Scuola Normale Superiore

CLASSE DI SCIENZE

PHD THESIS IN NEUROSCIENCE

From *in vitro* evolution to protein structure

Candidate:
Marco Fantini

Supervisors:
Annalisa Pastore
Antonino Cattaneo

Academic year 2019-2020

ABSTRACT

In the nanoscale, the machinery of life is mainly composed by macromolecules and macromolecular complexes that through their shapes create a network of interconnected mechanisms of biological processes. The relationship between shape and function of a biological molecule is the foundation of structural biology, that aims at studying the structure of a protein or a macromolecular complex to unveil the molecular mechanism through which it exerts its function. What about the reverse: is it possible by exploiting the function for which a protein was naturally selected to deduce the protein structure? To this aim we developed a method, called CAMELS (**C**oupling **A**nalysis by **M**olecular **E**volution **L**ibrary **S**equencing), able to obtain the structural features of a protein from an artificial selection based on that protein function. With CAMELS we tried to reconstruct the TEM-1 beta lactamase fold exclusively by generating and sequencing large libraries of mutational variants. Theoretically with this method it is possible to reconstruct the structure of a protein regardless of the species of origin or the phylogenetical time of emergence when a functional phenotypic selection of a protein is available. CAMELS allows us to obtain protein structures without needing to purify the protein beforehand.

Table of contents

Note:
Most of the experiments and results featured in this thesis were published on October 2019 on the journal “Molecular Biology and Evolution” (MBE) (Fantini et al., 2019) and some parts of this thesis are taken verbatim from the article. The datasets used in this study are publicly available with the SRA accession codes SRX5562455, SRX5562456 and SRX5562457 or with the BioProject accession code PRJNA528665 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA528665).

1 Introduction

Shape and function are linked. In our everyday life most of the physical objects we come in contact with are shaped in a way that allows them to exert a certain function on the surroundings. This is applicable to both manmade objects, like needles or gloves, and naturally selected creations, like thorns. This happens because shape is one of the most immediate ways to achieve a task, and it is used at the end of a designed process or naturally as a convenient way to gain an advantage under selective pressure.

Nature provides several examples of the adaptive characteristics connected to a particular shape, like wings or thorns. Even if bird and bat wings derive from different anatomical structures, they share the same shape because the trait is associated with the ability to fly. Identically for thorns: a rose prickle derives from shoots while a cactus spine was originally a leaf. Both evolved because the pointy shape acts as a deterrent for herbivores.

Inside the cell, the shape of proteins and other macromolecules is the most important attribute that allows them to exert the biological processes that form the machinery of life. Various organisms carry slightly different variations of the same proteins, but those protein variants will inevitably assume similar structures to express their intended function.

Everyday structural biologists use several techniques, like cryo-EM, crystallography or NMR, to reconstruct the structure of proteins. Studying the structure is what makes it possible to understand the functional details and the mechanism of action of the proteins.

I wondered if it were possible to do the reverse: to get the structure by analysing the variants that express the function.

Aim of the Thesis

The aim of this thesis' project is to prove that the variability obtained by in vitro evolution is sufficient to support the reconstruction of a protein structure or to identify some structural details of the protein.

I employed direct coupling analysis to infer the structural information and error prone PCR to generate the variability in vitro. TEM lactamase was chosen as model for the proof of principle since it is considered the gold standard reference for directed evolution studies and confer the host bacterium a very easy to screen antibiotic resistance. A limiting key aspect of the method is the collection of digital information from the biological library. I overcame this obstacle by using a third-generation sequencing platform that was able to produce reliable results that could be fed to the bioinformatic pipeline.

1.1 Evolutionary Couplings

Life is a complex phenomenon based on a precarious equilibrium of molecular transformations in a cellular environment mediated principally by proteins and their interactions. Proteins are encoded in a simple alphabet of only a few “letters”, but when combined in words, i.e. the peptide chain, they can express a variety of different functions that encompass signalling, transport and enzymatic catalysis. The function of a protein is an ability conferred by its shape, which in turn is dependent on the specific sequence and the sequence of the amino acids that form the peptide chain itself. It happens that sometimes the genetic information of an organism is corrupted by some mutations and that is reflected in small differences in the peptidic chains of some proteins. An alteration of the sequence of these proteins might affect their shape and in turn this will be reflected in an alteration of their function. If these proteins are important for the survival of the host organism and the modification generates a protein shape which damages the function of the protein, the host will likely die or have a serious disadvantage over its peers. For this reason, these types of mutations are unlikely to persist for more than a few generations and are probably lost during the course of evolution. This is why we can easily identify the same protein in different species, as the proteins are not allowed to variate indiscriminately but the function restricts the type of variants there might exist. Moreover proteins, depending on their functions are more or less able to variate, especially in their critical residues, and this information can be used in phylogenetic studies. Nevertheless, variation is very important, since it allows a greater chance to adapt to sudden changes of the environment. If we have a population with a single form of a given protein and this particular form is not able to function in the new environment, the population will go extinct. A population with more variants of the protein instead has a higher chance that one of these forms is able to function in the new environment and thus this allows part of the population to survive. Life is based on a regulated balance between mutations and adaptation that fuel a transforming driving force of cellular beings we call evolution.

Now let’s consider the case of a single deleterious mutation in a protein which happens at the same time of a second mutation in a spatially close part of the same protein. There is a chance that this second mutation can compensate the deleterious effect of the first. A charged residue can be compensated with an opposite charged one or a net loss of charge can be compensated with new charged residues appearing in the same area. Steric hindrance is the same, a new bulky residue can be accommodated if a spatially close one becomes smaller and vice versa. This co-occurring compensating mutations does not change nor the shape nor the function of the protein significantly and thus became part of novel variants of the protein that can persist during evolution. The key aspect of this phenomenon is that these joint mutation events leave an easy-to-spot trace in the evolution we can use to reconstruct which parts of the peptidic chain are close to each other. If a sufficient number of these are collected it is even possible to reconstruct the structure of the protein itself.

In a more technical explanation, when a set of variants of the same protein is retrieved it is possible to observe several positions that happen to mutate simultaneously during evolution. These positions are mutationally linked, since the single mutation might be harmful but the other one is able to compensate the former deleterious effect. They are thus “evolutionary coupled”, hence the name “evolutionary couplings”, and their co-occurrence analysed in a covariance

analysis is indicative of the spatial closeness in the structure of the protein of the amino acid positions that were altered by the mutations.

1.1.1 Mutual Information

The first attempts to rely on covariation to retrieve structural information belongs to the 80's and 90's (Altschuh et al., 1987; Göbel et al., 1994; Pazos et al., 1997) but the serious advancement in this field came with the advent of cheap genomic information in the sequencing era.
The earliest evolutionary couplings were calculated using Mutual Information (MI), a correlation estimator between variables that is not limited to numeric variables. Mutual Information as the name implies is a measure of the amount of information of a variable it is possible to capture from another.
Mutual information is a staple concept in information theory and I will only give a brief introduction of the topic. For further information, I encourage the reader to look for a book on the subject like "Elements of Information Theory" of Thomas M. Cover and Joy A. Thomas (Cover & Thomas, 1991).
While the concept of information is very abstract, for discrete variables information is commonly approximated with the entropy of the variable of interest. Mutual information can be considered an extension of this interpretation of the entropy: while entropy reports the information of a single variable, mutual information describes the amount of information one random variable contains about another. The entropy of a variable X is a function of the distribution p of the variable. In formula, the entropy H of the variable x ($x \in X$) is defined by:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

We can extend the concept to two or more variables. if we have two variables, X and Y, with a given distribution p, their joint entropy H is defined by:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

Since $0 \leq p \leq 1$, log(p) is always negative and thus H is always positive.
We can also define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning variable.

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \end{aligned}$$

The entropy of two variables, their jointed entropy and their conditional entropies are closely linked since the entropy of two variables is the entropy of one plus the conditional entropy of the

other:

$$H(X, Y) = H(X) + H(Y|X)\,.$$

which is very easy to demonstrate:

$$\begin{aligned}
H(X, Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\, p(x, y)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\, p(x)p(y|x)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\, p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\, p(y|x)\\
&= -\sum_{x\in\mathcal{X}} p(x)\log\, p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\, p(y|x)\\
&= H(X) + H(Y|X)\,.
\end{aligned}$$

Another important estimator is the relative entropy, which is a measure of the distance between two distributions. Mutual information is the distance of the joint probability distribution p(x,y) from the product distribution p(x)p(y). In formula:

$$I(X; Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}$$

where X and Y are two random variables with a joint probability function p(x, y) and marginal probability functions p(x) and p(y).
We can now observe an interesting relationship between mutual information and the entropies of the variables since:

$$\begin{aligned}
I(X; Y) &= \sum_{x,\,y} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}\\
&= \sum_{x,\,y} p(x, y)\log\frac{p(x|y)}{p(x)}\\
&= -\sum_{x,\,y} p(x, y)\log\, p(x) + \sum_{x,\,y} p(x, y)\log\, p(x|y)\\
&= -\sum_{x} p(x)\log\, p(x) - \left(-\sum_{x,\,y} p(x, y)\log\, p(x|y)\right)\\
&= H(X) - H(X|Y)\,.
\end{aligned}$$

Thus, mutual information is a measure of the information that one variable contains about another one. Numerically it equals to the reduction in the uncertainty of one random variable due to the knowledge of the other.
Mutual information has several characteristics, among them:

$$I(X;Y) = H(X) - H(X|Y),$$

$$I(X;Y) = H(Y) - H(Y|X),$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

$$I(X;Y) = I(Y;X),$$

$$I(X;X) = H(X).$$

The relationship between H(X), H(Y), H(X, Y), H(X I Y), H(Y I X) and I(X, Y) can be represented in a Venn diagram where the mutual information corresponds to the intersection of the information in X with the information in Y (**Figure 1.1.1**).

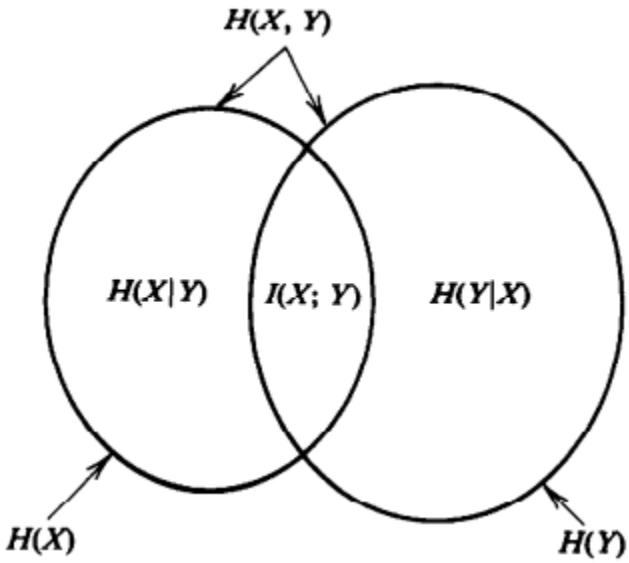


Figure 1.1.1 Venn diagram representing the relationship between the entropies.

In layman terms, we can comprehend these estimators using the weather as an example. We can have information about the weather in Rome and the weather in Milan. The weather in Rome give us also a little bit of information about the weather in Milan, because Rome and Milan are geographically close. This small amount of information is the mutual information. On the other hand, when the variables are more or less independent, like the weather in Rome and the weather in Beijing, the amount of information one can obtain from the other is close to zero. This analogy also holds for the residues of a protein. When the fold of the protein puts two residues close to each other, their entropies can capture this spatial closeness with a non-zero mutual information. On the other hand, residue spatially distant are less likely to have a

conspicuous mutual information. Reversing this paradigm, it is possible to infer the spatial closeness of two residues in a folded protein by looking at the mutual information of its residues.

1.1.2 Direct Coupling Analysis

The traditional correlation-based algorithms used for this analysis were effective but rather crude and suffered a lot of the confounding effects generated by indirect correlations.
Let's take, for example, a small network where amino acid A is linked to amino acid B and amino acid B is linked to amino acid C. In this network A and C indirectly covariate since they both respond to alterations of B, even if they might be far away in the protein structure. Another example might be a substitution in a faraway position of a protein that caused a conformational change in the protein that affect the interaction between other residues of the same protein. These residues are all correlated even if they lack a direct physical interaction.
The mutual information approach is defined *local*, since only a single residue pair is considered at a time. To overcome the problem, we require a *global* statistical model, where all the pairs are appraised simultaneously.
A sequence "*A*", aligned to the other members of the protein family, is defined as a vector $A = (A_1, A_2, A_3, \ldots, A_N)$ where A_i represents the amino acid of the sequence at position *i* of the alignment. The amino acids are encoded in an alphabet composed by the 20 standard amino

acids and a symbol for gapped regions. The simplest global model P for a maximum likelihood approach is a function of *A*, minimizing the number of parameters we have to determine:

$$P(A_1, \cdots, A_N) = \frac{1}{Z} \exp\left\{ -\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\}$$

Model parameters are the couplings $e_{ij}(A_i, A_j)$ between amino acid A_i in position i and amino acid A_j in position j, and local biases $h_i(A_i)$ describing the preference for amino acid A_i at position i.
Z is the partition function:

$$Z = \sum_{\{A_i\}} \prod_{i<j} \exp\{ - e_{ij}(A_i, A_j)\} \prod_i \exp\{h_i(A_i)\}$$

and is used for the normalization of P.
Obtaining the best fitting model for the coupling e_{ij} is a computational hard task and usually researchers use different approximation to simplify the task (Balakrishnan et al., 2011).
Once the e_{ij} and h_i parameters are retrieved, the direct information is calculated from the model.

$$DI_{ij} = \sum_{A_i, A_j = 1}^{q} P_{ij}^{Dir}(A_i, A_j) \ln\left(\frac{P_{ij}^{Dir}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right)$$

Direct information allows to disentangle the correlation network of the systems, favouring the identification of the strongest interactions to suppress the indirect correlations.
In an elegant way this was shown in a cardinal paper of Marks and colleagues (Marks et al., 2011) where both mutual and direct information were put in comparison in a self-explanatory graphical form (**Figure 1.1.2**).

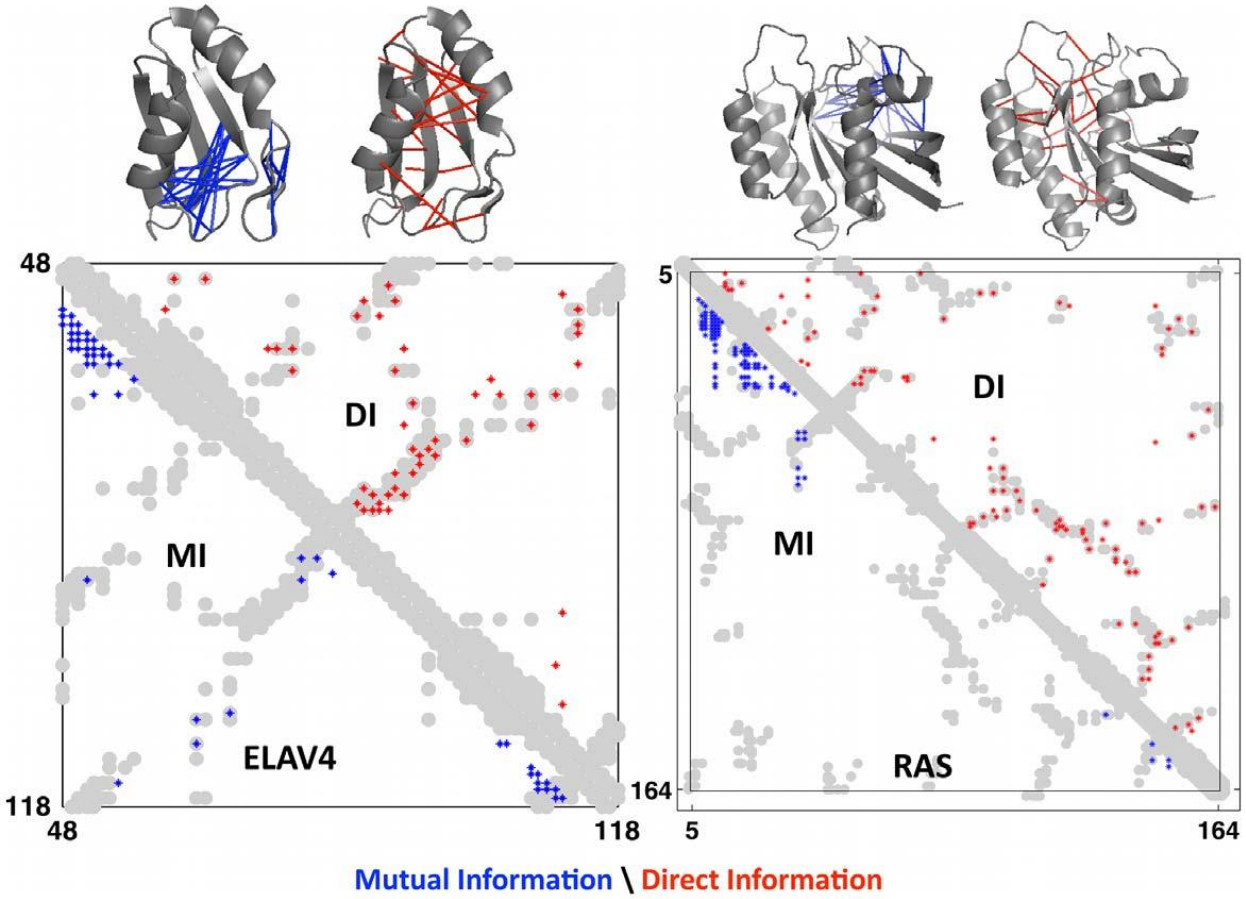


Figure 1.1.2 (from (Marks et al., 2011)**) Progress in contact prediction using the maximum entropy method.** Extraction of evolutionary information about residue coupling and predicted contacts from multiple sequence alignments works much better using the global statistical model (right, Direct Information, DI) than the local statistical model (left, Mutual Information, MI). Predicted contacts for DI (red lines connecting the residues predicted to be coupled from sequence information) are better positioned in the experimentally observed structure (grey ribbon diagram), than those for MI (left, blue lines), shown here for the RAS protein (right) and ELAV4 protein (left). The DI residue pairs are also more evenly distributed along the chain and overlap more accurately with the contacts in the observed structure (red stars [predicted], grey circles [observed] in contact map; centre, upper right triangle) than those using MI (blue [predicted], grey circles [observed]; centre, lower left triangle).

1.1.3 Fitness Landscape

Direct correlation analysis is not the only technique that is used to infer evolutionary couplings. In particular, in mutational studies emerged the strategy of analysing the protein mutational landscape by looking at the fitness of the variants in selective conditions.

In general, the fitness landscape of a protein describes how a set of mutations affect the function of a protein of interest (Hartman & Tullman-Ercek, 2019). The function of a protein is typically evaluated indirectly performing a phenotypic assay, usually by testing the growth of the

host organism under selective pressure but also, more recently, with some protein-dependent fluorometric/colorimetric assays. To quantify the advantage or disadvantage the mutations confer to a protein, a deep sequencing of a mutagenized library of variants of the target protein is performed, collecting samples of variants before and after the exposure to the selecting environment.
The selective pressure will increase the frequency of mutations that confer a functional advantage while deleterious mutations will decrease in frequency. The fitness is defined as the logarithm of the ratio of the frequencies before and after the selection of a given mutant, normalized to the ratio of the frequencies of the wild type variant in the same samples. When two or more mutations are independent, the fitness of the multiple-mutations variant is equal to the sum of the fitness of the single variants. In a paradigm where the fitness of a protein is determined by the overall function of the protein, like in a screening for the ability to grow in a selective media, it is possible to determine the evolutionary couplings by looking at the differences between the single mutant fitness and the combined mutant fitness. In general, by subtracting from the fitness of a double mutation the fitness of the single constituting mutations, one can infer if the interaction between these mutations is positively affecting the function of the protein or vice versa. A complete fitness landscape that analyses every possible variant of a protein is often used to study evolution, to identify proteins with new and useful properties, or to quantify mutability. The main difference between a fitness landscape and an evolutionary landscape is that the former tries to operate outside the paradigm of evolution, allowing the identification of characteristics of the protein that the constraints of a progressive evolution might have precluded.

1.2 Molecular Evolutions and Mutagenesis

Molecular evolution is the course of modifications of the sequence that constitute a biological molecule in succeeding generations.
In this manuscript however, I will use molecular evolution to indicate the *in vitro* molecular evolution a smaller umbrella term that includes a wide variety of techniques that perform *in vitro* mutagenesis of a cloned gene coupled with powerful screening methods to select functional variants of the protein it codes. Molecular evolution is especially useful to understand the role of functionally significant positions of a protein when the molecular mechanisms that regulate the function in not entirely understood. Molecular evolution is associated to a series of topics and analyses that relate to the impact of mutations in the population, to the development of new genes or new functions or to study evolutionary driving forces.
Historically the driving forces for molecular evolution mutagenesis were of physical nature, e.g. ultraviolet light or ionizing radiation induced DNA damage, or chemical nature, employing mutagens like mustard gas.
Nowadays, the most common methods to introduce random mutation in a sequence are chemical mutagenesis (Shortle & Botstein, 1983), alteration by mutation-inducing bacterial strains (Husimi, 1989; Toprak et al., 2012), incorporation of mutagenic nucleotide analogues (Zaccolo & Gherardi, 1999), DNA shuffling (Stemmer, 1994), incorporation of randomized synthetic oligonucleotides (Wells et al., 1985) and generation of mutations during the copy of the nucleic sequence by a low fidelity polymerase (Cadwell & Joyce, 1992).

1.2.1 The political events, war, and the chance proximity of key individual behind the discovery of mustard gas mutagenesis

Before delving into the serious part of the thesis, allow me to digress and tell the curious story that defined the history of mutagenesis, a tale that greatly inspired me during some difficult periods of my PhD internship. After all, it is impossible to discuss the origin of mutagenesis without narrating the extraordinary events and circumstances that revolved around the discovery of the first chemical mutagens: the mustard gas. This chapter will recount the story around the discovery of Charlotte Auerbach and John Michael Rabinovich, later known as Robson, as it is portrayed by the anecdotal report of the historical testimonies collected by Geoffrey Beale (Beale, 1993).
In 1940 the notion of gene as a unit of heredity was well established but the work of Griffith and Avery on the nucleic acid nature of the genetic information was still largely unknown.
One of the known properties of the gene was the ability to mutate, at the time defined as the transition from one an allelic form to another. Hermann Joseph Muller, future Nobel laureate for his work on the effects of radiation, demonstrated that mutation could occur both as a natural process or as a result of exposure to ultraviolet light or ionizing radiation such as X-ray. It was thought that X-ray induced mutation could be used to investigate the nature of the genes but soon it was evident that the type of mutations induced by radiation was not suited for such a

study and Muller itself hoped that a mutagenesis induced by chemicals could yield better information. For this reason, the discovery of chemical mutagenesis by Auerbach and Robson, in collaboration with Muller was treated as a sensational breakthrough in the history of genetics.
The key figure of this tale is Charlotte “Lotte” Auerbach, a scientist that was born in 1899 to a Jewish family in Krefeld, Germany. She was schooled in Berlin and attended lectures in the university of Berlin, Würzburg and Freiburg. At the age of 25 she started working under Otto August Mangold in Berlin-Dahlem. This situation was not congenial since Mangold, being a Nazi, was constantly antagonizing her for her Jewish origin. On one occasion discussing a change in her project, Mangold replied to a concern she raised saying: *“Sie sind meine Doktorantin. Sie müssen machen was ich sage. Was Sie denken hat nichts damit zu tun.”* that translate to *"You are my doctoral student. You have to do what I say. What you think has nothing to do with it."*. After Hitler was elected chancellor in 1933, Charlotte decided to leave the country to move to Britain with the help of an Anglo-German close family friend, chemistry professor Herbert Max Finlay Freundlich. She was shortly introduced to Francis Albert Eley Crew, head of the Institute of Animal Genetics at Edinburgh, who offered her a very modest position at his institute. Despite her financial situation was precarious, the intellectual atmosphere in the laboratory was very lively and with the help of her colleagues Lotte taught herself some genetics. The Institute was at the time one of the few places in Britain where genetic research was carried out and Crew himself collected a group of what Auerbach called “waif and strays” on minimal salaries from continental Europe (among them Peo Charles Koller, Guido Pontecorvo, Rowena Lamy, Bronislaw Marcel Slizynski, Helena Slizynska and Hermann Joseph Muller).
In 1939, with Crew’s help Charlotte was able to obtain the British nationality preventing the incarceration as an enemy alien following Britain’s declaration of war on Germany. At the time there was a paranoid spy fever in Britain, and it was expected that German parachutists would invade from the sky. On one occasion she was reported to the police because a mysterious tapping noise (her typewriter) was heard from her apartment, “from a room occupied by a lady with a strong German accent”. In 1940, when Italy entered the war, Guido Pontecorvo, a colleague and friend of hers, was unable to avoid the confinement as an enemy alien and was incarcerated in an internment camp on the Isle of Man.
In 1938, Crew peremptorily announced to her that she was to work with Muller. Muller suggested her to work with several known carcinogens to obtain chemical mutations in Drosophila, however they all failed to produce the expected results. Following these attempts, she stated to work on mustard gas as a collaboration with Robson and Alfred Joseph Clark. Experimentation on mustard gas originated from the observation by Clark of the similarities between radiation damage and the effect of mustard gas, especially in regard to eye injuries. Clark got the idea that these long-lasting effects might act on the genetic materials in cell nuclei. At the outbreak of the second world war it was expected that chemical warfare would be used in the coming battle, as it had been in the war of 1914-18, so much that soldiers were given gas masks in cardboard boxes to carry around wherever they went. Clark had a research contract with the government to research the effect of mustard gas, especially in eyes injuries.
Due to the strategic importance of the research for the war, the Muller team was forbidden to use the word “mustard gas” in any communication as it would have contravened the secrecy rules imposed by the government. In everyday communication, “substance H” was the term

used. Of course, since it was impossible to publish any results, there was very little interest on the subject.
The first experiments begun in November 1941 and were done on the roof of the pharmacology department in Edinburgh. Liquid mustard gas was heated in an open vessel and flies were exposed to the gas in a large chamber. All people involved quickly started to develop serious burn on their hand. Lotte was warned by her dermatology not to expose her hands to the gas again or she could develop serious injuries while Robson started wearing gloves. With the crude apparatus they employed it was impossible to control the amount of mustard gas to which the flies were exposed, and results were hard to replicate. After gas treatment the surviving flies were analysed by Muller ClB method, trying to identify X-linked mutations.
This tale would be incomplete without clarifying the struggle and the emotional distress that Lotte was facing during this period. In 1942 she wrote to Muller, her supervisor, who had moved to Amherst College in America:

> *"Dear Dr. Muller, I am afraid you will be disappointed how little further I have progressed with my work since I last wrote. It was not however my fault through laziness, but terrible difficulties with the dosage. From May to December I have done one experiment after the other without being able to reproduce the right dosage again . . . Robson insisted on a repetition of the original experiment for sex-linked lethals. As I expected, the result when it at last came off was completely confirmatory: 68 lethals and four semi-lethals in 790 tested chromosomes. In addition there were some visible mutations among the semi-lethals . . . .*
>
> *Yours very sincerely,*
> *Lotte Auerbach"*

However, Muller did not reply to Lotte's letter or perhaps his reply was sunk crossing the Atlantic Ocean so Lotte wrote again:

> *"Dear Dr. Muller, . . . I hope you won't think me ungrateful if I admit that in spite of my pleasure I was disappointed that there was no word from you or Thea [Muller's wife] . . . It is such a long time since we last heard from you, and I often wonder how life is treating you two just now. I hope, very kindly. In any case, here are my sincerest wishes that it will do so in 1943. I also was hoping for a word from you on my work. I am getting rather discouraged by the lack of interest I encounter everywhere. And the fact that you don't write about it makes me suspect that I have disappointed you very much by my various reports . . . All the same-hearty thanks once more, and the kindest regards for Thea and you.*
>
> *Yours,*
> *Lotte A."*

Luckily for everyone this tale does have a happy ending. In March 1942 the first experiment report was sent to the Ministry of Supply in London. In this first report Robson and Lotte

reported the generation of sex-linked lethal mutations generated by treatment with substance H (mustard gas), as well as breaks and rearrangements of chromosomes. In a later report they mention a difference in the susceptibility to mustard gas of different drosophila strains. The third report showed that the mustard gas acted directly on chromosomes rather than exerting an effect on cytoplasm. The fourth report was focused on the induction of visible mutations by mustard gas, the differences between X-ray damage and chemical damage and a draft of the possible genetic action of the gas. After the war, at the Eighth International Congress of Genetics in Stockholm in 1948, Lotte delivered a comprehensive review entitled “Chemical Induction of Mutations.” and was applauded by Muller, who was then the president, which gave Lotte enormous satisfaction. In 1948, she was awarded the Keith Prize of the Royal Society of Edinburgh. After receiving her D.Sc. degree in 1947, Lotte was appointed Lecturer in the Institute of Animal Genetics. From 1959 to 1969 she served as Honorary Director of the Unit of Mutagenesis Research which the Medical Research Council had established at Edinburgh. Finally, after her retirement in 1969, she was made a Professor Emerita.
At the end of his anecdotal report on Charlotte’s life and discovery Beale wrote his personal interpretation of the whole situation explaining to everyone the moral of this story:

> *“No doubt many factors contributed to the proposal (and successful prosecution) of the work. Among these we may mention: (1) the imminence of World War II, leading to the support for Clark’s work on mustard gas in Edinburgh; (2) the presence in Edinburgh at the same time of Clark, Robson, Lotte, and above all Muller, who met frequently for discussion of research; and (3) the existence of Crew’s Institute of Animal Genetics, at which there was a lively group creating an atmosphere favorable for scientific work. Lotte herself informed me that, speaking ironically, even Hitler could be held to some degree responsible, as he had forced her to leave Germany and abandon school teaching for scientific research. But there seems no doubt that it was the expertise in Drosophila research methods and perseverance of Lotte along with the intrepid, perhaps foolhardy handling of the chemical that were mainly responsible for the success of the work”*

1.2.2 Error prone polymerase chain reaction

Error prone PCR is a highly versatile mutagenesis technique that rely on the inaccurate copying by a DNA polymerase in a PCR reaction to produce a collection of mutated sequences. Error prone PCR is one of the oldest and most used strategies in mutagenesis thanks to the simplicity of the technique and the fact that most mutagenesis experiments are aimed to identify a small number of mutations that increase the stability or activity of the target protein. The error propensity that gives its name to the technique is generated and enhanced by a series of factors that alter the fidelity of the polymerase during *in vitro* DNA synthesis. In its original formulation (Cadwell & Joyce, 1992; Leung, 1989) the fidelity of the PCR was reduced by increasing the concentration of $MgCl_2$, adding $MnCl_2$, using an unbalanced
ratio of nucleotides (an increased 1 mM concentration of dGTP, dCTP, and dTTP together with standard 0.2 mM concentration of dATP), increasing the concentration of the polymerase, and increasing the extension time. Under these conditions the reported rate of mutagenesis per position could go as high as 2%. Cadwell and Joyce (Cadwell & Joyce, 1992) however argued

that in the setup proposed by Leung et al., the mutated sequences obtained after mutagenesis display a strong A→G and T→C transition bias that steered the DNA toward a high GC content. The nucleotide concentration unbalance Leung used in his work was drawn from a previous study which demonstrated that in their very simple model the mutation rate could only be increased by lowering the concentration of dATP relative to that of the other three dNTPs. Lowering the concentration of one of the other three nucleotides produced little effect on the mutation rate (Sinha & Haimes, 1981). The lack of deoxyadenosine as a substrate promoted several non-Watson-Crick base mispairs, in particular G-T mispairs, and this condition leads to an excess of A→G and T→C transitions on the final product. To correct this issue, Cadwell and Joyce tested several imbalanced ratios of deoxynucleotide and were able to identify a condition in which the frequencies of AT→GC and GC→AT mutations became identical, albeit producing a lower final yield of 0.6% mutations per nucleotide position. It is important to notice that even if this new protocol produced a balanced ratio of $\frac{AT \rightarrow GC}{GC \rightarrow AT}$mutations, there is still a significant preference for T→X changes (X ≠ T) and X→A changes (X ≠ A) and the observed 0.75:1 transitions to transversions ratio is different to the unbiased ratio of 0.5:1.
Several different types of polymerase have been tested to maximize the probability of a wrong base incorporation, among them the Klenow fragment (of E.coli polymerase I) (Keohavong & Thilly, 1989), T4 DNA polymerase (Sinha & Haimes, 1981), Sequenase (a modified T7 DNA polymerase) (Ling et al., 1991), Taq DNA polymerase (J. Chen et al., 1991; Eckert & Kunkel, 1990, 1991; Ennis et al., 1990; Tindall & Kunkel, 1988) and Vent (Thermococcus litoralis) DNA polymerase (Mattila et al., 1991).
In this group, the lowest fidelity is shown by the Taq polymerase, the one used by Cadwell and Joyce, with a “natural” mean error rate of 0.001-0.02% (Eckert & Kunkel, 1990, 1991) per nucleotide each time the polymerase travelled over the template DNA.

1.2.3 Mutagenic nucleotide analogs for random mutagenesis

An alternative strategy for PCR-based mutagenesis is employing mutagenic nucleotide analogs that promote nucleotide base mispairing during amplification. This strategy was pioneered by Zaccolo and Gherardi (Zaccolo et al., 1996) with the usage of triphosphates of mutagenic nucleosides as substrates for a PCR-based mutagenesis.
They identified four conditions that defined a good mutagenic nucleotide analog: the substrate should be stable under the conditions of PCR cycling, well incorporated, be of high mutagenic efficiency and be able to direct both transition and transversion point mutagenic changes when used in combination.
At the time, there were several base analogue mutagens whose 5' nucleoside triphosphates had been characterized (N^4-hydroxy-2'-deoxycytidine (Müller et al., 1978), 2-aminopurine-2'-deoxyriboside (Grossberger & Clough, 1981), 5-bromo-2'-deoxyuridine (Mott et al., 1984), O^6-methyl-2'-deoxyguanosine (Eadie et al., 1984; Snow et al., 1984), N^4-amino-2'-deoxycytidine (Negishi et al., 1985), N^4-methoxy-2'-deoxycytidine (Reeves & Beattie, 1985; Singer et al.,

1984), N^6-hydroxy-2'-deoxyadenosine (Abdul-Masih & Bessman, 1986), 5-hydroxy-2'-deoxycytidine and -uridine (Purmal et al., 1994)).
N^4-hydroxy-2'-deoxycytidine and N^4-methoxy-2'-deoxycytidine both display tautomeric constants between 10 and 30 (Brown et al., 1968; Morozov et al., 1982) (a tautomeric constant around 1 is required to induce transition mutations effectively) and are able to form a pair with both A and G, however the base pairing via the imino tautomers is sterically hindered due to the *syn* preference of the N^4-substituents (Morozov et al., 1982; Shugar et al., 1976).
The nucleoside dP (6-(2-deoxy-β-d-ribofuranosyl)-3,4-dihydro-6H,8H-pyrimido[4,5-c][1,2]oxazin-2-one) is an analogue of N^4-methoxy-2'-deoxycytidine but, due to its bicyclic ring structure, is restricted to the *anti* conformation and it is able to form stable base pairs with both A and G.
In 1996 Zaccolo and Gherardi (Zaccolo et al., 1996) demonstrated that the nucleoside triphosphate of dP (dPTP) would be a good substrate for the polymerases in PCR reactions. However, dPTP is only able to promote nucleotide transition mutations, and by itself is unable to explore the whole panorama of possible mutations since it cannot mutate a purine into a pyrimidine or vice versa.
There are very few nucleoside triphosphate analogues capable of transversion mutations (Pavlov et al., 1994; Purmal et al., 1994; Singer et al., 1986, 1989). Among them only 8-oxo-2'-deoxyguanosine triphosphate (8-oxodGTP), one of the major products of DNA oxidation generated during cellular oxidative stress, appears to be a substrate for DNA polymerases adequate for mutagenesis (Pavlov et al., 1994; Purmal et al., 1994).
8-oxodGTP assumes the *anti* conformation when paired with cytosine (Oda et al., 1991), while adopts the *syn* conformation with adenine (Kouchakdjian et al., 1991) therefore allowing A→C transversion mutations during DNA duplication (Cheng et al., 1992; Maki & Sekiguchi, 1992; Shibutani et al., 1991).
The mutations generated by these analogues are heavily biased towards a single mutation type. Of the transition mutations elicited by the presence of dPTP in the reaction, the vast majority are A→G (46.6%) and T→C (35.5%), while G→A (9.2%) and C→T (8%) are rather rare. A→G and T→C transitions are the result of the incorporation of dP instead of a T in either strand during DNA duplication followed by a mispairing of the incorporated dP with a G in the subsequent cycle. Incorporation of dP when paired to a G in either strand occur at a much lower frequency and thus produce the mutational bias.
8-oxodGTP is more biased than dP and virtually generates only a single type of transversion mutations: A→C (38.8%) and the complementary T→G (59%). These transversions are the result of the incorporation of 8-oxodGTP instead of a T in either strand during DNA duplication. The frequency of the reverse C→A transversions is 40 times lower than that of the A→C transversions (Cheng et al., 1992).
Random mutagenesis of DNA using nucleotide analogues relies on a PCR reaction carried out with both dPTP and 8-oxodGTP. The two analogues are efficiently incorporated into DNA in vitro by Taq polymerase and this allows for point mutations to be introduced with high efficiency

into the target sequence by PCR. The mutational load of the final product can be regulated by limiting or increasing the PCR cycles performed in mutagenic condition.
Clones mutagenized with the combination of dPTP and 8-oxodGTP in standard mutagenic conditions can reach a mutation frequency of 1×10^{-3} per nucleotide.

1.2.4 In vivo continuous evolution

Continuous Evolution is a different approach to molecular evolution, limiting the handling steps of the operators to a bare minimum and aimed to generate to a self-sufficient and fast paced evolution system.
Badran (Badran & Liu, 2015) simplifies the requirement for Darwinian evolution to four fundamental processes: translation (that is not simply the biological concept of the RNA producing process but more in general when the evolving molecule is not identical to the information carrier), selection, replication, and mutation. Traditional molecular evolution methods handle each of these processes separately while continuous evolution systems integrate all of these into an uninterrupted cycle.
The first formulation of the technology dates back to the sixties with the hallmark publication of Mills, Peterson and Spiegelman (Mills et al., 1967) where they proved the possibility to evolve a minimal Qβ bacteriophage genomic RNA in an enriched *in vitro* system. The authors raised the interesting question:

> *"What will happen to the RNA molecules if the only demand made on them is the Biblical injunction, multiply, with the biological proviso that they do so as rapidly as possible?"*

The authors provided an environment with plenty of nutrients and the replication apparatus and liberated the genome from the cellular confinement and the requirement of producing an infection machinery. The experiment was carried out as a set of serial reactions and as the experiment progressed, the genome multiplication rate increased and the product became smaller. By the 74th transfer, 83% of the original genome had been eliminated and the new genome was able to replicate up to 15 times faster. This new genome however could no longer direct the synthesis of new viral particles since this requirement was not implicit in the selection and many important genes were lost.
Additional studies, altering the tube environment and limiting different resources, extended this method to generate a continuous evolution of an RNA-mediated catalytic function (Wright, 1997), new promoters (Breaker et al., 1994), or other RNA related elements or function.
These early methods were exclusively carried out *in vitro* where the parameters involved were easy to control but on the other hand they gave up the ability to study the untapped complexity of a cellular system and set aside the biological results of an *in vivo* selection method could produce.
The earliest alternatives to *in vitro* methods came with the introduction of chemostats and auxostats, culture vessel automations where the bacterial growth is maintained through the continuous regulated inflow of fresh growth medium. These platforms were particularly suited to support a directed bacteriophage evolution by maintaining a fresh pool of bacteriophage infectable microbes where a fast evolution could develop. Bacteriophages have intrinsically high

mutation rates and are simple to collect and analyse, making them adequate model organisms for the initial *in vivo* directed evolution of selectable traits (Husimi, 1989).
One of the first studies that pioneered this method was the viral continuous evolution of the bacteriophage ΦX174 in *E. coli* and *S. typhimurium* to adapt to a high temperature environment (Bull et al., 1997). In this study an average of 1–2 mutations per 24 hours (corresponding to 0.2–0.5% of the viral genome, and 0.003–0.005 mutations per generation per kilobase) was generated in the absence of any added mutagens. In an expansion of this work, ΦX174 phage was propagated for six months, ~13,000 phage generations (Wichman et al., 2005), and proved the very long-term potential of application of the technology.
A novel adaptation of the bacteriophage mediated evolution is phage-assisted continuous evolution (PACE) (Esvelt et al., 2011). PACE exploits the coat protein pIII from the bacteriophage M13 to mediate infection and phage propagation. The bacterial host carry this protein under a transcriptional circuit regulated by the selection process. When the selective pressure produces a functional variant of the molecule of interest, pIII is produced and the virulence of the bacteriophage is restored, allowing a robust phage propagation in continuous culture. Meanwhile phages lacking functional pIII have no ability to propagate and are rapidly lost under continuous culturing conditions. The continuous flow of this system coupled with the steady infection of fresh cells allows mutations to accumulate only within the phage genome while the genome of the bacterial population is preserved.
The increasing popularity of bacteriophage mediated evolution promoted the development of several bacterial culture vessel for a monitored continual prokaryotic growth and in turn facilitated the study of continuous directed evolution of bacterial populations. Bacterial evolution compared to the phage assisted counterpart carry two main advantages, namely an even simpler system and a broader range of gene size (while viral fitness impedes the transfer of big genes).
Toprak (Toprak et al., 2012) used bacterial evolution to follow parallel populations of E. coli under differing antibiotic selection pressures using a morbidostat. This apparatus is a bacterial culture device in which the concentration of antibiotic is periodically adjusted to maintain a nearly constant selection pressure, *i.e.* it increases the antibiotic concentration to preserve the bacterial growth speed. The populations were subjected to an evolutionary pressure under the driving force of different antibiotics, and followed using whole-genome sequencing to collect the mutations that provoked antibiotic resistance.
In both bacteria and phagic evolution, some applications may require an extensive mutagenesis and the basal rate of bacterial mutagenesis might not be enough to provide access to necessary genetic changes on a practical timescale.
A popular method used to enhance the bacterial mutagenic propensity was introducing an orthogonal, low fidelity polymerase in the bacterial genome. A famous implementation of this idea uses a modified E. coli DNA polymerase I (Pol I) where the error rate is increased to

promote the mutagenesis of the target gene (Camps et al., 2003) while limiting the damage on the core genome of the cell.

1.2.5 Deep mutational scanning

If targeted mutagenesis examines a limited number of protein variants with a specific question in mind and other types of molecular evolution generate a limited pool of variants, deep mutational scanning (Fowler & Fields, 2014) is aimed to obtain an exhaustive collection of all possible mutants of a protein and verify their functional viability. As reference for the scale of this process, the average human protein is ~350 amino acids in length and it can yield 7,000 single mutations and over 22 million double mutations (Brocchieri & Karlin, 2005). The idea under the technology is rather old, after all alanine scan (Cunningham & Wells, 1989), in which each position of a protein is sequentially mutated to alanine, can be considered a limited ancestral version of this technology. A complete systematic mutagenesis instead was developed only relatively recently because it had to overcome two big challenges that could only be resolved with modern day technology. The first challenge was the generation of all of these variants in an efficient process, while the second was to be test and sequence the variant in parallel without the need to handle the individual mutant. In particular, the big leap forward in this context came with the development of mass parallel sequencing and high-throughput technologies.
Traditionally the most used mutagenesis method was random mutagenesis. It generated a very big library of 10^5-10^{11} protein variants and then this library was screened for the selected function bringing the diversity down to a small size so that it could be Sanger sequenced. It is evident that the limitation was not the original diversity but the requirement of a small pool of sequence for the sequencing. In the first formulation mutational scanning was just a technique that combined selection with high-throughput DNA sequencing. Because the sequenced variants were now several hundred times the number of variants obtained from traditional sequencing, a harsh selection pressure is not required anymore to collapse the size of the mutagenic library. A mild selection does not cause a drastic high pass filter on the function of the survivor protein and allows a bigger spectrum of variation. Moreover, instead of observing only a survival of the fittest, it was now possible to observe the variation in the mutant frequencies, and this variation is linked to the function of each variant. All variants harbouring beneficial mutations would increase in frequency, variants harbouring nearly-neutral substitutions would display little to no variation, whereas variants harbouring deleterious mutations would decrease in frequency.
To construct these libraries, several methods can be used (Fowler & Fields, 2014), but in practice the methods that became indissolubly tied with the deep mutational scans are oligonucleotide based directed mutagenesis methods like PFunkel (Firnberg & Ostermeier, 2012). Oligonucleotide directed mutagenesis allows a very efficient and fast construction of large libraries of singly mutated variants but has difficulties when used to construct libraries of multiply mutated variants. The double mutant space around the average-sized human protein can in theory be completely explored by systematic mutagenesis but since the number of possible sequences increases geometrically with the protein length, and the next generation

sequencing albeit able to cover a vast collection is still limited, the approach rapidly loses the power to be exhaustive.
Another advantage of oligonucleotide-based mutagenesis is being able to selectively restrict the number of positions to be mutagenized, *e.g.* mutagenizing only few residues in a substrate binding site or in the catalytic pocket. This strategy enables to reduce the overall diversity of the library and increase the read count and the accuracy of the critical residues for the study.

1.2.6 An artificial selection

Molecular evolution is a two-step pipeline where the first step involves generating the diversity, that is the phase when the variants are built, and then the second step is the selection process. If mutations are the driving force of the evolution, the selection poses the constraints that guarantee the protein function. While the mutations are generating random variability, the laboratory selection oversees the new variants authorizing only the conformations that follow the natural blueprint we are exploiting. The selection is always carried out as an evaluation of the target protein function or characteristic, and, as a consequence, it cannot be a generic task like mutagenesis but has to be designed on a case by case basis for each intended target. The most common selection processes are aimed to verify the preservation of the activity of an enzyme or the ability to bind an interaction partner. For instance, we can verify the activity preservation of a metabolic enzyme by analysing the conversion speed of the metabolites the enzyme catalyses or it is possible to verify the interaction ability in binding assay like in phage display. Given the scale of molecular evolution experiments, cumbersome assays are difficult to implement and most of the times the preferred assays are the ones that permit a rapid and parallel screening for the intended function. Common assays are survival assay to test the activity of essential genes, binding assays to test the interaction capability of the target protein and assays that exploit the ability of the protein to convert some chromogenic substrate.
Survival assays are the most straightforward type of assays, given that only the cells that carry a functional copy are allowed to survive. To have a rough approximation of the number of genes that can undergo this type of selection we can use the Venter list of essential genes for a minimal genome. In its final formulation, the minimal Genome of Syn3.0 had 473 genes (438 protein and 35 RNA-coding genes) classified as essential, ~31% of which were not assigned to any specific biological function. Ironically, even if the selection of these is based on their molecular function, it is not strictly necessary to understand the function of these genes, since the complex ability to survive is the actual phenotypic trait that is sieved for functionality. Gene essentiality is tightly linked to the cell type and to the environment. If the growth medium is devoid of tryptophan, the genes involved in the biosynthesis of tryptophan are essential, but for cells grown in a medium where the amino acid is available, these genes are not critical anymore. Some other genes, notably antibiotic resistance genes, do the opposite: in a normal environment do not confer any selective advantage but become essential when the bacterial cell is exposed to the antibiotic. The ampicillin resistance gene TEM-1 beta lactamase, which catalyses the hydrolysation of the β-lactam ring of many antibiotic penicillin derivatives and was the protein target of the mutagenesis done in this thesis, is part of this class of genes.
The other very common selection strategy in mutagenesis experiments is based on interaction/binding. This type of selection applies to proteins that directly interact to some other

molecules in solution. Notably, this group includes proteins involved in heteromeric complexes, protein involved in some functional transient interaction and receptor-ligand pairs. In order to select the protein variants where the interaction is still present, the protein is exposed to its interaction partner, often conjugated to some purification mechanism, and then purified by coprecipitation or retention exploiting the association to the interacting protein. Phage display and more in general most types of display technologies are the most popular selection platforms for functional interaction assays. They work by directly linking the genotype to the phenotype and are able to select for the functional interaction of the protein and at the same time collect the gene of these functional variants.
Selection can be either carried out in vitro (display technologies) or in vivo. Most of the organisms used for in vivo screening procedures are characterized by a fast replication time and an easy transformation strategy. Common organisms used for selection are fast growing bacteria, like Escherichia coli, or yeast, such as Saccharomyces cerevisiae. The main advantages of a bacterial selection system are the growth time and the transformation efficiency, while the yeast models have a eukaryotic genetic background that is preferred when the aim is to study a eukaryotic protein and can make use of the eukaryotic cell compartmentalization.

1.3 Sequencing

Molecular evolution creates the conditions to compel a given molecule to undergo an adaptive mutagenesis in order to survive the selective environment. If the desired output for the evolution is to obtain novel organisms that survive the selective pressure, then all the surviving cells achieve this purpose since each of them carry new functional variants of the molecule.
In other applications however the cell itself is not the desired output of the evolution but rather the desired information is the composition of the variant molecules that allowed the hosts to survive.
To obtain this information, sequencing the variant molecule is essential. Traditional experiments on synthetic evolution are focused on observing the mutational landscape of the protein of interest (K. Chen & Arnold, 1993; Esvelt et al., 2011) or study the trajectories of evolution. This was achieved by analysing the emergence of the patterns of mutations that appear progressively during the course of an evolution experiment in one or more samples subjected to the same selective pressure (Hillis & Huelsenbeck, 1992; Zaccolo & Gherardi, 1999).
Sanger sequencing is used when relatively few variants are screened (Jacquier et al., 2013) or when the newest sequencing technologies were not yet available (Zaccolo & Gherardi, 1999). In recent applications next generation platforms are employed to increase the throughput of the output so that several thousand or million variants are recovered in a single sample analysis (Olson et al., 2014). Second generation sequencing platform are however heavily constrained on the length of the molecule and cannot be used on any protein of interest.
Our application requires the sequencing of library made of hundred thousand to a few million different elements in size where each of the reads must cover the whole molecule variant, thus neither Sanger sequencing or traditional next generation platforms are sufficient.
We pioneered the usage of single molecule third generation sequencing on library of gene variants evolved by molecular evolution to remove the size restriction to our data collection.
The following sections will provide an introduction to both traditional and recent sequencing technologies to give the reader a complete panorama of the possible sequencing platforms that could be used to collect the data, discussing their advantages and drawbacks.

1.3.1 Traditional sequencing before high-throughput parallelization

The first attempt to create a widespread, easy and general sequencing technique began with the experimentation of Sanger and Coulson in 1974 (Sanger et al., 1974; Sanger & Coulson, 1975). The protocol involved partial DNA synthesis and a separation by electrophoresis through polyacrylamide gels to discriminate the polynucleotides lengths. The first step of the protocol is an initial asynchronous synthesis of the target DNA by polymerase I which extend the primer oligonucleotide and copy the template. This step generates a great number of oligonucleotides of different lengths, all starting from the primer. Afterwards, two second polymerisation reactions are performed: a ‘plus’ reaction (from (Englund, 1971, 1972)), in which only a single type of nucleotide is present, thus all extensions will end with that base, and a ‘minus’ reaction (from (Wu & Kaiser, 1968)), in which the other three nucleotides are used, which elongate the sequences up to the position before the next missing nucleotide. When all 4 ‘plus’ and 4 ‘minus’ reactions are completed and run on acrylamide gels, one can compare the patterns and deduce the sequence. For each position there should be a ‘plus’ band, corresponding to the nucleotide

in the position and a ‘minus’ band, corresponding to the next nucleotide in the chain. Deviations from this rule, showing differences in the ‘plus’ and ‘minus’ systems, allow the discrimination of adjacent identical nucleotides in the chain (**Figure 1.3.1**).
Also, in the mid-seventies Maxam and Gilbert (Maxam & Gilbert, 1977) developed a sequencing system based on chemical degradation at specific base types of the target sequence. This sequencing requires first to radiolabel the 5’ or 3’ end of the DNA fragment, then a chemical treatment partial digests the target sequence in the presence of specific nucleotide bases. There are several reactions that can target specific nucleotide or combinations, but commonly only four are used: purines (A+G) are depurinated using formic acid, guanines are methylated by dimethyl sulfate, cytosines are hydrolysed by hydrazine and sodium chloride, the pyrimidines (C+T) are hydrolysed using hydrazine without salt. The modified base promotes a site-specific cleavage when piperidine and heat are provided to the solution. The labelled fragments generated by the strand breaks are then migrated in an acrylamide gel and the sequence deduced by comparing the patterns of the four reactions (**Figure 1.3.1**).

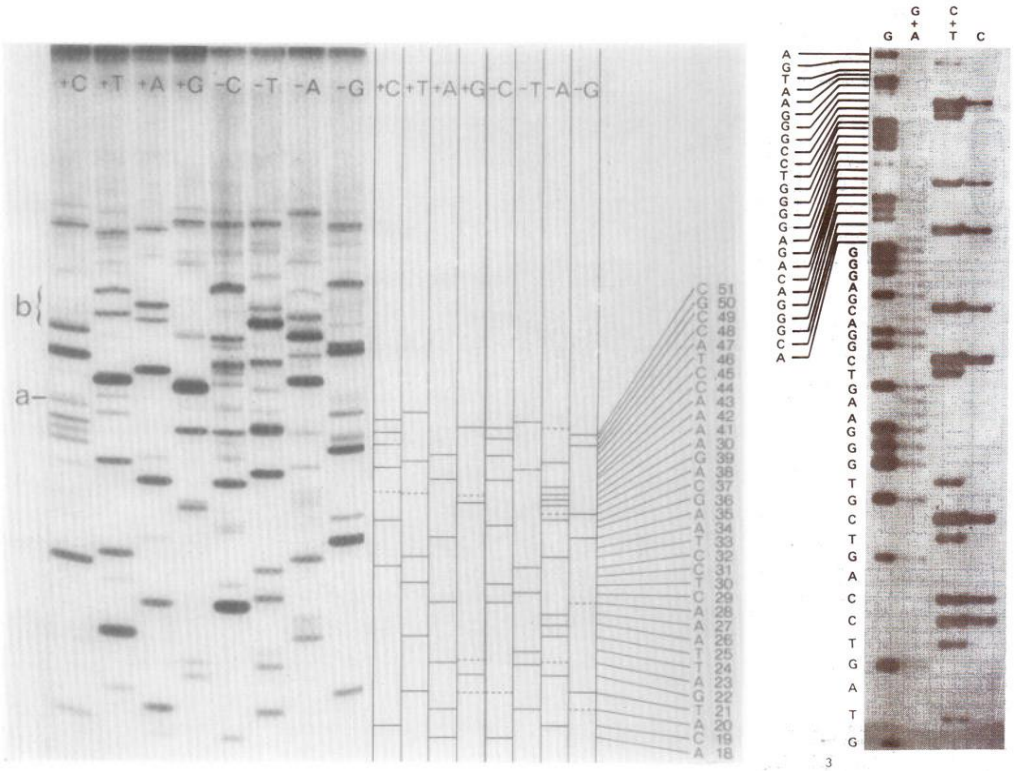


Figure 1.3.1 Examples of the earliest traditional sequencing and sequence assignment.
Left: Sanger and Coulson sequencing. Right: Maxam and Gilbert sequencing.

The major breakthrough in sequencing technology, so much that the technique is still used nowadays, came in 1977 with Sanger's chain-termination sequencing technique (Sanger et al., 1977).
The strategy exploits the incorporation in the nucleotide chain of the 2'3'-dideoxyribonucleotides (ddNTPs) version of the four nucleotides which prevents any further polymerization of the nascent chain in a DNA synthesis reaction.
A mixture of a radio labelled ddNTP and standard deoxynucleotides in a DNA extension reaction produces a collection of DNA strands of different lengths, all ending with the dideoxynucleotide which prevents a further extension of the chain.
Four different reactions, each involving a different dideoxynucleotide, generate a specific pattern when migrated in an acrylamide gel that can be used to infer the composition and sequence of the target nucleotide chain.

1.3.2 The massive parallel sequencing revolution

The first platform that went beyond traditional sequencing was released by 454 Life Sciences and allowed the first massively parallel sequencing of a human diploid genome (of professor JD Watson) in less than 2 months (Wheeler et al., 2008). In comparison the first draft of the human genome required around 10 years (1990-2000) before it was available to the public.
454 Life Sciences utilised a sequencing technology, pyrosequencing, pioneered by Pål Nyrén and colleagues (Nyrén & Lundin, 1985) and subsequently developed by Edward Hyman (Hyman, 1988).
Pyrosequencing, like Sanger's method, is another sequencing by synthesis (SBS), as it requires the synthesis of a new DNA strand to recover the sequence of the target DNA stretch. However, in pyrosequencing, the identification occurs in real time by measuring the pyrophosphate released during the polymerization process instead of using polyacrylamide gel migration and labelled nucleotides to reveal the incorporation of a specific base. When a nucleotide triphosphate is used to extend the DNA chain a pyrophosphate is released in the solution. This pyrophosphate can be used as a substrate in a series of reactions catalysed by the enzymes ATP sulfurylase and luciferase to produce inorganic phosphate and light. Cycling the addition of a single dNTP type to the solution (and subsequent washing) one can use the light produced by the luciferase to infer if this nucleotide was used to extend the nascent DNA chain. Nyrén expanded the technology in 1998 (Ronaghi, 1998) allowing the sequencing reaction to occur in droplets "printed" on a surface with lower background noise. 454 Life Science was soon acquired by Roche and the first commercial next generation sequencing (NGS) platform became available.
After the release of Roche 454 several new sequencing technologies were developed riding on the wave of the newborn genomic revolution, the most important of which is the Solexa-Illumina method. This method is also an SBS technique but the innovation resides in the methodology used to bind the DNA molecule to the solid substrate (Fedurco, 2006) and the 'reversible-terminator' dNTPs used for sequencing (Bentley et al., 2008).
The target DNA library had to be ligated to specific adapters that were designed to complement short oligonucleotides chemically bound to a surface. When this library was deposited on the sequencing surface (called sequencing cell or flowcell) at proper dilution it would sparsely bind

randomly to the lawn of oligos. These molecules were allowed to generate small separate clusters of identical sequences through a solid-phase amplification of DNA on the glass. This process was later called ‘bridge amplification’ for the characteristic arching of the bound DNA during the local amplification. Sequencing is achieved by using fluorescent ‘reversible-terminator’ dNTPs, which cannot bind further nucleotides as a fluorophore occupies the 3′ position of the deoxyribose. After the incorporated nucleotide-fluorophore pair in each cluster is excited by a laser and read by a CCD camera, the fluorophore is cleaved away and the polymerisation can continue (Turcatti et al., 2008). The concentration of identical sequences in clusters allows the reaction to proceed in synchronous and concerted small polymerization steps identical for every sequence of the cluster. This way, when the laser excites the fluorophores, each element of the cluster will emit a strong synchronous signal that can be collected by the camera.
The first machines were initially only capable of producing very short reads (up to 35 bp long) but the latest machines can reach up to 600 nucleotides per cluster. The main problem connected to Illumina sequencing is that the accuracy of the identification of the base (basecall accuracy) decreases with increasing read length (Dohm et al., 2008) for a phenomenon called dephasing. As the reaction proceeds, rare events such as misincorporation, over- or underincorporation of nucleotides, or problems in the removal of the blocking fluorophore can occur. After a while all these aberrations accumulate and affect the synchronicity of the polymerization of the sequences in the cluster reducing the purity of the signal emitted (Voelkerding et al., 2009).
Several other methods of next generation sequencing were developed and commercialized, noteworthy are SOLiD ligation-based sequencing or Ion Torrent sequencing, but Illumina was still the most widespread technique of the period.

1.3.3 The next “next generation”: 3rd generation NGS

Next generation technologies were a leap forward in sequencing methods that started the genomic era but are characterized by a series of constraints and limitations.
Time is a problem in these types of sequencing because these SBS methods always involve a halted polymerization process, and a typical run requires a couple of days.
The dephasing (or loss of synchronicity) is another typical limitation. When an ensemble of molecules of a cluster originated from a single molecule is extended in single nucleobase additions, the nascent DNA chain gradually diverge in length when the nucleobase addition does not occur with a perfect yield. Each dephased molecule generates a signal noise that decreases the quality of the basecall during sequencing and errors begin to appear in the read. This is the reason that makes next generation reads shorter and less reliable than the results of a traditional Sanger sequencing.
Another problem is the necessity to produce a strong signal for detection. Second generation technologies overcome the issue by relying on confined PCR amplification. The amplification process however, besides extending the time required, is not perfect and possess its own bias and issues. The most important concern is the fidelity of the DNA copy that insert a second

layer of processing between the sequence given as input and the sequence read by the machine.
The last constraint is the scale of the result. Sanger's method generates thousand nucleotide long sequences and therefore is very easy to connect and overlap them to other sequences in order to define a longer DNA sequence. These second-generation sequencing techniques instead generate much smaller fragments and require a huge amount of data to describe the same area.
There is not a unique definition to what constitutes the next 'next generation' of sequencing methods. Schadt (Schadt et al., 2010) tries to identify some commonalities of the new methods that counterpose some of the limitations of second generation sequencing technologies mentioned above: (i) higher throughput; (ii) faster turnaround time; (iii) longer read lengths; (iv) higher accuracy; (v) small amounts of starting material (theoretically only a single molecule); (vi) low cost.
The most relevant methods that define these new generation of sequencing technologies are two. The first is an SBS technology in which single molecules of DNA polymerase are observed as they synthesize a single molecule of DNA and the second is a nanopore-sequencing technology in which single molecules of DNA are threaded through a nanopore, and individual bases are detected as they pass through the hole (Weirather et al., 2017).

1.3.4 Nanopore Technologies

A nanopore is a very small hole. This is an iconic sentence that was advertised together with the first nanopore-based sequencer. A nanopore is indeed just a hole in the nanometre scale. Nanopores can be of biological origin, like transmembrane channel protein, or can be manmade pore, like pores made of graphene. The nanopore sequencing technology is a sequencing strategy that allows the detection and identification of the base composition and sequence of a DNA chain that is passing through a nanopore. A silica surface separates two conductive solution chambers, in one of which the DNA that is going to be sequenced is suspended. In this silica plate are embedded conductive nanopores that allow the flow of ions and small molecules. When an electric current is provided to these systems, charged molecules, in particular the DNAs, will begin to flow through the only opening between the chambers, the nanopores. This process can be facilitated or hindered by the addition of other proteins to the solution. The typical setup involves helicases or polymerases that bind to both the opening of the nanopore and to the DNA and regulate the speed of the nucleic chain translocation through the pore. During this process each conductive pore is connected to a detector that is able to quantify in real time the current that is passing through the hole. One of the main elements that influences the current is the shape of the pores, so when the DNA molecule passes through it the current is partially blocked. Different nucleobases have different shapes and properties that will influence the current that flows in the pore depending on which bases are passing in that moment. By decoding the characteristic alterations in the current (squiggles) it is possible to identify the base that disrupted the ionic current and, as the chain gets translocated, retrieve the

sequence. Of the many technologies described, this is the only one that allows a direct probing of the bases that form the DNA chain.
The main developer of this technology is Oxford Nanopore Technologies (ONT), a UK based company that is undertaking active research and development on new pores and other enhancement.
Their most successful sequencing device is the MinION, a sub 1000$ sequencing platform, which also is the smallest sequencing device currently on the market, as it is similar in size to an old USB stick (10 × 3 × 2cm) and weighs less than 100 g. With MinION and a laptop with modest specifications is possible to sequence in real time few Millions of Gigabases, sufficient to support projects like ZiBRA, a mobile genomics lab that travelled across Northeast (NE) Brazil in 2016 to study the epidemiology of the Zika virus (Faria et al., 2016).
Nanopore technology allows real time sequencing, long reads and high throughput but are also characterized by a very high error rate (accuracy ranging 65%−88%) (Lu et al., 2016), heavily biased towards insertions and deletions.

1.3.5 Pacific bioscience SMRT technology

The second most important third generation sequencing technology is the single-molecule real-time (SMRT) sequencing approach developed by Pacific Biosciences (PacBio) (Carneiro et al., 2012; Hestand et al., 2016).
While the nanopore counterpart is the only technology able to directly detect the physical composition of an existing chain, the PacBio SMRT is the first approach able to directly observe the composition of a single nucleotide strand of DNA as it passes through an immobilized polymerase. The technology works by binding a single polymerase at the bottom of a very small hole and track the synthesis of the new nucleotide chain in the DNA duplication process catalysed by the enzyme using nucleotides-linked fluorophores. The hardest obstacle that needed to be overcome is allowing the detection of a single molecule signal. The problem was solved by confining the reaction in a very small volume and designing a special detection strategy. Exploiting the biotin/streptavidin binding to glue the polymerase at the bottom of a very small well approximately the same size of the molecule allowed the entry only to very few nucleotides inside the reaction chamber and thus raised the signal to noise ratio and allowed the identification of the nascent chain. At the same time the detection was designed so that only the bottom 30 nm of the well are provided with the excitation wavelength for the fluorophores and can produce the signals captured by the instrument. This is called zero-mode waveguide (ZMW) technology. A ZMW is a hole, tens of nanometres in diameter, fabricated in a 100 nm metal film deposited on a glass substrate. Thousands of these holes are presents in the sequencing cell and are targeted by a laser that produces a ~600nm wavelength. Because of the relative size, the hole is too small to permit the passage of the wavelength but the first few nanometres at the bottom of the well are still illuminated by the exponential decay of the light that shines through the surface. This allows the fluorophore of the nascent chain to be easily detected because the polymerase bound to the base of the well force them to be very close to the bottom of the chamber. When the correct nucleotide is detected by the polymerase, it is incorporated into the growing DNA strand in a process that takes milliseconds, approximately

three orders of magnitude longer than simple diffusion. This difference in time results in higher signal intensity for incorporated versus unincorporated nucleotides, which creates a high signal to noise ratio.
Another improvement compared to other sequencing by synthesis methods is the position of the dye in the nucleotide. Since SMRT sequencing does not need to stop the polymerase to read the base, in PacBio's platform the fluorophore that is used in labelling is attached to the phosphate chain instead of being linked to the nitrogenous base. This way the DNA polymerase activity is facilitated and as a natural step in the synthesis process, the modified phosphate chain is cleaved when the nucleotide is incorporated into the DNA strand. Upon incorporation of a phospholinked nucleotide, the DNA polymerase frees the dye from the nucleotide when it cleaves the phosphate chain and allows the label to quickly diffuse away.
Despite all of these precautions to purify the signal, the raw read is still characterised by a very low accuracy (~85% per base), and the sequence is dominated by incorrect insertions and deletions (~14%). To solve this issue the Pacific Bioscience team designed a new library preparation for SMRT platform that exploited circular library elements and long reads to generate a high-quality intramolecular consensus from each read (**Figure 1.3.2**). This type of library, called SMRTbell, is created by ligating two small DNA hairpins to the extremities of the double stranded collection of sequencing targets, usually DNA fragments or amplicons.
This library is thus similar to a collection of double stranded DNAs but, since the two strands are joined together, each element is topologically a circular single strand chain.
After a primer matching a recognition sequence on the hairpin is added to the reaction during sequencing, the polymerase at the bottom of the well is capable to bind this molecule and start the synthesis of a new chain. Because the DNA is circular, the polymerase travels on both strands of the original molecule several times and produce a very long read that contain several alternating repetitions of each side. The multiple sequenced copies of the insert can be used *in silico* to create an intramolecular consensus for each sequence of the library in a process called circular consensus sequencing and thus reduce the global error rate of the results. This type of consensus read is usually very accurate and the precision increases with the number of repetitions in the raw read. Contrary to second generation sequencing, the residual error

distribution is uniformly distributed along the read, providing data with striking low error position bias.

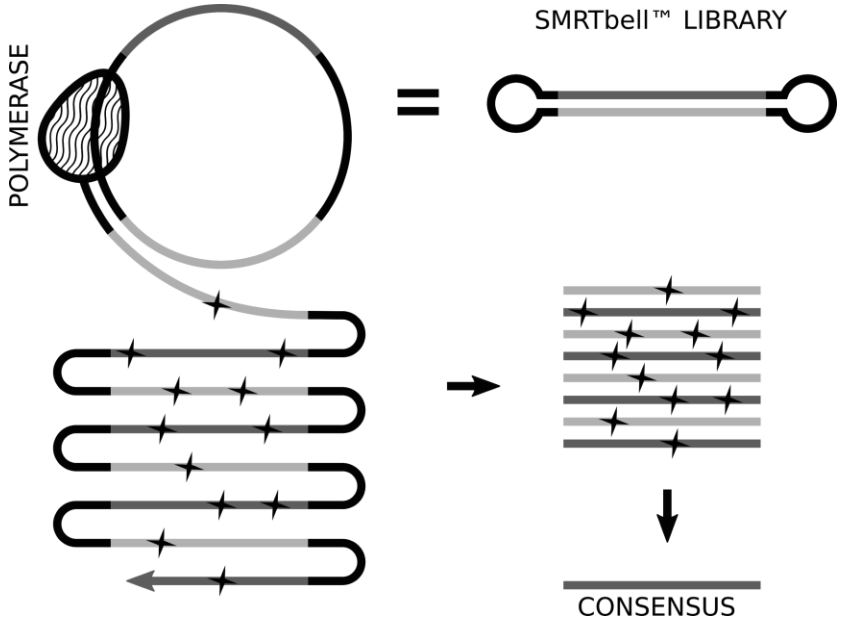


Figure 1.3.2 Schematic of the PacBio strategy of consensus sequencing. The immobilized polymerase at the bottom of the sequencing well catalyse the polymerization of a new DNA chain from the SMRTbell library element to which it is bound. The unusual circular conformation of the SMRTbell library allows the polymerase to copy both strands of the insert multiple times, which are then bioinformatically separated and merged to produce a high-quality consensus read.

Currently SMRT sequencing realizes longer read lengths than any other technology, producing average read lengths between 5-30 kb and maximum read lengths in excess of 200 kb (Kraft & Kurth, 2019). This was used to obtain de novo assembly of novel species (Ma et al., 2019) and to re-sequence the genome of known species to collect important variants (Audano et al., 2019). The throughput is the major disadvantage of this technology, being functionally linked to the number of ZMW wells. In the first functional prototype the throughput of SMRT sequencing was around 10K reads, but with the development of novel sequencing chemistries and the release of the newest Sequel platform, SMRT sequencing can produce about 300-600 K reads per sequencing cell with a theoretical maximum of 1 million reads.

1.4 TEM beta lactamase

1.4.1 History, nomenclature and numbering of beta lactamase

Beta lactamase is both a protein of critical clinical interest and an interesting molecule from an evolutionary perspective since it belongs to a family which contains several extremely similar members. The first scientific evidence of an enzyme capable of destroying the lactam antibiotics can be traced down to a paper from Abraham in 1940 (Abraham & Chain, 1940). Abraham and Chain described the *Bacillus (Escherichia) coli* "penicillinase" even before the therapeutic potential of penicillin was put on medical trials. The clinical use of penicillin was considered of low relevance since it was identified in *E. coli* and was not found in staphylococcal pus (Abraham & Chain, 1940) while at the time penicillin was targeted to treat staphylococcal and streptococcal infections (Abraham et al., 1941). In this pioneering work of Abraham and coworkers, the authors were not yet able to isolate the protein and even the fact that the physical agent that caused the recovery of a bacterial culture from the penicillin-derived inhibition was actually a protein was only a speculation based on strong evidence. The first success in isolating a penicillinase was reported by Kirby (Kirby, 1944) as the success in the extraction of a "highly potent penicillin inactivator" from strains of *S. aureus*. Kirby's report discloses, in contrast to what Abraham previously reported, that the strains from which he obtained the enzyme were "naturally" penicillin resistant and were all isolated from patients who had never received penicillin before. In 1961, Watanabe and Fukasawa (Watanabe & Fukasawa, 1961) reported that the ampicillin resistance can be transferred to other bacterial species and that the factors responsible for the transmissible drug resistance exist as cytoplasmic elements. Watanabe endorsed the usage of the term "episomes" for these factors of bacteria. In 1965, Datta (Datta & Kontomichalou, 1965) reported 2 novel antibiotic resistant strains of *S. typhimurium* from the United Kingdom and 1 new *E. coli* isolate from Greece. The British resistance (R) factors were designated R1818 and R7268 (originated in London and that encode the protein subsequently addressed as TEM-1), while the Greek one was named TEM. The TEM designation derived from the name of the Athenian patient, Temoniera, from whose feces the resistant *E. coli* strain was recovered in 1963 (Ruiz, 2018). In concomitance to the widespread use of penicillin and other similar lactam antibiotic in those years, beta lactamase quickly started proliferating among pathogenic bacterial strains in hospitals. Several medical institutions around the globe rapidly discovered numerous variants of the protein and gave them a name following the best nomenclature available. The first attempt to name and classify beta lactamases began when cephalosporinases were able to be differentiated from penicillinases. In 1968, Sawai and coworkers differentiated the two groups for their activity and complemented this information using the response to class specific antisera (Sawai et al., 1968). Richmond and Sykes classified in 1973 the beta lactamase in five major groups corresponding to different substrate profile (Richmond & Sykes, 1973), and in 1976 Sykes and Matthew extended the classification with an ulterior differentiation based on isoelectric focusing (Sykes & Matthew, 1976). Mitsuhashi and Inoue introduced a new category of beta lactamase able to hydrolyse cefuroxime in 1981 (Toda et al., 1981) and few years later Bush suggested another classification that, in addition to the substrate profiling, correlated substrate and inhibitory properties of beta lactamase with their molecular structure (Bush, 1989a, 1989b, 1989c). In

1995 Bush, extended her nomenclature (Bush-Jacoby-Medeiros Classification) (Bush et al., 1995) and further updated it in 2010 (Bush & Jacoby, 2010). Additional classification divided the beta lactamases into two broad sub-families according to the mechanism by which they perform hydrolysis, either through an active-site serine (Knott-Hunziker et al., 1979) or via essential zinc ions in the active sites of metallo-beta-lactamases (H. Zhang & Hao, 2011). When the sequence became available another category became a molecular class (also named Ambler classes A to D) based on molecular size and homology in active-site motifs (Ambler, 1980). The various classifications are summarized in **Figure 1.4.1**.

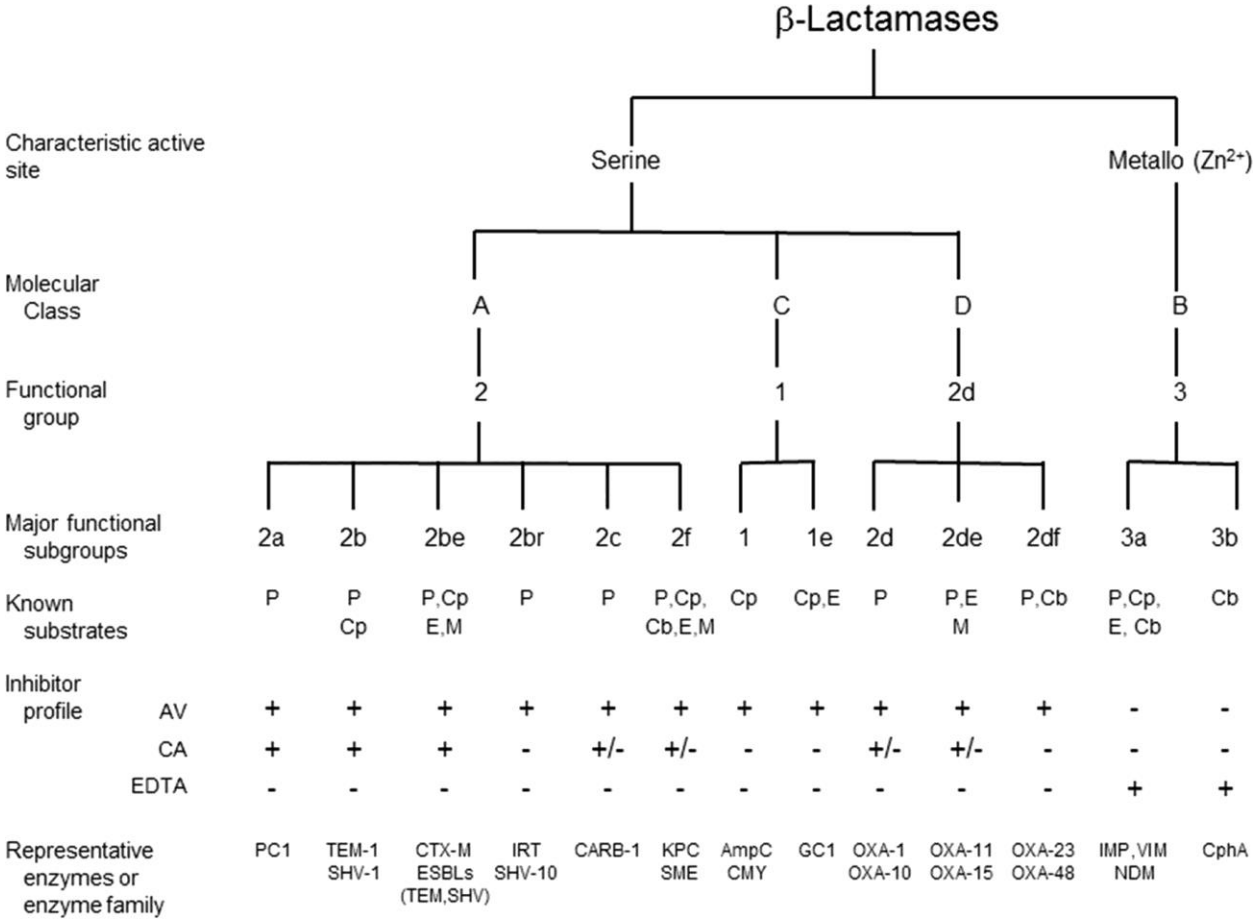


Figure 1.4.1 Molecular and functional relationships among β-lactamases. AV, avibactam; CA, clavulanic acid; Cb, carbapenem; Cp, cephalosporin; E, expanded-spectrum cephalosporin; M, monobactam; P, penicillin. (from (Bush, 2018)).

The chaos in the β-lactamase nomenclature was not limited to the general architecture of the family, but also several subfamilies were poorly organized. Groups around the world began discovered new forms of the enzyme and did not follow any specific rules for the nomenclature. The drastic increase in the number of the TEM family members made the situation even more chaotic so that Bush and Jacoby tried to patch the disorganized situation advocating to adopt a unique sequential nomenclature for the TEM lactamases in an iconic paper of 1997 (Bush, 1997):

"β-lactamase nomenclature has never been particularly rational. Enzymes have been named after a preferred substrate (CARB, FUR, IMP, OXA), other biochemical properties (SHV, NBC), genes (Amp, CepA), bacteria (AER, PSE), strains (P99), patients (TEM, ROB), hospitals (MIR, RHH), states (OHIO) and from the initials of their authors (HMS). A particular issue has arisen with the proliferation of naturally occurring TEM derivatives. TEM-1 and TEM-2 are biochemical twins differentiated by isoelectric point and now known to vary by substitution of lysine for glutamine at position 39. TEM13 resembles TEM-2 with a functionally silent methionine for threonine substitution at position 265. The other TEM enzymes listed in the Table either have an extended spectrum of hydrolysis, and hence are variably active on oxyimino-β-lactam substrates such as aztreonam, cefotaxime, ceftazidime and ceftriaxone (TEM-3 to -12, TEM-16 to -29 and TEM-42 and -43), or are resistant to inhibition by agents such as clavulanate, sulbactam and tazobactam (TEM-30 to -41). Missing TEM numbers (TEM-14 and 15, TEM-17 to -19, TEM-22 and TEM-23) either have yet to be fully sequenced or have been shown to be identical in amino acid sequence to others in the list. Before their defining alterations were known, many of these enzymes were given temporary but descriptive names such as CAZ-1 or CTX-1 to emphasize a preferred substrate (ceftazidime or cefotaxime, respectively) or, for example, IRT-2 to highlight an inhibitor-resistant TEM. This served a useful purpose by defining a functional characteristic of the enzyme. After the sequence was known, a TEM number was then assigned. The use of TEM numbers is now preferred because assignment of phenotypic names can be subjective. How rapidly must ceftazidime be hydrolysed in order to name an enzyme a 'CAZ' β-lactamase? What differential in activity is required for a 'CTX' designation rather than 'CAZ' for an enzyme that hydrolyses both substrates? How much must a K_i or an IC_{50} value be increased for an enzyme to be an 'IRT'? These criteria can be, and are, defined differently by individual groups. Since those working outside the β-lactamase field may legitimately argue that there are already too many names for β-lactamases, we urge that TEM derivatives be simply given a TEM number and not a further descriptive name."

Luckily the numbering of Class A beta lactamase, of which TEM-1 is a member, is not so confused. In 1991, Ambler proposed a numbering scheme based on sequence homology on few representative members (Ambler et al., 1991) (**Figure 1.4.2**) and to partially preserve a previous formulation used in the definition of Class A beta lactamase in 1980 (Ambler, 1980). Ambler specifies that this scheme is not meant to replace the individual protein numbering

scheme but to identify the residues in the context of homologous proteins comparison. This scheme is known as ABL (from “class A beta lactamase”) and it is still used nowadays.

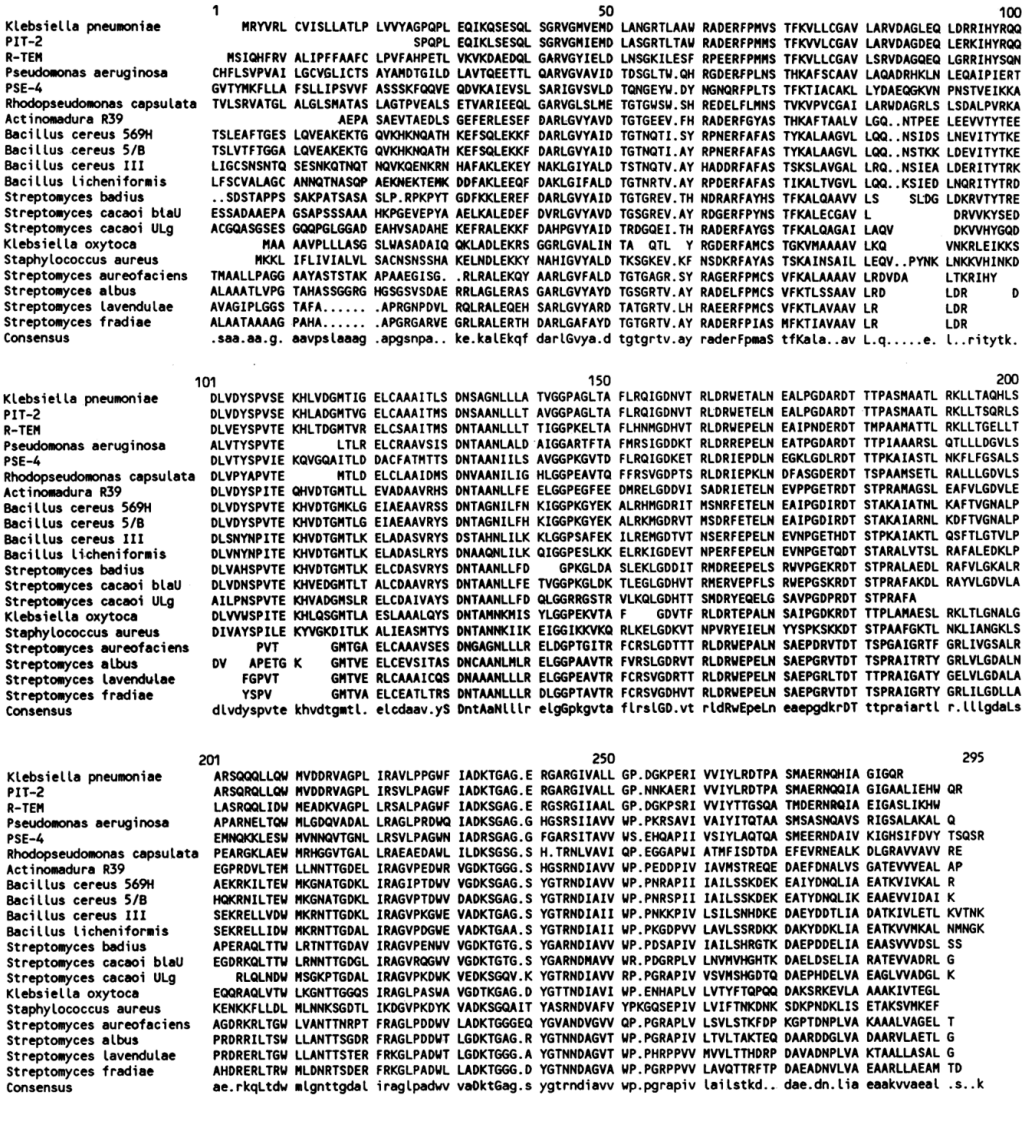


Figure 1.4.2 Alignment of the original 20 Ambler’s Class A beta-lactamases numbered according to the ABL scheme (from (Ambler et al., 1991))

1.4.2 Mechanism of antibiotic resistance

Bacterial cell wall is an essential component of the prokaryotic cell used by the organism to reinforce the membrane and protect it from osmotic hypertrophic stress and other environmental or biological hazards. The wall is structurally an amino sugar lattice called peptidoglycan organized in fibres composed by linear chains of two alternating sugar: N-acetylmuramic acid (NAM) and N-acetylglucosamine (NAG). Each NAM is linked to a short peptidic chain (usually a pentapeptide). By cross-linking two D-alanine of opposing NAM-peptides it is possible to join together two proximal amino sugar chains. This cross-linking of adjacent glycan strands confers the rigidity of the cell wall. The reaction is catalysed by the enzyme DD-transpeptidase which is a penicillin binding protein (PBP). β-Lactam antibiotics exhibit their bactericidal effects by inhibiting enzymes involved in this process of the cell wall synthesis. The β-lactam ring is sterically similar to the D-alanine-D-alanine at the terminal of the NAM pentapeptide, and acts as suicide inhibitors of the transpeptidase since the enzyme recognize it as a possible reaction substrate. The catalytic serine in the transpeptidase active site attacks the carbonyl of the β-lactam ring, resulting in the opening of the ring and formation of a covalent acyl-enzyme complex (Zapun et al., 2008). This complex is hydrolysed very slowly, thus effectively preventing further reactions. The level of transpeptidation of the wall components is a battleground where a constant transpeptidase crosslinking activity oppose the natural peptidoglycan autolysis. When β-lactam is added to the equation, the transpeptidase is not able to contrast the autolysis and the cell wall is rapidly compromised. The damage to the wall exposes the cell membrane to environmental hazards and causes cell lysis. To contrast the action of these antibiotics bacterial cells have evolved β-lactamases, enzymes that destroy the ring of β-lactam antibiotics before they suppress the activity of the wall transpeptidases. The antibiotic resistance of the protein derives from the β-lactamase enzymes hydrolysing the β-lactam ring of penicillin and derivatives (**Figure 1.4.3**).

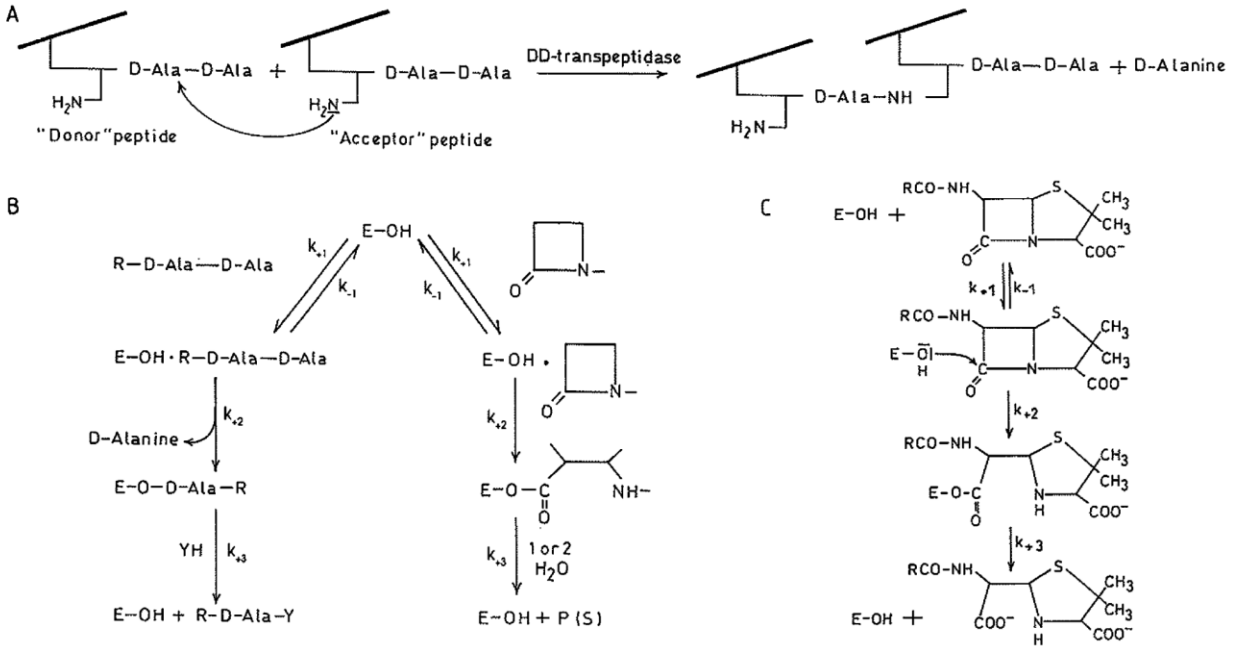


Figure 1.4.3 Mechanism of action of DD-transpeptidase and β-lactamase (adapted from (Lamotte-Brasseur et al., 1994))
A) The transpeptidation reaction catalysed by the bacterial DD-peptidases. The heavy lines represent the glycan moiety of the nascent peptidoglycan network. B) Interaction between the bacterial DD-transpeptidases (left branch) and their physiological substrates (R-DAla-DAla = donor substrate; YH = acceptor substrate) and β-lactams (right branch). C) The acylenzyme mechanism of β-lactamases.

1.4.3 TEM-1 β-lactamase structure

TEM-1 is a globular protein of ellipsoidal shape dimensioned around 30Å x 40Å x 50Å in size. TEM-1 is organized in a general α-β-α sandwich architecture with a five-stranded central beta sheet (Jelsch et al., 1993). The protein secondary structure elements are heavily biased towards helices, 39% of the residues are involved in α-helices and 5% in 3_{10} helices, while 17% percent are in β-strands.

The molecule can be divided into two structural domains, with the first including the helices at both termini of the protein (H1, H10, H11) and the central five-stranded β sheet into which they are packed, while the second is a domain made of eight helices (H2 to H9) located at the other side of the central sheet (**Figure 1.4.4**). The two domains are connected by two hinge regions which generate a large depression at their interface where the substrate binding site is located. Conformational changes around these hinges are prevented by a large number of hydrogen bonds and salt bridges that stitch the area in place. The main β-sheet is formed by five antiparallel strands and among them two belong to the N-terminal part of the chain (S1 and S2, residues 43-60), and the other three (S3, S4, S5) to the C-terminal (residues 230-266). In addition to the main β-sheet, there are two small two-stranded antiparallel sheets (SB and SC).

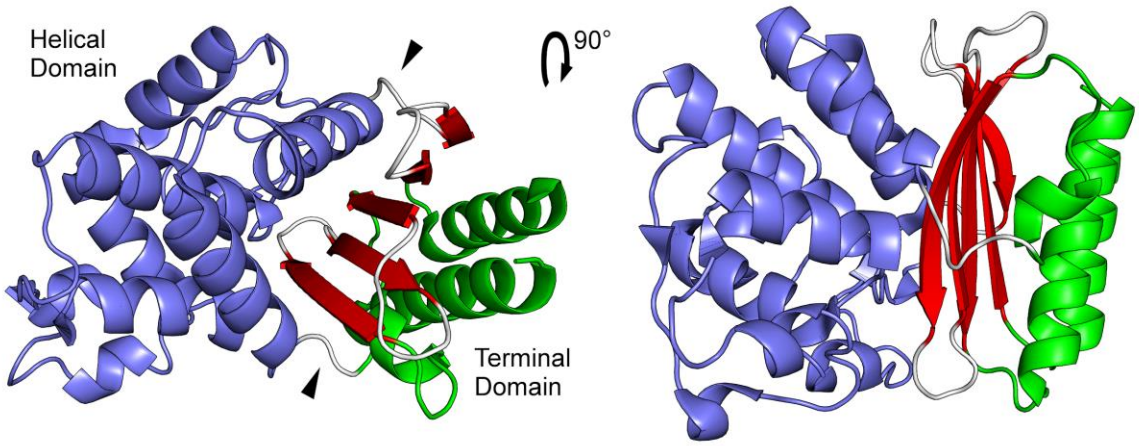


Figure 1.4.4 General structure of TEM-1 β-lactamase (PDB id 1ZG4) The protein is divided into two big subdomains: the green terminal helices and the red central sheet form the “terminal” domain while in blue on the other side of the beta sheet lies the “helical” domain. Black arrowheads indicate the two important hinge regions.

Strand S2 is made of five residues, while its adjacent strand S1 is eight residues long. This disparity in length means that S1 can form hydrogen bonds with S2 only on part of its length, leaving R43 and V44 at the N-terminal end of S1 without an obvious interaction partner. To accommodate this lack of interactions, the structure is stabilized by a hydrogen bond between the nitrogen in the main chain of V44 with E37 and a salt bridge between R43 and E64. Together with the contacts between E37 and R61 and the salt bridge R61-E64, these interactions form the core of the first hinge region. This hinge region locks the part of the interdomain crossover loop (residues 60-68), which links strand S2 to the catalytic helix H2. The

other hinge region (residues 212-222) is controlled by the salt bridge between R222 and the buried residue D233 (**Figure 1.4.5**).

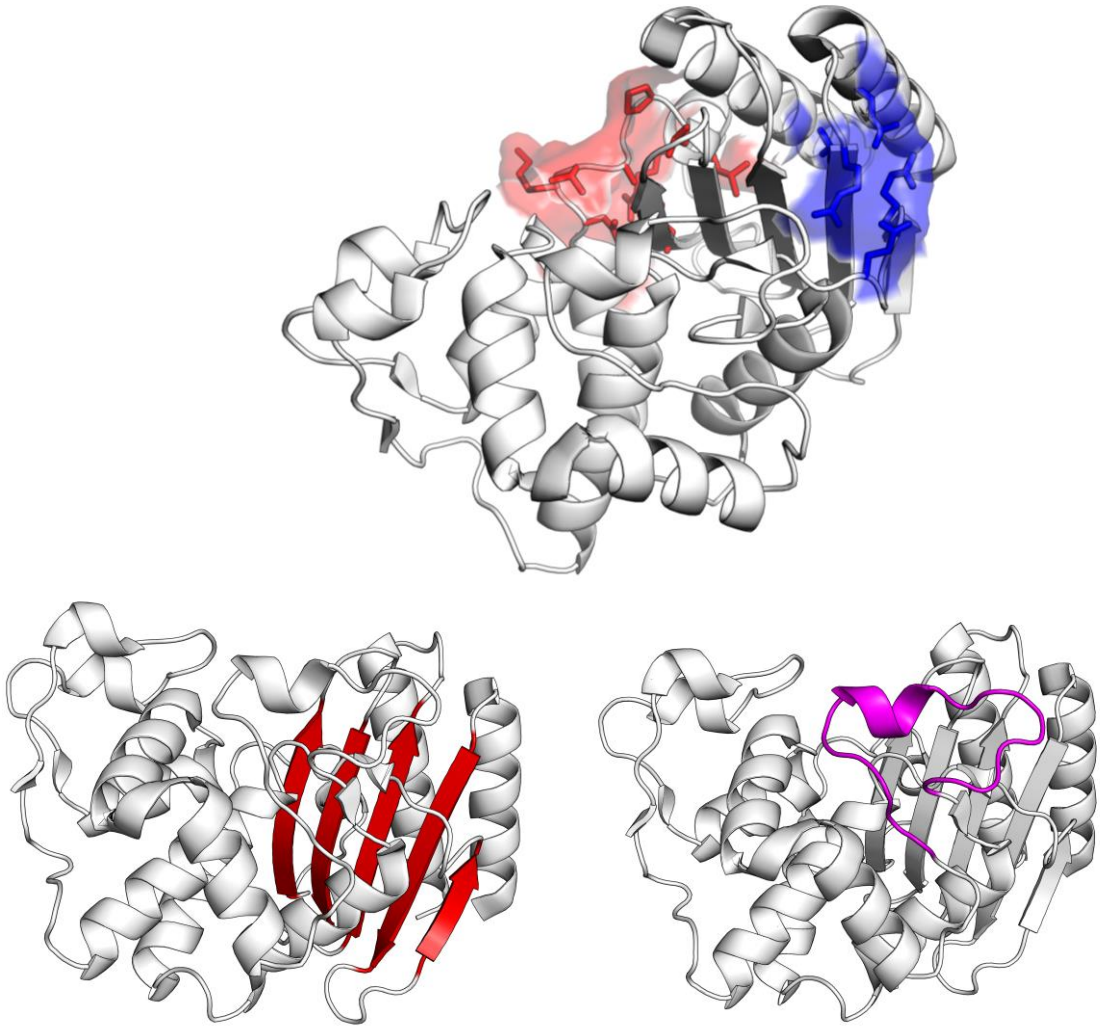


Figure 1.4.5 Features of TEM-1 β-lactamase (PDB id 1ZG4) Top: the residues interacting in the two hinge regions. Bottom left: the central beta sheet separates the two subdomains. Bottom right: the Ω-loop, characteristic of beta lactamases.

The Ω-loop (residues 161-179) forms another edge of the substrate binding site. The Ω-loop is a distinctive and important feature of class A β-lactamases. It contains the short 3_{10} helix H7 (residues 168-170) and the essential residue E166. The Ω in the name refers to the characteristic shape similar to the Ω letter of the Greek alphabet (**Figure 1.4.5**) and the two extremities of this peptide stretch are only 3.5Å apart. The few residues flanking the Ω-loop are held together by three hydrogen bonds which involve β-sheet SB. The Ω-loop is located at the protein-solvent interface but is strongly linked to the rest of the molecule.
Another notable part of β-lactamases is the innermost buries and catalytic helix H2. H2 in TEM-1 is one helix turn longer compared to other β-lactamases like PC1. In this turn, it is located the

characteristic residue D85 that is a specific residue present only in β-lactamases of Gram-bacteria. This helix does not have the same sequence nor the same characteristics in all the β-lactamases. In particular, the polarity of the cavity found in the vicinity of E166 of the Ω-loop is quite hydrophobic in TEM-1 and contains F72, L76, A135, L139, P145, L148, and L162 whose side chains display optimal van der Waals interactions. This helix also contains the disulfide bridge (C77-C123) that connects helices H2 and H4. This disulfide bridge is found in 7 out of 20 original Ambler's class A β-lactamases (**Figure 1.4.2**). If the bridge is removed by site-directed mutagenesis, the resulting enzyme is still able to function but has a heavily impaired thermostability.

Most of the catalytic residues of class A β-lactamases are invariant in the family. Among these S70, K73, S130, E166, and K234 are refined to low B-factors and, together with R244, few water molecules and a sulphate ion, are involved in complex hydrogen bond interactions.

The active site can be divided into two networks of interacting residues: On one hand the sulphate ion, S130, K234, WAT309, D214, D233, and R222 and on the other hand, K73, S70, WAT297, and E166. These two networks are connected through the WAT297-WAT323-sulphate ion interactions. WAT297, between S70 and E166, is in proper position as partner of

the deacylation step and WAT323 occupies the oxyanion hole for the lactam acyl of the drug (**Figure 1.4.6**).

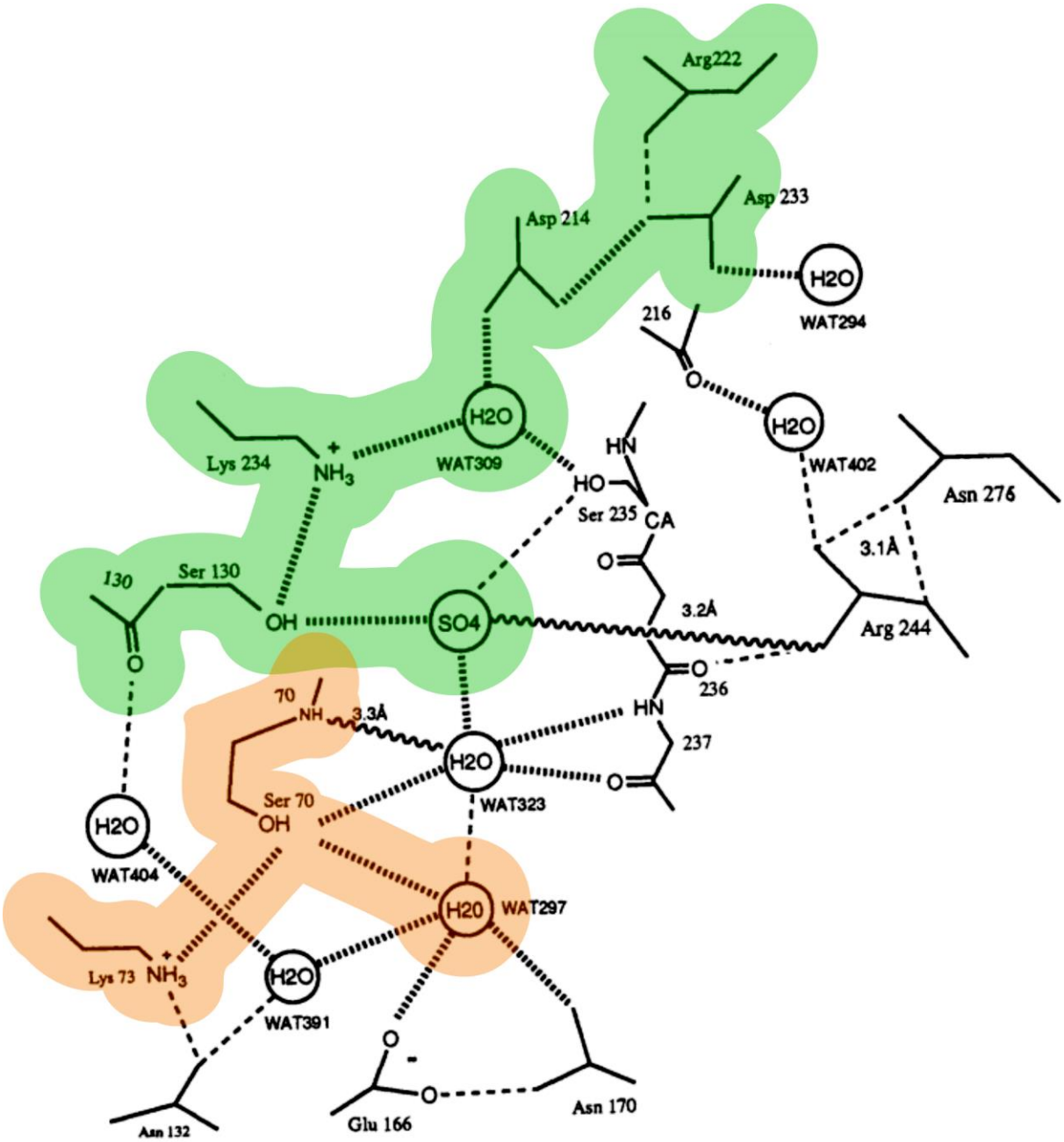


Figure 1.4.6 (adapted from (Jelsch et al., 1993)**) Interactions in the active site of TEM-1 β-lactamase (PDB id 1ZG4).** The two separated networks of interacting residues are coloured in green and red. WAT323 occupies the oxyanion hole for the lactam acyl of the drug.

1.4.4 A gold standard for molecular evolution

Bacteria that carry a functional beta lactamase can grow in a selective media supplied with lactam antibiotic like ampicillin. The selective growth is an extremely convenient phenotype to exploit to transfer genetic material in a bacterial cell. It is commonly used in cloning vectors for molecular biology to obtain strains that carry a desired plasmid. pUC19, the plasmid used in this work (Norrander et al., 1983), is one of the most common cloning vectors and carries a beta lactamase variant optimized for molecular biology tasks.

The simplicity of antibiotic selection caught the attention of other branches of biology and in particular it experienced a widespread popularity in the fields of biotechnology (Camps et al., 2003; Fujii et al., 2004; Long-McGie et al., 2000; Stemmer, 1994; Zaccolo & Gherardi, 1999) and evolutionary biology (Figliuzzi et al., 2016; Firnberg et al., 2014; Jacquier et al., 2013; Stemmer, 1994; Stiffler et al., 2015). Moreover, since the original discovery in 1963, several β-lactamase TEM variants began to appear, in particular coinciding with the introduction of a series of novel beta lactam antibiotics around the beginning of the 1980ies (Salverda et al., 2010). These variants typically contained one to three amino acid substitutions that caused an extension of the resistance spectrum to one or more of the new β-lactams. The surprising rapid rate of natural evolution of TEM-1 fuelled an interest in understanding the emergence of novel antibiotic resistance through evolution and through it predicting the evolution of said resistance to support the development of novel antibiotics. In 2010, Salverda (Salverda et al., 2010) collected the results of 25 independent publications on TEM-1 mutagenesis experiment of beta lactamase and 174 clinical TEM alleles proving that it is one of the most studied proteins both in mutational and evolutionary research.

2 Results

The retrieval of evolutionary couplings requires a very large number (at least some thousands) of analogues of a target protein to work properly. These sequences are collected from public sequence repository such as UniProt where scientists deposit a good portion of all the natural variants of the known or predicted proteins. However, the amount of protein variants that can be retrieved from databases is limited by the number of homologs deposited in them, thus when the target protein is rare or very recent, only few sequences could be obtained and the evolutionary couplings are not easily predicted. This limitation raises a major question: is it possible to use other types of data as a source for the evolutionary coupling analysis to circumvent the limitations posed by the number of protein variants deposited in sequence repositories? Theoretically any variant that maintains the fold of the original protein could be used as a source. However, since evolutionary coupling are essentially a tool to reconstruct the fold itself, this selection criterion is pointless. Instead, since the function of a protein is inseparable from its structure, it is possible to use a phenotypic selection on the protein function as a proxy for the collection of structurally similar variants. Protein variants generated with any type of agent or technique after being selected for the ability to perform a certain task are de facto identical to the ones produced by natural evolution. I am thus mimicking the driving force of natural evolution, and the couplings obtained from these data will still be fuelled by an evolution process, the only difference being that I will be using molecular evolution techniques instead of the classical natural evolution. To prove that molecular evolution could be used to generate variants for a structural analysis, a stereotypical model protein is needed. Due to the ease of selection, known structure, pre-existing vast collection of natural variants and extensive usage in mutagenesis experiments, TEM beta lactamase is an ideal choice for this proof of principle.

I defined the structural features observed in our experiment an in vivo determination of the structure, because the retrieved information is mirroring the conformation that beta lactamase

has in a living cell. To my knowledge, this is the first and only structure determination methodology able to do such an endeavour.
This work will include the first experiment where DCA was successfully applied to mutagenic data and can be used as a prototype for a series of experiments where structural information is extracted from multiple alignments of evolved proteins.

2.1 Project pipeline and rationale

The claim of this thesis is that it is possible to retrieve the evolutionary couplings of a protein, i.e. a statistical clue of the spatial proximity between two residues of a protein, from a laboratory-made collection of mutants.
The innovation presented is not the actual statistical analysis of the data to retrieve the couplings, since several techniques are already able to get the couplings from collections of the protein variants present in living organisms, but the type of data on which the analysis is imposed.
Being able to collect couplings from a laboratory-made library of sequences allows us to generalize the techniques when a collection of natural variants is not available.
Since the task is to prove the feasibility of the approach, I need to find the best conditions and methods to obtain a mutant library optimized for the subsequent coupling analysis.
I need, in this order: a good protein candidate to be our model system, a strategy to obtain mutations, a selection strategy to eliminate the non-functional variants, a sequencing platform to collect the sequences from the survivors and a statistical technique able to calculate the couplings.
The choice of the model protein and the selection strategy are tightly linked, since in a selection I need to test the function of the variants which is dependent on the protein chosen.
For example, at first, I considered as a model system the protein leu2, a protein able to catalyse the third step of the biosynthesis of leucine in yeast (that is the conversion of beta-ethylmalate into alpha-ketovalerate). I could select functional variants of leu2 by transforming a plasmid carrying the mutant gene in a yeast cell that lacks the leu2 enzyme. After transformation when the cells are grown in a medium lacking the amino acid leucine, only the yeast cells that

incorporate a plasmid with a functional leu2 gene are capable of the conversion of the leucine precursors and can produce the leucine needed to survive in the medium.
After serious consideration among several candidates, I chose the bacterial beta lactamase as a model system. Since beta lactamases are enzymes that catalyse the breakdown of the lactam ring present in many antibiotics like penicillin, ampicillin and cephalosporins, a functional beta lactamase can be easily selected by placing a bacterial culture transformed with variants of the enzyme in a medium supplemented with one of those antibiotics. Like the previous example, only the bacterial cells carrying a functional variant of the enzyme will be able to survive.
I chose beta lactamase because it has an easy selection strategy and for the hosting cellular system in which it is expressed, since bacterial cells are far easier to handle than eukaryotic cells. Moreover, it has a very strong portfolio of previous mutagenesis experiments where the protein was targeted for molecular evolution that could help us troubleshoot potential problems and can be used as a comparison during the analysis.
As a mutational strategy I first considered generating all possible double mutants of the protein by targeted mutagenesis, a technique that coupled with a selection process is better known as double mutant deep mutational scanning. However, generating such a library is extremely expensive, and since the number of double mutants grows with the square of the length of the protein, this procedure is even more expensive for longer protein and cannot be generalizable. Even though I was just testing a proof of principle and beta lactamase is a small protein on which deep mutational scanning could be applied, I wanted to propose a robust pipeline that would enable researchers to use the technology on any protein. Thus, it was imperative to find a mutational approach applicable to any protein of interest. Moreover, the coupling algorithms are proven to work with evolutionary data. In the case of deep mutational scanning instead the mutagenesis strategy would not take into consideration the evolutionary path a sequence must follow to obtain a certain mutation. For all these reasons, systematically test all double mutant combinations was not a safe bet.
I considered and tested as the source of mutations both error prone PCR, a technique that enhances the intrinsic low fidelity of certain polymerases in copying the DNA in a polymerase chain reaction (PCR), and a strategy that relies on adding to the PCR mix mutagenic nucleotide analogues that promote single nucleotide substitutions. Both of the methods work by promoting errors during DNA duplications, and thus they are similar to the process that would lead these errors to be generated in the natural course of evolution of the protein. Since I wanted to mimic the variants obtained from natural sources, I tested and analysed the similarity of our data to a collection of natural variants. This allowed me to validate the effects of the mutagenic technique and quantify the evolutionary landscape I obtained.
To streamline the data that were obtained from the molecular evolution to the coupling algorithm I needed to be able to digitalize the result in a collection of sequence variants. This was done by sequencing the DNA of the surviving organism after selection. Due to the amount of data required for an effective evolutionary coupling analysis, classical Sanger sequencing was not efficient enough to cover the high-volume demand and thus next generation sequencing was necessary.
The problem was that not many sequencing platforms are able to sequence these data. Because the technique requires sequence covariation, it was critical to obtain multiple mutations in the same sequence. However, most NGS platforms are unable to sequence more than a few

hundred nucleotides per sequence and therefore a complete read is achievable only in very small proteins.
To overcome these issues, the libraries needed to be sequenced with a specific type of single molecule sequencing technology called single molecule real time sequencing (SMRT).
The key advantage of single molecule real time sequencing (Eid et al., 2009) is the size of the DNA fragment sequenced (van Dijk et al., 2018), commonly over tens of thousands of base pairs, that undergo intramolecular circular consensus analysis to create a high quality read (introduction **chapter 1.3.5**) (Travers et al., 2010). For this project I sequenced all the libraries with Sequel, the latest SMRT sequencing platform of Pacific Bioscience.
Lastly, the analysis was performed applying Direct Coupling Analysis to the mutagenesis data.

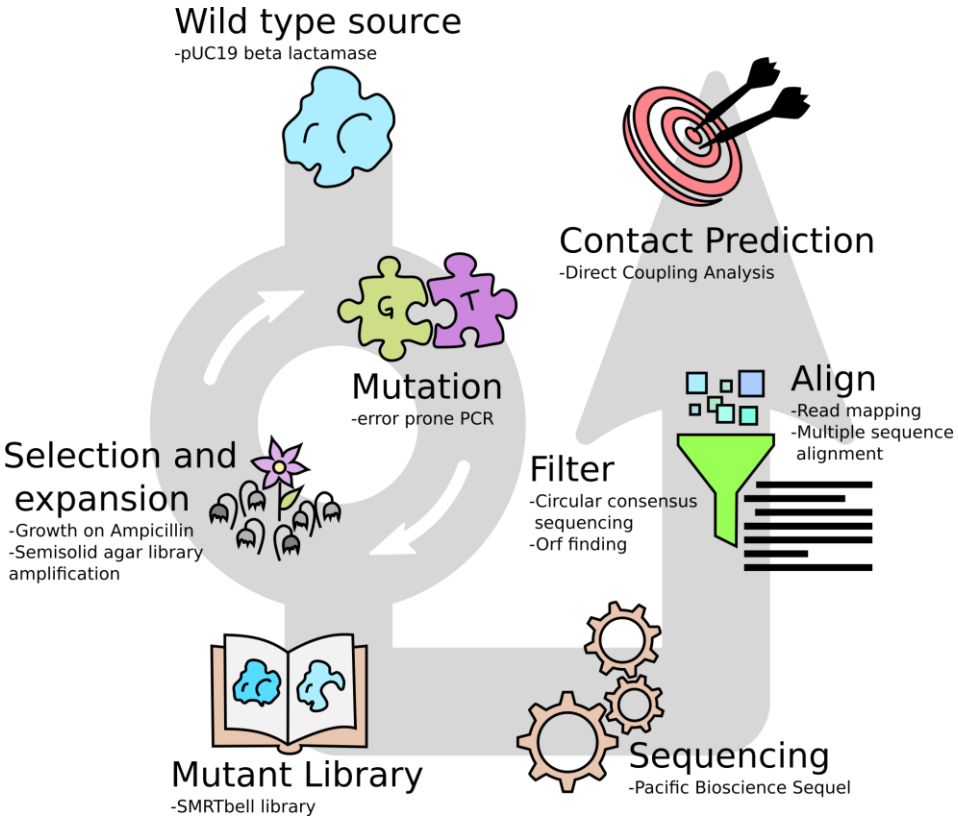


Figure 2.1.1 Schematic representation of the pipeline

2.2 Mutagenesis

2.2.1 Plasmid creation and primer design

I chose as the original beta lactamase for molecular evolution the TEM-1 beta lactamase of the plasmid pUC19 (Norrander et al., 1983). pUC19 is a plasmid with an exceptionally high transformation efficiency by electroporation that allowed me to easily obtain high complexity libraries. The sequence of beta lactamase in pUC19 is not the original TEM-1 beta lactamase (UniProt Q6SJ61) but presents two neutral aminoacidic substitutions in the coding sequence (Norrander et al., 1983). The protein has a well-defined crystal structure (PDB ID:1ZG4) (**Figure 2.2.1**).

I considered several types of mutational strategies, but only a few of them had the requisites to be the driving force of this molecular evolution, namely random mutagenesis with mutagenic nucleotide analogs and error prone PCR (Wilson & Keefe, 2001). Both solutions required a PCR reaction to generate the diversity and a ligation strategy to generate the plasmid library.

To simplify the library construction after mutagenesis I modified the original plasmid inserting two restriction sites flanking the lactamase gene. This allowed to perform a simple cut and paste of the lactamase gene from the PCR product to the backbone plasmid.

The position of these restriction sites is rather critical: the restriction sites must be included in the primers I use in the mutagenic PCR, so they have to be close to the coding region. However, any alteration in the 5' region of a gene could hinder the translation of the gene damaging the overall expression level. In particular, weakening the Shine-Dalgarno ribosomal binding site or damaging the transcription level by altering the RNA polymerase binding sites in the promoter can be potentially harmful for the selection. I designed and tested two alternative versions of the plasmid, one version (pUC19a) carrying the 5' restriction enzyme binding site just before the Shine Dalgarno motif (Hung et al., 1989), the other (pUC19b) carrying the restriction enzyme just before the transcription start site (Brosius et al., 1982) (**Figure 2.2.2**). Both plasmids exhibited a high electroporation efficiency and survival rate in competent cells similar to the original plasmid counterpart. Since pUC19a had the restriction enzyme closer to the transcription start site, this plasmid was preferred as the progenitor of the library. This has the advantage to require a smaller primer for mutagenesis.

The forward and reverse primers for the mutagenic PCR needed to contain the restriction sites to allow digestion of the PCR product for cloning, and, respectively, the start and stop codons of the beta lactamase to guarantee the correct translation of the protein.

By keeping the start and stop codons in the primers, these codons are protected from the mutagenesis, whilst all the nucleotides in between the primers will be subjected to amplification with reduced fidelity. In other words, since the new strands are generated elongating the annealed primer, the primer region that started the polymerization will never be copied and it is not a target of the error prone duplication catalysed by the polymerase. Thus, any error in the region will not be propagated in future duplications and the primer regions are protected from mutations.

In view of this and considering that the backbone of the plasmid is not subjected to any mutational pressure, the only part of the plasmid that underwent mutation spanned from the first nucleotide of the second codon to the last coding nucleobase before the terminator 'TAA'.

The first 23 amino acids of the beta lactamase code for to the leader peptide of the protein and they are cleaved for secretion. Hence the mature enzyme does not contain the immutable starting methionine and is fully mutagenized.
On the other hand, this strategy allows mutations to form in the leader peptide and that could hinder both the cleavage and secretion of the protein, limiting its function. I also considered an alternative primer design in which the forward oligo encompassed the whole leader peptide coding region. To do so I would have been forced to use long primers in high temperature PCR reactions, a condition that would have been unfavourable to the mutagenic amplification.
Moving the restriction enzyme inside the coding region of the beta lactamase is also a solution but could alter the protein and I decided against it. Moreover, the strategy I chose allowed the observation of mutations in the leader peptide and the relations the peptide forms with the rest of the protein which are per se interesting additional details that can be analysed.

```
UP_Q6SJ61_TEM1     1 MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRP
PDB_1ZG4_TEM1      1 ----------------------HPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRP
pUC19_TEM1         1 MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRP

UP_Q6SJ61_TEM1    61 EERFPMMSTFKVLLCGAVLSRVDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVREL
PDB_1ZG4_TEM1     38 EERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVREL
pUC19_TEM1        61 EERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVREL

UP_Q6SJ61_TEM1   121 CSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTM
PDB_1ZG4_TEM1     98 CSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTM
pUC19_TEM1       121 CSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTM

UP_Q6SJ61_TEM1   181 PAAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGS
PDB_1ZG4_TEM1    158 PVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGS
pUC19_TEM1       181 PVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGS

UP_Q6SJ61_TEM1   241 RGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW
PDB_1ZG4_TEM1    218 RGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW
pUC19_TEM1       241 RGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW
```

Figure 2.2.1 Comparison of beta lactamase sequences
Alignment of the pUC19 TEM-1 beta lactamase, the regular TEM-1 beta lactamase (UniProt id: Q6SJ61) and the TEM-1 lactamase from the crystal structure (PDB id:1ZG4)

bla promoter - bla Junction (*XhoI*)

```
pUC19    ...tacattcaaatatgtatccgctcatgagacaataaccct   gataaatgcttcaataatattgaaaa    aggaagagtATGAGTATTCAACATT...
pUC19a   ...tacattcaaatatgtatccgctcatgagacaataaccct   gataaatgcttcaataatattgaaaactcgaggaagagtATGAGTATTCAACATT...
pUC19b   ...tacattcaaatatgtatccgctcatgagacaataaccctcgagataaatgcttcaataatattgaaaa    aggaagagtATGAGTATTCAACATT...
```

bla - 3'UTR Junction (*NheI*)

```
pUC19    ...TAGGTGCCTCACTGATTAAGCATTGGTAActgtcaga      ccaagtttactcatatatactttagattgatt...
pUC19a/b ...TAGGTGCCTCACTGATTAAGCATTGGTAActgtcagagctagccaagtttactcatatatactttagattgatt...
```

Figure 2.2.2 Comparison between the modified pUC19 plasmids and the original in the flanking regions of the beta lactamase gene.
In purple and magenta: the -35 and -10 (Pribnow box) elements of the promoter. In bold black: the transcription start site. In green: the Shine-Dalgarno ribosomal binding site. In capital letters: the coding region of the beta lactamase gene. In brown: the stop codon of the coding region. Underlined: the restriction enzymes cut sites (*XhoI* at the 5' and *NheI* at the 3' of the gene).

2.2.2 Mutation strategy

Both nucleotide analogs and a low fidelity error prone PCR reaction are good sources of mutations that permit to obtain, quickly and cheaply, many variants of the protein of interest, but not without some intrinsic bias (see introduction **chapter 2.2.3** and **chapter 2.2.4**).
I wanted to see which of the two approaches was able to generate the most mutations, since for thus project more mutations are better than less. The other fundamental criterion for the choice is the number of survivors after the selection that the method is able to generate. This is a more complex parameter since it takes into account not only the sheer number of mutations but also their type. The number of survivors is naturally linked to the mutation rate, because the probability of the protein remaining functional gets lower and lower as it mutates. This means that the mutational power of the technique has to be modulated to guarantee both a good number of survivors and a sufficient amount of overall mutations.
The type of mutations also matters: while a single nucleotide substitution could have negative, neutral or even beneficial effect on the function of the protein, insertion and deletion will generate a frameshift that will inevitably result in a short non-functional peptide. The only exception to this rule is a second concerted insertion/deletion event that put back the protein frame on the correct track.
It was difficult to a priori decide which method was to be preferred between the two, so I decided to run some preliminary tests.

2.2.3 Mutagenic nucleotide analogs

The quantity of input DNA in the PCR reaction plays a crucial role for the generation of mutations. While this aspect plays no relevant role in theoretical unlimited reactions, this is very limiting in a small PCR reaction where the exponential growth of the product rapidly depletes primers and nucleotides. In reality, regardless of how many PCR cycles I subjected the samples, when the capacity of the reaction is saturated, i.e. when the reagents get depleted, no more duplications can occur and the reaction stops. This poses a limit to the total number of DNA duplications in a reaction. This limit is strictly dependent on the quantity of input DNA since the more copies I start from, the faster the reagents will be depleted. While the actual probability of nucleotide mutation is fixed, the more DNA duplication occurs, the more probable the occurrence of a mutational event became, thus the number of PCR cycles in non-saturating conditions is one of the valves that controls the final mutation rate.
Different groups use protocols that use different concentration of input DNA in the reaction mix, from 15 nM (Zaccolo et al., 1996) to under 500 pM (concentration suggested by the protocol of the reagents supplier, Jena Bioscience).
Considering a complete elongation of all the primers in the PCR mix, the amount of amplicon that could be obtained from the reaction is limited by the primer concentration (500 nM) in the reaction solution. This is an issue, because even for the lower condition of 500 pM input DNA, the fraction of the original input DNA, that is by definition without mutations, could be no less than 0.1%. This is not good for the previously discussed problem of scarcity of reagents,

meaning the larger the initial amount of DNA in the reaction, the fewer non-saturating amplification cycles can be completed and the lower the final mutation rate will be. Making things even worse, many mutated variants will likely be non-functional, further reducing the wild type to mutant ratio and the molecules with multiple mutations have a lower chance to be functional, with most disappearing after selection, thus further reducing the apparent mutation rate.

To balance between the PCR requirements and the input DNA concerns, I tried four different concentrations from 200 to 0.2 pM through serial 1/10 dilution of the input DNA that for our 870 bp beta lactamase gene fragment corresponded to a 5.74 ng to 5.74 pg (10 to 0.01 fmol) DNA input in a 50 µL reaction. Only the 200 and 20 pM concentration conditions produced a detectable PCR product. I then subcloned the DNA obtained at the lower concentration in the modified vector plasmid and electroporated the resulting library in competent bacterial cells. No growth could be observed in selective media.

This strategy resulted in a very low overall mutation rate, with over 1/25000 of carryover wild type input DNA from the PCR reaction and did not produce colonial growth in selective media. To the project means this protocol failed to fulfil the task required to act as our source of mutation and I moved to the error prone PCR protocol.

2.2.4 Error prone PCR

The protocol of error prone PCR I decided to employ (Wilson & Keefe, 2001) took into serious consideration the issue of the saturation of the reaction and depletion of reagents and solved it in a very simple manner. A small aliquot of the reaction solution is taken after a set number of cycles and diluted in fresh reaction mixture to reduce the amplicon level and maintain the duplication process in a continuous exponential phase. This strategy could probably be applied identically to the nucleotide analogues to obtain similar results, but as we shall see, the mutagenesis with error prone PCR already gave very satisfactory results.
Following the protocol, I diluted 1:10 the reaction in fresh reagents after every 4 PCR cycles. This was sufficient to maintain the amplicon level constant throughout the reaction, as it can be seen in a staining of polyacrylamide gel electrophoresis of various stages of the amplification process (**Figure 2.2.3**).

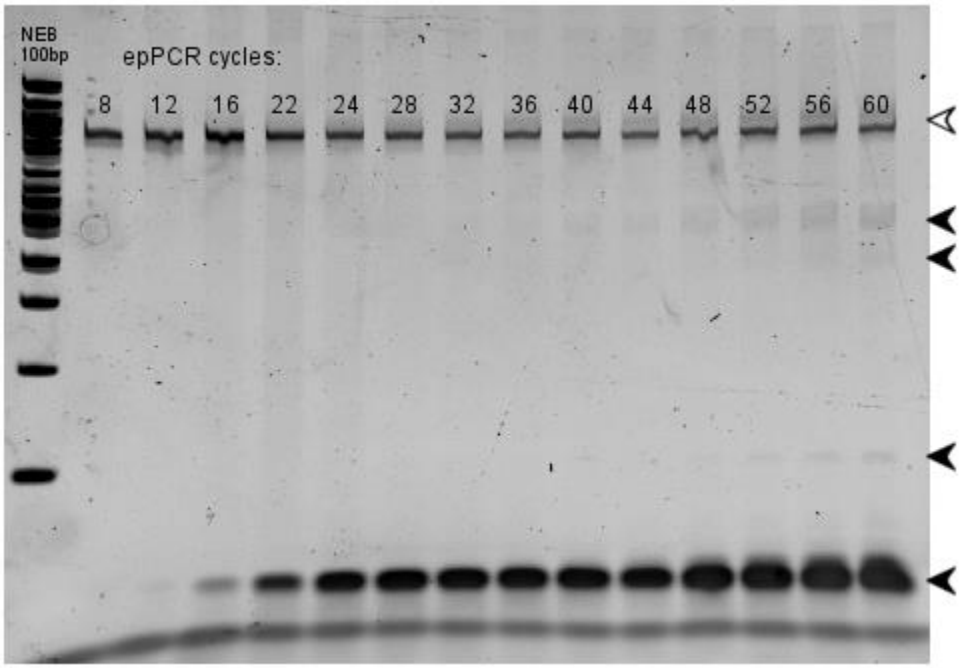


Figure 2.2.3 Amplicons present in the error prone PCR reaction solution at different time points of the PCR amplification.
White arrowhead: bands at the expected molecular weight for the beta lactamase amplicon.
Black arrowhead: secondary bands at unpredicted molecular weight.

The reaction also showed to have produced small secondary products at different size ranges, probably due to the accumulation of errors in the amplicons creating new regions where the primer could anneal. Sometimes small molecular weight byproducts of the PCR began to take over as the most abundant products of the reaction. This was a potential issue discussed in the protocol. However, this phenomenon was never an issue since it was spurious and seemed to be problematic only in the latest stage of the amplification. When this occurred, I gel-purified the band corresponding to the molecular weight of the expected product and proceed forward as normal.
At this stage I deviated from the protocol, inserting before the digestion of the amplicon with

restriction enzymes an extra amplification step with a high-fidelity PCR. This extra step increased the quantity of the available amplicon and separated the two strands that form the last amplicon in different molecular lineages. This is crucial because epPCR promotes a low fidelity copying of the template DNA and this in turn causes regions of the double stranded helix with a non-canonical pairing of the nucleobases to be rather common. If not corrected, the strands would have been segregated only once inside the bacterial host cell creating a mixed-plasmid population where non-functional plasmids could be carried over parasitizing the cells and exploiting the activity of the functional cognate plasmids.
Moreover, when sequencing plasmids with non-matching strands, the single molecule sequencing will read both of these strands several times, and a mismatch caused by a non-canonical pairing will create an inconsistency between the stands that will hinder the creation of a robust intramolecular consensus (we will discuss this in **chapter 2.3.1**).
After this second PCR the band of the correct molecular weight was purified, digested with the restriction enzymes (*XhoI* and *NheI*) and cloned in our vector to be transformed into competent cells.
I first tested the products of various steps of the error prone reaction, described as the total number of cycles done up to the point to understand which mutational load was still able to produce a sufficient number of survivors after library transformation. Of the four conditions tested, i.e. after 20, 32, 40, 60 cycles of mutagenic PCR, only the library derived from the reaction stopped after the 20th cycle was able to grow colonies in selective media (**Figure 2.2.4**).

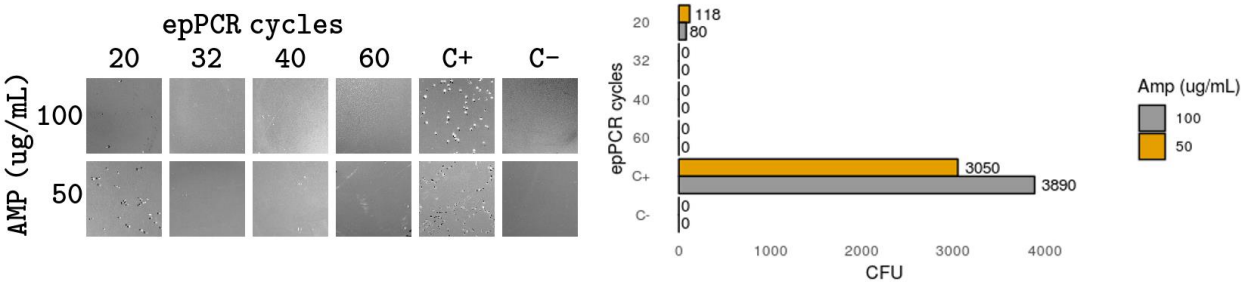


Figure 2.2.4 Effect of mutational load on cell survival under selective pressure
Left: survival in selective media of competent cells transformed with libraries producing beta lactamases obtained after different cycles of the error prone PCR. The positive control contains the original pUC19a TEM-1 lactamase, the negative control is an empty vector. Right: colony count after the selection.

Depending on the ampicillin concentration of the media, I observed different survival rates in respect to the wild type control (C+), 3.87% on ampicillin 50 µg/mL while 2.06% on ampicillin 100 µg/mL, supporting the idea of an effect of the mutations on the lactamase ability to destroy the antibiotic (see **chapter 2.2.6** for a discussion on antibiotic concentration).
From the growth in the 50 µg/mL ampicillin medium, ten colonies (A-J) were isolated and had their plasmid extracted and sequenced. The differences found in these sequences when compared to the wild type gene or protein are shown in **Table 2.2.1**. I obtained a significant amount of mutated DNA positions, with a mean of 6.67 mismatching nucleobases per sequence corresponding to a nucleotide mutation rate of 0.86%, and a good amount of differences in the

composition of the peptidic chain compared to the wild type. The sequences contained a mean of 3 aminoacidic substitutions corresponding to a mutation chance of 1.15% per amino acid position.

Sample	*DNA mutations (bp)*	*Protein mutations (aa)*
bla20_50A	11	5
bla20_50B	3	2 (1 STOP)
bla20_50C	9	3
bla20_50D	6	2
bla20_50E	5	2
bla20_50F	3	2
bla20_50G	5	4
bla20_50H	6	4
bla20_50I	6	4
bla20_50J	9	1
mean	*6.67*	*3*

Table 2.2.1 Error prone induced beta lactamase mutation rate observed after selection. Differences in the nucleic and amino acid sequence observed in the beta lactamase of a small sample of colonies (A-J) that survived the selection with 50µg/mL Ampicillin. The bacterial cells were transformed with a library of beta lactamases obtained after 20 cycles of error prone PCR. The samples containing early stop codons (greyed out in the table) were not used to calculate the means.

For this small test I did not obtain a complete coverage of the sequence for all the samples, therefore I limited the analysis to the first 780 nucleotides of the lactamase gene, corresponding to the first 260 amino acids of the peptide chain.
Comparing these numbers to the expected number of mutations under similar conditions for a random mutation in a DNA stretch of similar size (5.3 - 11 mismatching base pairs, 4 - 8.1 mismatching amino acids) (Wilson & Keefe, 2001), we can observe a concordance in the nucleotide mutation rate and a disagreement in the peptidic rate. This is likely due to the selection process. Most of the nucleotide mutations in this sample did not result in an amino acid substitution but only exist as silent mutations of the gene. Since any mutation has the chance to damage the function of the protein, a heavy presence of silent mutations serves to prove that the selection process is working as intended.

2.2.5 Ligation strategy optimization

Before doing any big scale transformation, the ligation process for the generation of the library had to be validated and optimized. During the preliminary experiments with mutagenic nucleotide analogs I observed a high degree of concatemerization of the reagents during ligation (**Figure 2.2.5**). The most evident concatemer was a byproduct weighing the same as a dimer of the backbone plasmid. These byproducts must be contained because, even if they do not contain the beta lactamase gene and thus cannot be selected, they would still damage the transformation efficiency by reducing the quantity of the effective properly ligated product.

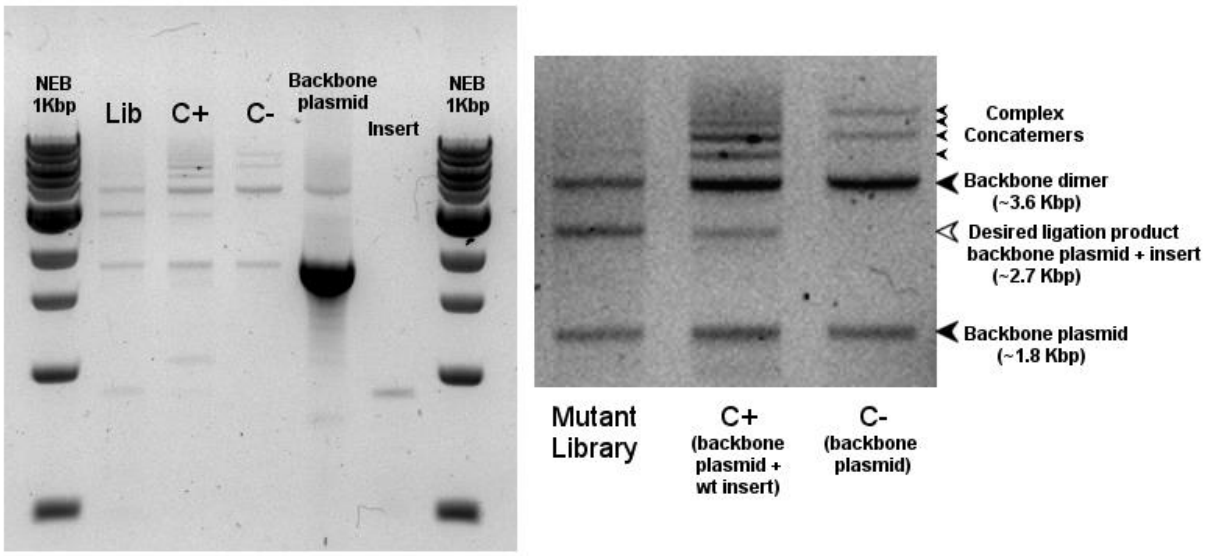


Figure 2.2.5 Concatemerization observed during the preliminary trials of library construction.
Left: Gel electrophoresis in agarose of the ligation products (library) of beta lactamases obtained after 20 cycles of error prone PCR to the backbone vector and controls. Right: magnification with increased contrast of the gel to show the different products at high molecular weight.

To reduce concatemerization during the ligation of the libraries from error prone PCR, the digested backbone plasmid was treated with Calf Intestinal Phosphatase (CIP), an enzyme that catalyses the removal of the terminal phosphate group of DNA fragments and prevents their self-ligation.
However, ironically, I observed an even higher degree of concatemerization in these libraries (**Figure 2.2.6**) compared to the previous attempts, because even if the backbone plasmid could not dimerize, the insert concatemerization is still a problem.

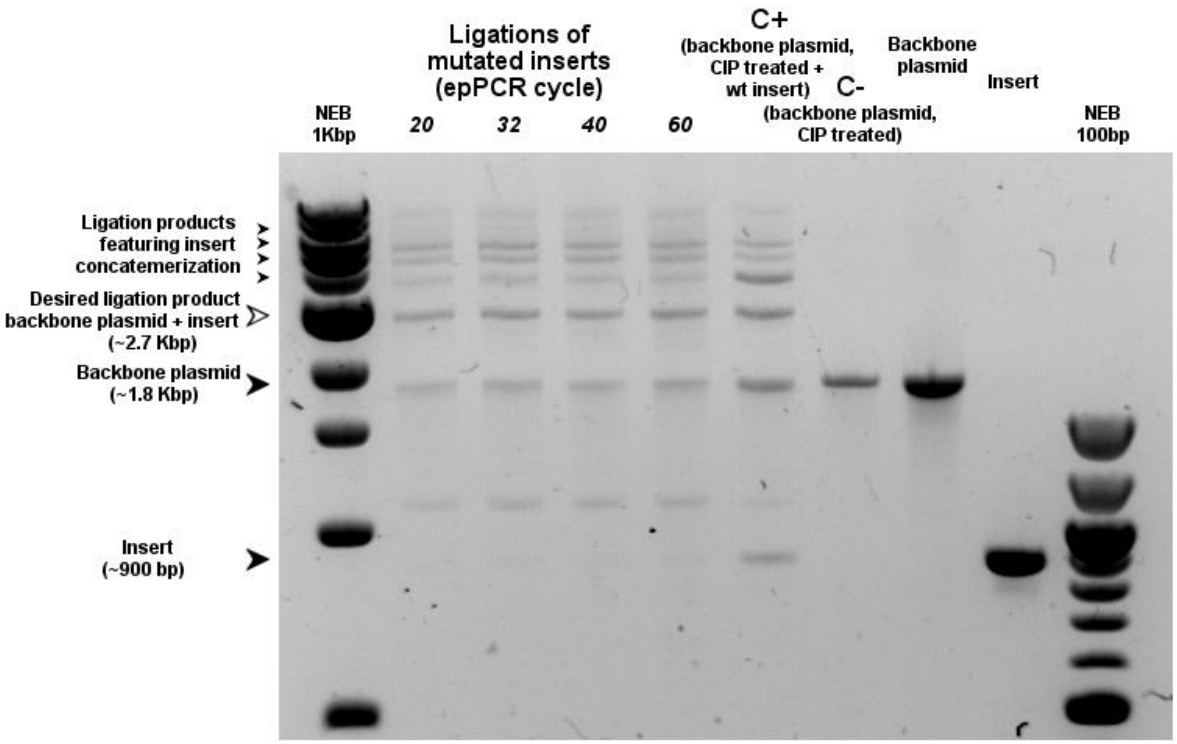


Figure 2.2.6 Concatemerization observed in ligation libraries of beta lactamases carrying different mutational loads and where the backbone plasmid was treated with CIP.
Gel electrophoresis in agarose of the libraries of beta lactamases obtained after 20, 32, 40 and 60 cycles of error prone PCR to the backbone vector and controls. The backbone plasmid fragment used in the ligation was treated with Calf Intestinal Phosphatase (CIP) after the restriction enzyme digestion.

Insert dimerization is even more problematic than backbone dimerization, because a plasmid with multiple copies of the insert would actually carry an antibiotic resistance gene that allows a cell to survive. After screening for functional variants, it would be impossible to discriminate which of the copies of the lactamase gene present in these objects was the one that conferred the bacteria the ability to survive under the antibiotic selective pressure. This condition would have been problematic and not a solution for our problem.
A positive result instead was that the ligation profile of the several libraries built from products of different stages of error prone PCR proved to be identical. This was good news because it implies that the ligation efficiency was not affected by the mutational load present in the insert. While this seems rather obvious, it was a very important control because the lactamase is subjected to a very strong mutagenesis during the course of the experiments. In the later stages of the project the mutational load on the lactamase will inevitably increase drastically, and if the ligation efficiency was damaged by the mutations then the bacteria surviving the selection would be very few and the libraries would not reach a sufficient diversity to support this project.
To enhance the efficiency of the ligation, I also checked which molar ratio of backbone plasmid would be optimal to obtain the relative maximal yield of the desired ligation product (**Figure 2.2.7**).

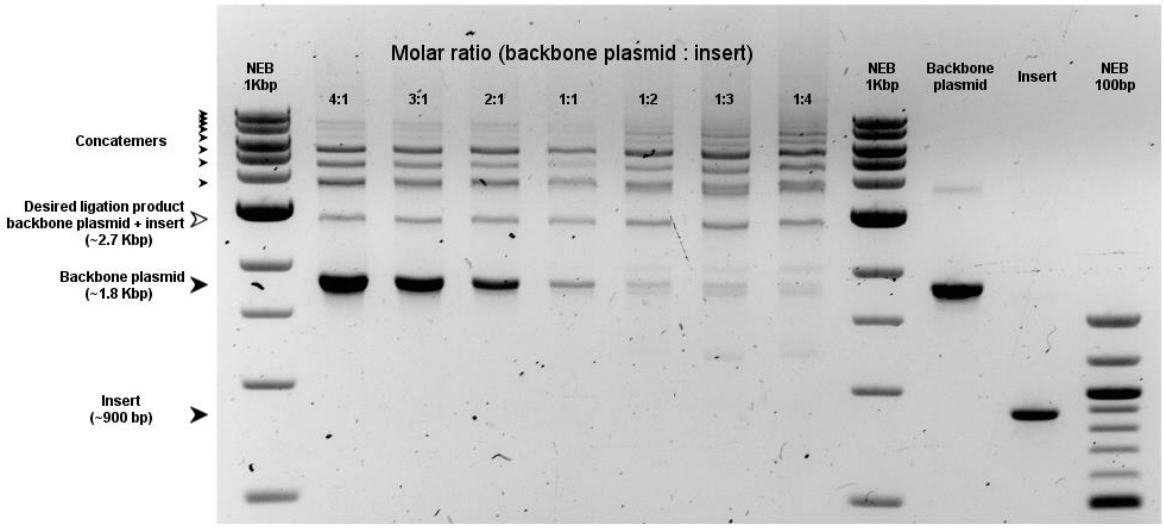


Figure 2.2.7 Concatemerization observed in ligation libraries of beta lactamases at different backbone plasmid: insert ratios.
Gel electrophoresis in agarose of the ligations of the beta lactamase insert obtained after 20 cycles of error prone PCR to the backbone vector at different ratios and controls.

Figure 2.2.7 shows that an increase of either the backbone plasmid or the insert from an equimolar 1:1 ratio did not have any significant effect on the amount of the desired products obtained. Increasing the quantity of backbone plasmid resulted in a strengthened signal of the bands produced by the concatemerization of the vector as well as the signal of the unligated vector itself. On the other hand, increasing the amount of insert did manage to reduce the signal of the unligated backbone plasmid but, instead of using this vector to produce the correct plasmid, it was employed to generate a great quantity of high molecular weight byproducts (i.e. a complex mixture of concatemers). The 1:1 ratio resulted to be optimal, but again did not solved the problem.
As a result of these preliminary tests I designed an unusual protocol where the insert was subjected to the CIP treatment instead of the common practice of CIP-treating the plasmid backbone. In addition, the insert was employed in an equimolar ratio to the backbone vector. The result of this procedure constituted the first generation error prone PCR library (**Figure 2.2.8**).

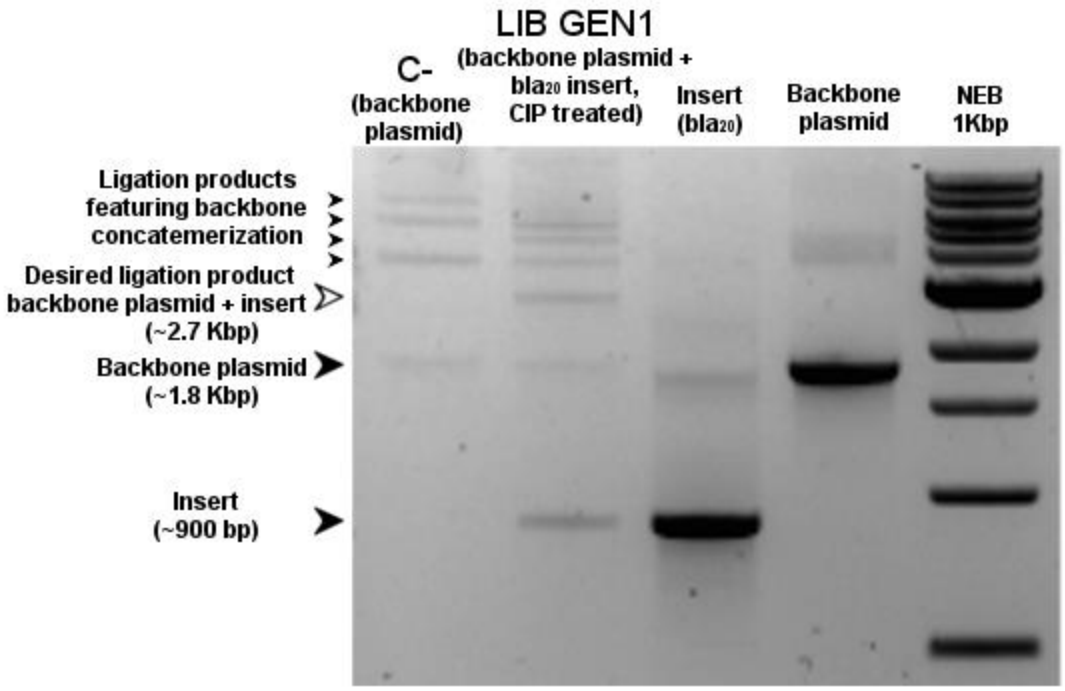


Figure 2.2.8 First generation of the error prone PCR libraries after ligation.
Gel electrophoresis in agarose of the ligation of the beta lactamases library obtained after 20 cycles of error prone PCR to the backbone vector and controls. This library has been labelled as the first generation of the error prone PCR libraries. The insert fragment used in the ligation was treated with Calf Intestinal Phosphatase (CIP) after the restriction enzyme digestion.

The library showed the strongest signal among all ligated products in correspondence to the desired plasmid and this protocol reduced the chance of a multiple incorporation of the lactamase gene in the vector. It was still possible to observe several byproducts generated by the concatemerization of the reagents, some of which were absent in the negative control.
To rule out the possibility of carrying over these byproducts after the functional selection in bacteria, I migrated on gel the plasmid extract obtained after the selection (see **chapter 2.2.7**) both in the circular and linearized forms (**Figure 2.2.9**).

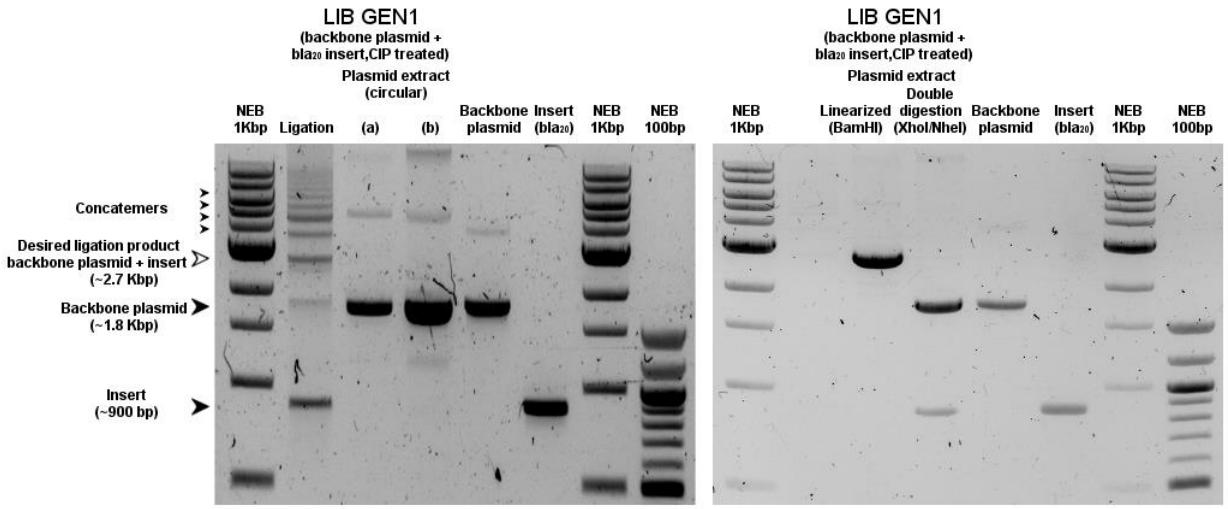


Figure 2.2.9 First generation beta lactamase library after selection.
Left: gel electrophoresis in agarose of first generation beta lactamase library after selection (25µg/mL Ampicillin) and controls. Right: The library sample observed after single and double digestion with restriction enzymes. BamHI is a single cutter restriction enzyme of the backbone plasmid while *XhoI* and *NheI* are the restriction enzymes flanking the insert used during ligation.

The plasmid after selection migrates predominantly as a single band, both in the circular and the linearized forms (the linearization was carried out with BamHI, a restriction enzyme able to digest a single site on the backbone).
The weight discrepancy between the linearized and the non-linearized circular plasmid extracts is due to the supercoiled topology of the circular plasmid extract. The secondary band in the circular sample could either be a relaxed (also known as open-circular) form or the plasmid or a ligation byproduct of higher molecular weight that was somehow able to produce a functional lactamase. Upon linearization, a single band of the expected molecular weight can be seen, with some extremely faint traces of accessory bands of different weights.
A double digestion reaction of the plasmid extract with the restriction enzymes flanking the insert recovered the original backbone and insert fragments.
This procedure is thus able to ligate efficiently the mutagenic insert in our backbone plasmid (pUC19a, see **chapter 2.1.1**) to generate the libraries that will be used for screening. Some byproducts of the ligation process were observed in the final library before the screening but, by forcing an enrichment in non-functional backbone plasmid concatemers, the library after selection predominantly displays only a single product corresponding to our desired ligation outcome.

2.2.6 Library transformation and antibiotic concentration

I chose to transform the mutagenized library by electroporation in ElectroMAX DH5α-E Competent Cells (ThermoFisher). The unmodified pUC19 plasmid is normally used as a control for transformation efficiency with these cells and shows an exceptional efficiency of up to 10^{10} transformants/µg. This is a best-case scenario and typically other plasmids show a lower efficiency, around 10^7 transformants/µg. The modified pUC19a plasmid showed a transformation efficiency comparable to the original pUC19, and produced a high number of colonies in media with a wide range of ampicillin concentration (from 10 to 1000 µg/mL).
The efficiency of the library transformation is additionally dependent on the ligation efficiency, on the mutation dosage and on the selective pressure of the medium.
From a few pilot experiments no bacteria survived among the cell transformed with a mutated plasmid library when grown in the medium with 1000 µg/mL ampicillin concentration and a there was a high mortality rate for the cells grown in 100 µg/mL ampicillin (that is the normal dosage for most molecular biology applications), a fact that is indicative of a very tight selection. This condition is not an ideal condition for this project since the harsh environmental constraints limit the availability of semi-neutral substitutions and narrow the mutational space that can be explored. I then considered ampicillin concentrations below the standard 100 µg/mL laboratory practice. I tested the bacterial growth of the library-transformed cells in media supplemented with 10, 25 and 50 µg/mL ampicillin. This range was significantly enriched in colonies able to survive the selection when compared to higher concentrations, and the number of survivors was directly dependent on the amount of antibiotic present in the medium.
This phenomenon is likely to be a consequence of the perturbation in the active site of the lactamase caused by the mutations, changing the catalytic speed and affinity and damaging the ability of the bacterium that express it to cope with an increasing amount of antibiotic.
This hypothesis was verified when the sham beta lactamase library containing only the wild type sequence was exposed to different concentrations of antibiotics and produced a very similar number or colonies independently to the concentration used.
The 10 µg/mL ampicillin condition was unable to consistently hold the stringency of the selection and sometimes few colonies could be observed even in the negative control where the plasmid lacked the beta lactamase gene.
I sequenced the plasmids extracted from few colonies grown in the 25 and 50 µg/mL ampicillin media, expecting to find a significant difference in the mutation rates between the groups. I reasoned that the more mutations are carried by one protein, the more likely the incorporation of a harmful substitution would become. However, the overall effect was rather bland, and, as

shown in **Table 2.2.2**, the two conditions featured very similar nucleotidic and peptidic mutation tallies.

Sample AMP25µg/mL	*DNA mutations (bp)*	*Protein mutations (aa)*
bla_25A	9	6
bla_25B	8	5
bla_25C	6	2
bla_25D	6	5
bla_25E	9	6 (1 STOP)
mean	*7.25*	*4.5*

Sample AMP50µg/mL	*DNA mutations (bp)*	*Protein mutations (aa)*
bla_50A	12	8
bla_50D	7	3
bla_50E	7	4
bla_50F	6	4
bla_50G	8	3
bla_50H	13	7
bla_50I	6	4
bla_50J	11	6
mean	*8.75*	*4.88*

Table 2.2.2 Error prone PCR induced beta lactamase mutation rate observed after selection with different concentrations of antibiotics.
Differences in the nucleic and amino acid sequence observed in the beta lactamase of a small sample of colonies (A-J) that survived the selection with 25 or 50µg/mL Ampicillin. The bacterial cells were transformed with a library of beta lactamases obtained after 20 cycles of error prone PCR. The samples containing early stop codons (greyed out in the table) were not used to calculate the means.

The logical conclusion was that I was seriously underestimating the presence of nearly-neutral substitutions. Both antibiotic conditions came from the same library and transformation, guaranteeing that before the selection process, the number of mutations would surely be identically distributed. I also expected the antibiotic to act as a strong selective pressure on the number of mutations sending the cells carrying the most altered enzymes to the chopping block. The data instead revealed a different scenario, with the two antibiotic conditions showing very similar substitution counts. It is possible to conclude that the expected dependency between the chance of survival and the harmful mutational load, if present, was masked by an underlying bigger distribution of neutral mutations with very little or no effect on the catalytic activity of the protein.
In **chapter 2.3.6** I will provide few cases of heavily mutagenized libraries where instead this dependency could be seen, possibly thanks to the sheer number of mutations involved.

2.2.7 Selection medium

Classical directed evolution performs the selection process in solid media and the results are usually limited to a few tens of thousands of colonies directly proportional to the number of petri dishes employed. Cultures grown on solid media are not easily scalable and the biomass they produce is quite limited. On the other hand, liquid media cultures are easily scalable and produce a lot of biomass but fail to preserve the library complexity and distribution. In liquid media fast growing phenotypes are not constrained and thus tend to dominate the culture while rare variants are prone to disappear.

This is a problem because neither a solid nor a liquid medium is suitable to generate a lot of diversity in the surviving bacteria. Deep sequencing is able to sequence millions of reads and it is in our interest to employ all these reads to collect as many different sequences as possible, as this would increase the diversity of the data and increase the precision on the subsequent analysis.

This is a very well-known paradox that can also be found in the field of antibody libraries, where the library complexity is related to the screening capability. Drawing from the solutions proposed for solving the problem in those high complexity libraries, I bypassed the issue by encapsulating the CFUs (colony-forming units) able to survive the selection in a matrix of a semisolid medium which allows local growth but prevents diffusion.

SeaPrep Agarose is a “soft agarose” product with ultralow gelling and melting temperature properties that is suitable for library cloning. When used in suitable concentrations, it can hold a few thousand bacterial colonies per litre of gelatinous media.

It is possible to control the number of transformants of the bacterial colonies that will originate the plasmid library after extraction by doing a few small-scale transformations and tuning a scale up procedure. The rule of thumb I chose when planning the selection was aiming to obtain around 10^5-10^6 CFU per bacterial growth. This in turn would match the depth of sequencing of the NGS platform and limit the number of duplicates in the reads.

During the course of the experiment 12 libraries were generated, each corresponding to a generation of mutagenesis. In **Figure 2.2.10** it is shown the colony counts I obtained for the 12 generations. When the transformants count did not match the expected value, the growth was discarded and I repeated the transformation increasing either the bacterial mass or the ligation product.

After colonial growth (**Figure 2.2.11**), the bacteria were collected and the plasmid library was extracted from the biomass. When a small linearized aliquot of the purified plasmid library was run on agarose gel it showed a single band at the expected molecular weight, in clear contrast to the several bands observed after ligation (**Figure 2.2.9**).

The first, fifth and twelfth generation libraries were sent to the Arizona Genomic Institute (AGI) to be sequenced on a Pacific Bioscience Sequel sequencing system. AGI performed quality control with Qubit, double digestion with XhoI and NheI restriction enzymes, band purification by electrophoresis, SMRTbell library creation and SMRT sequencing.

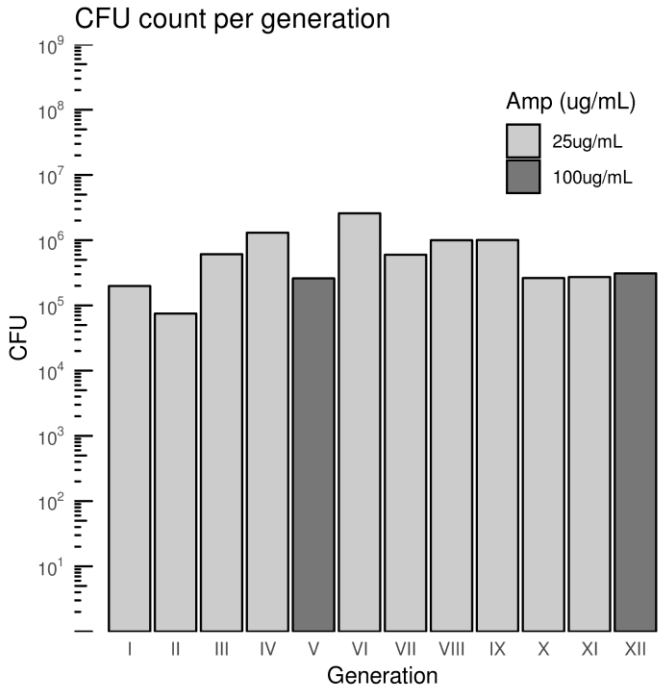


Figure 2.2.10 Number of transformants obtained each generation. The number of transformants is obtained by collecting and plating on solid medium a small 1mL aliquot of the 1L bacterial culture before the gelification of the growth medium. The number of colonies present in the semisolid medium can be calculated from the CFUs observed in the aliquot sample. Due to technical problems it was not possible to obtain a precise colony count for the seventh (VII) generation and the value shown is only a rough estimation.

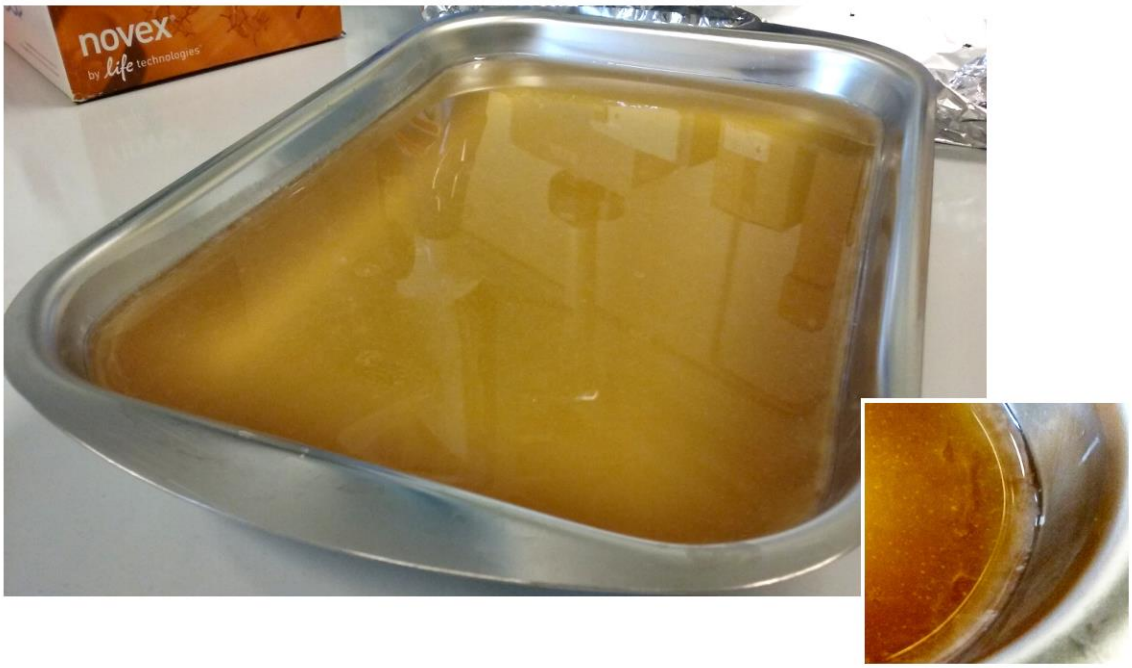


Figure 2.2.11 Bacterial culture setup and colonial growth on SeaPrep ultralow gelling soft agarose. Metallic tray containing 1L of bacterial culture grown in semisolid media. Small box on the right: high contrast image that reveal the colonial growth (the small white dots that can be seen in the growth medium).

2.3 Molecular Evolution

2.3.1 First generation of molecular evolution (GEN1)

From the sequencing of the first generation library (GEN1) I obtained 10.3 Gbp (Giga base pairs) total, divided in 622K reads with and average read length N50 of 1250bp. PacBio SMRT sequencing read circularly the DNA fragment several times (see introduction **chapter 1.3.5**) on both strands producing a very long low quality read. By splitting each DNA sequence in subread of repeated elements it is possible to align them to each other and build an intramolecular consensus.

NOTE: There is an issue in this step for some kind of algorithms that requires some further discussion. I originally considered sequencing, in addition to the library of functional variants, also the ligation product library that was used in transformation that is the library before the screening of functional variants. This extra sequencing would be necessary to employ an analysis based on fitness (see **chapter 1.1.3**), a staple technique used in evolution experiments, instead of relying on Direct Coupling Analysis (see **chapter 1.1.2**). This kind of pre-selection libraries however would have an issue during consensus building that needed to be solved during library construction: error prone PCR generates many non-canonical pairings of the nucleobases between the strands of the amplicon and require an extra PCR step before ligation to mitigate the issue (see the discussion on the error prone PCR protocol in **chapter 2.2.4**). Since both strands are sequenced several times to generate the circular consensus, anytime a strand difference is found by the algorithm, the resulting consensus would be very low quality since it would be an average of two different groups of subreads, one for each strand, each carrying a different version of the nucleotide. Segregating the strands enriched of non-canonical pairing with a regular PCR reaction, as I did for these samples, solves the problem. An alternative solution to the issue is to build a strand-specific circular average of the reads, generating a set of two consensus for each read, one for each strand. This however would raise the additional question if both strands should be considered equally in the analysis or not, thus the computational approach is subpar to the molecular one.

I built the intramolecular consensus of our first generation library with the CCS2 function of the official SMRTlink suite released by Pacific Bioscience.
This process additionally filters out bad quality reads as a function of two parameters that can be tweaked to influence whether a consensus read is included or not in the final output. Those two parameters are the minimal quality of the consensus (Phred score) and the number of repetitions that generate the consensus (called number of passes or simply “passes” as a

reference for the number of passes done by the polymerase on the intended target during SMRT sequencing).
In the first analysis the library was processed with the very lax default values requiring the consensus to have at least 3 passes and a minimum quality of Phred 10. This produced a rough estimate of the state of the sequences (**Table 2.3.1**).

CCS2 statistics on TEM-1 beta lactamase GEN1 library (passes > 3, Phred quality > 10)

Movie	*CCS reads*	*Number of CCS bases*	*CCS Read Length (mean)*	*CCS Read Score (mean)*	*Number of Passes (mean)*
m54138_171017_212252	428,267	387,192,271	904	0.995	22

Table 2.3.1 Statistics of the first generation library generated by CCS2 during consensus building.

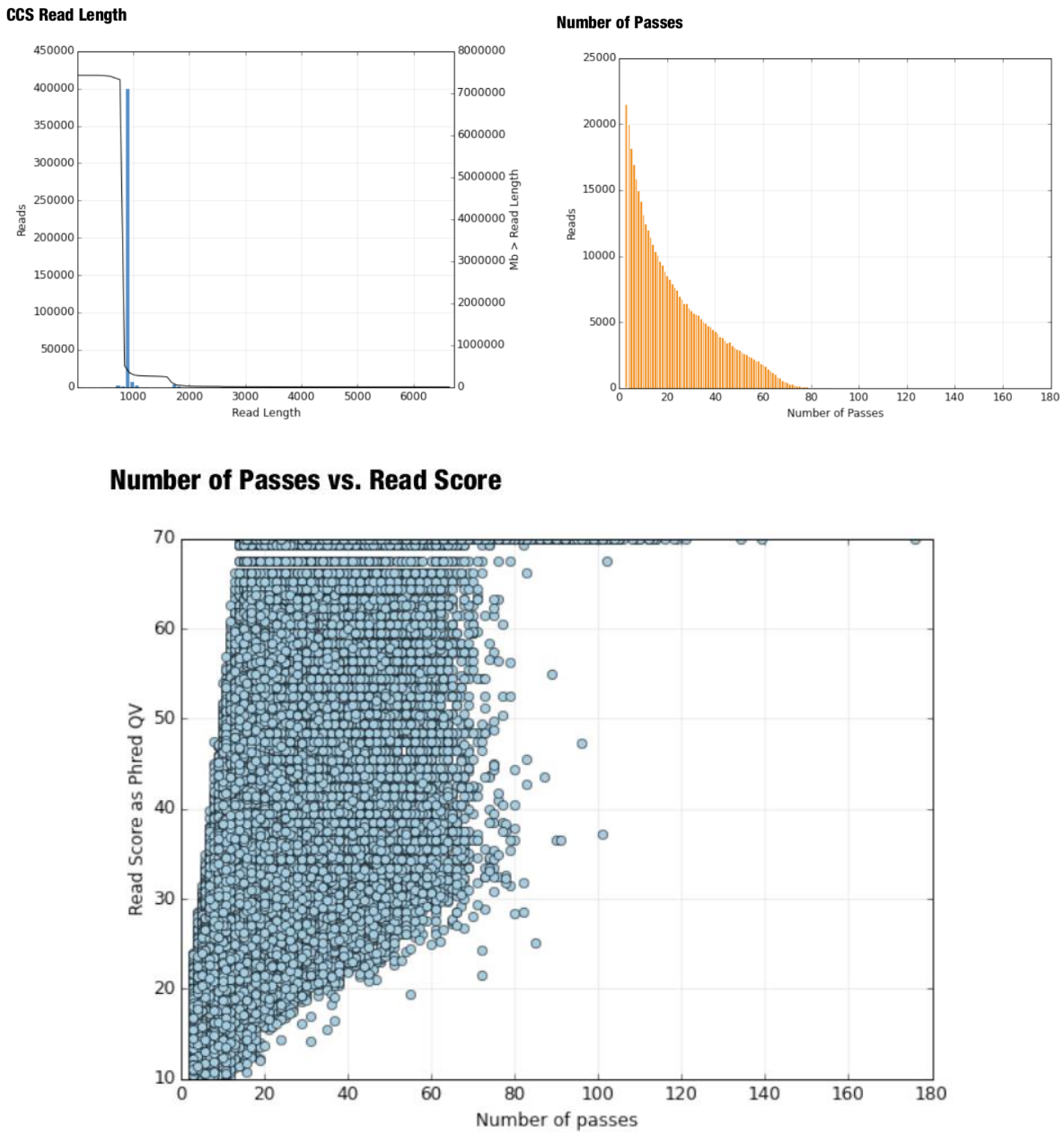


Figure 2.3.1 Graphical reports of the first generation library generated by CCS2 during consensus building.
A) Histogram of the lengths of the consensus reads. B) Histogram of the number of times the polymerase passed and read the sequence (i.e. the number of repetitions of the sequence present in the raw read). C) Scatterplot of the number of polymerase passes and read score associated to each raw read.

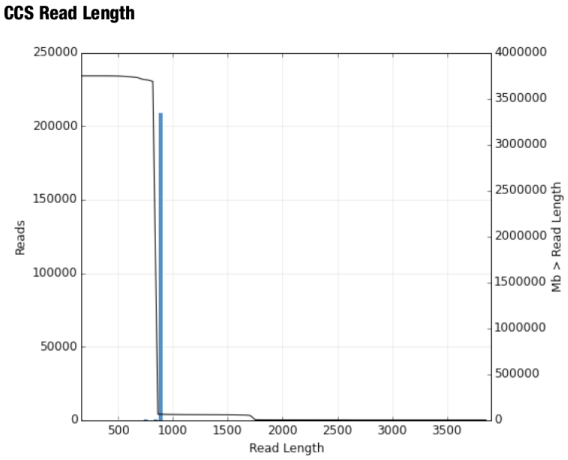
The sequences appeared to represent mostly the sequencing target, with the majority of them distributed around the intended length (~900bp) with a small cluster centred on double this length (**Figure 2.3.1A**). These double-length reads are likely a repetition of the intended

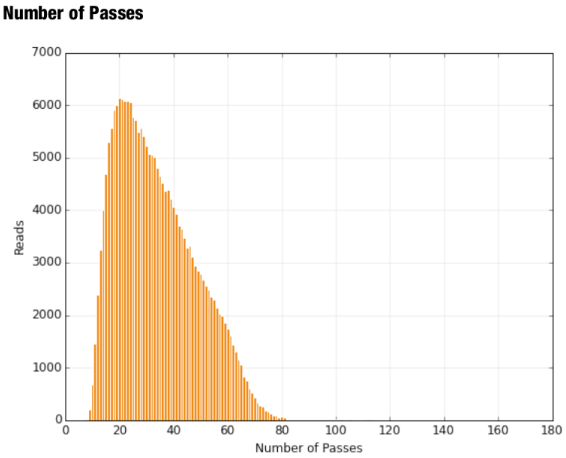
sequence due to sequencing or biological issues. The number of reads passing the quality criteria were around 420K but a lot of them feature a low number of passes (**Figure 2.3.1B**) and thus an overall low Phred quality score (**Figure 2.3.1C**).
I repeated the analysis increasing the stringency on the quality criteria to minimum quality of Phred 40 (**Table 2.3.2**).

CCS2 statistics on TEM-1 beta lactamase GEN1 library (passes > 3, Phred quality > 40)

Movie	*CCS reads*	*Number of CCS bases*	*CCS Read Length (mean)*	*CCS Read Score (mean)*	*Number of Passes (mean)*
m54138_171017_212252	216,130	192,280,768	889	1	34

Table 2.3.2 Statistics of the first generation library generated by CCS2 during consensus building using the additional parameter of Phred quality > 40.





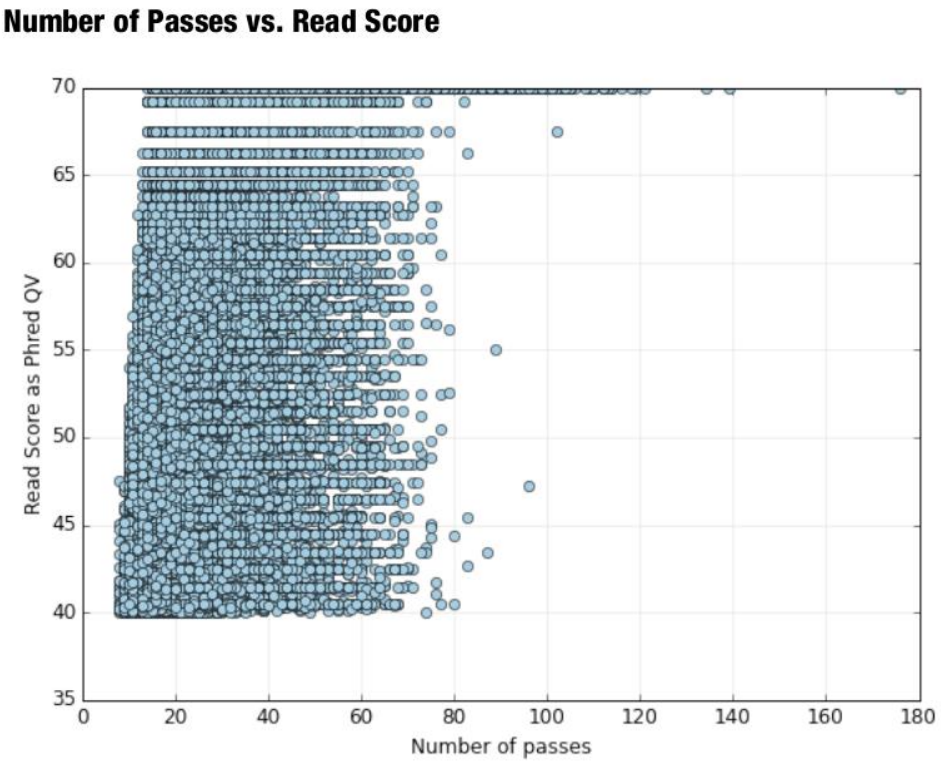


Figure 2.3.2 Graphical reports of the first generation library generated by CCS2 during consensus building using the additional parameter of Phred quality > 40
A) Histogram of the lengths of the consensus reads. B) Histogram of the number of times the polymerase passed and read the sequence (i.e. the number of repetitions of the sequence present in the raw read). C) Scatterplot of the number of polymerase passes and read score associated to each raw read.

In comparison to the previous analysis (**Table 2.3.1**) the usage of these new parameters resulted in a sharp drop in the number of passing sequences in association to a drastic increase in the mean quality (**Table 2.3.2**).
The sequences length distribution still shows a main peak around the intended length and a collateral small peak at double the size, but the data are significantly less dispersed around these peaks compared to the previous analysis (**Figure 2.3.1A**). This is caused by the removal of sequences where the basecaller of the Sequencing platform or the algorithm that creates the consensus produced some poorly supported insertions and deletions.
The distribution of the number of polymerase passes with the new threshold forms a nice bell-shaped curve (**Figure 2.3.2B**) and, in addition, practically no sequence has less than 10 passes (**Figure 2.3.2BC**).
The consensus reads were then mapped to a fragment of the reference wild type plasmid to obtain the position of the starting methionine of the lactamase variant. This step also allowed

the removal of the unrelated sequences and put all the sequences in the same strand orientation.
The reads were then in silico translated producing peptidic products of different lengths (**Figure 2.3.3**).

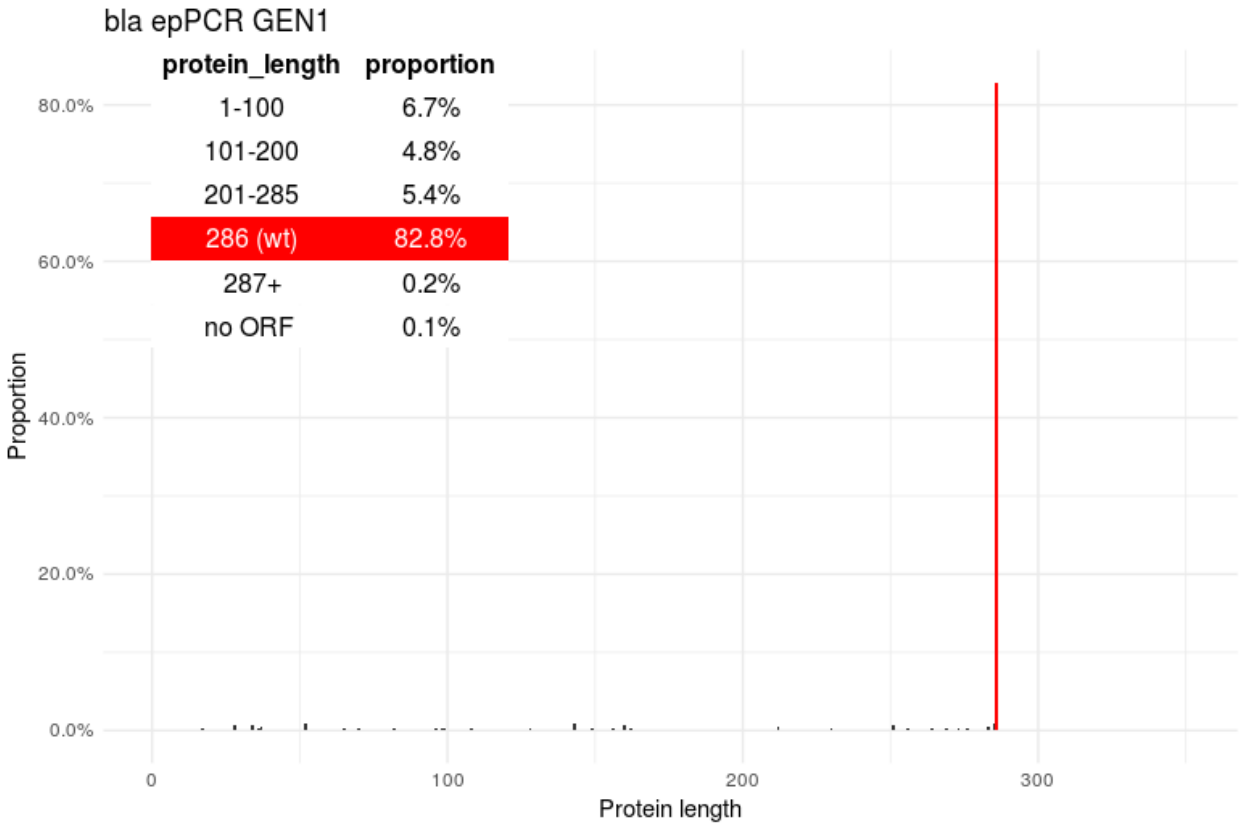


Figure 2.3.3 Proteins length distribution of the translated consensuses of the GEN1 library.

The vast majority of the sequences were 286 amino acids long, the same as the wild type TEM-1 progenitor lactamase. Longer sequences were barely observed, while shorter sequences were present, roughly uniformly distributed between 1 and the wild type protein length. The

longer variants have a good probability to still maintain the protein function while the shorter variants are likely non-functional and possibly a carryover of the selection process.
Aligning the proteins to the wild type lactamase allows us to calculate the number of mutations present in each protein and their distribution (**Figure 2.3.4**).

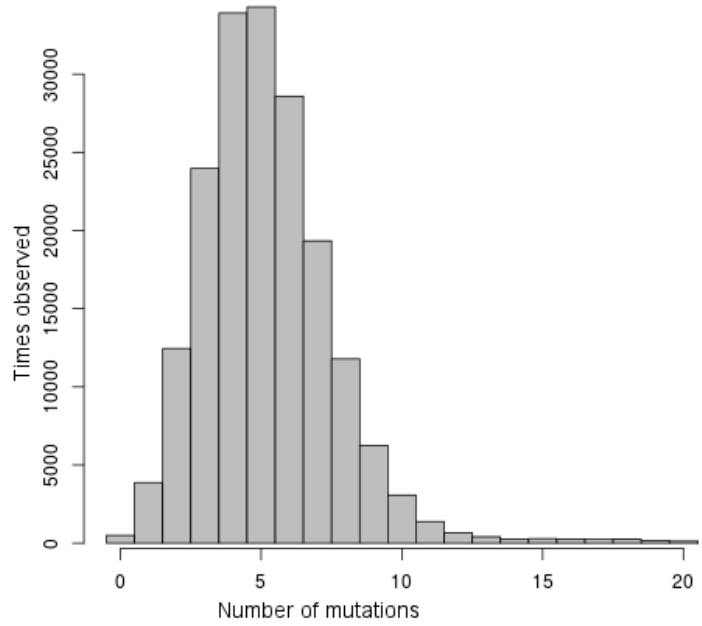


Figure 2.3.4 Number of mutations observed in the peptidic sequence of beta lactamase in the first generation library

The observed bell-shaped distribution is very similar to the expected Poissonian distribution and, as predicted from the previous experiments (**Table 2.2.2**), I found around five mutations per sequence (mean 5.11, median 5, variance 4.60).
Calculating the mutation propensity per position of the lactamase it is possible to observe that the rate of mutation is influenced by the position and it forms a pattern. The mutations of the library are not uniformly distributed along the sequence but instead many positions are mutation prone while others are much more conserved. Some entire regions, like the positions

corresponding to the second or sixth helix of the lactamase are far more conserved compared to other region like the leader peptide or the last helix.
I will discuss specifically these conserved regions in **chapter 2.4.2** of the thesis.

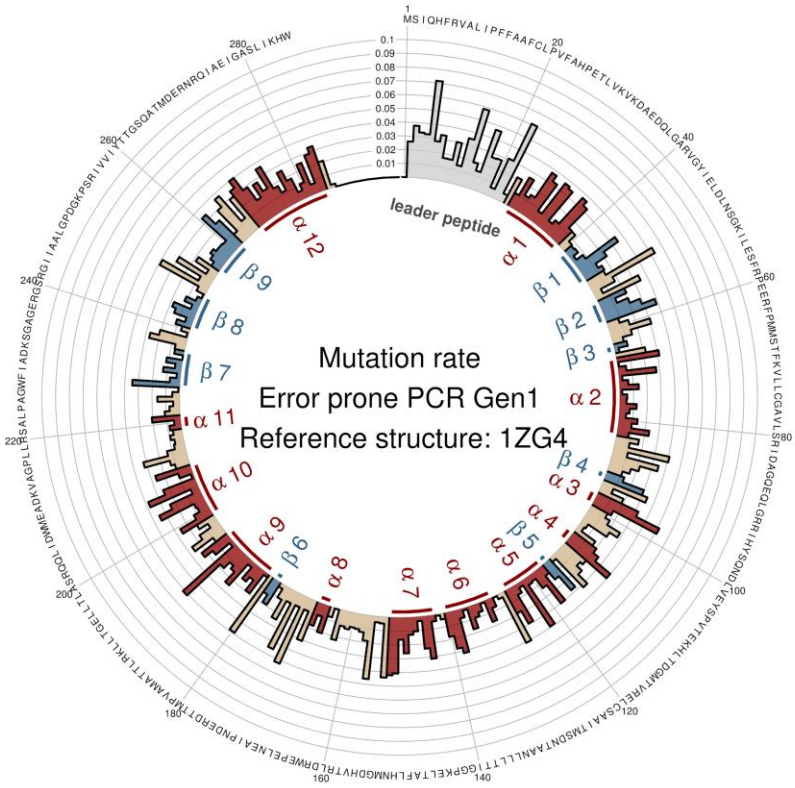


Figure 2.3.5 Mutation rate per residue position observed in the first generation library. The colours and annotations follow the secondary structure classification present in the PDB structure 1ZG4 (red: alpha helices, blue: beta strands, tan: coils). The leader peptide sequence (light grey) is missing in the structure.

2.3.2 Library complexity

It is always difficult to estimate with precision the complexity of genomic libraries (i.e. the number of different sequences in the collection), and these mutational libraries are not much different.
I will analyse only the complexity of this first mutational library, since several different factors influence the parameter and made the estimation unreliable in later time points.
Complexity (also called diversity) is an important parameter for evolution libraries since it is proportional to the overall mutational space that has been explored and is independent to the other quality parameter that is the number of mutations. The complexity indicates to what extent

I was able to variate the original sequence while the number of mutations is related to how drastic the modification of the protein will be. To have a good coverage of the conformational space around the structure, both a good number of mutations and a high complexity library are required.
The most important estimator of the library complexity is the number of transformants in the bacterial growth. Transformed bacteria usually are considered a monoclonal source of the transformed plasmids since the transformation event is rather rare and the fraction of polyclonality in the culture is usually so low that can be ignored. As a consequence, the number of CFUs in the culture can be considered the theoretical maximum of the complexity of the plasmid library, in the best-case scenario where each and every cell in the culture had taken a single unique plasmid. The transformant counts of the semisolid cultures used in this project were estimated consistently in the range of some hundred thousand to few million CFUs (**Figure 2.2.10**), and in this first generation I counted 198K transformants.
After sequencing the diversity can be estimated from the sequence data and the distribution of repeated elements. This is done by fitting with a truncated negative binomial distribution the number of groups composed by *n* identical sequences (cardinality), *n* being the times the sequence is found in the collection. This way the first element is the number of unique sequences, i.e. the sequences found only one time in the sequencing data (also called singletons) while the second element is the number of sequences found exactly two times (also called "twin pairs" or doubletons) etcetera (Bunge & Fitzpatrick, 1993). Through the fit with the negative binomial distribution it is possible to estimate the missing "zero element" group or the number of different sequences found exactly zero times (so not found) in our sequencing data. The sum of the number of different sequences, including the ones in the "unobserved" group that was just calculated, is the complexity of the library.
Due to the mutagenic nature of the library (that have an extremely high complexity before selection), it is very unlikely to observe identical sequences in the collection of plasmids that were incorporated in the cells. It is generally a good assumption to believe that all the doubletons observed during the sequencing would be a consequence of the random sampling of the plasmid library after the selection and expansion of the colonies in the grow broth. However, when comparing the sequencing data with a series of simulated samplings from libraries with different complexities, the sequencing appeared to fit better to libraries around one

million elements (**Figure 2.3.6**), much more than the estimated maximum complexity of the library calculated from the transformation efficiency (198K sequences).
Most likely the estimated complexity from the bacterial culture, due to an unknown factor, is far from the real diversity of the library, and must be considered an underestimation.
The same results were obtained from the other sequenced library (5th and 12th generation).

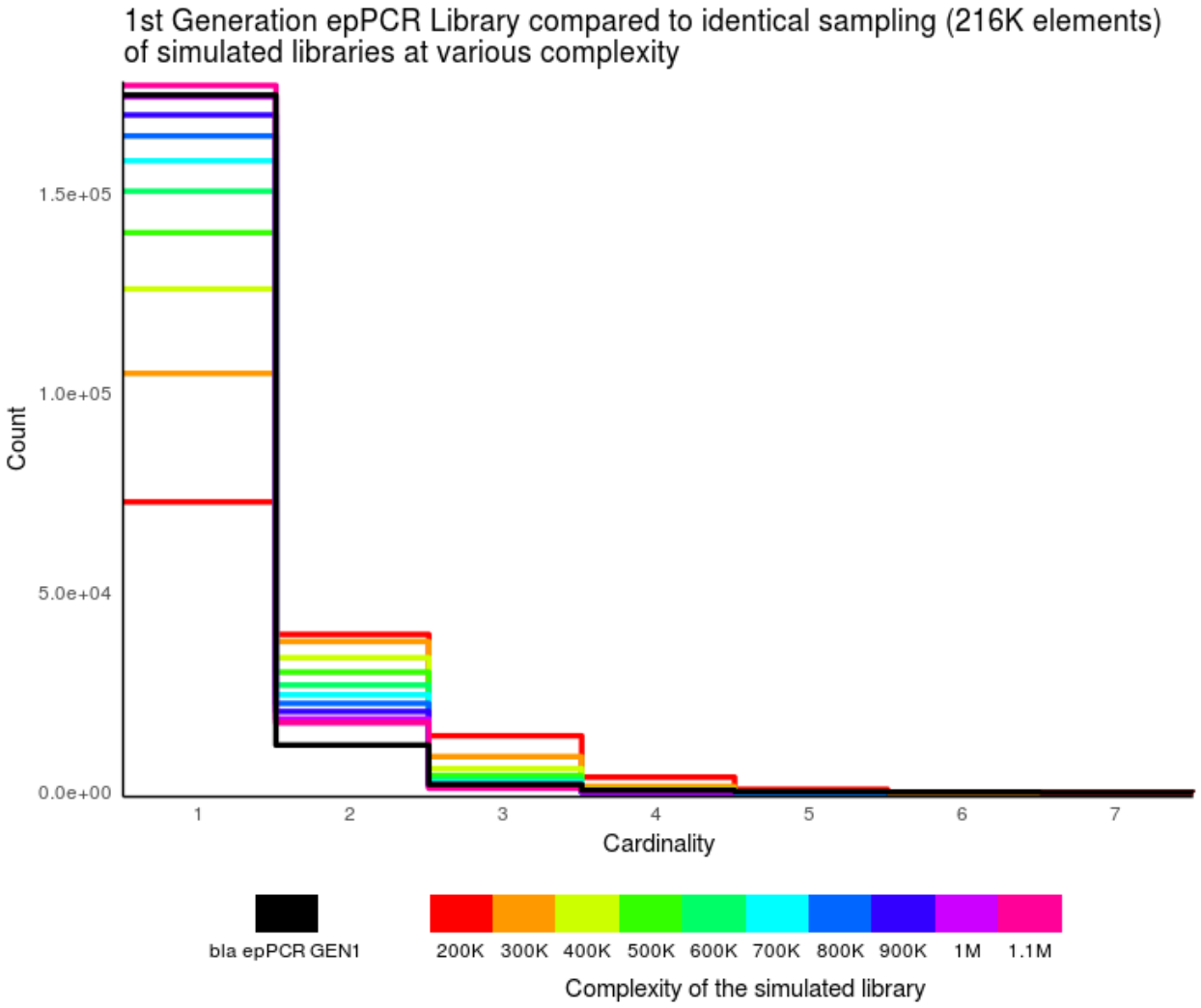


Figure 2.3.6 Comparison of the complexity observed in the first generation library to random sampling of simulated libraries of various complexity.
The simulations were random sampling with repetition of 216K elements, the same as the sequenced mutagenic library, from libraries that span a complexity range from 200K elements to 1.1 million elements.

I wondered if the type of library, being mutagenic and building errors by accumulation on previously mutated sequences, could be the reason for the discrepancy observed.
One of the previous assumptions was that the library is a randomly heavy-mutagenized collection of sequences, thus the probability of incorporating identical sequences in the transformed bacteria is infinitesimal. However, even if the process is defined as error-prone, it

still carries a high chance to produce a perfect copy. The earlier a mutation is generated, the higher it is represented in the collection, while errors that appear in the last few cycles have a good chance to be unique.
To prove this hypothesis, I calculated the correlation between the number of errors of a sequence and the number of times the sequence was observed (the cardinality). If it is true that the library is enriched of sequences carrying progenitor mutations, a high cardinality (a sequence seen multiple times) would be characterized by a low mutation frequency and vice versa, thus the two variables would be inversely correlated.
To avoid correlation from overrepresented outliers (like the wild type lactamase that was found a few hundred times), I limited the analysis to sequences below a small arbitrary threshold of cardinality 10.
As expected, the variables are inversely correlated (**Table 2.3.3**) proving the enrichment in sequences generated in the early stage of PCR amplification.
Another interpretation of these data could be that I collected a higher number of sequences associated to a lower number of mutations because every mutation has a chance to alter the protein and makes it inactive. The more mutations a protein possesses, the less likely it will remain functional.
The number of possible different beta lactamase sequences with N mutations is approximately $(L_{prot} \times (n_{aa} - 1))^N$, that for five mutations equals to $(286 * 19)^5 = 4.7e18$ different combinations. Even if factors such as mutation propensity and genetic code redundancy altered the probability to observe these combinations, without a high degree of duplication, it is very unlikely that the sequence carrying the same identical 5 nucleotide mutations is transformed into a cell more than once in the one million transformants culture used. Moreover, the incorporation happens before the stage where any possible advantage in survivability of the sequence can take place.
This relation between the number of mutations and cardinality of a sequence in each dataset is progressively lost with the increase of the mutational load of the library in the various sequenced generations (**Figure 2.3.7**).
I concluded that the effect is shrouded by a higher variability in the initial number of mutations. In other words, it is easier to accumulate identical sequences in the library that will be transformed starting from a single sequence in the first generation rather than a collection of sequences that were used as input of the error prone PCR in the fifth and twelfth generations. Moreover, as the mutational load increases, the spread of the number of mutations present in the sequences increases as well (since it follows a Poissonian distribution). This means that trying to correlate the number of total mutations to anything is less effective because it is unclear if the sequence has mutated thrice from a sequence carrying few mutations or only once from a more heavily-mutated sequence of the previous generation. To control that this relationship is specific to the library type and not a spurious correlation between the variables, I tested the correlation on the UniProt dataset and found no significant relationship.

```
        Pearson's product-moment correlation

data:  number of mutations and cardinality, library bla epPCR GEN1
t = -49.533, df = 167000, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
```

```
 -0.1250523 -0.1155991
sample estimates:
      correlation
-0.1203284
```

Table 2.3.3 Pearson correlation test between the number of mutations in the sequence and the number of times the sequence was found in the collection (cardinality).
Non parametric tests (Spearman and Kendall rank correlations) gave similar results.

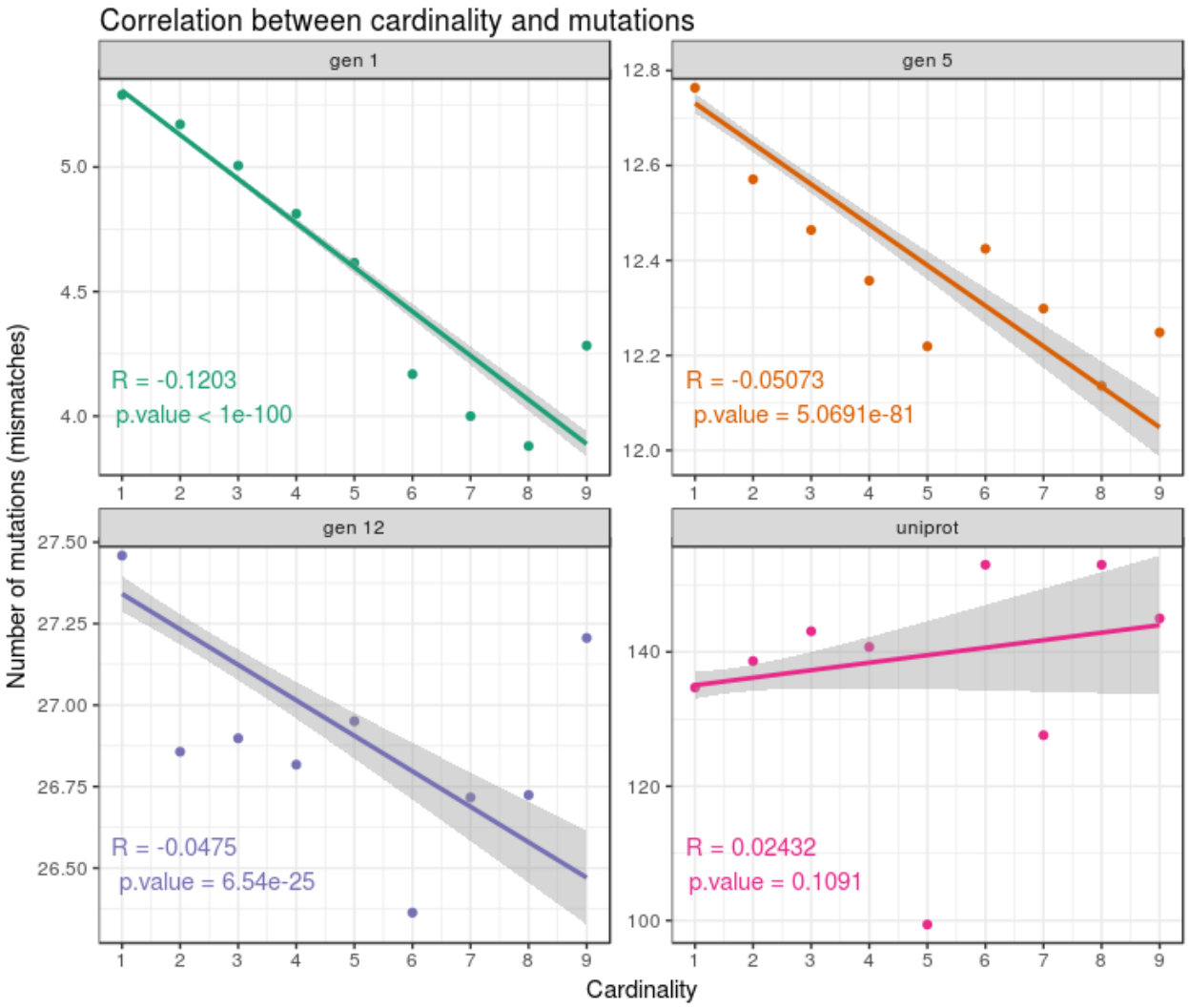


Figure 2.3.7 Mean number of mutations for each cardinality group in all the datasets.
The regression line with the confidence interval (shaded) is shown for each set. Pearson's R and the associated probability value can be found in the bottom left corner of each graph.

2.3.3 Direct coupling Analysis (GEN1)

From the 226K circular consensus reads obtained after sequencing I retained only the sequences that matched the reference ancestral sequence and, after in silico translation, the

protein whose length encompasses at least 80% of the wild type lactamase. This was done to reduce the impact of non-functional protein caused by the insurgence of an early stop codon. The remaining 177K sequences were aligned to the reference lactamase and used to build a multiple sequence alignment using mafft, keeping only the original 286 amino acid positions corresponding to the wild type enzyme.
To predict which of the residue pairs interact, I employed a custom implementation of Direct Coupling Analysis that apply to this MSA a pseudo-likelihood approximation (Balakrishnan et al., 2011) (see Materials and methods).
I selected the 286 residues pairs (equalling in number to the protein length, 0.72% of the total possible contacts) which showed the highest DCA score and were more than five residues apart in the MSA, and I compared them to the contact map of the reference structure (**Figure 2.3.8**).

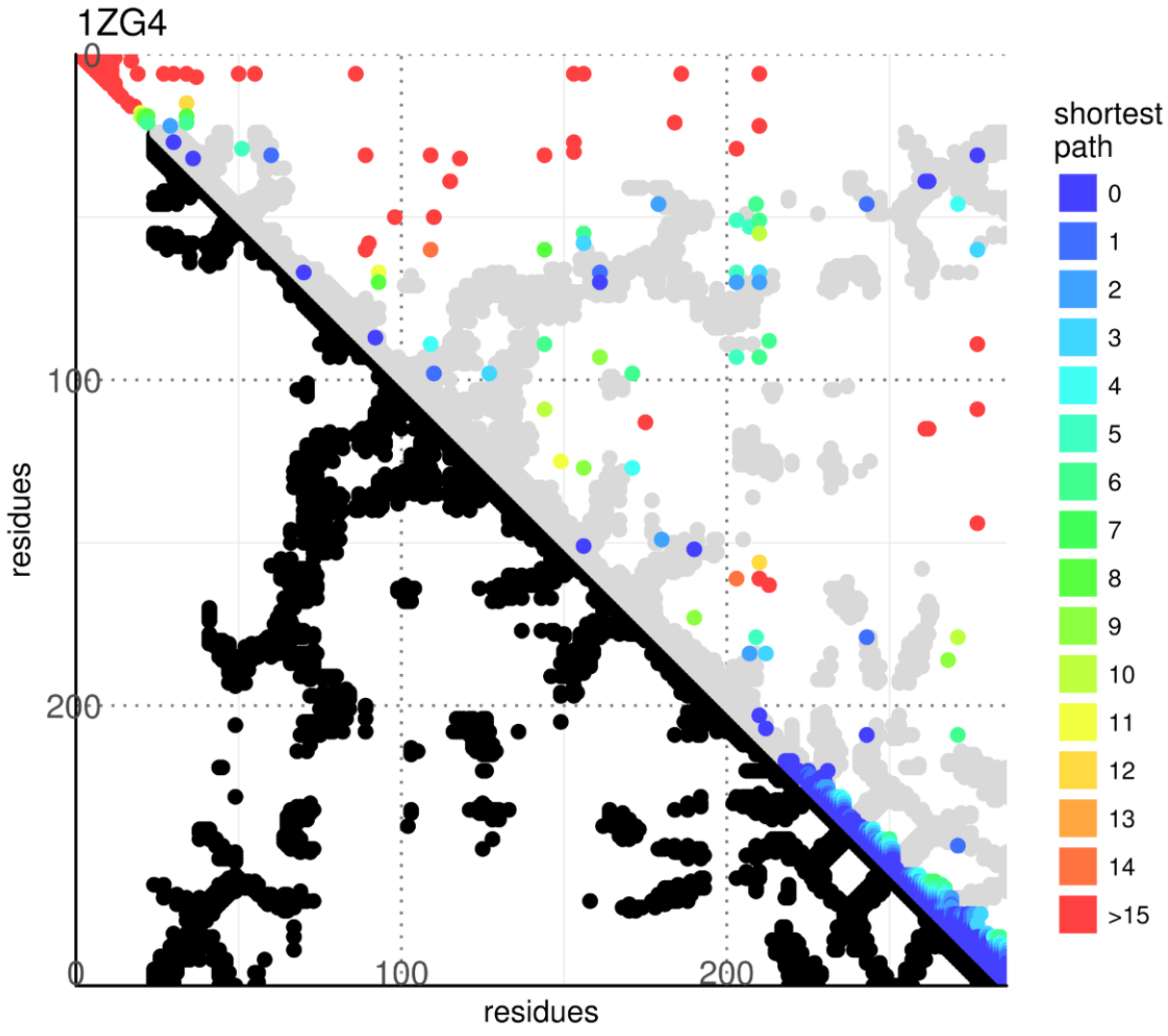


Figure 2.3.8 DCA plot of the first generation library.

DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the first generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

Even if the residues less than five residues apart were not included in the tally of the strongest 286 prediction pairs, I still displayed them on the graph to highlight a serious bias in the data. The problem lies in the C-terminal region of the protein, which displays a bizarre and clear-cut enrichment of contact density near the diagonal. After fiddling with the parameters, I noticed that the part of the C-terminal of the protein that presents this strange behaviour increases when I decreased the threshold of protein lengths kept in the analysis and vice versa (originally at least 80% of the wild type length). The solution to the enigma was actually simple: co-evolutionary traces are strongly biased by frameshift mutations.
The reason why shorter proteins are present in the first place can be either a single nucleotide mutation that generates an earlier stop codon in the open reading frame or a nucleotide insertion/deletion that generates a frameshift that almost always finds a termination codon after generating few improperly shifted amino acids.
An early termination signal is not particularly damaging to the analysis, since stops and gaps are ignored while we calculate the co-evolution signal, but the shifted amino acids that follow a frameshift are a problem and damage the analysis. A single nucleotide deletion, apart for the first initial amino acid, always produces a concerted change in the following protein sequence, a change that is identical every time this deletion appears. This logic identically applies to any insertion and every other type of frameshift. The result of this process is that in the sequence collection there are several sequences where some positions suddenly changed in a concerted manner which by definition generate a strong covariation signal.
It is interesting to notice that classical evolutionary data supplied to the algorithm do not have this problem, and thus it is a new issue brought by the mutagenic data. This is probably because frame shifted sequences do not carry a sufficient neutrality to the system to be retained throughout natural evolution, while in the small landscape generated by mutagenesis every protein that satisfies the selection criteria of activity will be part of the collection.
To bypass the issue, I decided to retain only the sequences that code for a protein identical in length to the wild type lactamase. This criterion is very constraining and far from optimal because it is seriously limiting the mutational landscape of the protein we will observe, but every other attempt I made to filter the data failed to remove the C-terminal diagonal density.

With this new criterion the density along the diagonal in the C-terminal region disappeared but nonetheless no significant evolutionary traces were found (**Figure 2.3.9**).

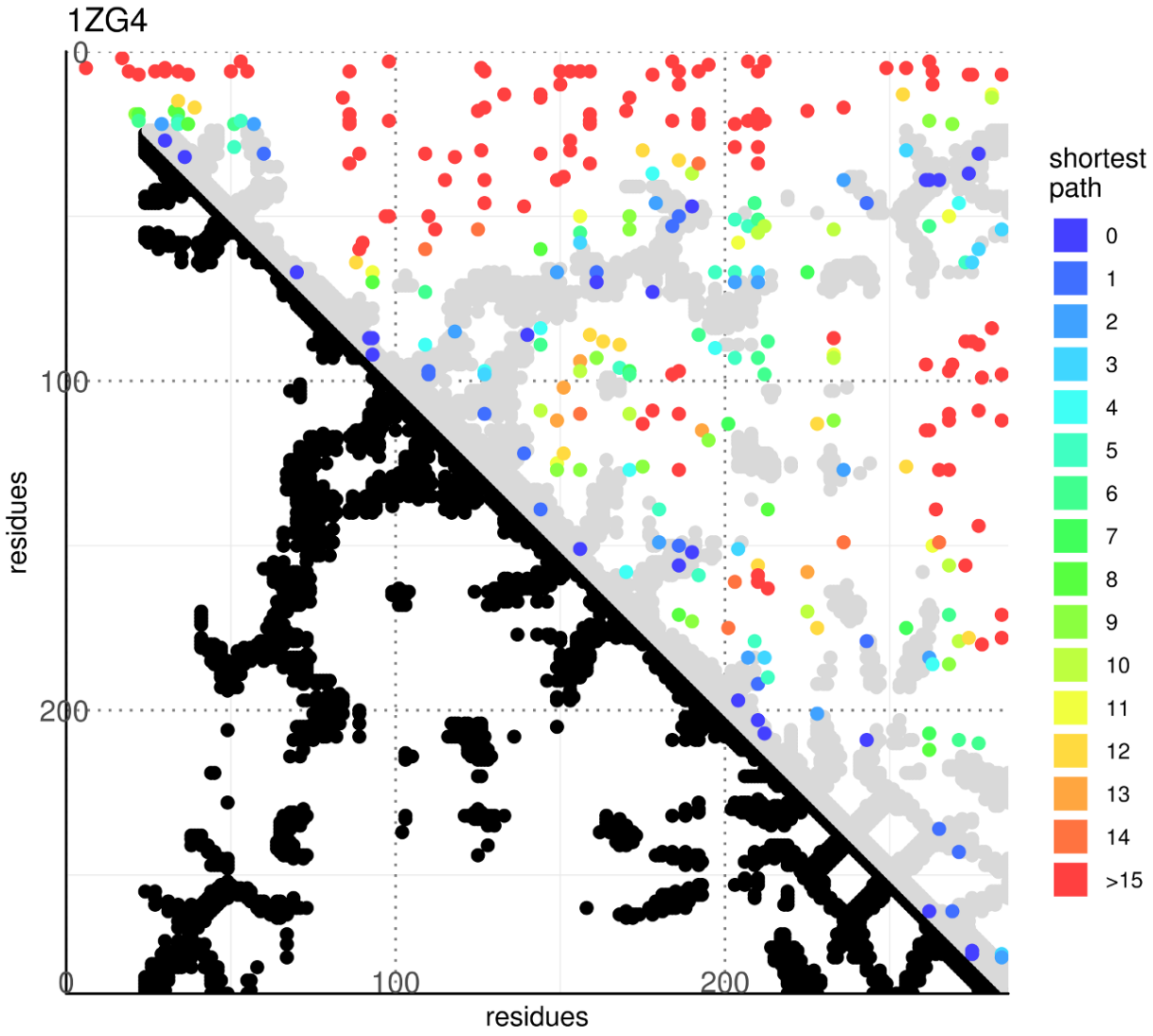


Figure 2.3.9 DCA plot of the first generation library after the removal of the sequences that could carry a frameshift.
DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the first generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph

the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

This generation featured five mutations per sequences, that is ~1.7% of the protein. These mutations were probably too few for our analysis so I decided to increase this number by repeating all the previous steps sequentially other four times, bringing the system from the first to the fifth generation of mutagenesis.

2.3.4 Fifth generation of molecular evolution (GEN5)

Like the first generation, the fifth generation ligation library was first subjected to a small scale preliminary transformation and selection to set up the conditions for the big scale transformation. This time the ligation was transformed into bacteria and selected in growth media with both the normal 25 mg/mL and an increased 100 mg/mL ampicillin concentration. From the surviving colonies, 10 were randomly selected and sequenced to calculate the mean number of mutations in the peptide chain (**Table 2.3.4**).

5TH GENERATION

Sample AMP25µg/mL	*DNA mutations (bp)*	*Protein mutations (aa)*	*Sample AMP100µg/mL*	*DNA mutations (bp)*	*Protein mutations (aa)*
blaV_25A	20	9	*blaV_100A*	19	13
blaV_25B	26	11	*blaV_100B*	33	15
blaV_25C	25	10	*blaV_100C*	26	13
blaV_25D	37	19	*blaV_100D*	26	15
blaV_25E	32	11	*blaV_100E*	26	13
blaV_25F	27	1	*blaV_100F*	20	10
blaV_25G	32	17	*blaV_100G*	25	13
blaV_25H	32	23	*blaV_100H*	31(1 Indel)	15(Frameshift)
blaV_25I	31	18	*blaV_100I*	40	18
blaV_25J	23	12			
mean	*28.5*	*14*	*mean*	*26.9*	*13.8*

Table 2.3.4 Mutations in the beta lactamase sequence observed after selection in the fifth generation library.
Mutations in the nucleic and amino acid sequence observed in the beta lactamase of a small sample of colonies (A-J) that survived the selection with 25 or 100µg/mL Ampicillin. The bacterial cells were transformed with a library of mutated beta lactamases. This library was obtained by applying 20 cycles of error prone PCR to the lactamase insert of the previous generation. The samples containing aberrant sequences (greyed out in the table) were not used to calculate the means.

In the first generation, I used the lowest antibiotic concentration because it showcased a mutation rate similar to the higher dose while the survival rate was greatly enhanced. In this

generation I did the opposite, with similar observed mutation rate I chose the highest antibiotic concentration to hinder the survival of the cells carrying variants of the lactamase coding for a protein with low activity (**Figure 2.3.10**, see result **chapter 2.2.6** for information on antibiotic concentration in the media).

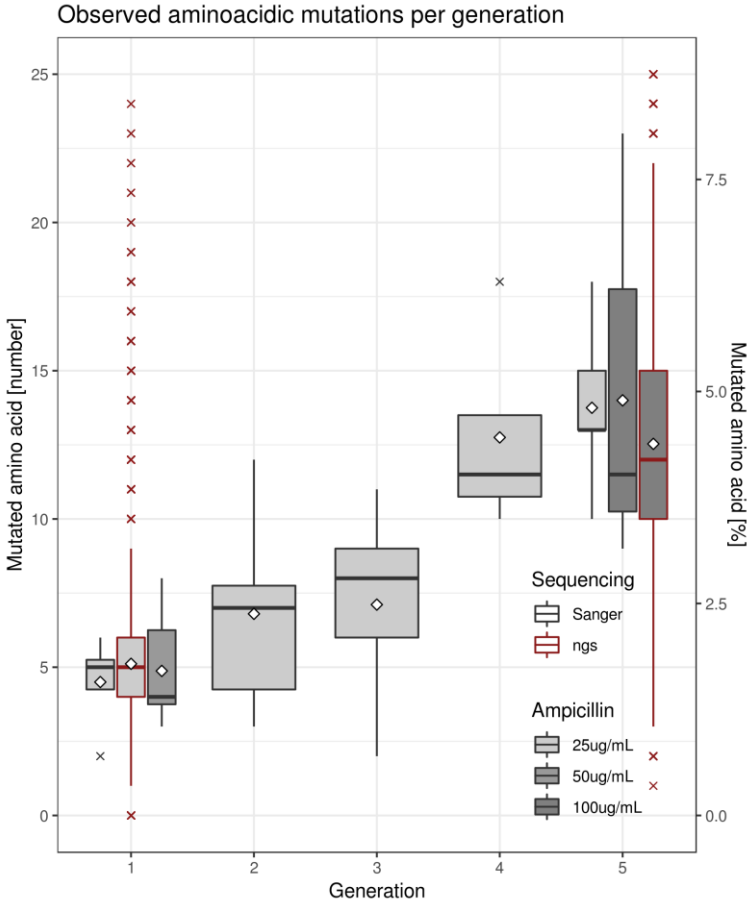


Figure 2.3.10 Mutations in the beta lactamase sequence observed during molecular evolution up to the fifth generation library.
Boxplot showing the number of amino acid mutations observed in the sample of clones sequenced after each generation (Sanger sequencing, black border) and after NGS (red border) up to the fifth generation. The white diamond dots indicate the mean**.**

From the sequencing of the fifth generation library (GEN5) I obtained 10.9Gb of raw sequences. I built the intramolecular consensus with the SMRTlink suite with similar parameters as for the first generation, requiring the consensus to have at least 10 passes and a minimum quality of Phred 40 (**Table 2.3.5**).

CCS2 statistics on TEM-1 beta lactamase GEN5 library (passes > 10, Phred quality > 40)

Movie	*CCS reads*	*Number of CCS bases*	*CCS Read Length (mean)*	*CCS Read Score (mean)*	*Number of Passes (mean)*
m54138_180418_211941	214,648	191,430,109	891	1	37

Table 2.3.5 Statistics of the fifth generation library generated by CCS2 during consensus building using the parameters: passes > 10, Phred quality > 40.

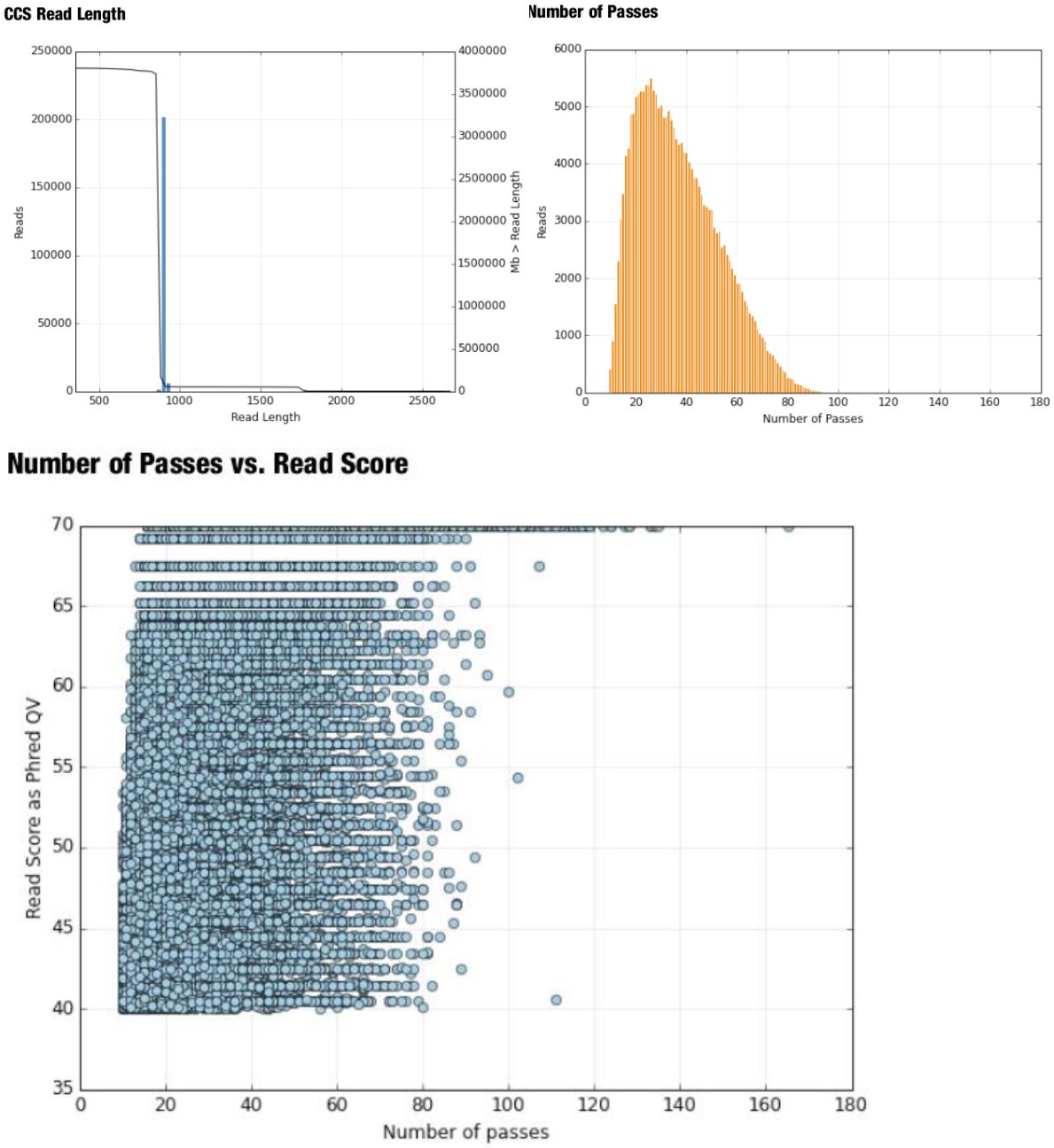


Figure 2.3.11 Graphical reports of the fifth generation library generated by CCS2 during consensus building using the parameters: passes > 10, Phred quality > 40.
A) Histogram of the lengths of the consensus reads. B) Histogram of the number of times the polymerase passed and read the sequence (i.e. the number of repetitions of the sequence present in the raw read). C) Scatterplot of the number of polymerase passes and read score associated to each raw read.

In general, the sequences obtained are very similar to those obtained from the first generation: both the number of passing sequences and the mean quality of this library match the ones obtained from the first generation library after consensus building (**Table 2.3.2**, **Table 2.3.5**); The sequences length distribution still shows a main peak around the intended length and a collateral small peak at double the size (**Figure 2.3.11A**); the distribution of the number of polymerase passes with the newest threshold of 10 passes form the same bell shaped curve observed in the first generation (**Figure 2.3.11BC**).
After removal of the unrelated sequences and after converting all the sequences to the same strand orientation (by mapping them to the same fragment of the reference wild type plasmid that was used in the first generation), the reads were in silico translated and produced peptidic products of different lengths as shown in **Figure 2.3.12**.

Figure 2.3.12 Proteins length distribution of the translated consensuses of the GEN5 library.

Again, the vast majority of the sequences were 286 amino acids long, as the wild type progenitor sequence. Longer sequences were scarce, while this time the group of shorter sequences was clearly enriched. This fact was puzzling but did not hinder the following analysis since the choice to keep only the protein with the same length as the wild type eliminates the problematic that arose in the previous generation caused by the presence of frame shifted sequences. This fact will become problematic and will be discussed in the last generation (**chapter 2.3.6**).

The alignment of these proteins to the wild type lactamase allowed us to calculate the number of mutations and their distribution (**Figure 2.3.13**).

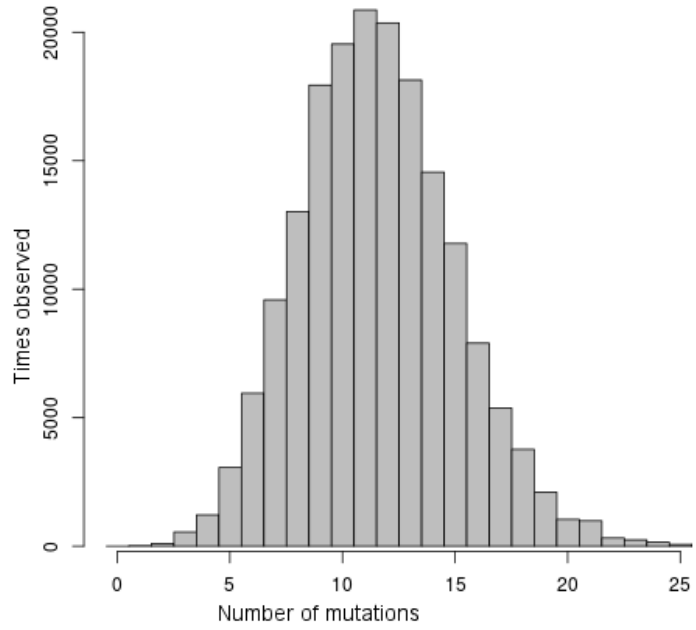


Figure 2.3.13 Number of mutations observed in the peptidic sequence of beta lactamase in the fifth generation library

The number of mutations per sequence found in this protein collection formed again a skewed bell-shaped distribution centred around 11 - 12 mutations per sequence (mean 12.55, median 12, variance 11.88).
The mutation propensity per residue of the lactamase was again not uniformly distributed along the sequence but instead clearly dependent on the position (**Figure 2.3.14**). In particular the position 180 was a clear outlier, with more than 30% of the sequences carrying a mutation. Analysing the sequences that carried a mutation in this position, I found that almost every sequence was switching the methionine 180 to a threonine. This mutation is well known in the literature and it will be discussed along with other overrepresented mutations in **chapter 2.4.2** of the thesis.

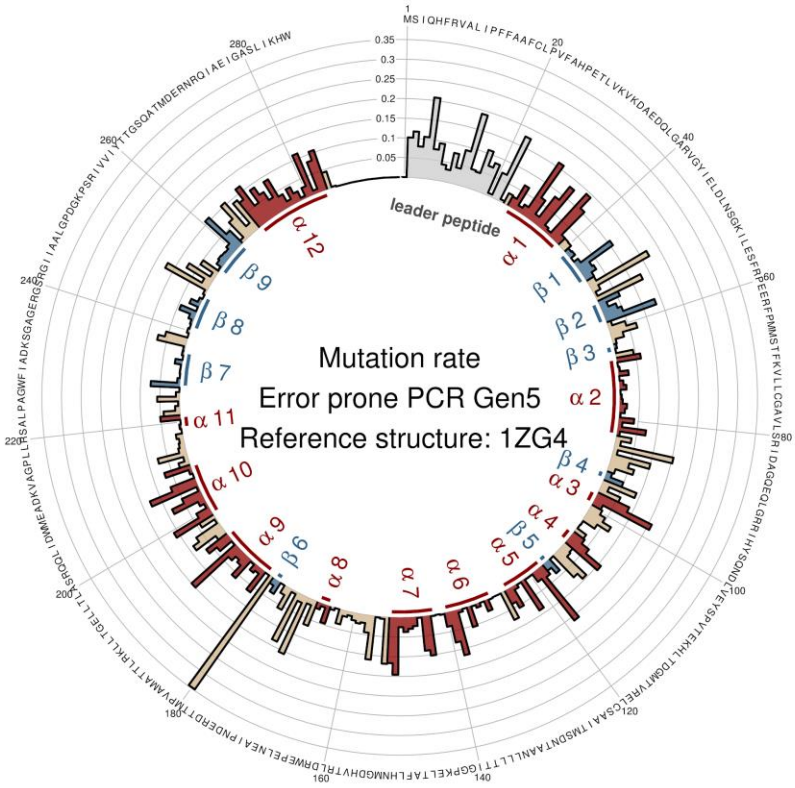


Figure 2.3.14 Mutation rate per residue position observed in the fifth generation library. The colours and annotations follow the secondary structure classification present in the PDB structure 1ZG4 (red: alpha helices, blue: beta strands, tan: coils). The leader peptide sequence (light grey) is missing in the structure.

2.3.5 Direct coupling Analysis (GEN5)

From the 214K circular reads obtained after consensus building, and after filtering out the unrelated sequences and the reads coding for a different length (compared to the wild type) lactamase, I obtained 178K sequences suitable for my purpose.
The sequences were aligned to the reference lactamase and used to build a multiple sequence alignment, keeping as usual only the original 286 amino acid positions of the wild type enzyme. I applied the Direct coupling Analysis to this multiple sequence alignment in the same way I did for the previous generation and compared the prediction to the contact map of the reference structure (**Figure 2.3.15**).

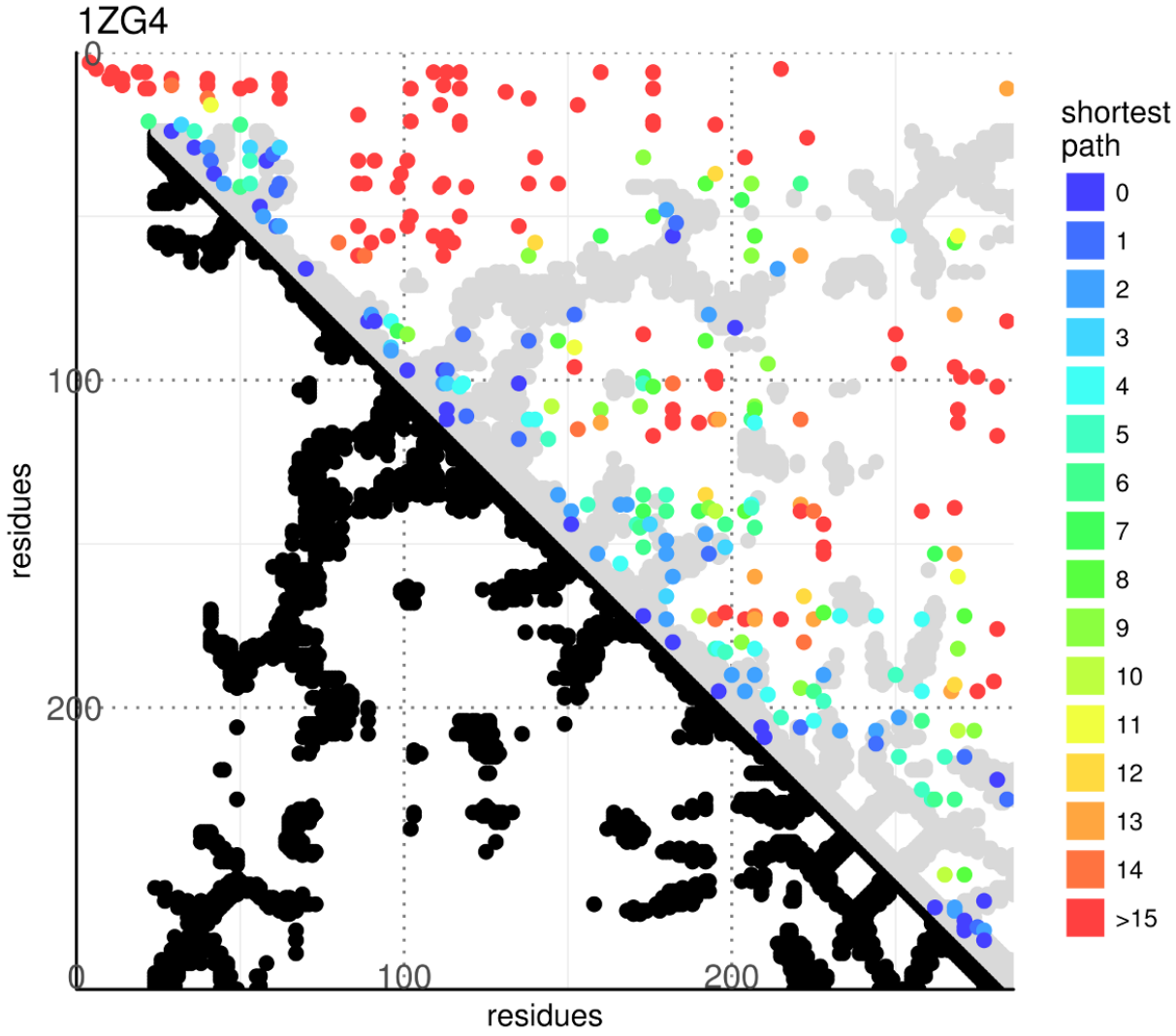


Figure 2.3.15 DCA plot of the fifth generation library.
DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the fifth generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph

the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

Similarly to what happened in the first generation, the prediction failed in creating proper clusters and seemed to be uncorrelated to the underlying structural data.
This time, however, the sequences had around 4% mutated residues compared to the wild type progenitor. With this mutation rate a phylogenetic correction can be applied during the analysis to balance the bias in the data. To remove this parental inheritance (the “phylogenetic” bias created during mutagenesis) and sampling biases in the MSA, each sequence contribution was reweighed during the analysis on the number of similar sequences present in the dataset. The number of similar sequences is intended as the number of sequences in the dataset that has at most x% of amino acid difference between each other, with x being a user defined threshold. The first generation was very similar to the wild type and any threshold drastically reduced the number of effectively non redundant sequences, while the fifth generation had a mean of 11-12 mutation (~4%) in respect to the wild type and around double this number (~8%) if cross-checked one versus the other. In this generation, I tested several thresholds and obtained the best results with a threshold set to at most 5% of discordance in the sequences peptidic compositions (**Figure 2.3.16**).

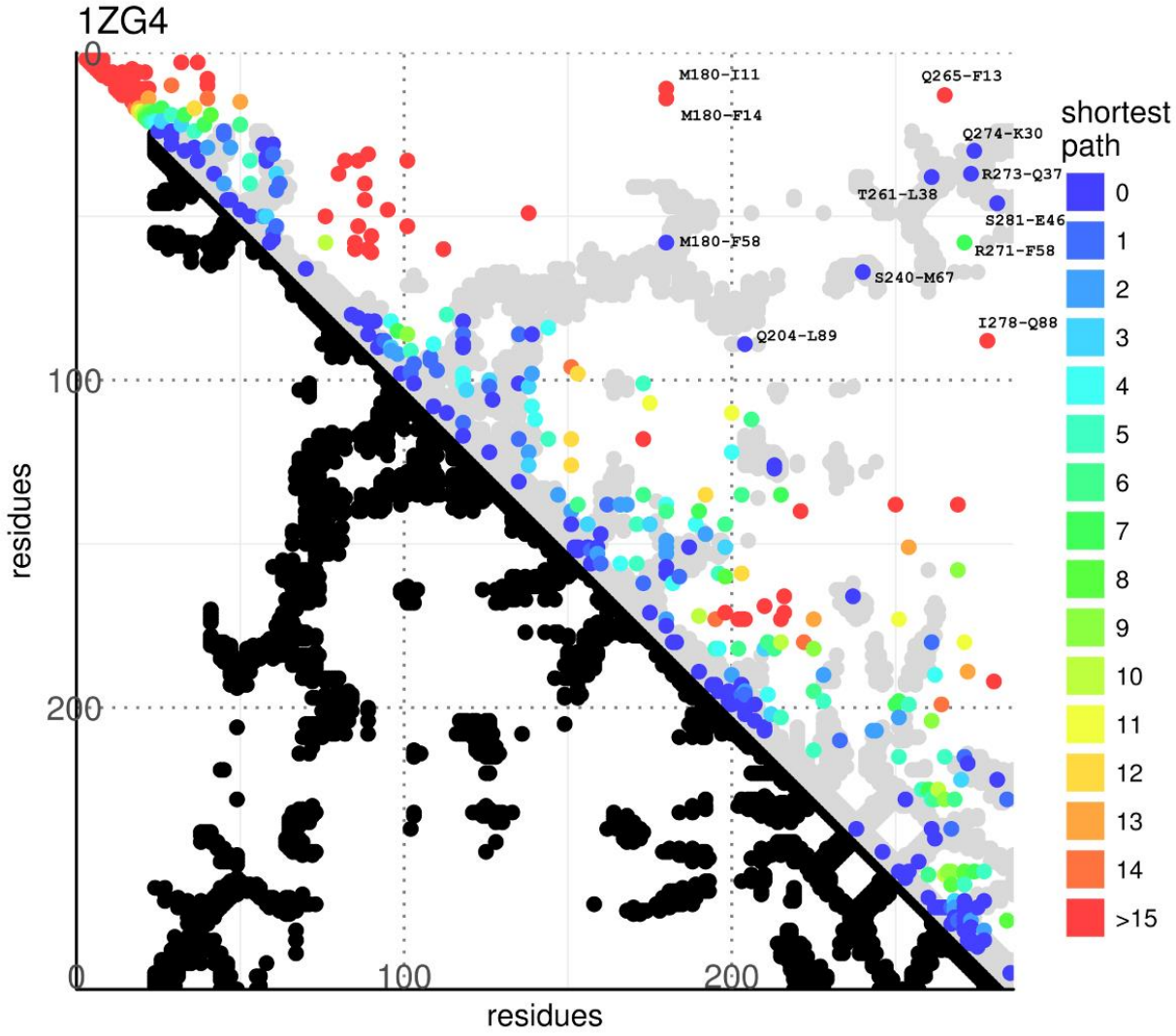


Figure 2.3.16 DCA plot of the fifth generation library after reweighting for similarity (95%). DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the fifth generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph

the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

After the phylogenetic correction the agreement in the structural information is greatly improved, the predictions tend to cluster and are crowded in the area near the diagonal. Of particular interest are the elements at the beginning of the protein (residues 0-60) where the prediction clearly overlaps with the interactions made by the first few N terminal secondary structure to each other (Helix H1 with the first two strands of the central beta sheet, see introduction **chapter 1.4.3**). Other important clusters can be seen in correspondence to the branching points from the diagonal (near the diagonal around residues 100, 160, 200 and 260). This set of contacts running perpendicular to the diagonal are quite common in contact maps and are created when the protein chain takes a sharp turn and a set of residues form a “hairpin” of antiparallel secondary structures. The clustering at the branching point should be expected because the branching point is where the loop takes place. Since loops are more conformationally flexible and solvent exposed as compared to secondary structure elements, their compositions tends to be less critical for the fold of the protein and allows a broader range of variation. More variations in amino acid composition increase the probability to observe a covariation pattern during the DCA.
Another interesting feature of this prediction is this contact prediction point distance from the diagonal. It can be clearly seen that the predictions around the N terminal secondary elements (residues 0-60) are tightly packed against the diagonal, like the underlying structural traces, while in the rest of the protein the contacts are more dispersed, in a similar way to the interactions observed in the crystal.
Talking of short- and long-range contacts, the long-range contacts, even if few in number, are nonetheless very interesting because long range predictions are the most important to reconstruct the tertiary structure of the protein (these prediction points can be found annotated in **Figure 2.3.16**). Most of the long range predicted interactions overlap the structural trace, with only five exceptions that do not match the reference. One of these exceptions, R271-F58, is predicting a contact between helix 12 and the second beta strand of the protein. This interaction is incorrect, but it has to be said that these secondary structure elements are both part of the “terminal domain” (see introduction **chapter 1.4.3**) and are not far from each other. Moreover, the prediction point itself is located right in the middle of a cluster of interacting residues that could act as a confounding factor.
Other two of the incorrect predictions involve the methionine at position 180: the predicted interaction between M180 and the isoleucine in position 11 and the interaction with the phenylalanine in position 14. In the long-range interaction predicted by the DCA there are actually three interactions that involve M180, the other being the (correct) interaction with the phenylalanine in position 58 that, ironically, is the position we observed in the previous exception. As already stated in the previous section, M180 was a clear outlier of this generation, where ~30% of the sequences of the library were carrying a substitution to a threonine. Most probably this unbalance in the frequency of variation had generated a stronger background for the position 180 compared to the other residues of the chain, thus making the position noisy and inaccurate.

The correct predicted long-range interactions are instead seven, four of which specific for the terminal domain. Two of these mediate the interaction between the N- and C- terminal helices (Q274-K30, R73-Q37) while the other two mediate the interaction of the terminal helices with the beta sheet (T261-L38 mediate the interaction between the middle strand of the sheet B9 and the N terminal helix H1 while S281-E46 mediate the interaction between the C terminal helix H12 with the first beta strand B1).
M180-F58, the prediction involving M180 previously discussed, is a contact between the coil between B6 and H9 and the lateral beta strand B2. This is an extremely interesting area because F58 stands at the very beginning of one of the two hinge regions that connect the two domains of the protein.
The last two contacts are at the opposite ends of the innermost helix of the protein, the helix H2 of the helical domain, a helix that carries both a catalytic and structural functions. Q240-L89 is an interaction between H10, the last helix of the helical domain, and the coil at the end of H2, while S240-M67 mediate the interaction of H2 with a conserved serine in the coil between the beta strands B7 and B8.
Overall this prediction is indubitably a good result. Compared to the first few attempts, this correction cleaned the prediction quite nicely creating clusters in regions of interest. This new prediction also reflected lots of the underlying structural data. However, a critical assessment of the result can only conclude that the prediction is still insufficient to answer our biological problem, that is to retrieve the protein structural information. The data are still far too sparse to define clear-cut interaction zones and tend to cluster around the diagonal. This analysis provided a good description of contacts in proximal position along the protein sequences while long range contacts, even if describing critical zones, are far too few to give us a panorama of the protein tertiary structure. The protein prediction also shows a clear cluster of points (residues in positions 40-70 against residues around positions 100) that do not reflect any structural contact.
I hypothesized that this cluster of points could reflect a dimeric interaction of the protein, but there is very little support for this theory in the literature (Braswell et al., 1986). Other types of lactamases (most importantly type D beta lactamases) are known to create oligomeric structures (Danel et al., 2001) and I wondered if what I observed could be a remnant of an ancestral dimeric form of the protein. I assumed that if such an interaction had to occur it would probably resemble the interaction in the crystal packing. Thus, I calculated all the possible contacts of the beta lactamase with its nearest crystallographic repetitions in the reference structure to verify if a similar cluster could be observed. Of all the crystal-packing dimers only one paring showed a significant surface of interaction (**Figure 2.3.17**). This putative interaction cluster of points, by luck or by design, has a fairly good match with our data. It is really difficult to demonstrate if this cluster reflects a real long-lost dimeric interaction or these predicted points are in fact an artifact that, by chance or by luck, overlap with the contacts made by crystal packing. It could even be a non-natural interaction that the protein developed by molecular evolution that brings a cryptic advantage against the laboratory conditions. Right now, this is still an open question and this correlation can only be considered a noteworthy observation.

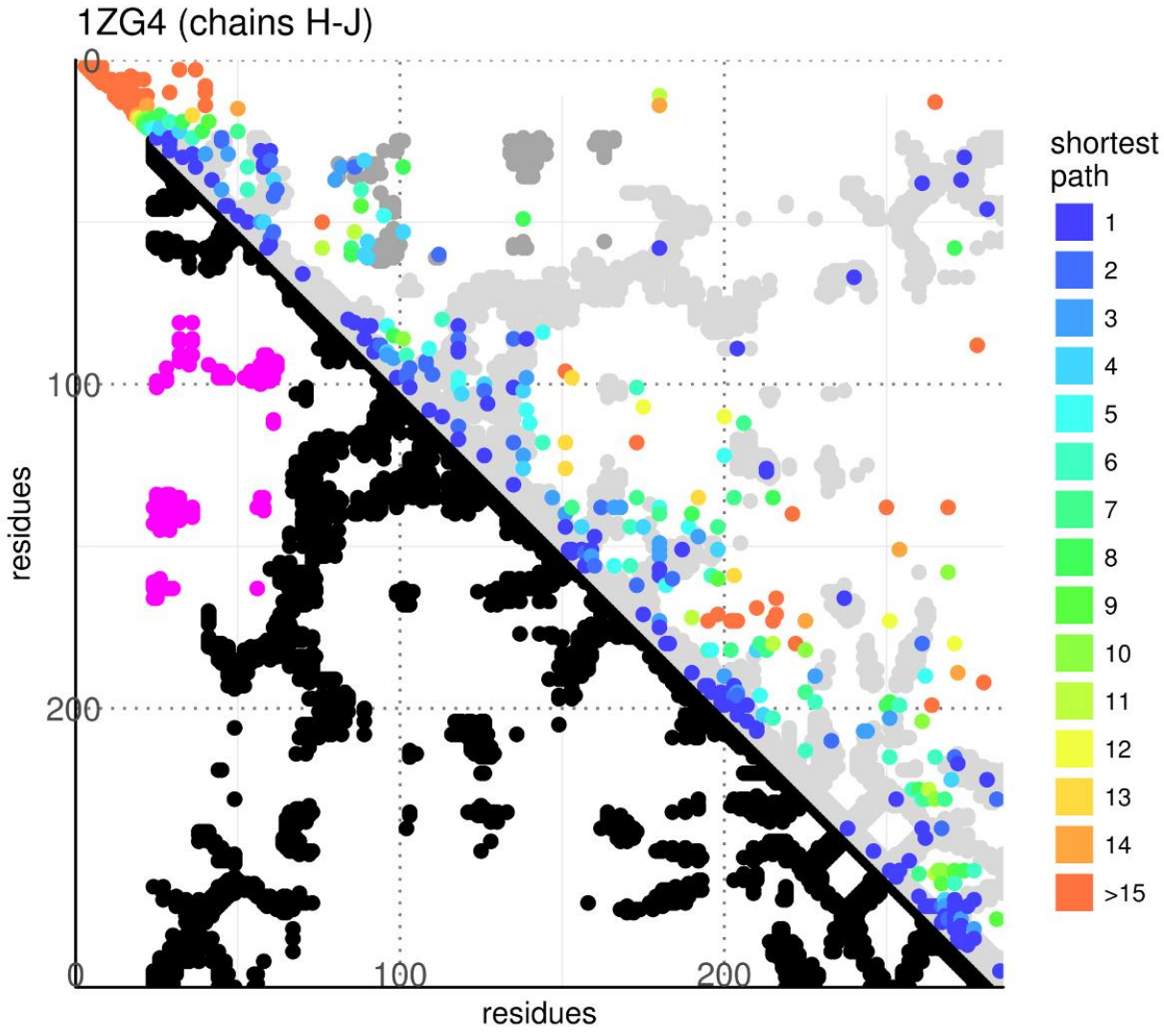


Figure 2.3.17 Reweighted DCA plot of the fifth generation library including the strongest interchain contacts between crystal dimers in the reference structure.
DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the fifth generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment

are excluded from this ranking to promote visualization of long range interactions. In the graph the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure). The contacts between the crystal dimers in the crystal structure that have the largest interaction surface are coloured in bright magenta and in dark grey in the mirrored part of the graph.

As a final remark for this generation, it is important to notice that the second helix H2 shows essentially no predicted contacts with the other helices of the domain. In the crystal structure this helix is the most important helix of the helical domain since it is the central innermost helix onto which the other helices are packed on. The problem lies in the number of mutations observed in this area. Since a high substitution rate (M180 in **Figure 2.3.14**) is associated with a stronger coevolution signal (**Figure 2.3.16**) it is logical to assume that a low substitution rate is associated to a weaker one. Helix H2, being very deep in the protein hydrophobic core and being very important for the folding and function of the protein, was very reluctant to any alteration pressure and presented us with fewer mutations (**Figure 2.3.14**) that translated into fewer predictions (**Figure 2.3.16**).
To overcome all of these problems and try to obtain a more accurate prediction, new generations of mutated beta lactamase were produced and the last library reached the final milestone of twelve generations of mutagenesis.

2.3.6 Twelfth generation of molecular evolution (GEN12)

The twelfth and last generation of mutagenesis was also subjected to a few tests to evaluate the best conditions to apply to optimize the big scale transformation. In addition, because the ninth generation seemed to have reached a plateau in the number of mutations (**Figure 2.3.18**), I also collected the same data for the tenth generation. Both generations were transformed into bacteria and selected in growth media with the normal 25µg/mL and the increased 100 µg/mL ampicillin concentration, 10 colonies were randomly selected from the surviving growth and sequenced to calculate the mean number of mutations in the peptide chain (**Table 2.3.6** and **Table 2.3.7**).

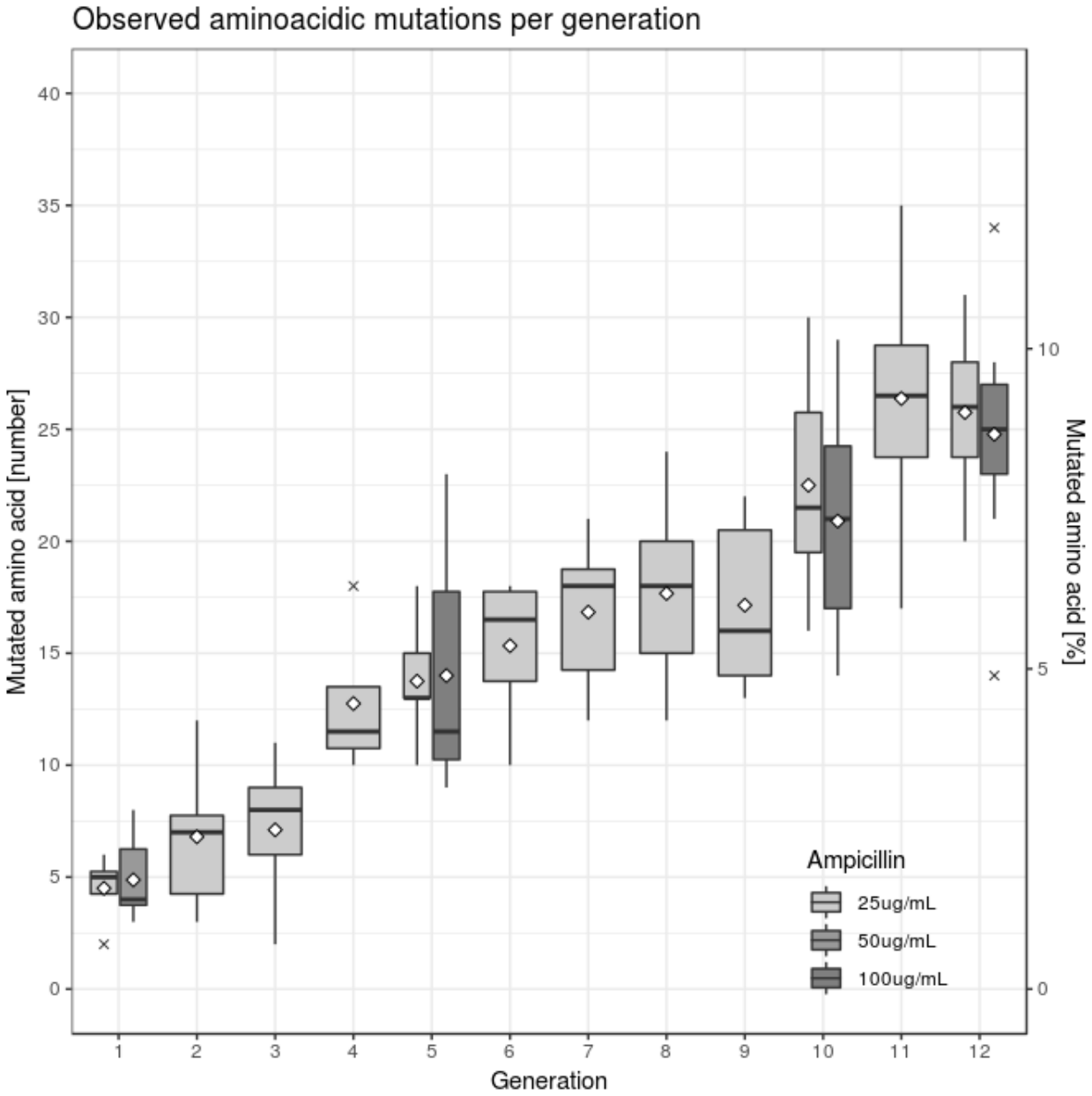


Figure 2.3.18 Mutations in the beta lactamase sequence observed in a small sample of clones sequenced after each generation of molecular evolution.
Boxplot showing the number of amino acid mutations observed in the sample of clones sequenced after each generation (Sanger sequencing). The white diamond dots indicate the mean**.**

10TH GENERATION

Sample AMP25µg/mL	*DNA mutations (bp)*	*Protein mutations (aa)*	*Sample AMP100µg/mL*	*DNA mutations (bp)*	*Protein mutations (aa)*
			blaX_100A	42	14
blaX_25B	*56*	*28*	*blaX_100B*	53	20
blaX_25C	*38*	*18*	*blaX_100C*	49	26
blaX_25D	*49*	*21*	*blaX_100D*	43	22
			blaX_100E	35	15
blaX_25F	*52*	*25*	*blaX_100F*	49	25
blaX_25G	*58*	*30*	*blaX_100G*	46	21
blaX_25H	*47*	*20*	*blaX_100H*	50	21
blaX_25I	*46*	*22*	*blaX_100I*	38	16
blaX_25J	*43*	*16*	*blaX_100J*	52	29
mean	*48.6*	*22.5*	*mean*	*45.7*	*20.9*

Table 2.3.6 Mutations in the beta lactamase sequence observed after selection in the tenth generation library.

Mutations in the nucleic and amino acid sequence observed in the beta lactamase of a small sample of colonies (A-J) that survived the selection with 25 or 100µg/mL Ampicillin. The bacterial cells were transformed with a library of mutated beta lactamases. This library was obtained by applying 20 cycles of error prone PCR to the lactamase insert of the previous generation. The samples containing aberrant sequences (greyed out in the table) were not used to calculate the means.

12TH GENERATION

Sample AMP25µg/mL	*DNA mutations (bp)*	*Protein mutations (aa)*	*Sample AMP100µg/mL*	*DNA mutations (bp)*	*Protein mutations (aa)*
blaXII_25A	49	20 (1 STOP)	*blaXII_100A*	65	27
blaXII_25B	66	28 (1 STOP)	*blaXII_100B*	52	28
blaXII_25C	60	32 (1 STOP)	*blaXII_100C*	65	25
blaXII_25D	61	27	*blaXII_100D*	55	23
			blaXII_100E	50	26
blaXII_25F	57	20	*blaXII_100F*	60	25
blaXII_25G	42	19 (1 STOP)	*blaXII_100G*	49	21
blaXII_25H	57	31	*blaXII_100H*	69(12 Indel)	33 (4 Ins)
blaXII_25I	47	25	*blaXII_100I*	58	34
blaXII_25J	59	28 (1 STOP)	*blaXII_100J*	46	14
mean	*55.5*		*mean*	*55.6*	*24.8*

Table 2.3.7 Mutations in the beta lactamase sequence observed after selection in the twelfth generation library.

Both generations show a clear effect of the antibiotic concentration.
In the first generation I chose the lowest antibiotic concentration because with similar mutation rate results I obtained a drastic difference in transformation efficiency. In the fifth generation I chose the highest antibiotic concentration to hinder the variants of the lactamase coding for a protein with low activity. I tested the 10th generation with two antibiotic concentrations because in the ninth generation the number of observed mutations per sequence seemed to have reached a saturation point, but the 10th generation proved it was not the case (**Figure 2.3.18**). Since the mutational load plateau was not yet reached, I continued the creation of more mutated generations with the lower 25 µg/mL ampicillin concentration to maintain a lower selective pressure. The last generation showed a strong incidence of deleterious mutations in the sample selected with a lower antibiotic concentration, while the stronger selection seemed more adequate for our purpose.
To sum up, the complete set of measurements for the number of mutated amino acids per sequence in our libraries, keeping only the samples I used to create the generational chain from the first to the last generation, looked like this (**Figure 2.3.19**):

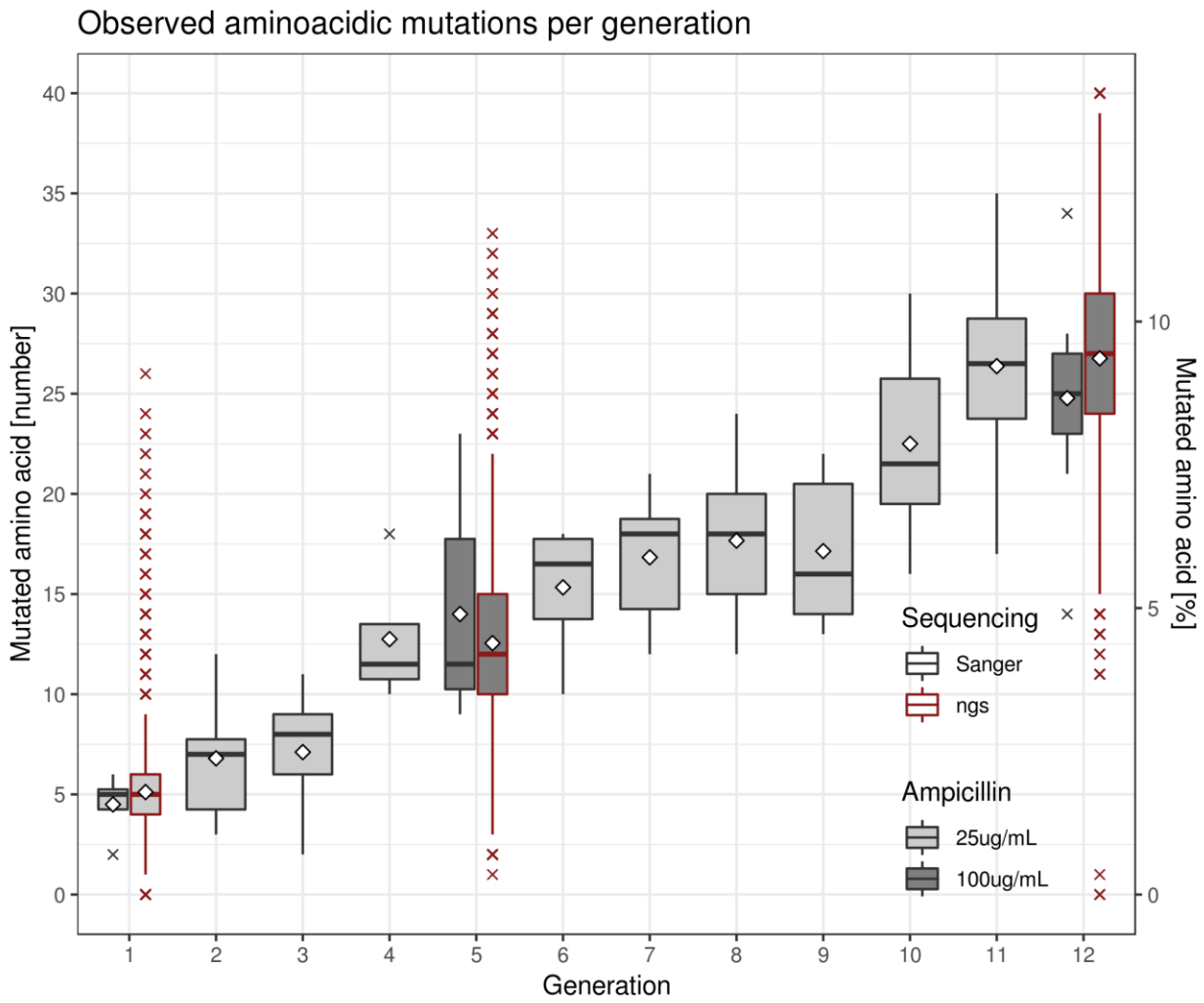


Figure 2.3.19 Mutations in the beta lactamase sequence observed during molecular

evolution.
Boxplot showing the number of amino acid mutations observed in the sample of clones sequenced after each generation (Sanger sequencing, black border) and after NGS (red border). The white diamond dots indicate the mean**.**

After the sequencing of the twelfth generation library (GEN12), the dataset contained 8.59Gb of sequences distributed in 783777 Raw reads. I built the intramolecular consensus with the same parameters as for the fifth generation, requiring the consensus to have at least 10 passes and a minimum quality of Phred 40 (**Table 2.3.8**).

CCS2 statistics on TEM-1 beta lactamase GEN12 library (passes > 10, Phred quality > 40)

Movie	*CCS reads*	*Number of CCS bases*	*CCS Read Length (mean)*	*CCS Read Score (mean)*	*Number of Passes (mean)*
m54138_181012_135712	99,395	91,135,751	916	1	32

Table 2.3.8 Statistics of the last generation library generated by CCS2 during consensus building using the same parameters used in generation 5.

There is sharp drop in the number of sequences obtained at the end of consensus building. Even if the dataset contained ~20% less material (8.59Gb) compared to the previously selected generations (~10Gb), I did not expect that to translate into a halving of the number of effective sequences (from 170 K to less than 100K).
The mapping and the *in silico* translation served only to aggravate this problem, showing that only ⅓ of the proteins were able to maintain the length of the wild type archetype (**Figure 2.3.20**).

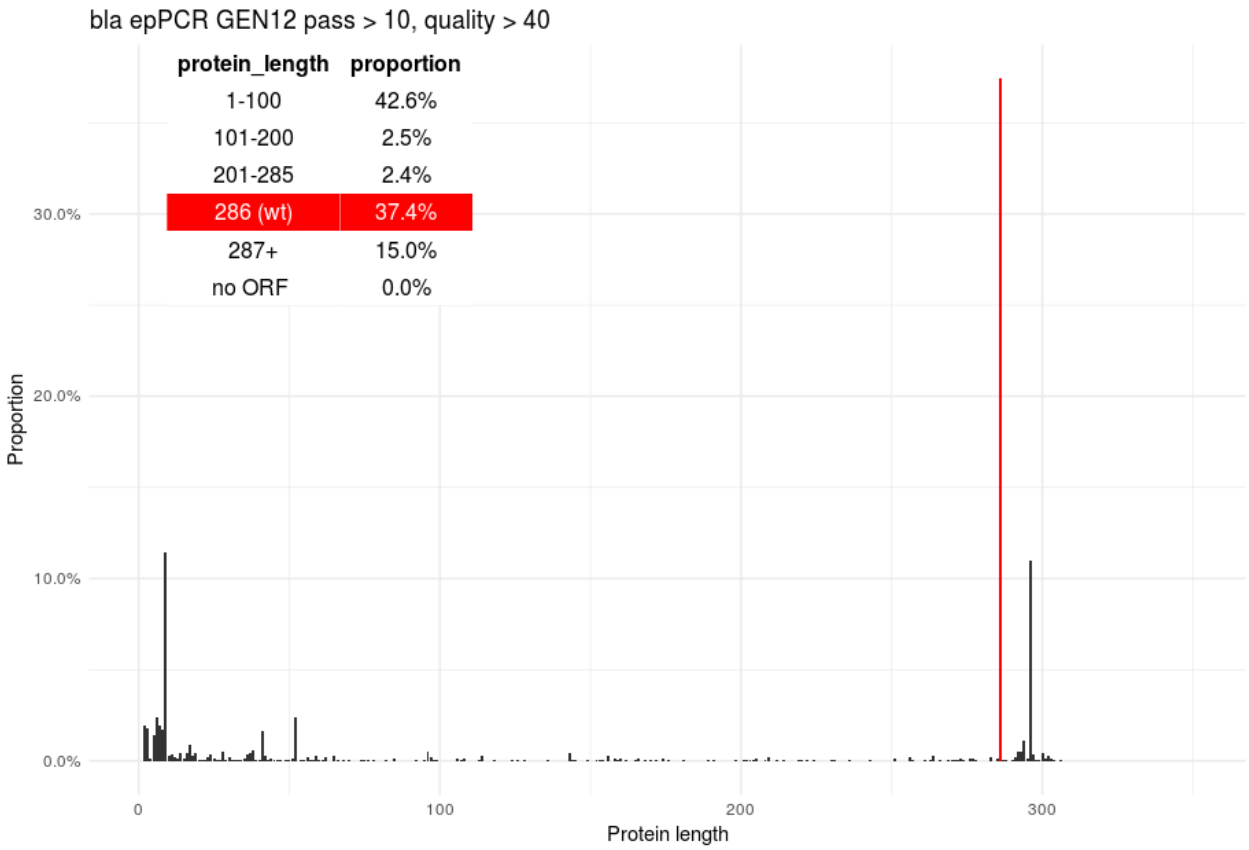


Figure 2.3.20 Proteins length distribution of the translated consensuses of the GEN12 library.

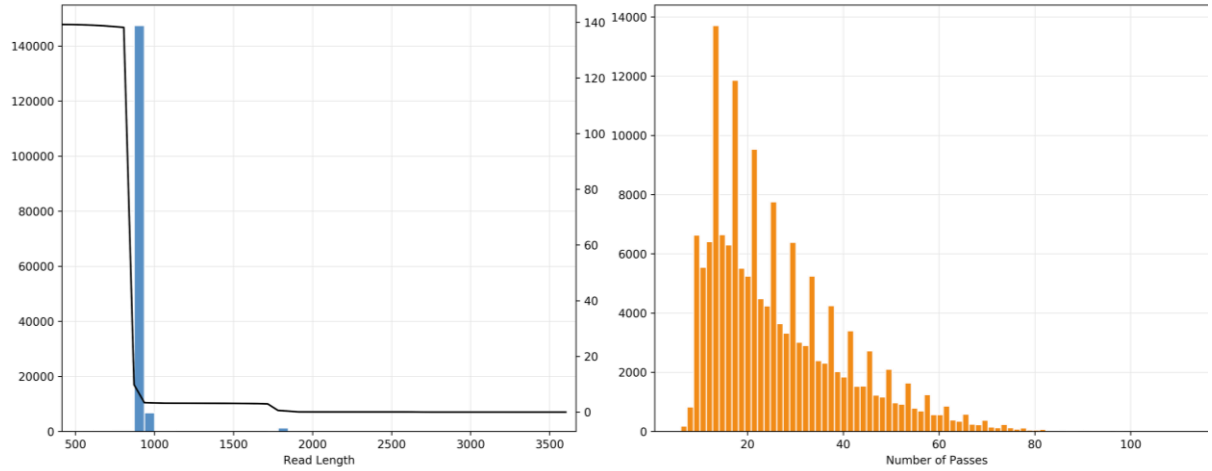
The origin of the issue lies in the presence of the small truncated peptides that form up to 40% of the sample. How could these truncated forms be functional and be selected, being only less than a third the length of the functional protein? The way the in silico translations were generated in the first place was to map each sequence to the original TEM-1 beta lactamase DNA sequence to retrieve the ATG codon corresponding to the starting methionine, and from there onward the nucleotide sequence was translated into the peptidic counterpart. However, during the Sanger sequencing I made to control the advancement of mutagenesis, I observed a very frequent and progressive increase in the fraction of sequences carrying a degeneration of the regions flanking the beta lactamase.
These degenerations of the flanking sites were small insertions and duplication of the surrounding sequence. They were likely created during the enzymatic digestion of the library or during the ligation of the insert, and they were progressively accumulated throughout the generations. The degeneration is fairly neutral for the function of the gene and gave no significant harm or advantage to the sequence. In contrast, the degeneration before the gene could cause the appearance of an earlier starting site and thus affect the length of the protein. In particular, if the genomic mapping was tricked by a small duplication in the 5' of the gene to align the starting ATG codon to an earlier start site, we would either observe a longer protein, if the frame was maintained, or a truncated frame shifted byproduct otherwise.
This was in line to what observed in the graph above (**Figure 2.3.20**), with an accumulation of

both very short fragments and proteins longer than the reference lactamase sequence. This does not mean that the sample is rich in variants that translate in short truncated peptides, but it means there is an issue in the pipeline that generates problematic translations that have to be addressed.
I decided to restart the analysis from the consensus building. My target was to obtain a number of sequences comparable to the other two sequenced generations and to do so I lowered the requirements needed to pass the consensus building process. More lax parameters such as keeping the sequences with at least 5 polymerase passes and that showed at least a mean quality of Phred 30 were sufficient to achieve this goal (**Table 2.3.9**).

CCS2 statistics on TEM-1 beta lactamase GEN12 library (passes > 5, Phred quality > 30)

Movie	*CCS reads*	*Number of CCS bases*	*CCS Read Length (mean)*	*CCS Read Score (mean)*	*Number of Passes (mean)*
m54138_181012_135712	157,899	145,528,553	921	1	26

Table 2.3.9 Statistics of the last generation library generated by CCS2 during consensus building using the parameters: passes > 5, Phred quality > 30.



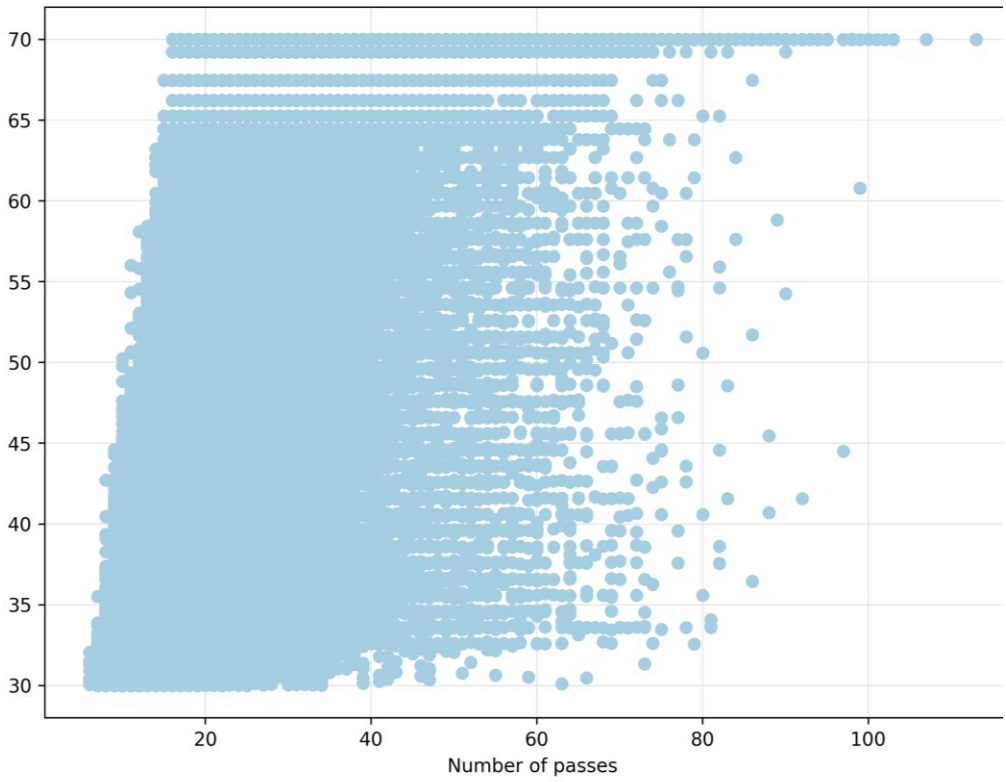


Figure 2.3.21 Graphical reports of the last generation library generated by CCS2 during consensus building using the parameters: passes > 5, Phred quality > 30.
A) Histogram of the lengths of the consensus reads. B) Histogram of the number of times the polymerase passed and read the sequence (i.e. the number of repetitions of the sequence present in the raw read). C) Scatterplot of the number of polymerase passes and read score associated to each raw read.

The profiles of read lengths, number of passes and number of passes in relation to the mean quality score were in line with the previous generations (**Figure 2.3.21**). The unusual distribution in the histogram of the number of passes per sequence observed in the analysis report is a graphical artifact caused by the data binning (in specific binning two classes together) used in the SMRTlink css2 program to generate the plot. The actual distribution is a smooth curve.
To correct the second problem, the inflation of the truncated sequences in the library, I had to add a couple of assumptions to this model: i) if the cell was able to survive this was probably due to the expression of a functional beta lactamase and ii) these functional lactamases are probably coded by the longest and most complete open reading frame present in our sequenced fragments. With these assumptions the aminoacidic sequences of the beta lactamases of the sequenced library are obtained by *in silico* translating the longest open reading frame present in the reads. As expected, the new profile of protein lengths vastly differs from the previous, in particular in the shorter protein class (**Figure 2.3.22**).

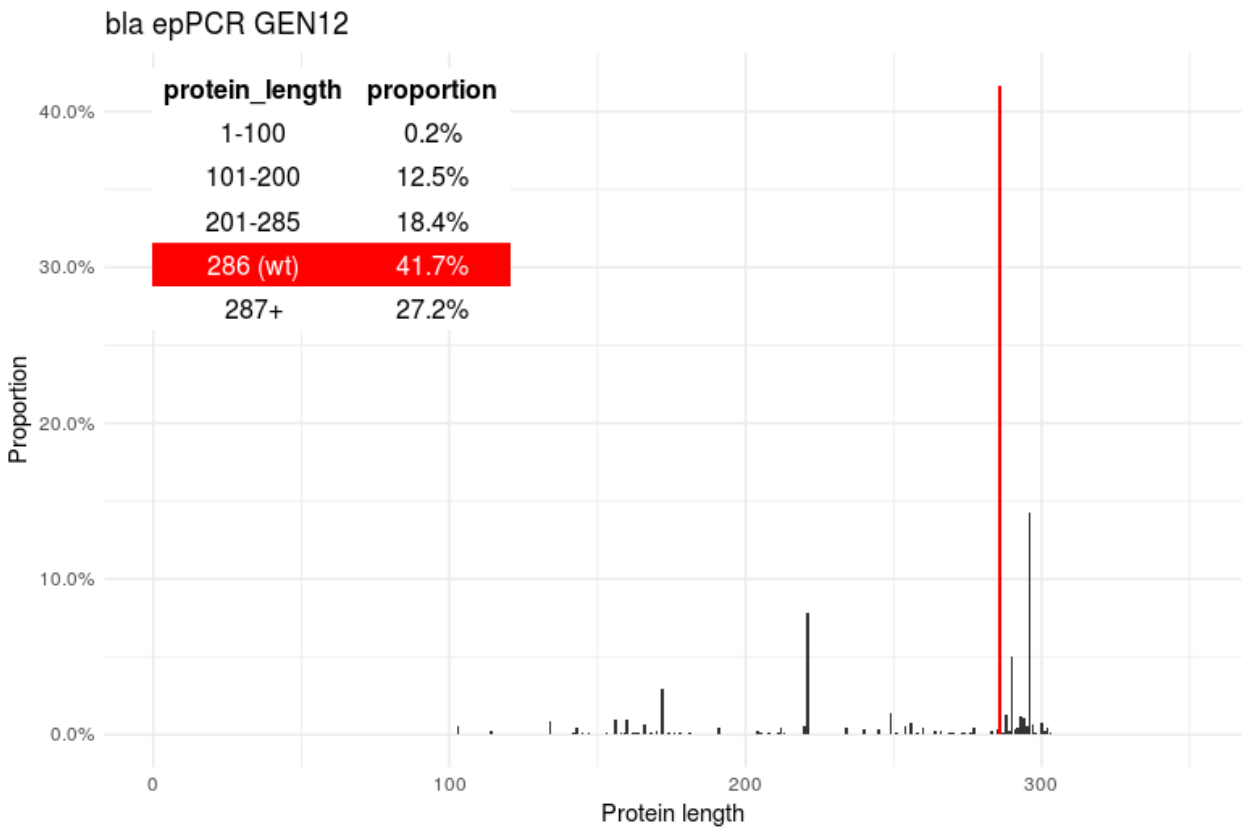


Figure 2.3.22 Proteins length distribution of the translated longest ORFs found in each consensus sequence of the GEN12 library.

Another effect of using this method is an increase in the fraction of proteins longer than the original archetype. This is likely caused by the emergence of an earlier translation starting point in those sequences caused by the degeneration of the 5' flanking region of the gene.
The sequences coding for proteins shorter than the TEM-1 lactamase must be dropped because it was proved in the first generation that the frameshifts heavily affected the coevolutionary strength (**Figure 2.3.8**). There is no reason however not to keep the sequences that code for proteins longer than the wild type sequence since it would be very unlikely for a sequence carrying a frameshift to produce a longer product.
All the translated proteins that were at least long as the wild type were collected (106K in total) and aligned to each other to obtain an MSA where only the columns corresponding to the original amino acid sequence were retained. The alignment was then used to calculate the number of mutation and their distribution (**Figure 2.3.23**).

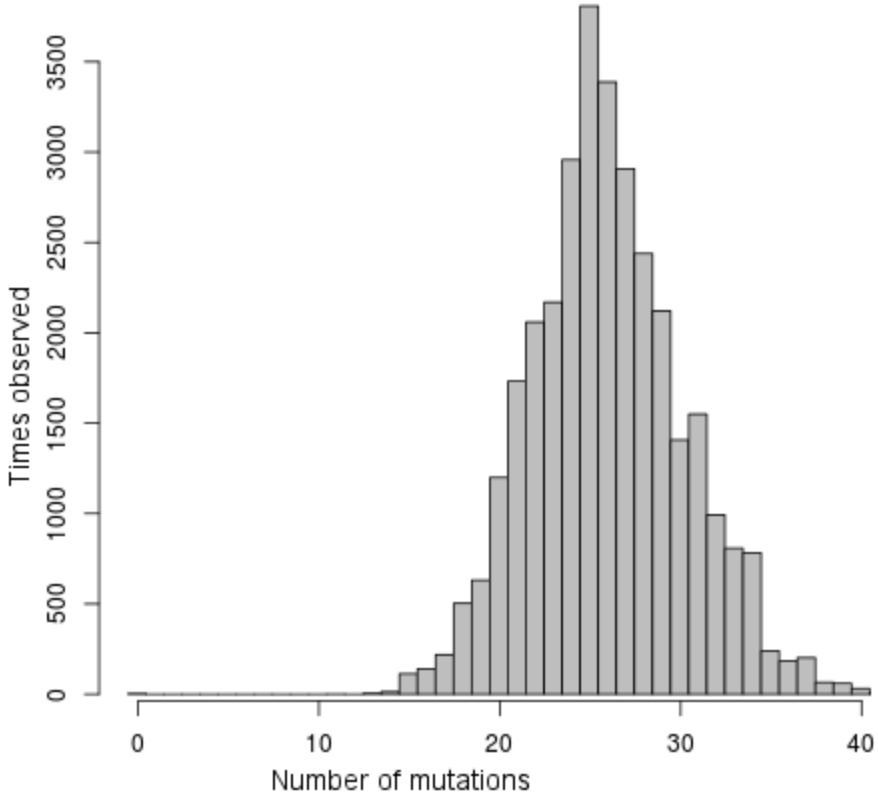


Figure 2.3.23 Number of mutations observed in the peptidic sequence of beta lactamase in the twelfth generation library

The bell-shaped distribution peaks around 25 mutations per sequence (mean 26.90, median 27, variance 21.97). This library is one of the most mutated molecularly evolved TEM beta lactamase libraries ever produced, where its elements diverge from the ancestral protein for around 1/10 of their original peptide chain composition.

This is our last generation of mutagenesis and needs to be discussed in detail and be compared with the previous generations. This will be done in standalone chapters to explain the similarities and dissimilarities to the features observed in a collection of natural variants.

2.4 Molecular evolution in DCA

2.4.1 Molecular evolution libraries meet the expected quality and mimic natural variability.

To recapitulate a little before going into the actual data, this project is about using molecular evolution instead of natural evolution as an alternative source of variability to retrieve the evolutionary coupling of a protein. I employed error prone PCR to drive the mutagenesis that allowed the generation of several libraries of mutational variants of the beta lactamase gene. These libraries underwent *in vivo* phenotypic selection in a growth functional assay to retain only the variants coding for a functional protein. I noticed that the mutational load generated by the error prone PCR was unable to produce variants that were at the same time vital and heavily mutated, so I adopted the generational approach of mutating the library that was phenotypically selected after a previous round of mutagenesis. This way new mutations began to accumulate on a mutated-but-functional scaffold of the previous generation and the results were allowed to be both functional and vital and be progressively more divergent from the ancestral sequence.

To verify the progress of molecular evolution and maintain libraries with a fair amount of complexity, I controlled three parameters throughout the twelve generations: the number of transformants in the bacterial growth, the number of mismatching amino acids in a small sample of clones and the information entropy at each amino acid position.

Since each bacterial colony in the selection medium expresses a single functional variant of the protein, the number of transformants poses a theoretical ceiling to the library diversity. I kept the number of transformants at least in the same order of magnitude of the sequencing capacity of NGS platform (between 100 thousand and 1 million, see **chapter 2.2.7**) to guarantee a good library complexity. In the last few generations I raised this limit to 400 thousand clones to increase the probability of sequencing unique variants.

After each generation a small sample of clones underwent Sanger sequencing to retrieve an estimation of the number of mismatching nucleobases and amino acids with respect to the ancestor sequence (**Figure 2.3.18** and **Figure 2.3.19**).

To complement this information, the same parameter was estimated from the sequencing results of the three NGS sequenced generations. The distribution of the number of mismatches per sequence fitted the theoretical Poissonian model expected for a mutagenesis (gen1: λ=5.12; gen5: λ=12.54; gen5: λ=26.9) (**Figure 2.4.1**).

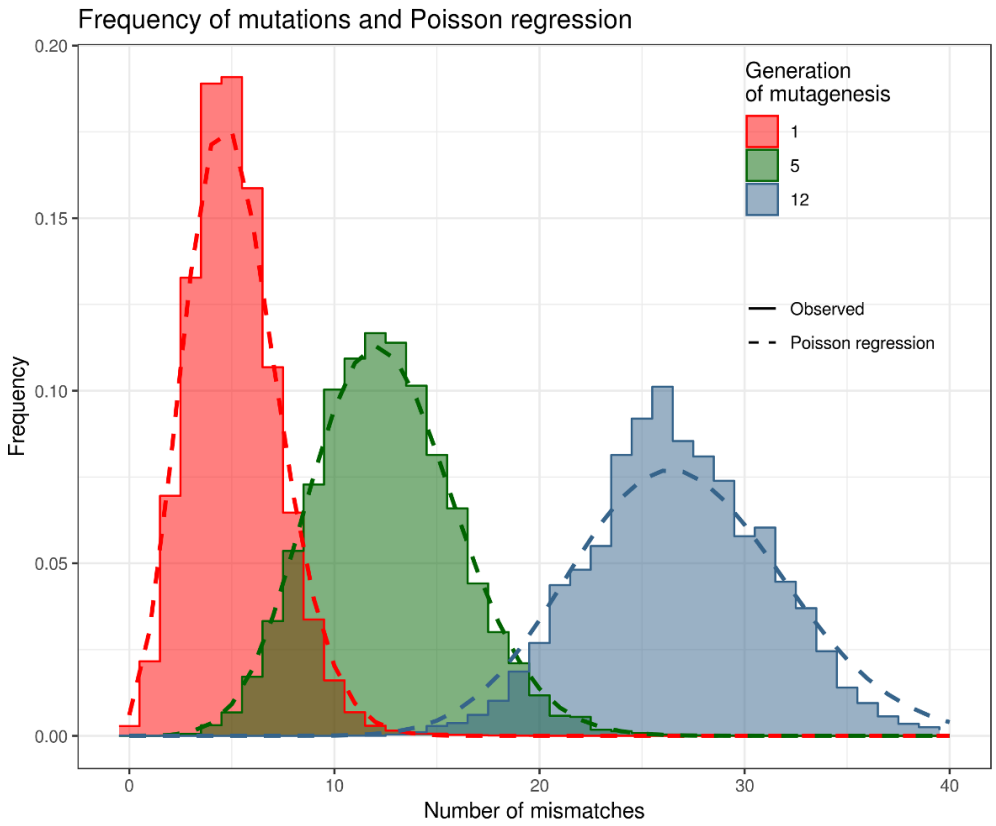


Figure 2.4.1 Mutations observed in the peptidic chains of the sequenced libraries. Frequency distribution of the number of aminoacidic mutations observed in the sequenced libraries (solid lines) and their respective Poissonian regressions (dotted lines: gen1: λ=5.12; gen5: λ=12.54; gen5: λ=26.9).

The median number of mutated residues observed when the picked colonies matched perfectly that obtained from next generation sequencing (**chapter 2.3.6**, **Figure 2.3.23**) and what was expected from a Poissonian model, proving that the handful of colonies picked provided a good representation of the mutations present in the library. The observed steady increase in the number of mutations throughout the molecular evolution supports the idea that the final mutation rate of the evolved protein library can be regulated by increasing the number of generations. Sequencing data also allowed us to calculate the mutation rate per amino acid position, defined as the frequency of the observed mismatching amino acids compared to the original pUC19 beta lactamase sequence (**Figure 2.4.2**).

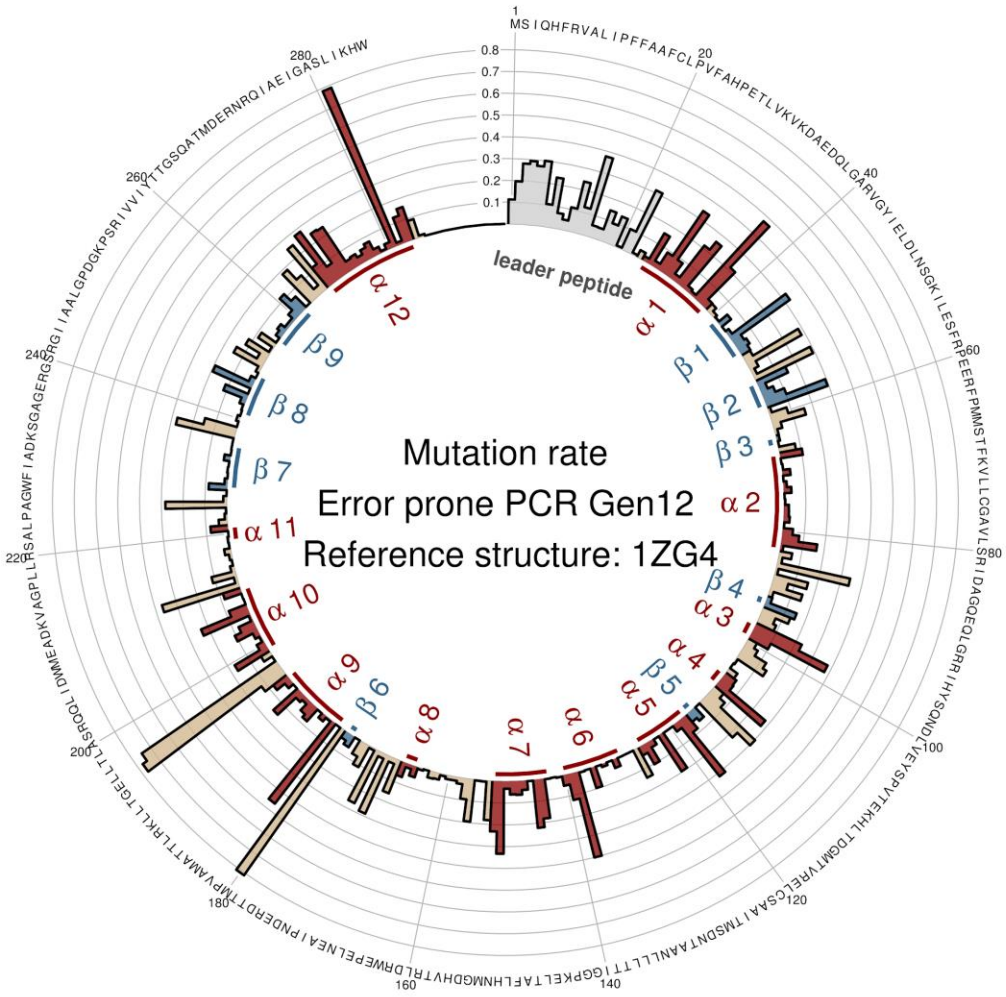


Figure 2.4.2 Mutation rate per residue position observed in the twelfth generation library. The colours and annotations follow the secondary structure classification present in the PDB structure 1ZG4 (red: alpha helices, blue: beta strands, tan: coils). The leader peptide sequence (light grey) is missing in the structure.

After 12 generations of molecular evolution there were several instances where the mutations became more common than the original residue for a given position. To be more specific, these positions were not extremely more variable than the rest, but the most common amino acid that could be found in the position changed to another one (M180T, E195D, L196I, S281T). This phenomenon made the mutation rates less informative, since they involved a comparison to the original residue that is now a minority. To circumvent the problem, I measured the Shannon information entropy of each residue, obtaining an approximation of the impact of the

mutagenesis for each position, without the need of a reference sequence (**Figure 2.4.3**).

Figure 2.4.3 Shannon information entropy per residue position observed in the twelfth generation library.
The colours and annotations follow the secondary structure classification present in the PDB structure 1ZG4 (red: alpha helices, blue: beta strands, tan: coils). The leader peptide sequence (light grey) is missing in the structure.

Shannon entropy also tells us that the protein features a double standard in the mutation propensity of its positions, with some positions remaining practically unchanged while others accommodate the mutations more easily.
The proportion of mutants and the information entropy of each residue in the NGS sequenced generations were strongly correlated one to the other (**Figure 2.4.4**) and with those observed

from the UniProt dataset (mutant frequency: rho 0.624, p < 1e-15; entropy: rho 0.632, p < 1e-15) (**Figure 2.4.5**).

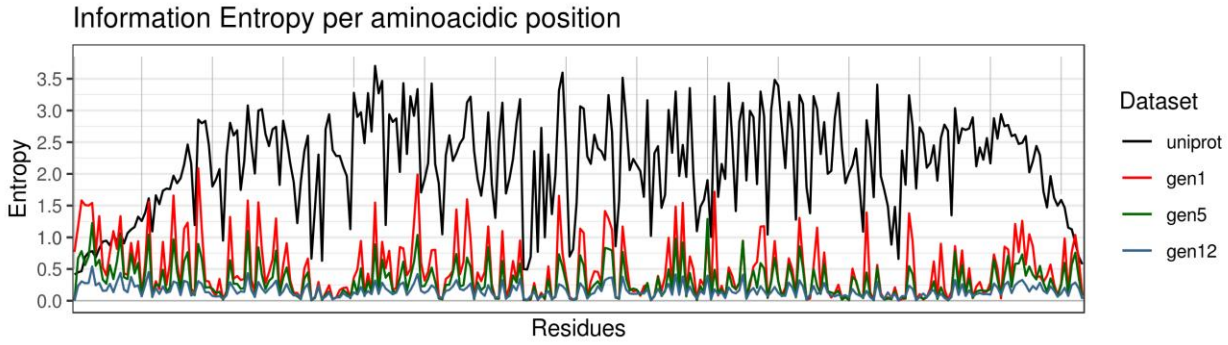


Figure 2.4.4 Shannon information entropy in the UniProt and the in vitro evolved datasets.

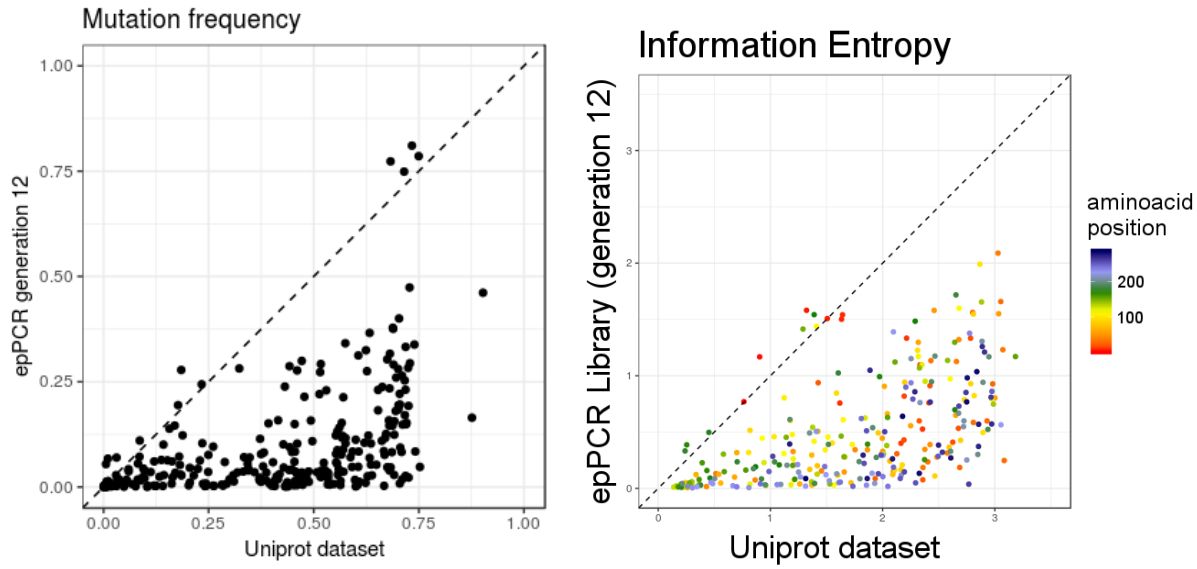


Figure 2.4.5 Comparison of the UniProt dataset to the twelfth generation of molecular evolution.

Correlation of the Shannon entropies (left) and mutation rates (right) observed between positions of the UniProt dataset and the same position of the twelfth generation of molecular evolution. In the graph on the right the points were coloured based on the order they appear on the peptide chain (position).

Moreover, entropies and mutation frequencies of the evolved libraries are almost always lower than corresponding entropies and mutation frequencies of the reference “natural” UniProt dataset (**Figure 2.4.4**, **Figure 2.4.5**). This supports the idea that the latter poses a limit to which a molecular evolution library would tend, given enough mutagenesis rounds. Interestingly, the positions which correlate the most with the UniProt dataset are the N-terminal ones, where the leader peptide is. This could be either caused by some biological properties of the leader peptide that somehow can support a faster evolution or could be due to an increased variability of the area due to its closeness to the degenerate flanking region.

The data corroborate the hypothesis that molecular evolution took a very similar pathway as the

one taken by natural evolution, suggesting that the former could effectively substitute the latter as a source of genetic variants. The cycles of mutations and selections determine the population bottlenecks that alter the genetic variability of the system. This will inevitably affect the genetic variability of the population. The drastic reduction in the population size after catastrophic events will establish a founder effect and in these smaller communities genetic drift and fixation are far more common.
In our molecular evolution, mutations build the variability and the selection generates a rapid drop in the population size. Thus, by analogy, the mutagenesis mimed the time that builds variability whilst selection mimed the catastrophic events, migrations, isolations and all the events in which the population size is functionally reduced.

2.4.2 The mutational landscape of the evolved library reflects the structural features of TEM beta lactamases

The beta lactamase structure 1ZG4 from PDB (https://www.rcsb.org) (**Figure 1.4.4**) was used as a reference structure to assess the contact prediction and the accuracy of the prediction analysis. The PDB structure lacks the first 23 amino acids, corresponding to the leader sequence for secretion, which is cleaved during protein maturation to allow protein release. Detailed information of the structural feature of the TEM-1 beta lactamase can be found in the introduction **chapter 1.4.3**)
The profiles of the mutation rate and entropy per residue observed in our molecular evolution libraries is conserved and increases across generations, in line with what is observed in the UniProt dataset of the naturally evolved beta lactamase family (**Figure 2.4.4**).
This profile reflects the different mutation propensities of the various residues as well as the interactions with the solvent and the polarity of the local environment. There is a high degree of conservation in the presence of bulky nonpolar residues like tryptophans and methionines and cysteines involved in the sulfur bridge, whilst small residues show in general an increased variability (**Figure 2.4.6**).

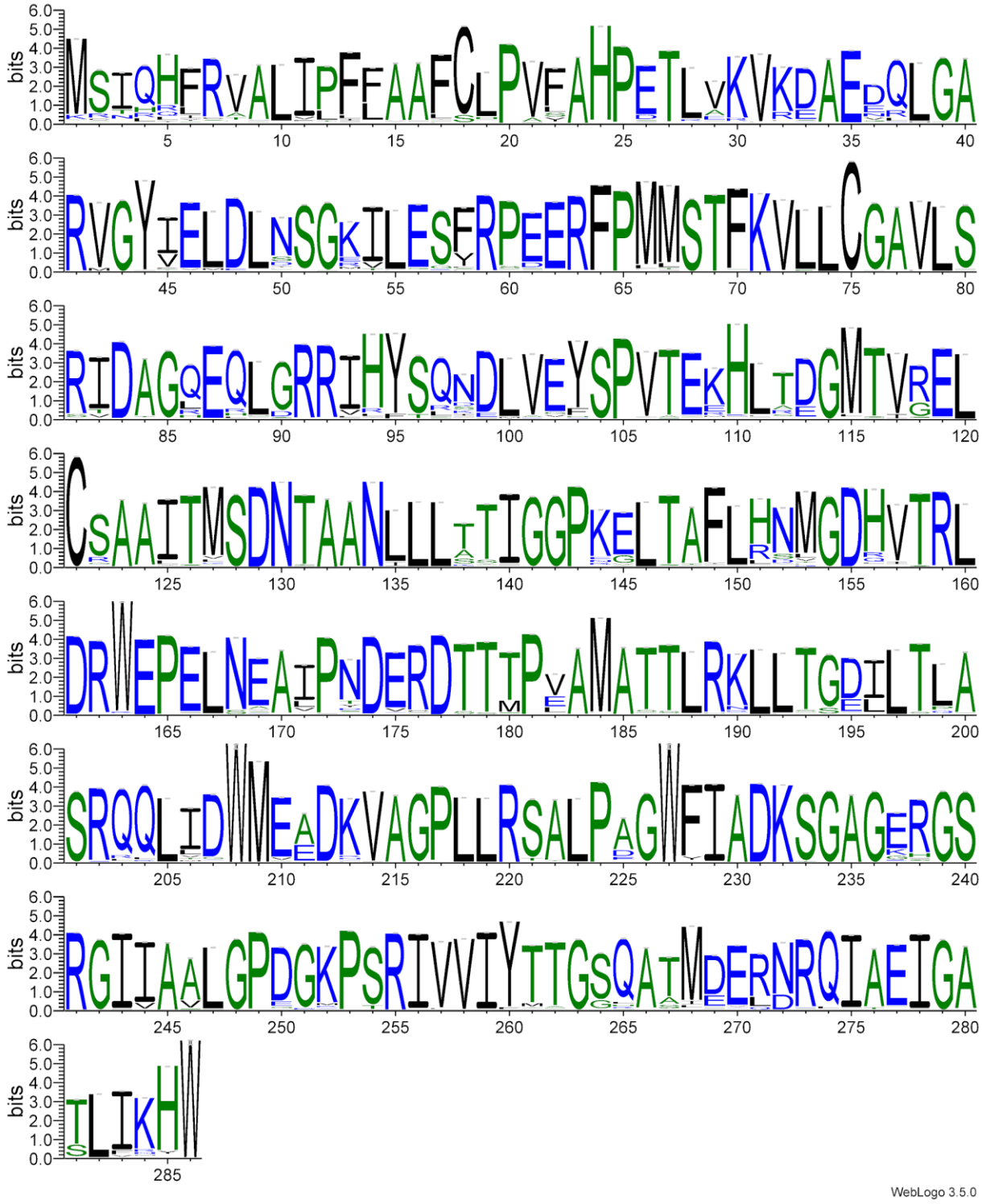


Figure 2.4.6 Logo representation of the conservation in the amino acid sequence after 12 generations of molecular evolution.

A periodic alternating pattern of high to low entropy can be seen in the long alpha helices H1, H9, H10 and H12 (**Figure 2.4.3**). This reflects the nature of the two halves of the helices, one being hydrophilic partially exposed to the solvent, the other containing hydrophobic residues packed

against the protein core.
H2 is different from the other helices since it is located deep inside the hydrophobic core of the protein and mediates most of the hydrophobic interactions of the protein. This parallels the lower mutation frequency and entropy observed in all our libraries (**Figure 2.3.5**, **Figure 2.3.14**, **Figure 2.4.2**, **Figure 2.4.3**), since mutations in the hydrophobic core have a high chance to damage the fold and thus impair the function of the protein.
The correlation (spearman correlation: rho 0.53, p < 1e-15) observed between the mean crystallographic B factor of residues in the reference structure and the information entropy retrieved from the evolved library is also noteworthy (**Figure 2.4.7**).

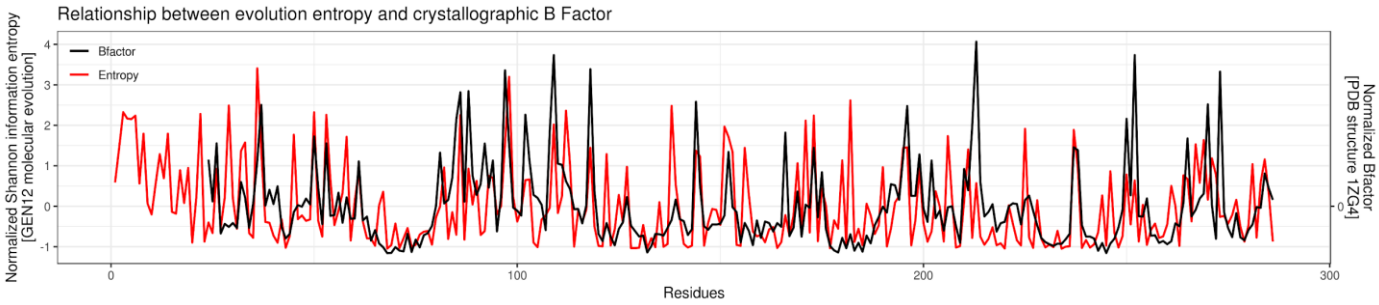


Figure 2.4.7 Molecular evolution entropies compared to the crystallographic B factors. Relationship between the entropy of the residues obtained in molecular evolution and the mean B factor of the residues observed in the reference structure 1ZG4.

This correlation likely reflects the tendency of residues that are part of ordered structures to be averse to mutation.
While the mutational landscape of TEM-1 beta lactamase covers a broad range of substitutions, four mutations in particular became by genetic drift more frequent than the original sequence in the last generation of molecular evolution: M180T, E195D, L196I, S281T (**Figure 2.4.2**). Among these M180T, that corresponds to M182T in the standard numbering scheme of class A beta lactamases (ABL) (Ambler et al., 1991), is a well-documented mutation known to contribute to the protein stability and found very commonly both in natural variants (Huang & Palzkill, 1997; X. Wang et al., 2002) and in mutagenesis experiments (Goldsmith & Tawfik, 2009). E195D and L196I (E197D and L198I in ABL) are mutations in the H8/H9 turn which are commonly found during mutagenesis (Salverda et al., 2010). Significantly, D197 is the consensus amino acid for this position (197) in the original alignment of class A beta lactamase (Ambler et al., 1991).
I used principal component analysis (PCA) on the Shannon entropies associated to every position of each dataset, to evaluate the evolution of the mutagenized libraries towards the natural diversity (**Figure 2.4.8**).

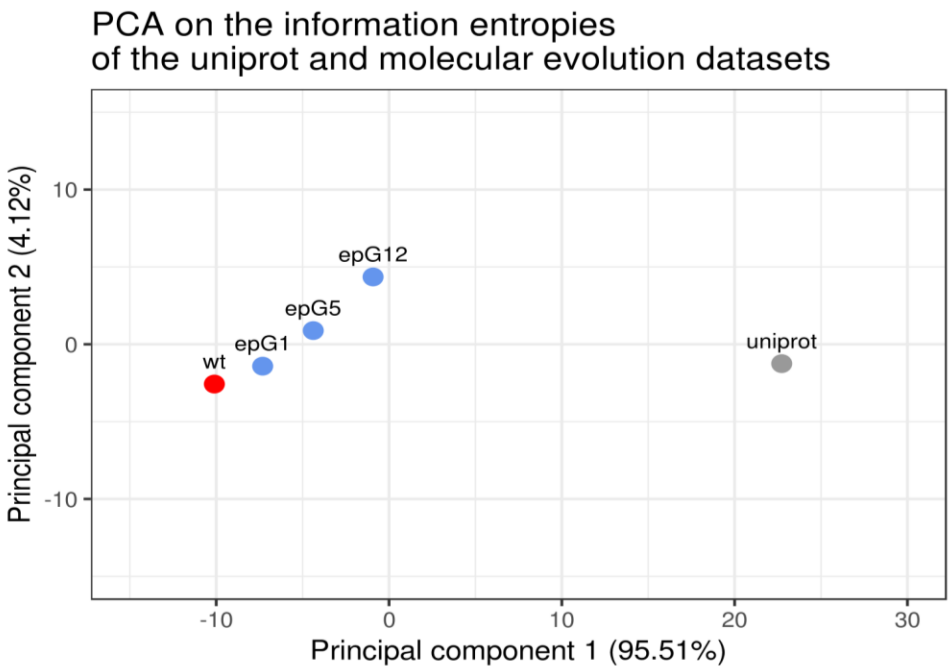


Figure 2.4.8 Principal component analysis applied to the UniProt and molecular evolution dataset.
Each point represents a dataset. Shannon information entropy was calculated for each position of each dataset and then subjected to PCA. Euclidean distance was used as distance metric. Grey represent the UniProt dataset, cyan the molecular evolution libraries (ep: error prone PCR). Numbers in the labels above the data points indicate the molecular evolution generation. The original pUC19 TEM-1 beta lactamase (wt: “wild type”, in red) was added as additional datapoint before the analysis as a zero vector. The percentage of variance (POV) of the component is shown in brackets on the axis label.

I also applied PCA (B. Wang & Kennedy, 2014) (**Figure 2.4.9**) and t-SNE (**Figure 2.4.10**) on the sequences themselves, to evaluate the degree of dispersion for each generation in comparison to the natural variants.

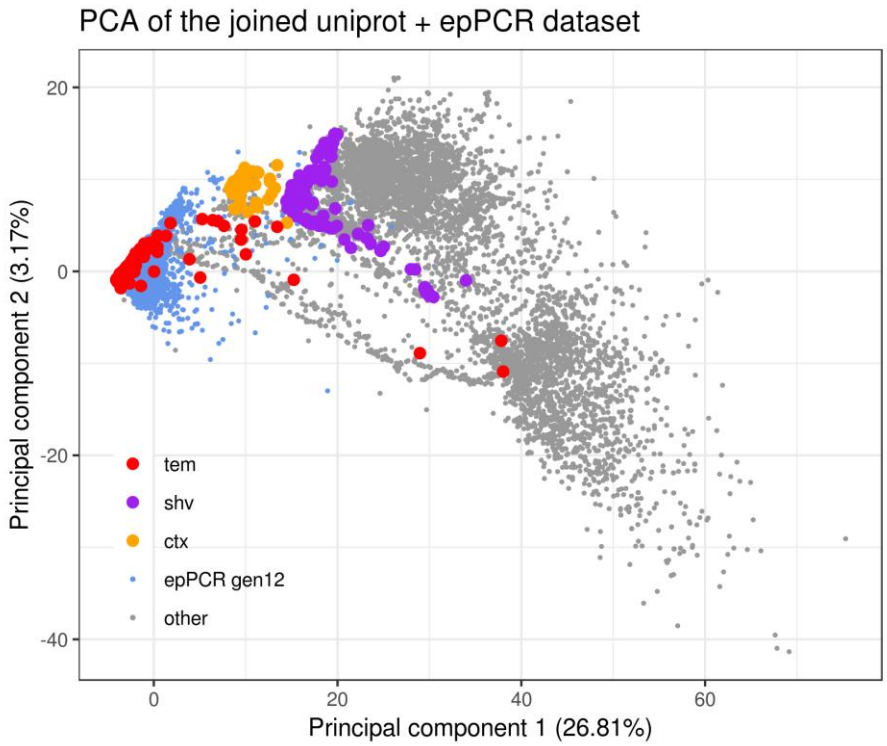


Figure 2.4.9 Principal component analysis applied to the joined UniProt / error prone PCR 12th generation library dataset. Each point represents a sequence of the joined dataset. Each amino acid in the sequence was encoded as the frequency of that amino acid in that position in the entire joined dataset. The frequency value of each position was ranked and then subjected to PCA. Euclidean distance between ranks was used as distance metric. Grey and cyan represent the original dataset (grey UniProt, cyan epPCR library). Overlayed on top, the UniProt sequences' membership to one of the three main families of type A beta lactamases retrieved from the corresponding UniProt annotation are displayed in bright colours. The original pUC19 beta lactamase before molecular evolution is classified as a TEM beta lactamase (red). The percentage of variance (POV) of the component is shown in brackets on the axis label.

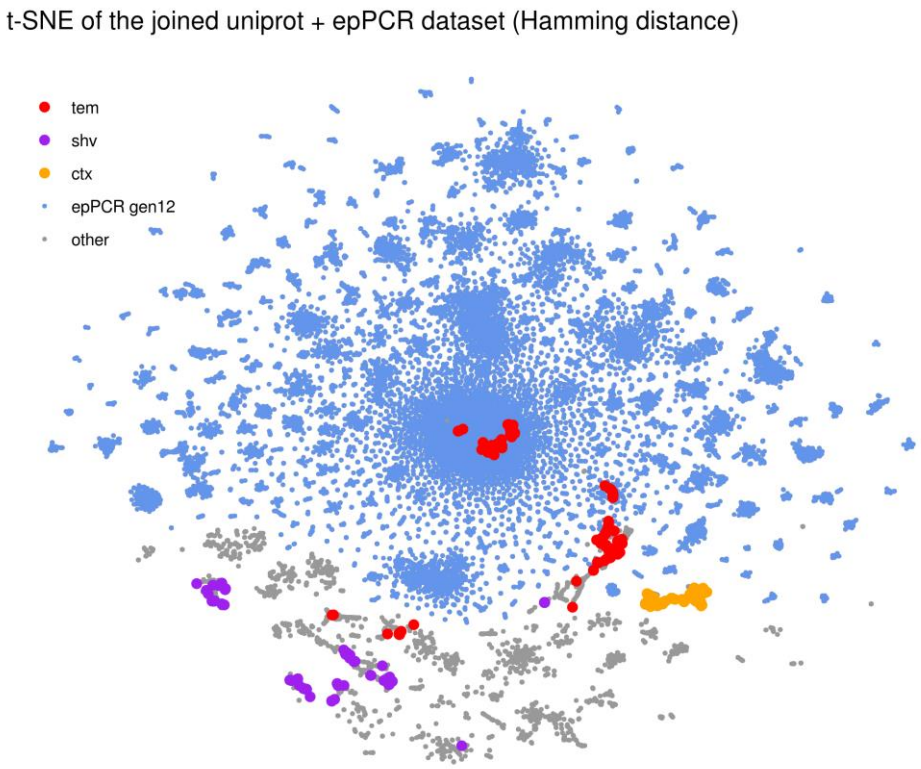


Figure 2.4.10 t-SNE dimensionality reduction applied to the joined UniProt / error prone PCR 12th generation library dataset. Hamming distance between sequences was used as the distance metric. Grey and cyan represent the original dataset (grey UniProt, cyan epPCR library). Overlaid on top, the UniProt sequence membership to one of the three main families of type A beta lactamases retrieved from the corresponding UniProt annotations are displayed in bright colours. The original pUC19 beta lactamase before molecular evolution is classified as a TEM beta lactamase (red)

These analyses suggest that the subsequent mutagenesis cycles consistently evolve the sequences in a concerted direction that is similar to that observed in the natural dataset. t-SNE also suggest that the cluster of the evolved lactamase is only an extension of the TEM family and does not represent other members of class A beta lactamase (**Figure 2.4.10**). Thus, the molecular evolution libraries describe the mutational space of a specific protein and not of a protein family. From this analysis we may conclude that the library has retained most of the characteristics of a collection of natural beta lactamase variants and represents only the mutational landscape around the protein of interest. This means that the library provides a picture of the early stages of the evolution of a protein, neither too similar nor too diverse from the original version, but exploring the landscape of mutational substitutions in a direction dictated by natural selection.

2.4.3 The predominant Direct Coupling Analysis predictions are short range interactions where the co-evolution effect is stronger

The longest open reading frame was extracted from each of the 150.000 circular consensus reads obtained after sequencing the last generation of mutagenesis and the proteins shorter than the archetype were removed. I built a multiple sequence alignment (MSA) from the remaining 106487 (68.9%) translated peptides and kept only the original 286 positions related to the wild type enzyme. To predict which residue pairs interact, I employed a custom implementation of DCA that applies a pseudo-likelihood approximation to this MSA as well as to the MSA obtained similarly from the other two sequenced generations of mutagenesis.
The 286 residue pairs (0.72% of the total possible contacts) which showed the highest DCA score and were more than five residues apart in the MSA were compared to the contact map of the reference structure (**Figure 2.4.11**).

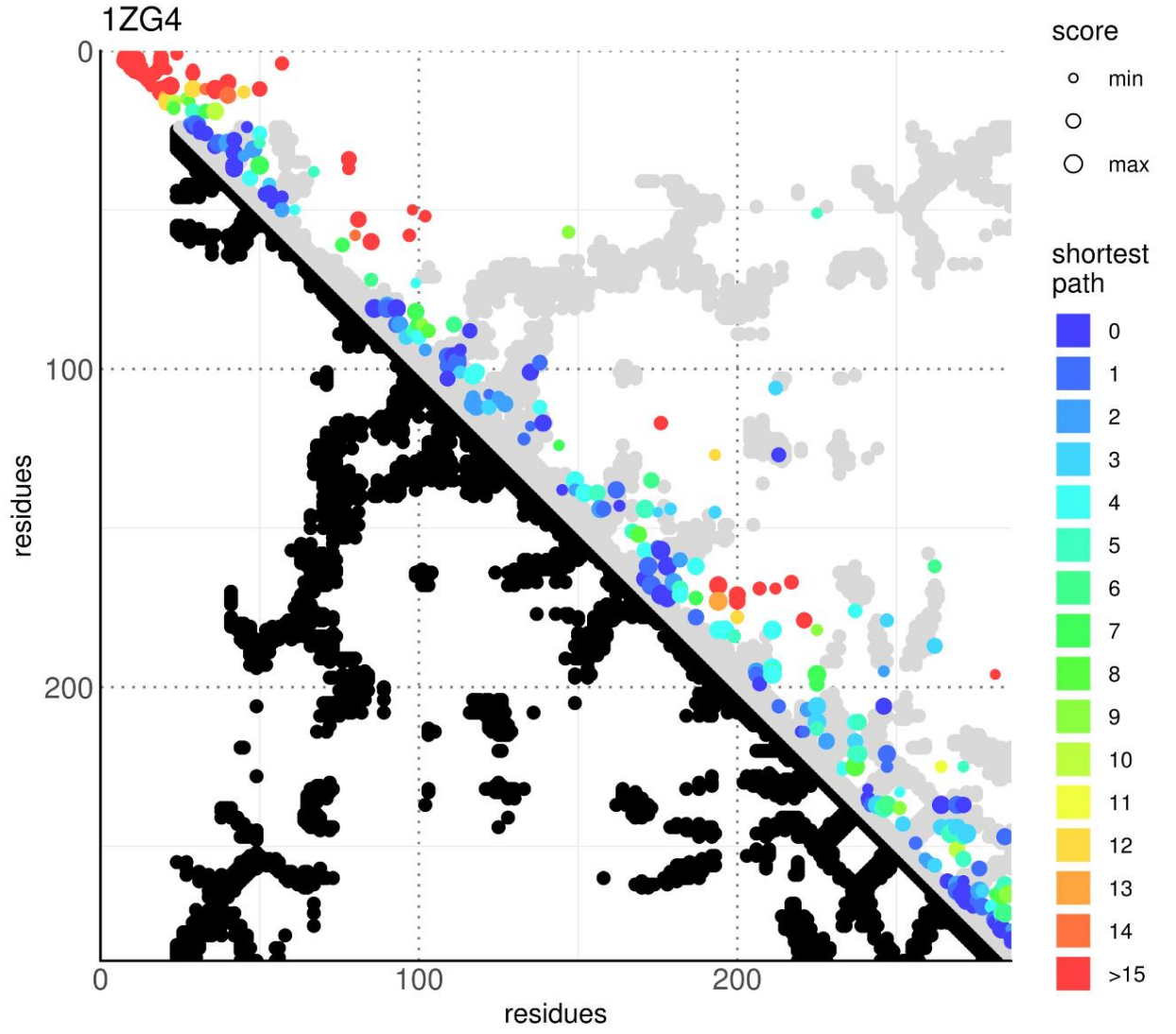


Figure 2.4.11 DCA plot of the twelfth generation library.
DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact

predictions by DCA obtained from the twelfth generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph, the dot size indicates the ranking of the prediction score while the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

Despite removal of the predicted contacts where the position pairs were too close to each other in the MSA, there is an enrichment of proximal interactions (near the diagonal of the contact map), at the expense of long range contacts. This was a good indicator that the analysis was extracting co-evolving positions since proximal residues using standard evolutionary dataset are typically associated with a high DCA score. In general, the predicted contact distribution was non-random and contacts tended to crowd at both extremities of the helices ignoring highly conserved areas like H2 (residues 67-83). Other minor crowding could be observed around two big looping regions (90-100 and 160-170). N-terminal crowding is likely due to the degeneration and duplication around the starting site that was already observed during Sanger sequencing, while the C-terminal density is probably due to sequential mutated positions in sequences where a C-terminal frameshift creates a block of strongly correlated positions without significantly affecting the functionality of the protein. The propensity to avoid conserved areas like H2 instead reflects the difficulty in creating a robust prediction when observing an inadequate number of mutations. We thus faced an interesting problem: the more contacts a residue mediates the more harmful a mutation becomes and thus we will observe a limited set of variations. However, since the mutational space at each position is tightly linked to the prediction power, the contacts formed by the most important residues are also the ones harder to predict.

2.4.4 Improving the prediction power in key areas and retrieval of long range interactions

To improve the accuracy and the spread of the predictions I applied a correlation-based approach identical to that proposed for fitness by Schmiedel and Lehner (Schmiedel & Lehner, 2019). Residues in structural proximity are often deeply interconnected and likely to share the same environment, consequently producing similar interaction patterns with other positions. Exploiting this similarity, I could obtain interactions from conserved positions by calculating partial correlation of the protein positions on the DCA patterns, because highly interconnected positions will have a

characteristic association pattern across the protein easily recognizable by partial correlation, even if the original DCA predictions are fairly inaccurate.
To validate this approach, I calculated the partial correlation with the UniProt dataset (**Figure 2.4.12A**). As expected, the predictions from partial correlation are very similar to the coupling score obtained by DCA (**Figure 2.4.12B**), and are in general less densely packed around the diagonal albeit showing a few more incorrect predictions.

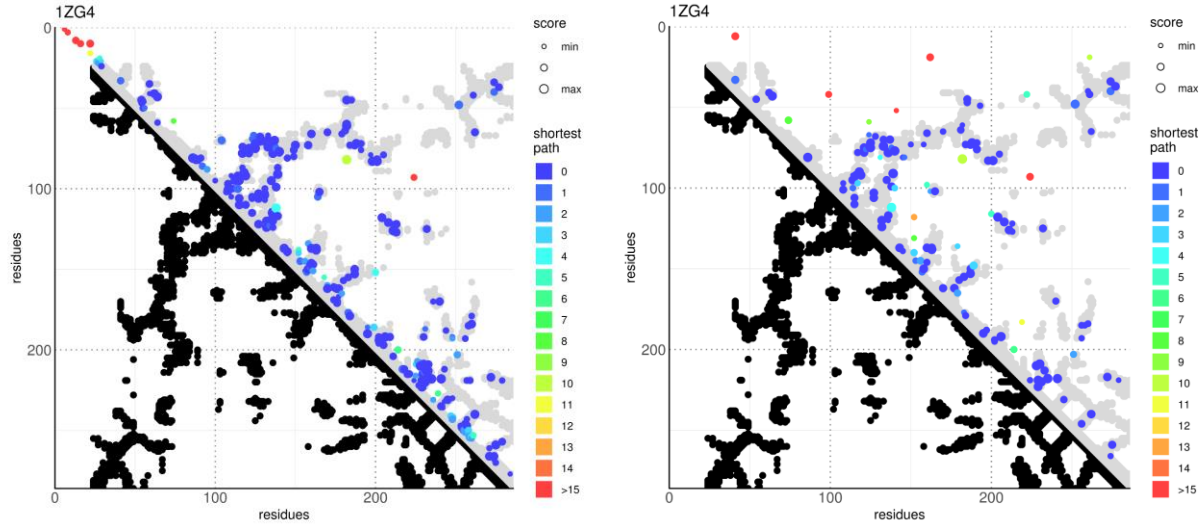


Figure 2.4.12 DCA plots of the dataset of the "natural variants" obtained from UniProt. A) DCA plot showing the top L (L = 286, the length of the protein amino acid chain) contact predictions by DCA obtained from the UniProt dataset. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph, the dot size indicates the ranking of the prediction score while the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure). B) Partial correlation of the UniProt dataset. Plot of the top L/2 partial correlations of residue positions on DCA score obtained from the UniProt dataset.

The partial correlation approach applied to the molecular evolution dataset (**Figure 2.4.13**) gave instead very different results compared to the original coupling score (**Figure 2.4.11**) and resulted more similar to what observed in the UniProt dataset, where the predicted interactions were more broadly distributed and both the terminals and the diagonal were far less crowded (**Figure 2.4.12**).

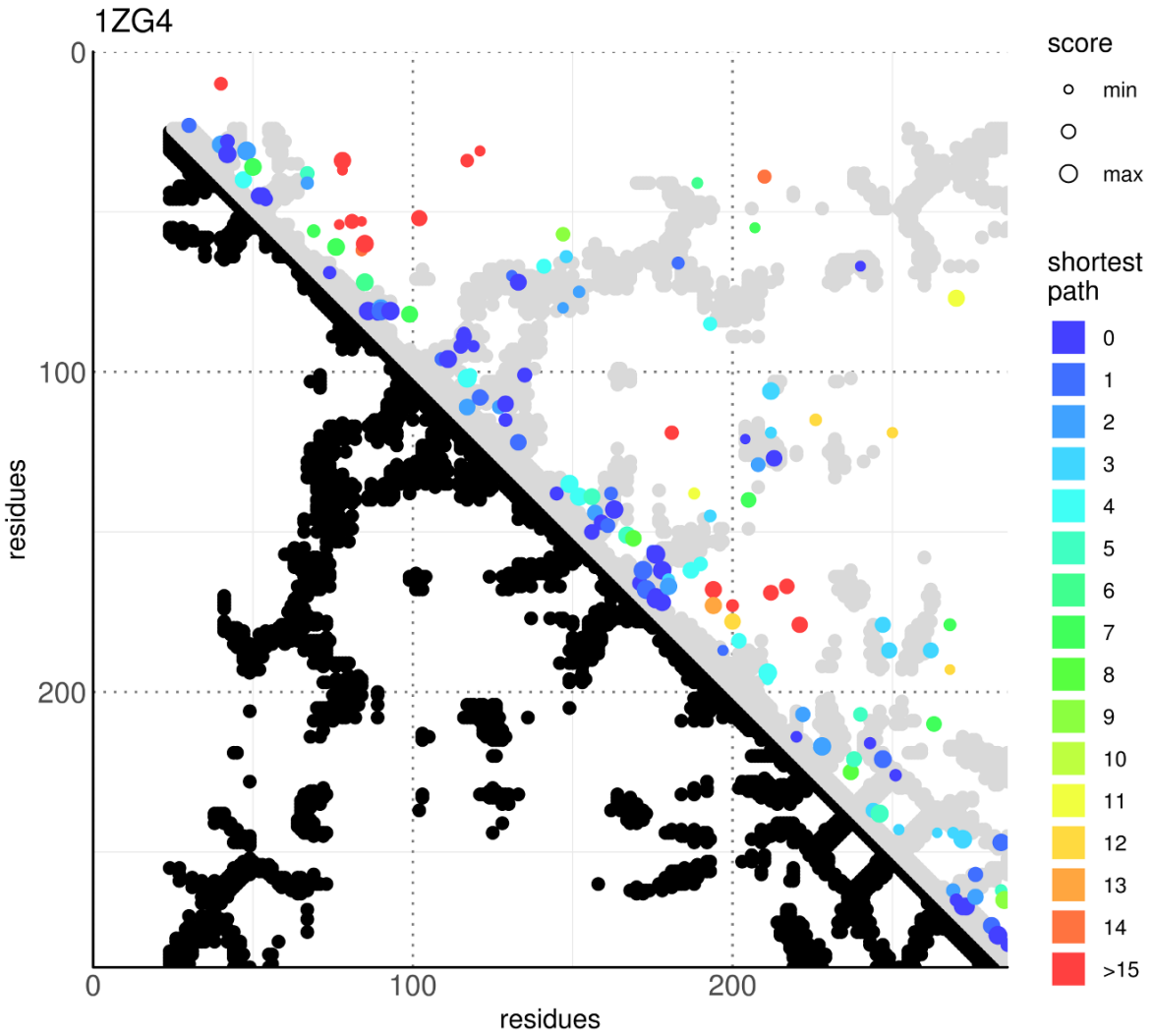


Figure 2.4.13 Partial correlation of the DCA score of the twelfth generation library. Plot of the top L/2 partial correlations of residue positions on DCA score obtained from the 12th generation of molecular evolution. The graph is an LxL grid where each axis represents the amino acid positions of the lactamase chain, from the N- to C-terminals. Each point represents the pair of residues described by its coordinates. The graph is separated in two halves. In the lower half black dots represent pairs of residues that have at least a pair of their respective non-hydrogen atoms less than 8.5 Å apart in the reference crystallographic structure (PDB id: 1ZG4). These positions are considered residues in contact with each other. In the upper half the top L DCA predictions from the molecular evolution dataset are plotted above the grey mirrored silhouette of the crystallographic contacts. Pairs where the respective residues are less than 5 positions apart in the lactamase alignment are excluded from this ranking to promote visualization of long range interactions. In the graph, the dot size indicates the ranking of the prediction score while the colour indicates the shortest path (as the lowest L1 norm in the graph grid space) connecting the point to a contact pair position (a pair of residues that have non-hydrogen atoms less than 8.5 Å apart in the reference structure).

The accuracy of the prediction was relatively low (**Figure 2.4.14A**), even if several times bigger

than the random expectation. However, this inaccuracy was caused by low precision and not by a low trueness to the underlying values as proven by the low value of shortest path from a true contact observed for the predicted pairs (**Figure 2.4.14B**).

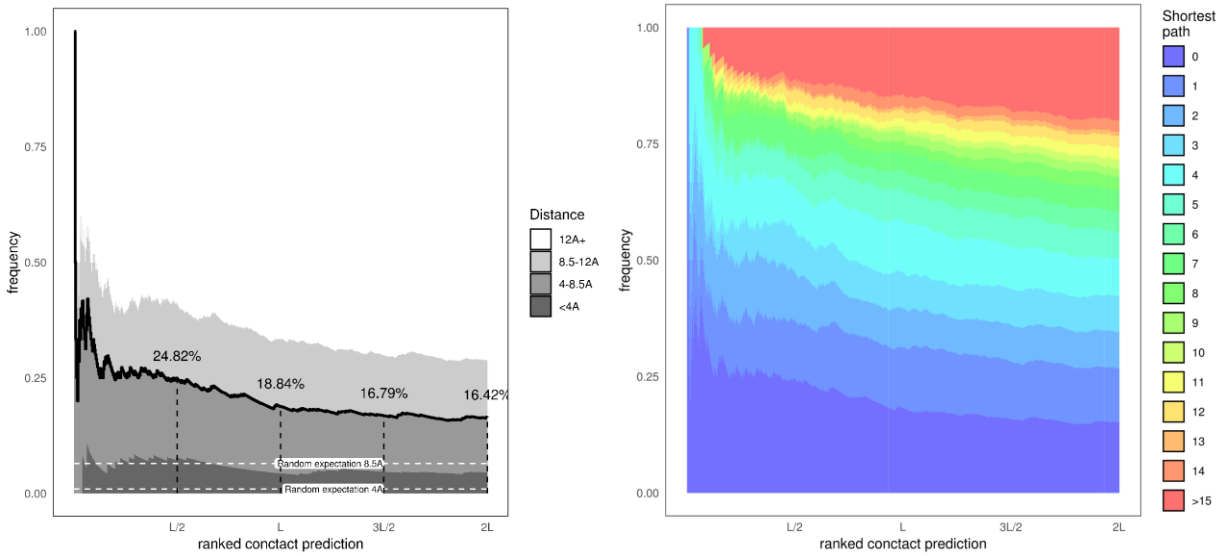


Figure 2.4.14 Accuracy of the partial correlation prediction.
A) Partial correlations between the positions of the DCA score at the 12th generation of molecular evolution sorted by their value. From the top: the first x elements were extracted, x increasing along the X axis. The graph represents the fraction of residue pairs that have atoms less than 4, 8.5 or 12 Å apart in the reference structure. Dotted white lines represent the frequency of position pairs expected to be under 4 and 8.5 Å apart in a random sampling (random expectation). B) Partial correlations at the 12th generation of molecular evolution. The correlations were sorted and the top elements extracted as described in panel E. The graph represents the frequency of shortest path distances from a true contact (non-hydrogen atoms that are less than 8.5 Å apart in the reference structure) observed in the contact pairs of these samples.

Along the contact map diagonal there were several instances of densities in correspondence to strong secondary structure interactions, like the proximity between N-terminal sheets and the H1 helix represented in the graph with the cluster of points found near residues 25 to 60. Other subdiagonal crowding (around residues 90-170) could be observed in the helical domain in correspondence to the interactions formed by the bending of the peptide chain in a turn. These and similar interactions, formed between the C-terminal half of the five stranded sheet and the C-terminal helices of the protein (200-285), could also be seen in the original DCA score (**Figure 2.4.11**) and were the most evident areas along the diagonal of the UniProt dataset where the predicted interactions clustered (**Figure 2.4.12**). Long range contacts, represented in the contact map by data points in regions far from the diagonal, were significantly different when the interactions obtained by partial correlation were compared to those predicted by the original DCA score. Partial correlation prediction showed several off-diagonal prediction points, mainly associated with highly interconnected regions or between elements of the hydrophobic core. In particular, several interactions of the H2 helix (67-83) with other elements of the helical domain

(H2 to H10, residues 65-210) were found, demonstrating the centrality of the helix, even though the region *per se* is characterized by a very small mutational landscape (**Figure 2.4.2**, **Figure 2.4.3**). The analysis identified also another cluster in the helical domain describing the proximity of helix H10 (199-210) to helix H5 (117-126).
Overall, with this correction I was able to obtain a predicted contact map that matches effectively the contact map of the reference crystal structure without any prior structural information. Our analysis demonstrated the possibility to obtain evolutionary couplings from a collection of sequences evolved in vitro. DCA highlighted the strongest evolutionary signal of proximal interactions (around the diagonal of the contact map) while partial correlation extracted information on the centrality of the H2 helix and the relations between secondary structure elements. These results demonstrate that molecular evolution can be used as an easy and powerful tool for structural analysis, by compressing the millions of years of natural selection into the couple of months of in vitro mutagenesis and selection.

3 Discussion

The idea of linking protein structure to the amino acids covariation is an old concept that took root in the late 90s (Thomas et al., 1996). Yet the study of evolutionary couplings is still an emerging frontier of bioinformatics, able to retrieve the network of interactions that dictate protein fold and function (Ekeberg et al., 2013; Kamisetty et al., 2013; Marks et al., 2011; Morcos et al., 2011; Ovchinnikov et al., 2014, 2017; Weigt et al., 2009). The innovation brought by the more recent implementation of this technique is the ability to produce structural information without the need of experimental structure determination, relying only on the traces left by evolution on protein sequence. The correlations are obtained from the continuous polishing process that the flow of time exerts on sequence to optimize/retain function. This makes any structural information retrieved by the analysis like a fossil imprint of an in vivo interaction.

The current computational techniques based on evolutionary couplings require thousands of sequences to provide statistically meaningful results (Ekeberg et al., 2013; Marks et al., 2012; Morcos et al., 2011). Thus, current evolutionary coupling methods are limited to ancient and universal protein families, for which sequence data are available across a huge variety of species. This is a major limitation: a large number of human proteins, for instance, do not have ancient phylogenetic origin (Lander et al., 2001). They are therefore not amenable to evolutionary coupling methods based on phylogenetic databases and can only be tackled by experimental approaches. Another advantage of mutational libraries with respect to the classical phylogenetic data is the representation of a sequence instead of a family, since the Markovian models that retrieve the sequences for the alignments in the standard analysis do not differentiate close paralogs from true orthologs. This poses a serious limitation for protein families rich in paralogs like globulins, for which it is nearly impossible to obtain information for a specific member of the family. The ability to represent a protein instead of a family is a new feature that can enable to distinguish a different level of details during the biological interpretation of the data.

In this thesis, I presented a strategy that was published under the name CAMELS (Coupling Analysis by Molecular Evolution Library Sequencing) that overcomes this limitation (Fantini et al., 2019). CAMELS lays the bases to develop a general method to gather structural information on protein contacts without performing experimental structural studies or the need for thousands of natural variants of the target protein across natural evolution. In this work I provided a unique pipeline from the molecular to the computational levels using the most advanced techniques and solved a number of crucial technical problems. Because DCA is good at capturing compensating mutations, a high mutational load in the collection of functional sequence variants is recommended. When a single harmful mutation appears in the sequence, the protein will likely not be functional and bacteria that carry that specific variant will die. However, if a second mutation able to compensate the damage is also present in the sequence, the function of the protein can be restored and the host cell survives. At the time that the selection is introduced, both mutations must already be present in the sequence, hence the more mutations are inserted in each round of mutagenesis, the better. This coincidence was favoured by increasing the selective pressure in the generations that were going to be sequenced. This way, if a single mutation is harmful but still barely allows survival in a low selective pressure, there will still be

few generations in which the second compensating mutation could occur before the strong selection of the last generation reaps all the sequences carrying mutations that are not compensated.
The CAMELS method is based on the power of phenotypic selection. I produced one of the largest and most diversified molecular evolution libraries that shows high single molecule sequencing quality and sequence divergence of nearly 10% (i.e. 25 amino acid mutations and around 55 mismatching nucleobases) from the original ancestral protein. It is also the first library to have been sequenced at the full-length protein level by third NGS. Other databases of TEM-1 mutagenic variants are available, some of which were created using epPCR (Jacquier et al., 2013) or deep mutational scanning (Firnberg et al., 2014) as the mutagenic mechanism. These public datasets cannot be used to infer structural information because they are mostly composed of single amino acid variants and thus cannot generate evolutionary coupling. The sequencing reads do not always cover the full-length molecule thus losing long range information (Firnberg et al., 2014). The method described in this work is different, since it produces a deep, high quality and full-length sequencing of a prolonged selection-driven evolution of TEM-1 lactamase instead of focusing on the effects of single mutations.
The libraries produced in this project were used to obtain structural information, by creating sequence diversity through mutation and analysis of artificial evolutionary couplings. The predicted contact map matches successfully that of the reference crystal structure even though at the cost of a bias towards short and medium range contacts. These results show that the described pipeline is effectively simulating the course of evolution, even if it is not entirely compressing the millions of years of natural selection into the couple of months of in vitro mutagenesis and selection. It was nevertheless possible to successfully follow the early stages of the landscape exploration of the evolving protein and use this to extract direct information about protein folding.
Pilot work on structure prediction from molecular evolution experiments have been published during the development of this project (Figliuzzi et al., 2016; Rollins et al., 2019; Schmiedel & Lehner, 2019). CAMELS offers several advantages as compared to these methods. The strategies previously used can only be applied to proteins strictly under 200 amino acids and can thus be used solely on a small fraction of the proteome from all three domains of life (J. Zhang, 2000). In particular, crucial for the success of the method described in this thesis is the growth of the library in a matrix of a semisolid medium, which allows local growth but prevents diffusion. Importantly, the use of third generation sequencing is a strong advantage that can be used to easily overcome the sequence read length limitations of traditional sequencing platforms. A key advantage of CAMELS is the absence of protein length constraints, since both the mutagenesis strategy and the sequencing allow processing of proteins of any length.
Another limit of previous techniques is the impossibility to exhaustively sampling all double mutants in the limited libraries that can be screened in complex systems like human tissue cultures. Structure determination with the previous methods would only be achievable if the libraries were biased to massively reduce diversity. Therefore, previous strategies are best suited to small proteins, or protein systems where the directed evolution strategy can handle large libraries, such as in the case of the GB1 domain (Olson et al., 2014). CAMELS employs instead multiple rounds of mutation enrichment to compress the variability in a library of a few hundred thousand elements. This solves the problem of limited library diversity and has the

potential to be of practical value to investigations of moderately sized proteins in systems where the library size is an important constraint.
The CAMELS method is in principle generalizable: by generating hundreds of thousands of mutagenic functional variants, it permits to focus on any protein and builds the foundation for a targeted structural analysis. This may allow the investigation by DCA-like methods of evolutionary younger proteins, like eukaryotic-only or vertebrate-only proteins or human proteins of neurobiological interest, ultimately solving species-specific questions that need species-specific answers.
What are the current limits of CAMELS? The most important one is the inability to fully reconstruct the protein structure in the current proof of concept formulation of the technology. Nevertheless, CAMELS provide local and long-range contacts. Improvements might be envisaged to overcome this limitation in future work. The main factors that could allow a complete structure determination are likely the number and distribution of mutations, the sequencing depth and the strength of the selective pressure. All these parameters can be easily scaled up, hopefully, as follow up of the present proof-of-concept study. The data revealed a strong correlation between the mean crystallographic B factor of residues in the reference structure and the information entropy retrieved from the evolved library. This happens because residues that are part of ordered regions as in the protein core are averse to mutation and thus the most conserved ones. Harming these zones would affect the fold and function. A drastic increase in the number of mutations would help to generate variability in these conserved key residues that could be translated in better evolutionary couplings. The same logic applies by forcing a distortion in the mutation propensities to favour the generation of mutations in key areas. The problem of this approach is that altering the propensities can distort the landscape obtained by unbiased epPCR. Modifying the driving force of the mutagenesis from epPCR to deep mutational scanning could provide a diverse landscape that might produce more precise couplings for the reasons mentioned. A critical comparison between the landscapes and couplings produced by epPCR and deep mutational scanning may be important to improve the technique in future implementations. Increasing the sequencing depth and the corresponding scale of selection is also an easy albeit currently expensive solution. This would likely increase the statistical power and allow low-entropy regions to show enough variation to be translated in couplings.
The fifth generation produced more long-range contacts in the standard DCA prediction in respect to the twelfth. This was something unexpected and rather puzzling. A possible explanation could be that the fifth generation produced nearly twice the number of reads of the twelve generation that in turn could alter the number of mutations in key areas. Another interpretation is that accumulation of mutations over the generations could create a broad compensatory effect that partially hinders the results. The correlated mutations accumulating throughout evolution are a mixture consisting of directly and indirectly interacting mutations. Since most random mutations are destabilizing, under the severe mutational load exerted on TEM-1 by selective pressure, most of the variants in the library should have a compromised stability. This strong selection pressure towards fixation of stabilizing mutations might thus favour the accumulation of correlated mutations of residues that do not directly interact. A more systematic investigation of this point using different target proteins will be important in the future.
An important requirement of CAMELS is phenotypic selection, a necessity that makes the

method truly evolutionary: like in the natural environment, selection is always based on the target protein function. As a consequence, selection must be designed on a case-by-case basis. For some proteins (such as TEM-1 beta lactamase), a phenotypic selection scheme is readily designed. More in general, selection schemes based on interactions could be considered that is probably the best approach to generalize the method. CAMELS could easily be modified, for instance, for the study of protein-protein interactions, exploiting selection schemes for interacting proteins coupled to SMRT sequencing, which would allow observation of protein pairs in a single sequencing read. Selection schemes based on signalling by the mutated target protein could also be envisaged.

The next obvious step will be to exploit standard and generic selection methods that rely on the folding and binding properties of the mutant proteins in the library, regardless of their functional activity. In our laboratory we have, for instance, already planned to use selection schemes to select for interacting partners, using a strategy we already pioneered for screening more stable antibodies against a given target (Chirichella et al., 2017; Visintin et al., 1999). CAMELS could be applied to two covariant interacting proteins, which could then be co-selected by a two-hybrid scheme for preserving their mutual binding. This strategy will provide information on the direct or indirect structural determinants for protein-protein interacting domains. This would be a revolutionary breakthrough that is not restricted to specific cases. It should be noted however that to successfully apply CAMELS, a good selection is critical. Although testing the function of a protein may seem a simple requirement on paper, the difficulty of developing an effective screening strategy cannot be underestimated especially for understudied proteins.

An elegant recent study successfully used deep mutagenesis to attempt determination of an unknown structure of a large complex human receptor in a physiologically relevant active conformation (Park et al., 2019). However, this example was somewhat limited to the specific structure of the protein. It will be interesting to apply CAMELS to members of the GPCR family, exploiting the signal transduction propriety of receptors coupled to a screenable selection readout.

Finally, one of the biggest obstacles to an in vitro evolution approach was the precarious equilibrium between mutagenic strength and selection survival rate. This issue was solved with a generational approach. I can further envisage future applications of the method to a continuous evolution in a specialized bioreactor. Overall, the CAMELS method provides a solid methodology that bypasses the most limiting factors of evolutionary coupling analysis techniques and opens a new page in structural biology and evolution.

4 Materials and methods

4.1 Plasmid construction & cloning

In order to easily clone in later steps, the mutagenized Amp^R, the backbone plasmid vector pUC19 (Norrander et al., 1983) (ATCC 37254) from ThermoFisher Scientific (SD0061) was modified to add flanking *XhoI* and *NheI* restriction sites to the already present Amp^R ORF to be able to easily clone in later steps the mutagenized Amp^R. To construct the plasmid, both the β-lactamase gene and the complementary plasmid vector fragments were amplified with oligonucleotides carrying the *XhoI* and *NheI* restriction sites (XhoI_bla_fw: tgaaaactcgaggaagagtATGAGTATTCA, NheI_bla_rv: acttgggctagctctgacagTTACCAATGC; NheI_backbone_fw: gtcagagctagccaagtttactcatatat, XhoI_backbone_rv: ctcttcctcgagttttcaatattattgaag), were then digested with the restriction enzymes and ligated with T4 ligase (**Figure 2.2.2**). The 5' restriction site was placed just behind the Shine-Dalgarno sequence and the ability to metabolize ampicillin was assessed by growth of *E. coli* carrying the plasmid in selective media. The new plasmid is named pUC19a. The Amp^R gene of pUC19 (GenBank: M77789.2) express a TEM-1 (class A) β-lactamase whose structure can be viewed in the 1ZG4 PDB entry.

4.2 Error prone PCR

Mutagenesis of the Amp^R gene was achieved with error prone PCR (Wilson & Keefe, 2001) in a mutation prone buffer with manganese ions, low magnesium, unbalanced dNTPs concentrations and a low fidelity DNA polymerase. Both low magnesium and the presence of manganese ions affect the efficiency of magnesium ions as cofactors of the polymerase by competition or by sheer low availability, while the unbalanced dNTP concentration favours mutations by scarcity of substrate and the deliberate usage of a low fidelity polymerase further increases the mutation rate. The reaction mix contained Tris pH 8.3 10 mM, KCl 50 mM, MgCl2 7 mM, dCTP 1 mM, dTTP 1 mM, dATP 0.2 mM, dGTP 0.2 mM, 5′ primer (bla_mut_fw: tgaaaactcgaggaagagtATG) 2 µM, 3′ primer (bla_mut_rv: acttgggctagctctgacagTTA) 2 µM, template DNA 20 pg/µl, MnCl2 0.5 mM (added just before reaction starts), Taq G2 DNA polymerase (Promega M784A) 0.05 U/µl (added just before reaction starts).
The error prone PCR was carried out in serial reactions of 4 cycles in 100 µL in the recommended supplier reaction conditions and with an annealing temperature of 62° C. In the first reaction tube, the DNA template was a gel purified *XhoI/NheI* digested β-lactamase fragment 20 pg/µl, while subsequent reactions were fed with 10 µL of the previous PCR product.

4.3 Library construction

The purification and digestion protocols before library construction changed slightly among generations. However, the optimized version of the pipeline employed in the last generations proceeded as follows: ~80 µL of the PCR reaction mixture underwent a cumulative amount of

20 cycles of error prone PCR, avoiding carrying over other reaction byproducts as much as possible. PCR was performed in standard reaction conditions to amplify the product and guarantee that the two strands of the amplicons did not contain mismatching base pairs. This step helped reducing the ambiguity in base calling during the circular consensus analysis. The purified PCR product was digested with *XhoI* and *NheI* restriction enzymes for 3h in CutSmart buffer (NEB). One hour before the end of the reaction, an appropriate amount of calf intestinal phosphatase (CIP) (NEB M0290S) was added following the supplier's instruction. Adding CIP during insert digestion strongly reduced the formation of insert concatemers, guaranteeing a single β-lactamase variant per plasmid. After gel purification to remove the CIP, the ligation between the fragment and the *XhoI/NheI* digested backbone of pUC19a was performed in a 1:1 insert:vector ratio. Formation of backbone concatemers was expected and unavoidable, but did not hinder the selection efficiency.

4.4 Selection

The ligated library was purified and then transformed by electroporation in ElectroMAX DH5α-E competent cells (Invitrogen #11319019). I employed ultralow gelling agarose (SeaPrep, Lonza #50302) 0.3% in Luria Broth (LB) medium with ampicillin 25 µg/mL to grow the bacteria (Elsaesser & Paysan, 2004; Fantini, Pandolfini, et al., 2017) obtaining between 0.4 and 3 million surviving colonies per litre. The bacterial growth in the fifth and twelfth generation was performed in LB medium with ampicillin 100 µg/mL to increase the stringency of the selection before the sequencing. After 40h growth, the bacterial pellet was retrieved by centrifugation 7500RPM at RT and the plasmids extracted by a maxi prep.

4.5 Sequencing

Construction of the libraries and sequencing on PacBio Sequel platform were carried out by Arizona Genomics Institute (AGI). After sequencing, the library was processed with the PacBio official analysis software SMRTlink to obtain the circular consensus (using ccs2) of the reads. In this step, the sequences where the consensus was built from less than 10 sequencing polymerase passes or when the predicted accuracy was less than 100 ppm (Phred 40) were filtered out from the dataset. The result was then mapped to the wild type β-lactamase *XhoI-NheI* digestion fragment of pUC19a with bowtie2 (Langmead & Salzberg, 2012) to retrieve the coding strand and the start site of the lactamase. After in silico translating the dataset, protein collection was further refined keeping only the elements coding a protein of 286 amino acids (as the wild type) and then aligned using MAFFT (http://mafft.cbrc.jp/alignment/software/) (Katoh, 2002) to construct the MSA. The 12th generation had issues with the *in silico* translation step caused by degeneration of the N-terminus as well as the starting site, resulting in a big amount of sequence with a premature termination codon. To circumvent the problem, the longest open reading frame, identified with a custom script, was considered the correct genetic sequence and the translated products were filtered to keep the sequences coding for proteins of at least the wild type length. This procedure was required to remove from the alignment bad quality reads, unrelated sequences and protein variants carrying a frameshift which would generate a strong

correlation noise between adjacent amino acid positions. It is interesting to notice that classical evolutionary data supplied to the algorithm do not have this problem, and thus this is a new issue brought by the mutagenic data. This is probably because frame shifted sequences do not carry a sufficient neutrality to the system to be retained throughout natural evolution, while in the small landscape generated by mutagenesis, every protein that satisfies the selection criteria of activity will be part of the collection. To compare our data to the natural occurring mutations of TEM beta lactamase, I created a reference dataset by running a small seed of TEM beta lactamases in Hmmer (Finn et al., 2011) on the UniProt database (https://www.uniprot.org). An alternative dataset that could be used as a control is the Pfam family of beta-lactamase2 (PF13354). Since the Hmmer dataset from UniProt is a collection of sequences that specifically matched the profile of the TEM family while Pfam family is a more general beta-lactamase collection I preferred the former as reference for our analysis.

4.6 Direct Coupling Analysis

The predicted contact pairs were obtained using a custom implementation (Fantini, Malinverni, et al., 2017) of the asymmetric version of the DCA (Morcos et al., 2011; Weigt et al., 2009) that applies
the Pseudo-likelihood method to infer the parameters of the Potts model (Balakrishnan et al., 2011; Ekeberg et al., 2013):

$$P(x) = \frac{1}{Z} e^{\sum_i^N h_i(Xi) + \sum_{i,j}^{N,N} J_{ij}(X_i, X_j)} \qquad (1)$$

where X is a sequence of the MSA and Z is the partition function.
Sequences were reweighed using an identity threshold that reflects the mutation rate of the generation analysed to remove parental inheritance (intended as “phylogenetic” bias created during mutagenesis) and sampling biases in the MSA. The first generation was too similar to the wild type to apply any sampling correction without unreasonably reducing the number of effectively non-redundant sequences (Morcos et al., 2011). The fifth generation used a 95% identity threshold and the twelfth generation a 90%. A standard L2 regularization was added following the original regularization described in Ekeberg et al., 2013 (Ekeberg et al., 2013) (λ = 0.01). The code used a scoring scheme for contacts where the DCA scores were computed as the Frobenius norm of the local coupling matrices of the Potts model. Dunn et al. average product correction (APC) was subtracted to remove background correlation (Dunn et al., 2008). The N top scoring contact predictions (N equals the MSA sequence length) were compared with the contact map of the reference structure (1ZG4) constructed considering two residues to be in contact if at least a pair of their respective heavy atom (non hydrogens) was less than 8.5 Å apart (Ekeberg et al., 2013). As it is standard practice, I removed predictions along the diagonal of the contact map if the residue pairs were less than five positions apart to promote enrichment of long-range predictions. Increasing this threshold to 8 or more did not change significantly the contact map prediction. I used the shortest-path (SP) distance (Malinverni et al., 2015) defined

as the L1 norm in the contact map lattice to join DCA predictions and the closest structural contact to visualize the agreement between predictions and empirical observations.

4.7 Partial Correlation

Partial correlation is a measure of the correlations between two variables after removing from both the possible correlations they might have with another set of confounding variables. In other words, partial correlation is the correlation between the residuals obtained from the regression of the variables of interest to the confounding variable set. In the present work, partial correlation is used to infer the correlation between each set of two rows of the DCA score matrix, removing the correlation these rows might have to all other rows of the matrix. Each row of the DCA matrix is a vector of the strength of association between the residue represented by the row and every other residues of the protein. Thus, the partial correlation of the DCA matrix represents the similarity of the profile of these vectors, considering and removing the correlation with the profile of all the other residues. To obtain the partial correlation matrix from the symmetrical DCA score matrix, I first set all the diagonal elements of the matrix to 1 and then approximated the partial correlation between rows with the pcor.shrink function of the corpcor R package. The package implements a James-Stein estimator for the covariance matrix. The details of the method are explained in (Schäfer & Strimmer, 2005) and in the manual of the package (https://cran.rproject.org/web/packages/corpcor/corpcor.pdf).

4.8 Other bioinformatic tools

Graphs generation was performed with R version 3.2.3 (2015-12-10). Poisson regression was performed with the fitdistrplus R package, while PCA was performed with the base R prcomp function. Mutation rates and Shannon information entropies were calculated with custom scripts. t-SNE was performed in MATLAB version R2018b using the Hamming distance as metric.

4.9 Dataset

All sequencing data that support the findings of this study have been deposited in the National Centre for Biotechnology Information Sequence Read Archive (SRA) and are freely accessible through the SRA accession code PRJNA528665
(http://www.ncbi.nlm.nih.gov/sra/PRJNA528665).

References:

Abdul-Masih, M. T., & Bessman, M. J. (1986). Biochemical studies on the mutagen, 6-N-hydroxylaminopurine. Synthesis of the deoxynucleoside triphosphate and its incorporation into DNA in vitro. *The Journal of Biological Chemistry*, *261*(5), 2020–2026.

Abraham, E. P., & Chain, E. (1940). An Enzyme from Bacteria able to Destroy Penicillin. *Nature*, *146*(3713), 837–837.

Abraham, E. P., Chain, E., Fletcher, C. M., Gardner, A. D., Heatley, N. G., Jennings, M. A., & Florey, H. W. (1941). Further Observations on Penicillin. *The Lancet*, *238*(6155), 177–189.

Altschuh, D., Lesk, A. M., Bloomer, A. C., & Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, *193*(4), 693–707.

Ambler, R. P. (1980). The structure of β-lactamases. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *289*(1036), 321–331.

Ambler, R. P., Coulson, A. F., Frère, J. M., Ghuysen, J. M., Joris, B., Forsman, M., Levesque, R. C., Tiraby, G., & Waley, S. G. (1991). A standard numbering scheme for the class A beta-lactamases. *The Biochemical Journal*, *276*(Pt 1), 269–270.

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, *176*(3), 663-675.e19.

Badran, A. H., & Liu, D. R. (2015). In vivo continuous directed evolution. *Current Opinion in Chemical Biology*, *24*, 1–10.

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., & Langmead, C. J. (2011). Learning generative models for protein fold families. *Proteins*, *79*(4), 1061–1078.

Beale, G. (1993). The discovery of mustard gas mutagenesis by Auerbach and Robson in 1941. *Genetics*, *134*(2), 393–399.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59.

Braswell, E. H., Knox, J. R., & Frère, J. M. (1986). The association behaviour of beta-lactamases. Sedimentation equilibrium studies in ammonium sulphate solutions. *The Biochemical Journal*, *237*(2), 511–517.

Breaker, R. R., Banerji, A., & Joyce', G. F. (1994). Continuous in Vitro Evolution of Bacteriophage RNA Polymerase Promoters? In *Biochemistry* (Vol. 33).

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, *33*(10), 3390–3400.

Brosius, J., Cate, R. L., & Perlmutter, A. P. (1982). Precise location of two promoters for the beta-lactamase gene of pBR322. S1 mapping of ribonucleic acid isolated from Escherichia coli or synthesized in vitro. *The Journal of Biological Chemistry*, *257*(15), 9205–9210.

Brown, D. M., Hewlins, M. J. E., & Schell, P. (1968). The tautomeric state of N(4)-hydroxy- and of N(4)-amino-cytosine derivatives. *Journal of the Chemical Society C: Organic*, 1925.

Bull, J. J., Badgett, M. R., W1chmaqf, H. A., Huehenbeck, J. P., Hillis, D. M., Gulati, L. A., Ho, C., & Molineuxtp, I. J. (1997). *Exceptional Convergent Evolution in a Virus*.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. *Journal of the American Statistical Association*, *88*(421), 364.

Bush, K. (1989a). Characterization of beta-lactamases. *Antimicrobial Agents and Chemotherapy*, *33*(3), 259–263.

Bush, K. (1989b). Classification of beta-lactamases: groups 1, 2a, 2b, and 2b'. *Antimicrobial Agents and Chemotherapy*, *33*(3), 264–270.
Bush, K. (1989c). Classification of beta-lactamases: groups 2c, 2d, 2e, 3, and 4. *Antimicrobial Agents and Chemotherapy*, *33*(3), 271–276.
Bush, K. (1997). Nomenclature of TEM beta-lactamases. *Journal of Antimicrobial Chemotherapy*, *39*(1), 1–3.
Bush, K. (2018). Past and Present Perspectives on β-Lactamases. *Antimicrobial Agents and Chemotherapy*, *62*(10).
Bush, K., & Jacoby, G. A. (2010). Updated Functional Classification of -Lactamases. *Antimicrobial Agents and Chemotherapy*, *54*(3), 969–976.
Bush, K., Jacoby, G. A., & Medeiros, A. A. (1995). A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrobial Agents and Chemotherapy*, *39*(6), 1211–1233.
Cadwell, R. C., & Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis. *Genome Research*, *2*(1), 28–33.
Camps, M., Naukkarinen, J., Johnson, B. P., & Loeb, L. A. (2003). Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proceedings of the National Academy of Sciences*, *100*(17), 9727–9732.
Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., & Depristo, M. A. (2012). *Pacific biosciences sequencing technology for genotyping and variation discovery in human data*.
Chen, J., Sahota, A., Stambrook, P. J., & Tischfield, J. A. (1991). Polymerase chain reaction amplification and sequence analysis of human mutant adenine phosphoribosyltransferase genes: The nature and frequency of errors caused by Taq DNA polymerase. In *Mutation Research* (Vol. 249).
Chen, K., & Arnold, F. H. (1993). Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences*, *90*(12), 5618–5622.
Cheng, K. C., Cahill, D. S., Kasai, H., Nishimura, S., & Loeb, L. A. (1992). 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. *The Journal of Biological Chemistry*, *267*(1), 166–172.
Chirichella, M., Lisi, S., Fantini, M., Goracci, M., Calvello, M., Brandi, R., Arisi, I., D'Onofrio, M., Di Primio, C., & Cattaneo, A. (2017). Post-translational selective intracellular silencing of acetylated proteins with de novo selected intrabodies. *Nature Methods*, *14*(3), 279–282.
Cover, T. M., & Thomas, J. A. (1991). Elements of Information Theory. In *Elements of Information Theory*.
Cunningham, B., & Wells, J. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, *244*(4908), 1081–1085.
Danel, F., Frère, J. M., & Livermore, D. M. (2001). Evidence of dimerisation among class D β-lactamases: Kinetics of OXA-14 β-lactamase. *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology*, *1546*(1), 132–142.
Datta, N., & Kontomichalou, P. (1965). Penicillinase Synthesis Controlled By Infectious R Factors In Enterobacteriaceae. *Nature*, *208*(5007), 239–241.
Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105.
Dunn, S. D., Wahl, L. M., & Gloor, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, *24*(3), 333–340.
Eadie, J. S., Conrad, M., Toorchen, D., & Topal, M. D. (1984). Mechanism of mutagenesis by O6-methylguanine. *Nature*, *308*(5955), 201–203.

Eckert, K. A., & Kunkel, T. A. (1990). High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Research*, *18*(13), 3739–3744.
Eckert, K. A., & Kunkel, T. A. (1991). DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.*, *1*(1), 17–24.
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., … Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*(5910), 133–138.
Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *87*(1).
Elsaesser, R., & Paysan, J. (2004). Liquid gel amplification of complex plasmid libraries. *BioTechniques*, *37*(2), 200–202.
Englund, P. T. (1971). Analysis of nucleotide sequences at 3' termini of duplex deoxyribonucleic acid with the use of the T4 deoxyribonucleic acid polymerase. *The Journal of Biological Chemistry*, *246*(10), 3269–3276.
Englund, P. T. (1972). The 3′-terminal nucleotide sequences of T7 DNA. *Journal of Molecular Biology*, *66*(2), 209–224.
Ennis, P. D., Zemmour, J., Salter, R. D., & Parham, P. (1990). Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: Frequency and nature of errors produced in amplification (polymorphism/gene families/recombination/histocompatibility). In *Proc. Nail. Acad. Sci. USA* (Vol. 87).
Esvelt, K. M., Carlson, J. C., & Liu, D. R. (2011). A system for the continuous directed evolution of biomolecules. *Nature*, *472*(7344), 499–503.
Fantini, M., Lisi, S., De Los Rios, P., Cattaneo, A., & Pastore, A. (2019). Protein Structural Information and Evolutionary Landscape by In Vitro Evolution. *Molecular Biology and Evolution*.
Fantini, M., Malinverni, D., De Los Rios, P., & Pastore, A. (2017). New techniques for ancient proteins: Direct coupling analysis applied on proteins involved in iron sulfur cluster biogenesis. *Frontiers in Molecular Biosciences*, *4*(JUN).
Fantini, M., Pandolfini, L., Lisi, S., Chirichella, M., Arisi, I., Terrigno, M., Goracci, M., Cremisi, F., & Cattaneo, A. (2017). Assessment of antibody library diversity through next generation sequencing and technical error compensation. *PLOS ONE*, *12*(5), e0177574.
Faria, N. R., Sabino, E. C., Nunes, M. R. T., Alcantara, L. C. J., Loman, N. J., & Pybus, O. G. (2016). Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, *8*(1), 97.
Fedurco, M. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*, *34*(3), e22–e22.
Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., & Weigt, M. (2016). Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular Biology and Evolution*, *33*(1).
Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2).
Firnberg, E., Labonte, J. W., Gray, J. J., & Ostermeier, M. (2014). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, *31*(6), 1581–1592.
Firnberg, E., & Ostermeier, M. (2012). PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS ONE*, *7*(12), e52031.
Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, *11*(8), 801–807.
Fujii, R., Kitaoka, M., & Hayashi, K. (2004). One-step random mutagenesis by error-prone rolling circle amplification. *Nucleic Acids Research*, *32*(19), e145–e145.

Göbel, U., Sander, C., Schneider, R., & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*, *18*(4), 309–317.
Goldsmith, M., & Tawfik, D. S. (2009). Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proceedings of the National Academy of Sciences*, *106*(15), 6197 LP – 6202.
Grossberger, D., & Clough, W. (1981). Incorporation into DNA of the base analog 2-aminopurine by the Epstein-Barr virus-induced DNA polymerase in vivo and in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(12 II), 7271–7275.
Hartman, E. C., & Tullman-Ercek, D. (2019). Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution. In *Current Opinion in Systems Biology* (Vol. 14, pp. 25–31). Elsevier Ltd.
Hestand, M. S., Van Houdt, J., Cristofoli, F., & Vermeesch, J. R. (2016). Polymerase specific error rates and profiles identified by single molecule sequencing. *Mutation Research*, 39–45.
Hillis, D. M., & Huelsenbeck, J. P. (1992). Signal, Noise, and Reliability in Molecular Phylogenetic Analyses. *Journal of Heredity*, *83*(3), 189–195.
Huang, W., & Palzkill, T. (1997). A natural polymorphism in β-lactamase is a global suppressor. *Proceedings of the National Academy of Sciences*, *94*(16), 8801–8806.
Hung, A., Thillet, J., & Pictet, R. (1989). In vivo selected promoter and ribosome binding site up-mutations: Demonstration that the Escherichia coli bla promoter and a Shine-Dalgarno region with low complementarity to the 16 S ribosomal RNA function in Bacillus subtilis. *Molecular and General Genetics MGG*, *219*(1–2), 129–136.
Husimi, Y. (1989). Selection and evolution of bacteriophages in cellstat. *Advances in Biophysics*, *25*, 1–43.
Hyman, E. D. (1988). A new method of sequencing DNA. *Analytical Biochemistry*, *174*(2), 423–436.
Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., Gros, P.-A., & Tenaillon, O. (2013). Capturing the mutational landscape of the beta-lactamase TEM-1. *Proceedings of the National Academy of Sciences*.
Jelsch, C., Mourey, L., Masson, J.-M., & Samama, J.-P. (1993). Crystal structure ofEscherichia coli TEM1 β-lactamase at 1.8 Å resolution. *Proteins: Structure, Function, and Genetics*, *16*(4), 364–383.
Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15674–15679.
Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.
Keohavong, P., & Thilly, W. G. (1989). Fidelity of DNA polymerases in DNA amplification. *Proceedings of the National Academy of Sciences*, *86*(23), 9253–9257.
Kirby, W. M. M. (1944). Extraction of a Highly Potent Penicillin Inactivator From Penicillin Resistant Staphylococci. *Science*, *99*(2579), 452–453.
Knott-Hunziker, V., Waley, S. G., Orlek, B. S., & Sammes, P. G. (1979). Penicillinase active sites: Labelling of serine-44 in β-lactamase I by 6β-bromopenicillanic acid. *FEBS Letters*, *99*(1), 59–61.
Kouchakdjian, M., Bodepudi, V., Shibutani, S., Eisenberg, M., Johnson, F., Grollman, A. P., & Patel, D. J. (1991). NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7H-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7H-dG(syn).cntdot.dA(anti) alignment at lesion site. *Biochemistry*, *30*(5), 1403–1412.
Kraft, F., & Kurth, I. (2019). Long-read sequencing in human genetics. *Medizinische Genetik*,

31(2), 198–204.
Lamotte-Brasseur, J., Knox, J., Kelly, J. A., Charlier, P., Fonzé, E., Dideberg, O., Frère, J.-M., Lamotte-brasseur, J., Charlieri, P., Fonze, E., & Frerei, J. (1994). The Structures and Catalytic Mechanisms of Active-Site Serine β-Lactamases. *Biotechnology and Genetic Engineering Reviews*, *12*(1), 189–230.
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.
Leung, D. (1989). A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique*.
Ling, L. L., Keohavong, P., Dias, C., & Thilly, W. G. (1991). Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and vent DNA polymerases. *Genome Research*, *1*(1), 63–69.
Long-McGie, J., Liu, A. D., & Schellenberger, V. (2000). Rapid in vivo evolution of a β-lactamase using phagemids. *Biotechnology and Bioengineering*, *68*(1), 121–125.
Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279.
Ma, J.-E., Jiang, H.-Y., Li, L.-M., Zhang, X.-J., Li, H.-M., Li, G.-Y., Mo, D.-Y., & Chen, J.-P. (2019). SMRT sequencing of the full-length transcriptome of the Sunda pangolin (Manis javanica). *Gene*, *692*, 208–216.
Maki, H., & Sekiguchi, M. (1992). MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis. *Nature*, *355*(6357), 273–275.
Malinverni, D., Marsili, S., Barducci, A., & de Los Rios, P. (2015). Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS Computational Biology*, *11*(6), 1–15.
Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., & Sander, C. (2011). *Protein 3D Structure Computed from Evolutionary Sequence Variation*. *6*(12).
Marks, D. S., Hopf, T. a, & Sander, C. (2012). Protein structure prediction from sequence variation. *Nature Biotechnology*, *30*(11), 1072–1080.
Mattila, P., Korpela, J., Tenkanen, T., & Pitkämem, K. (1991). Fidelity of DNA synthesis by the Thermococcus litoralis DNA polymerase—an extremely heat stable enzyme with proofreading activity. *Nucleic Acids Research*, *19*(18), 4967–4973.
Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560–564.
Mills, D. R., Peterson, R. L., & Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences*, *58*(1), 217–224.
Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), E1293-301.
Morozov, Y. V., Savin, F. A., Chekhov, V. O., Budowsky, E. I., & Yakovlev, D. Y. (1982). Photochemistry of N6-methoxyadenosine and of N4-hydroxycytidine and its methyl derivatives I: spectroscopic and quantum chemical investigation of ionic and tautomeric forms: syn-anti isomerization. *Journal of Photochemistry*, *20*(3), 229–252.
Mott, J. E., Van Arsdell, J., & Platt, T. (1984). Targeted mutagenesis in vitro: lac repressor mutations generated-using AMV reverse transcriptase and dBrUTP. *Nucleic Acids Research*, *12*(10), 4139–4152.

Müller, W., Weber, H., Meyer, F., & Weissmann, C. (1978). Site-directed mutagenesis in DNA: Generation of point mutations in cloned β globin complementary DNA at the positions corresponding to amino acids 121 to 123. *Journal of Molecular Biology*, *124*(2), 343–358.
Negishi, K., Takahashi, M., Yamashita, Y., Nishizawa, M., & Hayatsu, H. (1985). Mutagenesis by N4-aminocytidine: induction of AT to GC transition and its molecular mechanism. *Biochemistry*, *24*(25), 7273–7278.
Norrander, J., Kempe, T., & Messing, J. (1983). Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene*, *26*(1), 101–106.
Nyrén, P., & Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry*, *151*(2), 504–509.
Oda, Y., Uesugi, S., Ikehara, M., Nishimura, S., Kawase, Y., Ishikawa, H., Inoue, H., & Ohtsuka, E. (1991). NMR studies of a DNA containing 8-hydroxydeoxyguanosine. *Nucleic Acids Research*, *19*(7), 1407–1412.
Olson, C. A., Wu, N. C., & Sun, R. (2014). A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, *24*(22), 2643–2651.
Ovchinnikov, S., Kamisetty, H., & Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *ELife*, *2014*(3), 1–21.
Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, *355*(6322), 294 LP – 298.
Park, J., Selvam, B., Sanematsu, K., Shigemura, N., Shukla, D., & Procko, E. (2019). Structural architecture of a dimeric class C GPCR based on co-trafficking of sweet taste receptor subunits. *Journal of Biological Chemistry*, *294*(13), 4759–4774.
Pavlov, Y. I., Minnick, D. T., Izuta, S., & Kunkel, T. A. (1994). DNA Replication Fidelity with 8-Oxodeoxyguanosine Triphosphate. *Biochemistry*, *33*(15), 4695–4701.
Pazos, F., Helmer-Citterich, M., Ausiello, G., & Valencia, a. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, *271*(4), 511–523.
Purmal, A. A., Kow, Y. W., & Wallace, S. S. (1994). 5-Hydroxypyrimidine deoxynucleoside triphosphates are more efficiently incorporated into DNA by exonuclease-free Klenow fragment than 8-oxopurine deoxynucleoside triphosphates. *Nucleic Acids Research*, *22*(19), 3930–3935.
Reeves, S. T., & Beattie, K. L. (1985). Base-pairing properties of N4-methoxydeoxycytidine 5'-triphosphate during DNA synthesis on natural templates, catalyzed by DNA polymerase I of Escherichia coli. *Biochemistry*, *24*(9), 2262–2268.
Richmond, M. H., & Sykes, R. B. (1973). The beta-lactamases of gram-negative bacteria and their possible physiological role. *Advances in Microbial Physiology*, *9*, 31–88.
Rollins, N. J., Brock, K. P., Poelwijk, F. J., Stiffler, M. A., Gauthier, N. P., Sander, C., & Marks, D. S. (2019). Inferring protein 3D structure from deep mutation scans. *Nature Genetics*, *51*(7), 1170–1176.
Ronaghi, M. (1998). DNA Sequencing: A Sequencing Method Based on Real-Time Pyrophosphate. *Science*, *281*(5375), 363–365.
Ruiz, J. (2018). Etymologia: TEM. *Emerging Infectious Diseases*, *24*(4), 709–709.
Salverda, M. L. M., De Visser, J. A. G. M., & Barlow, M. (2010). Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiology Reviews*, *34*(6), 1015–1036.
Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448.
Sanger, F., Donelson, J. E., Coulson, A. E., Kössel, H., & Fischer, D. (1974). Determination of a

nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *90*(2), 315–333.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467.
Sawai, T., Mitsuhashi, S., & Yamagishi, S. (1968). Drug resistance of enteric bacteria. XIV. Comparison of beta-lactamases in gram-negative rod bacteria resistant to alpha-aminobenzylpenicillin. *Japanese Journal of Microbiology*, *12*(4), 423–434.
Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–R240.
Schäfer, J., & Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. In *Statistical Applications in Genetics and Molecular Biology* (Vol. 4, Issue 1).
Schmiedel, J. M., & Lehner, B. (2019). Determining protein structures using deep mutagenesis. *Nature Genetics*, *51*(7), 1177–1186.
Shibutani, S., Takeshita, M., & Grollman, A. P. (1991). Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature*, *349*(6308), 431–434.
Shortle, D., & Botstein, D. (1983). Directed mutagenesis with sodium bisulfite. *Methods in Enzymology*, *100*, 457–468.
Shugar, D., Huber, C. P., & Birnbaum, G. I. (1976). Mechanism of hydroxylamine mutagenesis Crystal structure and conformation of 1,5-dimethyl-N4-hydroxycytosine. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, *447*(3), 274–284.
Singer, B., Chavez, F., & Spengler, S. J. (1986). O4-methyl-, O4-ethyl-, and O4-isopropylthymidine 5'-triphosphates as analogs of thymidine 5'-triphosphate: kinetics of incorporation by Escherichia coli DNA polymerase I. *Biochemistry*, *25*(6), 1201–1205.
Singer, B., Chavez, F., Spengler, S. J., Kusmierek, J. T., Mendelman, L., & Goodman, M. F. (1989). Comparison of polymerase insertion and extension kinetics of a series of O2-alkyldeoxythymidine triphosphates and O4-methyldeoxythymidine triphosphate. *Biochemistry*, *28*(4), 1478–1483.
Singer, B., Fraenkel-Conrat, H., Abbott, L. G., & Spengler, S. J. (1984). N 4 - Methoxydeoxycytidine triphosphate is in the imino tautomeric form and substitutes for deoxythymidine triphosphate in primed poly d[A-T] synthesis with E. coli DNA polymerase I. *Nucleic Acids Research*, *12*(11), 4609–4619.
Sinha, N. K., & Haimes, M. D. (1981). Molecular mechanisms of substitution mutagenesis. An experimental test of the Watson-Crick and topal-fresco models of base mispairings. *The Journal of Biological Chemistry*, *256*(20), 10671–10683.
Snow, E. T., Foote, R. S., & Mitra, S. (1984). Kinetics of incorporation of O6-methyldeoxyguanosine monophosphate during in vitro DNA synthesis. *Biochemistry*, *23*(19), 4289–4294.
Stemmer, W. P. (1994). DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proceedings of the National Academy of Sciences*, *91*(22), 10747–10751.
Stiffler, M. A., Hekstra, D. R., & Ranganathan, R. (2015). Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell*, *160*(5), 882–892.
Sykes, R. B., & Matthew, M. (1976). The β-lactamases of Gram-negative bacteria and their rôle in resistance to β-lactam antibiotics. *Journal of Antimicrobial Chemotherapy*, *2*(2), 115–157.
Thomas, D. J., Casari, G., & Sander, C. (1996). The prediction of protein contacts from multiple sequence alignments. *"Protein Engineering, Design and Selection," 9*(11), 941–948.
Tindall, K. R., & Kunkel, T. A. (1988). Fidelity of D N A Synthesis by the Thermus aquaticus D N A Polymerase. *J. Am. Chem. SOC. Zuberbuhler, A. D*, *27*(1), 4036–4046.
Toda, M., Inoue, M., & Mitsuhashi, S. (1981). Properties of cephalosporinase from Proteus

morganii. *The Journal of Antibiotics*, *34*(11), 1469–1475.
Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D. L., & Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics*, *44*(1), 101–105.
Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159–e159.
Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A.-P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis †. *Nucleic Acids Research*, *36*(4), e25–e25.
van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–681.
Visintin, M., Tse, E., Axelson, H., Rabbitts, T. H., & Cattaneo, A. (1999). Selection of antibodies for intracellular function using a two-hybrid in vivo system. *Proc Natl Acad Sci U S A*, *96*(21), 11723–11728.
Voelkerding, K. V, Dames, S. A., & Durtschi, J. D. (2009). *Next-Generation Sequencing: From Basic Research to Diagnostics*.
Wang, B., & Kennedy, M. A. (2014). Principal components analysis of protein sequence clusters. *Journal of Structural and Functional Genomics*, *15*(1), 1–11.
Wang, X., Minasov, G., & Shoichet, B. K. (2002). The Structural Bases of Antibiotic Resistance in the Clinically Derived Mutant β-Lactamases TEM-30, TEM-32, and TEM-34. *Journal of Biological Chemistry* , *277*(35), 32149–32156.
Watanabe, T., & Fukasawa, T. (1961). Episome-mediated transfer of drug resistance in Enterobacteriaceae. I. Transfer of resistance factors by conjugation. *Journal of Bacteriology*, *81*, 669–678.
Weigt, M., White, R. a, Szurmant, H., Hoch, J. a, & Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(1), 67–72.
Weirather, J. L., De Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). *Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; referees: 2 approved]*.
Wells, J. A., Vasser, M., & Powers, D. B. (1985). Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene*, *34*(2–3), 315–323.
Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X., Liu, Y., … Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), 872–876.
Wichman, H. A., Millstein, J., & Bull, J. J. (2005). Adaptive Molecular Evolution for 13,000 Phage Generations. *Genetics*, *170*(1), 19–31.
Wilson, D. S., & Keefe, A. D. (2001). Random Mutagenesis by PCR. In *Current Protocols in Molecular Biology* (Vol. 51, Issue 1, pp. 8.3.1-8.3.9). John Wiley & Sons, Inc.
Wright, M. C. (1997). Continuous in Vitro Evolution of Catalytic Function. *Science*, *276*(5312), 614–617.
Wu, R., & Kaiser, A. D. (1968). Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, *35*(3), 523–537.
Zaccolo, M., & Gherardi, E. (1999). The effect of high-frequency random mutagenesis on in vitro protein evolution: A study on TEM-1 β-lactamase. *Journal of Molecular Biology*, *285*(2), 775–783.
Zaccolo, M., Williams, D. M., Brown, D. M., & Gherardi, E. (1996). An Approach to Random

Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. *Journal of Molecular Biology*, *255*(4), 589–603.
Zapun, A., Contreras-Martel, C., & Vernet, T. (2008). Penicillin-binding proteins and β-lactam resistance. *FEMS Microbiology Reviews*, *32*(2), 361–385.
Zhang, H., & Hao, Q. (2011). Crystal structure of NDM-1 reveals a common β-lactam hydrolysis mechanism. *The FASEB Journal*, *25*(8), 2574–2582.
Zhang, J. (2000). Protein-length distributions for the three domains of life. *Trends in Genetics*, *16*(3), 107–109.