 20th century, and with it the change of perspective from phenotypic characterisation to a more genome-oriented reflection on biological processes. Based on the near-infinite number of variations exhibited in nature, the most useful traits are beneficial for its organism, promoting a better survival and reproduction. These traits would perhaps be inherited through Mendel's principles by the offspring, which would then itself be subjected to the same process.

It is interesting to note that this theory is the fruit of centuries of research, documentation, and study from dozens of naturalists and scientists across the world. Several changes in Western thought were necessary, quite often in domains decorrelated from actual sciences. As can be seen nowadays, these evolutionary biology concepts and mechanisms are now extended to numerous other fields such as economy, sociology, or computer sciences.

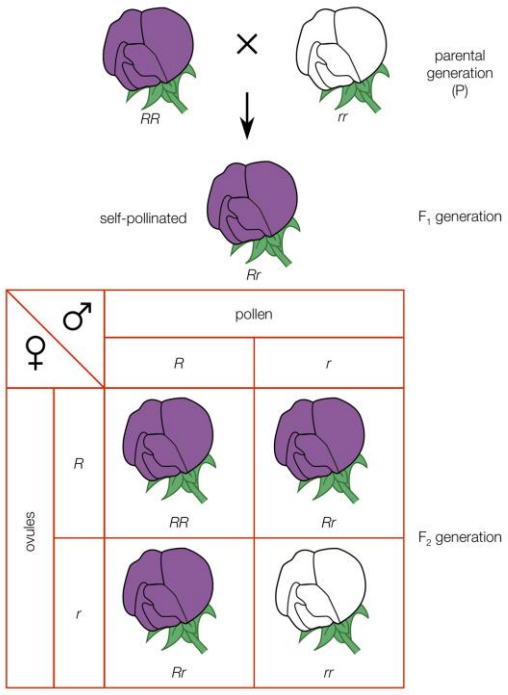



Figure B.1.2: Mendel Principles for heredity.
R is presented as the dominant allele, while r is recessive. Only "rr individuals" of the flower population would display the white petals phenotype, while all others are purple. If the parental generation is of RR and rr genotypes, then their gametes will be of R and r, respectively. Therefore, the first generation is necessarily Rr, and purple, by mixing of these sexual cells. In the second generation however, when the Rr genotype individuals reproduce, one would observe a quarter of each phenotype described in the adjacent table, only one of them being of the white petals, due to the recessive nature of its allele compared to that of the purple one. From Encyclopaedia Britannica, 2013.

## 2. Quantitative Biology

As we mentioned previously, Mendel's tremendous experiments with pea plants already constituted an extensive intent to study the mechanisms of heredity from a genotypic and quantitative standpoint, contrarily to the more common phenotypic and empiric approach adopted by biologists at the time. The exact mechanism upon which evolution was acting upon genetic variation was unknown, and biologists were thus divided mainly between considering that evolution operates either by sudden, large mutational leaps (saltationism), or by smaller cumulative changes over time (gradualism). It was still doubtful that such minute and random variations could fully explain the appearance of inherently different species, albeit small scale changes could be responsible for finer adaptations in already established populations.

However, in order to fully reunite Mendel's principles with Darwin and Russel's natural selection, there was still to determine how discrete hereditary units such as genes could manifest into the continuous range of phenotypic variations that was observed in nature. Indeed, quantitative traits such as height or weight seemed to be almost completely hereditary at the time, but could not be explained through Mendel's principles for single genes or alleles. It will not be until the mathematisation and formalisation of these biological processes by, most notably, Ronald Fisher, John B. S. Haldane and Sewall Wright, that began a truly quantitative understanding of genetics.

### 1) Population Genetics

This heated debate will eventually begin to settle when, in 1918, British statistician Ronald Aylmer Fisher published his paper "The Correlation between Relatives on the Supposition of Mendelian Inheritance"[37]. It is noteworthy to mention that the term and concept of variance were first introduced in this article, which will be later further developed[38]. Using experimental data from Francis Galton and Karl Pearson[39,40], Fisher examined the hypothesis that phenotypic features are "determined by a large number of Mendelian factors", *i.e.* that the continuous variation of complex traits such as the ones recorded by biometricians at the time could be explained by the contributive and imperceptible influence of each and every one of the genes implied. Indeed, the random sampling of the multiple gene variants that takes places at each generation (according to Mendel's principles) could produce a continuously distributed phenotype in the population.

This will be later known as the infinitesimal or polygenic model, and his theories will culminate in his 1930 book “The Genetical Theory of Natural Selection”, partly establishing the identity and importance of population genetics and quantitative genetics as new disciplines[41]. Where the former will be focused towards understanding the genetic differences within and between populations, the latter will try to determine the inheritance motifs of complex phenotypic traits across generations. In this book, Fisher will among other things mathematically demonstrate that natural selection operates through variations of allele frequencies in populations, thereby promoting the more adapted variants over others that are less so, effectively proving that Mendelian genetics and the natural selection theory are compatible.

Another leading figure of this newly founding domain of population genetics was also British, and around the same time as Fisher. Starting in 1924, the geneticist John Burdon Sanderson (J. B. S.) Haldane will write a series of papers leading up to their synthesis in his 1932 book “The Causes of Evolution”[42]. Redacted from a series of lectures, the publication summarises his results in quite of an informal way, the appendixes being the most formal and mathematical parts of the book. Although praised for its scientific interest, Haldane tends to blend personal opinions with his formulated theories and compelling results, which, according to Fisher’s words in his review of said book, makes it so that “one receives the impression more of able conversation on a series of interesting topics, than of a considered treatise on genetical theory”.

Nonetheless, Haldane managed to apply statistical analysis techniques to many real-life examples of natural selection, and in particular to the evolution of peppered moths in England during the 19th century. Indeed, at the start of the Industrial Revolution in England, most of these moths were white, peppered with black spots, in order to efficiently camouflage themselves when set to rest on the lightly-coloured bark of birch trees, a very widely spread species in the English countryside.

However, because of the numerous and newly installed coal-burning factories, rural areas and their birch trees began becoming blanketed with dark soot, and sulphur dioxide emissions led to lichen dying off trees, drastically modifying the environment in less than a century. This rapid biome transformation induced a substantial shift in moth populations: while the light-coloured variant was hugely favoured before towards hiding from bird predation and the dark-coloured phenotype was not, this relative advantage was completely inverted under this new polluted habitat, a process that would be referred to as “industrial melanism” (Fig. B.2.1).

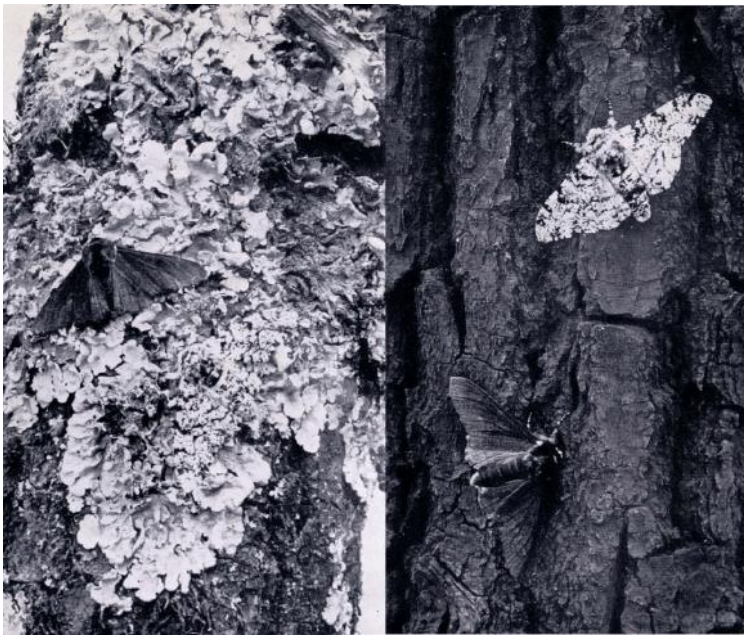


Figure B.2.1: Comparative photography of moth populations in rural England.
On the left, we can see both phenotypes of moth (light and dark colours) on a lichen-covered bark birch tree. Whereas the original phenotype is almost invisible, the darker colouration is very noticeable.
On the right, same situation but on soot-covered birch tree. The natural camouflage of the lighter variant of moth is incredibly flashing on the bark, on the contrary of the newer, more adapted dark moth variant.

Thus, entomologists observed that at the end of the century, nearly all of the phenotypes observed in these modified areas were of that dark-coloured type[43]. Using the recorded data, Haldane showed that for such a situation, in order to invade the whole population in less than a 100 years, the melanic (dark) phenotype would have needed to be at least 50% more adapted - stealthy in this case - than the previously dominant light-coloured variant[42]. This example all the more solidified the validity of mathematical models combining Mendelian genetics with natural selection, and their necessity if one were to try to quantitatively understand evolutionary processes.

## 2) Adaptive Landscapes

Finally, the last founder of population genetics is considered to be the American biologist Sewall Wright. He too developed many statistical tools, but most notably towards the representation of the interplay between genotype (or phenotype) and reproductive success of a population. To this end, Wright introduced the concept of evolutionary landscapes in his 1932 paper "The roles of mutation, inbreeding, crossbreeding and selection in evolution"[44]. If one was to try to enumerate all the possible combinations of parameters (alleles, genes, etc.) that would result in a viable, observable phenotype in nature, the list would be endless. As such, Wright thought of reducing the polydimensional array of genes into a single axis, plotted against the overall "adaptiveness" of the corresponding genotype (Fig. B.2.2). This way, impossibly complex systems can be easily represented, and movement of a population on this landscape reflects change in gene frequencies, towards a greater - or lesser - "evolutionary adaptiveness".

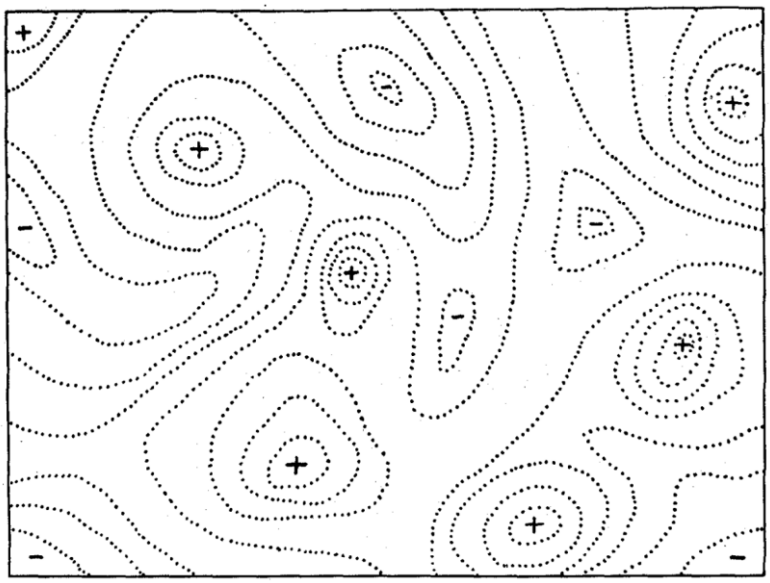


Figure B.2.2:
Sewall Wright's evolutionary landscapes.
The 2D surface (x- and y-axis) is described by all possible genotypes for the population. Just like in any sort of topographical map, the z-axis, representing overall "adaptiveness", defines the third dimension of the diagram.
The "+ zones" reflect high peaks, and "- zones" valleys. Here, adaptiveness refers to the reproductive success, or fitness, of the population. The greater it is, the more adapted the population is to its environment.

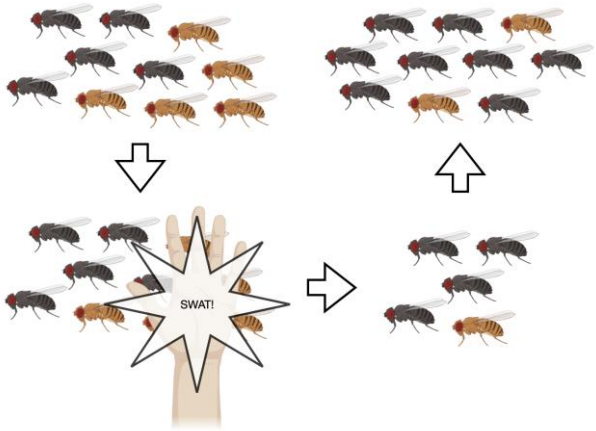
This adaptiveness can be considered as equivalent to reproductive success or fitness, *i.e.* the probability that an organism will succeed in passing its genetic material to its offspring, which is in turn based on several parameters, both intrinsic and environmental. Using this representation, it is quite simple to visualise the possible evolutionary trajectories of populations, either towards peaks or valleys. Moreover, the landscape can be considered smooth or rugged, depending on the shape of its topography. Indeed, if a small number of mutations result in a greater change in overall fitness, this could be pictured as a coarse, uneven terrain. On the contrary, if moving through the space barely has any effects on adaptiveness, the landscape will be quite flat and levelled.

Wright argued that although it surely is natural selection that drives populations to the nearest adaptive mountains, some other processes were necessary in order to explore new zones. Indeed, survival and reproduction is not always linked to fitness[45]. In the case of small populations that were quite isolated from each other, he observed that due to this random sampling, distributions of allele frequencies could change from one generation to the next, and as such, leading to a drastic reduction of genetic variation when inbreeding was too strong (Fig B.2.3). Because these effects can only remove variants and cannot add any, in the long term, they would thus tend to homogenise the gene pool in the population. With the sole modifying influence of mutations, this translates to an "erratic drift" around the landscape's peaks, contrarily to the "steady directional drift" of selection. Wright would qualify this as a trial-and-error non-adaptive wandering, allowing species to get away from their nearest mountains, moving throughout the valleys and plateaus in order to find other relatively higher adaptive peaks. Although the question of its relative strength and importance compared to selection fuelled many debates for decades afterwards (one of the most fervent critic being none other than Fisher himself), the existence of this competing evolutionary force was unequivocal.

However, Wright only considered landscapes where population genotypes are represented against their mean fitness. A few years later, his concept will be expanded by George Gaylord Simpson, an American palaeontologist who, even though not a mathematician, will be heavily influenced by Wright's work. In order to explain patterns of equine evolution in fossil records, Simpson proposed a reinterpretation of these adaptive landscapes, which would link, instead of genotypes, phenotypes to fitness[46]. This approach proved to be relevant, even in the context of the phenotypical data that was gathered by palaeontologists, and traditionally ignored by geneticists. Because this method was based on the examples of speciation in Equidae lineage, the landscape was necessarily dynamic in order to account for changes in ecological pressures, contrarily to the static nature of Wright's.

Figure B.2.3: Genetic drift analogy.

In this figure, we can see a small starting population of around 10 flies, with two predominant phenotypic variations (alleles), black and yellow colours. During some random cataclysmic event (hand swat), the greater part of the yellow subpopulation is wiped out. Even though both phenotypes were exactly as vulnerable to the event, one is still there while the other is almost extinct, purely by chance. Thus, the frequency of yellow individuals among the population will be much lower than in the next generation than it was before the random event. In this example, starting from a 50/50 ratio, we end up in an 80/20 proportion. It is noteworthy and quite intuitive to mention that the smaller the population, the greater the effects of the drift, and vice versa.




Although this landscape metaphor is a very powerful tool to aid biologists visualise how evolution operates, it is not in any shape or form an actual mathematical model that holds explanatory value. The most outspoken critics were William Provine, Sergey Gavrilets and Jonathan Kaplan, who held many grudges against the use of adaptive landscapes in order to envision evolution processes[47], and for the most part proposed to abandon the metaphor once and for all. In trying to give mathematical sense to the diagrams, *i.e.* trying to fit them to available models, one would often stumble upon contradictions that weaken the whole interpretation[48]. Moreover, as available computational power expanded through the years, the necessity of reducing the high-dimensionality of these systems was not so justified, and it was proved to actually be misleading in many cases[49]. Indeed, the "real shape" of these multidimensional landscapes would often be very different from what we can see and imagine, potentially misguiding the unknowing biologist to fruitless investigations[50].

Nevertheless, several examples show that this metaphor has been successfully expanded by multiple teams of researchers[51], and is still widely used to this day for its heuristic value.

## 3. Molecular Evolution

Reconciliation of the natural selection theory with Mendelian genetics in order to explain how evolution operates would be called the "Modern Synthesis", in an effort to unify from-macro-to-micro evolutionary perspectives across multiple biological disciplines. By the middle of the 20th century, evolution was thus analysed through the lens of a wide collection of fields, such as genetics[52], ecology[53], paleontology[46], embryology[54], and botany[55].

However, the whole evolutionary perspective that was developed through population genetics during the early 1900s will eventually have to adapt to the development of new technologies. As we mentioned previously (A.4, "On the methods and importance of studying proteins"), novel techniques such as X-ray crystallography or electrophoresis[56] quickly allowed scientists to investigate at a molecular scale, blending previously distinct domains like genetics, biochemistry, biophysics and microbiology. Known as molecular biology, these new experimental techniques and computational models will be crucial to this modern discipline, and thus reshape the way we considered the links between the different actors of biological systems, most notably the relationship between genes and proteins. With the numerous discoveries that were made, it quickly became evident that DNA was the main physical support of genetic information in living organisms, encoded in the specific sequences of nucleobases.

Once the structure of the informational molecule was discovered, the mechanism of replication promptly followed[57] along with the infrastructure of codons (nucleotides triplets) that underlies the genetic code[58]. In an effort to describe the flow of information in living systems, all this and more led to the formulation of the central dogma of molecular biology by Francis Crick in 1958[59], as follows (Fig B.3.1) : "Once 'information' has passed into protein *it cannot get out again.* In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible."

This hypothesis thus led the scientific community to focus on the foundation on this system, where the true information lies, in the genes. Indeed, in such a scheme, the phenotype of an individual would be determined by the state of the translated proteins, as the product of the interaction between genotype and environment. The phenotype, depending on its adaptiveness or fitness, is then selected for/against by natural selection, and its corresponding allele(s) will increase in frequency in the population. This view, mainly developed by George C. Williams in his book "Adaptation and Natural Selection"[60] and later popularised by Richard Dawkins[61], thus holds the genes as units of selection, instead of the canonical organism or even species-level that was widely used since population genetics were established. This approach would be known as the "gene-centred view of evolution". To this day, there is still debate over the relative importance of kin and group selection compared to that of genes[62].

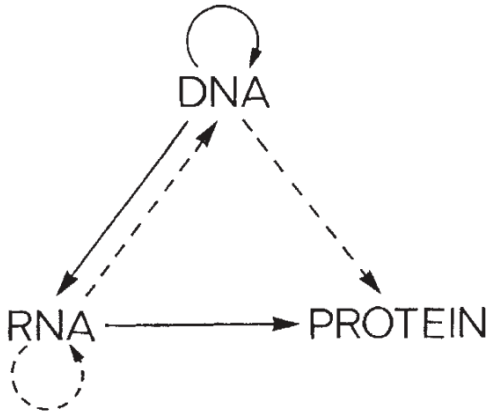



Figure B.3.1: Central dogma of molecular biology.
In most cases (solid lines), DNA can be copied into DNA (replication), and converted into RNA (mRNA to be precise, transcription). Proteins are then synthesised from these RNAs (translation).
However, in some instances (dotted lines), RNA can replicate itself (RNA replication), as well as revert back into DNA (reverse transcription). Proteins can also rarely be translated directly from DNA[63].
From Central Dogma of Molecular Biology[64].

Such change of perspective started withdrawing attention - and funding - from traditional evolutionary biology researchers to the favour of these newer and more promising projects. In particular, the advent of DNA and protein sequencing enabled the beginnings of molecular phylogenetics. Thanks to the works of Sanger on insulin[65], Pauling and Zuckerkandl on hemoglobin[66], and Margoliash on cytochrome c[67], the amino-acid sequences of these proteins were determined and compared between different species and lineages. Divergence between sequences led to time estimates between mutations, effectively acting as a "molecular clock".

By aligning these sequences next to one another, it was indeed possible to calculate the number of mutations that differentiate each of these lineages, and the result was quite puzzling: it seemed that the rates of evolutionary change were fairly constant over time and over different species, using paleontological data as "absolute benchmarks". Although these facts were indisputable, they were in contradiction with a few elements of the Modern Synthesis. Indeed,

it would seem rather bizarre that living organisms adapt at a constant rate, given the often-random nature of changes in evolutionary pressures driven by the environment. For example, the wide distribution of phenotypes in Darwin's finches would hint at a rapid speciation from a common ancestor, following the respective adaptations of each population according to its biome.

This molecular clock hypothesis would indeed imply a differential rate of evolution between proteins in an organism, and the phenotype or morphology of this individual. Needless to say, fierce debates ignited once more among the evolutionary biology community. This controversy crystallised around the neutral theory of molecular evolution, developed by Motoo Kimura and Tomoko Ohta and published in 1968[68], and supported a year later by the works of Jack King and Thomas Jukes in 1969[69]. Their proposition rekindled the argumentation around the relative importance of genetic drift compared to selection processes.

In order to explain the quickly changing patterns of molecular divergence within and between species, the hypothesis was that most mutations that appear and reach fixation in populations are neutral in regard to their fitness effect. Indeed, as the founders of population genetics showed, fixations of beneficial mutations are exceedingly rare, and cannot account for the amount and rates of change observed in DNA or protein sequences by the end of the 20th century. Moreover, the fact that the rates of evolutionary change "depend on time measured in years but are almost independent of generation time, living conditions, or even the genetic background"[68] would strongly hint at the predominant role of genetic drift in the process given its stochastic nature, contrarily to the conventional Darwinian forces of selection. Furthermore, if the vast majority of these mutations held little to no effect on the overall phenotypes and fitness, then natural selection would be almost powerless anyway, except for the occasional culling of the most unfit alleles that randomly emerges through mutation during the neutral variation[70].

Although the debates between supporters of this theory - the "neutralists" - and the classical darwinian "selectionists" lasted for decades, most researchers agreed to the key role of genetic drift in evolutionary processes.

## C. Enzyme Engineering

By the end of the 20th century, the numerous discoveries and advances in molecular sciences allowed a deeper understanding of protein structures, their functions, and in particular the relationship between the two. As mentioned previously, this time period saw the development of enzymology and structural biology, which were crucial in trying to uncover the secrets of enzymatic catalysis and protein folding. Although these beginnings focused on natural products, researchers quickly tried to go beyond what was found in nature by modifying and even building new proteins, in order to study the interaction between activity and structure in enzymes[71]. To this end, mainly two strategies emerged: rational design and directed evolution.

### 1. Rational Design

This first approach is heavily dependent on one's ability to predict how an amino-acid sequence will fold in its three-dimensional structure. Indeed, given that the spatial conformation of an enzyme is inherently linked to its biochemical activity, designing the function of an enzyme necessarily means designing the appropriate structure for it. Therefore, understanding the mechanisms upon which the linear chain reorganises into the final compound is key, and as such, the development of computer-based tools would eventually become fundamental to the whole process.

However, the first experiments could not rely on these computational methods, and the artificial blueprints were thus manually designed and optimised, based on protein sequences that were already available. Using solid-phase peptide synthesis[72], Bernd Gutte and Robert Merrifield managed to create several truncated proteins based on natural compounds[73,74]. These early successes jumpstarted the interest in the field in the following years, eventually leading to the inclusion of theoretical simulations for better results. Imitating what was found in nature is one thing, but creating entirely novel structures and functions is a whole other matter. As many scientists realised quite early in history, the possible variation and complexity observed in living organisms is nearly infinite[75]. If we consider a quite small 100 amino-acid protein, the total number of possible combinations with the 20 available amino-acids would be $20^{100}$ (roughly $10^{130}$), which is more than the total estimated number of particles in the universe. Such a sequence space is impossibly gigantic to effectively sample, so rules and guidelines must be set-up in order to explore this expanse as efficiently as possible.

Usually, the artificial target is heavily based on an already known protein structure, and only a handful of residues are mutated to other amino-acids in order to see the - hopefully beneficial, and otherwise limited - effects of these changes on the overall structure and activity of the protein. Such methods are known as site-directed mutagenesis, targeting specifically certain parts of the protein's DNA sequence, so that a modified version of the polypeptide is created through the transcription/translation processes. This drastically helps alleviate the extensivity of the sequence space sampling, as much of the macromolecule is then identical to its natural counterpart. Moreover, the final three-dimensional shape and flexibility of the protein mainly depends on the combination of its residues' properties. Some sequences of amino-acids are thus sometimes more frequent than others, reflective of their ability to shape the protein into specific conformations[76], reducing once more the explorable space around the target of interest.

The advent of molecular dynamics was originally quite decorrelated from any biological applications, but the benefits and pertinence of such methods in the fields of enzymology and structural biology quickly showed to be immense. As already mentioned in the first chapter, the core concepts of computational biology rely on the ability to effectively predict the 3D structure of the protein of interest, either by comparing it to similar-shaped proteins of known structure (homology modelling), or from scratch, based on its available amino-acid sequence (*de novo* modelling).

1) Comparative protein modelling (Homology Modelling).

This approach remains the most reliable method to predict three dimensional structures of proteins, with an accuracy that can be comparable to low-resolution, experimentally determined ones[77,78]. Using computational algorithms (FASTA[79] or BLAST[80]), it is possible to align two or more sequences in order to determine their similarities and differences, *i.e.* their sequence homology. Then, based on the assumption that proteins sharing a high percentage of homology in sequence will share very similar structures, after scanning the amino-acid sequence of a target protein, it is possible to use similar sequences' known structures as templates to model its 3D folding. Indeed, structures have been proved to be more conserved evolutionarily than their peptidic counterparts[81]. Due to how evolution mostly operates through single-point changes, crucial residues for function tend to be quite conserved, while others can coevolve due to their proximity to one another[82](Fig C.1.1). Once the template structure(s) has been established, the model is generated and subsequently assessed for its quality, and thus

plausibility. Usually, an energy will be assigned to the model, based on either statistical potentials[83] or quantum mechanics calculations[84], the aim being to have it the lower possible.

Whereas this technique is quite accessible, it necessarily relies on prior structural information on the targets, such as protein sequence and structure databases. Moreover, the quality of the generated models heavily depends on the sequence identity between the target and the template(s). The lesser the percentage of homology is, the more unreliable the predicted structure will be, with a threshold around 20-30% of sequence identity where errors become too severe[85].

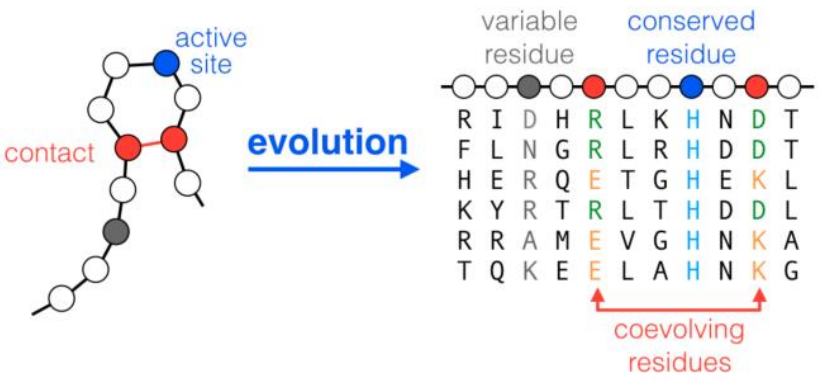



Figure C.1.1:
Coevolution and conservation of residues in protein sequences.
Six sequences are aligned next to each other. While many residues change between them, some are linked and evolve more frequently together. Others, often implied in crucial functions (active site, etc), are rigorously conserved through lineages.
Adapted from Cocco *et al.* [86].

However, recent advances in this field drastically changed the quality of results obtained through homology modelling. With the advent of artificial intelligence and deep learning, protein structure prediction algorithms are now – and increasingly so - able to propose reliable models for unknown proteins. Most notably, the AlphaFold software[87], developed by Google's Deepmind company, managed to predict nearly 100 protein structures with a median 92% accuracy during the 14$^{th}$ Critical Assessment of protein Structure Prediction challenge experiment in November 2020, similarly to experimental techniques such as x-ray crystallography[88]. By feeding and training the deep neural network with the tens of thousands of protein amino-acid sequences and their corresponding structures that are presently available in the PDB, the software is able to compute and predict the spatial constraints between each proximal residues in a new sequence. First breaking down an entire structure in smaller parts, the algorithm starts from small groups of amino-acids, solves their spatial conformation, and then pieces these small clusters to each other, refining the model as it goes, in order to end up with the final structure. This software and the efficiency it displayed was stunning for the structural and computational biology community, and thus holds out a great many prospects for the future of the domain, and the protein folding problem as a whole.

2) *De novo* modelling.

This technique is much more ambitious, in the fact that it tries to infer the tertiary structure of proteins directly from their sequences. By simulating the physical principles and interactions that would dictate the folding of the amino-acid chain, the aim is to predict which preferential 3D configuration(s) the protein will adopt. This assumes that these preferential structure (-s, there can be several) will have the lowest free-energy of all possible conformations.

Although it does not need any additional information apart from the target amino-acid sequence, contrarily to the comparative model, the number of computational resources needed for this exploration are tremendous, even for small proteins, and especially for large proteins. Indeed, due to the near-infinite number of possible conformations and variants that are still very similar to the target protein, sampling and modelling of all the possibilities is unreasonable. Therefore, in a step-by-step process, the computed conformations are slowly but steadily brought to the most thermodynamically stable state, through numerous iterations of folding and refolding[89]. One of the most famous and successful example of application of this method surely is the Rosetta Platform set up by David Baker and his laboratory in Seattle[90,91] (Fig. C.1.2), even though more and more of these initiatives aim to use more efficiently the often idle computational power that resides in most households nowadays (most notably Folding@home, Human Proteome Folding Project).

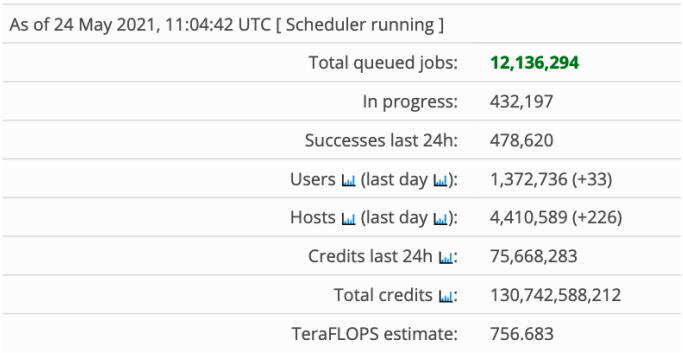



Figure C.1.2: Rosetta@Home traffic details. Rosetta@Home is a distributed computed project aiming at protein folding prediction. Idle computer processing resources are used from volunteers' computers to perform calculations.
Figure from Rosetta@Home website, Baker Lab, University of Washington.

It is worth noting that before any method is applied, it is always efficient to prepare the targeted structure by splitting it into its potential different domains. This could be considered as a pre-processing step that helps the algorithms in performing more efficiently. To this end, the two previous approaches can be considered, either by comparing said domains to other proteins in the databank, or by *de novo* modelling from the sequence. These domains are then regrouped as one to form the final structure[92].

## 2. Directed Evolution

Contrarily to the rational approach for protein design that was described previously, directed evolution processes do not require extensive knowledge on the target's structure, function or catalytic mechanism[93]. Indeed, the method relies on mimicking natural selection, where the investigator can themselves set the fitness parameters, adjust the pressure of selection as they see fit, and select the best variants that fit their needs. Just like in nature, the whole endeavour is a black box, the experimenter "blind" to the inner workings of the system, the only objective being improved overall fitness of the considered organism, protein, gene, etc.

The first directed evolution experiment took place quite a few years before the upturn of molecular sciences, in the 1960s, when Sol Spiegelman managed to set-up the first *in vitro* RNA-based self-replicating system[94]. By incubating RNA replicases (RNA-dependent RNA polymerases) with the Qβ bacteriophage (a RNA virus that targets and uses bacteria to reproduce) in aqueous solution with nucleotide bases and salts, consistent RNA replication was observed, and reproduced through serial dilutions[95]. Due to the thermodynamic constraints of the system, *i.e.* that shorter RNA strands are easier and faster to replicate than longer ones, the initial strand of several thousands of nucleotides always ends up collapsing into dwarfed versions of itself, only a few hundred bases long in less than a hundred generations. Albeit one could consider this drastic reduction in complexity more akin to regression than evolution, this bare-bone system retained its ability for auto-replication, and the sequence changed, *evolved*, in order to optimise the process resources and time-wise.

However, applying similar methodologies to enzymes will take several decades, until American researcher Frances Arnold developed directed evolution strategies for improved and novel catalytic activities in enzymes[96]. Although some prior examples of protein optimisation exist before her works[97], Arnold is still considered to have spearheaded the jumpstart of the field, and was later awarded with the 2018 Nobel Prize in Chemistry for her pioneering influence in the use of directed evolution for the discovery of new enzymes. The target protein was subtilisin E, a protease (cleaving protein) from the bacteria *Bacillus subtilis*, the objective to engineer it to be able to perform its catalytic activity in "highly nonnatural environments", in this case high concentrations of organic solvents[98]. The gene of interest was randomly mutated and expressed in bacteria, which were subsequently subjected to testing: plated on agar media containing their substrate and the organic solvent (dimethylformamide, DMF), only the bacteria exhibiting functional subtilisin would produce

a visible halo around the colonies. These were then collected, randomly mutated, replated, retested, etc. With four sequential rounds of this process, a variant able to efficiently catalyse the chemical reaction was isolated: only 6 additional mutations allowed the variant to perform the reaction 256 times faster than the original protein when in 60% DMF.

Such results were incredibly promising, and paved the way for ever greater achievements in the following decades. Even though there are nowadays numerous possible declinations of this method, the concept stayed the same than in this earliest experiment, as the system basically requires three things: variation between each generation, which induces notable and detectable fitness differences, that are heritable between cycles. If the directed evolution platform follows these guidelines, the targets can then be put through iterative rounds of mutagenesis, selection/screening, and amplification, in order to obtain fitter and fitter genes/proteins at the end of each cycle (Fig. C.2.1).

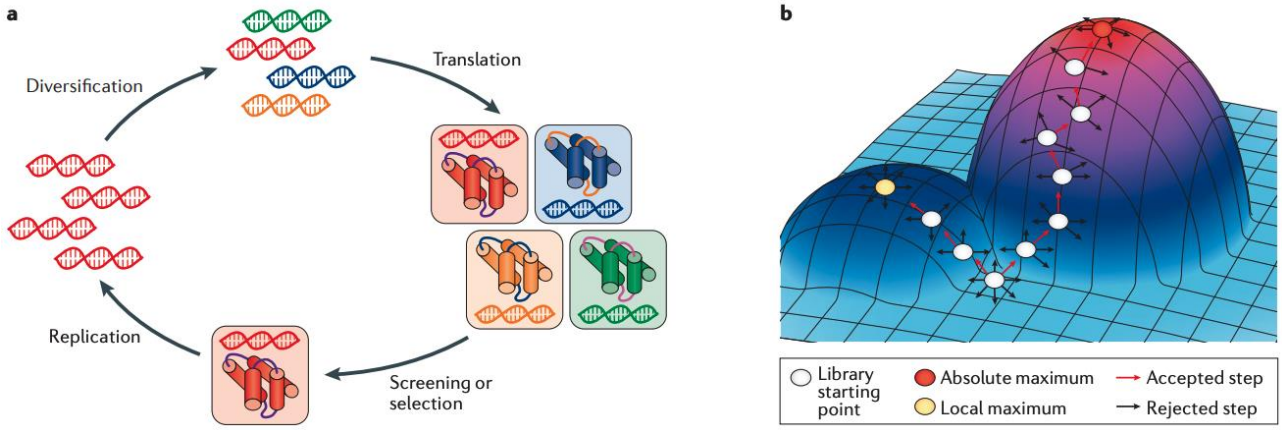



Figure C.2.1: Directed evolution schematics.
A) Basic principle of a directed evolution cycle. Rounds of diversification, expression, testing and amplification allow the user to evolve a gene/protein towards a desired phenotype (heat or antibiotic resistance, specific activity, etc).
B) Representation of numerous directed evolution cycles on an arbitrary fitness landscape. Several "peaks" can be reached from the initial "valley", some relatively higher than others.
Adapted from Liu *et al*. [99]

For the diversification step, the target gene sequence is usually declined in a highly diverse library of hundreds, thousands, or even more variants. The size and diversity of this library is important in two regards. First, quite obviously, the bigger the haystack, the harder it is to find the needle. But on the other hand, the wider one would cast its net, the greater its chances to stumble upon a treasure. There is thus a trade-off between the wideness we wish to explore on the landscape, counterbalanced by the ease - or difficulty rather - of searching through it. In the case of vast libraries, high-throughput assays are naturally preferable in

this endeavour, considering the meagre chances of producing a fitter variant[100]. To this end, one of the most used techniques consists in amplifying the natural error rates of DNA polymerases while amplifying the gene of interest. In a process known as error-prone polymerase chain reaction (ep-PCR), mistakes are made when duplicating the DNA strands, effectively producing numerous variants[101,102]. In directed evolution experiments, rates of few mutations per genes are appropriate to generate diversity without nullifying the original function of the protein[103]. More information on this topic will be detailed later in D. Study of noise in biological systems.

Enzyme engineering aims at either enhancing pre-existing properties of proteins or discovering novel types of catalytic activities. Henceforth, being able to efficiently evaluate the fitness of the numerous variants generated in the process is of crucial importance. To this end, two different approaches can be considered, the screening or the selection of variants. Discussion on the respective pros and cons of each strategy will be developed in the next sub-chapter. However, the concept is always the same: testing variants and selecting them for their increased fitness compared to the original phenotype.

Finally, one of the main differences between these artificial selection systems and how nature operates resides in the heritability factor. Indeed, as we explained earlier (B.1: Theory of Natural Selection), in living organisms the link between generations is simply provided by Mendelian genetics, as the genotype of the offspring is based on the ones of its parents. Selective advantages present in an individual are anchored in its genes, and with luck and time these will eventually propagate to the rest of the population. However, in the case of directed evolution, the genotype and the phenotype of an individual can be isolated from each other, contrarily to natural conditions. Especially for *in vitro* experiments, but also relevant *in vivo*, the proteins expressed are not necessarily tested in presence of their respective coding genes. Therefore, to be able to relate the random mutations that are generated to the most fit variants obtained after each cycle of evolution, a genotype-phenotype linkage is required at all times. Without a convenient way to associate each gene variant with its corresponding protein and thus phenotype, the process cannot be iterative. In this endeavour, researchers developed several solutions[104,105] : either by physical linkage (covalent or non-covalent), or by encapsulating both in the same compartment (droplets-based strategies).

In Arnold's experiments for example, this link was consistently maintained by the bacteria, which held genetic information at all times, and expressed it into the cognate

proteins. Notable examples of such methods include : phage display[106], bacterial display[107], ribosome display[108], mRNA display[109] (Fig. C.2.2).

Even though both approaches present their respective strengths and weaknesses, they are quite often used in conjunction, in "combinatorial" or "semi-rational" strategies[110]. Drawing from rational design actually allows the creation of more focused libraries, for example by incorporating beneficial mutations that were already characterised in the starting gene of interest. In a sense, this equals to getting a head-start in the search of fit variants, instead of spending a lot of time and resources scanning and testing through countless uninteresting ones.

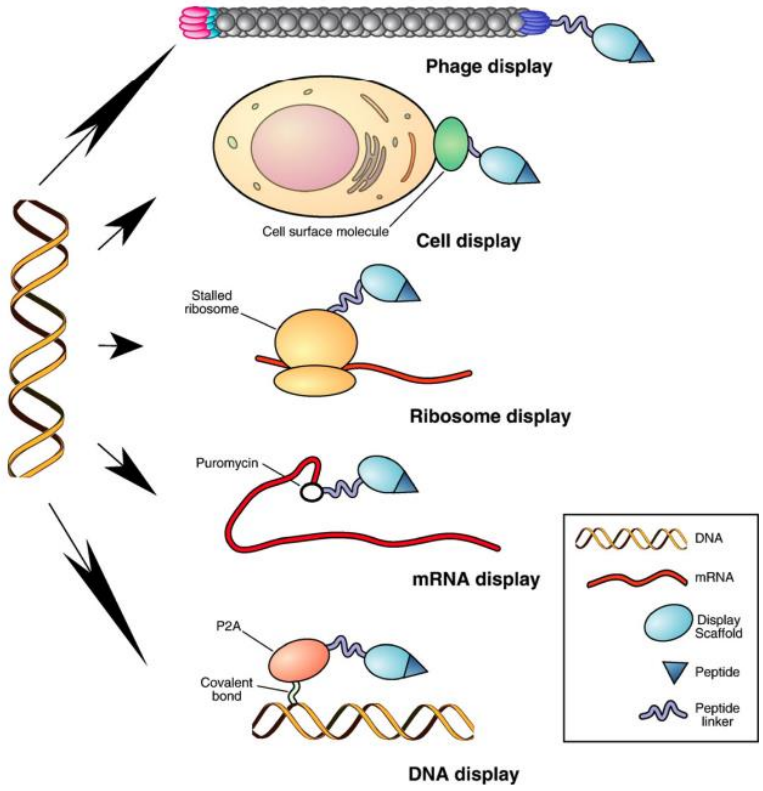



Figure C.2.2: Common display techniques.
In the most common display techniques, the genetic material (DNA) is physically linked to its corresponding phenotype, of various and diverse forms. In most cases, the polypeptide expressed from the gene is fused to another protein, except in the case of mRNA display, where a puromycin linker is used to covalently bind genotype and phenotype. In any case, these techniques allow the recovery of the DNA strand and its corresponding phenotype, based on the desired physico-chemical properties of the latter, such as binding efficiency to a target, or even resilience towards physical (temperature) or chemical (small molecules) perturbations. Adapted from Sergeeva *et al.*[111]

## 3. Selection & Screening

In order to detect the fitness differences between the multiple variants and to identify which are most presenting the desired properties, a screening or selection strategy must be set up. While screening methods are based on the individual evaluation of every member of the library and subsequent physical sorting of the best variants, selection approaches couple the presence of the desired property to survival, so that only functional proteins are kept for the next cycle. Whether it is for rational design or directed evolution, both types of evaluation can be applied, even though the greater size and diversity of directed evolution libraries often makes selection methods more appropriate. We will now address the benefits and drawbacks of each strategy, along with some examples.

### 1) Screening methods

Screening platforms rely on the spatial seclusion of every variant, followed by the quantitative characterisation of their activity. This necessarily sets some constraints on the throughput of these methods, as it is more experimentally tedious and heavy to separate each and every member of the library. Developing fast and autonomous ways to perform the assays is thus much more peremptory than in selection systems. To this end, using physical parameters as proxies for levels of fitness is an efficient way to facilitate evaluation and ensuing proper sorting of the variants. For example, linking the variants' activity to the synthesis of fluorogenic, colorogenic, light-generating or absorbance-modifying compounds allow for the quantitative comparison of individuals from the library. These molecules can either be directly the outcome of the enzyme's activity, or a by-product of the physico-chemical interaction between the variant and another substrate. The instigator can then set a threshold for the detected parameter - fluorescence intensity for example, directly linked to the activity of the protein - to determine which variants will be collected and injected into the next round.

However, these methods still present several limitations. Most notably, it is sometimes very complex and even unfeasible to establish such proxies between fitness and quantitative, detectable physical parameters of the system. Whether it is because these exogenic compounds are toxic for the cells, for the biochemical reaction of interest itself, or simply because no sensors can be efficiently coupled to the enzyme's activity, alternatives have to be considered. In this regard, *in vitro* compartmentalisation (IVC) can alleviate and even dismiss such issues, providing an agnostic, artificial environment for screening[112]. The development of microfluidics-based techniques were a defining milestone towards cell-free strategies, effectively replacing the natural compartments with droplets in water-in-oil emulsions[113,114].

Several different processes were developed in order to enhance the relatively low throughput of screening methods, such as the use of microplates[115], eventually combined with robotic-assisted automation[116]. These platforms enable the screening of hundreds to several thousands of variants thanks to arrays of wells that can hold up variable volumes of samples. From the millilitre scale - easily handled by humans - to the nanolitre scale - more adequately manipulated by robots - these methods are relatively simple to upscale, leading to greater and greater throughputs[117]. Another potent technique developed at the end of the 20th century that found use in many different fields of biological and medical research is flow cytometry[118]. Especially in the context of detecting particles or cells, the Fluorescence-Activated Cell Sorting (FACS) enabled researchers to reach even higher throughputs, up to tens of thousands of variants screened per day[119]. In this system, particles are flushed in a single-file and individually passed through a laser in order to detect eventual fluorescence (*i.e.* presence of fit variant). Droplets are then generated, and depending on the fluorescence level recorded each compartment is sorted into different categories at the end of the column (Fig. C.3.1).

Even though screening strategies generally present lesser throughputs than selection methods, the quantitative data that is obtained from the phenotypic characterisation of each variant is invaluable for the overall description of the library (activity distribution, fitness landscape, etc), which cannot be obtained through selection.

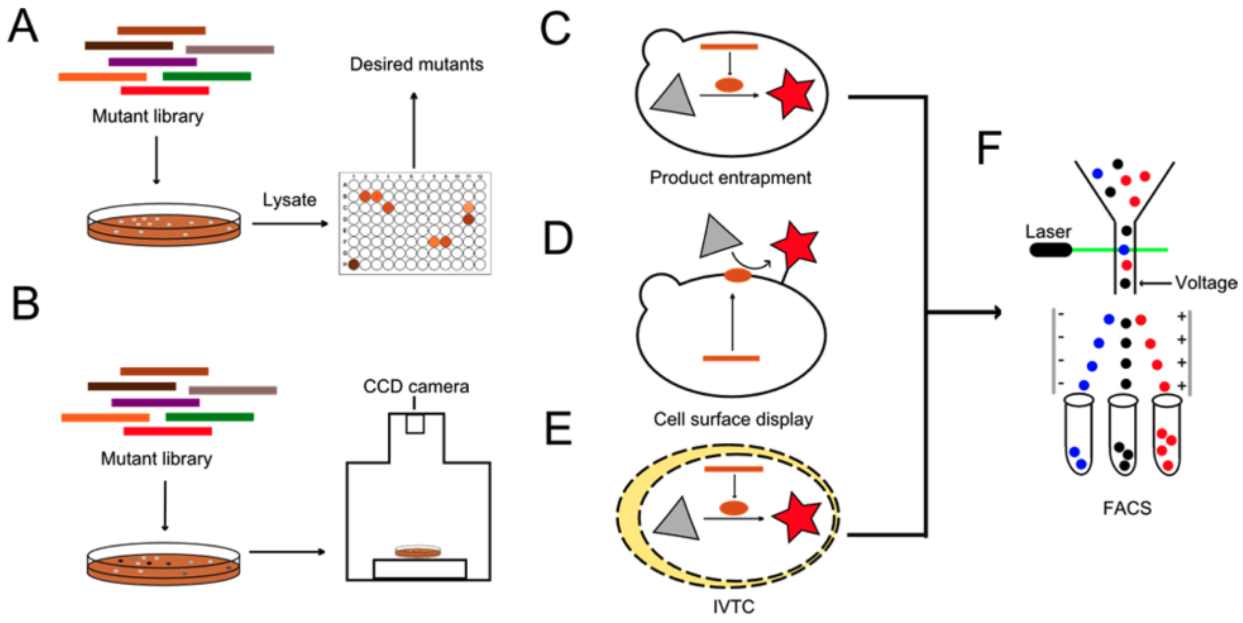



Figure C.3.1: Overview of screening methods.
A) Microplate screening. After transformation of a library of variants in bacteria and plating, each colony was lysed and tested in the wells of a microplate. Followed by subsequent screening based on the synthesis of colorogenic compounds. B) Digital imaging. Same principle as A, but for light-emitting colonies.
C) Product entrapment. The gene of interest (orange bar) is isolated in a cell, which produces the cognate protein (orange oval). The substrate (triangle) is then converted into a fluorescent compound (star) if the variant is adapted. The cell is then screenable due to is fluorescence.
D) Cell surface display. Instead on isolating the gene and protein inside the cell, the latter is attached to its surface, while the fluorescent product is also linked to it via enzymatic reaction.
E) *In vitro* compartmentalisation. Same as C, but instead of a living cell, an artificial compartment is used (microfluidic-based droplets for example). F) FACS. Cells or particles are sorted depending on their fluorescence levels. Adapted from Xiao *et al*. [120]

### 2) Selection methods

Selection systems, on the other hand, allow the evaluation of much larger libraries. Indeed, in such methodologies, one does not need to individually analyse every variant to be able to filter the best ones, as these are autonomously selected for by the system. In the context of protein engineering, the desired enzyme properties[121] - activity[122], stability[123] or even specificity[124] - must relate to the fitness of the individual, either translating to more frequent survival or to greater replication rates compared to other variants.

One of the easiest and earliest applications of this principle is the selection for binding affinity. Protein domains, antibodies, but also oligonucleotides strands (DNAs, RNAs, mRNAs, etc), can be tested for the strength of their binding to specific substrates or molecules, which are immobilised on surfaces[125,126]. Libraries are variants are flushed over the targets, and subsequently washed away. In this case, survival means staying bound to the target, as only the fittest variants will exhibit greater binding affinities. These are then collected with their respective genetic material for more in-depth analysis if their features are satisfactory, or injected into new rounds of variation and selection if the properties are still lacklustre.

More often than not however, the enzyme/trait of interest is not directly linked to the cell survival or tolerance to exogenous species. Indeed, towards the study of more complex systems, establishing such links between the studied trait of a protein and its overall fitness can be quite tedious. One approach is to apply extremely stringent selective pressures on the targets, such as the presence of antibiotics in the media. In this regard, the aim is then to engineer the host and its metabolism in order to correlate the fitness of a variant to its expression of antibiotic resistance, and thus promote the survival of the fittest[127]. Another complementary method consists in "sabotaging" every host of the library so that the default phenotype is bound to die except if a fit variant saves it. Experimentally, this translates to auxotrophy, *i.e.* the inability of the organism to synthesise a vital compound for its growth and survival[128]. In the presence of an active enzyme, the missing metabolite is produced, and survival ensues. Moreover, with this method, one can then measure survival, *i.e.* fitness, and use it as a proxy for function.

It is however noteworthy to mention that living organisms often find ways to bypass the constraints of systems engineered to steer them in specific evolutionary directions, and whether it is through simple genetic recombination or through more complex mechanisms, these eventualities need to be taken into account when designing experiments[129]. As mentioned previously, one very effective way to solve these issues is to transfer the whole strategy to an *in vitro* setting. One specific example, quite similar to the *In Vitro* Compartmentalisation mentioned previously, and conceptually at the basis of the experimental platform that was developed through this thesis, is the Compartmentalised Self-Replication (CSR) from Holliger[130] (Fig. C.3.2).

In this process, the *Taq* DNA polymerase is the target of the directed evolution process, the objective being to improve its catalytic activity for DNA synthesis. Originally extracted and isolated from *Thermus aquaticus*[15], a thermophilic bacteria, this enzyme quickly became a major component of common molecular biology methodologies - and most notably for polymerase chain reactions (PCR) - because of its heat-resistant properties[131].

In CSR, although bacteria are needed for their transcriptional and translational machinery, they are simply used as an expression vector for the *Taq* genes and their cognate proteins. The selection step is actually performed *in vitro*, which allows for a less biased sampling of DNA polymerase properties. Indeed, the expressed phenotypes do not need to be viable for their host in the long-term, they just need to be efficient at replicating DNA strands. A range of conditions can thus be applied to the system without implications for the organisms - even higher temperatures, the presence of inhibitors/disruptors in the reaction buffer, etc - which makes the platform incredibly versatile for the selection and subsequent evolution of DNA polymerases under a wide array of constraints.

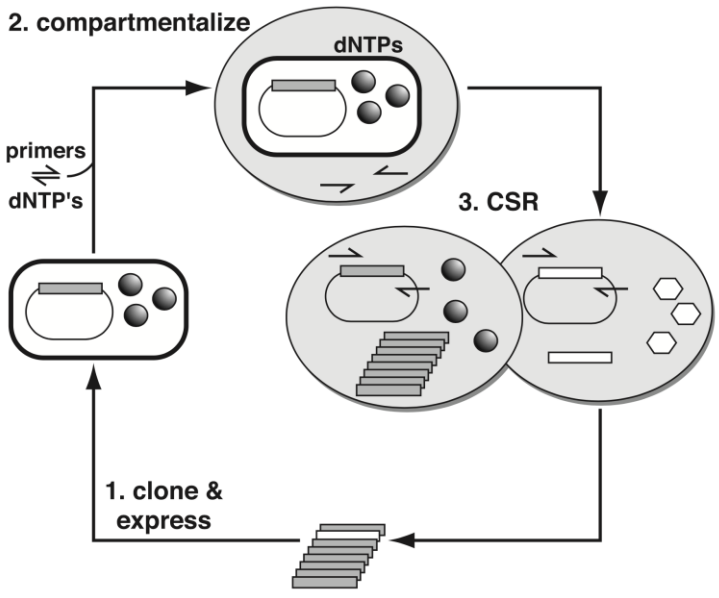



Figure C.3.2: Compartmentalised Self-Replication (CSR) strategy.
Starting from a library of DNA polymerase genes, each cycle consists in: 1) Cloning and expression in *E. coli* strains. 2) Encapsulation of individual bacteria in droplets along with material for PCR (reaction buffer, primers and dNTPs). 3) Auto-selection via PCR. The heat of the process lyses bacteria, releasing the polymerases in the droplets. The most active enzymes replicate their own genes a lot, while inferior variants fail to do so. After this step, the emulsion is broken and the genetic material retrieved. The resulting library is enriched in the fitter *Taq* genes, and injected into another cycle.
Adapted from Ghadessy *et al*. [130]

As previously stated, *in vitro* strategies present several advantages compared to living organisms. First and foremost, the cloning process is a natural throughput bottleneck, as the transformation efficiency into cells is always imperfect.

## D. Study of noise in biological systems

Although living organisms gradually complexified through consistent and reliable mechanisms, the fundamental role of randomness and stochastic processes in these complex systems cannot be undermined. Fluctuations exist at every scale in biology, from macroscopic populations to microscopic concentrations of proteins in cells. Random variations in biological systems became a subject of study itself, far from what was simply considered as artefacts and exceptions in overall trends, or even nuisances for the experimentalists. As the example of the genetic drift process discovery shows, even though such events are negligible in effect when considered at the macroscopic scale and in the case of large populations, the influence of random fluctuations should not be neglected in any molecular setting, for their substantial consequences towards evolution. Along with the advent of quantitative and synthetic biology during the last two decades, more and more evidence suggests that this stochasticity is actually beneficial for individuals, populations, and species.

### 1. Noise in Nature

As discussed in length in the second chapter, natural selection needs phenotypic differences between individuals in order to effectively promote the survival and reproduction of the more adapted variants. An important source of diversity resides in the mutations that randomly arise at the genetic level, erratic changes in the sequence of informational biomolecules like DNA. However, other forms of phenotypic variation have been investigated long before the development of molecular biology. Indeed, what was first described as "non-genetic diversity"[132,133] steadily unveiled to encompass multiple types of natural fluctuations. For example, human twins share the same genetic material, but very often present distinct phenotypes. Although these instances were thought to be the result of environmental adaptation, it quickly appeared that inner sources of variation existed beyond what genes were encoding. In a similar fashion, bacteria inside greater colonies are genetically identical to one another, and yet when grown in the same environmental conditions exhibit "characteristic behavioural differences"[11]. Whether it is through variable reactions to chemical stimuli or different lengths for their division cycles, an "individuality" can be assigned to each bacterium.

Through the development of both computational sciences and engineering methodologies applied to biological problems, our understanding of these stochastic events grew entwined with a more systemic approach of biochemical processes. As heralded by quantitative biology, the aim of systems biology is to elucidate the overarching principles behind the architecture of living organisms and the innumerable complex interactions that take place within them. In this regard, being able to substantially simplify these circuits and networks in order to study them apart from each other has been a boon, through the growth of synthetic biology as well[134,135].

From these new experimental strategies[136,137] and mathematical analysis[138], we learned that what we named "noise" can be caused by multiple factors, and can actually be separated into two different types of stochastic fluctuations : intrinsic or extrinsic. The former is considered inside individual cells, and relates to the inherent probabilistic nature of biochemical reactions, of substrates and enzymes randomly colliding with each other before binding or reacting. The latter, however, is linked to the differences between cells, *i.e.* how resources allocation in the whole organism influence several distinct pathways. A compelling example is to consider a dual-reporter set-up, with the expression of two different fluorescent proteins in cells, one green and the other red (Fig D.1.1). Where extrinsic noise would be responsible for the distinct levels of expression for each protein between cells, intrinsic noise would relate to the inner differences of gene expression inside an individual cell.

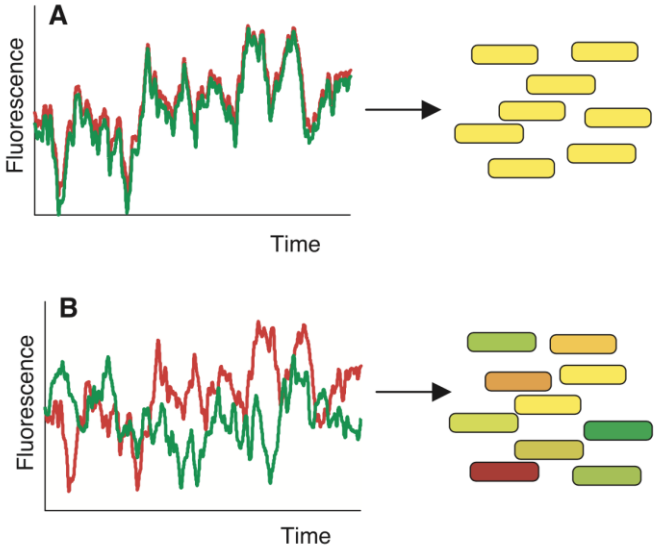



Figure D.1.1: Differences between intrinsic and extrinsic noise sources.
Bacteria are grown expressing two fluorescent proteins, one red and the other green, both regulated identically.
A) In the absence of intrinsic noise, the green and red proteins are expressed in equal amounts, resulting in an overall yellow colour for the cells. Although every individual is yellow, the absolute quantities of proteins between them is not necessarily the same, because of extrinsic noise.
B) In the presence of intrinsic noise, the ratio of green and red proteins is not the same between cells, resulting in a range of different colours in the colony.
Adapted from Elowitz *et al.* [137]

One of the most common manifestation of noise at the molecular level is the finite number effect. In the case of protein synthesis in cells and their subsequent eventual translocation, we can intuit that because of the smaller size of the nuclei compared to the cytoplasm, the concentration of protein changes more abruptly in the former than in the latter. One can thus observe that the smaller a system is, the more susceptible it becomes to random fluctuations and events, *i.e.* noise[139]. However, biological noise also has another fundamental impact on gene expression itself, which, just like protein synthesis, is governed by the probabilistic nature of numerous biochemical reactions : binding of promoters, repressors, etc[140]. Several works have showed that both transcription[141,142] and translation[136,143] processes can operate through burst windows because of this finite number effect, thus increasing noise in the gene expression process. Indeed, the cascade of processes for gene expression generally involve small numbers of molecules, whether it is the genes themselves, the mRNAs, etc.

First, in the case of transcription. The efficiency of mRNA production - and thus of the transcription step - is dictated by the kinetics of promoter activation. If these are slow to bind and unbind, models predict an all-or-nothing synthesis of mRNAs, *i.e.* “bursts” of transcription. Intuitively, we can see that the faster the promoter kinetics are, the smoother mRNAs are produced, and the lesser the noise in the system. As one would expect, the effects of fluctuations generated at this step of the overall gene expression would cascade downstream and affect every subsequent step of the process. Secondly, for the translation step. Because most mRNAs have much shorter lifetimes than proteins, bursts of mRNAs induce immediate and proportional protein translation in the system: “translational bursting” corresponds to a similar pattern of erratic production. Protein abundance levels will thus reflect this bimodal regime, with either very high or very low concentrations in cells, which would lead to a highly heterogeneous populations of cells in terms of protein content.

In both cases, the kinetics and magnitude of these bursts can be modulated via greater decay rates for mRNAs, and to a lesser extent for proteins. The effects of transcriptional bursts on translational dynamics would be considerably lowered if mRNAs are quickly degraded in the system. In the same fashion, protein level heterogeneities between cells are bound to flatten the faster protein decay is. The overall consequences of these noisy effects towards gene expression can thus be, in theory, buffered[144].

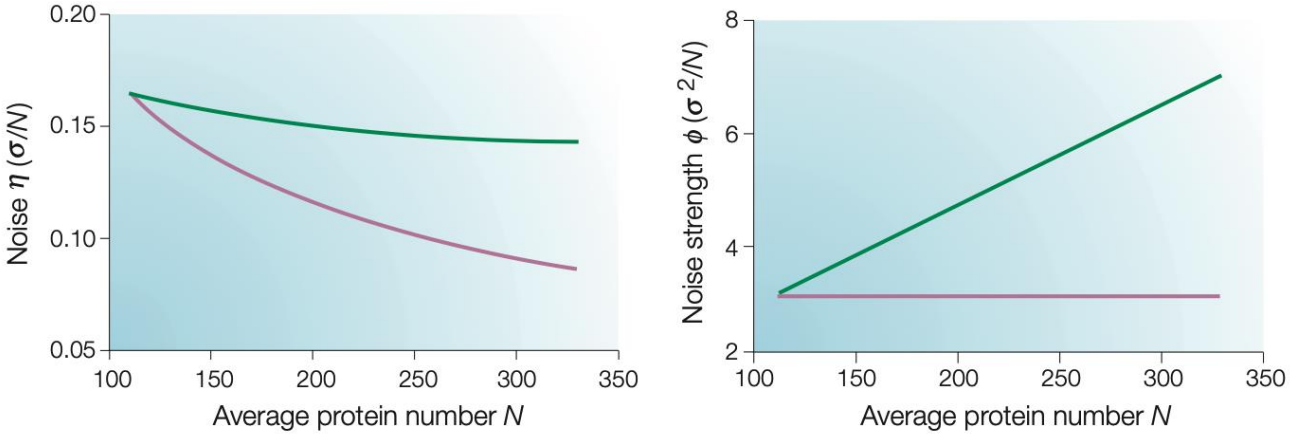



Figure D.1.2: Noise dependency on protein abundance.
One common measure of noise in biological systems is the coefficient of variation, η or CV. This parameter is the ratio of the standard deviation σ to the mean N (σ/N), and scales along 1/√N. On the other hand, noise strength is defined as φ= σ2/N.
Left - Noise dependency on average protein abundance when transcription (pink) or translation (green) rates are increased. We can see that transcription has a much more defined effect on noise than translation. Right - Same as Left, but for noise strength. Transcription has almost no effect on noise strength, whereas it increases linearly with the protein abundance when translational efficiency is increased. Low transcription/high translation regimes are thus characterised by overall increased levels of noise, contrarily to the other end of the spectrum, high transcription/low translation rates.
Adapted from Kaern *et al*. [139]

## 2. Mechanisms for Noise Control

As we just mentioned, the random fluctuations that manifest in gene expression, albeit independently of living systems, can be buffered and altered. However, as extensively discussed in the previous chapters, life seems to always finds a way to make use of its constitutive physico-chemical processes. We will now describe several mechanisms adopted by living systems in order to accommodate the irrevocable presence of noise in biological processes, and even how they manage to actively use it to their ends.

In regards to gene expression, a very common way to modulate the outcome of the network is to implement feedback regulation. On the one hand, negative feedback loops are quite intuitively considered to be noise-reducing mechanisms, as they essentially tend to settle systems into homeostasis or stable states, and thus dampen the effects of stochastic perturbations[145–147]. On the other hand, positive feedback strategies generally amplify the random fluctuations and their effects on population diversity. Moreover, these also often result in bistable cellular states with high and low expression profiles, effectively transitioning from a graded response to a binary one, promoting phenotypically very distinct variants in genetically identical populations.

At the molecular scale, noise is characterised through the error rates of biological processes. Whether it is for the synthesis of oligonucleotides such as DNA strands or peptides and proteins, the physico-chemical similarities between building blocks (nucleotides or amino-acids in the former examples) often vary, and it is possible for the machinery of living organisms to make mistakes during their respective biochemical reactions. As we discussed previously, errors, or mutations, often lead to deleterious effects on the overall fitness of an organism. As such, nature thus found ways to modulate how noise operates *in vivo*, in order to enhance the resilience of living organisms to the stochastic and unpredictable nature of the millions of molecular interactions that sustain them.

As the case of protein expression is particularly interesting to us in the context of this project, we will illustrate this concept with the examples of kinetic and conformational proofreading[148–150]. In living organisms, this mechanism most notably allows to drastically decrease the error rate of the translation step, compared to other precise and reliable processes such as DNA replication (Fig. D.2.1), although it is also used to the same end for other processes, such as DNA repair[151] or antigen discrimination by T-cells[152]. Indeed, during protein biosynthesis, the differences between the "right" and the "wrong" tRNAs to bind in the ribosome - in order to collect the cognate amino-acid - are so minute that even one base difference in the mRNA codons can acutely change the nature of the translated protein. This possibly leads to deleterious or malfunctioning proteins, an issue because of the time and resources wasted in the synthesis, and the potentially disruptive or even toxic effect of such polypeptides on biological systems.

| Biological process | Average error rates / Noise |
|---|---|
| DNA Replication | $10^{-8}$ - $10^{-10}$ |
| RNA Transcription | $10^{-4}$ - $10^{-5}$ |
| Protein Translation | $10^{-3}$ - $10^{-4}$ |

Figure D.2.1: Average error rates of fundamental biological processes.
Error rates are indicated as the number of mutations inserted over the total number of units utilised (DNA or RNA bases, and amino-acids). For example, a rate of $10^{-3}$ during protein translation means that one amino-acid over a thousand inserted in a protein sequence is a wrong one. DNA replication, of all the processes presented, is by far the most precise and reliable. Intuitively, one can understand the utmost importance of a very low error-rate in the replication of the genetic material between different generations of cell lineages in living organisms.

To this end, an additional irreversible step is inserted in the error-corrected mechanism. In the case of protein expression, the amino-acylated tRNAs are actually brought and linked to the ribosomes by a protein complex, EF-Tu•GTP, the elongation factor Tu being a GTPase with its substrate. Through tRNA binding to the mRNA that is being translated, the GTP of the EF-TU is hydrolysed into GDP, and the elongation factor protein is discarded from the ribosome. This simple, energy-consuming step allows for a finer discrimination between the right and wrong tRNAs, as the ribosomes will not trigger this hydrolysis for complexes carrying non-cognate tRNAs. This mechanism should not be confused with conformational proofreading, that also takes place in the ribosomes and other proteins, but does not require any energy expenditure[150].

While the efficiency of kinetic proofreading depends on additional steps and energy consumption, conformational proofreading uses the inherent differences of free-energy between the correct and incorrect tRNA-protein complexes that randomly try to bind to the translating mRNA (Fig. D.2.2). Right before the activation of the EF-Tu GTPase, there already are spatial discrepancies between those two variants, that lead to a difference in chemical stability for the corresponding complexes with the A-site of the ribosome. The wrong aminoacyl-tRNAs are thus more likely to be discarded even before the kinetic proofreading step, without energy expenditure. However, it is noteworthy that both methods still rely on setting an artificial bottleneck on both right and wrong substrates, the "wrong one" being much more stringent than the "right one". This effectively allows for a better specificity in the molecular system, although at the cost of either time or energy.

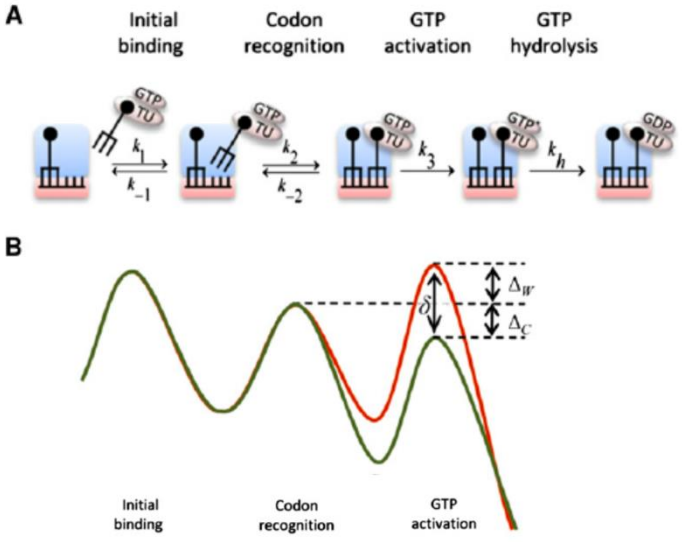



Figure D.2.2: Conformational proofreading during protein translation.
A) Steps of molecular recognition between the ribosome and the aminoacyl-tRNA•EF-TU•GTP complex. Kinetic proofreading is the irreversible "GTP hydrolysis" step. Before that, the ribosome needs to activate the EF-TU GTPase, using conformational proofreading. In the worst-case scenario, even if free-energy differences between the correct (C, green) and wrong (W, red) tRNAs were indistinguishable up to this step, spatial deformations of the ribosome allow for more precise discrimination between the two complexes.
B) Free-energy landscape of codon recognition for the successive steps shown in A).
Adapted from Savir *et al*[153].

On the other hand, at the population scale, more often than not, nature and living organisms have adapted to integrate stochastic fluctuations to their intended functioning in order to actually benefit from them, when it is too complex or inefficient to modulate them at the molecular level. One of these strategies is known as “bet-hedging”, a trade-off between the mean and the variance of fitness in a population[154]. Such examples include differential seed germination in plants[155], female multiple mating[156], and bacterial persistence[157]. In the case of quickly-changing environmental conditions, chances of survival for fit organisms tend to decrease, and can even lead to populations going extinct if alterations are too dire and/or too swift. Of course, these events also depend on the adaptability of the population, and its response rate towards environmental change.

One solution to this issue that appeared in many different orders of life thus consists in creating “fallback individuals”, in order to survive many possible contingencies. In each of these systems, phenotypic variation in a population is a mean to ensure that some of the offspring will always be adapted to its environment, even when it fluctuates a lot. Several variants of this strategy exist, reflective of their evolutionary boldnesses[158] :

- Conservative bet hedging: “A bird in the hand is worth more than two in the bush”

This corresponds to the safest way to ensure that the population will never go extinct, which is to exhibit quite low diversity in the population, well adapted overall. Organisms never are as fit as they could theoretically be, but they show robustness to their environmental conditions, *i.e.* a low-risk/low-reward strategy, that is fruitful in most cases but insufficient in the case of rare and extreme background fluctuations. These populations thus lean towards reducing their mean fitness in order to lower the variance in fitness across generations.

- Diversified bet hedging: “Don’t put all your eggs in one basket”

The opposite approach is more similar to high-risk/high-reward gambits. This would relate to spreading the survival efficiency of phenotypes in the population, which would possibly pay off in the eventuality of some environmental fluctuations. Although the fitness variance of the offspring is greater than in most cases, it is at the cost of the adaptiveness of most individuals in the population. When considered in long-term perspectives, it is a quite efficient way to shield the species from extinction, as the probability of creating adapted individuals at any generation is never null.

Intuitively, we can expect that diversified strategies are much more potent in the case of quickly and strongly changing environments, as the safe bet of conservation does not always hold its promises. Moreover, these precautions necessarily have a cost in time, resources and energy for the organisms, as one would expect. Bet-hedging and the phenotypic heterogeneity it implies in a population therefore always lead to an overall drop in mean fitness for its usual environmental conditions, but to an increase in fitness for highly variable environments, and the faster the changes, the greater the increase[159].

## 3. Consequences for protein evolution & Perspectives

As we have seen through the previous examples, living systems developed a plethora of complex physical, chemical and biological mechanisms and networks in order to either benefit from random fluctuations, or to buffer its eventual prejudicial effects. Recent advances in systems biology and molecular genetics raised several hypotheses towards our understanding of evolution in organisms, concurring on the predominant role of network architecture on the apparition and selection of novel traits[160,161]. Indeed, such works postulate that most of the innovations that gave rise to complex living organisms would not come from the spontaneous generation of new functional components or processes such as proteins or biochemical reactions, but rather innovative ways to couple regulatory elements inside a system (molecular circuits, feedback loops, signalling pathways) as a means to generate phenotypic diversity. How complex biological systems handle noise at the genetic or phenotypic level would thus heavily rely on their specific architecture of regulation.

However, in the case of core processes such as transcription and translation, that are at the basis of simpler organisms like prokaryotes and hence removed from a complicated system of regulatory elements, it has been shown that noise minimisation could be considered as an evolvable trait in and out of itself[162]. Admittedly, important and sensitive genes which could lead to their host's death when subjected to important fluctuations of expression are very often found to show high transcription and low translation rates, minimising the impact of fluctuations on the gene overall expression (cf 1) Noise in Nature). The same can be said for genes encoding proteins involved in stochiometric complexes, as ratios between each and every component is crucial to their role, thus leading to a high sensitivity to noise. Therefore, the ability of a system to manage noise is a trait than can be selected for and against, and has been for a long time in nature.

In the context of proteins, one of the smallest scale of systems subjected to evolution, there are two mains properties that describe the propensity and quality of an individual or a population when reacting to such fluctuations: robustness and evolvability. Where the former can be defined as the persistence of traits under stress and perturbations, the latter corresponds to the ability to exhibit novel functions following genetic variation (Fig. D.3.1).

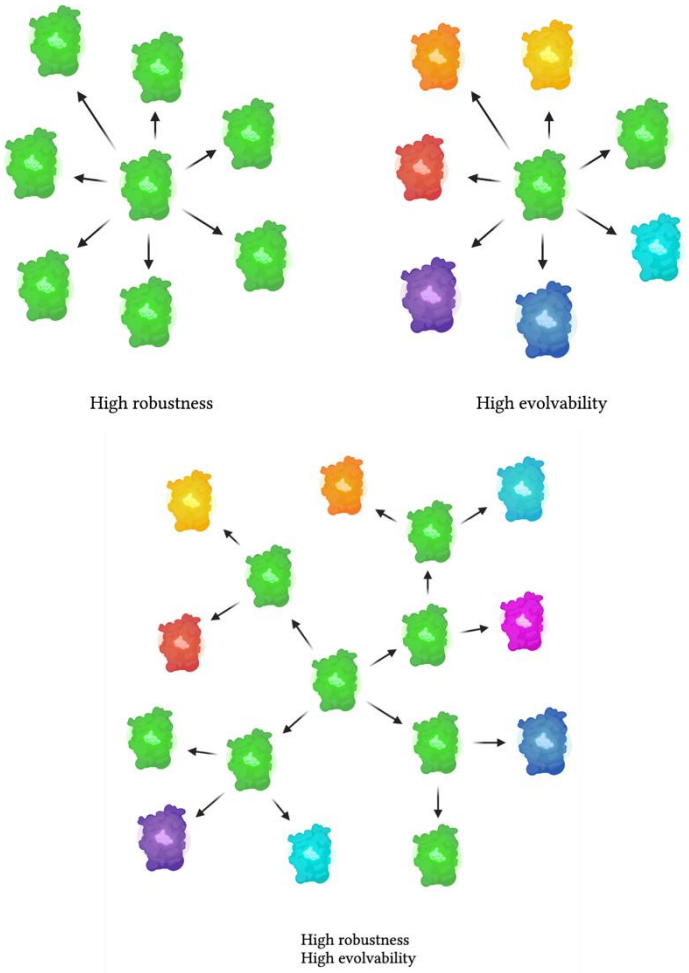



Figure D.3.1: Schematic for the representation of robustness and evolvability of proteins.
Let us consider a system (protein at the centre) exposed to changing conditions, such as new viable variants appear. For proteins, mutational robustness implies minimal to non-existent modification of the function (high robustness case). In the opposite case, novel functions can be generated from the initial phenotype quite easily (high evolvability case). When proteins show both high robustness and evolvability, the fitness landscape is shaped as a plateau (neutral landscape), and both cases are merged. While the population can change genotype along the way, phenotypes do not necessarily change as well, as showed by the potential change of colours.

The relationship between robustness and evolvability is crucial in our understanding of evolution, as it dictates the manner in which random mutations impact protein function, fitness, and to an extent any novel properties that could arise as the result of these variations. Indeed, in order to survive and reproduce under changing environmental conditions, organisms must maintain their functions, most notably at the biochemical level, which is reflective of their robustness to mutations. On the other hand, living systems also need to adapt to new conditions, need to be alterable and to develop original functions in order to survive on the long term, a potential for change indicative of their evolvability, which also needs to be maintained to some extent in a population.

As such, there would seem to be a conflict between robustness and evolvability, as the former would tend to limit the scope of phenotypical effects provided by mutations, diminishing the observable diversity in a population and thereby its potential to exhibit novel, viable and adaptive modifications. However, several recent works showed that phenotypic robustness may actually be linked to increased adaptability[163,164]. Admittedly, populations that display stronger phenotype resilience against perturbations would tend to navigate flat fitness landscapes, also called neutral networks[165]. Such plateaus are characterised by very similar phenotypes, although the genotype space can be quite vast. Each variant that composes this landscape is separated from the others by neutral point mutations, resulting in a wide and flat topography. Moreover, through the selection of promiscuous activities displayed in the population, random sampling of this considerable genetic diversity allows for a higher likelihood of novel functions emerging.

In essence, populations showing high phenotypic robustness also tend to be very adaptable – *i.e.* evolvable – in the long term, because of this tremendous space of phenotypically similar but genotypically distinct configurations that can be freely explored.

In the case of our experimental platform, which aims at studying the effects of additional, artificial translation noise on protein evolution, we thus expect to confirm this trend. By inserting non-heritable phenotypic variation in a large population of DNA polymerases that are selected for their ability to survive and reproduce (auto-replication), we anticipate to select for conservation of this auto-replicative activity, which is linked to a resistance to noise/phenotype robustness in our population. Along relatively high mutation rates (compared to natural conditions), our libraries of variants should converge to these plateau-ish landscapes, *i.e.* neutral networks. Although experiments have been conducted in order to collect data from experimental evolution under high phenotypic noise of proteins like cytochrome P450s[166] and β-lactamases[167], no similar studies have been conducted towards DNA polymerases to our knowledge.

Therefore, we expect that by selecting under high phenotypic noise, we indirectly select for robustness, and thus for evolvability, a very interesting characteristic for DNA polymerases, fundamental molecular biology tools that are notoriously challenging to modify and to adapt to one's will. Such populations of polymerases could then be selected towards other stimuli, such as even hotter environments, or activity in presence of small inhibitory molecules.

Still, we ought to mention several endeavours that began paving the way towards a better understanding of how phenotypic noise affects biological systems. Most notably in the case of transcription[168,169], it has been shown that activity of error-prone RNA polymerases can lead to improved robustness for downstream translated proteins. As for the translation, modularity of ribosomal accuracy can be achieved through targeting various subunits of the protein complex[170,171], leading to a spectrum of differential effects for protein expression. The investigations aiming at modifying the architecture of the translation process, by either insertion of non-canonical amino acids loaded onto tRNAs[172,173] or alteration of the genetic code itself[174], are also crucial to our global knowledge regarding more conceptual and complex engineering of such systems. These endeavours, much like ours, try to probe into the systemic aspect of biological processes, in order to examine the interactions between each component, and their effects towards fundamental inquiries such as the emergence of novel catalytic properties for enzymes, evolutionary trajectories for populations, etc.

# An *in vitro* directed evolution platform for KlenTaq DNA Polymerase

Before getting into the experimental details of the platform, we will first present an overview of the method. Inspired by the CSR of Holliger[130] that was previously described, the first iteration of this process in our laboratory was designed by Adèle Dramé-Maigne[175], using the expertise in molecular programming that was brought back from Japan by Yannick Rondelez[176,177], in order to broaden the self-selecting CSR concept to other enzymes than only DNA polymerases. This experimental process, just like the original CSR, was only semi-*in vitro*, using *E. coli* bacteria to transform the genes of interest, and to translate them into the proteins that would be subsequently trialled for their properties.

In parallel, another fully *in vitro* version of this method was developed by Rémi Sieskind[178], as a possible solution to bypass the constraints and limitations of using live organisms in the endeavour of directed evolution for enzymes. Indeed, the absence of any transformation step would necessarily improve the throughput of the method, along with removal of any biases that would only pre-select the variants that are not toxic for their hosts. Overall, this transition from partially to fully *in vitro* would therefore mainly help in a more comprehensive exploration of the enzymes' fitness landscapes. However, these improvements come at a cost, as replicating what nature does is always more complex and tiresome than using it in the first place. We will explore these numerous challenges in the following chapters.

My project itself started during this extensive development of the experimental platform, and I had the opportunity to help the instigator, Rémi Sieskind, in optimising the numerous processes and steps. Although our objectives with the platform were not the same, a vast majority of my work overlaps with his, and should thus be considered as an extension of the tremendous effort that birthed this set-up. In terms of goals, my project went back to the original purpose of Holliger, *i.e.* the self-selection of DNA polymerases. Indeed, one of the main advantages of working with a completely *in vitro* set-up was to explore how molecular evolution could operate when decorrelated from *in vivo* biases. In our case, it was studying how noise perpetuates through generations, and how it influences the selection of specific characteristics in enzymes, and such experiments are not feasible if using living systems like bacteria or eukaryotic cells.

The target of the study is the KlenTaq polymerase, an exonuclease deficient derivative of the Taq DNA polymerase, originally extracted by Chien et al. in 1976 from the thermophilic bacterium *Thermus aquaticus*[15]. Similar to the Klenow Fragment produced by the truncation of the DNA pol I from *E. coli*[179], KlenTaq is a N-terminally truncated Taq polymerase. The first 278 amino acids are absent, and the protein therefore lacks the 5'→ 3 exonuclease activity of its original counterpart[180]. Moreover, KlenTaq shows improved fidelity and thermostability[181] relative to the wild type, as well as enhanced resistance to PCR inhibitors found in whole blood or soil samples[182].

This property was actually the reason this enzyme was used by Adèle Dramé-Maigne in her PhD thesis[175], in order to develop a CSR-like process, as successful amplification in bacteria lysate was not reproducible with the regular Taq polymerase. Although similar toxicity issues were not particularly present in the *in vitro* setting developed by Rémi Sieskind[178], the cell-extract nature of our cell-free protein synthesis system supported the continuation of KlenTaq usage in our platform, as residual components of such mixtures could still impede the selection process. Moreover, in an effort to be able to compare the efficiency of the whole *in vitro* directed evolution process to that of the *in vivo*, *E. coli*-based platform, keeping the same target was preferable.

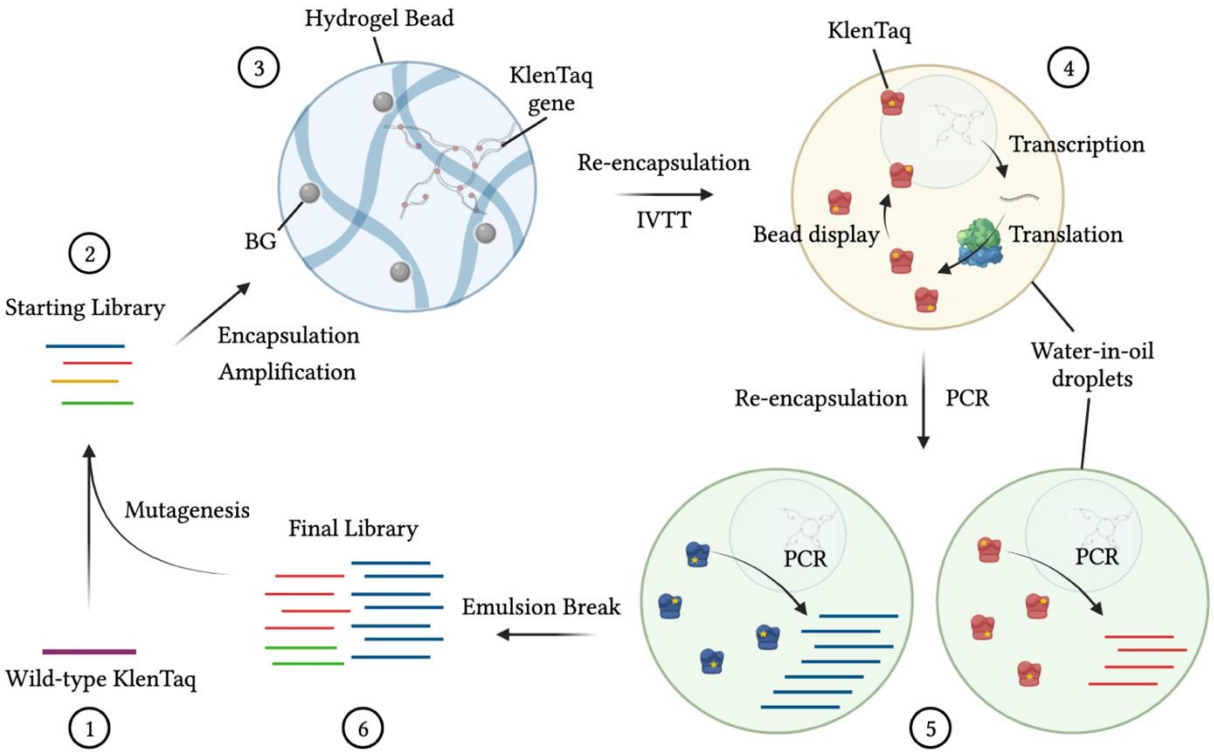



Figure 1.0 : Schematic for the *in vitro* directed evolution of KlenTaq DNA polymerase.
Starting from the wild-type KlenTaq gene (1), a highly diverse library is created through error-prone PCR (2). Each variant is circularised and encapsulated with the proper reactants in order to perform a Rolling-Circle Amplification (RCA), while a matrix of hydrogel polymerises in the droplet (3). Once the RCA is done and the hydrogel is fully formed, the emulsion is broken, beads are washed and re-encapsulated along with an IVTT (*In Vitro* Transcription Translation) Mix. The DNA polymerases cognate to the amplified genes are thus expressed *in situ*, and through SNAP-tagging these are covalently bound to the Benzylguanine (BG) moieties previously inserted in the hydrogel matrix. Using aminoglycoside antibiotics to interfere with ribosomal accuracy, additional mutations are inserted at random in the protein sequences, represented by the star symbols on the proteins (4). After subsequent emulsion break and beads wash, these are re-encapsulated with PCR buffer and adequate primers in order to perform compartmentalised self-replication (5). The better the enzymes, the greater amount of their genes are synthesised, leading to an enrichment of the best variants in the final library obtained after PCR and emulsion break (6). A fraction of this library can then be used for analysis after each round, through qPCR or Next-Generation Sequencing (NGS), while the rest is re-injected into another round of diversification, expression and selection.

## 1. Diversification

At the basis of any directed evolution process is the insertion of mutations, similarly to how nature operates in living systems. Indeed, evolution is a driving force that can only act on the diversity present in a population, in order to select and promote the fittest variants through generations. In our experimental platform, this diversity is artificially inserted through random mutagenesis, and more specifically through error-prone PCR. To this end, the gene of interest (KlenTaq in our case) is amplified by the Taq polymerase, but its fidelity is impaired through several different means. Some factors that help in producing mutations while replicating DNA are, non-exhaustively: increasing the concentrations of DNA polymerase, metal ions cofactors (notably $Mg^{2+}$, but adding $Mn^{2+}$ also works), mutagenic dNTPs analogs, as well as increasing the extension time of the PCR protocol[183].

However, one very common issue of such protocols lies in the multiple biases of nucleotide substitutions that arise at the end of the mutagenesis. Indeed, libraries of variants produced with these methods usually contain a greater amount of A→T and G→C substitutions, eventually increasing the GC content of the genes making up the library. Although the industrial random mutagenesis kits that are nowadays developed mostly even out such biases, the process is still unperfect, and such considerations should always be taken into account when proceeding to create libraries of variants using error-prone PCR.

With this in mind, we cloned our KlenTaq constructs in pIVEX vectors, optimised for *in vitro* expression. To this end, a T7 promoter and a RBS motive are necessary before the beginning of the gene. At the end of the KlenTaq sequence, the SNAP-tag gene is also inserted in the constructs, so that every variant can be put through our bead display system. Separated by a linker, the SNAP-tag moiety can be cleaved from the protein of interest using the Thrombin protease. Finally, the T7 terminator is set after the SNAP-tag sequence. Apart from the KlenTaq gene, the entire construct is standardised, so that every variant of the library can be inserted into the same circularised constructed (Figure 1.1). Indeed, the variants need to be circular for the RCA step to work, a necessary amplification process for proper transcription and translation of the genes afterwards. The insertion of the mutants resulting from the mutagenesis is operated through Gibson assembly[184], an isothermal reaction that allows the reunion of several DNA fragments, here our KlenTaq variants and standardised backbones.

These processes are set in order to perform the actual cycles of directed evolution, but on the road to getting that platform properly working, we simplified this system of libraries to a "mock" equivalent. Instead of an ensemble of several hundreds of millions of variants with their individual characteristics, we first optimised the experimental setup with a Yes/No library. Indeed, to assess the efficiency of the differential amplification that we are aiming towards, beginning with a library composed of only variants - an active one (KlenTaq) and inactive one (inactivated KlenTaq or inert protein in PCR like GFP) – is a much simpler system to handle experimentally. Starting from a mixture largely made up of inactive variants (90% inactive / 10% active), we expect the proportions to flip after a round of selection, if the amplification of the most active variants of the library is indeed more efficient in our system.

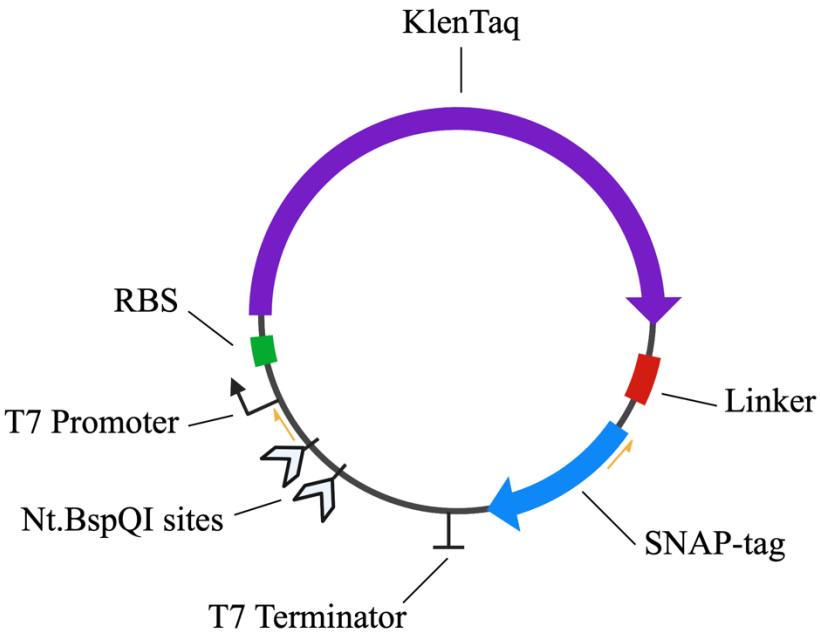



Figure 1.1: KlenTaq construct for *in vitro* directed evolution.
After random mutagenesis, the KlenTaq gene is circularised by Gibson assembly for the subsequent RCA process. Apart from the polymerase gene itself, the rest of the structure is constant, and the same for every variant of the library. Each mutant is simply inserted in this backbone before being reinjected in a cycle of directed evolution. In orange are represented the two primers that are used to perform the self-selecting PCR step at the end of the process, the forward right before the T7 Promoter, and the reverse at the beginning of the SNAP-tag gene. As such, only the variants that conserved the tag are allowed to amplify themselves.

Although we mainly used GFP as our inactive variant - most notably because the protein is much easier to observe through its fluorescence, which is useful for assessing its efficient cell-free expression for example – we still created a non-functional KlenTaq mutant, so that the differences between the two versions of our target protein are as limited as possible in a further test of our platform. To this end, a single amino acid change in the active site is sufficient to remove all catalytic activity from the KlenTaq polymerase: Asp332Gly, simply replacing the GAC sequence by a GGC through site-directed mutagenesis[185]. This variant's inability to amplify DNA has been tested, compared to its active counterpart, so that the inactivation has been conducted successfully.

## 2. Hydrogel Beads Generation

### 1) Hydrogel formulation

The second step is the encapsulation of the numerous variants of the KlenTaq gene obtained from the error-prone PCR in hydrogel beads, and their respective amplification within these compartments. To be able to independently assess the fitness of each variant of the library, it is necessary to isolate them from one another. In the case of Holliger's CSR, bacteria fulfil this role, but in a fully *in vitro* set-up, this is made possible through the synthesis of artificial compartments. The process was already set-up by Rémi Sieskind when I arrived in the laboratory, and only few adjustments had to be done to adapt it for my endeavours. Based on the works of Thiele *et al.*[186], protocols were reproduced in order to develop the hyaluronic acid-based hydrogel beads.

Hyaluronic acid, also known as hyaluronan, is a linear disaccharide polymer, made from the successive repetition of uronic and amino sugars, namely the D-glucuronic acid and the N-acetyl-D-glucosamine. These polymers usually contain several tens of thousands of disaccharides repeats, to sizes of up to millions of Da. *In vivo*, the polymer has many crucial roles either structurally in connective, epithelial, and neural tissues, or for its properties towards complex processes such as cell migration, skin healing and wound repairs. This polymer is thus highly biocompatible, and as such grew quite popular in recent years in light of its possible applications in medical research, most notably when its properties are modified to fit the need[187]. Strategies for additional crosslinking of this biopolymer allow for a finer control of the hydrogel properties, and in our case, the physico-chemical characteristics of our compartments.

Monomers of the polymer are thus to be thiolated in order to perform crosslinking between chains. To this end, solid hyaluronic acid (50kDa in average) is diluted in MES buffer, and several compounds are added so that the thiolation takes place (Fig. A.2.1). Once the reaction is over, dialysis is performed to purify the thiolated product and the latter is subsequently assessed through an Ellman's test, as to quantify the fraction of free thiol groups in the polymer, *i.e.* its ability to crosslink. Ellman's reagent (DTNB) reacts quickly with thiol moieties, the resulting compound of yellow colour absorbing in visible light. Based on Rémi's experiments, a reaction time of 6h seemed to yield the expected 20% of free thiols that were reported in Thiele's paper.

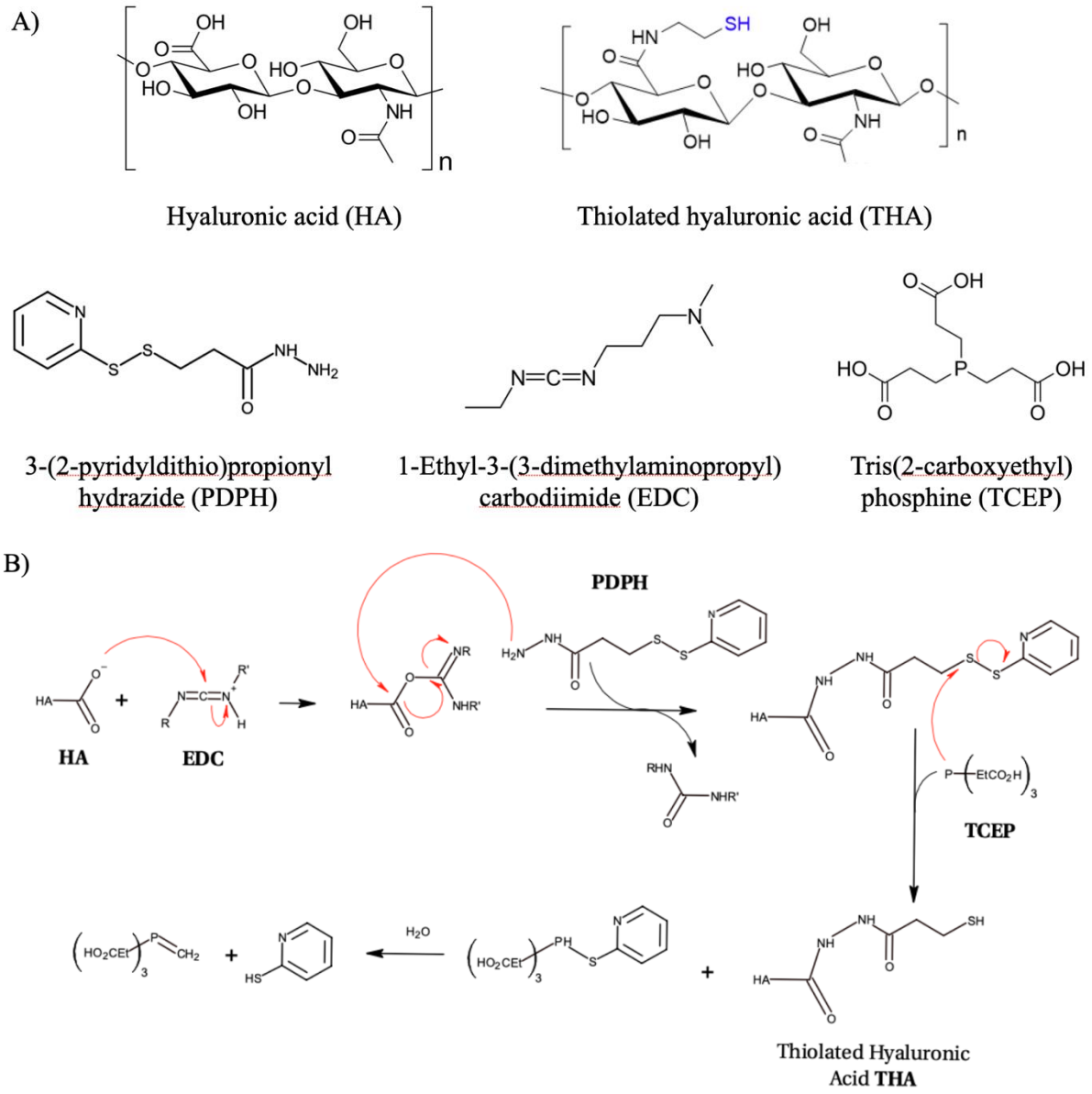



Figure 2.1: Thiolation of hyaluronic acid polymer.

A) List of reagents used to generate the hydrogel beads. The EDC is a carbodiimide used as a carboxyl activating agent for the coupling of primary amines, in conjunction with PDPH, a crosslinker used to reversibly conjugate sulfhydryl groups to carbonyl moieties. Lastly, the TCEP is a reducing agent, used to break disulfide bonds. At the end of the reaction, around a fifth of the HA polymer repeat units are thiolated. B) The reaction mechanism for the thiolation of the hyaluronic acid polymer. By-products of the reaction are purified through membrane dialysis.

The strategy of crosslinking is then based on the Michael reaction between our newly formed THA (Thiolated Hyaluronic Acid) and PEG-DVS (Polyethylene glycol divinylsulfone), one molecule of the latter reacting with two different thiol moieties, effectively acting as a bridge between polymer chains. Faster kinetics and larger resulting pores defined the choice of PEGDVS[188] among other possible crosslinking reagents, such as PEG-DA[186] (Polyethylene glycol diacrylate).

However, the PEGDVS is not the only thing reacting with the thiol groups of our THA. Indeed, as we mentioned previously, in directed evolution experiments, a stable linkage between genotype and phenotype must always be present in the system if we are to efficiently select the most active variants of the library. In our platform, the hydrogel compartments are meant to provide this connection through bead display. Although we will properly address this step of the process in the next subchapter, we will simply explain the overall concept here, for it has consequences on the formulation of our hydrogel. Each of the KlenTaq variants is expressed fused to a SNAP-tag, which can react with many different substrates, but in our case with Benzylguanine-maleimide (BG-Mal) moieties previously bonded to the surrounding hydrogel matrix. Thanks to a similar chemistry to that of the Michael reaction, the THA can react with maleimide groups to form covalent bonds, allowing us to place “protein hooks” in the compartments, effectively immobilising the DNA polymerases in the gel after being translated. Therefore, during the preparation of our THA-based hydrogel, various concentrations of BG-Mal are added to the reaction mix, depending on the final concentration of protein that ought to be obtained in the media for the selection step. It is noteworthy to mention that the concentration of BG-Mal used is not reflective of the concentration of protein displayed on the hydrogel beads, it only sets the maximum concentration of DNA polymerase attainable in the compartments. Moreover, in order to lessen the fluctuations in protein expression between compartments, a lower concentration of BG-Mal would help in that regard. In any case, the amount of thiol groups in the hydrogel made to react with BG-Mal groups is negligible compared to the total amount of thiol moieties left for crosslinking.

Finally, one crucial property of this hydrogel is to break down when heated during the final PCR step of our process, eventually becoming a low-viscosity aqueous solution that can be manipulated with micropipettes without issue. The DNA contained in this solution can then be gathered, purified and quantified, in order to assess the amplification ratios of each of the variant encapsulated in the hydrogel beads in the beginning of the cycle.

2) Microfluidic-based generation of beads

In the context and constraints of our experimental system, microfluidic-based techniques have been chosen to perform the encapsulation step of our process, most notably because of their high throughput and monodispersity. Moreover, the generated droplets can then be used as scaffolds for the hyaluronan hydrogel, in order to form distinct compartments for every variant of the library.

To this end, numerous microfluidic chips were developed by Rémi Sieskind. However, in order to work efficiently with the greatest library we could, the best option proved to be an adaptation of the millipede, a microfluidic device able to produce droplets consistently with a very high throughput and monodispersity[189]. This device is characterised by its large central channel for the aqueous phase, and the two smaller channels on its sides for the fluorinated oil phase transporting the emulsion (Novec™ HFE-7500 3M Engineered Fluid with 2% FluoSurf surfactant). Between those, several hundred V-shaped microchannels (around 500) are used to generate the water-in-oil droplets (Fig. A.2.2).

However, several modifications had to be enacted in order to use it in ideal conditions. The geometry of the device was already optimised when I arrived, leading to the desired size of around 15µm for the droplets. Moreover, the fast kinetics of crosslinking between THA and PEGDVS often clogged the capillaries of the microfluidic chips in minutes at room temperature, requiring to set-up a two-aqueous-phase-inlet flow focusing architecture. The two phases are thus blended *in situ* at the last moment thanks to a mixing chamber with a magnetic stirrer placed before the nozzles. Additionally, working on ice or cooling pads seemed to be necessary in the case of droplet productions longer than an hour, for the same reasons. This process quickly proved to be much more consistent and efficient in order to generate a greater volume of beads.

Because of the differential viscosity between the THA solution and the PEG-DVS one, traditional pressure pumps were not ideal, usually leading to an inaccurate 1:1 ratio in the microfluidic device. Using a syringe pump to control the flowrate of both solutions also proved to be effective to achieve equal parts of each in the mixing chamber, while we kept control of the oil flow with a regular pressure pump.

In the end, we manage to produce emulsion at a rate of 6µL/min, which roughly makes $10^8$ droplets per hour, a sufficient volume to theoretically compartmentalise and test every variant of our library. This process follows Poisson's Law, as the probability of encapsulating a variant does not depend on anything other than the absolute concentration of variants in the hydrogel master mix. In our endeavour, compartments with more than one gene are highly detrimental to the selection efficiency, as fitter variants can then amplify unfit ones, nullifying the "purifying" nature of the successive cycles of directed evolution. In order to mitigate such cross-contamination cases, a Poisson's parameter λ of 0.5 is chosen, so that very few hydrogel beads present multiple variants. As such, most of the generated beads are empty, so that such events are exceedingly rare. However, because of the very high-throughput of the millipede device, we can afford this relative inefficiency without worry.

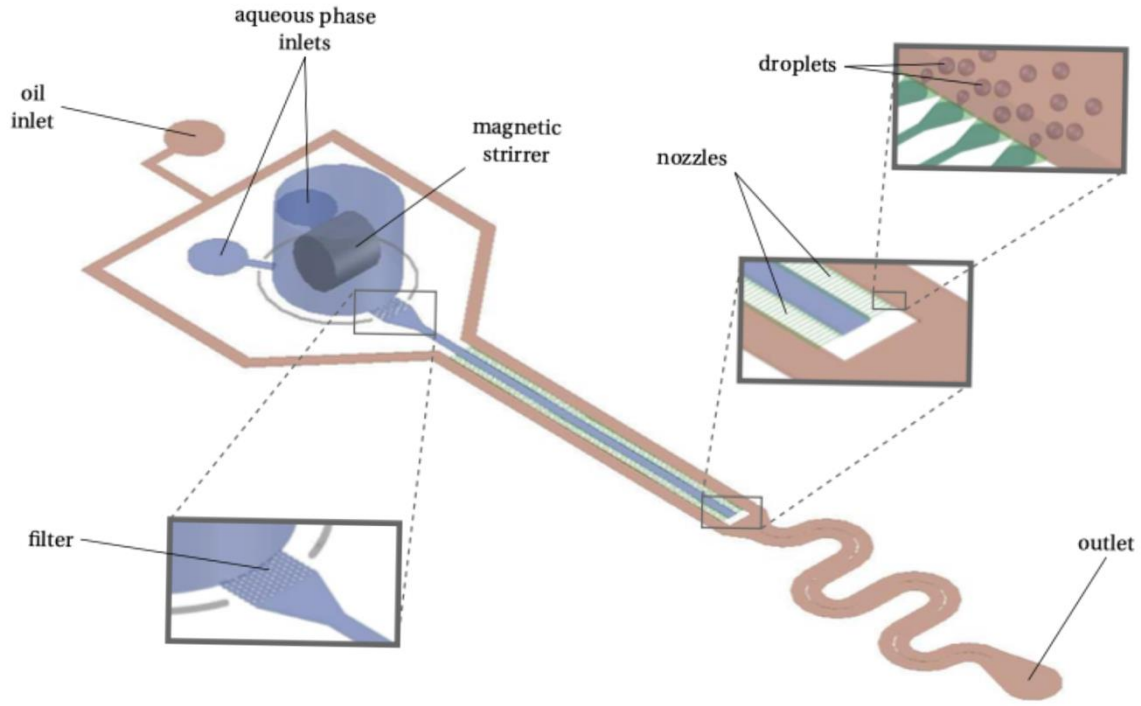



Figure 2.2: Millipede device isometric sketch.
In blue, the aqueous phase is made from the reunion of the two inlets (THA and PEG-DVS) inside the mixing chamber. Once near the numerous nozzles, monodisperse droplets are formed. In brown, the oil phase goes through the channels and carries the resulting emulsion to the outlet. Adapted from Rémi Sieskind's PhD Thesis[178].

## 3. Rolling Circle Amplification

Simultaneously to the reticulation process, an amplification step has to be performed on the individual gene variants that are compartmentalised in the hydrogel matrix. Indeed, a ~15µm droplet represents a volume of ~1pL, and the absolute concentration of a single copy of encoding DNA in each individual hydrogel bead is then roughly 1pM. However, the Cell-Free Protein Synthesis (CFPS) method that was chosen for our platform - the *In Vitro* Transcription Translation (IVTT) process that will be detailed in the next subchapter – requires around 1nM of genetic material in order to efficiently produce the corresponding proteins. A thousand-fold amplification is thus needed so that the subsequent steps can be functional. Many different amplification processes were investigated, but the one that fitted the most our process proved to be the Rolling Circle Amplification (RCA).

First, this isothermal reaction operates at 30°C, which is not detrimental to the structural integrity of the hydrogel, contrarily to most PCR-based amplification methods. Moreover, this temperature is also the optimal condition for the concurrent hydrogel reticulation.

Secondly, due to the 5'→ 3' strand-displacement activity of the φ29 DNA polymerase, the amplification of an initial circular gene results in a concatemer of copies of this template (Fig. A.3.1). Such condensed products – of around 1 µm in diameter at the end of the process – are too large to diffuse out of the hydrogel beads, ensuring functional monoclonality of variants in the compartments. Indeed, as stated previously, cross-contamination of gene variants between beads would result in drastic decrease of the selection efficiency of our cycles.

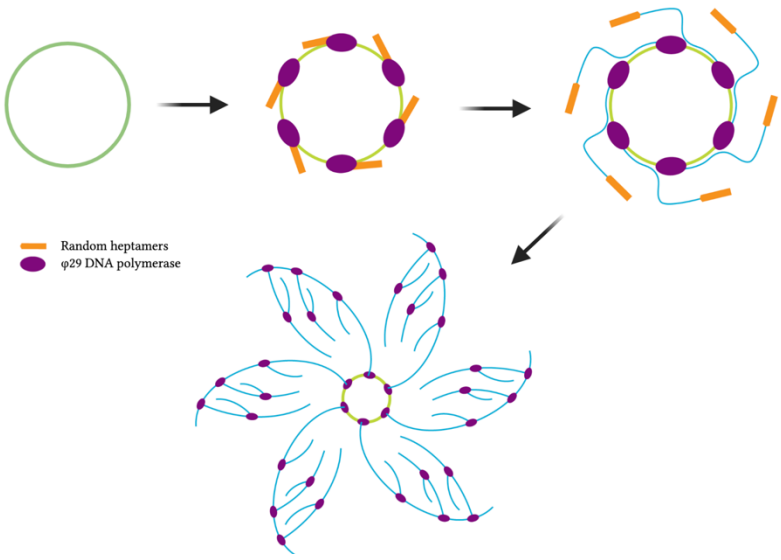



Figure 3.1: Rolling Circle Amplification reaction scheme.
In green, the starting template, our circular, twice-nicked KlenTaq gene. Random primers bind anywhere on the ssDNA and trigger φ29 binding. Due to 5'→ 3' strand-displacement activity of the DNA polymerase, φ29 can continue the replication, producing multiple linear copies still attached to the original template. Additional primers and polymerases can then bind to those single-strands, effectively leading to an exponential amplification and a highly-branched, star-like DNA product. Adapted from RCA DNA Amplification Kit description, Molecular Cloning Laboratories.

Although this additional genetic material is necessary for the expression step of our overall process, several issues still have to be taken into account. Most notably, the fact that this amplification would actually tend to flatten discrepancies between fit and unfit variants if still present in the compartments after the self-selection step. Indeed, let us consider a thousand-fold amplification of each unique, individual variant in the beads, and that those copies are themselves replicated based on the efficiency of their cognate polymerases. In the case of fit variants, we will end up with – for example - 10 times the 1000 copies, whereas unfit variants will keep the same amount of material. If we compare these two resulting quantities at the end of the process, we would get a 10-fold enrichment in the fit variant. However, if we remove the RCA product - the 1000 additional copies replicated from the original template in each compartment by φ29 – we would end up with a 9000-fold enrichment, thus massively improving the selection efficiency of the cycle.

To this end, the mix of dNTPs used for the RCA is modified so that the resulting amplified material can be selectively degraded by adding a solution of $I_2$ to the samples, similarly to the protocol described by Gish *et al*[190]. In our case, 30% of the dCTPs are replaced by α-S-dCTPs, 2'-deoxycytidine-5'-O-(1-Thiotriphosphate), that are incorporated just like regular dCTPs, but form phosphorothioate bonds. The Iodine is able to break down those bonds, so that the heavily-condensed genetic material in the beads is split into small pieces of dsDNA. Through gel purification, such fragments are easily separated from the starting template and its copies at the end of the self-selecting PCR step.

At the end of the process, after incubating the samples for 3h at 30°C, the RCA is over and the hydrogel reticulated. The emulsion is broken with 1H,1H,2H,2H-perfluoro-1-octanol and beads are collected from the aqueous phase, the oil droplets acting as scaffolds for the hydrogel. The RCA products can be seen inside the hydrogel beads using a dye that becomes fluorescent when bound to dsDNA (Evagreen) and fluorescence microscopy, confirming that the KlenTaq genes are effectively amplified by φ29 (Figure A.3.2).

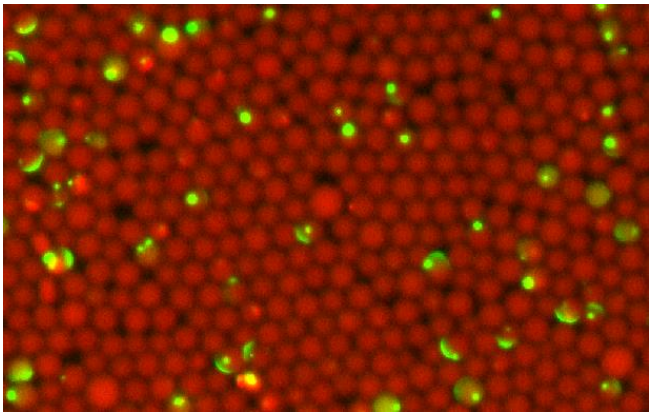


Figure 3.2: Hydrogel beads and RCA products.
Microscopy picture of THA beads after overnight incubation at 30°C. Cy5 red fluorescent dye is previously bound to the hydrogel, while Evagreen dye is used to tag the RCA products.

## 4. Protein Expression and Bead display

Once the beads generation and clonal amplification is over, the following step consists in the *in vitro* expression of the corresponding proteins, in the compartments, from the RCA product, and their subsequent binding on the hydrogel beads. Indeed, linking the genotype (amplified variant) to its phenotype (translated cognate protein) is necessary for artificial selection to operate. We will first present the bead display method.

### 1) The Genotype-Phenotype linkage

As we mentioned previously, we need to obtain distinct compartments that contain both the gene variant and the corresponding proteins. To that end, we relied on the SNAP-tagging method, a protein fragment that can be fused to any other protein. This 20kDa tag was obtained through directed evolution of the human $O^6$-methylguanine-DNA methyltransferase enzyme (MGMT)[191,192]. Originally crucial for genome stability and DNA repair in mammals, this engineered version can specifically and covalently react with benzylguanine derivatives, usually themselves linked to fluorescent probes, other proteins, etc. This tagging method presents several advantages, namely: its versatility and fast reactivity, largely independent of what is bound to the benzylguanine moiety; its chemical inertness towards other proteins and biomolecules; and its permeability to cell membranes, allowing intracellular tagging.

In our case, we use SNAP-tagged versions of our KlenTaq variants, so that they can react with $O^6$-benzylguanine-maleimide groups that are bound to the hydrogel matrix during the bead generation step. As the maleimide moieties are covalently linked to the free thiols of our hydrogel, the KlenTaq-SNAP proteins also end up bound to the bead matrix (Figure. 4.1).

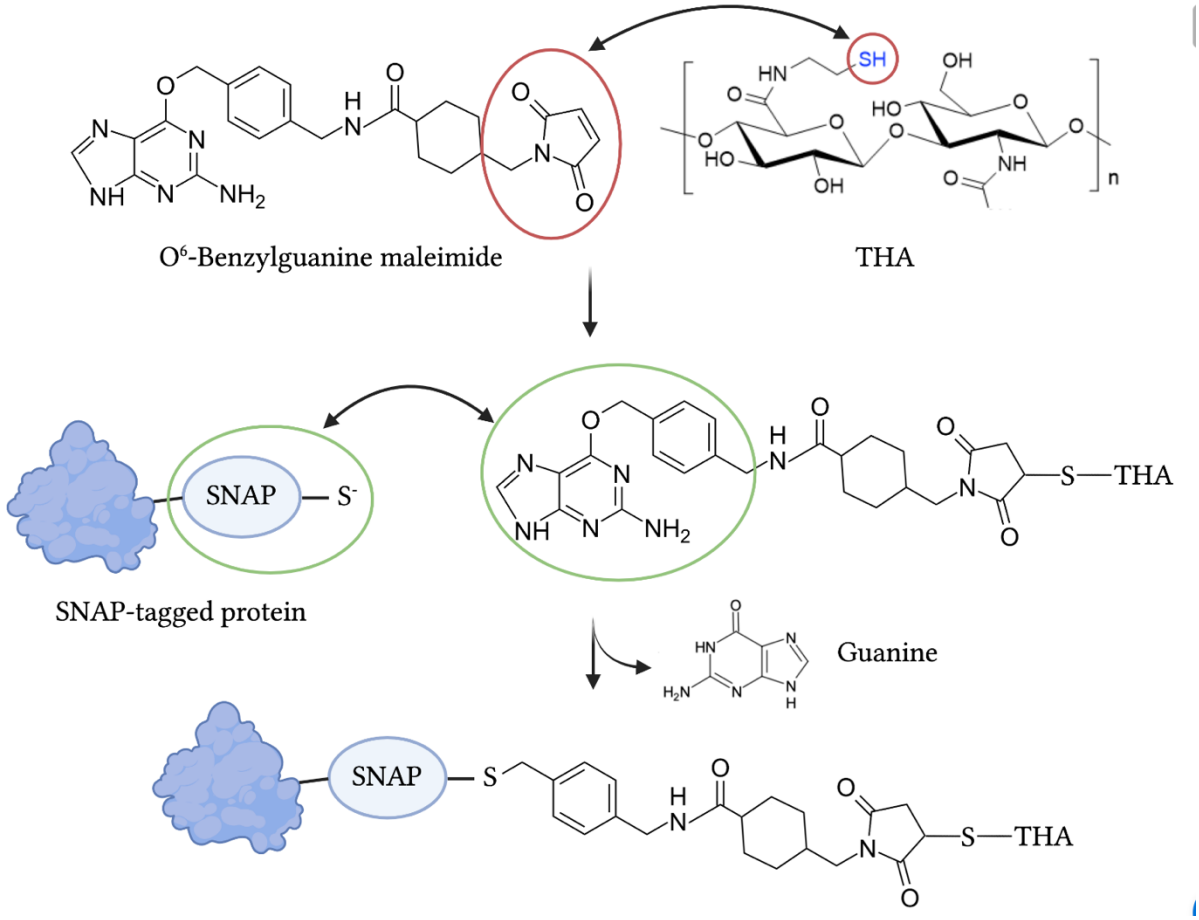



Figure 4.1: Bead display reaction scheme.
First, the free thiols in our THA polymer react with the $O^6$-benzylguanine-maleimide (BG-Mal) moieties, in order to provide the hydrogel matrix with "protein hooks". Once proteins are translated, the SNAP-tagged constructs can bind to the free benzylguanine, releasing guanine molecules in the media, and resulting in a covalent bond between the protein of interest and the hydrogel bead.

### 2) *In vitro* KlenTaq expression

The Cell-Free Protein Synthesis (CFPS) system that was chosen in order to express the studied KlenTaq proteins is based on cell extracts[193], for several reasons. First, this method is drastically cheaper than the other technique, PURE, that uses purified recombinant proteins[194]. Indeed, the cell extract can be prepared in our laboratory, contrarily to the PURE system that must be bought from manufacturers (although home-made versions have been reported). Second, because of the higher yield and longer production times[195]. Following the optimisation of Filippo Caschera and Rémi Sieskind with the S17 Cell extract system, robust protocols were established for the *In Vitro* Transcription and Translation (IVTT) of the proteins of interest. Most notably, the working concentration of genetic material in order to efficiently express proteins was fixed around 1nM.

Starting from BL21*(DE3) *E. coli* strains, cultures are grown, lysed and centrifuged. This extract is then heat-incubated to precipitate superfluous proteins and degrade genomic bacteria DNA. In the end, this cell lysate contains the expression machinery from *E. coli*, albeit not completely pure. However, it is still adapted to the specific production of our proteins. By inserting a polyhistidine tag at the end of the proteins of interest, purification of IVTT samples through Ni-NTA chromatography columns showed that expression of SNAP-tagged KlenTaq was indeed possible, as efficiently as expressing SNAP-tagged GFP (Figure 4.2, Left). As for the expression of proteins in our final set-up, previous works have shown that IVTT production is viable with RCA products.

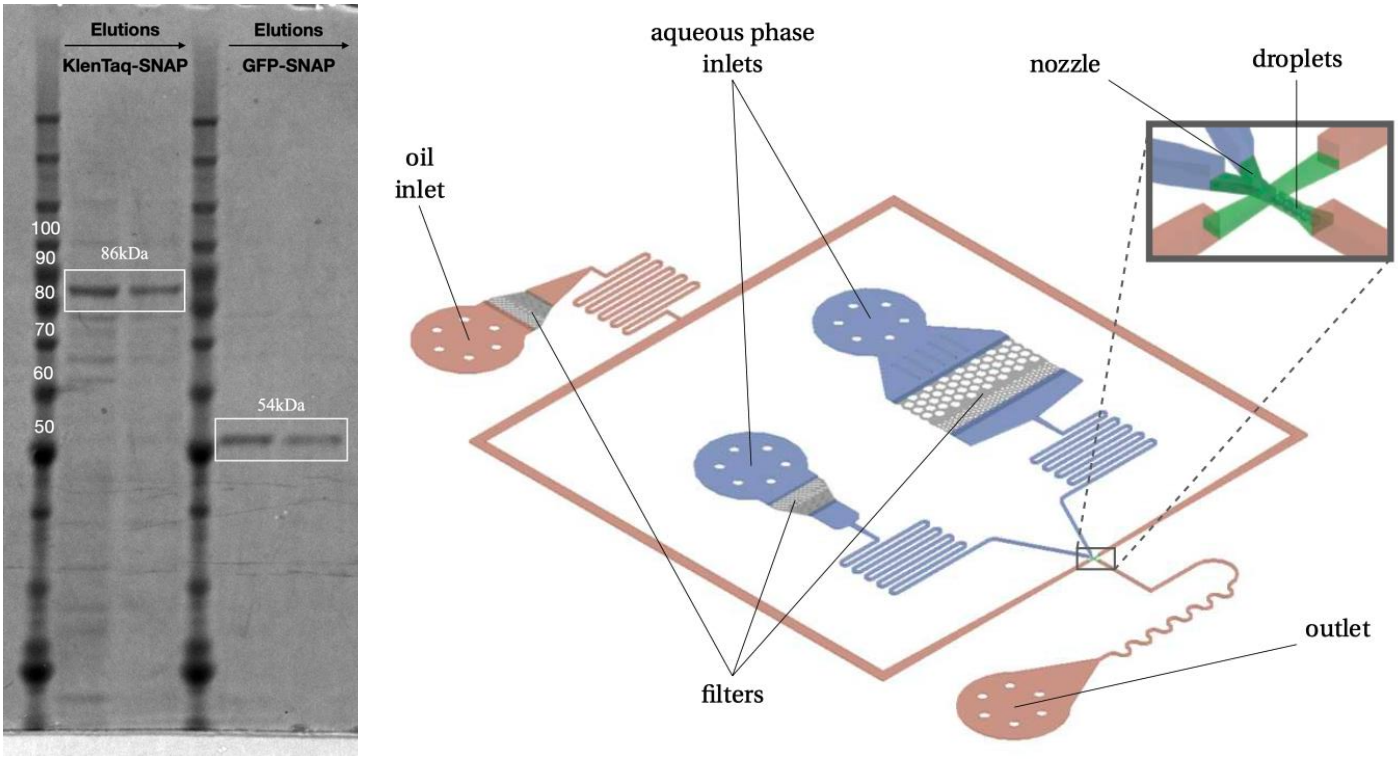



Figure 4.2:
Left: Purification of SNAP-KlenTaq and GFP-SNAP from IVTT.
From an IVTT mixture left overnight with the SNAP-KlenTaq and GFP-SNAP genes, the proteins were purified following a standard Ni-NTA column chromatography. Polyhistidine tags were inserted at the end of the protein sequences beforehand. Samples of the elution fractions were loaded on acrylamide gel for SDS-PAGE. Both proteins are obtained with concentrations within the same order of magnitude, indicating that GFP can safely be used as proxy for KlenTaq expression in IVTT. Protein ladder in kDa.
Right: Re-encapsulation device isometric sketch.
This microfluidic device was first developed by Shim *et al*[196], and then adapted by Adèle Dramé-Maigne[175]. The nozzle width and filter structure were later modified by Rémi Sieskind to accommodate the re-encapsulation of hydrogel beads. Adapted from Rémi Sieskind's PhD Thesis[178].

In order to preserve compartment clonality through the whole process, protein expression cannot be conducted in bulk, as it would mean cross-contaminating each and every distinct compartment with the contents of the others. The selective power of the whole process would be undermined, nullifying the entire endeavour. As such, the hydrogel beads containing the amplified variant must be re-encapsulated in droplets during the IVTT process. To this end, the microfluidic device developed by Shim *et al*.[196] was modified to accommodate the passage of hydrogel beads, so that ~30 μm droplets are created (Figure 4.2, Right).

### 5. Self-Selecting PCR

The last step of this whole process is the self-selecting PCR process. It is the most important part of the experimental platform, as it is at the end of the PCR that fit variants are more amplified than unfit ones, but also the hardest one to optimise. Indeed, the efficiency of this differential amplification depends on many parameters, which themselves are mostly based on the conditions of the previous steps, namely:

- The extent of the amplification in a compartment depends on the starting concentration of genetic material inside it. The duration of the RCA step is thus important towards controlling the initial – and final - amount of DNA in the hydrogel beads.

- The efficiency of the PCR itself depends on the concentration of DNA polymerases in the compartment. This, in turn, revolves quite evidently around the duration of the IVTT step, but also around the amount of BG-Mal inserted in the hydrogel matrix, both being critical towards the control of enzyme concentration, and thus subsequent success of the PCR.

After optimisation of the numerous previous steps in our cycle, we could then proceed to the assess if the final PCR step was indeed self-selecting for the best variants or not. As a starting experiment towards a proof of principle, we performed a mock selection from a library consisting of only two variants: the KlenTaq-SNAPf and the GFP-SNAPf constructs (see 1) Diversification).

Beginning the process of RCA amplification with a plasmid mix of 90% GFP-SNAPf / 10% KlenTaq-SNAPf, we expect to reverse the two proportions after a few cycles of selection. Indeed, during the final auto-PCR step, the GFP protein cannot replicate its own gene, contrarily to the active polymerase KlenTaq, leading to a progressive enrichment of the latter through the successive rounds of selection. Using qPCR with specific primers for KlenTaq and GFP, we can then quantify the amount of each plasmid present in the beads before and after the "auto-selection PCR". In order to alleviate eventual issues of co-encapsulation during the initial compartmentalisation before the RCA, and to present the beginning of a proof of principle for the method we developed, we started a cycle with two different batches of THA beads, one where the active variant (KlenTaq) was amplified by RCA, and the other where the inactive one (GFP) was as well. At the end of the process, the relative fractions of both variants were quantified (Figure 5.1).

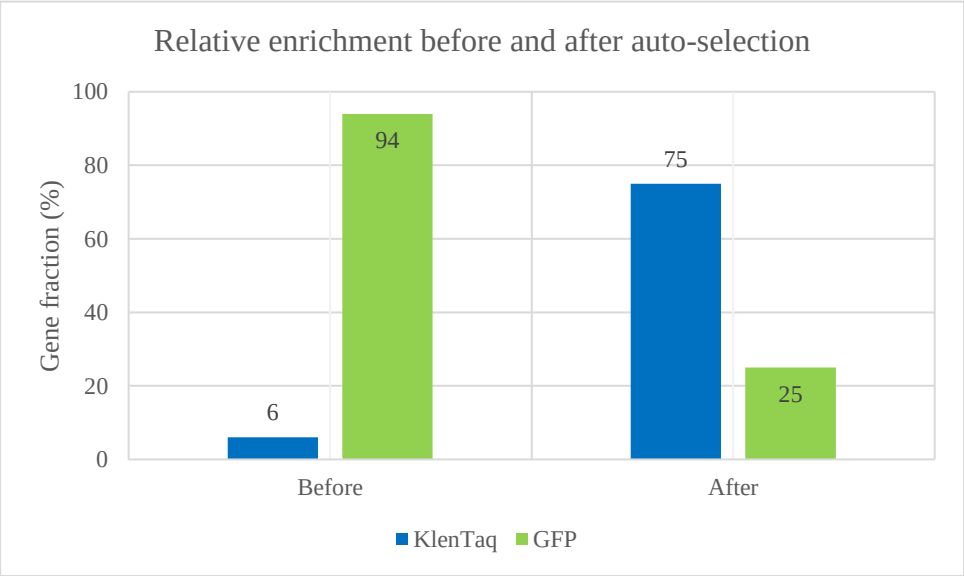



Figure 5.1: Fractions of active (KlenTaq) and inactive (GFP) variants before and after final PCR.
After the final PCR step, variants present differential amplification ratios based on their respective activity. Starting from an active/inactive ratio of 6/94, we end up with a 75/25 ratio at the end of one round of the whole process.

Although these results seem largely sufficient to perform proper selection - and thus subsequent evolution - on our polymerases variants, such selection efficiency ended up quite difficult to replicate when starting from an actual mix of KlenTaq and GFP plasmids, amplified through RCA and so on. Because we always record positive amplification for the inactive variants in such experiments, we suspect the first compartmentalisation step to be the source of the problem, albeit we did not manage to solve the issue at this time.

## 6. On the study of noise during protein expression

As we non-exhaustively mentioned previously, there are numerous mechanisms that life has developed in order to modulate, control or even dissipate the sources and effects of molecular noise. The aim would be to insert such noisy effects in our system through the addition of aminoglycoside antibiotics during protein synthesis, molecules that have been proved to interfere with ribosomal accuracy[197–199]. We expect that adding sources of phenotypic noise in the set-up would change the fitness landscapes and thus the evolutionary response of the system. In such conditions, we would predict that fitness is intrinsically linked to characteristics such as mutational robustness - the ability to withstand stress (sequence and structural changes) while maintaining proper function, in this case, self-replication – and indirectly evolvability, *i*.e. the potential of a protein to develop new and/or improved functions.

### 1) Aminoglycosides

In order to mimic the addition of phenotypic noise in our directed evolution system, we decided to use aminoglycoside antibiotics. This category of chemicals has been used for its medicinal and bactericidal activity against Gram-negative aerobic bacteria, most of them isolated from *Streptomyces* bacteria. Most molecules of this family exhibit an amino-modified glycoside moiety that plays a role in their activity, although one of the most famous antibiotic of this class, streptomycin, lacks this chemical group (Figure 6.1). Their bactericidal properties actually find their origins in the same feature that is of interest to us.

Although the exact mechanism is still unclear, aminoglycoside antibiotics are known as protein synthesis inhibitors, interfering with bacterial ribosomes. Binding to the 30S subunit, they disturb peptide elongation and translational proofreading, leading to higher translational inaccuracy, as well as a higher number of premature terminations[200]. Using such chemicals to simulate the addition of translational noise in our experimental platform thus appears well feasible, at least in theory. Moreover, being able to directly link a concentration of disruptive molecules to a number of mutations in the proteins resulting from the impaired translation is the closest thing to a “noise knob” that we could feature in our set-up, not to mention the quite simple investment that using such antibiotics represent compared to the rest of the platform. To this end, we began our assays with streptomycin and kanamycin, two of the most readily available and used antibiotics of this family.

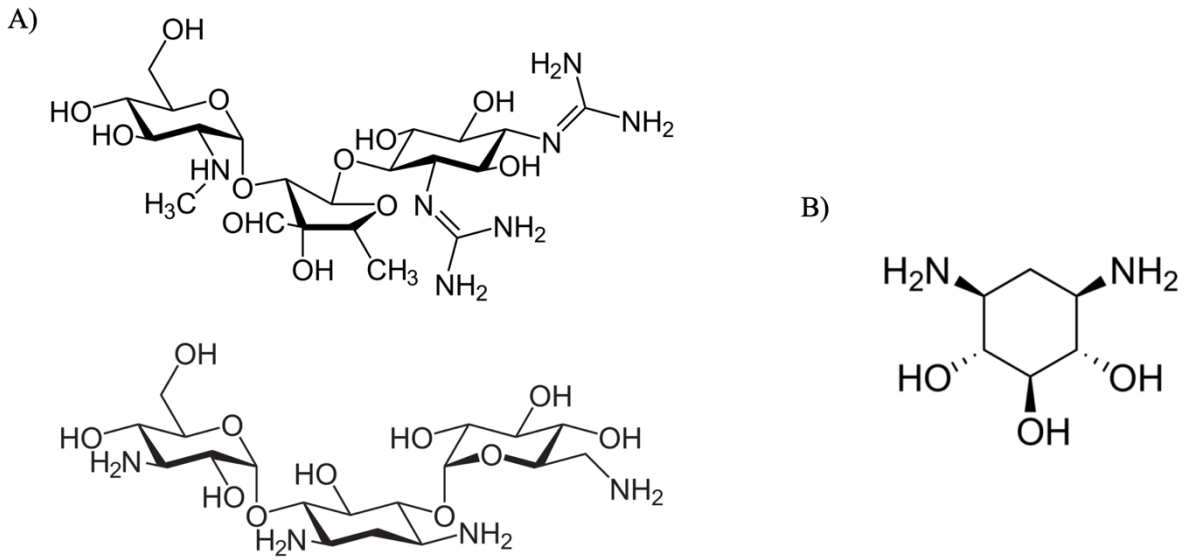



Figure 6.1: Aminoglycoside antibiotics chemical structures.

A) At the top, topological structure of streptomycin. At the bottom, topological structure of kanamycin. These antibiotics are very widely used as selective agents in cell cultures, as well as treatment for several types of bacterial infections. Both were tested and assayed for their potential use in our set-up.
B) 2-deoxystrept-amine moiety, present at the centre of the kanamycin molecule but absent in streptomycin. This chemical group has been linked to translational inaccuracies and ribosomal translocation inhibition.

### 2) Noise assay

In order to test the disruptive efficiency of these antibiotics on our IVTT cell-free protein synthesis system, we used the translation of GFP as a proxy. Indeed, any impact on the translational accuracy of the ribosomes would result in a lesser number of functional proteins, and being able to measure the expected loss of fluorescence in our samples is extremely convenient as a starting point of our investigation.

Given that such antibiotics are toxic for cells around the millimolar range (most notably when used as selective agents for cell culture), we set-up a range of kanamycin and streptomycin concentrations in IVTT samples to assess the magnitude of disruption in the system, namely from 5 nM to 5 μM. Dozens of these runs were completed, each time recording the fluorescence in real-time using a CLARIOstar Plus plate reader, because of the intrinsic variability of expression between experiments. Results are presented in Figure 6.2, where a snapshot of the recorded (normalised) fluorescence at the plateau of expression (Δt~8h) is plotted against the antibiotic concentration.

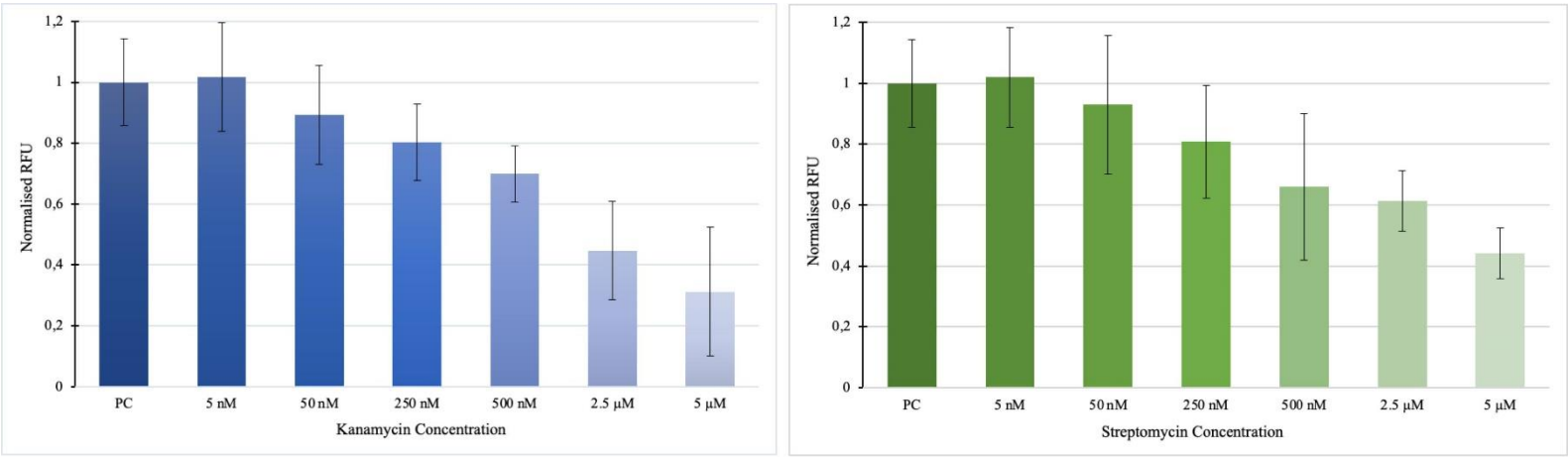
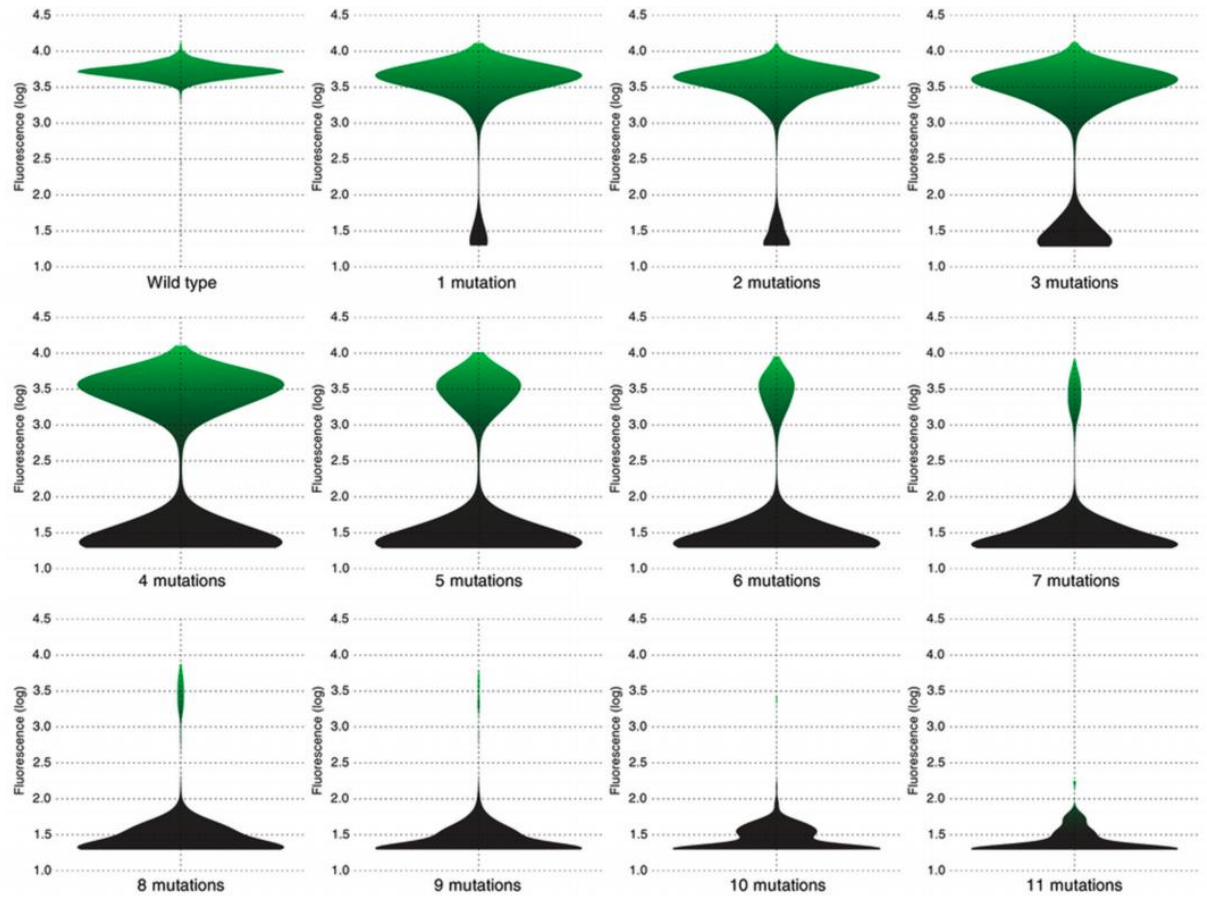



Figure 6.2: Effects of aminoglycoside antibiotics concentrations on GFP translation.

IVTT protein synthesis of GFP. Experiments were run with varying concentrations of aminoglycoside antibiotics: kanamycin (left) and streptomycin (right). The positive controls (PC) did not contain any antibiotic.

In both cases, the effect of the aminoglycoside antibiotic is pretty straightforward and monotonous: the greater the concentration, the greater the disruptive effect. Nonetheless, we decided to focus on the kanamycin for now, as the activity of both antibiotics seem to be quite equivalent, in order to alleviate some experimental constraints. Further study of the eventual differences between those two antibiotics - if there are any beyond the same disruptive effect on GFP fluorescence that was investigated here, notably in terms of the quality of the mutations produced in the translated proteins – could be interesting, in order to potentially develop different kinds of "noise knobs" for the system.

To begin a simulacrum of quantifying the amount of mutations that occurred during the translation of the GFP proteins in our samples, we based our considerations on the work of Sarkisyan *et al.*[201], a work that experimentally and theoretically investigated the local fitness landscapes of GFP, and particularly the effects of an increasing number of mutations in its gene towards its fluorescence. By comparing their results to our own, we managed to roughly determine what would qualify as "low", "median" and "high" noise regimes, depending on the proportion of fluorescence that was recorded compared to the controls, along with the approximative number of missense mutations that would suffice in order to gradually disrupt GFP fluorescence (Figure 6.3).




| Noise regime | Low | Median | High |
|---|---|---|---|
| [Kanamycin] | 1 nM – 100 nM | 100 nM – 2.5 µM | > 2.5 µM |
| % Fluorescence | % > 75 % | 25 % < % < 75 % | % < 25 % |
| # Mutations | 0 - 3 | 4 - 6 | 7 + |

Figure 6.3: Influence of the number of mutations on GFP fluorescence.

At the top, figure adapted from Sarkisyan *et al*[201] that relates the influence of missense mutations on the average fluorescence recorded in a GFP population. At the bottom, we transposed these results with our data, in order to establish the concentration ranges of several noise regimes: low, median, and high.

As we can see, the concentrations of antibiotic that were set in our experiments almost completely fill out the range of noise profiles that we can expect. However, an important take away of the various successful directed evolution experiments that were performed over time is that the number of mutations inserted during the diversification step must not be too high, as it often means losing hardly-acquired protein stability and enzymatic function. Therefore, we expect that a low regime of noise – around and below 100 nM in kanamycin - would be acceptable in the conditions of our platform, although higher concentrations could very well be used in separate experiments to assess the extent of such consistent disruption on the protein fitness landscapes.

3) Quantification of noise

Even so, one of the still obscure aspects of this noise assay is the quantification of mutations, and more precisely the frequency of premature translation termination compared to insertion of mutations. Indeed, the rough characterisation we established previously is solely based on the latter rather than the former, and we don't have any information on the proportion of early stoppage randomly happening through this antibiotic-induced increase of translational inaccuracy. To this end, we assessed and quantified the differential translation of two GFP variants, GFP- HisTag and HisTag-GFP, after affinity chromatography purification. Because of the small volumes of IVTT reaction mix that we use, protein concentrations that are gathered after elution are equally minute.

We thus decided to perform silver staining on polyacrylamide gels after SDS-PAGE, a highly-sensitive, easy to undertake, fast and cost-effective method for total protein quantification[202,203]. In ideal conditions, masses down to 0.25 ng of protein can be observed. Considering that with our native GFP in overnight IVTT and control-like conditions, we usually obtain dozens if not hundreds of micrograms of protein, this precision is largely enough for our purposes, even when accounting for eventual drastic impairment of protein translation. Similar to previous experiments, real-time fluorescence monitoring of the samples is also conducted with our CLARIOstar during IVTT in order to see if the kinetics are matchings the ones detailed previously. The whole process is represented in Figure 6.4.

In this experiment, the amount of purified GFP-HisTag is a proxy for the effective ratio of translation termination, as only the fully translated proteins that exhibit a functional polyhistidine tag can be purified. On the other hand, purified HisTag-GFP represents the amount of initiated protein, but not necessarily terminated. We expect that being able to assess the effects of increasing concentrations of kanamycin on the fluorescence of GFP (functionality) and on the mass of purified GFP (expression/stability) for both variants would be quite informative. Moreover, the evolution of the former would help us get a better idea of the proportion of inserted mutations – that would only disrupt fluorescence, but not translation continuation – while the latter would allow to roughly estimate the frequency of premature termination. Results are presented in Figure 6.5.

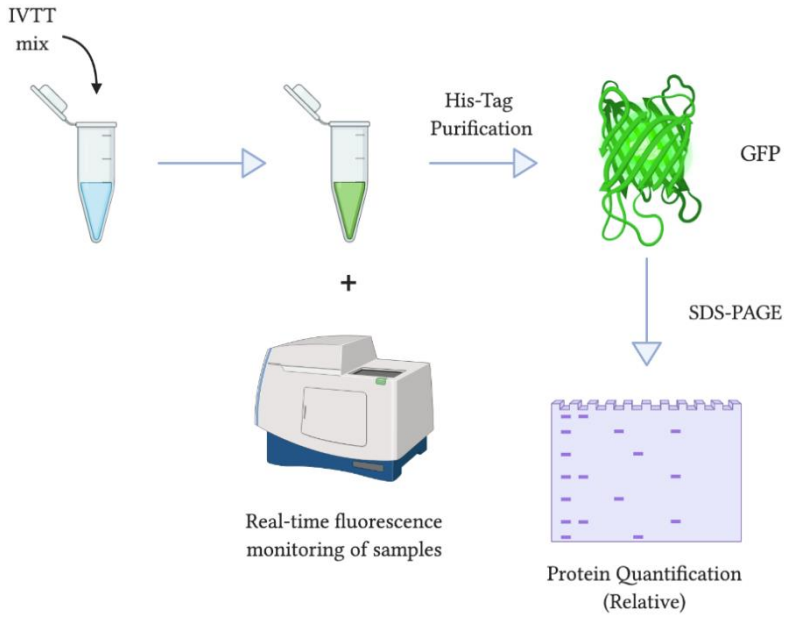



Figure 6.4: Process for protein quantification by silver staining.

Starting from IVTT mixtures with varying concentrations of kanamycin, GFP-HisTag and HisTag-GFP are separately expressed and purified through affinity chromatography columns. SDS-PAGE on polyacrylamide gels is then carried, before proceeding to the silver staining. Note that the protein quantification is relative to the standards of known protein concentrations loaded on the gels (BSA).

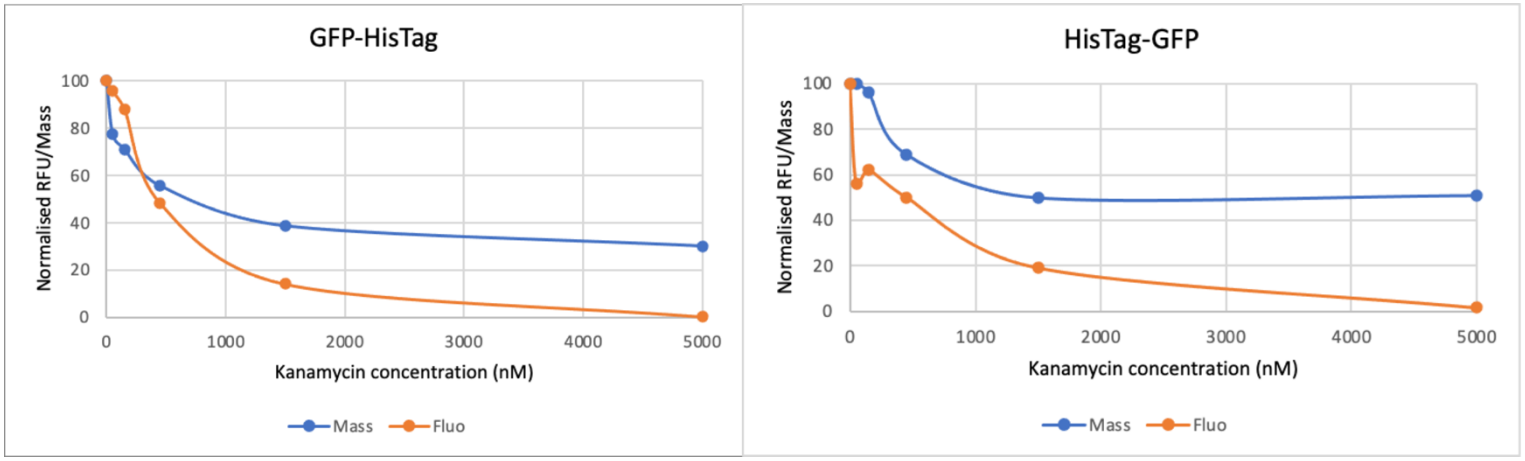



Figure 6.5: Influence of kanamycin concentrations on fluorescence and purified protein mass in GFP.

Starting from IVTT mixtures with varying concentrations of kanamycin, GFP-HisTag and HisTag-GFP are separately expressed and purified through affinity chromatography columns. Fluorescence levels are extracted from real-time monitoring during IVTT, while protein masses are quantified through silver staining. The “0” concentration in both graphs is the positive control without any kanamycin in the IVTT reaction mixture.

First, we can learn from the GFP-HisTag experiments that in a low noise regime (50 to 150nM kanamycin), there seems to be few mutations that disrupt fluorescence when inserted in the proteins. However, this trend flips around the median regime, to set heavily in the high noise regime, as the frequency of early termination seems to stabilise. Whereas the recorded fluorescence is almost null starting from 1.5 μM, there are still fully terminated protein being expressed in the IVTT (although non-functional). Secondly, in the case of the HisTag-GFP experiments, we can see that mutations are quite frequent starting from the median regime, as the sudden drop in the mass of purified proteins reflects random changes in the polyhistidine tag of the construct and its functional inactivation. We hypothesised that the higher amount of proteins (compared to the GFP-HisTag construct) obtained at higher antibiotic concentrations is due to a bias in the quantification method through silver staining, as constructs around the appropriate molar weight are also taken into account, even though they constitute a fraction of incomplete HisTag-GFP.

All in all, these experiments confirm the trends that were previously observed, but are still a form of relative quantification, the samples of interest being compared to the protein standards of known concentrations (dilutions of commercial BSA) that we load on the polyacrylamide gels before electrophoresis. In order to obtain additional information, we gathered that a much more precise and finer method would be necessary. To this end, we decided to work with Joelle Vinh's "Spectrométrie de Masse Biologique et Protéomique" laboratory at ESPCI, which specialises in particular towards protein Mass Spectrometry (MS). After a digestion treatment (based on the activity of the trypsin protease), proteins are broken down into smaller peptides that are subsequently injected into the mass spectrometer, producing a "peptide fingerprint" of the population of proteins. This technique allows for an extremely precise characterisation of protein residues, down to the mutations that could statistically arise in each protein fragment.

Although we managed to run a few mass spectres of our GFP samples that were translated in different regimes of antibiotic concentrations, at the time of writing this manuscript, we did not have enough time to perform more experiments and extract valuable information from the gathered data. Several adjustments of the purification protocols were necessary in order to fine-tune the experimental conditions of the MS experiments, as will be detailed in the next chapter. However, our IVTT and purification protocols still proved to be compatible with MS analysis, which is encouraging towards further steps of characterization in this regard.

# Discussion & Conclusion

We developed in this thesis an extensive framework based on the experimental platform previously designed by Rémi Sieskind[178]. With the aim of studying the influence of translational noise on protein evolution, we converted the methodology that he meticulously planned to our purpose, conserving some elements, changing and adding others.

In summary, this comprehensive strategy for the *in vitro* directed evolution of proteins begins with the creation of a large library of variants, subtle declinations of the original DNA polymerase of interest, KlenTaq. Inserting random point mutations in its sequence, we end up with a collection of mutants which are subsequently individually encapsulated in hyaluronan-based hydrogel beads, using very high-throughput microfluidic devices. During its cross-linking, an isothermal amplification process known as RCA is performed on each gene, inside their respective hydrogel compartment. Once fully reticulated, the beads are then re-encapsulated in a cell-free protein synthesis mixture established from *E. coli* cell-extract, in order to perform transcription and translation of the genes captured in each compartment. At the same time, the SNAP-tag that has been inserted inside the sequence of each variant beforehand is subjected to a bead display reaction, covalently bounding the resulting proteins to the surrounding hydrogel matrix. After gathering the ensuing water-in-oil emulsion, breaking it and thoroughly washing it, the beads are once again re-encapsulated but this time with a PCR mixture, so that every DNA polymerase variant contained in the hydrogel beads can perform the amplification of its own starting gene contained in the compartment, *i.e.* its self-replication.

At the end of the process, we thereby wind up with another library of genes, enriched in the variants most efficient at replicating their genes, the most active ones. This collection of DNA strands can then be analysed through sequencing, and utilised as a starting block for another round of directed evolution. Cycle after cycle, we expect to create and select DNA polymerases with new and/or improved structures, functions, etc. This platform presents several advantages to most directed evolution methodologies: firstly, its verry high throughput, that allows the evaluation and selection of several tens of thousands of variants at the same time. Secondly, its fully *in vitro* nature, that allows for an evolution unbiased by the eventual regulative systems that can be found *in vivo*. A direly needed factor towards the study of translational noise in protein evolution, as it would not have been possible to consider such elements in the presence of living organisms such as bacteria.

To this point, as a means to implement a “noise knob” onto this experimental platform, we explored the use of aminoglycoside antibiotics, chemicals that impact the ribosomal accuracy during the translation of proteins. As a first investigation of the mutation-inducing properties of kanamycin and streptomycin in our proteins, we studied the effects of these compounds on the expression of GFP, a useful proxy to assess the extent of missense of non-sense mutations, easily observable through the loss of fluorescence. Most notably, we established the profile of these error-inducing antibiotics in function of their concentrations, in order to get a better idea of the regimes of noise when using such chemicals. We then tried to assess more precisely the nature of these mutations, using at first relative quantification methods such as silver staining of polyacrylamide gels, and then started investigating using much finer methods based on protein mass spectrometry. Unfortunately, we did not have enough time to extract considerable amounts of data on this end.

All in all, the fundamental inquiry that is at the basis of this project is unfortunately still left unanswered, although much work has been done towards the concretisation and establishment of an experimental framework that could hope to investigate the complex processes that are nowadays under study. While there is much left to characterise on the matter of translational noise, we could envision that this directed evolution platform would be used to explore similar research topics as the one we set for ourselves. Albeit promising, one of the major limitations of the whole process lies in its many different steps, and the interplay between each and every one of them, making it quite difficult to evaluate the causes of failure when not thoroughly mindful of these interactions. Undoubtedly, finding a way to simplify and streamline the strategy as a whole - maybe through more efficient or less taxing compartmentalisation techniques instead of microfluidics - would certainly be a boon to the experimentalist.

# Step-by-step characterisation

In this chapter, the numerous investigations which were pursued in order to optimise each and every step of the process will be detailed. Most notably, an important part of the overall work was to harmonize the different reactions so that they could work without hindering the subsequent ones. As such, changes in the experimental protocols are presented in the chronological order of study, even though we were often required to go back and forth between the various stages of the process. Unless otherwise specified, the PCR experiments in the following chapter are always conducted on the KlenTaq gene.

## *A. In Vitro* directed evolution platform for KlenTaq polymerase

### 1. Starting from bacteria

Because of the previous works done in the laboratory with KlenTaq (by Adèle Dramé-Maigne), and the high yields of protein expression using engineered strains of *E. coli*, the first matter was to determine whether the KlenTaq protein could be expressed with the SNAP-tag from bacteria, and if they were active with the SNAP-tag during PCR. To this end, KlenTaq-SNAP and SNAP-KlenTaq (SNAP-tag either a C- or N-terminus) gene constructs were cloned in the pIVEX2.3d vector, which is optimised for *in vitro* expression. Performing a colony PCR allows us to see if the gene of interest is successfully inserted in the bacteria genome. Moreover, as we do not add any DNA polymerase to the PCR mix, the colony PCR also acts as an auto-PCR experiment, as the KlenTaq polymerases expressed by the proteins are meant to amplify the KlenTaq-specific amplicon themselves. Using KRX Competent Cells from Promega, we managed to clone both pIVEX-KlenTaq-SNAP and pIVEX-SNAP-KlenTaq constructs into *E. coli* bacteria, and to express the respective proteins, active in PCR (Figure 1.1).

As a first preliminary test for the potential toxicity of THA in KlenTaq PCR, we decided to set an experiment where increasing amounts of hydrogel beads were added to the PCR mix, in the same conditions as the previous experiment. Considering that in the real process, the PCR step will be conducted in ~30µm radius droplets, which in turn contains the ~15µm beads, the hydrogel roughly accounts for 12% of the reaction volume. The range was thus set between 5 and 20%, to see if the hydrogel concentration had any effect on the PCR itself. Results showed that the beads were fortunately non-toxic for the process, independently of the concentration in that range (Figure 1.2).

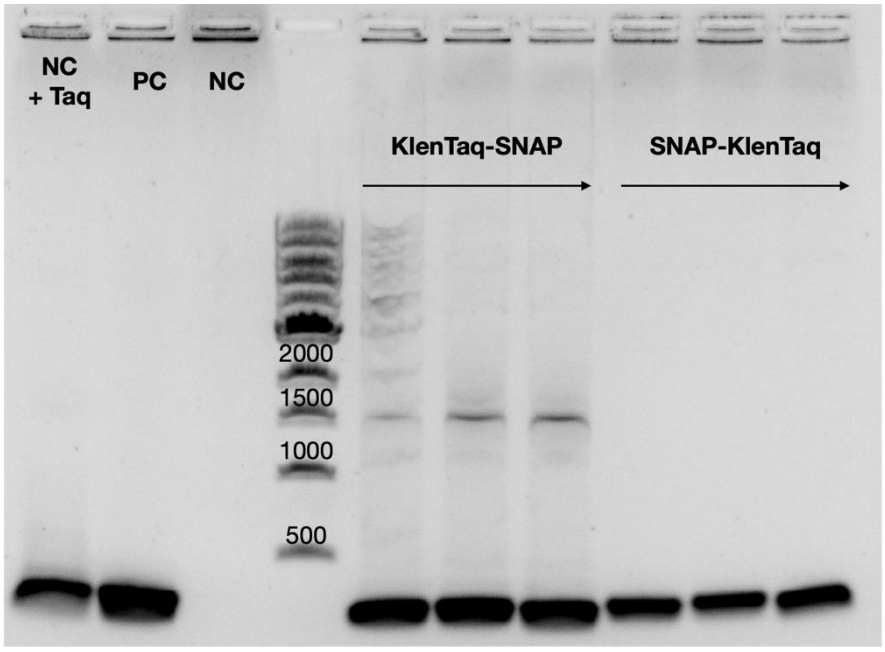



Figure 1.1: KlenTaq expressing bacteria auto-PCR.
Electrophoresis agarose gel of PCR performed by *E. coli* bacteria. Such bacteria were grown according to the protocol described in Materials and methods. 3h after induction, they were washed in resuspension buffer (50mM TrisHCl pH 7.5, 100mM NaCl), and diluted to an $OD_{600}$ of around 10, so that 1µL of bacteria in 20µL of PCR mix represents an average concentration of 100 bacteria/nL, which is optimised for the PCR (too much is toxic for the process, too few leads to no amplification). The Negative Control (NC) is set with our inactive variant of the KlenTaq DNA polymerase, while the positive control is set with 1% Vent DNA polymerase. The amplicon is 150 bp long and specific to the KlenTaq gene.

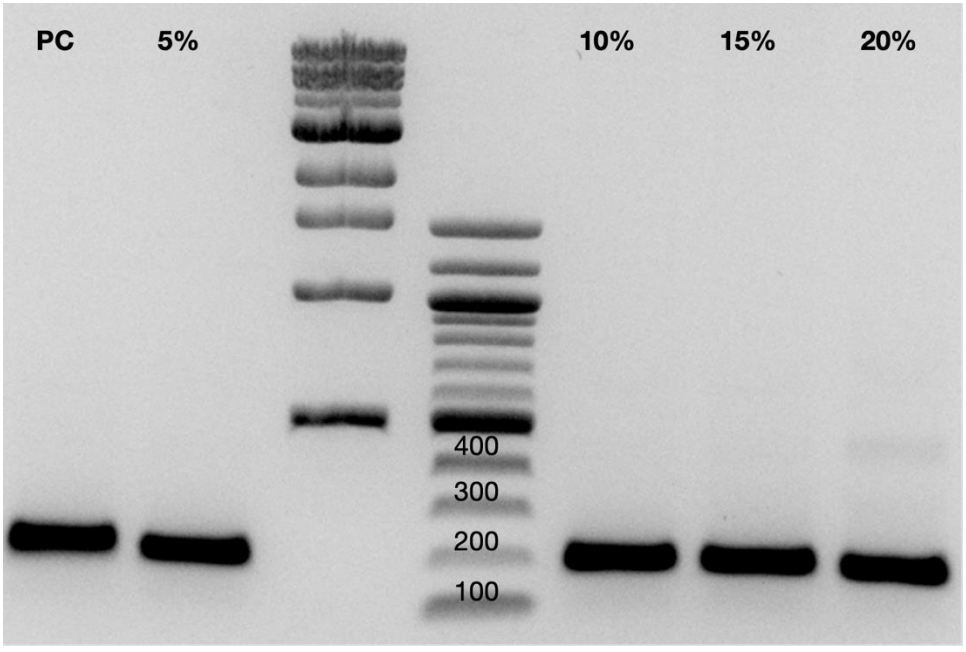



Figure 1.2: KlenTaq PCR with increasing amount of THA Beads
Electrophoresis agarose gel of PCR samples. Range of empty THA beads in KlenTaq PCR, using KRX bacteria to express the DNA polymerases. The PC is set with mQ water instead of THA beads. The amplicon is 150 bp long and specific to the KlenTaq gene.

In order to set up relevant controls, and for the future optimisation experiments of the KlenTaq itself, we purified both constructs via metal affinity chromatography, inserting polyhistidine-tags at the end of each protein. Bacteria expressing the KlenTaq mutants were lysed, and the proteins were then washed and gathered using HisTag purification columns. The corresponding 86kDa proteins were successfully observed on acrylamide gels after purification. However, even though subsequent PCR tests with such purified proteins showed amplification, it was unspecific of the 1.6kb long KlenTaq gene (Figure 1.3).

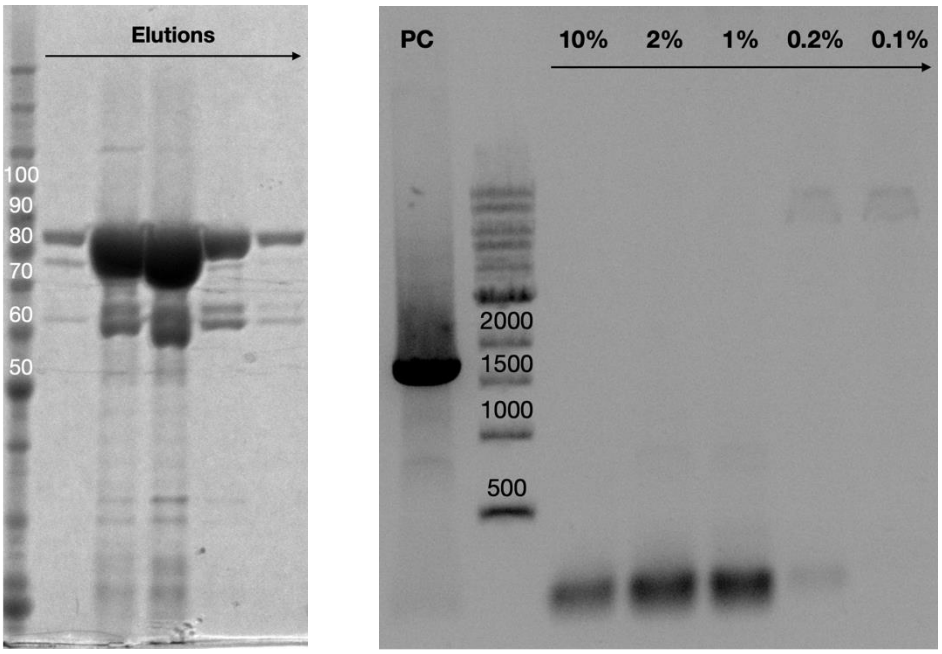



Figure 1.3: Bacterial expression and KlenTaq PCR activity test.
(Left) SDS-PAGE for the elution fractions obtained through metal affinity chromatography purification of KlenTaq-SNAP-Histag from KRX bacteria. The construct (~86kDa) is successfully overexpressed and purified. (Right) Electrophoresis agarose gel of PCR performed with dilutions of KlenTaq expressed from KRX. Below a 1% concentration in KlenTaq, the PCR does not work anymore. Moreover, the amplification is not specific, as products are only hundred-bases long instead of the expected 1.6kb.

Hypothesising that the presence of the protein tags could hinder the polymerases activity, we inserted a Thrombin protease recognition site between the sequences of KlenTaq and its His-tag. This protease is widely used for the controlled cleavage of fusion proteins, recognizing specifically the protein sequence LVPR\GS, cleaving the peptide bond between the Arginine (R) and the Glycine (G) residues. This way, we were able to precisely cut the polyhistidine tag from the KlenTaq mutants with the Thrombin enzyme. Due to the very small size of the polyhistidine tag (~2.5kDa), we conducted our cleavage experiments on the KlenTaq-Thr-SNAP-HisTag constructs, as cleavage of the SNAP-tag is much more easily seen on acrylamide gels (~20kDa). Study of kinetics showed that within 2h at room temperature, the KlenTaq proteins were entirely cleaved of their SNAP tags (Figure 1.4).

However, noting that the SNAP-tag theoretically could, at some point during the evolution of the polymerase, lead to some issues with our target protein, we decided to insert another cleavage recognition site between the sequences of KlenTaq and the SNAP-tag, a TEV (Tobacco Etch Virus) site. This protease is also very specific and efficient for the separation of fusion proteins, recognizing the ENLYFQ\S sequence, cleaving between Glutamine (Q) and Serine (S). Of the possible problems that could arise, the one we hope to circumvent with being able to remove the SNAP-tag is a potential bias in the selection of the protein characteristics (stability, activity).

We thus had the ability to specifically cleave each tag separately from the protein, depending on the construct (KlenTaq-TEV-SNAP-Thr-Histag or SNAP-TEV-KlenTaq-Thr-HisTag). In most cases however, we settled with the KlenTaq-Thr-SNAP-Thr-HisTag protein, as it was enough to cleave the entirety of the protein tags before testing its activity during PCR.

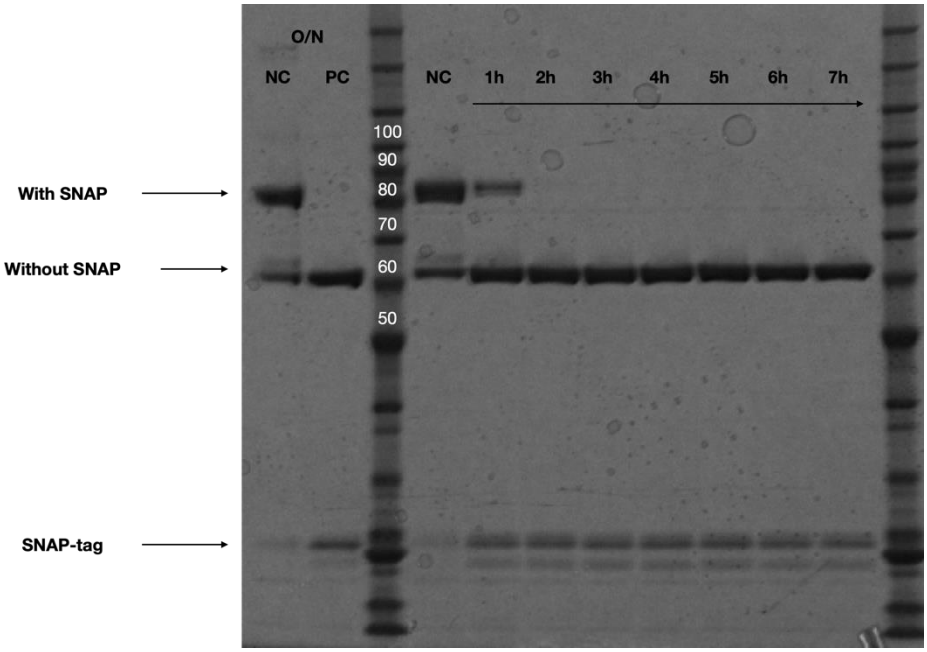



Figure 1.4: Thrombin cleavage kinetics for KlenTaq-Thr-SNAP-HisTag.
SDS-PAGE for samples obtained after an increasing amount of cleavage time with the Thrombin protease. ~20µg of KlenTaq proteins were set in each sample with one unit of Thrombin protease, in 1x Thermopol Buffer. We obtain ~63kDa proteins, which matches the size of KlenTaq. While both controls are set as overnight samples, the NC is set without Thrombin, while the PC also allows us to check if cleavage is indeed specific.

Following PCR experiments with such proteins showed that the polymerase activity was not modified with the His-tag and the SNAP-tag, both showing no sign of hinderance for KlenTaq (Figure 1.5). Moreover, although all experiments were set with the same serial dilutions of DNA polymerases obtained from bacteria-based expression and subsequent purification, we can see that there are strong fluctuations in the KlenTaq concentrations of such samples. Indeed, where 1% of the elution fractions was enough to perform PCR in previous experiments, 10% is more suitable in the later ones.

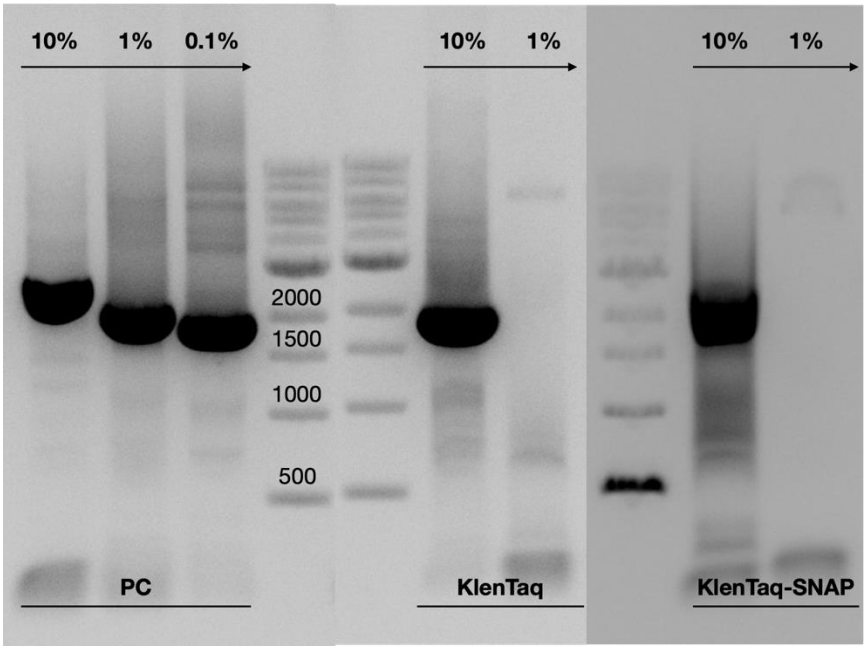



Figure 1.5: KlenTaq PCR test after HisTag cleavage

Electrophoresis agarose gel of PCR samples. The KlenTaq proteins used for the PCR were obtained after Thrombin protease cleavage of two purified constructs: KlenTaq-Thr-SNAP-Thr-HisTag (KlenTaq after cleavage), and KlenTaq-TEV-SNAP-Thr-HisTag (KlenTaq-TEV-SNAP after cleavage). The amplicon is 1.6kb long. The positive control is set with 1% Vent DNA polymerase and does not contain any component of the hydrogel. The negative control does not contain the KlenTaq gene. On the furthest gel (right), the SNAP-tag is conserved to assess if its presence hinders the polymerase activity of KlenTaq.

Finally, we investigated if the proteins were active in the hydrogel during PCR, and if so, whether the activity was impeded or not. Indeed, considering that the polymerases will be bound to the hydrogel in the beads, their free movement in the medium - and thus activity - may be hindered. We set up two almost identical experiments of PCR in hydrogel with the KlenTaq-SNAP mutants, expressed and purified from bacteria. In the first, the hydrogel was reticulated without Benzylguanin, preventing the SNAP-tag from bonding with the hydrogel matrix. In the second, 1 mM BG were added to the PCR/hydrogel mix that allowed the fused proteins to form a covalent bond with the THA via SNAP-tag reaction (bead display step, 1h at 37°C before PCR). We found the same results for both experiments, indicating that the polymerase activity is not hindered by its link with the hydrogel bead, and a further confirmation that the SNAP-tag also does not seem to affect polymerase activity (Figure 1.6).

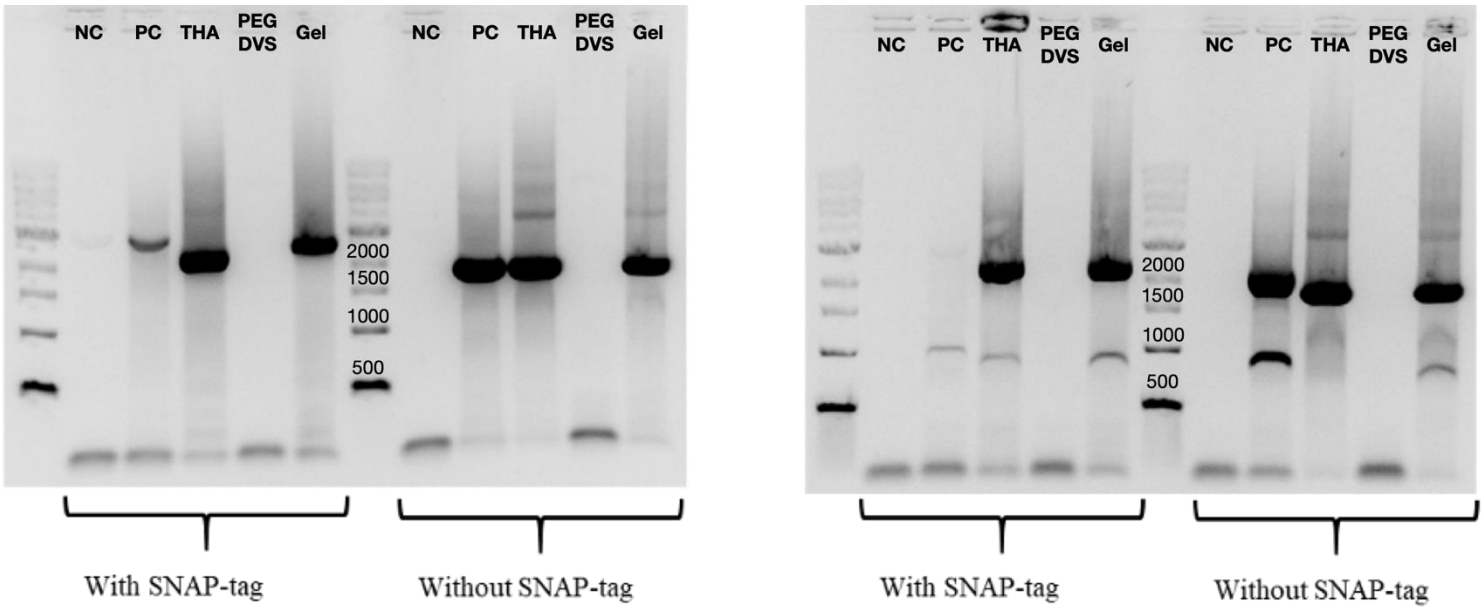



Figure 1.6: THA hydrogel effect on KlenTaq PCR

Electrophoresis agarose gel of PCR samples. KlenTaq PCR is conducted in various environments: THA alone, PEGDVS alone, and both together (hydrogel). The amplicon is 1.6kb long. The positive control does not contain any component of the hydrogel. The negative control does not contain the KlenTaq gene. (Left) The hydrogel is reticulated without any BG-Mal. The polymerases are “free” in the media. (Right) The hydrogel is reticulated with 10 µM BG-Mal. The polymerases are covalently bound to the hydrogel matrix. In both cases, the KlenTaq protein was also sometimes preventively cleaved of its SNAP-tag, to assess if its link to the hydrogel matrix was hindering its polymerase activity.

## 2. IVTT expression & PCR activity

Once our first experiments with bacteria-expressed proteins were successful, we turned our attention to the IVTT process, with similar investigations in mind: whether the KlenTaq-SNAP proteins could be expressed, active during PCR, etc. Using previous optimisation performed by former members of our team, we conducted the IVTT expression of KlenTaq polymerases with a concentration of genetic material (KlenTaq-SNAP-HisTag plasmid) around 1nM. With a large volume of IVTT mix (~200µL) set at 37°C overnight, we managed to express and purify our KlenTaq-Thr-SNAP-Thr-HisTag and SNAP-Thr-KlenTaq-Thr-HisTag (~86kDa), along with the GFP-SNAP protein (~54kDa), which would be essential for a number of controls down the road. Subsequent activity tests were set with such purified polymerases, in comparison with our previously well-established PCR from either bacteria themselves (auto-PCR) or bacteria-expressed proteins. We discovered that there seemed to be an issue of activity with the KlenTaq-Thr-SNAP-Thr-HisTag and SNAP-Thr-KlenTaq-Thr-HisTag proteins, even after SNAP and HisTag cleavage with Thrombin (Figure 2.1).

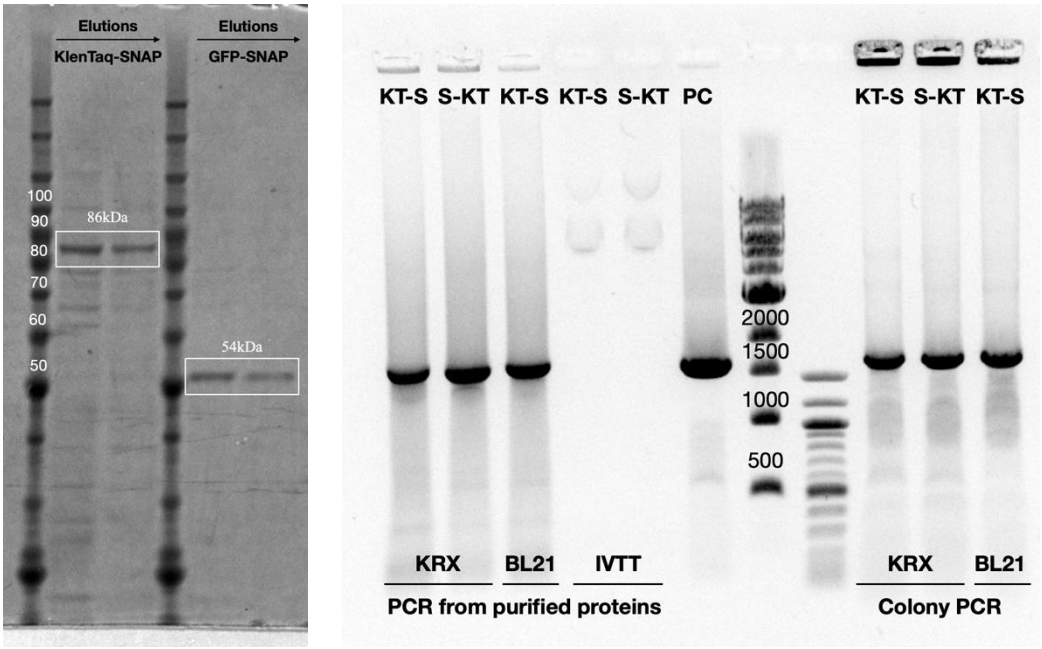



Figure 2.1: IVTT expression and KlenTaq PCR activity test.
(Left) SDS-PAGE for the elution fractions obtained through metal affinity chromatography purification of KlenTaq-SNAP-Histag and GFP-SNAP-HisTag from IVTT systems. The constructs (~86kDa and ~54kDa) are successfully overexpressed and purified. (Right) Electrophoresis agarose gel of PCR performed with KlenTaq in several conditions. We detect amplification in the case of proteins expressed or purified from bacteria (KRX and BL21), but not from polymerases purified from IVTT. The amplicon is 1.6kb long. The positive control is set with 1% Vent DNA polymerase.

A first approach to investigate the issue was to see if traces amount of *In Vitro* Transcription Translation mix (IVTT) were still present after protein purification, which could cause the issue in PCRs, due to its potential toxicity for the reaction. Experiments showed that the PCR was unsuccessful when more than 0.1% of the reaction mix was IVTT mix (Figure 2.2). We reckon that the numerous components of the bacteria lysate may disrupt the polymerases activity when too concentrated in the PCR mix. Although it was not the issue *per se* for our experiments with proteins purified through affinity chromatography, as these are much purer after elution (at least under 0.1% of IVTT mix in the PCR after dilution), it could have been a problem in our overall "real" process. Hence, for the rest of the experiments, we thoroughly washed the hydrogel beads after the IVTT step with Thermopol (PCR Buffer) mix, to prevent further potential PCR disruption.

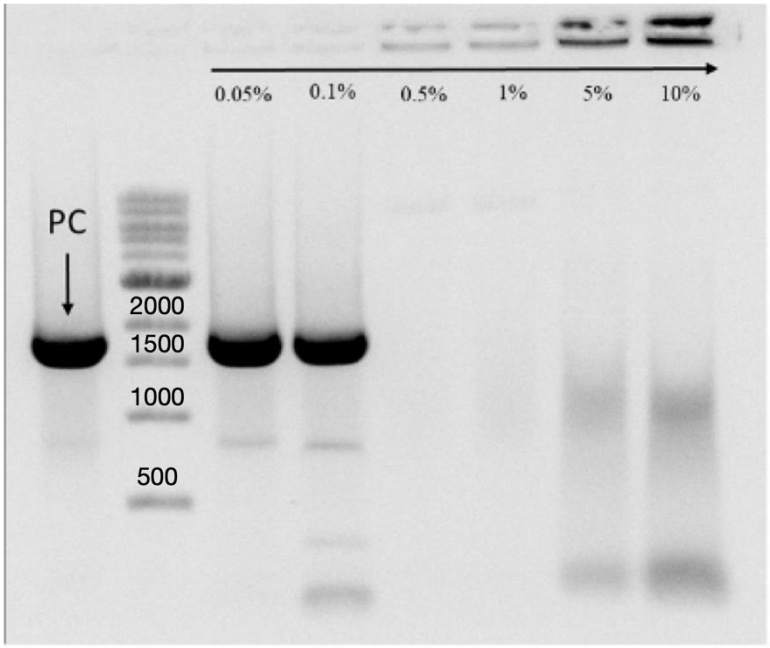



Figure 2.2: KlenTaq PCR with increasing amount of IVTT mix
Electrophoresis agarose gel of PCR samples. Samples are set with purified KlenTaq DNA polymerase and increasing amounts of IVTT mixture in the PCR mix. Above 0.1% of IVTT in the PCR, no amplification is observed. The amplicon is 1.6kb long. The positive control is set with 1% Vent DNA polymerase and 0% IVTT mix.

More importantly, as previous experiments foreshadowed, the control of DNA polymerase concentration in the PCR mix is crucial for the optimal amplification of the template. To this end, two parameters must be taken into account: the duration of the IVTT step, which will produce varying amounts of our polymerases; and the concentration of BG-Mal set in the hydrogel matrix, which reflects the maximal concentration of enzyme that can be displayed on the beads.

On the one hand, towards the understanding of the effects of IVTT duration on the PCR step, we proceeded to perform RCA on THA beads with 1µM of BG-Mal, before using such beads in several conditions, but most importantly with 1, 2 or 3 hours of IVTT (Figure 2.3). From this experiment, we can conclude that although the amplification is still unspecific, 2h of IVTT seems optimal, and the SNAP-tag cleavage heavily improves KlenTaq activity in the reaction, contrarily to what was discovered in the case of bacteria-expressed proteins.

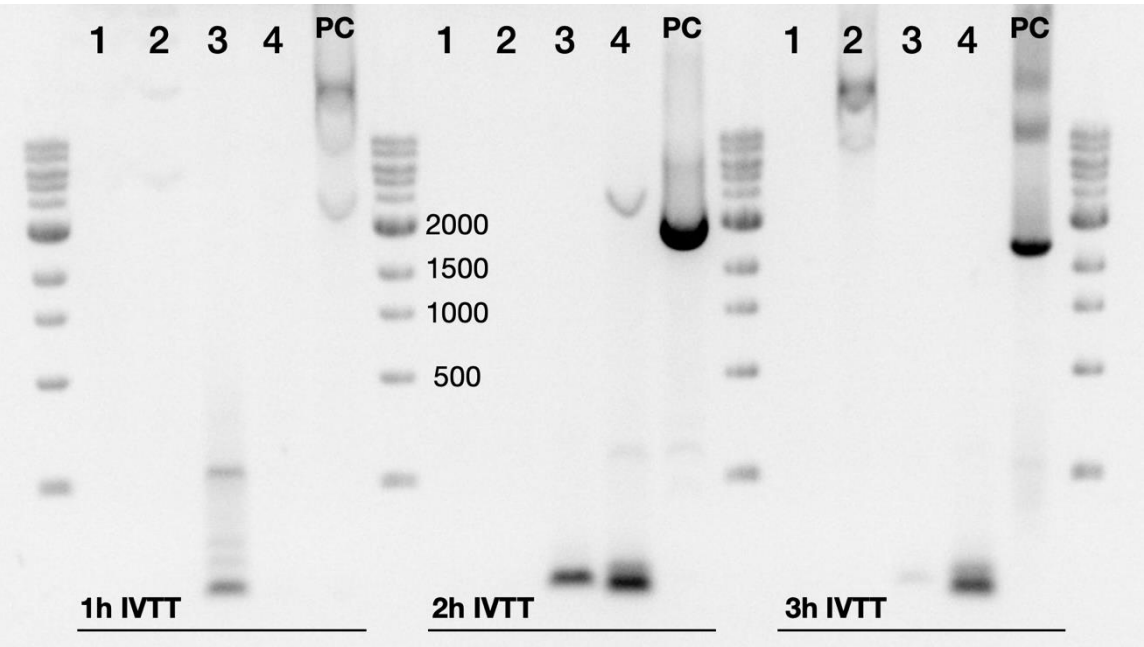



Figure 2.3: Influence of IVTT duration on KlenTaq PCR
Electrophoresis agarose gel of PCR samples. PCR performed on THA beads that underwent RCA and varying durations of IVTT, in several different conditions. 1) No additional components. 2) With additional template plasmid. 3) With SNAP-cleavage. 4) With 1% Vent DNA polymerase. The take-away result is that 2h of IVTT seems better compared to 1h and 3h. The amplicon is 1.6kb long. The positive control is set with 1% Vent DNA polymerase and the template.

On the other hand, in order to investigate the adequate concentration of BG-Mal to use in the THA, we set up an experiment where we use hydrogel beads that already underwent RCA with our KlenTaq gene, with 10µM of BG-Mal inserted in the matrix before reticulation. Those beads are then incubated in the dark at 37°C for 1h (SNAP-tag reaction) with varying quantities of KlenTaq-TEV-SNAP purified from bacteria, before being thoroughly washed in Thermopol buffer to remove the excess of unbound DNA polymerases. Results show that a working concentration of around 1µM in KlenTaq seems to be optimal for the PCR step, as less enzyme results in too weak of an amplification, and more in a smear of unspecific products (Figure 2.4). Hence, we decided to use 1µM of BG-Mal in our THA beads, in order to achieve such a maximal concentration of DNA polymerases after the IVTT step.

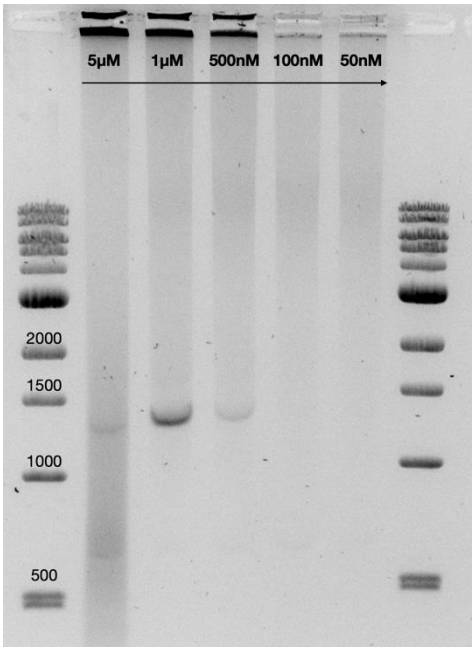



Figure 2.4: Range of [DNA polymerase] in KlenTaq PCR.
Electrophoresis agarose gel of PCR samples. Before PCR, THA beads that underwent RCA but not IVTT are incubated with varying concentrations of purified KlenTaq from bacteria for the SNAP-tag reaction (1h in the dark at 37°C, agitated at 500RPM). The amplicon is 1.6kb long.

Taking all of these results into account, starting from 1µM BG-Mal THA beads that underwent RCA, we replicated the KlenTaq expression from IVTT with a total duration of 2h, and consolidated the previous results and amplification (Figure 2.5). The PCR is still unspecific, but the bands that are obtained are much darker and neater than before, which constitutes progress nonetheless.

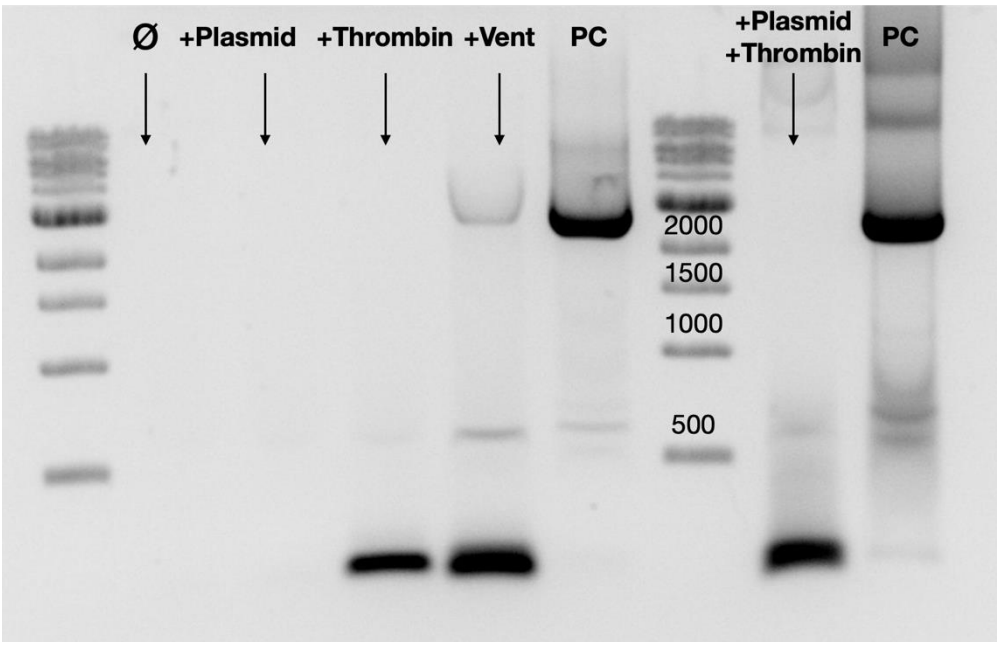



Figure 2.5: KlenTaq PCR with 2h of IVTT in several conditions
Electrophoresis agarose gel of PCR samples. The various individual experimental conditions of the same as in Fig. 2.5. The amplicon is 1.6kb long. The positive control is set with 1% Vent DNA polymerase and the template.

While all of this investigation was taking place, Rémi Sieskind discovered that a similar issue arose in the IVTT expression of its protein, and his set-up. Indeed, because the RCA step produces linear templates through its exponential amplification process, these strands can be degraded by some of the enzymatic components of the IVTT mixture, and most likely by the RecBCD complex, acting as an exonuclease in the media. As such, the longer the IVTT, the lesser template is left at the end to amplify, which could explain the observed unspecificity of the products and weak efficiency of the PCR. In order to counteract this process, several inhibitors of this degradation were investigated, and the better one was found out to be the $\chi_6$-sequence, which delays the formation of the protein complex and the degradation of DNA[204,205]. We thus tested the disruptive performance of this sequence during the IVTT on our THA beads by looking at the amplification efficiency of the following PCR (Figure 2.6). It still did not solve this unspecificity, but it drastically improved the overall amplification observed at the end of the PCR, which is always appreciable.

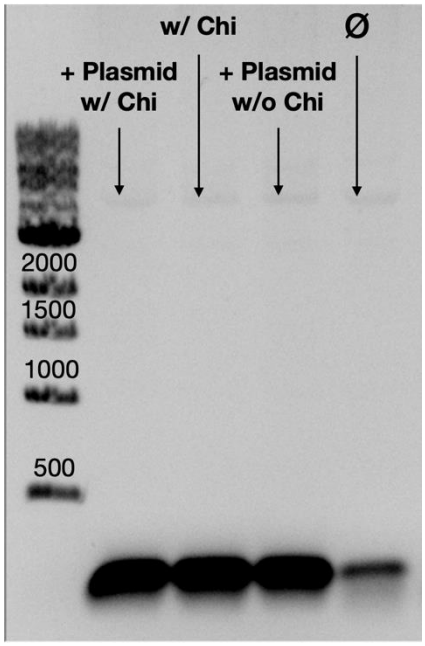



Figure 2.6: Effect of adding the $\chi_6$-sequence during the IVTT step prior to KlenTaq PCR.
Electrophoresis agarose gel of PCR samples. The PCR is set with samples that underwent 2h of IVTT either with or without the $\chi_6$-sequence. The template plasmid was also added in the PCR master mix of some samples to counterbalance DNA degradation during IVTT. The amplicon is 1.6 kb long.

In the face of this unspecific amplification after all these optimisation steps, we investigated possible pathways of improvement that were still left unexplored, and targeted the PCR itself. Indeed, the PCR protocol and primers were the same as the ones Adèle Dramé-Maigne designed for her intent and purposes, which grew quite dissimilar from mine through the successive optimisations of my *in vitro* set-up. A number of changes were set-up by following the Sigma Aldrich "KlenTaq LA DNA Polymerase Mix Technical Bulletin". Although this document refers to the PCR optimisation for their commercial KlenTaq polymerase, it would seem that their information holds true even for the enzyme expressed through bacteria, IVTT, and with the SNAP-tag.

We changed the PCR protocol from a three-step cycling protocol to a two-step one, with annealing and extension at the same temperature, and we heavily modified the KlenTaq PCR primers: the annealing temperature was too low and increased (60°C to 72°C), and the sequences were designed so that a final CG or CC motive was inserted at 3' end of the primers, in order to increase priming efficiency[206].

With those new settings, RCA, IVTT and PCR were conducted on THA beads, and specific amplification was finally recovered (Figure 2.7). Once again, this experiment is further confirmation that 2h of IVTT seems optimal compared to 1h and 3h.

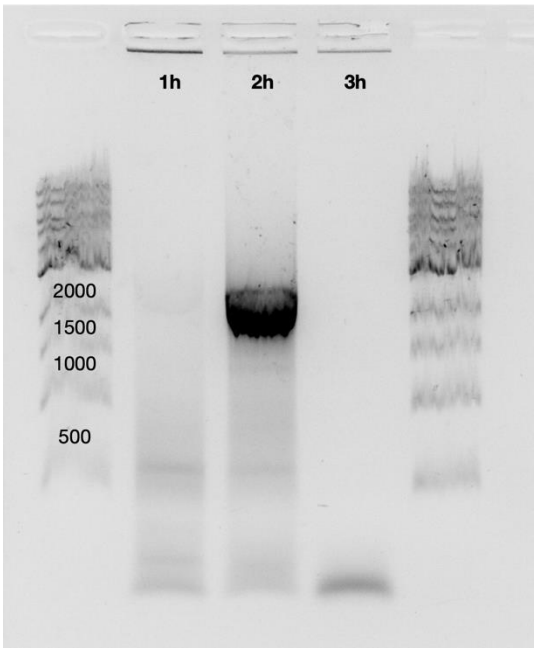



Figure 2.7: Influence of IVTT duration on KlenTaq PCR
Electrophoresis agarose gel of PCR samples. PCR performed on THA beads that underwent RCA and varying durations of IVTT, in the new and improved conditions that were previously set-up. The amplicon is 1.6kb long.

However, one interesting - and crucial - information of this optimisation process we discovered is that the self-selecting PCR at the end of our cycle is a very finely tuned reaction. In order to obtain a satisfying amplification efficiency, a minute and sensitive balance must be found between all of the previous steps, a multiple trade-off between protein expression, template degradation, polymerase concentration in the media, etc.

Although the RCA and IVTT steps were performed in droplets, all of this PCR optimisation was done in bulk, meaning that we still had to reproduce these results in droplets. Fortunately, the conditions we set in bulk – most importantly, the proportion of THA beads in the PCR master mix - were quite close to those in the real, "in droplets" process.

## 3. Optimisation in droplets

First experiments proved unsuccessful, because of the instability of the emulsion when heating during the denaturation step at 95°C. A first step was thus to add several agents in order to stabilise the droplets: Yeast RNA (1%), Pluronic F-127 (0.4%) and Ficoll PM70 (20%). Ficoll acts as a crowding agent to facilitate penetration of reagents in the droplets[207]. The Yeast RNA prevents the adsorption of DNA strands on the droplets interface with the surfactant, and Pluronic F-127 acts the same for proteins. However, even after the addition of these compounds, our KlenTaq PCR was still not working in droplets, using THA beads that previously underwent RCA and IVTT.

As an exploratory endeavour, we investigated the influence of THA beads on KlenTaq PCR when compartmentalised in droplets. Indeed, even though we knew the compound was not toxic for the reaction, it did seem possible that the relative proportion of beads inside a droplet could have an effect on the amplification kinetics that characterise the PCR. We thus decided to set up a PCR experiment, in bulk, with varying concentrations of THA beads that underwent RCA and IVTT, according to the protocols of the final platform (Fig. 3.1).

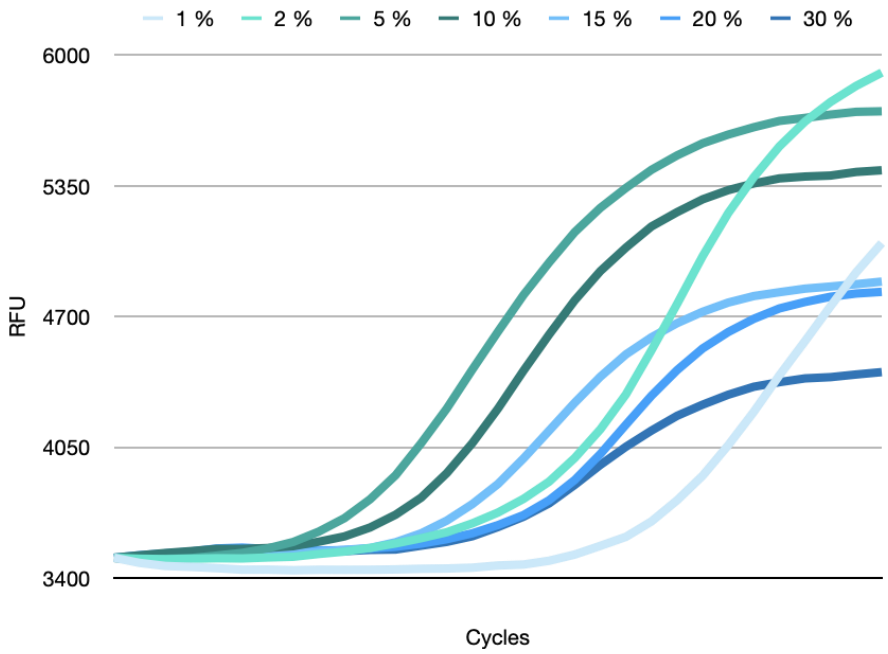



Figure 3.1: Influence of % THA beads in PCR mix on KlenTaq PCR.
Fluorescence graph of KlenTaq PCR through cycles. PCR performed on THA beads that underwent RCA and 2h of IVTT. Varying concentrations of THA Beads seem to be instigate differential amplification kinetics of the KlenTaq gene. The amplicon is 1.6kb long. Graph obtained with a CFX96 qPCR machines from Biorad.

Based on the previous experiment, we assessed that concentrations of THA beads between 2% and 15% were the most efficient towards the better kinetics of amplification for KlenTaq. We decided to replicate the experiment with arbitrary concentrations of 8% and 16% THA beads respectively, in order to estimate if the amplification was indeed appreciable in those conditions, which turned out to be the case for both conditions (Fig. 3.2).

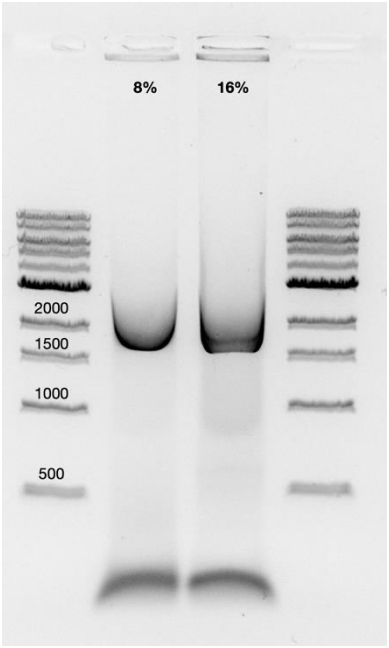



Figure 3.2: Influence of THA beads % on KlenTaq PCR.
Electrophoresis agarose gel of PCR samples. PCR performed on THA beads that underwent RCA and 2h of IVTT. Both concentrations of THA Beads seem to be leading to an efficient amplification of the KlenTaq gene. The amplicon is 1.6kb long.

However, the experimental set-up we used consisted of a re-encapsulation step before the PCR, which compartmentalises our ~15µm THA beads in ~20µm water-in-oil droplets. This would entail that roughly 40% of the reaction volume inside the droplet is occupied by the hydrogel bead, and as we can see on the previous graph, the kinetic is much more efficient below the 20% mark.

To solve this issue, we switched the 20µm re-encapsulation device for a similar ~30µm one, as the ratio of volumes between hydrogel bead and water-in-oil droplet would go down to 12%, much closer to the most efficient ratios that were previously determined. In these conditions, we managed to obtain a properly functioning PCR in droplets with THA beads that underwent RCA with the KlenTaq gene, followed by 2h of IVTT (Fig. 3.3).

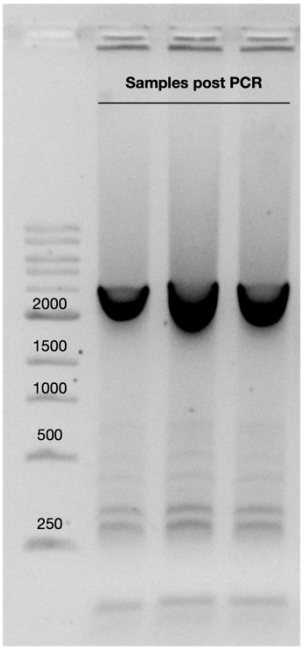



Figure 3.3: Samples after KlenTaq PCR in 30µm droplets.
Electrophoresis agarose gel of PCR samples. PCR performed on THA beads that underwent RCA and 2h of IVTT. 3 different replicas were analysed on gel after PCR and emulsion break. Although the bands are around the 2kb ladder band, the amplicon is 1.6kb long (the gel was probably overloaded).

The optimisation of this final PCR was the last step of the process. Although these optimisations steps were performed using the active KlenTaq DNA polymerase, we expect that the process will still function properly with variants of the enzyme, which would exhibit varying degrees of activity compared to the original protein.

## *B.* Study of noise during protein expression

### 1. Aminoglycoside antibiotics

In the context of translation and protein synthesis, as a first step of our investigation of the efficiency of aminoglycoside antibiotics towards their error-prone properties, we initially screened a wide range of concentrations both for kanamycin and streptomycin. The following graphs are obtained from the fluorescence recorded in the CLARIOstar, plotted against the respective antibiotic concentration at $\Delta t = 8h$.

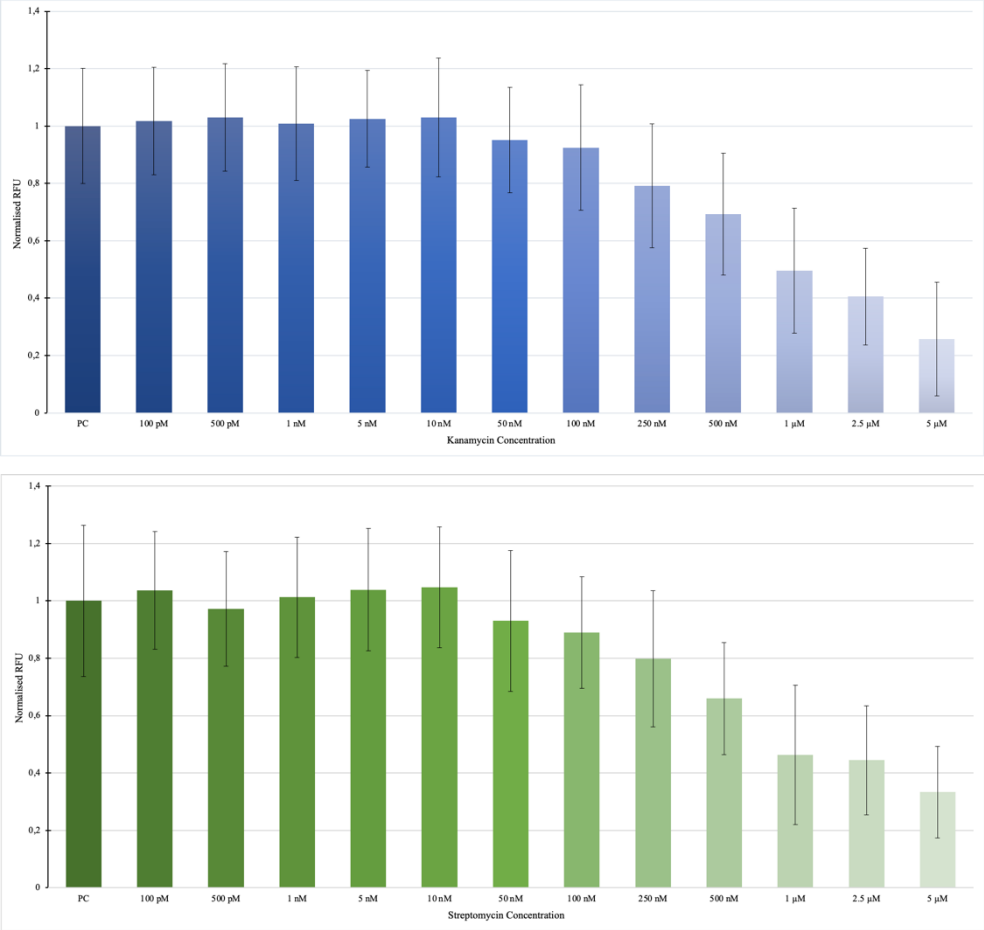



Figure 1.1: Effects of aminoglycoside antibiotics concentrations on GFP translation.

IVTT protein synthesis of GFP. Experiments were run with varying concentrations of aminoglycoside antibiotics: kanamycin (blue) and streptomycin (green). The positive controls (PC) did not contain any antibiotic. Data obtained through 12 different replica experiments.

For future experiments, we decided to narrow the concentration range to [5 nM, 50 nM, 250 nM, 500 nM, 2.5 µM, 5 µM]. As we can see, there is quite a lot of variability between each run, and multiple replicas of each experiment were necessary in order to capture the trends.

Next, as both antibiotics have to be prepared in large quantities - due to the minute working concentrations – we wanted to see if batches of diluted kanamycin and streptomycin could be prepared in advance, frozen, and later thawed and used in our experiments. We thus measured the effects of two weeks old, -80°C frozen batches, compared to freshly prepared ones of antibiotics on GFP protein synthesis.

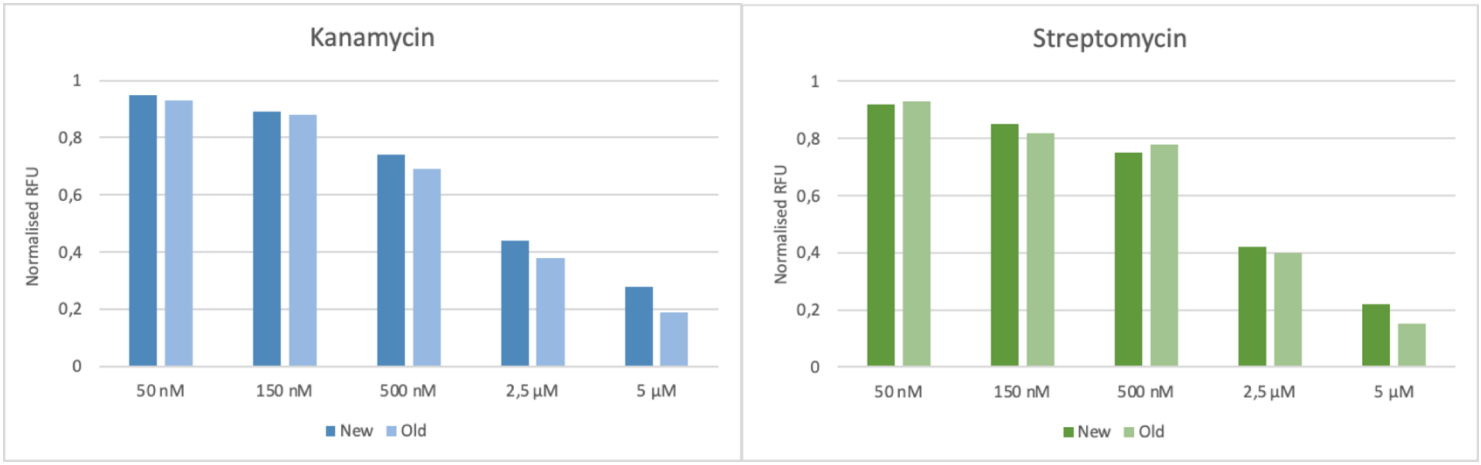



Figure 1.2: Effects of aging and freezing on aminoglycoside antibiotics.

IVTT protein synthesis of GFP. Fluorescence recordings are normalised compared to the Positive Control, which does not contain any antibiotic. Experiments were run with varying concentrations of aminoglycoside antibiotics: kanamycin (blue) and streptomycin (green). Old samples are run with two weeks old, prepared-and-frozen batches of antibiotics, while the new ones are freshly prepared.

The effects of freezing and eventual deterioration of the antibiotics seem to be minimal, as the differences in effect of old and new batches of kanamycin and streptomycin are negligible in our set-up. We continued performing our experiments with such frozen dilutions of antibiotics.

## 2. Quantification of mutations

In order to perform a relative - albeit rough - quantification of non-sense or missense mutations inserted through the use of aminoglycoside antibiotics during protein synthesis, we relied on the silver staining kit and protocol provided by Thermofisher (24612). However, it quickly became clear that the aforementioned conditions were not precise enough to accurately distinguish bands from one another, and that we had to adapt the protocol to our goals.

Most of the optimisation consisted in fine-tuning the dilution factor of the samples on the polyacrylamide gels, and the development duration (staining of silver ions on the proteins contained in the gel). In the end, we managed to capture pictures that were subsequently put through protein gel analysis software.

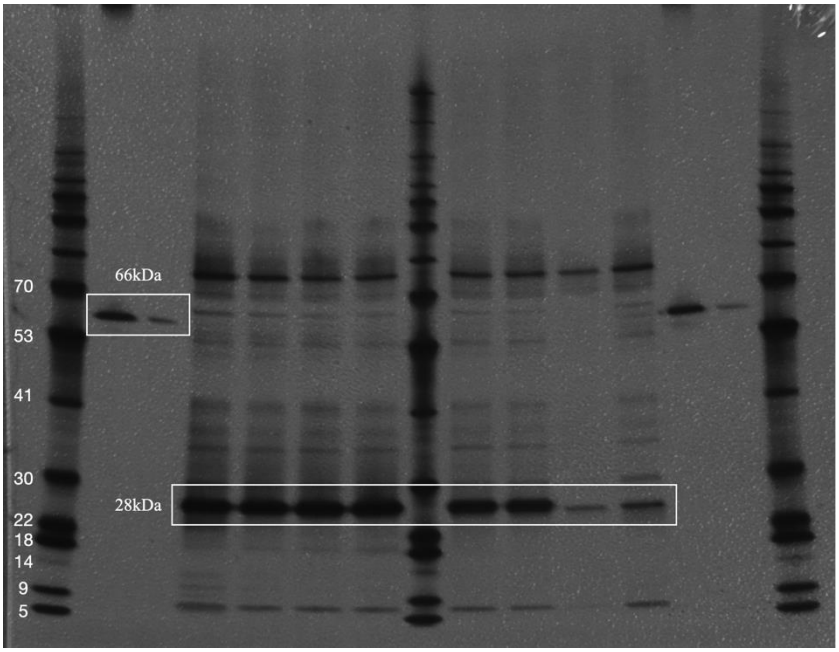



Figure 2.1: Silver staining after SDS-PAGE.
As a standard for relative quantification, serial dilutions of commercial BSA (~66kDa) are set on the polyacrylamide gel with the GFP-HisTag (~28kDa) samples obtained after IVTT purification (HisTag affinity chromatography), before running a SDS-PAGE, and silver staining.

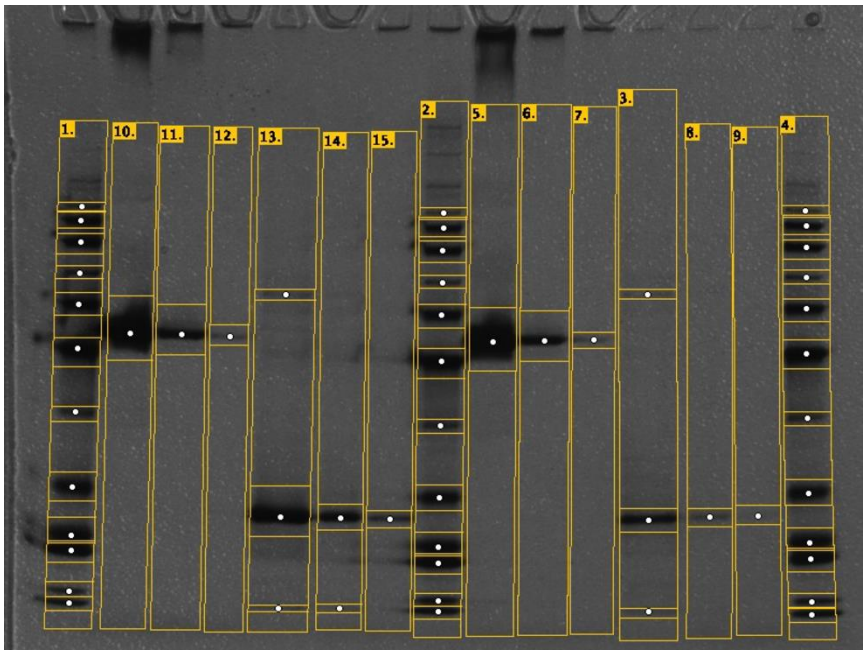



Figure 2.2: Gel analysis after silver staining / SDS-PAGE.
With the analysis software, a relative quantification of the samples can be achieved. Density on the gel reflects the amount of protein loaded and migrated, so that bands can be compared between each and other.

### 3. Mass spectrometry

As a means to get finer and more informative data on the nature of the many random mutations that are inserted during protein synthesis through the use of aminoglycoside antibiotics, we decided to rely on mass spectrometry methodologies. However, to attain these superior levels of detail and accuracy on the characterisation of each change in peptidic fragments, samples have to be rigorously prepared before such analysis.

Most notably, the presence of any detergent is highly detrimental to the integrity of the mass spectrometer when loading samples, so thorough washing steps are mandatory in our set-up, as we use Tween 20 for the proper resuspension of our THA beads. Although it is not an issue in the case of our preliminary experiments with GFP-HisTag, it is still a very important condition to keep in mind. However, for our experiments with samples purified through affinity chromatography columns after IVTT, we can see in previous silver-stained polyacrylamide gels that even though our protein of interest is quite clearly purified, it is not neatly isolated. Consequently, a high number of E. coli proteins were recorded in the MS analysis, which dilutes the accuracy towards our targets, GFP-based peptidic fragments.

To this end, after investigating several ways to achieve higher purity in our samples, we ended up tweaking our purification protocol towards a gradient of imidazole between our washing and elution buffers (5mM and 150 mM imidazole respectively). Washing the affinity chromatography columns with intermediate buffers containing 25, 50, 75, 100 and 125 mM imidazole, we managed losing roughly half of the unspecific proteins still bound, at the cost of a 30% fraction of our GFP. All in all, this would result in an increase of the GFP/Noise ratio, which was confirmed by MS analysis of the samples.

# Material & Methods

## 1. Microfluidic devices

For the preparation of microfluidic wafers:

Starting from the masks designed by Rémi Sieskind, the 2-layer moulds were fabricated in the white rooms of the Institut Pierre-Gilles de Gennes pour la Microfluidique. SU-8 2005 photoresistant resin is used for the first layer (5µm-high motif) and SU-8 2050 for the second (50µm-high motif), both subsequently spin-coated (at 3000 and 3250 RPM respectively) on the silicon wafers, before illumination with UV light as to reticulate both motives onto the wafer.

Every microfluidic device was prepared identically, as follows:

40g of liquid PDMS elastomer was poured with 4g of curing agent (Silicone Sylgard 184, Dow Corning, 1317318) on a mould lined with rigid, aluminium foil. After mixing, the mould is left in a vacuum chamber, the air extracted from the PDMS mixture, and then left to incubate at 70°C for 2 hours. The resulting slab is then detached from the mould, and covered in tape to prevent dust from sticking to it. Using a 1.5mm biopsy punch (Integra Miltex, 33-31A-P/25), the inlets and outlet are punched out of the slab. The slab and a 1mm-thick glass slide are then extensively washed with acetone, isopropanol and ultra-pure water, before being covered in tape again. In a white room, the PDMS slab and the glass slide are bonded covalently with a plasma cleaner. The resulting device is then incubated at 200°C for 5 hours.

Finally, in order to better prevent droplet polydispersity, a hydrophobic coating (3M Novec 1720) is flushed through the device by the outlet, before drying off at 90°C for 1 hour. It is to be noted that this step was performed each and every time before use of any microfluidic device.

For the mixing millipedes, several steps are however added to the process:

First, during inlet/outlet punching, the mixing chamber is punched as well with a 4mm biopsy punch (Integra Miltex, 33-34-P/25), but the PDMS stub is kept for later use. In this hole, a 2mm-large magnetic stirring bar (Bel-Art Spinbar, F37119-0002) is. Part of the PDMS stub retrieved earlier is then used to plug the hole, being watchful as to not hinder the rotation of the magnetic stirrer in any way. 1g of PDMS elastomer is mixed with 0.1g of the curing agent, degassed in the vacuum chamber, and set on top of the mixing chamber stub, as to ensure proper sealing during use of the device. The latter is then finally set for 2 hours at 70°C.

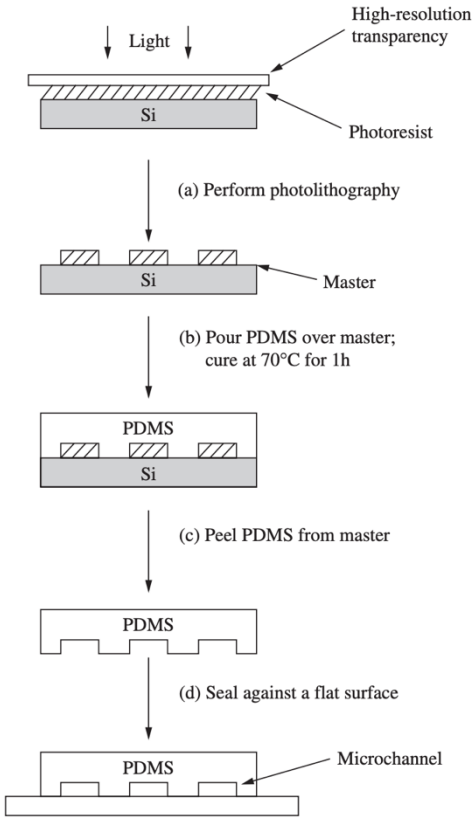



Figure 1: Microfluidic devices preparation scheme.
(a) Photolithography is performed at the IPGG white rooms, in order to create the wafers from which the microfluidic devices are later based. The high-resolution transparency filter is the mask of the device, set to reticulate only part of the photoresistant resin below it. After exposure to UV light, the master wafer is set on the silicone disk, forming a "negative" of the microchannels that are desired on the chips. (b) Liquid PDMS is poured on the wafer, along with a curing agent. After incubating at 70°C for a number of hours, a solid, transparent slab of PDMS is obtained, moulded from the previous relief. (c) The slab is carefully peeled from the wafer, as to not recklessly break the fragile silicone disk. (d) Finally, the slab is sealed on a glass slide after plasma treatment. Adapted from Tang *et al*.[208]

## 2. RCA and hydrogel beads

Thiolation of hyaluronic acid and subsequent Ellman's test are performed according to Rémi Sieskind's protocols[178]. Aliquots of THA are then stored at -80°C for a few months.

For the preparation of hydrogel beads:

Before anything else, our circular DNA is twice nicked, using the Nt.BspQI enzyme (NEB R0644S) and the appropriate protocol. The nicking sites, distant of 40bp, have been shown that to improve RCA overall efficiency[178]. Also beforehand, a premix of dNTPs is prepared from a stock (NEB N0477L), and composed in the end of 30% α-S-dCTPs and 70% regular dCTPs for the future $I_2$-based degradation of the RCA product. The DMSO (D8418-100ML), Pluronic F-127 (P2443-250G) and Yeast RNA (10109223001) are supplied by Sigma-Aldrich, whereas the random primers (SO181), which initiate the RCA and are protected from Φ29's 3'→ 5' exonuclease activity by phosphorothioate bonds at the 3' end, are supplied by Thermofisher. Finally, the fluorescent dye used to detect the amplification products from the RCA (ds-DNA) is Evagreen (Biotium, 31000).

Due to the quickly-reticulating nature of the THA hydrogel at room temperature, the RCA mixture is prepared through two separate solutions, one for the THA and one for the PEG-DVS. Each of these is twice concentrated in THA and PEG-DVS, so that the *in-situ* blend of the two actually produces a solution at the right concentration. However, a number of precautions have to be undertaken in the preparation of said pre-mixes. First, DTT, a reducing agent usually present in the RCA reaction buffer, cannot be used here as it would react with the thiol moieties of several compounds in the media. Moreover, a number of reagents cannot be mixed with the PEG-DVS, due to its highly reactive nature. Namely, the Maleimide-bound compounds (BG-Mal: NEB S9153S; Dylight405-Mal: Thermofisher 46600), the BSA (NEB B9000S) and the Φ29 DNA polymerase (NEB M0269L). These are thus only added with the THA, at twice the concentration required.

The THA and PEGDVS solutions are preventively filtered through 0.2μm filters (Corning, 431212). Two 250μL syringes (Hamilton, 81120) are first both filled with 50 μL of Novec 7500 fluorinated oil, then 150 μL of either the THA or the PEG-DVS solution are collected in each syringe. The oil is manually centrifuged to gravitate towards the bottom of the syringes, the furthest away from the exit, the THA/PEG-DVS solutions nearest from it, while any residual air is pushed away. 30 cm of PTFE tubing (outer diameter: 1.6mm, inner diameter: 0.5mm, BOLA, S1810-09) is then adapted onto 25G×1" needles (Terumo, AN*2425R1). The tubing is then filled with the respective aqueous phase by pushing the piston.

Syringes are then finally set on the syringe pump, with a pushing speed of 3 μL/min, while the fluorinated oil channel is controlled through regular pressure pumps, set around 180/200mbar. Tubings are placed in the mixing millipede inlets, as well as a magnetic stirrer under the microfluidic device, so that the magnetic bar can rotate as fast as possible. For the generation of hydrogel beads, a 2%w/w (32mg/mL) fluorinated oil phase is used (Fluosurf surfactant). The resulting emulsion is collected in a 1 mL low-bind pipet-tip at the outlet of the device, put into a 2 mL-low bind Eppendorf tube, and incubated in the dark at 30°C for at least 3 hours, so that the hydrogel fully reticulates and the RCA completes. The emulsion is finally collected and broken. To this end, the fluorinated oil phase below the emulsion is first carefully removed, and ~1 mL of aqueous buffer (PBS with 0.1% v/v Tween 20) is added to the emulsion, as well as ~200 μL of PFO, 1H,1H,2H,2H-perfluoro-1-octanol (Sigma-Aldrich, 370533-25G). The tube is vortexed extensively and then set to decant for a few minutes. Once again, the aqueous phase (containing the hydrogel beads) rises to the top, and can be collected with a low-bind tipped micropipette.

As mentioned previously, multiple washing steps are necessary between the different processes of the platform. Once the hydrogel beads are collected, these are centrifuged in an Eppendorf, the supernatant removed and replaced with fresh washing buffer (PBS with 0.1% v/v Tween 20), and the contents of the tube homogenised by vortexing. The operation is then repeated several times (at least three or four, depending on the visual purity of the hydrogel pellet). The Tween 20 is a surfactant that prevents beads from sticking to one another, which usually leads to hydrogel aggregates.

| | Initial conc. | Final conc. | THA solution (36 mg/mL) | PEGDVS solution (20 mg/mL) |
|---|---|---|---|---|
| **Φ29 DNA pol RB (No DTT)** | 10x | 1x | 33.3 µL | 33.3 µL |
| dNTPs mix | 10 mM | 500 µM | 33.3 µL | - |
| DMSO | 20% | 1% | 16.65 µL | 16.65 µL |
| Evagreen | 20x | 0.5x | 8.33 µL | 8.33 µL |
| Circular DNA | 1 nM | 10 pM | 6.66 µL | - |
| Random primers | 500 µM | 12.5 µM | 8.33 µL | 8.33 µL |
| **TCEP** | 10 mM | 500 µM | 33.3 µL | - |
| Yeast RNA | 100ng/µL | $1\%_{v/v}$ | 3.33 µL | 3.33 µL |
| Pluronic F-127 | $8\%_{w/w}$ | $0.4\%_{w/w}$ | 16.65 µL | 16.65 µL |
| BG-Mal | 100 µM | 2 µM | 13.32 µL | - |
| Dylight405-Mal | 1 mM | 5 µM | 3.33 µL | - |
| BSA 9000S | 100x | 1x | 3.33 µL | 3.33 µL |
| Φ29 DNA polymerase | 10 U/µL | 0.2 U/µL | 13.3 µL | - |
| **mQ** | - | - | 153.2 µL | 243.4 µL |
| Total | - | - | 326.3 µL | 333 µL |

Figure 2: THA and PEG-DVS solution recipes for in-gel RCA.
Due to viscous nature of the THA, its solution is prepared at room temperature. 12 mg-aliquots of THA are solubilised with a pre-mix of the compounds highlighted in bold. The rest of the components are added in order, from top to bottom. BSA and Φ29 DNA polymerase are added in the THA mix around 30 min after the addition of the maleimide-bound reagents. The mixture is then set on ice. On the other hand, preparation of the PEG-DVS solution is straightforward, and on ice from the beginning. Moreover, the Φ29 Reaction Buffer is prepared from NEB M0269S recipe, without the DTT.

### 3. Protein expression

For the IVTT-based expression, S17 Cell extract was prepared according to Rémi Sieskind's protocols, along with the PEP-based energy buffer, and the amino acids mix[178]. Aliquots of the three preparations are then flash-frozen and stored at -80°C for several months.

Before each new preparation of IVTT mixture, a fresh solution of maltodextrin has to be prepared (Sigma-Aldrich, 419672-100G), 18 mg dissolved in 100 µL mQ (15.3%$_{w/w}$). Used as a substrate for glycolysis, it drives the reprocessing of pyrophosphates. The PEG$_{8000}$ (Sigma-Aldrich, 89510-250G-F) solution is taken from a stock at 36%$_{w/w}$ in mQ, stored at 4°C for several months. The PEG acts as a crowding agent, and enhances expression yields. The $\chi_6$-sequence, used to inhibit the degradation of DNA by the RecBCD protein complex, is obtained through the annealing of a 1:1 mixture of the following sequence and its reverse complement, at 500 µM each. Once heated at 98°C, a slow decrease in temperature down to 20°C ensures the duplex-state of the sequences, and can be directly used in the IVTT mix.

5'-TCACTTCACT**GCTGGTGG**CCACT**GCTGGTGG**CCACT**GCTGGTGG**
CCACT**GCTGGTGG**CCACT**GCTGGTGG**CCACT**GCTGGTGG**CCA-3'

The amino acid mix contains the necessary 20 amino acids for protein translation, whereas the energy mix consists of pretty much everything else that the expression needs: NTPs, NAD, tRNAs, enzyme cofactors for glycolysis, polycations for T7 RNA polymerase stabilisation, phosphoenolpyruvic acids (PEP) for ATP synthesis, and oxalates (K-Ox) to inhibit the consumption of ATP in the system.

Due to the swelling of THA beads in low ionic strength solutions[178], the salts necessary to the IVTT - potassium glutamate and magnesium glutamate - are extracted from the master mix, in order to prepare a buffer for the hydrogel beads. Prior to the protein expression step, the beads can be extensively resuspended and washed with this buffer, without any risk of shrinking/swelling due to the change of ionic strength between successive media. During our directed evolution process, THA beads obtained after RCA are washed several times with the buffer, so that a corresponding volume of beads in their buffer is used in the IVTT reaction (~30 µL for 100 µL of total mix). The final mix is incubated at 34°C, for two hours in the case of being re-encapsulated in droplets, in our experimental platform.

| | Initial concentration | Final concentration | Volume |
|---|---|---|---|
| S17 Cell extract | - | | 33 µL |
| Amino acids mix | 17 mM | 3 mM | 17.7 µL |
| Energy Buffer | 14x | 1x | 7.1 µL |
| PEG 8000 | 36% | 2% | 5.6 µL |
| Maltodextrin | 500x | 35x | 7 µL |
| $\chi_6$-sequence | 250 µM | 50 µM | 2 µL |
| Potassium Glutamate (KGlu) | 3 M | 80 mM | 2.7 µL |
| Magnesium Glutamate (MgGlu) | 200 mM | 4 mM | 2 µL |
| DNA template | - | 1 nM | 25 µL |
| mQ | - | - | |
| Total | | | 100 µL |

Figure 3.1: S17 Cell-extract based *In Vitro* Transcription Translation (IVTT) mixture recipe.

| | Initial concentration | Final concentration | Volume |
|---|---|---|---|
| Potassium Glutamate (KGlu) | 3 M | 275 mM | 91.8 µL |
| Magnesium Glutamate (MgGlu) | 1 M | 13.5 mM | 13.5 µL |
| Tween 20 | - | $0.1\%_{v/v}$ | 1 µL |
| mQ | - | - | 893.7 µL |
| Total | - | 14x | 1000 µL |

Figure 3.2: *In Vitro* Transcription Translation wash recipe.

For the bacteria-based expression, KRX (Promega, L3002) and BL21(DE3) (NEB, C2527H) strains were used for the bacterial expression of proteins. After following the transformation protocols appropriate for the strain, samples are set on agar selection plates overnight at 37°C. Individual clones are then picked and resuspended in 5 mL of LB medium, along with the adequate dilution of antibiotic. These liquid cultures are then incubated overnight at 37°C. Upon retrieval, they are diluted 100 times with LB and antibiotic again, and set to grow at 37°C. After reaching an OD600nm around 0.5-0.6, protein expression in cultures is induced with either L-rhamnose (Sigma-Aldrich, R3875-25G) for the KRX cells or IPTG (Sigma-Aldrich, I5502-5G) for the BL21(DE3) ones, at 37°C for a few hours.

## 4. Protein purification

In order to be able to purify proteins, whatever the system, polyhistidine tags have to be inserted at the C terminus of the proteins. Indeed, all methods that were used are based on the affinity of this tag with metal ions, most notably cobalt ions with the affinity chromatography columns used here.

In the case of bacteria-based expression, cultures are centrifuged in pre-weighed 50mL Falcon tubes, the supernatant is removed and the pellet weighed. 2mL of sonication buffer are added per 100mg of bacteria pellet, the latter resuspended by vortexing. An ultrasonic cell disruptor (Branson Ultrasonics Sonifier S-250A, 101063196R) is then used to break up the cells, with 5 cycles of switching between sonication and cooling steps of 30s each (Power 4, duty cycle 40%). A fraction of the mixture is kept for further analysis. In order to separate the lysate from the solid debris, the samples are centrifuged at 18,500×g for 15 minutes, at 4°C. The supernatant is then put through His GraviTrap TALON (GE Healthcare Life Sciences, 29-0005-94) affinity chromatography columns.

In the case of IVTT- based expression of proteins, the samples are simply diluted in Binding Buffer, up to 1 mL, and put through His SpinTrap TALON (GE Healthcare Life Sciences, 29-0005-93) affinity chromatography columns.

In both cases, the columns are preventively equilibrated with either sonication or binding buffer. The samples are then injected in the columns. These are washed with the same sonication/binding buffer two to three times, followed by a similar operation with the washing buffer. All of these flow-through fractions are kept for further analysis. Finally, elution buffer is put through the columns several times, and every fraction is kept separated.

Samples of the gathered fractions are then analysed with protein gels, the protocols being detailed afterwards.

| | Sonication Buffer | Washing Buffer | Elution Buffer |
|---|---|---|---|
| [$NaH_2\ PO_4$] | 40 mM | 40 mM | 40 mM |
| [NaCl] | 300 mM | 300 mM | 300 mM |
| [Imidazole] | 10 mM | 20 mM | 300 mM |
| [DTT] | 1 mM | 1 mM | 1 mM |
| [Triton X-100] | $0.05\%_{v/v}$ | - | - |

Figure 4.1: Bacteria-based protein purification buffers recipe.
Triton X-100 is a surfactant that helps cell membrane dismantle during sonication.

| | Binding Buffer | Washing Buffer | Elution Buffer |
|---|---|---|---|
| [$NaH_2$ $PO_4$] | 50 mM | 50 mM | 50 mM |
| [NaCl] | 300 mM | 300 mM | 300 mM |
| [Imidazole] | - | 5 mM | 150 mM |

Figure 4.2: IVTT-based protein purification buffers recipe.
The buffers are prepared from stock solutions: 2M Imidazole, 200 mM NaCl. The sodium phosphate solution is prepared from solid monobasic and dibasic compounds, the pH being then set to 7.4.

Protein gels: SDS-PAGE, Coomassie blue and silver staining analysis

The protein gels used were bought from Thermo Fisher (NP0323BOX), and run with the corresponding MOPS-buffer (Thermofisher, NP0001). They are then either stained using the SimplyBlue (Thermofisher, LC6065) staining protocol, or the kit and protocol from Thermofisher (24612) for silver staining.

## 5. DNA Amplification

In the case of the PCR performed by the KRX bacteria expressing KlenTaq, we worked with samples diluted to concentrations of around 100 bacteria / nL, something that was already optimised by a previous PhD student in the laboratory, Adèle Dramé-Maigne. This concentration is sufficient for proper auto-amplification, without being toxic for the PCR reaction.

| | Initial concentration | Final concentration | Final volume |
|---|---|---|---|
| Thermopol | 10x | 1x | 10 µL |
| dNTPs | 10 mM | 200 µM | 2 µL |
| Primer Forward | 10 µM | 200 nM | 2 µL |
| Primer Reverse | 10 µM | 200 nM | 2 µL |
| $MgSO_4$ | 75 mM | 1.5 mM | 2 µL |
| Bacteria | - | 100 bacteria / nL | 10 µL |
| mQ water | | | 72 µL |
| Total | | | 100 µL |

Figure 5.1: PCR mixture recipe for KlenTaq expressing bacteria (colony PCR).

| Step | Temperature | Time |
|---|---|---|
| Annealing | 95°C | 2-5 minutes |
| 20 Cycles | 95°C<br>55-65°C<br>72°C | 15-30 seconds<br>15-30 seconds<br>1 minute per kb |
| Final Extension | 72°C | 5 minutes |
| Hold | 4°C | |

Figure 5.2: PCR protocol for KlenTaq-expressing bacteria (auto-PCR).

For the final step of the process, the self-selecting PCR in the hydrogel beads, the following protocol is used. The only difference being that the DNA template and DNA polymerase fractions are replaced by an equivalent volume of THA beads, washed a number of times in 1x Thermopol Buffer (NEB, B9004S). The resulting master mix is 1.11x concentrated in Thermopol, but we ensured that the PCR process was not hindered by the excess of salts. After encapsulation, but before PCR cycling, an incubation step at 4°C for 1h is set to ensure that the Thrombin protease manages to cleave all of the KlenTaq proteins in the hydrogel beads.

| | Initial concentration | Final concentration | Final volume |
|---|---|---|---|
| Thermopol Buffer | 10x | 1x | 10 µL |
| dNTPs | 10 mM | 200 µM | 2 µL |
| Forward Primer | 10 µM | 200 nM | 2 µL |
| Reverse Primer | 10 µM | 200 nM | 2 µL |
| $MgSO_4$ | 75 mM | 1.5 mM | 2 µL |
| DNA template | 4 nM | 400 pM | 10 µL |
| DNA polymerase | - | 1% | 1 µL |
| Pluronic F-127 | $8\%_{w/w}$ | $0.4\%_{w/w}$ | 5 µL |
| Yeast RNA | 100ng/µL | $1\%_{v/v}$ | 1 µL |
| Thrombin protease | 1 U/µL | - | 1 µL |
| mQ water | - | - | 65 µL |
| Total | - | - | 100 µL |

Figure 5.1: PCR mixture recipe for KlenTaq self-selecting step.
$MgSO_4$ is supplied by Sigma-Aldrich (M5921- 500G).

| Step | Temperature | Time |
|---|---|---|
| Annealing | 95°C | 2-5 minutes |
| 20 Cycles | 95°C<br>68°C | 30 seconds<br>7 minutes |
| Final Extension | 68°C | 10 minutes |
| Hold | 4°C | |

Figure 5.2: PCR protocol for KlenTaq self-selecting step.

At the end of the whole process, a qPCR reaction is set in order to quantify the amount of genetic material present in the beads, and *in fine* the relative amplification of each variant.

| | Initial concentration | Final concentration | Final volume |
|---|---|---|---|
| Standard Taq Buffer | 10x | 1x | 10 µL |
| dNTPs | 10 mM | 200 µM | 2 µL |
| Primer Forward | 10 µM | 200 nM | 2 µL |
| Primer Reverse | 10 µM | 200 nM | 2 µL |
| Evagreen | 20x | 0.5x | 2.5 µL |
| HS Taq | - | - | 0.5 µL |
| mQ water | | | 61 µL |
| Selection Samples | | | 10 x 2 µL |
| Total | | | 10 x 10 µL |

Figure 5.3: qPCR mixture recipe.
Hot Start Taq DNA polymerase is supplied by NEB (M0495). Samples gathered from the self-selection step were diluted into ranges of concentration, to be compared with ranges of standard concentration (from 1fM to 10pM). Reactions were monitored in CFX96 qPCR machines from Biorad.

## 6. Study of antibiotic-induced ribosomal noise

When studying the effects of aminoglycoside antibiotics on protein translation, we used the same IVTT protocols that are previously described, the only difference being the addition of diluted antibiotic (either kanamycin or streptomycin). The only used plasmids were either from GFP-HisTag or HisTag-GFP constructs.

Individual samples consisted of at least 10 µL of IVTT mix, containing 1 µL of ten times concentrated antibiotic solution, in order to obtain the desired final concentration. Samples were then incubated overnight in a CLARIOstar Plus plate reader at 34°C, 250 RPM, in order to record emitted fluorescence. In each experiment, at least 2 replicas were set for each antibiotic concentration, and for each antibiotic, as well as 3 positive controls replica that did not contain any antibiotic (replaced by mQ water in the sample recipe).

| | Initial concentration | Final concentration | Volume |
|---|---|---|---|
| S17 Cell extract | - | | 33 µL |
| Amino acids mix | 17 mM | 3 mM | 17.7 µL |
| Energy Buffer | 14x | 1x | 7.1 µL |
| PEG 8000 | 36% | 2% | 5.6 µL |
| Maltodextrin | 500x | 35x | 7 µL |
| $\chi_6$-sequence | 250 µM | 50 µM | 2 µL |
| Potassium Glutamate (KGlu) | 3 M | 80 mM | 2.7 µL |
| Magnesium Glutamate (MgGlu) | 200 mM | 4 mM | 2 µL |
| Antibiotic | - | - | 10 µL |
| DNA template | - | 1 nM | 15 µL |
| mQ | - | - | |
| Total | | | 100 µL |

Figure 6.1: *In Vitro* Transcription Translation (IVTT) mixture recipe for noisy translation.

After purification through affinity chromatography columns (GE Healthcare Life Sciences, 29-0005-93), samples were loaded on polyacrylamide gels (Thermo Fisher NP0323BOX) and subsequently stained with silver staining (Thermo Fisher 24612). Analysis was performed using the free GelAnalyzer software.

## 7. Miscellaneous

Agarose gels were prepared with 0.8%$_{w/v}$ agarose powder from Sigma Aldrich (A3139). Electrophoresis was run in 1x TBE (Tris base, boric acid, EDTA) for 45min at 120V. Samples were 50% diluted with SDS-Loading dye before being loaded onto the gels.

Thrombin cleavage protocol:

In a sterile Eppendorf tube, the target protein is mixed with bovine thrombin protease (GE Healthcare, 27-0846-01) and Thermopol Buffer. The mix is gently pipetted up and down and left at room temperature. In order to assess the kinetics of the reaction, samples are taken out of the mix every hour and denatured before being loaded on a protein gel. According to the purveyor, 1 unit of Thrombin can cleave 100 µg of GST fusion protein in 16h at 22°C when in 1xPBS.

| | Target protein (100µg) |
|---|---|
| Target Protein | 10 µl |
| Thermopol Buffer | 39 µl |
| Thrombin Enzyme (1 U/µL) | 1 µl |

Figure 7.1: Thrombin cleavage reaction mix.

TEV protease cleavage protocol:

In a sterile Eppendorf tube, the target protein is mixed with TEV protease (NEB, P8112S) and TEV protease reaction buffer. The mix is gently pipetted up and down and left 1h at 30°C. According to the purveyor, 10 units of Thrombin can cleave 15 µg of protein in 1h at 30°C or overnight at 4°C.

| | Volume | Final conc. |
|---|---|---|
| Fusion protein | 1 µL | 20 µg |
| TEV Protease (10U/µL) | 1 µL | 10U |
| 20X TEV Protease Reaction Buffer | 7.5 µL | 1x |
| 100 mM DTT | 1.5 µL | 1 mM |
| Nuclease-free Water | Up to 150 µL | - |

Figure 7.2: TEV (Tobacco Etch Virus) cleavage reaction mix.

| Name | Sequence (5’ → 3’) |
|---|---|
| primf_qKT_v2: | TTGGCTGCTGGTTGCACTGG |
| primr_qKT_v2: | GCTTCACGCGGAACACCAAAC |
| iR274 | TTATTCTTTGGCGCTCAGCCAATC |
| iR279 | ATGCGTCTGCTGCATGAATTTGG |
| iV019 | CGTAGAGGATCGAGACCTCGATCCCGCG |
| iV020 | TGTTCGCACCCAGACAGTTCCAGCTTGCC |

Figure 7.3: List of primers used for PCR of KlenTaq gene.

For the first optimisation steps of the process (KlenTaq auto-PCR in bacteria), we used the primer pair primf_qKT_v2/primr_qKT_v2, that were designed by a previous PhD student of the team, Adèle Dramé-Maigne, which corresponds to a 150 bp amplicon. The next step was to optimise the PCR step for the complete amplification of the polymerase gene, so we switched to primers for a 1.6 kb amplicon (iR274/iR279), which is the entirety of KlenTaq’s gene, these primers binding at the extremities of it. However, due to the issues of unspecificity that we faced when studying cell-free approaches, we redesigned the primer pair along the guidelines of the Sigma Aldrich “KlenTaq LA DNA Polymerase Mix Technical Bulletin”, and ended up with the pair iV019/iV020.

# References


1. Radzicka, A. & Wolfenden, R. A proficient enzyme. *Science* **267**, 90–93 (1995).
2. Kurland, C. G. Molecular characterization of ribonucleic acid from Escherichia coli ribosomes: I. Isolation and molecular weights. *J. Mol. Biol.* **2**, 83–91 (1960).
3. Nissen, P. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. *Science* **289**, 920–930 (2000).
4. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci.* **37**, 205 (1951).
5. Chou, K.-C. & Cai, Y.-D. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins Struct. Funct. Bioinforma.* **53**, 282–289 (2003).
6. Moutevelis, E. & Woolfson, D. N. A Periodic Table of Coiled-Coil Protein Structures. *J. Mol. Biol.* **385**, 726–732 (2009).
7. Johnson, K. A. & Goody, R. S. The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. *Biochemistry* **50**, 8264–8269 (2011).
8. Fraser, T. R. On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebata, coedia, morphia, and nicotia. 19 (1865).
9. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Dtsch. Chem. Ges.* **27**, 2985–2993 (1894).
10. Bennett, J. M. & Kendrew, J. C. The computation of Fourier synthesis with a digital electronic calculating machine. *Acta Crystallogr.* **5**, 109–116 (1952).
11. Spudich, J. L. & Koshland, D. E. Non-genetic individuality: chance in the single cell. *Nature* **262**, 467–471 (1976).

12. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223–230 (1973).

13. Stetter, K. O. History of discovery of the first hyperthermophiles. *Extremophiles* **10**, 357–362 (2006).

14. Wang, K. *et al.* Morphology and genome of a snailfish from the Mariana Trench provide insights into deep-sea adaptation. *Nat. Ecol. Evol.* **3**, 823–833 (2019).

15. Chien, A., Edgar, D. B. & Trela, J. M. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J. Bacteriol.* **127**, 1550–1557 (1976).

16. Morton, R. K. [6] Methods of extraction of enzymes from animal tissues. in *Methods in Enzymology* vol. 1 25–51 (Elsevier, 1955).

17. Nason, A. [8] Extraction of soluble enzymes from higher plants. in *Methods in Enzymology* vol. 1 62–63 (Elsevier, 1955).

18. Gunsalus, I. C. [7] Extraction of enzymes from microorganisms (bacteria and yeast). in *Methods in Enzymology* vol. 1 51–62 (Elsevier, 1955).

19. Porter, R. R. [12] The partition chromatography of enzymes. in *Methods in Enzymology* vol. 1 98–112 (Academic Press, 1955).

20. Hirs, C. H. W. [13] Chromatography of enzymes on ion exchange resins. in *Methods in Enzymology* vol. 1 113–125 (Academic Press, 1955).

21. Heppel, L. [15] Separation of proteins from nucleic acids. in *Methods in Enzymology* vol. 1 137–138 (Academic Press, 1955).

22. Hayaishi, O. [14] Special techniques for bacterial enzymes. Enrichment culture and adaptive enzymes. in *Methods in Enzymology* vol. 1 126–137 (Academic Press, 1955).

23. Jelsch, C. *et al.* Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin. *Proc. Natl. Acad. Sci.* **97**, 3171–3176 (2000).

24. Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **8**, 3 (2001).

25. Fiaux, J., Bertelsen, E. B., Horwich, A. L. & Wüthrich, K. NMR analysis of a 900K GroEL–GroES complex. *Nature* **418**, 207–211 (2002).

26. Adrian, M., Dubochet, J., Lepault, J. & McDowall, A. W. Cryo-electron microscopy of viruses. *Nature* **308**, 32–36 (1984).

27. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **161**, 438–449 (2015).

28. Schwede, T. Protein Modeling: What Happened to the "Protein Structure Gap"? *Structure* **21**, 1531–1540 (2013).

29. Zhang, Y. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* **18**, 342–348 (2008).

30. Stevenson, D. C. The Internet classics archive. http://classics.mit.edu/ (1994).

31. Shah, D. M. S. Pre-Darwinian Muslim Scholars' Views on Evolution. 18 (2013).

32. Bowler, P. J. *Evolution : the history of an idea*. (2003).

33. Larson, E. J. *Evolution : the remarkable history of a scientific theory*. (Modern Library, 2004).

34. Russel. Contributions to the theory of natural selection : a series of essays. http://www.biodiversitylibrary.org/bibliography/46265 (1871).

35. Bateson, W., Mendel, G. & Wheeler, W. M. *Mendel's principles of heredity; a defence*. (University press, 1902).

36. Alter, S. G. Evolution: The History of an Idea. Third Edition. By Peter J Bowler. *Q. Rev. Biol.* **79**, 305–305 (2004).

37. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1919).

38. Fisher, R. A. & others. 014: On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. (1921).

39. Galton, F. Regression Towards Mediocrity in Hereditary Stature. *J. Anthropol. Inst. G. B. Irel.* **1**, 246–263 (1886).

40. Pearson, K. Downloaded from https://royalsocietypublishing.org/ on 21 April 202. 27.

41. Fisher, R. A. *The genetical theory of natural selection,*. (The Clarendon Press, 1930).

42. Haldane, J. B. S. A MATHEMATICAL THEORY OF NATURAL AND ARTIFICIAL SELECTION--I. 32.

43. Clarke, C. A. Evolution in reverse: clean air and the peppered moth. 11.

44. Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. (1932).

45. Hagedoorn, A. L. & Hagedoorn, A. C. (Vorstheuvel L. B. *The relative value of the processes causing evolution,*. (M. Nijhoff, 1921).

46. Simpson. Tempo and mode in evolution. *Columbia Univ. Press N. Y.* (1945).

47. Dietrich, M. R. & Skipper, R. A. A Shifting Terrain: A Brief History of the Adaptive Landscape. in *The Adaptive Landscape in Evolutionary Biology* (eds. Svensson, E. & Calsbeek, R.) 3–15 (Oxford University Press, 2013). doi:10.1093/acprof:oso/9780199595372.003.0001.

48. Provine, W. B. The role of mathematical population geneticists in the evolutionary synthesis of the 1930s and 1940s. *Stud. Hist. Biol.* **2**, 167–192 (1978).

49. Gavrilets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312 (1997).

50. Kaplan, J. The end of the adaptive landscape metaphor? *Biol. Philos.* **23**, 625–638 (2008).

51. Ruse, M. Are Pictures Really Necessary? The Case of Sewell Wright's 'Adaptive Landscapes'. *PSA Proc. Bienn. Meet. Philos. Sci. Assoc.* **1990**, 63–77 (1990).

52. Dobzhansky, T. & Gould, S. J. *Genetics and the Origin of Species*. (Columbia University Press, 1982).

53. Ford Edmund Brisco. *Ecological genetics / E. B. Ford, ...* (Methuen J. Wiley, 1964).

54. De Beer, G. *Embryos and Ancestors*. (Clarendon Press, 1958).

55. Stebbins, G. Variation and Evolution in Plants. in (1950).

56. Tiselius, A. A new apparatus for electrophoretic analysis of colloidal mixtures. *Trans Faraday Soc* **33**, 524–531 (1937).

57. McQuillen, K. THE REPLICATION OF DNA IN ESCHERICHIA COLI* BY MATTHEW MESELSON AND FRANKLIN W. STAHL. **44**, 12 (1958).

58. Brenner, S., Benzer, S. & Barnett, L. Distribution of Proflavin-Induced Mutations in the Genetic Fine Structure. *Nature* **182**, 983–985 (1958).

59. Crick, F. H. On protein synthesis. in *Symp Soc Exp Biol* vol. 12 8 (1958).

60. Williams, G. C. *Adaptation and natural selection: a critique of some current evolutionary thought*. (Princeton Univ. Press, 1996).

61. Dawkins, R. The Gene as the Unit of Selection. 287.

62. Wade, M. J. *et al.* Multilevel and kin selection in a connected world. *Nature* **463**, E8–E9 (2010).

63. McCarthy, B. J. & Holland, J. J. Denatured DNA as a direct template for in vitro protein synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **54**, 880–886 (1965).

64. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

65. Sanger, F. Determination of the structure of insulin opens the way to greater understanding of life processes. **129**, 5.

66. Zuckerkandl, E., Jones, R. T. & Pauling, L. A COMPARISON OF ANIMAL HEMOGLOBINS BY TRYPTIC PEPTIDE PATTERN ANALYSIS. *Proc. Natl. Acad. Sci. U. S. A.* **46**, 1349–1360 (1960).

67. Margoliash, E. & Schejter, A. Cytochrome c. in *Advances in Protein Chemistry* (eds. Anfinsen, C. B., Anson, M. L., Edsall, J. T. & Richards, F. M.) vol. 21 113–286 (Academic Press, 1966).

68. Kimura, M. THE RATE OF MOLECULAR EVOLUTION CONSIDERED FROM THE STANDPOINT OF POPULATION GENETICS. *Proc. Natl. Acad. Sci.* **63**, 1181–1188 (1969).

69. King, J. L. & Jukes, T. H. Most evolutionary change in proteins may be due to neutral mutations and genetic drift. **164**, 11 (1969).

70. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci.* **99**, 16134–16137 (2002).

71. Richardson, J. S. & Richardson, D. C. The de novo design of protein structures. 6 (1989).

72. Merrifield, R. B. **Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide**. *J. Am. Chem. Soc.* **85**, 2149–2154 (1963).

73. Gutte, B. A synthetic 70-amino acid residue analog of ribonuclease S-protein with enzymic activity. *J. Biol. Chem.* **250**, 889–904 (1975).

74. Moser, R., Thomas, R. M. & Gutte, B. An artificial crystalline DDT-binding polypeptide. *FEBS Lett.* **157**, 247–251 (1983).

75. Dryden, D. T. F., Thomson, A. R. & White, J. H. How much of protein sequence space has been explored by life on Earth? *J. R. Soc. Interface* **5**, 953–956 (2008).

76. Mandell, D. J. & Kortemme, T. Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420–428 (2009).

77. Martí-Renom, M. A. *et al.* Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).

78. Orry, A. J. W. & Abagyan, R. Preparation and Refinement of Model Protein–Ligand Complexes. in *Homology Modeling: Methods and Protocols* (eds. Orry, A. J. W. & Abagyan, R.) 351–373 (Humana Press, 2012). doi:10.1007/978-1-61779-588-6_16.

79. Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).

80. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

81. Kaczanowski, S. & Zielenkiewicz, P. Why similar protein sequences encode similar three-dimensional structures? *Theor. Chem. Acc.* **125**, 643–650 (2010).

82. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).

83. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins Struct. Funct. Genet.* **17**, 355–362 (1993).

84. Lazaridis, T. & Karplus, M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, 477–487 (1999).

85. Baker, D. Protein Structure Prediction and Structural Genomics. *Science* **294**, 93–96 (2001).

86. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. *Rep. Prog. Phys.* **81**, 032601 (2018).

87. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

88. Callaway, E. 'IT WILL CHANGE EVERYTHING': AI MAKES GIGANTIC LEAP IN SOLVING PROTEIN STRUCTURES. 2.

89. Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D. & Voelz, V. A. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* **17**, 342–346 (2007).

90. Baker, D. A surprising simplicity to protein folding. *Nature* **405**, 39–42 (2000).

91. Bonneau, R. & Baker, D. Ab Initio Protein Structure Prediction: Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189 (2001).

92. Wollacott, A. M., Zanghellini, A., Murphy, P. & Baker, D. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci.* **16**, 165–175 (2006).

93. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).

94. Mills, D. R., Peterson, R. L. & Spiegelman, S. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 217–224 (1967).

95. Kacian, D. L., Mills, D. R., Kramer, F. R. & Spiegelman, S. A Replicating RNA Molecule Suitable for a Detailed Analysis of Extracellular Evolution and Replication. *Proc. Natl. Acad. Sci.* **69**, 3038–3042 (1972).

96. Cirino, P. C. & Arnold, F. H. Exploring the Diversity of Heme Enzymes through Directed Evolution. in *Directed Molecular Evolution of Proteins* (eds. Brakmann, S. & Johnsson, K.) 215–243 (Wiley-VCH Verlag GmbH & Co. KGaA, 2002). doi:10.1002/3527600647.ch10.

97. Hall, B. G. Experimental evolution of a new enzymatic function. II. Evolution of multiple functions for ebg enzyme in E. coli. *Genetics* **89**, 453–465 (1978).

98. Chen, K. & Arnold, F. H. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci.* **90**, 5618–5622 (1993).

99. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

100. Brakmann, S. & Schwienhorst, A. *Evolutionary methods in biotechnology: clever tricks for directed evolution*. (John Wiley & Sons, 2006).

101. Leung, D. W. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique* **1**, 11–15 (1989).

102. Cadwell, R. C. & Joyce, G. F. Randomization of genes by PCR mutagenesis. *Genome Res.* **2**, 28–33 (1992).

103. Hermes, J. D., Blacklow, S. C. & Knowles, J. R. Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. *Proc. Natl. Acad. Sci.* **87**, 696–700 (1990).

104. Doi, N. & Yanagawa, H. Genotype-Phenotype Linkage for Directed Evolution and Screening of Combinatorial Protein Libraries. *Comb. Chem. High Throughput Screen.* **4**, 497–509 (2001).

105. Leemhuis, H., Stein, V., Griffiths, A. & Hollfelder, F. New genotype–phenotype linkages for directed evolution of functional proteins. *Curr. Opin. Struct. Biol.* **15**, 472–478 (2005).

106. Smith, G. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).

107. Freudl, R., MacIntyre, S., Degen, M. & Henning, U. Cell surface exposure of the outer membrane protein OmpA of Escherichia coli K-12. *J. Mol. Biol.* **188**, 491–494 (1986).

108. Yan, X. & Xu, Z. Ribosome-display technology: applications for directed evolution of functional proteins. *Drug Discov. Today* **11**, 911–916 (2006).

109. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci.* **94**, 12297–12302 (1997).

110. Lutz, S. Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **21**, 734–743 (2010).

111. Sergeeva, A., Kolonin, M. G., Molldrem, J. J., Pasqualini, R. & Arap, W. Display technologies: Application for the discovery of drug and gene delivery agents☆. *Adv. Drug Deliv. Rev.* 34 (2006).

112. Tawfik, D. S. & Griffiths, A. D. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* **16**, 652–656 (1998).

113. Theberge, A. B. *et al.* Microdroplets in Microfluidics: An Evolving Platform for Discoveries in Chemistry and Biology. *Angew. Chem. Int. Ed.* **49**, 5846–5868 (2010).

114. Catherine, C., Lee, K.-H., Oh, S.-J. & Kim, D.-M. Cell-free platforms for flexible expression and screening of enzymes. *Biotechnol. Adv.* **31**, 797–803 (2013).

115. Takátsy, G. The Use of Spiral Loops in Serological and Virological Micro-Methods. *Acta Microbiol. Immunol. Hung.* **50**, 369–82; discussion 382 (2003).

116. Attene-Ramos, M. S., Austin, C. P. & Xia, M. High Throughput Screening. in *Encyclopedia of Toxicology* 916–917 (Elsevier, 2014). doi:10.1016/B978-0-12-386454-3.00209-8.

117. Antypas, H., Veses-Garcia, M., Weibull, E., Andersson-Svahn, H. & Richter-Dahlfors, A. A universal platform for selection and high-resolution phenotypic screening of bacterial mutants using the nanowell slide. *Lab. Chip* **18**, 1767–1777 (2018).

118. Fulwyler, M. J. Electronic Separation of Biological Cells by Volume. *Science* **150**, 910 (1965).

119. Black, C. B., Duensing, T. D., Trinkle, L. S. & Dunlay, R. T. Cell-Based Screening Using High-Throughput Flow Cytometry. *ASSAY Drug Dev. Technol.* **9**, 13–20 (2011).

120. Xiao, H., Bao, Z. & Zhao, H. High Throughput Screening and Selection Methods for Directed Enzyme Evolution. *Ind. Eng. Chem. Res.* **54**, 4011–4020 (2015).

121. Turner, N. J. Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* **5**, 567–573 (2009).

122. MacBeath, G. Redesigning Enzyme Topology by Directed Evolution. *Science* **279**, 1958–1961 (1998).

123. Gatti-Lafranconi, P. *et al.* Evolution of Stability in a Cold-Active Enzyme Elicits Specificity Relaxation and Highlights Substrate-Related Effects on Temperature Adaptation. *J. Mol. Biol.* **395**, 155–166 (2010).

124. Toscano, M. D., Woycechowsky, K. J. & Hilvert, D. Minimalist Active-Site Redesign: Teaching Old Enzymes New Tricks. *Angew. Chem. Int. Ed.* **46**, 3212–3236 (2007).

125. Hawkins, R. E., Russell, S. J. & Winter, G. Selection of phage antibodies by binding affinity. *J. Mol. Biol.* **226**, 889–896 (1992).

126. Rebar, E. & Pabo, C. Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671–673 (1994).

127. Santoro, S. W., Wang, L., Herberich, B., King, D. S. & Schultz, P. G. An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nat. Biotechnol.* **20**, 1044–1048 (2002).

128. Borsuk, S. *et al.* Auxotrophic complementation as a selectable marker for stable expression of foreign antigens in Mycobacterium bovis BCG. *Tuberculosis* **87**, 474–480 (2007).

129. Meredith, H. R., Srimani, J. K., Lee, A. J., Lopatkin, A. J. & You, L. Collective antibiotic tolerance: mechanisms, dynamics and intervention. *Nat. Chem. Biol.* **11**, 182–188 (2015).

130. Ghadessy, F. J., Ong, J. L. & Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci.* **98**, 4552–4557 (2001).

131. Saiki, R. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).

132. Delbrück, M. Statistical Fluctuations in Autocatalytic Reactions. *J. Chem. Phys.* **8**, 120–124 (1940).

133. Delbrück, M. The Burst Size Distribution in the Growth of Bacterial Viruses (Bacteriophages). *J. Bacteriol.* **50**, 131–135 (1945).

134. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).

135. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).

136. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nat. Genet.* **31**, 69–73 (2002).

137. Elowitz, M. B. Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).

138. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci.* **99**, 12795 (2002).

139. Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).

140. Orphanides, G. & Reinberg, D. A Unified Theory of Gene Expression. *Cell* **108**, 439–451 (2002).

141. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123**, 1025–1036 (2005).

142. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional Pulsing of a Developmental Gene. *Curr. Biol.* **16**, 1018–1025 (2006).

143. Blake, W. J., KÆrn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).

144. Bokes, P., King, J. R., Wood, A. T. A. & Loose, M. Transcriptional Bursting Diversifies the Behaviour of a Toggle Switch: Hybrid Simulation of Stochastic Gene Expression. *Bull. Math. Biol.* **75**, 351–371 (2013).

145. Simpson, M. L., Cox, C. D. & Sayler, G. S. Frequency domain analysis of noise in autoregulated gene circuits. *Proc. Natl. Acad. Sci.* **100**, 4551–4556 (2003).

146. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).

147. Orrell, D. & Bolouri, H. Control of internal and external noise in genetic regulatory networks. *J. Theor. Biol.* **230**, 301–312 (2004).

148. Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci.* **71**, 4135–4139 (1974).

149. Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57**, 587–595 (1975).

150. Savir, Y. & Tlusty, T. Conformational Proofreading: The Impact of Conformational Changes on the Specificity of Molecular Recognition. *PLoS ONE* **2**, e468 (2007).

151. Reardon, J. T. & Sancar, A. Thermodynamic Cooperativity and Kinetic Proofreading in DNA Damage Recognition and Repair. *Cell Cycle* **3**, 139–142 (2004).

152. McKeithan, T. W. Kinetic proofreading in T-cell receptor signal transduction. *Proc. Natl. Acad. Sci.* **92**, 5042–5046 (1995).

153. Savir, Y. & Tlusty, T. The Ribosome as an Optimal Decoder: A Lesson in Molecular Recognition. *Cell* **153**, 471–479 (2013).

154. Seger, J. & Brockmann, H. What is Bet-Hedging. *Oxf Surv Evol Biol* **4**, (1987).

155. Cohen, D. Optimizing reproduction in a randomly varying environment. *J. Theor. Biol.* **12**, 119–129 (1966).

156. Yasui, Y. Female multiple mating as a genetic bet-hedging strategy when mate choice criteria are unreliable. *Ecol. Res.* **16**, 605–616 (2001).

157. Kussell, E., Kishony, R., Balaban, N. Q. & Leibler, S. Bacterial Persistence. *Genetics* **169**, 1807–1814 (2005).

158. Philippi, T. & Seger, J. Hedging one's evolutionary bets, revisited. *Trends Ecol. Evol.* **4**, 41–44 (1989).

159. Thattai, M. & van Oudenaarden, A. Stochastic Gene Expression in Fluctuating Environments. *Genetics* **167**, 523–530 (2004).

160. Kitano, H. Biological robustness. *Nat. Rev. Genet.* **5**, 826–837 (2004).

161. Gerhart, J. & Kirschner, M. The theory of facilitated variation. *Proc. Natl. Acad. Sci.* **104**, 8582–8589 (2007).

162. Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. & Eisen, M. B. Noise Minimization in Eukaryotic Gene Expression. *PLoS Biol.* **2**, e137 (2004).

163. Lenski, R. E., Barrick, J. E. & Ofria, C. Balancing Robustness and Evolvability. *PLoS Biol.* **4**, e428 (2006).

164. Wagner, A. Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B Biol. Sci.* **275**, 91–100 (2008).

165. van Nimwegen, E., Crutchfield, J. P. & Huynen, M. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci.* **96**, 9716–9720 (1999).

166. Bloom, J. D. *et al.* Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.* **5**, 29 (2007).

167. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).

168. Brakmann, S. & Grzeszik, S. An Error-Prone T7 RNA Polymerase Mutant Generated by Directed Evolution. 8 (2001).

169. Goldsmith, M. & Tawfik, D. S. Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc. Natl. Acad. Sci.* **106**, 6197–6202 (2009).

170. Agarwal, D., Gregory, S. T. & O'Connor, M. Error-Prone and Error-Restrictive Mutations Affecting Ribosomal Protein S12. *J. Mol. Biol.* **410**, 1–9 (2011).

171. VanNice, J., Gregory, S. T., Kamath, D. & O'Connor, M. Alterations in ribosomal protein L19 that decrease the fidelity of translation. *Biochimie* **128–129**, 122–126 (2016).

172. Murakami, H., Ohta, A., Ashigai, H. & Suga, H. A highly flexible tRNA acylation method for non-natural polypeptide synthesis. *Nat. Methods* **3**, 357–359 (2006).

173. Ohuchi, M., Murakami, H. & Suga, H. The flexizyme system: a highly flexible tRNA aminoacylation tool for the translation apparatus. *Curr. Opin. Chem. Biol.* **11**, 537–542 (2007).

174. Kawahara-Kobayashi, A. *et al.* Simplification of the genetic code: restricted diversity of genetically encoded amino acids. *Nucleic Acids Res.* **40**, 10576–10584 (2012).

175. Dramé-Maigné, A. & Rondelez, Y. Directed Evolution of Enzymes based on in vitro Programmable Self-Replication. 18.

176. Montagne, K., Plasson, R., Sakai, Y., Fujii, T. & Rondelez, Y. Programming an *in vitro* DNA oscillator using a molecular networking strategy. *Mol. Syst. Biol.* **7**, 466 (2011).

177. Fujii, T. & Rondelez, Y. Predator–Prey Molecular Ecosystems. *ACS Nano* **7**, 27–34 (2013).

178. Sieskind, R. Evolution dirigée d'enzyme in vitro, auto-sélection par programmation moléculaire. 322.

179. Klenow, H. & Henningsen, I. Selective Elimination of the Exonuclease Activity of the Deoxyribonucleic Acid Polymerase from Escherichia coli B by Limited Proteolysis. *Proc. Natl. Acad. Sci.* **65**, 168–175 (1970).

180. Barnes, W. M. PCR amplification of up to 35-kb DNA with high fidelity and high yield from A bacteriophage templates. *Proc Natl Acad Sci USA* 5 (1994).

181. Barnes, W. The fidefity of Taq polymerase catalyzing P C R is improved by an N-terminal deletion. 7.

182. Kermekchiev, M. B., Kirilova, L. I., Vail, E. E. & Barnes, W. M. Mutants of Taq DNA polymerase resistant to PCR inhibitors allow DNA amplification from whole blood and crude soil samples. *Nucleic Acids Res.* **37**, e40–e40 (2009).

183. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random Mutagenesis by Error-Prone PCR. in *In Vitro Mutagenesis Protocols* (ed. Braman, J.) vol. 634 103–109 (Humana Press, 2010).

184. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

185. Shinkai, A., Patel, P. H. & Loeb, L. A. The Conserved Active Site Motif A of Escherichia coliDNA Polymerase I Is Highly Mutable. *J. Biol. Chem.* **276**, 18836–18842 (2001).

186. Thiele, J. *et al.* DNA-functionalized hydrogels for confined membrane-free in vitro transcription/translation. *Lab. Chip* **14**, 2651 (2014).

187. Segura, T. *et al.* Crosslinked hyaluronic acid hydrogels: a strategy to functionalize and pattern. *Biomaterials* **26**, 359–371 (2005).

188. Rakszewska, A., Stolper, R. J., Kolasa, A. B., Piruska, A. & Huck, W. T. S. Quantitative Single-Cell mRNA Analysis in Hydrogel Beads. *Angew. Chem. Int. Ed.* **55**, 6698–6701 (2016).

189. Amstad, E. *et al.* Robust scalable high throughput production of monodisperse drops. *Lab. Chip* **16**, 4163–4172 (2016).

190. Gish, G. & Eckstein, F. DNA and RNA sequence determination based on phosphorothioate chemistry. *Science* **240**, 1520–1522 (1988).

191. Juillerat, A. *et al.* Directed Evolution of O6-Alkylguanine-DNA Alkyltransferase for Efficient Labeling of Fusion Proteins with Small Molecules In Vivo. *Chem. Biol.* **10**, 313–317 (2003).

192. Mollwitz, B. *et al.* Directed Evolution of the Suicide Protein $O^6$-Alkylguanine-DNA Alkyltransferase for Increased Reactivity Results in an Alkylated Protein with Exceptional Stability. *Biochemistry* **51**, 986–994 (2012).

193. Jewett, M. C., Calhoun, K. A., Voloshin, A., Wuu, J. J. & Swartz, J. R. An integrated cell-free metabolic platform for protein production and synthetic biology. *Mol. Syst. Biol.* **4**, 220 (2008).

194. Shimizu, Y., Kanamori, T. & Ueda, T. Protein synthesis by pure translation systems. *Methods* **36**, 299–304 (2005).

195. Gregorio, N. E., Levine, M. Z. & Oza, J. P. A User's Guide to Cell-Free Protein Synthesis. *Methods Protoc.* **2**, 24 (2019).

196. Shim, J. *et al.* Ultrarapid Generation of Femtoliter Microfluidic Droplets for Single-Molecule-Counting Immunoassays. *ACS Nano* **7**, 5955–5964 (2013).

197. Magnet, S. & Blanchard, J. S. Molecular Insights into Aminoglycoside Action and Resistance. 22.

198. O'Sullivan, M. E. *et al.* Aminoglycoside ribosome interactions reveal novel conformational states at ambient temperature. *Nucleic Acids Res.* **46**, 9793–9804 (2018).

199. Ren, L., Rahman, M. S. & Humayun, M. Z. *Escherichia coli* Cells Exposed to Streptomycin Display a Mutator Phenotype. *J. Bacteriol.* **181**, 1043–1044 (1999).

200. Mingeot-Leclercq, M.-P., Glupczynski, Y. & Tulkens, P. M. Aminoglycosides: Activity and Resistance. *Antimicrob. Agents Chemother.* **43**, 727–737 (1999).

201. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

202. Switzer, R. C., Merril, C. R. & Shifrin, S. A highly sensitive silver stain for detecting proteins and peptides in polyacrylamide gels. *Anal. Biochem.* **98**, 231–237 (1979).

203. Lelong, C., Chevallet, M., Luche, S. & Rabilloud, T. Silver Staining of Proteins in 2DE Gels. in *Two-Dimensional Electrophoresis Protocols* (eds. Tyther, R. & Sheehan, D.) vol. 519 339–350 (Humana Press, 2009).

204. Spies, M., Amitani, I., Baskin, R. J. & Kowalczykowski, S. C. RecBCD Enzyme Switches Lead Motor Subunits in Response to χ Recognition. *Cell* **131**, 694–705 (2007).

205. Marshall, R., Maxwell, C. S., Collins, S. P., Beisel, C. L. & Noireaux, V. Short DNA containing χ sites enhances DNA stability and gene expression in *E. coli* cell-free

transcription-translation systems: Enhancing TXTL-Based Expression With χ-Site DNA. *Biotechnol. Bioeng.* **114**, 2137–2141 (2017).

206. Lowe, T., Sharefkin, J., Yang, S. Q. & Dieffenbach, C. W. A computer program for selection of oligonucleotide primers for polymerase chain reactions. *Nucleic Acids Res.* **18**, 1757–1761 (1990).

207. Dauty, E. & Verkman, A. S. Molecular crowding reduces to a similar extent the diffusion of small solutes and macromolecules: measurement by fluorescence correlation spectroscopy. *J. Mol. Recognit.* **17**, 441–447 (2004).

208. Tang, S. K. Y. & Whitesides, G. M. Basic Microfluidic and Soft Lithographic Techniques. 26.

## RÉSUMÉ

Les processus évolutifs prennent place à de nombreuses échelles différentes, de l'échelle macroscopique des populations et des espèces jusqu'aux phénomènes microscopiques agissant sur les protéines et les molécules des organismes vivants. Quel que soit le système, l'évolution naturelle a toujours besoin de diversité pour effectuer une sélection, afin de favoriser la survie des organismes les plus adaptés. En outre, l'une des caractéristiques communes de ces mécanismes multi-échelles est la présence constante de bruit. En effet, ses effets sont visibles de par la stochasticité des réactions moléculaires jusqu'à la dynamique des populations. Si l'on a pendant longtemps pensé que ces perturbations aléatoires étaient préjudiciables aux systèmes biologiques, un nombre croissant d'études ont émis l'hypothèse que ces fluctuations avaient en fait de nombreux impacts positifs sur le vivant. Notamment, en ce qui concerne le lissage des paysages adaptifs, et les effets sur la robustesse et la capacité à évoluer des populations.

## MOTS CLÉS

paysages adaptatifs ; bruit biologique ; évolution dirigée

## ABSTRACT

In this PhD project, we developed an experimental platform in order to quantitatively characterize the influence of artificial noise on protein populations. Based on Holliger's Compartmentalised Self Replication, the objects of the study are DNA polymerases, and in particular the KlenTaq polymerase, fundamental molecular biology tools, notoriously challenging to evolve. In this endeavour, libraries of variants of these proteins are generated, and then put through a fully in vitro directed evolution process, consisting of cycles of diversification, compartmentalisation of these variants in droplets using microfluidic devices, and selection for fitness using their ability to replicate and amplify their own genetic material. The resulting library could then be sequenced using NGS, and the data interpreted through Statistical Physics inspired analysis or Machine Learning methods. With the aim of adding an additional source of noise in the system, we also studied the effects of aminoglycoside antibiotics on the in vitro translation of the proteins, these being known to interfere with ribosomal accuracy. Being able to efficiently tune the extent of error-prone behaviour during protein synthesis would help in our understanding of the effects of background noise on biological processes such as protein translation and evolution.

## KEYWORDS

fitness landscapes; biological noise; directed evolution