Characterization and evolution of artificial RNA ligases

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Aleardo Morelli

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Adviser: Burckhard Seelig

June 2015

Abstract

Enzymes enable biocatalysis with minimal by-products, high regio- and enantioselectivity, and can operate under mild conditions. These properties facilitate numerous applications of enzymes in both industry and research. Great progress has been made in protein engineering to modify properties such as stability and catalytic activity of an enzyme to suit specific processes. On the contrary, the generation of artificial enzymes *de novo* is still challenging, and only few examples have been reported. The study and characterization of artificial enzymes will not only expand our knowledge of protein chemistry and catalysis, but ultimately improve our ability to generate novel biocatalysts and engineer those found in nature.

My thesis focused on the characterization of an artificial RNA ligase previously selected from a library of polypeptide variants based on a non-catalytic protein scaffold. The selection employed mRNA display, a technique to isolate *de novo* enzymes *in vitro* from large libraries of 10^{13} protein variants. The artificial RNA ligase catalyzes the formation of a phosphodiester bond between two RNA substrates by joining a 5'-triphospate to a 3'-hydroxyl, with the release of pyrophosphate. This activity has not been observed in nature. An initial selection carried out at 23°C yielded variants that were poorly suitable for biochemical and biophysical characterization due do their low solubility and poor folding. We hence focused our studies on a particular improved ligase variant called ligase 10C, isolated from a subsequent selection performed at 65°C.

Here we report the structural and biochemical characterization of ligase 10C. We solved the three-dimensional structure of this enzyme by NMR. Unexpectedly, the original structure of the parent scaffold used for building the original library was abandoned. The enzyme instead adopted a novel dynamic fold, not previously observed in nature. The structure was stabilized by metal coordination, yet lacked secondary structural motifs entirely. We also compared the catalytic and thermodynamic properties of ligase 10C to enzyme variants previously selected at lower temperature (23°C). Ligase 10C displayed a remarkable increase in melting temperature of 35°C compared to its

mesophilic counterpart. In addition, its activity at 23°C was about 10-fold higher compared to the mesophilic variants. This work was the first mRNA display selection for catalytic activity at high temperature, and further highlighted the capacity of the technique to select for proteins with rare properties.

To facilitate detailed mechanistic studies of this unnatural enzyme, a crystal structure would be essential. Unfortunately, ligase 10C did not form crystals likely due to its highly dynamic regions. With the goal of identifying a truncated less flexible version of the enzyme that would be more suited for crystallization, we generated a library of random deletion variants of ligase 10C and performed an mRNA display selection to identify shorter active variants.

Finally, we describe the attempted selection of an enzyme for the same RNA ligation reaction from a completely random polypeptide library. The long-term goal of the overarching project in the Seelig lab is to elucidate and compare the structure and mechanism of enzymes generated from different starting points, yet catalyzing the same reaction, to obtain insights into potential evolutionary pathways.

In summary, our work revealed the unusual structural and biophysical properties of the artificial ligase 10C, and thereby demonstrated the power and flexibility of mRNA display as a technique for the selection of *de novo* enzymes.

Table of contents

List of tables

List of supplementary tables

List of figures

List of supplementary figures

Chapter 1 : Introduction

Sections 1.3 and 1.4 were adapted from the review article: Golynskiy, M. V., Haugner III, J. C., Morelli, A., Morrone, D., and Seelig, B. (2013) *In vitro* evolution of enzymes. *Meth. Mol. Biol.* **978,** 73-92 with permission from Springer Science and Business Media. All authors contributed significantly to writing the article. Dr. Burckhard Seelig planned, reviewed and edited the manuscript before submission.

Hyperlink to original publication
http://link.springer.com/protocol/10.1007%2F978-1-62703-293-3_6

Figures 1.3, 1.4, and 1.5 were reprinted from Seelig B, Szostak JW (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448**:828-831 with permission from Macmillan Publishers Ltd.

1.1 Thesis overview

Chapter 1 introduces the use and benefits of enzymes in research and industrial applications, and highlights the importance of directed evolution in developing tailored biocatalysts. Several methods for the evolution of natural enzymes and for the generation of *de novo* enzymes are described. Particular attention is given to an artificial RNA ligase previously selected by Dr. Seelig with mRNA display.

Chapter 2 describes the selection and characterization of a thermostable variant of the artificial RNA ligases. Additional rounds of mRNA display on the variants initially isolated at 23°C were performed at 65°C, and a soluble variant termed ligase 10C was characterized by biochemical and biophysical assays. A melting temperature of 72°C was observed, 24-35 degrees higher compared to two mesophilic counterparts also studied in this chapter. Ligase 10C was also active at 65°C, and 8 to 10-fold more active than the mesophilic variants at 23°C.

Chapter 3 reports on the solution of the three-dimensional structure of ligase 10C by NMR. We observed a new and dynamic fold, which also bound zinc but in a different manner than the parent scaffold. We proposed two putative zinc binding sites and a putative substrate binding site which corresponded to a region conserved among ligase 10C and other ligases selected at 23°C.

Chapter 4 details the development of a method to create libraries of random deletion mutants. The protocol was easier to implement and delivered more consistent results than existing procedures employed for the same purpose. The method was tested on DNA sequences encoding ligase 10C and the results were analyzed by both next generation sequencing and mRNA display. We found that the method created single deletions of varying size and position with only a slight bias for deletions at the 3' end. The obtained deletion library of ligase 10C was used as input in an mRNA display selection that led to the isolation of two active N-terminal deletion variants.

Chapter 5 describes experiments aimed at selecting an RNA ligase from a completely random library. We first optimized the efficiency of all the steps of the selection protocol for high yield and purity of the mRNA displayed proteins and then performed a selection for 13 rounds. Selection progress indicated that no active variants were enriched. Results were verified by sequencing the DNA from the final round of selection and screening of isolated clones. We are currently in the process of repeating the selection under modified conditions that might increase the chance for folding and activity, yet, as this project is work in progress the results will not be part of this thesis.

The final section of the thesis presents conclusions and future directions for mRNA display and artificial enzyme selections in general, and for specific projects in the Seelig lab.

1.2 Significance

1.2.1 The benefits of new enzymes

Enzymes are catalysts capable of performing reactions with minimal by-products, and regio- and enantioselectivity, under very mild conditions. Given these favorable

characteristics, biocatalysis has been employed in various industries, from food to pharmaceutical manufacturing, as a cheaper and environmentally friendly alternative to chemical catalysis [1].

For example, enantiopure β-amino acids are important building blocks of drugs such as the anti-cancer taxol, and β-lactams, which are employed in the production of antibiotics. The chemical synthesis of β-amino acids involves a high metal catalysts load, and hazardous materials. On the other hand, enzymatic catalysis with transaminases offers high selectivity, broad substrate specificity [2], and an enantiomeric excess as high as 99% [3] without hazardous waste. While biocatalysis already plays an important role in the pharmaceutical industry, it was argued that in the future its importance will increase in this sector [4]. Biocatalysis, along with metabolic engineering, will have central roles in multidisciplinary platforms aimed at the manufacturing and discovery of new drugs directed at previously unaddressed targets [4].

Enzymes also have several applications in the food industry. Amylases, are commonly used to break down starch, in the production of high fructose corn syrup, for brewing, baking [5], and in detergents [6]. Proteases among other enzymes are added in detergents, and are also employed in cheese production and meat tenderizing. Pectinases increase yield and favor juice clarification in wine and fruit juice production [5].

Another area of increasing importance of biocatalysis is energy production: demand is expected to increase by 53% in year 2030, and the decrease in fossil fuel availability, climate change, and increasing oil price has driven the exploration of alternatives such as biofuels [7]. Enzymes like cellulases and lipases are crucial in the biofuel manufacturing from cellulose and microalgae. While production processes still need improvement, there is great potential to provide energy sources cheaper and greener than fossil fuels and first generation biofuels [7,8] .

Finally, enzymes have been employed as tools in molecular biology, allowing major advances in basic science. Application of a thermo stable polymerase [9] simplified the execution of the polymerase chain reaction [10], because it was no longer necessary to replace the enzyme at end of each cycle. Ligases and nucleases allowed the

development of recombinant DNA technology [11], and proteases are crucial in proteomics applications from sequencing to protein-protein interaction network discovery [12]. High fidelity polymerases have enabled more accurate DNA amplification, hence they are employed in sample preparation for genome sequencing [13,14]. In addition, the introduction of nucleases such as CRISPR and TALENS [15] permitted accurate editing of large eukaryotic genomes. In particular, CRISPR hold great therapeutic promise for curing diseases at the genetic level [16].

1.2.2 Biocatalysis and the impact of directed evolution

All the applications previously described have been possible because of several discoveries and technological improvements that occurred in the last fifty years; however the use of enzymes for practical purposes can be dated farther back in time. Biotechnology and biocatalysis have been employed since the early history of humanity in alcohol fermentation, bread baking and tanning, as witnessed by Egyptian drawings reporting of such applications [17], or by treatises describing wine production in the roman empire during the second century BC [18].

Nonetheless, the foundations of modern biocatalysis were laid only in the early 19^{th} century with several observations and discoveries that led to the use of isolated enzymes and cell extracts in biotechnology [19]. During the first half of the 20^{th} century enzymes purified from their host organisms were employed in biotechnology applications, however extraction was tedious, yields were often low, and methods to improve enzymatic properties were not available [20]. Later, the introduction of recombinant DNA technology [11] allowed the production of large amounts of the desired protein through heterologous expression. Subsequently, the advent of site-directed mutagenesis [21], enabled the first examples of rational engineering, with success in increasing thermal stability [22] and altering substrate specificity [23]. Random mutagenesis with chemical agents or mutator strains was also employed in enzyme engineering [24-26]. However this approach presented several disadvantages

such as harm to the host organism, off-target mutations [27], and modification of a limited set of amino acids [28].

Despite success achieved with engineering several enzymes and their commercialization, in the early 1990's there were still technical limitations [20]. Building libraries of mutants through random mutagenesis was difficult, and it was argued [29,30] that properties such as catalytic activity and stereo selectivity were too complex and not understood well enough to be engineered systematically and reproducibly through site-directed mutagenesis.

However, random mutagenesis was facilitated by the development of mutagenic PCR [31]. Soon after the introduction of this procedure, pioneering papers were published by Francis Arnold, describing several rounds of mutagenic PCR, each followed by screening, which led to the evolution of a protease to function in an organic solvent [32]. This iterative process of mutation and selection was termed directed evolution and became soon a standard approach for enzyme engineering.

Usually directed evolution was conducted *in vivo*. However, towards the end of the decade, *in vitro* methods were developed. These methods had two main advantages over *in vivo* methods: they were not constrained by the need to maintain the host organism viability, permitting a wider range of conditions, and library size was not limited by transformation efficiency, hence a larger sequence space could be sampled. These methods and their applications will be the focus of the following sections of this chapter, and particular attention will be given to mRNA display. A more general discussion of directed evolution can be found elsewhere [19,20,27,33,34]

1.3 Evolving enzymes *in vitro*

In vitro enzyme evolution offers a means to engineer enzymes by exploring enormous libraries of protein variants that exceed the capabilities of *in vivo* methods. The development of cell-free protein production systems made it possible to evolve enzymes outside of cells, in a test tube. *In vitro* evolution techniques have been used to improve existing enzymes and, in addition, have enabled the generation of biocatalysts *de novo* from a non-catalytic protein library [35].

All methods used for enzyme evolution require that each protein in a pool of mutants can be traced back to its encoding gene for identification, and potentially for the purpose of amplification, expression and further evolution [36]. A stable genotype-phenotype linkage allows for many enzyme variants to be mixed in a single reservoir while maintaining the ability to amplify genes of individual desired variants. Those variants are isolated from the reservoir using suitable screening or selection approaches. In the case of *in vivo* evolution methods, the genotype and phenotype are linked as the protein and its gene are contained in the same cell. With partial *in vitro* methods, proteins are translated by the host's cellular machinery and then displayed in an extracellular fashion, for example on the surface of a phage in the phage display approach. In contrast, the methods described here are carried out entirely *in vitro* and do not require any step to be performed inside a host cell. The crucial genotype-phenotype link is maintained through either a direct physical link or through artificial compartmentalization.

1.3.1 Benefits of in vitro *evolution*

In vitro methodologies have several advantages over *in vivo* and partial *in vitro* methods because they are not limited by cell survival, growth, or function. The three main advantages are: (1) the ability to work with larger libraries of variants, (2) the tolerance to conditions that would be deleterious to cell survival, and (3) the ability to directly manipulate the DNA after each round of evolution.

As *in vitro* evolution is not dependent on library transformation into a host, the number of unique sequences that can be evaluated in a single experiment exceeds *in vivo* approaches. The largest reported *in vitro* libraries contain 10^{14} DNA sequences [37]. By comparison, phage display libraries produce up to 10^{10} unique variants in a single transformation [38]. Library sizes up to 10^{12} variants were reported for phage display by the pooling of dozens of separate transformations, but such scale-up may not be feasible for most laboratories [39]. Most typical library sizes for *in vivo* selections are between 10^{6} and 10^{8} variants. Because *in vitro* evolution can search a larger sequence space, it is particularly well suited for isolating beneficial enzyme mutations that may be very rare.

The evolution of enzymes *in vitro* greatly expands the range of substrates and environmental conditions that can be investigated. The presentation of substrate to the enzymes is simplified as no cell walls have to be crossed, which are impermeable to many potential substrates. Most importantly, substrates and enzymes can be used that would be toxic to a cell [40]. Furthermore, enzymes can be engineered with *in vitro* methods for increased stability under extreme conditions of pH, temperature, ion concentration, or in the presence of denaturants or organic solvents. *In vitro* evolution also allows for a more accurate representation of enzyme performance. Cellular evolution, in contrast, can generate complex phenotypes that falsely suggest increased activity through increased enzyme accumulation, rather than improved catalysis [41].

Finally, *in vitro* evolution allows for direct manipulation of the DNA library between each round of evolution. Unlike *in vivo* methods that require time-consuming purification of the target gene, DNA from *in vitro* evolution is amplified directly through PCR. This facilitates the introduction of diversity through methods like error-prone PCR or *in vitro* recombination. In comparison, *in vivo* methods may introduce genetic diversity by using microbial strains deficient in DNA repair pathways to eliminate the need for DNA purification. However, these mutations may occur anywhere in the genome, necessitating a low mutation rate for continued survival [42]. Thus, by combining a selection or screen with methods to add genetic diversity, full Darwinian evolution can be carried out more conveniently *in vitro*.

While *in vitro* evolution greatly expands the tools available for the creation and engineering of new enzymes, *in vivo* approaches have certain advantages, too. As *in vitro* methods require purification of the genotype/phenotype components, *in vivo* evolution may involve fewer discrete steps. Furthermore, some enzymes are being developed for *in vivo* use, such as enzymes that need to function within a metabolic pathway. Those enzymes could initially be evolved *in vitro*, but ultimately need to be evolved in their native environment to optimize their intracellular compatibility. Thus, the two approaches can complement each other.

1.3.2 General workflow for in vitro *methods*

All *in vitro* directed evolution methods follow a similar scheme. The initial DNA library encoding the protein variants is transcribed and translated, either sequentially or in a one-pot reaction. Next, the genotype-phenotype link and then genotype-phenotype-substrate link is established. This may be accomplished through a physical connection (ribosome display, mRNA display, and DNA display), or through compartmentalization (IVC, IVC-based DNA and microbead displays) (Figure 1.1). Active enzyme variants that convert substrate to product result in co-localization of genotype-phenotype with product and are then isolated by screening or selection. Finally, the genotype is recovered and either analyzed directly by sequencing or subjected to additional diversification for subsequent rounds of evolution.

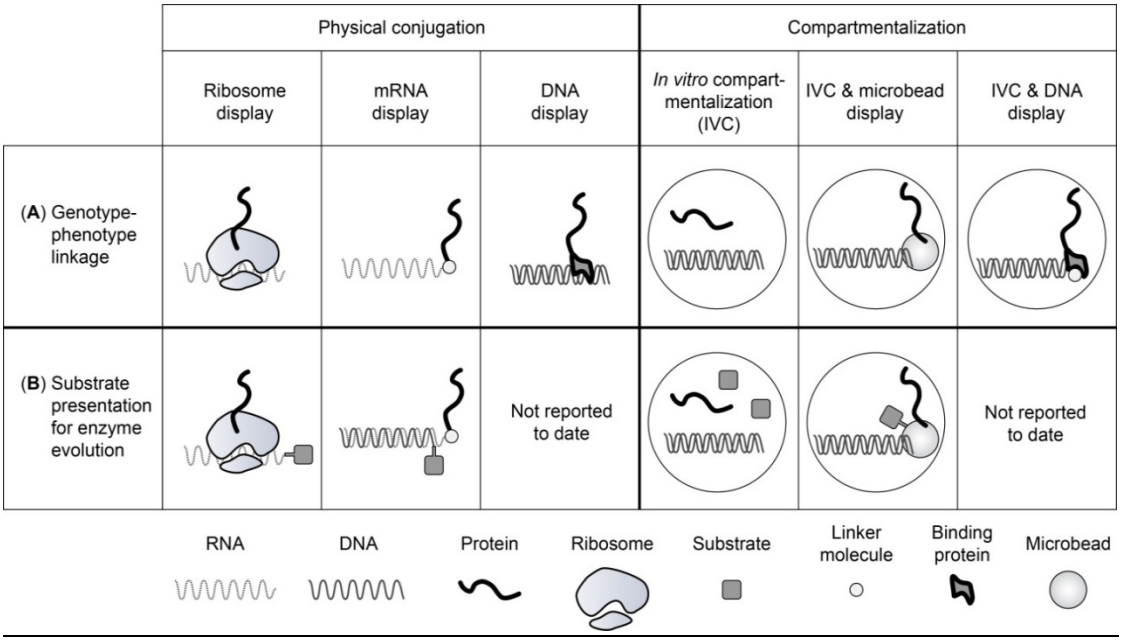

Figure 1.1-Overview of methods for the *in vitro* selection or screening of proteins discussed in this review.
(A) The top row shows different strategies to establish the crucial linkage between gene and protein. **(B)** The bottom row illustrates the introduction of substrate into the selection scheme to enable the evolution of enzymes.

1.3.3 Library construction

In any directed evolution procedure, the size and quality of the starting DNA library are of great importance as they affect the probability of finding the desired

mutant. Although *in vitro* selection methods can sift through comparably large libraries of trillions of mutants, the sheer size of the protein sequence space prevents us from sampling more than an exceedingly small fraction of all possibilities. For example, the largest protein libraries used to date contain about 10^{13} variants (Table 1.1). This vast number of mutants will just be enough to include one molecule of all possible combinations for a sequence of ten amino acid positions. In comparison, most natural proteins are more than 100 amino acids in length. Therefore, libraries of mutants should be designed wisely to increase the chances of success in a directed evolution experiment. Accordingly, one should consider randomizing specific amino acid positions by using degenerate codons. Instead of randomizing positions with NNN codons (N=A,C,G,T), NNK codons (K=G,T), NNS codons (S=C,G) or even a reduced alphabet of NDT codons (D=A,G,T) can be used to reduce oversampling caused by codon degeneracy [43]. The use of degenerate codons can also reduce the likelihood of introducing unintended stop codons. For example, the NNN codon includes three stop codons whereas the NNK or NNS codons include only one. Alternatively, a given library can be assembled from fragments that have been pre-selected to decrease the occurrence of premature stop codons [37]. More recently, DNA synthesis via phosphoramidite trinucleotides has become commercially available [44]. Codon by codon synthesis using trinucleotides offers full control of the library composition by defining the set of desired amino acid mutations at any position while avoiding stop codons [45].

Table 1.1- Comparison of *in vitro* technologies.

Method	**Genotype-phenotype link**	**Reported variants in single experiment**	**Results**
Ribosome display	Non-covalent complex of mRNA-ribosome-protein	~10^{13}	Proof of concept selections for sialyltransferase [46], beta-lactamase [47], dihydrofolate reductase [48], DNA ligase [49] and sortase [50]
mRNA display	Covalent fusions of mRNA-protein via puromycin	~10^{13}	Selection for *de novo* RNA ligase [51,52]

DNA display	Covalent or non-covalent complex of DNA-protein	~10^{12}	Proof of concept selection of binders, but no enzymes [53,54]
In vitro compartment-talization (IVC)	Spatial confinement	~10^{9} (selection) ~10^{6}-10^{8} (screening by FACS/microfluidics)	Selection for methyltransferase [55] and restriction nuclease [56]; Proof of concept screening for β–galactosidase [57,58]
IVC & microbead display	Non-covalent complex of DNA-microbead-protein	~10^{9} (selection) ~10^{6}-10^{8} (screening by FACS/microfluidics)	Screening for phosphotriesterase [59]; Proof of concept screening for hydrogenase [60]; Proof of concept selection of biotin ligase [61]
IVC & DNA display	Covalent or non-covalent complex of DNA-protein	~10^{8}-10^{9}	Selection of antibodies as heterodimers, but no enzymes [62]

In order to use a DNA library for a specific *in vitro* evolution technique, the sequences at both termini of the DNA have to be made compatible to the method of choice. The 5'-end includes promoter and enhancer sequences necessary to facilitate transcription and translation, respectively. The nature of these sequences depends on the type of transcription and translation system used. Other sequence elements might be included such as a terminator, stabilizing hairpins, affinity purification tags or sequences that are specific to the particular *in vitro* evolution method [63].

1.4 Methods for *in vitro* directed evolution

1.4.1 Ribosome display

The ribosome display technology creates the genotype-phenotype linkage through a ternary complex of a stalled ribosome, the translated protein and its encoding mRNA (Figure 1.1). The complex is stabilized by high magnesium concentrations and low temperatures. Ribosome display was initially described for the purification of specific mRNA sequences based on immunoprecipitation of the encoded protein [64]. Subsequently, this method was developed further to select and evolve peptides and proteins [65,66]. Although ribosome display has mostly been used for selection of

binders, several model selections for enzymatic activity have been reported and will be reviewed here in more detail.

Several criteria must be met in order to generate ribosome-displayed proteins. Most importantly, the terminal stop codon of the gene of interest must be removed. This will prevent the ribosome from dissociating and releasing the nascent protein and will instead promote stalling of the ribosome and therefore maintain the ternary complex. Stem-loop structures are often added to flank the gene on both termini to increase RNA stability during translation and subsequent manipulations. Since the protein is not released from the ribosome, a C-terminal protein spacer (>100 amino acids) is added to ensure that the displayed protein has exited the protein-conducting channel of the ribosome and can fold properly. Typically, ribosome-displayed proteins are generated through sequential transcription and translation, as coupled transcription/translation systems can result in 100-fold reduced protein yield [65,67]. The translation is stopped by decreasing the temperature and increasing the Mg^{2+} concentration to stabilize the ternary complex. To maintain the genotype-phenotype linkage, the subsequent selection process also has to be performed at low temperatures and in presence of elevated Mg^{2+} concentrations. The ribosome-displayed proteins are mostly used in selections without any additional purification. The RNA is recovered after the selection by dissociating the ternary complex through chelation of Mg^{2+} with EDTA.

Ribosome display has been utilized in a number of model selections for enzymatic activity. Most selections were performed by selecting for binding to an immobilized substrate, substrate analog, or inhibitor. These model selections demonstrated enrichment of the desired enzyme (10 to 100-fold per round of selection) compared to an inactive control (Table 1.1) [46,48-50]. While enzyme selection strategies based on binding can be successful in isolating enzymes with known properties (e.g. searching through metagenomic libraries for a desired activity), they are not well suited for changing substrate specificity or substantially improving activity [33,68]. In one example of a truly product-driven model selection, ribosome display has been employed for isolation of a T4 DNA ligase [49]. Active enzymes able to ligate a DNA adaptor to the 3'-end of their

encoding mRNA were selectively amplified via an adaptor-specific primer and were enriched 40-fold over known inactive mutants. Similar to this selection approach, the 3'-end of the mRNA could be used for the attachment of alternative substrates which would allow for a selection of other catalysts by ribosome display.

1.4.2 mRNA display.

mRNA-displayed proteins are covalently attached to their encoding mRNA via the small linker molecule puromycin (Figure 1.1) [69,70]. Central to the mRNA display method is the modification of the stop codon-free 3'-end of the messenger RNA with a puromycin-containing DNA linker prior to translation [71,72]. During the subsequent *in vitro* translation, the ribosome synthesizes the polypeptide until it reaches the DNA-puromycin-modified 3'-end of the mRNA where it stalls. Puromycin, which is an antibiotic that mimics the aminoacyl end of tRNA, enters the ribosome and becomes covalently attached to the C-terminus of the nascent polypeptide. The resulting mRNA-displayed proteins are typically purified from unfused proteins and mRNA using purification tags. The mRNA-displayed proteins are reverse transcribed to produce the cDNA. Reverse transcription also minimizes potential RNA secondary structure and increases RNA stability. Detailed protocols on mRNA display have been published recently [52,73,74]. Through slight modifications of the mRNA display protocol, covalent fusions of protein and encoding cDNA can be generated (cDNA display) [75,76].

In section 1.5.2 of this chapter we will describe in more details the *de novo* selection of an enzyme by mRNA display, starting from a zinc finger non catalytic scaffold [51]. This is so far the only published report of *de novo* enzyme selection by *in vitro* directed evolution, and the characterization and evolution of the resulting enzyme constitutes the basis of my thesis.

1.4.3 In vitro *compartmentalization (IVC)*

Directed evolution by *in vitro* compartmentalization mimics *in vivo* evolution inside a cell by using water-in-oil emulsions to enclose proteins and their encoding DNA within the same droplet compartment thereby creating the genotype-phenotype link through spatial confinement [77]. IVC has been employed not only in several model enzymes selections, but also to improve the performance of existing enzymes through screening and selection methods.

Compartmentalization by droplet formation is achieved by stirring an aqueous solution of genes and a coupled transcription/translation (TS/TL) system into a mixture of mineral oil and surfactants [78]. The DNA concentration is chosen such that the average droplet contains no more than a single gene. The low volume of the droplets (5-10 femtoliters) corresponds to a low nanomolar concentration of the single DNA molecule, which is efficiently transcribed and translated inside the droplet [58,77,79]. Although droplet composition is similar across different IVC experiments, in some cases the oil/surfactant mixtures need to be optimized for compatibility with the specific TS/TL solution used and the enzymatic activity that is being evolved [77,80]. It has been shown that the droplets are stable up to 100°C for many days and do not exchange DNA or protein between each other [77,81].

IVC-based selections have been used to evolve enzymes that process nucleic acid substrates. Here, the encoding DNA is also the substrate for the enzyme and the selection is dependent on successful DNA modification. In one approach, the activity of the methyltransferase (M.HaeIII) was improved toward a non-native, although already recognized, DNA sequence [55]. A library of variants of M.HaeIII was made by mutating the DNA contacting residues. The 3'-end of the DNA library was modified with a biotin moiety and connected to the remaining gene via the target methylation site that can be cleaved by endonuclease NheI unless the site is has been methylated by M.HaeIII. Therefore, only methylated genes were not cleaved by NheI and were captured on streptavidin beads. A similar approach was used for the model selection of a restriction endonuclease activity from a randomized library of the restriction enzyme FokI. Three

specific residues were randomized in the catalytic domain, and cleavage sites for FokI were introduced in the 3'-UTR [56]. Only the genes coding for an active FokI variant were cleaved and captured on beads after incorporation of biotinylated deoxyuracil triphosphate at the cohesive ends generated by the restriction enzyme.

The IVC methodology has also been used in combination with screening approaches. This allows for the evolution of enzymes for non-nucleic acid related reactions, but also reduces the number of mutants that can be interrogated compared to selection strategies. In the screening approach, either fluorescence activated cell sorting (FACS) or microfluidics-based droplet sorting are used to separate active and inactive enzymes based on the conversion of non-fluorescent substrate into fluorescent product. For FACS mediated screening, water-in-oil-in-water emulsions (double emulsions) are generated since FACS instrumentation is incompatible with oil as the main medium [82]. Exploiting this principle, the very low β-galactosidase activity of the Ebg enzyme from *E. coli* was increased at least 300-fold by *in vitro* evolution using a commercially available fluorogenic substrate [58]. Recently, the same researchers reported a model enrichment of β-galactosidase using a home-made microfluidic system [57]. Although the throughput in the microfluidic system is about 10-fold less than in FACS-based screening, this loss is offset by other advantages. First, the microfluidic system generates highly monodisperse droplets, enabling quantitative kinetic analysis [57,83]. Second, the authors utilized microfluidic components that allowed them to fuse droplets together and introduce new content into droplets. This conferred multiple benefits as the authors were able to perform emulsion PCR in droplets and then merge them with droplets containing the TS/TL mix. By generating about 30,000 gene copies per droplet prior to TS/TL, low enzymatic activity is more likely to be detected due to the elevated enzyme concentration [57]. Furthermore, reagents can be readily added to the droplets after translation, in case the translation conditions are not compatible with enzymatic assay [84]. The use of microfluidics is a promising route for IVC-based enzyme engineering due to the modularity and potential for customization of individual components. However, in

contrast to commercially available FACS instruments, assembly of microfluidics devices still requires substantial expertise.

IVC has also been used in conjunction with *in vivo* enzyme evolution by generating compartments that contain cells. To keep the focus of this review we are not discussing this *in vivo* application.

1.4.4 DNA display

Strategies that either directly or indirectly establish a physical link between the DNA and the encoded protein are referred to as DNA display (Table 1.2). Although several different DNA display methods have been developed, only the IVC-mediated microbead display has been used to evolve enzymes. This method generates the genotype-phenotype link through the capture of DNA and its translated protein onto the same streptavidin-coated microbeads inside a droplet (Figure 1.1) [59,60]. This approach requires multiple biotinylated reagents such as primers, antibody and reaction substrate in order to capture the template DNA, the protein modified with an epitope tag and the substrate onto the microbead, respectively.

Using microbead display, Tawfik and Griffiths improved the catalytic performance of an already very efficient phosphotriesterase enzyme 63-fold ($k_{cat} > 10^5$ s^{-1}) through FACS-based screening [59]. This work demonstrated the ability to generate, break and regenerate the IVC droplets and purify the genotype-phenotype-product attached to the microbeads. Furthermore, a substrate was used that carried a photo-caged biotin. Therefore, the substrate stays in solution until the biotin is uncaged, which causes the immobilization of substrate and resulting product on the beads. Incubation with a fluorescent product-specific antibody enabled the specific labeling and isolation by FACS of only those microbeads to which functional enzymes and their coding DNA were attached [59].

In a different proof of concept experiment, a modified microbead display protocol was performed as a selection instead of a screen, thereby potentially harnessing larger library sizes [61]. In this experiment, an active biotin ligase was enriched from a mixture

of inactive genes. Following product formation and immobilization, the purified microbeads were incubated with product-specific antibodies that were conjugated to a cleavable, gene-specific PCR primer instead of a fluorophore. Re-emulsification and droplet PCR with a solution lacking this primer resulted in a 20-fold enrichment of the desired genes.

Another microbead display model screen employing FACS used an indirect readout for activity to isolate [FeFe] hydrogenases [60]. Because the hydrogenase activity (H_2 breakdown) is difficult to measure directly, the authors employed a redox-sensitive dye that can generate a fluorescent signal. Purified microbeads carrying the immobilized DNA and enzymes were re-compartmentalized in the presence of the redox dye. This dye was modified with a C12-alkyl chain and therefore interacts non-specifically with the hydrophobic polystyrene beads. Hydrogenase activity resulted in fluorescence of the dye and enabled flow cytometric sorting of the microbeads to recover the DNA of active enzymes, yielding a 20-fold enrichment over inactive genes. This proof of concept study used microfluidics to generate mono-disperse droplets and microbeads with a larger diameter (5.6 µm rather than 1 µm) to increase the bead surface allowing more fluorescent substrate to bind, thereby improving the signal to noise ratio. The indirect readout as described here could be applied to other screening strategies if environmentally sensitive fluorophores are available (pH, redox potential).

Presently, only microbead display has been employed to evolve enzymes. Yet other DNA display methods could potentially be used for this purpose. In contrast to microbead display, all other DNA display methods directly attach the protein to its encoding gene via a fusion protein which binds to a specific DNA sequence within the parent gene or to a small molecule attached to the parent gene (Table 1.2). The IVC method is often used in conjunction with DNA display as the physical genotype-phenotype linkage allows for the microcompartments to be broken up and generated again in order to introduce new components into the system (e.g. substrates). However, two proof-of-concept studies conducted without IVC demonstrated the production of

DNA-displayed proteins solely by incubating templates with the *E. coli* cell extract [53,54].

Table 1.2- DNA display methods
Only the microbead display has been used to evolve enzymes.

Method	**Principle of attachment**	**DNA - point of attachment**	**Protein fusion partner**
Microbead display [59-61]	Non-covalent binding of DNA to streptavidin microbead and of HA-tagged protein via anti-HA antibody to same bead, IVC is needed	Biotinylated	HA-tag
STABLE [62,85]	Non-covalent attachment of protein to DNA, IVC is needed	Biotinylated	Streptavidin
CIS-display [53]	Non-covalent attachment of protein to DNA	RepA gene	DNA replication initiator (RepA)
Covalent DNA display [86,87]	Covalent attachment of enzyme to suicide inhibitor that is linked to DNA, IVC is needed	Modified with 5-fluoro-deoxycytidine	HaeIII methyltransferase
Covalent antibody display [54]	Covalent attachment of enzyme to DNA	P2A gene	Endonuclease P2A
SNAP display [88,89]	Covalent attachment of enzyme to suicide inhibitor that is linked to DNA, IVC is needed	Modified with benzyl guanine	SNAP-tag

1.4.5 General principles and comparison of different in vitro *methods*

The types of reactions catalyzed by enzymes can be divided into transformation reactions, bond-forming reactions and bond-breaking reactions (Figure 1.2A). Depending on the reaction type, the strategy by which enzymes can be selected varies slightly. In general, affinity selections are used to isolate enzymes by methods that create a physical link between phenotype and genotype such as ribosome display, mRNA display and DNA display (Figure 1.1 and Figure 1.2B). To enable an enzyme affinity selection, the substrate has to be linked to the gene-enzyme complex. Enzymes for a transformation

reaction can then be isolated if a product-specific affinity reagent, such as an antibody, is available (reaction type 1). Via the antibody, the ternary complex of product, active enzyme and gene is separated from inactive variants through immobilization. In the case of an affinity selection for bond-forming enzymes (reaction type 2), the second substrate carries a selectable moiety. Only proteins that catalyze the bond formation between two substrates will attach this moiety to the gene-protein-substrate complex and can therefore be isolated. For bond-breaking reactions (reaction type 3), the whole complex of gene, protein and substrate is immobilized via the substrate and only variants that cleave the bond will be released and selected. In contrast to affinity selections, the IVC methodology mostly employs fluorescent screening to isolate evolved enzyme variants either by FACS or microfluidics (Figure 1.2C and D). This can be achieved if the product of the reaction becomes fluorescent or a fluorescent product-specific antibody is available. Alternatively, the second substrate, which will be attached in a bond-forming reaction to the gene-microbead-substrate-complex, is fluorescent.

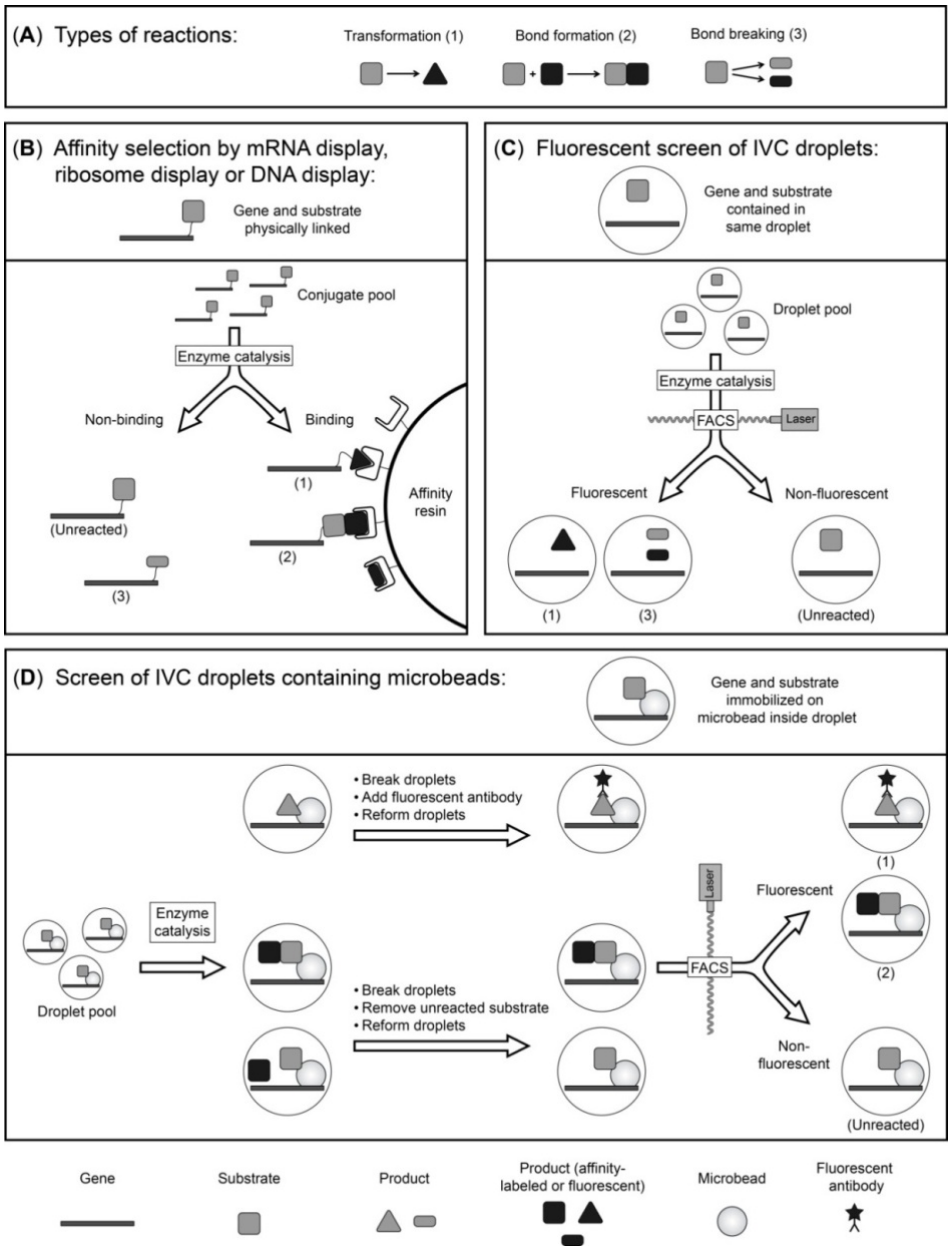

Figure 1.2-Isolation of enzymatic activities using *in vitro* technologies.
(A) Types of enzymatic activities that can be evolved using *in vitro* approaches. **(B)** Affinity selection of physically linked gene-substrate/product conjugates. The enzyme itself is also linked to the gene-substrate complex, but is omitted from the figure for improved clarity. **(C)** Screen of IVC droplets that become fluorescent as a result of catalysis by the enzyme (not shown) contained in same compartment. Separation is achieved through fluorescence activated cell sorting (FACS) or microfluidics. **(D)** Screen for enzyme catalysis by FACS of IVC droplets containing microbeads. The enzyme contained in each compartment is not shown to improve clarity. Numbers in brackets refer to the type of activity as shown in (A).

For any enzyme evolution experiment regardless of which methodology is used, the specific selection or screening strategy has to be customized with respect to the underlying reaction. In the case of affinity selections, the need to link the substrate to the gene-complex without substantially changing the nature of the substrate can be challenging especially for small substrates. On the other hand, suitable fluorophores that enable the screening of IVC droplets might not be compatible with some types of chemical reactions.

Two important questions have to be considered when deciding on which enzyme evolution strategy to use: Is the desired mutant potentially very rare such as a mutant exhibiting a novel activity? Or, alternatively, is the goal of the evolution experiment to generate a highly proficient enzyme? Selection strategies can search larger libraries and are therefore more likely to discover rare mutants, compared to screening approaches. At the same time, affinity selections only select for a single turnover event and cannot evolve an enzyme for high substrate affinity as the substrate is linked to the enzyme and therefore present at a high local concentration. In contrast, IVC-based screening methods can directly evolve an enzyme for high turnover and substrate affinity, yet, the library size of screening methods is several orders of magnitude smaller than those of selections. Therefore, it might be most beneficial to combine the two strategies and first use an affinity selection method to isolate potentially rare enzyme variants with altered activity or substrate specificity and then switch to an IVC-based screening method to optimize enzymatic proficiency.

1.5 *De novo* enzymes by computational design and *in vitro* directed evolution

In most of the directed evolution experiments, mutations are introduced into an existing enzyme to improve its properties: activity or stability. However, if a protein with the desired activity does not exist in nature, the activity needs to be generated *de novo*. Two approaches have been employed so far to obtain novel enzymes: *in vitro* selection [51], and computational design [90-97].

1.5.1 De novo enzymes by computational design

Computational design is a rational approach which requires detailed knowledge of the targeted reaction. The elucidation of the catalytic mechanism constitutes the starting point of any computational design experiment. Subsequently, different active site conformations are designed to fit the transition state. Finally, the active site conformations are matched combinatorially to several protein scaffolds that can accommodate the active site, to generate hundreds of thousands of structures. Powerful software, such as the Rosetta software [98], allow to generate these structures quickly, and to evaluate them for optimal packing and geometry of the active site. The procedure described above is usually performed in four steps [99]:

1) Definition of target reaction and catalytic mechanism
2) Modeling of an ideal active site, called theozyme, by quantum mechanics, in order to stabilize the transition state
3) Docking of the theozyme into a scaffold protein, followed by *in silico* mutagenesis to further stabilize the transition state and improve packing
4) Computational evaluation of the obtained structures regarding geometry of the active site and binding energy of the transition state, with subsequent experimental evaluation of the top scoring structures

Several *de novo* enzymes, such as a retro-aldolase [92], Diels-Alderase [94], and Kemp eliminase [93], have been obtained through computational approaches. These enzymes usually possessed low activity. However, improvements could be obtained, by further modeling [100] and by *in vitro* directed evolution [95]. For example, the K_{cat}/K_m of several retro-aldolase designs was improved as high as 88-fold by changing design strategy, and subsequent directed evolution led to improvements of 1100 and 1300-fold [95]. In another report, the catalytic efficiency of a Diels-Alderase activity was improved 18.5-fold, through further computational modeling, by addition of 24 residues which turned a 13 residue loop into a helix-turn-helix motif [100]. These examples show that combination of *de novo* enzyme design and directed evolution offer the potential to improve novel enzymes with low activities.

1.5.2 De novo enzymes by in vitro directed evolution

The *in vitro* selection technology mRNA display was used to generate the first *de novo* enzyme through directed evolution [51]. The starting library was composed of 10^{13} members. On the other hand most directed evolution techniques sample libraries made of to 10^{10} proteins, severely limiting the possibility of finding rare variants.

The input library of protein variants was based on a zinc finger non-catalytic scaffold. Two loops composed of nine and twelve amino acids each were randomized to introduce genetic variation. A 5'-triphosphate dependent RNA ligase was selected, which catalyzes the ligation of two RNA strands, forming a 5'-3' phosphodiester bond, and relies on a splint oligonucleotide to align the two substrates (Figure 1.3).

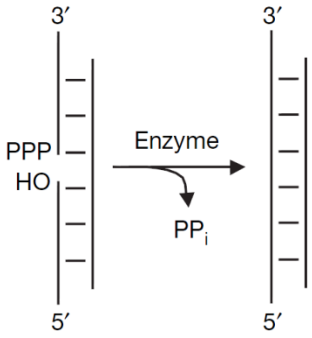

Figure 1.3-Splinted ligation of RNA with a 5' triphosphate releasing pyrophosphate [51]

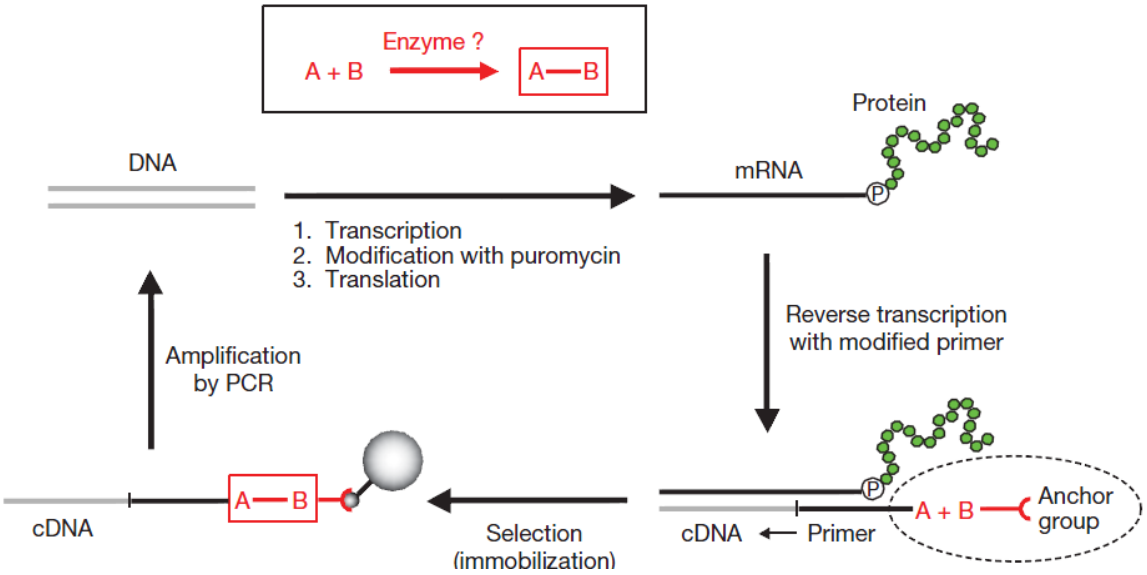

Figure 1.4-General scheme for the selection of bond-forming enzymes. [51]
Substrate A is attached to the reverse transcription primer and substrate B is modified with an anchor group. Ligation of A and B results in a covalent link between the cDNA and the anchor group. The cDNA of active clones can then be separated from that of non-active clones because they are not capable of ligating A and B.

The mRNA display procedure involved several steps (Figure 1.4). The input library was transcribed into RNA, which was modified with puromycin, and upon translation a stable link was obtained between the RNA and the protein it encoded for, as explained in section 1.4.2. The obtained mRNA displayed proteins (fusions) were then purified, and reverse transcribed with a primer attached to substrate A, in this case a 5'-PPP RNA oligo (see Figure 1.3). After further purification, fusions were incubated with the second substrate, a 3'-OH RNA modified with biotin. Ligation of the two substrates by active enzymes resulted in a covalent link between the cDNA and the biotin group. The cDNA of active clones was immobilized on a streptavidin resin and separated from the cDNA of inactive clones, which were incapable of forming the crucial link between their encoding cDNA and the biotin group. The DNA recovered at the end of each round was then PCR amplified to reintroduce the promoter region and the whole cycle of selection and amplification was repeated until the library was enriched with the active variants.

After 17 rounds of selection and evolution (by error prone PCR and recombination), all but 8 out of 49 clones (83%) had mutated or deleted the original cysteine pattern consisting of 4 CX_nC motifs, and contained deletions of either 17 or 13 amino acids (Figure 1.5). The presence of major sequence changes suggested major structural rearrangements as well. We solved the tree-dimensional structure of a ligase enzyme variant which is described in chapter 3 and compared it to the original scaffold.

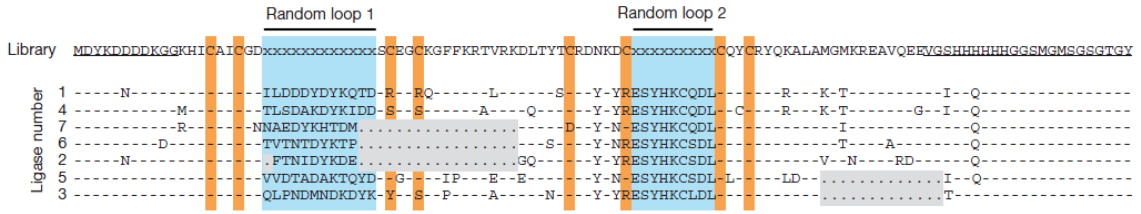

Figure 1.5-Sequences of starting library and selected artificial RNA ligases [51].
Randomized loops are highlighted in blue and Cys residues in orange. Loop 2 was highly conserved among all active enzymes isolated suggesting a key role in the new enzyme. Most clones have mutated or lost at least two of the original 8 cysteines, and contain major changes such as deletion of 13 and 17 amino acids.

1.5.3 Molecular biology applications of the de novo RNA ligase

All known natural ligases rely on ATP as a cofactor and on a 5' monophosphate to perform ligation. These enzymes are used in several molecular biology applications. In particular, a common procedure for the analysis of RNA transcripts, involves ligation of known sequences (adaptors) at both RNA termini to enable reverse transcription and then PCR amplification. The resulting DNA is then characterized by several methods such as sequencing or microarray.

Some classes of RNA, including prokaryotic messenger RNA or secondary short interfering RNAs (siRNAs) in *C. elegans*, are characterized by their triphosphate group at the 5' end. Since existing RNA ligases require a monophosphate, the 5'-PPP must be dephosphorylated prior to ligation. This additional step complicates sample preparation and data analysis [101].

In 2013, the artificial RNA ligase we characterized was shown to have broad substrate specificity and to ligate adaptors to the 5'-PPP end of two siRNAs sequences

[101]. These findings demonstrated that the enzyme was potentially suitable for the isolation and characterization of any 5'-PPP group, and showed the ability of in vitro directed evolution methods to generate useful biocatalysts.

1.6 Selection of functional proteins from random libraries with mRNA display

It has been hypothesized that the earliest functional proteins arose from random sequences. Both theoretical [102] and experimental work [103] has been conducted to verify the plausibility of this hypothesis. For example, in 2009 [102] the structures of 18465 random protein sequences generated *in silico* and extracted from an initial pool of $2x10^4$ proteins, were predicted using the Rosetta software. It was found that these proteins possessed well folded secondary and tertiary structures comparable to those of natural proteins, and not significantly different biophysical properties. Previous experimental work [103] suggested that, given a pool of 10^9 random 50 amino acids peptides displayed on phage, 20% are properly folded, as assessed by results of a proteolysis resistance assay performed on 79 clones picked randomly. Other experimental work [104] showed that when inserting completely random sequences into a functional scaffold, variants maintaining the initial activity could be found.

One study [105] addressed experimentally the question of what is the frequency of functional proteins in random libraries. In 2001, Keefe and Szostak [105] starting from a pool of $6x10^{12}$ random proteins of 80 amino acids each, selected for an ATP binding protein, using mRNA display. Four different families (A-D) of proteins with sequences unrelated to each other and to natural proteins were found. In particular, one of the isolated variants had a very high affinity for ATP of (K_d= 100 nM)

Furthermore, X-ray crystallography studies showed a novel fold characterized by three β-strands and two α-helices, which was stabilized by coordination of a zinc atom [106]. No other examples of zinc stabilized nucleotide binders were observed in nature. Interestingly, zinc was not added during the selection step, but was present during translation as a component of the translation lysate. This suggested that metal coordination is a strategy to readily generate folded structure of polypeptides, while

decreasing the amount of sequence information needed to generate folded functional proteins.

These results indicate that random libraries are a source of folded and functional proteins. However, the idea that catalytic activity is too complex to arise from completely random libraries [107] limited the attempts to select enzymes from this type of libraries. Only few studies are available which involved prescreening for solubility [108,109], or the use of random binary patterned libraries to enrich for secondary structures [110]. However, while most known three-dimensional structures of modern enzymes fall in the category of well-folded structures [111], a view biased by the structure of proteins that can be crystallized, the possibility that primordial enzymes lacked this level of complexity cannot be discounted. In order to enable a low level of activity, a functional binding site, with more rigid requirements for binding the substrate and catalysis, are needed while the rest of the sequence could be under lower selection pressure, increasing the chances of selecting for catalytic activity from unbiased library.

This reasoning, in combination with evidence for functionality in libraries of random sequences, prompted us to start the selection of an RNA ligase from a random library [37] that was previously used as input for the selection of the ATP binder described above [105].

Chapter 2 : Thermostable artificial enzyme isolated by *in vitro* selection

This chapter is a reprint of the article: Morelli, A., Haugner III, J. C., and Seelig, B. (2014) Thermostable artificial enzyme isolated by *in vitro* selection. *PLOS ONE*. No permission from the editor is required to reuse the article as long as the original article is cited. I purified the artificial ligases by size exclusion chromatography and characterized thermal stability and secondary structure by circular dichroism. Haugner expressed and purified ligases by Ni-NTA chromatography and performed activity assays. Dr. Seelig designed and performed the selection of the heat stable ligases.

2.1 Summary

Artificial enzymes hold the potential to catalyze valuable reactions not observed in nature. One approach to build artificial enzymes introduces mutations into an existing protein scaffold to enable a new catalytic activity. This process commonly results in a simultaneous reduction of protein stability as an undesired side effect. While protein stability can be increased through techniques like directed evolution, care needs to be taken that added stability, conversely, does not sacrifice the desired activity of the enzyme. Ideally, enzymatic activity and protein stability are engineered simultaneously to ensure that stable enzymes with the desired catalytic properties are isolated. Here, we present the use of the *in vitro* selection technique mRNA display to isolate enzymes with improved stability and activity in a single step. Starting with a library of artificial RNA ligase enzymes that were previously isolated at ambient temperature and are therefore mostly mesophilic, we selected for thermostable active enzyme variants by performing the selection step at 65°C. The most efficient enzyme, ligase 10C, is not only active at 65°C, but is also an order of magnitude more active at room temperature compared to related enzymes previously isolated at ambient temperature. Concurrently, the melting temperature of ligase 10C increased by 35 degrees compared to these related enzymes. While low stability and solubility of the previously selected enzymes prevented a structural characterization, the improved properties of the heat-stable ligase 10C finally allowed us to solve the three-dimensional structure by NMR. This artificial enzyme

adopted an entirely novel fold that has not been seen in nature, which was published elsewhere. These results highlight the versatility of the *in vitro* selection technique mRNA display as a powerful method for the isolation of thermostable novel enzymes.

2.2 Introduction

Protein stability is often a limiting factor for the application, engineering and structural studies of proteins. Low protein stability can result in aggregation, susceptibility to protease degradation and poor yields in the expression of soluble protein, thereby complicating the study and use of these proteins. For commercial applications, proteins commonly need to be particularly stable to increase their tolerance to process conditions like high temperatures or organic solvents [34]. Furthermore, proteins with low stability are less tolerant to mutations thereby limiting further engineering because even slightly destabilizing mutations can lead to unfolding. This can create situations where mutations that would improve enzyme activity in a protein engineering project appear ineffective because the enzyme was not stable enough to remain folded [112]. Conversely, improved thermal stability correlates with mutational robustness and evolvability [113].

Methods to increase the thermodynamic stability of proteins include rational design, consensus-based design, directed evolution, and commonly some combination of these approaches [114]. Rational design introduces mutations predicted to enable additional stabilizing interactions [115]. However, this approach requires extensive structural knowledge, substantial computing power and is technically challenging, which still limits the accessibility of this method. Consensus based-design utilizes phylogenetic information to determine which amino acids are preferred at certain positions [115]. This method can also be used to reconstruct thermostable ancestral proteins or, be combined with structural knowledge, which likely further improves the prediction of stabilizing mutations. However, these approaches are dependent on the quality of the constructed phylogenetic tree, which is non-trivial to accurately assemble. Directed evolution is a combinatorial approach that introduces mutations at random and then screens for desired properties such as improved activity or stability [116-118]. High throughput screens are

often performed *in vivo*, utilizing colorimetric [119] or fluorescent [120] reporters to measure levels of soluble expression as readout for stability or *in vitro* using protease resistance and phage display [121,122]. Protein variants are also commonly assayed directly for thermostability and activity as purified proteins, but these methods have a relatively low throughput [114,123]. As mutations are introduced randomly, the chance of success increases with the number of mutants sampled. This favors high throughput methods which can sample millions to trillions of mutants [124,125]. Individual methods aimed to generate more stable protein variants can also be combined for best results as was demonstrated by consensus design that used the sequence output of a library selection [126].

We previously reported the *in vitro* selection of *de novo* RNA ligase enzymes that catalyze a reaction not observed in nature [51]. These artificial enzymes ligate RNA with a 5'-triphosphate to the 3'-hydroxyl of second RNA forming a native 5'-3' linkage and releasing pyrophosphate. These artificial ligases are zinc dependent metalloenzymes of about 10 kDa. Several enzymes resulting from this *in vitro* selection experiment were analyzed in more detail. All examined enzymes were soluble when expressed as fusion proteins with maltose-binding protein (MBP), but most enzymes were poorly soluble when expressed on their own. NMR HSQC spectroscopy of the most soluble clone, ligase #6, revealed that a significant portion of the protein was well-folded, yet the overall resolution of the data was insufficient to solve the three-dimensional structure [51]. To overcome this issue, we again utilized *in vitro* selection. We modified the conditions of our original procedure and continued the selection to isolate ligase variants with improved stability in order to facilitate structural and mechanistic studies of these artificial enzymes.

Here, we describe in detail the *in vitro* selection of RNA ligases with increased stability. For this directed evolution experiment we utilized the mRNA display technology, an *in vitro* display method, which covalently links each protein to its encoding mRNA [69,70]. Using this technology, up to 10^{13} unique proteins can be sampled in a single experiment, which is orders of magnitude more than most other

selection strategies [124]. To isolate enzymes with increased thermodynamic stability, we modified parts of the selection procedure and performed the ligation step at 65°C. For the selection reported here, we used the output library from our previous selection at room temperature [51] as starting material. We hypothesized that enzymes, which are active at elevated temperature, will have a more stable protein fold that in turn will facilitate structural characterization. We also hoped that the increased structural stability would correspond to increased solubility and expression *in vivo*. After several rounds of selection, representative ligase clones were sequenced and tested for soluble expression in *E. coli*. The soluble and most active ligase 10C was characterized further and its activity and stability were compared to two closely related sequences from the previous selection at room temperature. The experiments revealed that ligase 10C is both more stable and more active than either of these ligases. The structure of ligase 10C will be described in chapter 3, and an application for this enzyme was previously published [127,128]. This is the first report of an mRNA display selection at high temperature. These results demonstrate the efficacy of mRNA display for isolating thermostable enzymes as stability and activity are selected simultaneously in a high throughput experiment.

2.3 Results

2.3.1 Setup of selection procedure

Sequence analysis of the artificial RNA ligase enzymes that resulted from the final round of the previous *in vitro* selection performed at 23°C [51] revealed substantial sequence diversity. The DNA encoding those diverse ligases was used as the starting library for the selection at 65°C described in this paper without introducing further sequence diversity. The RNA ligation reaction catalyzed by the previously selected enzymes is dependent on a complementary splint oligonucleotide that base-pairs to the two substrate RNAs [51] (Figure 2.1). During the selection at 23°C, this splint base-paired to eight nucleotides of each substrate (Figure 2.1B). In order to ensure stable base-pairing during a splinted ligation at elevated temperatures, a longer splint was chosen to

extend the region complementary to each substrate to twenty nucleotides (Figure 2.1C). The 40-nucleotide-long splint resulted in a melting temperature of 76°C and 69°C with the PPP-substrate and the HO-substrate, respectively (Figure S2.1).

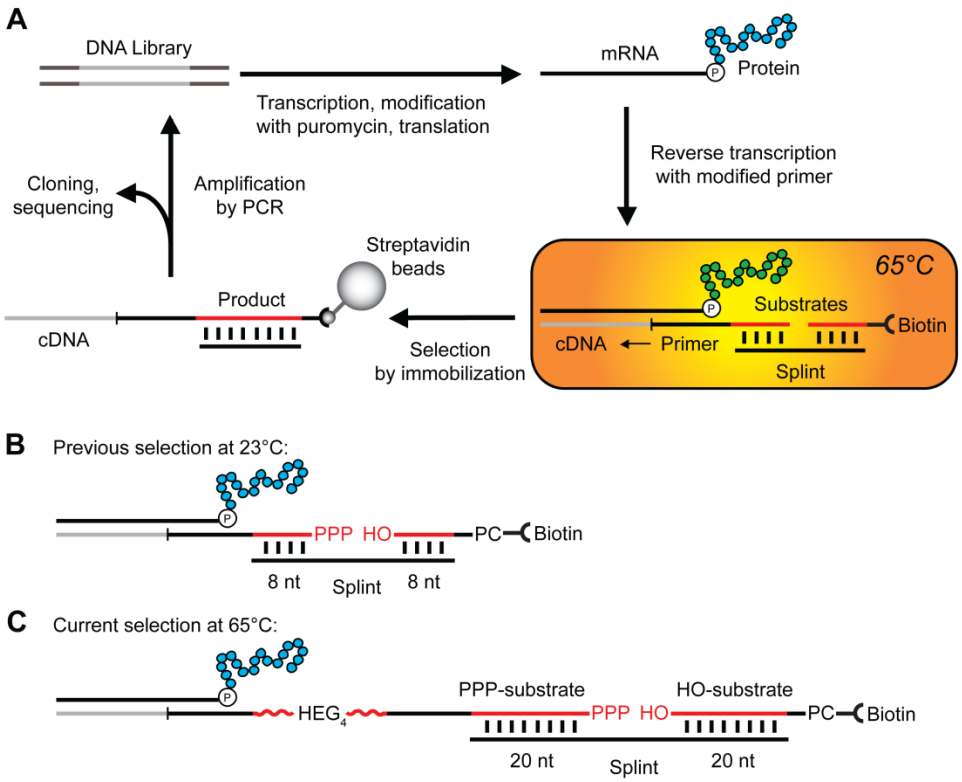

Figure 2.1-*In vitro* selection of artificial ligase enzymes with increased stability.
(**A**) Schematic of the isolation of ligase enzymes. The DNA library encodes the library of proteins that resulted from the original selection of ligase enzymes at 23°C [51,52]. The DNA is transcribed into RNA, modified with puromycin at the 3'-end and translated *in vitro* yielding a library of mRNA-displayed proteins [52]. Reverse transcription with a primer containing one RNA substrate shown in red results in a complex of protein, mRNA, cDNA and substrate. This complex is incubated at 65°C with the second RNA substrate (red) and the complementary splint as highlighted in the orange box. The cDNA of ligases active at this temperature is immobilized on streptavidin beads and amplified for subsequent rounds of selection, or identified by cloning and sequencing. (**B**) Detailed view of ligation reaction substrates in complex with the mRNA-displayed protein. The two strands of RNA in red, the 59-triphosphate RNA (PPP-substrate) and 39-hydroxyl RNA (HO-substrate), are joined in a template-dependent ligation reaction. The PPP-substrate is part of the reverse transcription primer. The photocleavable site (PC) is used to release the cDNA that encodes active enzymes from streptavidin immobilization by irradiation at 365 nm. The splint acts as template of the ligation and base pairs with 8 nucleotides of each RNA substrate during the previously published selection at 23°C [51,52], and with (**C**) 20 nucleotides of each substrate during the current selection at 65°C. HEG_4 represents the linker of four hexaethylene glycol units (red wavy line).

To enable the selection of active enzymes, the PPP-substrate was linked to the mRNA-displayed proteins via the reverse transcription (RT) primer that initiates the cDNA synthesis (Figure 2.1A). This linkage resulted in a high local concentration of substrate in the vicinity of each protein. In order to reduce this local concentration and thereby favor the selection of enzymes with an increased substrate affinity, we lengthened the RT primer by an additional eighteen non-complementary nucleotides and four flexible hexaethylene glycol linker units (HEG_4, Figure 2.1C). The hexaethylene glycol linker simply acted as a long unstructured tether to increase the average distance between protein and substrate. The use of the longer RT primer in combination with the splint of 40 nucleotides (nt) in length (Figure 2.1C) resulted in a ligase activity of about 50% compared to a ligation using the shorter RT primer and the 16 nt splint (Figure 2.1B).

We then evaluated the ligase activity of the starting library at increasing temperatures in order to determine a temperature at which the majority of the library members are inactive. Using the 40 nt splint and the HEG_4-RT primer, at 65°C no ligation was detectable (< 10%), whereas at 60°C the ligation activity was about half of the activity measured at 23°C. Therefore, we decided to carry out the selection for higher stability at 65°C.

During the previous selection for ligases, 57% of the isolated enzymes had acquired a second FLAG binding sequence (DYKXXD) in addition to the FLAG binding sequence that was part of the N-terminal constant region. This was likely a result of a selection bias caused by two FLAG affinity purification steps per round of selection. In order to counteract this FLAG purification bias, we changed the selection protocol to using the E-tag affinity purification instead. Therefore, we replaced the FLAG tag coding sequence in the N-terminal constant region of the library with an E-tag sequence by PCR. The ligation activity was unaffected by the change of tags.

2.3.2 In vitro *selection at 65°C*

To enrich for RNA ligase enzyme with increased thermostability, we performed a total of six rounds of selection and amplification (Figure 2.1A). After reverse transcription, the mRNA-displayed proteins were incubated with the HO-substrate-65 and the RNA splint for 60 min and/or 5 min. The percentage of cDNA that was immobilized on streptavidin beads after each round of selection is shown in Figure 2.2. In the case of the 60 minute incubation, the percentage of immobilized cDNA increased steadily over the course of the selection, from 0.61% after round 1 to 6.6% after round 6. In order to increase the selection pressure by favoring enzymes with faster ligation rates, in round 4, we incubated a second aliquot of the mRNA-displayed proteins for only 5 min yielding 0.66% immobilized cDNA. This cDNA was used as input for following round, but no increase in the amount of immobilized cDNA after 5 min incubation was observed in round 5 (amount decreased to 0.41%). Therefore, we performed the sixth and final round of selection, again with 60 min incubation. The resulting DNA was cloned and sequenced for further analysis.

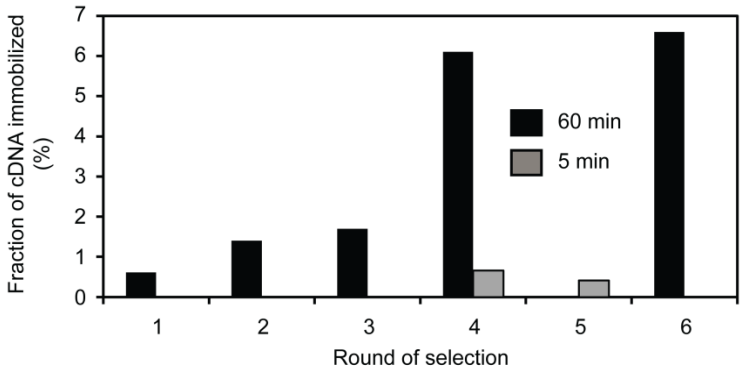

Figure 2.2-Progress of selection for ligases at 65°C.
The fraction of ^{32}P-labelled cDNA that bound to streptavidin agarose after each round of selection is shown. The reaction time was either 60 min or 5 min as indicated by black or gray bars, respectively.

2.3.3 Sequence analysis and expression of selected ligases

The sequence alignment of 32 clones from the sixth round of selection at 65°C revealed two protein families (Figure S2.2). One representative clone from each family

was cloned and expressed in *E. coli* to examine soluble expression (Figure S2.3). While both clones expressed well, ligase 10C was highly soluble whereas ligase 10H was largely insoluble. Furthermore, native Ni-NTA affinity purification of ligase 10H yielded no soluble protein (data not shown) and, therefore, ligase 10H was not characterized further.

The sequence of ligase 10C shares similarities to ligases #6 and #7 from the original selection with #7 being more similar (Figure 2.3). All three ligases are almost identical in sequence in the formerly randomized region 2, and all three share the deletion of 17 amino acids following region 1. Ligases 10C and #7 also share the sequence in region 1, but 10C contains a second deletion of 13 amino acids near the C-terminus. This C-terminal deletion is also found in other clones from the selection at 23°C [51], but these proteins were poorly soluble when expressed without an maltose-binding protein fusion and therefore unsuited for a direct comparison.

Figure 2.3-Sequence alignment of the library used as input for original ligase selection [129] with ligases #6, #7 [51] and 10C that were selected at 23°C and at 65°C, respectively [130,131].
The amino acids in regions 1 and 2 of the original library were randomized prior to the selection at 23°C and are shown as "x". Dashes symbolize amino acids that are identical to the starting library. A period highlighted in gray represents a deletion. The underlined N-terminal amino acids of the library and ligase 10C represent a Flag epitope tag and an E epitope tag, respectively.

2.3.4 Activity of ligase enzymes

To compare the enzymatic activity of ligase 10C to ligases #6 and #7, we assayed the three enzymes at 23°C and 65°C (Figure 2.4). Ligase 10C was the only enzyme active at 65°C. In comparison, ligases #6 and #7 were active at room temperature as expected, but had no measurable activity at 65°C. In addition to its activity at 65°C, ligase 10C was also active at room temperature. To compare the activity of the three enzymes more accurately, we measured the k_{obs} for each ligase at 23°C. At a subsaturating substrate concentration of 10 μM, ligase 10C had a k_{obs} of 0.165 ± 0.015 h^{-1} while ligases #6 and

#7 had k_{obs} of 0.0174 ± 0.0066 h^{-1} and k_{obs} of 0.0207 ± 0.0045 h^{-1}, respectively (Table S2.1). This represents an 8 to 10-fold increased activity of ligase 10C compared to ligases #6 and #7 even at 23°C. While the main goal of the selection was to isolate an enzyme with greater thermostability, as an added benefit, the most stable enzyme also featured an improved catalytic rate at room temperature.

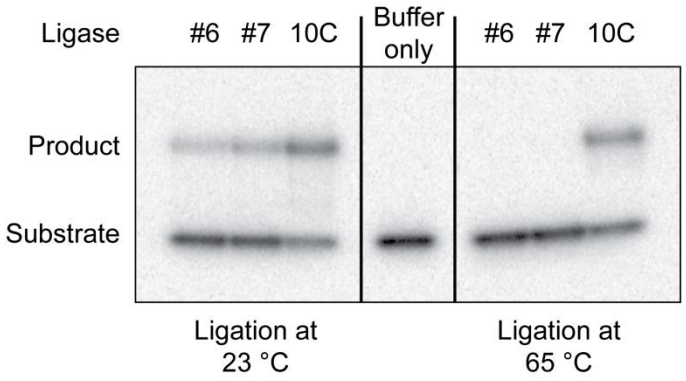

Figure 2.4-Activity of ligase enzymes assayed at different temperatures.
Ligases #6 and #7 had been selected previously at 23°C [51,52] and ligase 10C was selected at 65°C. In this assay, the ^{32}P-labeled PPP-substrate-65, HO-substrate-65 and 40 nt splint were incubated with the individual enzymes for 16 h and activity monitored by gel-shift.

2.3.5 Characterization of thermal stability by circular dichroism (CD)

In order to assess if the unique enzymatic activity of ligase 10C at 65ºC was correlated to increased structural stability, we measured thermal denaturation curves of all three ligases by circular dichroism. In preparation of the thermal unfolding experiment, we measured the CD spectra of the three enzymes (Figure S2.4). All three spectra exhibited two minima of negative ellipticity at 205 nm and between 220 and 225 nm, respectively. While those minima suggested α-helical secondary structural content [132], the 205 nm minimum was substantially more negative than the second minimum, which differs from purely alpha helical proteins that have similar absolute values for both minima. Nevertheless, we used the strong negative ellipticity of all three ligases at 222 nm to monitor thermal unfolding of the proteins over a temperature range from 5 to 91 ºC. We found all three enzymes to give the characteristic single sigmoidal transition

corresponding to a two-state unfolding reaction (Figure 2.5). As determined from the curves, the enzymes showed very different melting temperatures. Ligase 10C had the highest melting temperature (T_m = 72ºC), which was 35 degrees higher than the T_m of ligase #6 (37ºC), and 24 degrees higher than the T_m of ligase #7 (48ºC). The high melting temperature of 72ºC for ligase 10C was in agreement with its retained enzymatic activity at 65ºC as the enzyme has not undergone unfolding yet. In contrast, ligases #6 and #7 were fully denatured at 65ºC, and, therefore, their complete lack of enzymatic activity at 65ºC could be explained by their unfolding.

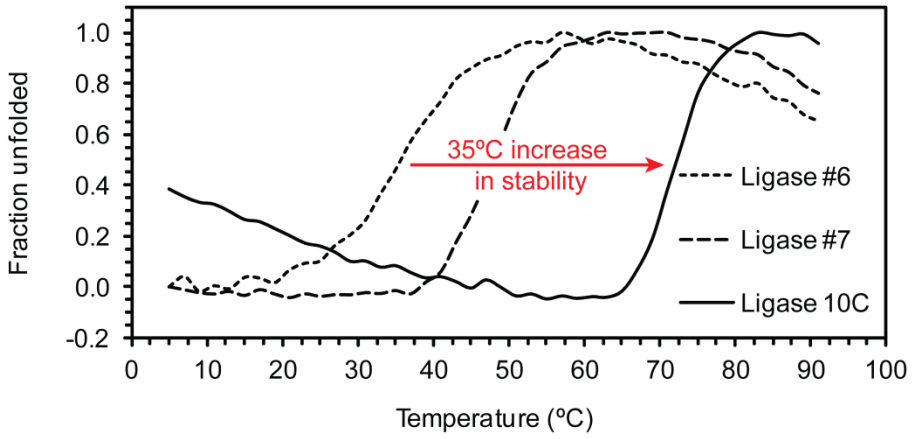

Figure 2.5-Thermal unfolding curves of ligases #6, #7 and 10C.
Thermal unfolding was monitored by circular dichroism at 222 nm. For each measurement 10 accumulations were acquired.

2.4 Discussion

We isolated a thermostable artificial RNA ligase enzyme by *in vitro* selection at 65°C of a library of artificial ligases that were originally generated at 23°C. The isolated ligase 10C was more thermostable and more active than the two most closely sequence-related ligases #6 and #7 identified during the selection at 23°C. Ligase 10C has a melting temperature (T_M) of 72°C corresponding to a stability increase of 24 degrees compare to #7, and 35 degrees compared to ligase #6. Previously reported T_M improvements through protein engineering are commonly between 2 to 15 degrees [115]. The T_M increase by 35 degrees reported here favorably compares with those rare examples of 'record-setting

stabilizations' [114,133-135]. While the ligases #6 and #7 have no measurable enzymatic activity at 65°C, ligase 10C ligates RNA at 65°C with an activity that is similar to its activity at 23°C. Furthermore, the activity of ligase 10C at 23°C is about an order of magnitude higher than the activity of the ligases #6 and #7 at the same temperature.

The increased thermostability of ligase 10C is likely due to improved intermolecular contacts within the protein compared to the mesophilic ligases #6 and #7. In contrast to these enzymes isolated at 23°C, the properties of ligase variant 10C were suitable to solve its three-dimensional solution structure by NMR [20]. The structure featured a small, well-folded core coordinated by two Zn^{2+}-ions. In addition, the folding core also contained a highly dynamic internal loop and was framed by unstructured termini. In order to discuss a potential correlation between differences in primary sequence and altered thermal stability, we mapped sequence differences between ligase #7 and 10C onto the structure of 10C (Figure 2.6). We chose ligase #7 for comparison because despite the high sequence similarity it showed a large difference in thermostability. All differences between these two ligases are found in or near the structured region responsible for zinc coordination. We previously demonstrated by NMR that residues Ile68 and His69 near the C-terminus of ligase 10C made long range NOE contacts with several residues at the N-terminus (Lys17, His18, Ala27 and Glu28)[127]. Notably, His18 is one of the zinc coordinating residues in 10C and mutating this position to Ala resulted in a drastically reduced solubility of the protein [127]. In addition, ligase #7 contained an additional 13 amino acids located between the residues corresponding to Ile68 and His69 in ligase 10C, which likely moved His69 and prevented its contacts with Lys17, His18 at the N- terminus. Presumably, all these mutations could compromise the intramolecular interactions in these positions reported for ligase 10C and decrease the stability of ligase #7 at high temperature. Ligase 10C also differed from ligase #7 in two additional positions (Ser54 and Asp65) which may further influence protein stability. A direct comparison of the overall flexibility of ligase 10C and the two mesophilic ligases would require solving also the structures of ligases #6 and #7 by

NMR. This would be beyond the scope of our characterization, and preliminary experiments suggested that ligase #6 is not amenable to detailed NMR studies.

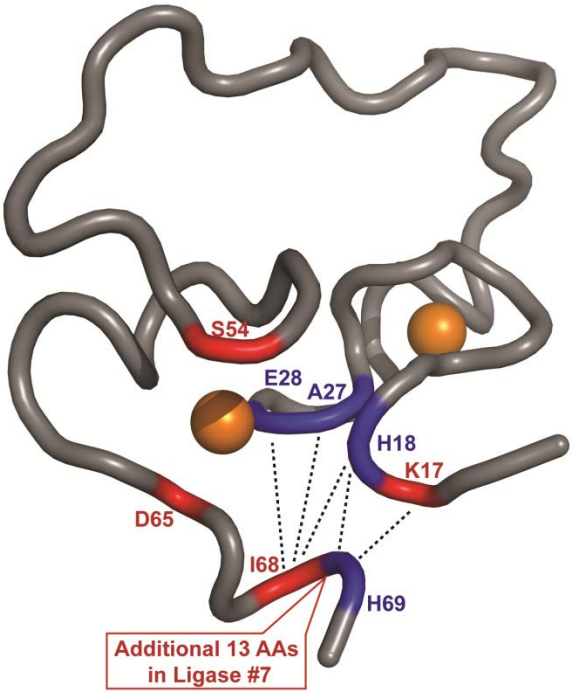

Figure 2.6-Sequence differences between ligase #7 and ligase 10C mapped onto the NMR structure of ligase 10C [127].
Mutations are shown in red. Residues potentially perturbed by the mutations are labeled in blue and long range NOEs are shown as dashed black lines. The two coordinated zinc ions as depicted as orange spheres and the residue numbers refer to ligase 10C. The unstructured termini of ligase 10C were omitted for clarity.

The in vitro selection at 65°C not only yielded the family A of related sequences that included ligase 10C (Figure S2.2), but also a second family B represented by ligase 10H which could not be expressed solubly in E. coli. During the original selection at 23°C, we noted that of the seven ligases characterized, only #6 and #7 were soluble without being expressed as a MBP fusion. While ligase 10C is closely related to #6 and #7, ligase 10H is most similar to ligase #1 which also did not express solubly. Isolating proteins like ligase 10H and ligase #1 is not surprising because mRNA display uses a eukaryotic in vitro translation system and therefore soluble expression in E. coli was never directly selected for. Additionally, the covalently linked RNA increases protein

solubility which can also contribute to this result. In general, this solubilizing effect is a favorable feature of mRNA display because it allows identifying proteins that might be lost during other selection techniques due to poor solubility. It is possible that ligase 10H could be solubilized through MBP fusion like ligase #1, but such a modification would have complicated subsequent structural studies.

Considering the high melting temperature of 72°C for ligase 10C, it is particularly surprising to discover the lack of secondary structural motifs like α-helices or β-strands combined with highly dynamic regions [127]. The structure of this artificial enzyme does appear to match with any known protein folds. While it is increasingly appreciated that catalytic activity of enzymes can require conformational flexibility [136-138], thermal stability is usually associated with tight packing and rigidity. Generally, thermophilic enzymes possess well packed hydrophobic cores [139], few exposed surface loops [140] and additional stabilizing interactions such as salt bridges [141] and a high number of hydrogen bonds [142]. These features lead to an increased rigidity that, while favoring stability at higher temperature, often appears to decrease activity at lower temperature. This observation has been interpreted to mean that stability, dynamics and catalysis are a tradeoff, but this common notion has recently been called into question [143]. The structure of the ligase 10C [127] combines a high flexibility and the absence of a packed hydrophobic core with thermostability, and is equally active at 65°C and at ambient temperature. The structure of this *de novo* enzyme challenges the common view of how enzymes are supposed to look - a view that is biased by proteins amenable to crystallization.

The high degree of disorder and flexibility present in 10C might be a feature that favors its evolvability. For example, the presence of disordered regions and a loosely packed structure found in viral proteins, structural characteristics similar to those found in 10C, may allow for increased evolvability because each mutation, due to a lower amino acid interconnectivity, would lead to a slower loss in stability, compared to the more packed structures of thermophilic enzymes [144]. Similarly, ligase 10C might also be highly evolvable because of its flexible structure and disordered regions. Yet, this

artificial enzyme was generated *de novo* and, unlike biological proteins, has not been shaped by billions of years of evolution. As its structure and function has just come into existence, ligase 10C could be considered a model protein for primordial enzymes. For these reasons, properties of this enzyme like its evolutionary potential will be interesting to study, however comparisons to natural proteins might be challenging.

The starting library for this selection at elevated temperature was a mixture of protein variants that was final the output of the previously described selection for artificial ligases at 23°C [51]. No further genetic diversity had been introduced. Sequencing of the starting library showed a diverse mixture of unrelated sequences and sequence families. Ligase 10C had not been observed during the sequencing of 49 individual clones and was only sufficiently enriched and detected after the subsequent selection at 65°C. It is conceivable that future mutagenesis and directed evolution of ligase 10C using the same selection strategy will further improve thermal stability and activity. These studies will help us understand the evolutionary potential of this artificial enzyme and also yield improved catalysts for a variety of applications [128].

2.5 Conclusions

The discovery of this thermostable enzyme and its unusual structure emphasizes the value of directed evolution approaches that do not require a detailed understanding of protein structure-function relationships, but instead randomly sample sequence space for functional proteins. In contrast, it would have been impossible to construct this particular artificial enzyme by rational design despite recent advances in rational protein engineering. In the current project, we employed the *in vitro* selection technique mRNA display [69,70]. This method uses product formation as the sole selection criterion and is independent of the mechanism of the catalyzed reaction. The technique has several advantages over other selection strategies [35]. The mRNA display technology enables to search through large libraries of up to 10^{13} protein variants. This feature is beneficial because the chance of finding a desired activity increases with the number of variants interrogated. Furthermore, the in vitro format of this method allows selecting for activity

under a wide range of conditions, which is similar to the common approach of screening much smaller libraries of purified proteins, but in contrast to in vivo selection strategies where maintenance of cell viability limits the experimental possibilities. Previous reports on mRNA display include the improvement of folding and stability of proteins by selecting for resistance to protease degradation [35], or by selecting in the presence of increasing amounts of the denaturant guanidine hydrochloride [145,146]. Interestingly, in parallel to our successful selection for RNA ligases at elevated temperature, we also attempted a similar selection in presence of guanidine hydrochloride, but no enrichment was observed even after six rounds (data not shown). Nevertheless, to our knowledge the work presented here is the first description of an mRNA display selection at elevated temperatures yielding thermostable proteins. The *in vitro* format of mRNA display should facilitate other selections at a variety of pH, temperatures, ionic strength, or in the presence of co-solvents, inhibitors or other chemicals. Such experiments will help to study the coevolution of protein stability and activity, and also has the potential to produce proteins that are more stable in industrial or biomedical applications.

2.6 Materials and methods

2.6.1 Preparation of oligonucleotides

^{32}P-labeled PPP-substrate-23 used in original selection at 23°C (5'-PPP–GGAGACUCUUU) and PPP-substrate-65 for selection at 65°C (5'-PPP–GGAGAUUCACUAGCUGGUUU) were prepared through T7 transcription as reported previously [51,52]. The HO-substrate-23 (5'–CUAACGUUCGC), HO-substrate-65 (5'-UCACACUGUCUAACGUUCGC) and HO-substrate-65-Bio (5'-(PC)-UCACACUGUCUAACGUUCGC, (PC) represents PC biotin phosphoramidite from Glen Research, Sterling VA) were purchased from Dharmacon (Lafayette, CO) and prepared according to the manufacturer's protocol. The DNA splint (5'–GAGTCTCCGCGAACGT) complementary to the substrates-23 and RNA splint (5'-AAACCAGCUAGUGAAUCUCCGCGAACGUUAGACAGUGUGA) complementary to the substrates-65 were purchased from Integrated DNA Technologies (Coralville, IA).

The reverse transcription primer (HEG_4-RT) was produced by ligating the PPP-substrate-65 to BS75P-HEG_4 in the presence of BS76 as template using T4 DNA ligase [147] and purified by denaturing PAGE. All oligonucleotides were dissolved in ultra-pure water and concentrations determined by UV absorbance.

2.6.2 Selection of RNA ligases at 65°C

The mRNA display selection was performed as previously published [51], with the following exceptions. Primers BS99 and BS24RXR2 were used to amplify the DNA by PCR. Primer BS99 replaces the N-terminal FLAG affinity tag that was used in the previous selection at room temperature [51] with the E-tag. Accordingly, both FLAG affinity purification steps in the previous protocol were substituted by E-tag affinity purifications. For the first E-tag purification, the mRNA-displayed proteins eluted from the oligo(dT)cellulose were mixed with binding buffer (same as Flag binding buffer [51]) and then incubated for 30 min at 4°C with rotation with 25 µL Anti-E affinity gel (from Anti E-tag affinity column, GE healthcare Biosciences; prewashed with E clean buffer (100 mM glycine, pH 3.0, 0.05% Tween-20) and binding buffer). The Anti-E tag affinity gel was then washed with binding buffer and eluted with binding buffer containing two equivalents of E-peptide (Bachem, Osteocalcin (7-19, human); one equivalent of E-peptide saturates the antigen sites of the antibody resin) for 3 min at 4°C. The second E-tag purification was performed in a similar fashion using 50 µL Anti-E affinity gel and 6 equivalents of E-peptide to elute. The elution from the second E-tag affinity purification was incubated with the HO-substrate-65-Bio and the RNA splint in presence of 2 mM $MgCl_2$ and 100 µM $ZnCl_2$ for 1 hour at 65°C in selection rounds 1, 2, 3 and 5. In round 4, the sample was divided into two aliquots, one of which was incubated for 1 h, and the other aliquot was incubated for 5 min. The reaction was quenched and purified on streptavidin beads as described previously [51], and the photocleaved DNA was amplified by PCR and used as input for the following round. For the starting material in round 5, the photocleaved DNA from round 4 was used that resulted from the 5 min incubation.

2.6.3 Expression and purification of RNA ligases

RNA ligases were expressed and purified as previously described [127].

2.6.4 Screening for ligase activity by gel-shift assay

5 μM ^{32}P-labeled PPP-substrate-65, 6 μM RNA splint, 7 μM HO-substrate-65, 20 mM HEPES pH 7.5, 100 mM NaCl, 100 μM $ZnCl_2$ and 1.7 μM enzyme (purified by Ni-NTA affinity chromatography [127]) were combined and incubated for 16 hours at 23°C and 65°C. Reactions were stopped by the addition of EDTA to a final concentration of 10 mM. Immediately following, the RNA was denatured for 40 min at 65°C in 7.5% formaldehyde, 58% formamide and 11.6 mM MOPS pH 7.0. Samples were separated by 20% denaturing PAGE gel containing 2% formaldehyde. The gel was analyzed using GE Healthcare (Amersham Bioscience) Phosphorimager and ImageQuant software (Amersham Bioscience). The amount of radiation in both the substrate and product bands was measured and % ligated was determined by dividing the intensity of the product band by the sum of the product and substrate bands.

2.6.5 Determination of observed rate constants (k_{obs})

5 μM enzyme (purified by Ni-NTA affinity and size exclusion chromatography [127]) was incubated with 10 μM ^{32}P-labeled PPP-substrate-23, 15 μM DNA splint, 20 μM HO-substrate-23 and ligation was monitored up to 2 hours at 23°C. Reactions were quenched with two volumes of 20 mM EDTA in 8 M urea after 0, 15, 30, 60 and 120 minutes, heated to 95°C for 4 min and separated by 20% denaturing PAGE gel. The gel was analyzed using GE Healthcare Phosphorimager and ImageQuant software (Amersham Bioscience). The rate constant (k_{obs}) was determined by taking the slope of the linear fit of % ligated over time and correcting for enzyme concentration by multiplying by the ratio of PPP-substrate to enzyme (10 uM : 5 uM or 2) resulting in a value with the unit per hour (h^{-1}). The reported values are an average of 3 independent replicates ± the standard deviation. Total conversion was < 10% for all cases.

2.6.6 Circular dichroism and thermal denaturation

Ligase enzymes (purified by Ni-NTA affinity and size exclusion chromatography [127]) were concentrated to 50 μM and dialyzed against CD buffer (150 mM NaCl, 2 mM HEPES, 0.5 mM 2-mercaptoethanol, 100 μM $ZnCl_2$). Circular dichroism spectra and thermal denaturation curves were recorded on a JASCO J-815 spectropolarimeter at 30 μM or 50 μM protein, respectively. The following parameters were used for both measurements: 1.5 nm band width, 2 seconds response time, standard sensitivity, 10 accumulations. The ellipticity at 222 nm was monitored to determine thermal denaturation curves over a temperature range from 5 to 91°C with a ramp rate of 1°C/min and a temperature pitch of 2°C.

2.7 Supplementary information

Table S 2.1- Data for determining k_{obs}.

Replicate	Ligase 10C		Ligase 6		Ligase 7	
	Slope [a]	R2	Slope	R2	Slope	R2
1	0.0894 h^{-1}	0.973	0.0124 h^{-1}	0.910	0.0128 h^{-1}	0.989
2	0.0828 h^{-1}	0.982	0.00768 h^{-1}	0.951	0.00984 h^{-1}	0.993
3	0.0750 h^{-1}	0.981	0.00599 h^{-1}	0.834	0.00840 h^{-1}	0.991

[a] Five time-points were used for ligases #6 and #7. Only the first four time-points were used for ligase 10C as the final point had begun to plateau and would have skewed the analysis.

Table S 2.2- Sequences of ligases 10C and 10H selected at 65°C; and ligases #6 and #7 selected previously at 23°C for comparison.

Ligase 10C	MGAPVPYPDPLEPRGGKHICAICGNNAEDYKHTDMDLTYTDRDYKNCESYHKCSDLCQYCRYQKDLAIHHQHHHGGSMGMSGSGTGY
Ligase 10H	MGAPVPYPDPLEPRGGKHICAICGDILDDDYDYKQTDSREGRQGFFKRTLRKDLTYSCRDYKYRESYHKCSDLCQSCRYQKALAIHHQHHHGGSMGMSGSGTGY
Ligase #6	MDYKDDDDKDGKHICAICGDTVTNTDYKTPDLTSTCRDYKNRESYHKCSDLCQYCRYQKALAMGTKREAAQEEVGSHHQHHHGGSMGMSGSGTGY
Ligase #7	MDYKDDDDKGGRHICAICGNNAEDYKHTDMDLTYTDRDYKNCESYHKCQDLCQYCRYQKALAMGIKREAVQEEVGSHHQHHHGGSMGMSGSGTGY

Table S 2.3- Sequences of oligonucleotides.

BS75P-HEG_4	5'-P-TGTACGATTCGATGACGA-HEG_4-TTTTTTTTTTTTTTTCCCAGATCCAGACATTC ("P" represents the 5'-phosphate group, "HEG_4" represents four hexaethylene glycol units (Spacer18 from Glen Research, Sterling, VA)
BS76	5'-TCGTCATCGAATCGTACAAAACCAGCTAGTGAATC
BS99	5'-TCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGAGCACCAGTCCCTTACCCTGATCCGCTGGAACCGCGTGGCGGAAAGCACATCTGC
BS24RXR2	5'-TTAATAGCCGGTGCCAGATCCAGACATTCCCATAGAACCGCCATGATGATG

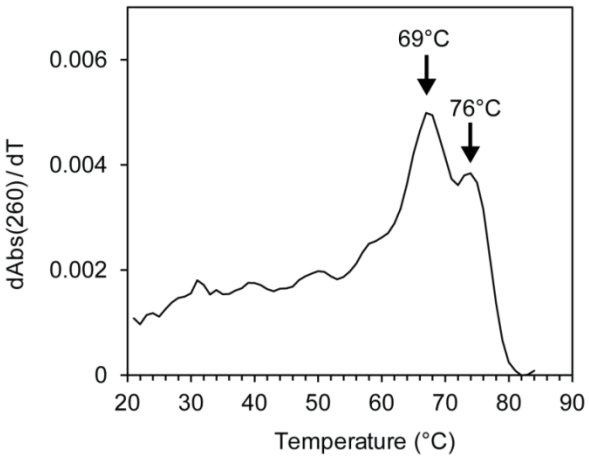

Figure S 2.1-Thermal denaturation of substrate and splint oligonucleotides used in the selection and activity assays at 65°C.

The first derivative of the melting curve for the 40 nt splint in the presence of both PPP-substrate-65 and HO-substrate-65 RNA oligonucleotides is presented. The concentration of each oligonucleotide was 0.5 µM in a buffer containing 70 mM KCl, 100 µM $ZnCl_2$, 5 mM 2-mercaptoethanol and 20 mM HEPES at pH 7.5.

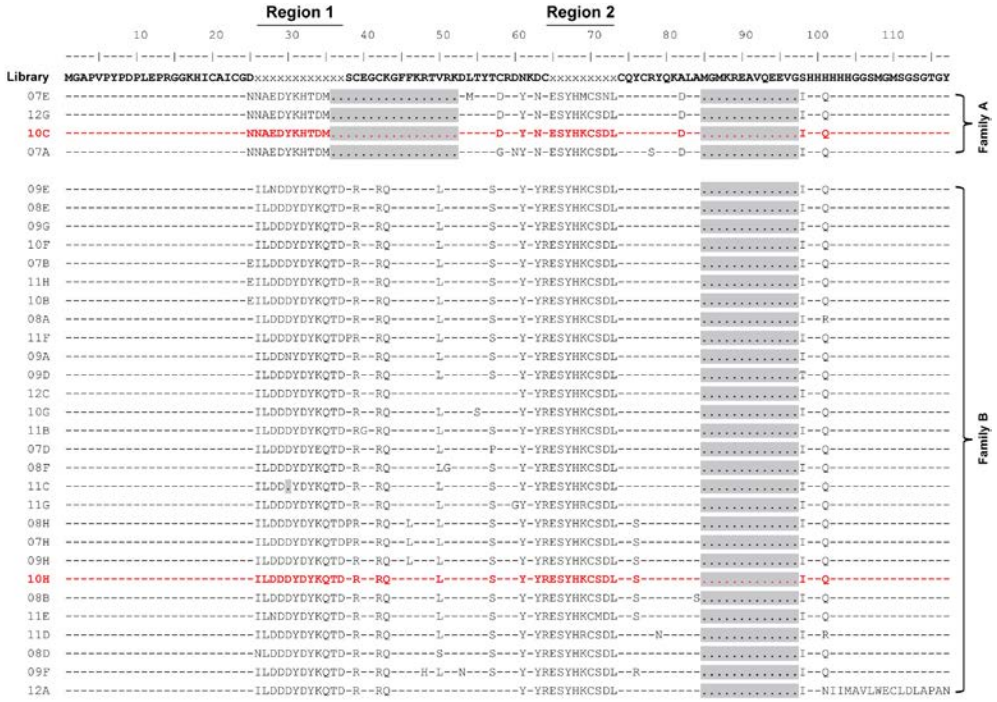

Figure S 2.2- Clones identified from round 6 of the *in vitro* selection at 65°C.
Two protein families (A, B) were identified and a representative clone from each family was chosen for further characterization (10C and 10H, shown in red).

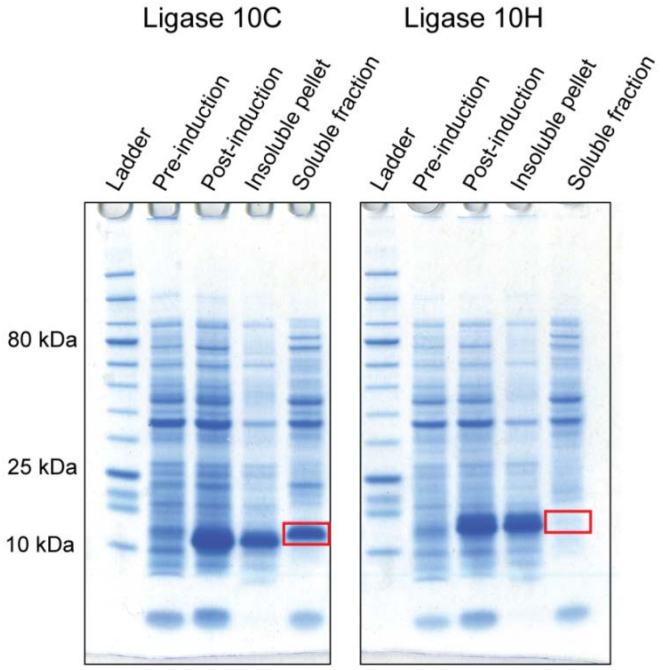

Figure S 2.3-Protein expression in E. coli of representative ligases selected at 65°C.
A Coomassie-stained SDS-PAGE gel shows samples of whole cells pre- and post-induction and the insoluble and soluble fractions after cell lysis and centrifugation. Red boxes in the lane "Soluble fraction" indicate the presence or absence of soluble ligases 10C and 10H, respectively.

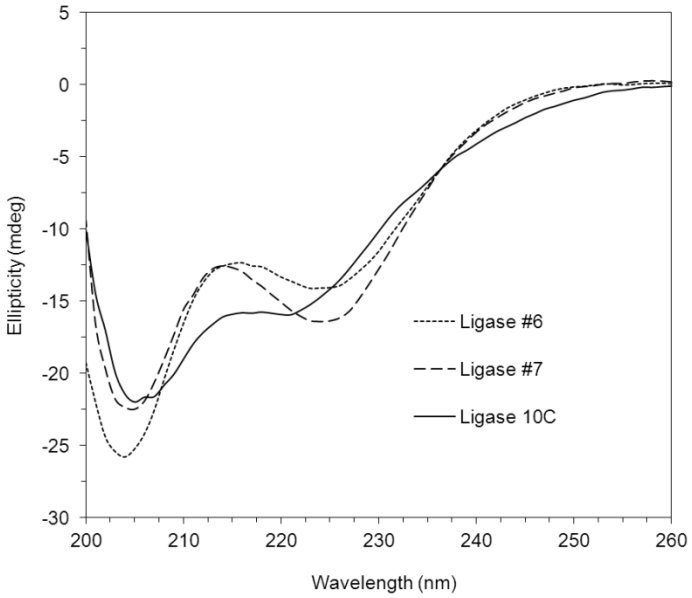

Figure S 2.4-Circular dichroism spectra of ligases #6, #7 and 10C at 25°C.

Chapter 3 : Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution

The following is a reprint of the article: Chao, F.-A., Morelli, A., Haugner III, J. C., Churchfield, L., Hagmann, L. N., Shi, L., Masterson, L. R., Sarangi, R., Veglia, G., and Seelig, B. (2013) Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution. *Nat. Chem. Biol.* **9,** 81-83. The article is reprinted here with permission from Macmillan Publishers Ltd. Chao solved the structure of the ligase by NMR, supported by Masterson and Shi. I prepared protein samples for NMR and biochemical characterization, and characterized the enzyme by CD and mass spectrometry. Churchfield and Haugner characterized mutants by activity assays and CD. Sarangi characterized the zinc coordination by EXAFS.

Hyperlink to original publication:
http://www.nature.com/nchembio/journal/v9/n2/full/nchembio.1138.html

3.1 Summary

Engineering functional protein scaffolds capable of carrying out chemical catalysis is a major challenge in enzyme design. Starting from a non-catalytic protein scaffold, we recently generated a novel RNA ligase by *in vitro* directed evolution. This artificial enzyme lost its original fold and adopted an entirely novel structure with dramatically enhanced conformational dynamics, demonstrating that a primordial fold with suitable flexibility is sufficient to carry out enzymatic function.

3.2 Introduction

The known structures of naturally occurring proteins can be assigned to an apparently finite number of different fold families [148,149]. Starting from an existing fold, divergent evolution through a combination of gene duplication and mutations is a common path for proteins to acquire new functions while retaining their original fold [150,151]. However, the origin of those biological folds remains subject to debate [152,153]. Only a few examples have been described in which new function acquisition

is accompanied by a simultaneous change in the protein fold. Those examples have largely been generated by rational design or involve protein binders [91,105,154-158].

Recently, we created artificial RNA ligase enzymes by *in vitro* evolution [51,52]. These enzymes catalyze the joining of a 5'-triphosphorylated RNA to the 3'-hydroxyl group of a second RNA, a reaction for which no natural enzyme catalysts have been found. We began with a non-catalytic small protein domain consisting of two zinc finger motifs from the DNA binding domain of human retinoid-X-receptor (hRXRα) [159] (Figure 3.1). Two adjacent loops of this protein were randomized to generate a combinatorial library of mutants as input for the selection and evolution process [129]. Although zinc fingers are common structural motifs, they are not known to take part in catalysis in natural proteins. In contrast, we isolated from the zinc finger library active enzymes that exhibit rate accelerations of more than two-million-fold [51].

3.3 Results

Sequence analysis of the artificial enzyme showed that several amino acids essential to maintaining zinc finger structure integrity were mutated or deleted, suggesting that the original scaffold may have been abandoned during the process of mutagenesis and evolution. The original hRXRα scaffold consisted of two *loop-helix* domains, each containing a zinc ion tetrahedrally coordinated by four cysteines [159]. However, during evolution of the ligase enzyme, only half of the zinc-coordinating cysteines had been conserved. In the starting scaffold, two helices were packed perpendicularly to form the globular fold and build the hydrophobic core and an additional helix was located at the C-terminus (Figure 3.1b). In the ligase enzyme, seven residues of the former DNA recognition helix and ten residues of the C-terminal helix were deleted from the original hRXRα scaffold.

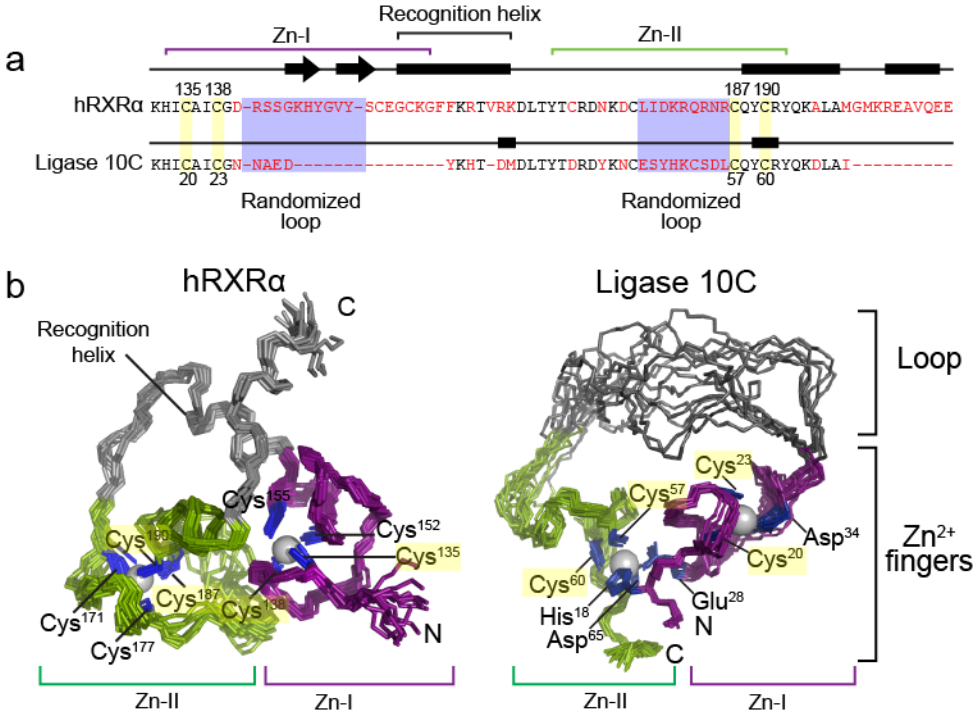

Figure 3.1-Changes in primary sequence and three-dimensional structure upon directed evolution of the hRXRα scaffold to the ligase enzyme 10C.
(a) Comparison of the primary sequences of hRXRα[159] (residues 132-208) and the artificially evolved ligase (residues 17-68). The two zinc finger regions are highlighted with purple and green brackets. The red letters denote residues that are not conserved between the two sequences. **(b)** Three-dimensional structure of hRXRα[159] and NMR ensemble of ligase 10C (for clarity, flexible termini are not shown). Although both proteins contain two zinc fingers, the overall structures are substantially different. Only two zinc-coordinating cysteines of each zinc finger in hRXRα are still coordinating zinc in ligase 10C (highlighted in yellow; also shown in **a**) whereas all other ligands differ in the two structures. Zinc-coordinating residues are labeled and shown in blue. In contrast to that in hRXRα, zinc finger Zn-II in ligase 10C comprises residues of both N- and C-terminal sequences, imposing a cyclic structure to the enzyme. Notably, the new ligase lost both helical domains of hRXRα (gray), replacing the recognition helix with a long unstructured loop (gray).

NMR structural analyses of the ligase 10C, chosen for its superior solubility and thermostability, revealed that the evolved ligase lost the original zinc finger scaffold, adopting an entirely novel structure (Figure 3.1b). This new 3D structure still contained two zinc sites that constitute the folding core of the protein, however, the two Zn^{2+} ions were coordinated by several new ligands with a different register. The deletion of two N-terminal cysteines during directed evolution resulted in the concomitant rearrangement of the local geometry of the zinc-binding loop. Additionally, the short stretch of anti-parallel β-sheet within the first zinc finger (Zn-I) was also deleted. The C-terminal loop-helix

domains and the recognition helix of hRXRα responsible for binding to the DNA groove [160] were lost completely; the latter was replaced by an unstructured loop of twenty amino acids connecting the two new zinc fingers. The zinc fingers made up the most structured region, as demonstrated by the presence of short- and long-range nuclear Overhauser Effect (NOE) contacts. Moreover, several long-range NOEs indicated that the two metal-binding loops are in close proximity, while most of the protein presented only short range NOE contacts (Figure S3.1 and Table S3.1). The conformational ensemble resulting from simulated annealing calculations showed two well-defined regions (residues 17-35 and 49-69) with root-mean-square-deviation from the average of less than 1 Å, whereas the large loop encompassing residues 36-48 was completely unstructured (RMSD greater than 6 Å). The three-dimensional structure of the enzyme was compounded by residual dipolar coupling measurements, which also helped to better define the local geometry around the zinc binding sites (Figures S3.2 and S3.3).

The two metal centers were responsible for the overall fold of the ligase. In the absence of Zn^{2+}, the NMR fingerprint spectrum of the enzyme displayed broad and mostly unresolved resonances, typical of a molten globule. Titration of Zn^{2+} to the metal-free protein first saturated the C-terminal Zn^{2+} binding site (Zn-II) and induced a substantial structural rearrangement with sharper and more dispersed resonances (Figures S3.4 and S3.5). The transition between the unfolded and folded states of the ligase involved multiple intermediate species. For selected resonances, we could discern the presence of two distinct states in slow exchange in the NMR time scale. Complete saturation with Zn^{2+} funneled the enzyme into a more defined structure, with the complete resolution of fingerprint resonances showing only one population of peaks. Elemental analysis by inductively coupled plasma mass spectrometry revealed 2.74 ± 0.01 equivalents (± s.d.) of bound zinc per ligase molecule. We were able to fit the thermocalorimetry data using models with two or more Zn^{2+} binding sites, however, the fit does not improve significantly with n>2 (Figure S3.6 and Table S3.2). Assigning two sites in accordance with the NMR titration data leads to one binding site Zn-II with higher affinity (Kd ~ 3μM), and a second binding site Zn-I with lower affinity for Zn^{2+}

(Kd ~ 93 μM). These values were further supported by the zinc concentration dependence of the enzyme activity, showing a steep drop in activity at concentrations below 100 μM Zn^{2+} (Figure S3.7). Notably, the ligase affinity for Zn^{2+} was substantially lower than those reported for natural zinc-containing proteins which commonly have dissociations constants of 10^{-8}-10^{-13} M [161]. Structure calculations were carried out in the absence of explicit Zn^{2+} ions to avoid conformational search bias and converged toward a structural ensemble with two distinct Zn^{2+} binding sites: the tetracoordinated N-terminal site (Zn-I) with weaker binding affinity, and the hexacoordinated C-terminal loop (Zn-II) with higher binding affinity (Figure S3.2 and S3.3). Extended X-ray absorption fine structure (EXAFS) data corroborated these results, showing that both Zn sites coordinated with two Sulfur ligands (cysteines) with a Zn-S distance of 2.3 Å, and at least one site had four Zn-N/O ligands while the other site had two to four. These atom ligands can be either protein based or water molecules (Figure S3.8).

The directed evolution process that yielded the artificial enzyme was based only on product formation without structural constraints [51]. As a result, the ligase enzyme evolved into a new structure with substantially increased conformational dynamics compared to the original DNA-binding scaffold [162]. In fact, ligase 10C showed an overall increase in structural plasticity and malleability. Although the two zinc fingers had heteronuclear NOEs similar to those of the original scaffold, the loop region that replaced the recognition helix had much higher flexibility, with heteronuclear NOEs below 0.5. These data indicate augmentation of conformational dynamics in the ps-ns time scale supported by longitudinal (T_1) and transverse (T_2) nuclear spin-relaxation measurements as well as hydrogen/deuterium exchange data (Figures 3.2a, S3.9 and S3.10a).

A distinct signature of the hRXRα structure is slow (microsecond-millisecond) conformational dynamics [162] (Figure S3.10b), which may be correlated with the protein's ability to optimize protein-DNA interactions. The *in vitro* evolution of hRXRα into the RNA ligase redistributed those conformational dynamics, particularly in the region N-terminal to the Zn^{2+} binding site Zn-II (residues 46-53, Figure 3.2b).

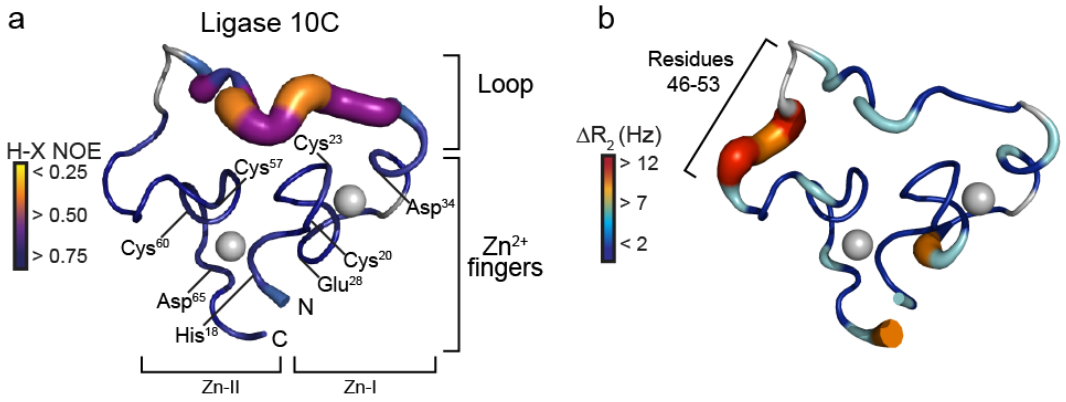

Figure 3.2-Conformational dynamics of the ligase enzyme 10C.
(a) Mapping of the heteronuclear NOEs (proxy for fast dynamics on a ps-ns time scale) on ligase 10C. Zinc-coordinating residues are labeled. **(b)** Mapping of the exchange rates (R_{ex}) obtained from relaxation dispersion measurements as proxy for slow dynamics (μs-ms timescale). Color gradient and thickness of backbone indicate that the ligase fast dynamics is located mostly in the unstructured loop, whereas the slow dynamics is located mostly in the region N-terminal to the Zn-II site (residues 46–53) and is potentially correlated to catalytic activity.

To probe the substrate binding surface, we carried out an NMR titration with a pseudo-substrate that lacked the 2'-hydroxyl group, preventing enzyme turnover (Figure S3.11). Chemical shift perturbation mapping of the ligase structure (Figures 3.3a and S3.12) indicated that one of the highly perturbed regions in the substrate bound form (residues 46-53) corresponds to high values of chemical exchange (slow conformational dynamics) in the substrate-free form (Figure 3.2b). Notably, most alanine mutations in this region decreased or completely obliterated the enzyme's activity (Figures 3.3b). Specifically, mutations E48A, Y50A and H51A abolished enzymatic activity, whereas C47A and C53A caused a 97% reduction in ligase function. These residues' high conservation among evolved ligase variants further demonstrated their importance (Figure 3.3c). The combined results suggest that this protein region (residues 46-53) is important for substrate recognition and binding and may contain the active site of the enzyme. Four of those five mutation-sensitive residues (C47, E48, H51, C53) are good potential metal ligands. Many natural enzymes, such as polymerases, that catalyze chemical reactions similar to the specific RNA ligation described here use a mechanism

involving catalytic divalent metal ion cofactors, which are coordinated jointly by the nucleic acid substrates and active site residues of the enzyme [163]. One may speculate that upon forming the enzyme-substrate complex some of the mutation-sensitive residues in ligase 10C are involved in binding additional Zn^{2+} ions that facilitate catalysis, but are not bound by the protein alone. However, additional experiments studying the enzyme in complex with substrate are needed to elucidate the catalytic mechanism of our artificial enzyme.

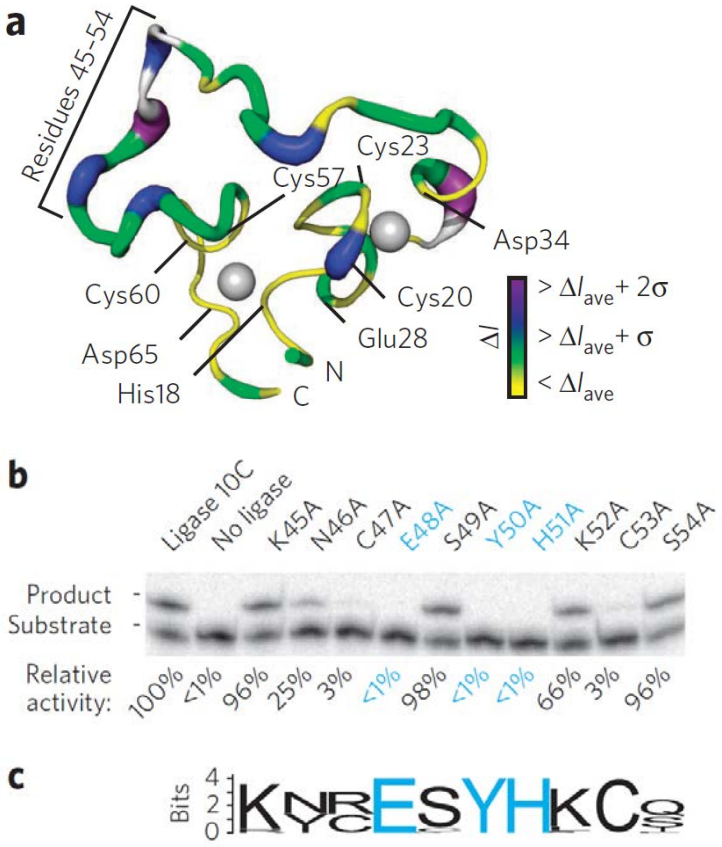

Figure 3.3-Substrate binding surface of ligase 10C probed by NMR and alanine scanning.
(a) Mapping of NMRchemical shift perturbations (intensity changes, Δ*I*) for ligase 10C shows regions affected by formation of the complex with RNAsubstrate. The most perturbed regions are indicated by thicker lines and darker colors. Zinc-coordinating residues are labeled. **(b)** Activity assay of ligase 10C and alanine mutants by gel shift. Ligation activity is normalized to the activity of ligase 10C and represents the mean value from two independent experiments. Residues with activity below detection limit are shown in blue. **(c)** Sequence conservation analysis of region 45-54 for ligase enzymes generated by directed evolution [51]. The sequences were analyzed using the web based application WebLogo (V 2.8.2) from a dataset of 49 enzyme sequences. Residues E48, Y50, and H51 (blue) which completely lost activity during alanine scanning (see above) were conserved among all sequences. The two residues with 97% reduced activity of their alanine mutant were either conserved (Cys53) or had one alternative amino acid (Cys47). None of the other residues in this region were conserved.

3.4 Discussion

The increased flexibility of the new ligase structure relative to the parent hRXRα could potentially originate from their different functional roles. hRXRα is a DNA binder and has been proposed to work through an induced-fit mechanism [159]. In contrast, the RNA ligase has evolved to function as a catalyst. This role requires additional flexibility to optimize interactions with a target molecule and carry out chemical catalysis using

transient interactions that occur in excited conformational states rather than through a stable, low energy complex [136,164,165]. This argument is supported by an independent directed evolution experiment in which the same hRXRα library yielded proteins that bind ATP and maintain the original, non-catalytic DNA binding scaffold, but have no catalytic function [129]. In contrast, the evolution of the ligase enzyme resulted in a different structure and increased dynamics.

Compared to natural enzymes which evolved over billions of years, the laboratory-evolved ligase enzyme contained substantially fewer secondary structure elements such as α-helices and β-strands, and instead exhibited increased flexibility. The complete reorganization of the starting scaffold during *in vitro* evolution may have led to the loss of these structural elements. This novel structure has not been subjected to extensive selection pressure which shaped contemporary enzymes during their natural evolution and can therefore be considered an early or primordial catalytic fold. Further evolution of this enzyme *in vitro* or inside a cell will explore if incremental mutations lead to structural and dynamic properties more similar to natural enzymes. Although flexibility has been suggested to increase the probability of developing new functions [152], it also reduces overall protein stability; a trade-off which enzymes must balance during evolution.

3.5 Conclusions

This report describes what is to our knowledge the first new protein structure emerging simultaneously with a novel enzymatic function. This ligase evolved in the absence of selection pressure to maintain the protein's original function (DNA binding). Would proteins evolving in nature also more readily adopt new folds and functions if they were freed from maintaining their original function? Although the search for such examples in nature is still ongoing, the simplified environment of *in vitro* evolution enables us to generate precedents and study basic principles of complex natural evolution. Finally, *in vitro* directed evolution has the potential to produce novel biocatalysts for a wide range of applications. The unique structure of the artificial ligase

enzyme demonstrates that this approach can successfully generate novel enzymes without being limited to known biological folds [166].

3.6 Materials and methods

All chemical compounds used in this study were purchased from Sigma-Aldrich unless noted otherwise and were of Molecular Biology Grade, and certified for the absence of ribonucleases when used for ligation reactions.

3.6.1 Sequence of RNA ligase 10C.

MGAPVPYPDPLEPRGGKHICAICGNNAEDYKHTDMDLTYTDRDYKNCESYHKC
SDLCQYCRYQKDLAIHHQHHHGGSMGMSGSGTGY

All ligase protein preparations consisted of the sequence above except for point mutations in the case of ligase mutants. Note that the sequence HHQHHH functions similarly to a 6xHis-tag.

3.6.2 Expression and purification of ^{15}N-labeled ligase protein for NMR studies.

Ligase samples were expressed in *E. coli* BL21-DE3 Rosetta strain cells (Novagen). Cells were grown in LB with 36 μg/mL kanamycin overnight at 37˚C. This culture was then used to inoculate 1 L of LB medium containing 36 μg/mL kanamycin. The cultures were grown to an OD_{600} of 0.6-0.8 at 37˚C, spun down, and resuspended in M9 minimal medium (50 mM Na_2PHO_4, 22 mM KH_2PHO_4, 8.5 mM NaCl, 2 mM $MgSO_4$, 1 mg/L thiamine, 1 mg/L biotin, 60 μM $ZnSO_4$, 10 g/L dextrose, 1 g/L $^{15}NH_4Cl$, and 36 μg/mL kanamycin, pH = 7.3). Cultures were shaken for 1 h at 37˚C, induced with 1 mM IPTG, and shaken overnight at room temperature before being spun down and stored at -20˚C.

Frozen cell pellets were resuspended in lysis buffer (20 mM HEPES, 400 mM NaCl, 100 μM $ZnCl_2$, 100 mg/L Triton X-100, 5 mM β-mercaptoethanol, pH = 7.4) and lysed using a S-450D Digital Sonifier (Branson). Cell debris was removed by centrifugation and the His-tagged ligase protein was purified by affinity chromatography using Ni-NTA Superflow resin (QIAGEN). The protein was eluted with acidic elution

buffer (20 mM NaOAc, 400 mM NaCl, 0.1 mM $ZnCl_2$, 100 mg/L Triton X-100, 5 mM β-mercaptoethanol, pH = 4.5) into 1 M HEPES at pH = 7.5, and immediately mixed to adjust the pH. Protein purification was evaluated by SDS-PAGE on Ready Gel precast gels (Bio-Rad). Elution fractions containing ligase protein were concentrated under high pressure in a stirred-cell concentrator unit with a 5,000 MWCO Ultracel Ultrafiltration cellulose membrane (Millipore) and dialyzed into FPLC buffer (20 mM HEPES, 150 mM NaCl, 0.1 mM $ZnCl_2$, and 0.5 mM β-mercaptoethanol, pH = 7.5).

Monomer ligase protein was isolated by size-exclusion chromatography using the AKTA FPLC system (GE Healthcare) equipped with a 10 mm x 300 mm column (Tricorn) and Superdex 75 resin (GE Healthcare). The separation was carried out in FPLC buffer. Fractions containing monomer protein were pooled and concentrated using 10,000 MWCO Ultra-4 Centrifugal Filter units (Millipore). Purity was assessed by SDS PAGE gel (Figure S3.13).

3.6.3 Expression and purification of $^{15}N/^{13}C$-labeled ligase samples for NMR studies.

Ligase samples were expressed in *E. coli* BL21-DE3 Rosetta strain cells (Novagen). Cells were grown in LB with 36 μg/mL kanamycin overnight at 37˚C, spun down, and resuspended in M9 minimal medium (contents as described above, except with 2 g/L [^{13}C]-dextrose). The resuspended cells were used to inoculate 100 mL of M9 minimal medium and were grown to an OD_{600} of 0.6 at 37˚C, at which time the culture was used to inoculate 900 mL of M9 minimal medium. The 1 L culture was grown to an OD_{600} of 1.0 at 37˚C, induced with 1 mM IPTG, and shaken overnight at 37˚C before being spun down and stored at -20˚C. The $^{15}N/^{13}C$-labeled protein was purified in the same manner as the ^{15}N-labeled protein samples.

3.6.4 Expression of selectively labeled ligase protein for NMR studies.

Ligase samples were expressed in *E. coli* BL21-DE3 Rosetta strain cells (Novagen). Cells were grown in LB with 36 μg/mL kanamycin overnight at 37˚C, spun down, and used to inoculate 1 L of selectively labeled M9 medium (40 mM Na_2PHO_4, 22

mM KH_2PHO_4, 8.5 mM NaCl, 1 mM $MgSO_4$, 50 μM $CaCl_2$, essential vitamins and minerals, and 36 μg/mL kanamycin, pH 7.0). To the medium was also added 250 mg of a single ^{15}N-labeled amino acid (cysteine, leucine, lysine or tyrosine), 600 mg of the remaining 19 unlabeled amino acids and, except for labeling [^{15}N]cysteine, one of the following additional amino acid supplements: 900 mg glutamine, asparagine and arginine when labeling [^{15}N]lysine; 900 mg valine and isoleucine when labeling [^{15}N]leucine; and 900 mg phenylalanine, tryptophan, alanine, serine, glycine and cysteine when labeling [^{15}N]tyrosine. Cultures were grown to an OD_{600} of 1.0 at 37˚C, induced with 1 mM IPTG, and shaken for 6 h at 37˚C before being spun down and stored at -20˚C. Selectively labeled protein was purified in the same manner as the ^{15}N-labeled protein samples.

3.6.5 Generation of ligase mutants.

Ligase mutants were obtained by site-directed mutagenesis (QuikChange Lightning, Agilent). Plasmid DNA was purified using the QIAprep Spin Miniprep kit (QIAGEN). The ligase mutants were verified by DNA sequencing. The primer sequences used to generate the indicated mutations in the ligase were designed in accordance with the QuikChange Primer Design tool (Agilent) and were as follows:

```
K45A F:     5'-CTACACCGATCGAGACTACGCGAATTGTGAGAGCTACC
K45A R:     5'-GGTAGCTCTCACAATTCGCGTAGTCTCGATCGGTGTAG
N46A F:     5'-CCGATCGAGACTACAAGGCTTGTGAGAGCTACCATAAGTG
N46A R:     5'-CACTTATGGTAGCTCTCACAAGCCTTGTAGTCTCGATCGG
C47A F:     5'-CCGATCGAGACTACAAGAATGCTGAGAGCTACCATAA
C47A R:     5'-TTATGGTAGCTCTCAGCATTCTTGTAGTCTCGATCGG
E48A F:     5'-GACTACAAGAATTGTGCGAGCTACCATAAGTGCTCGG
E48A R:     5'-CCGAGCACTTATGGTAGCTCGCACAATTCTTGTAGTC
S49A F:     5'-AGACTACAAGAATTGTGAGGCCTACCATAAGTGCTCGGAC
S49A R:     5'-GTCCGAGCACTTATGGTAGGCCTCACAATTCTTGTAGTCT
Y50A F:     5'-CTACAAGAATTGTGAGAGCGCCCATAAGTGCTCGGACTTGTG
Y50A R:     5'-CACAAGTCCGAGCACTTATGGGCGCTCTCACAATTCTTGTAG
H51A F:     5'-CTACAAGAATTGTGAGAGCTACGCTAAGTGCTCGGACTTGTG
H51A R:     5'-CACAAGTCCGAGCACTTAGCGTAGCTCTCACAATTCTTGTAG
K52A F:     5'-ACAAGAATTGTGAGAGCTACCATGCGTGCTCGGACTTGTGC
K52A R:     5'-GCACAAGTCCGAGCACGCATGGTAGCTCTCACAATTCTTGT
C53A F:     5'-GTGAGAGCTACCATAAGGCCTCGGACTTGTGCCAGT
C53A R:     5'-ACTGGCACAAGTCCGAGGCCTTATGGTAGCTCTCAC
S54A F:     5'-GTGAGAGCTACCATAAGTGCGCGGACTTGTG
S54A R:     5'-CACAAGTCCGCGCACTTATGGTAGCTCTCAC
```

3.6.6 Expression and purification of ligase mutants.

Ligase mutants were expressed in *E. coli* BL21-DE3 Rosetta cells (Novagen). Cells were cultured in 1 l of LB medium with 36 μg/mL kanamycin to a D_{600} of 0.8–1.0 at 37°C. Cultures were induced with 1 mM IPTG and shaken for 6 h at 37°C before being spun down and stored at −20°C. Ligase mutant proteins were purified by Ni-NTA affinity chromatography in the same manner as described for the ^{15}N-labeled protein samples.

3.6.7 Analysis of metal content by ICP-MS.

Ligase 10C was purified as described for the ^{15}N-labeled protein samples and then dialyzed three times against buffer (100 mM NaCl, 10 mM β-mercaptoethanol and 20 mM TrisHCl at pH 7.5; pretreated with Chelex 100 beads (Bio-Rad) for 2 h and filtered) at a ratio of 1/1,000. The metal content of 14 μM protein was measured by ICP MS (Thermo Scientific XSERIES 2 ICP-MS with ESI PC3 Peltier-cooled spray chamber at the Department of Earth Sciences, University of Minnesota).

3.6.8 Ligase activity assay for zinc dependence.

Zinc was removed from ligase 10C by treatment with ion-exchange resin (Chelex 100, Bio-Rad). Ligase 10C (5 μM) was incubated with 20 μM HO-substrate, 10 μM ^{32}P-labeled PPP-substrate base-paired to splint, 20 mM HEPES (pH 7.5), 150 mM NaCl, 500 μM β-mercaptoethanol and the concentrations of $ZnCl_2$ as indicated in (Figure S3.7) for 6 h at room temperature. The ligation reactions were quenched with 20 mM EDTA in 8 M urea, heated to 95°C for 4 min and separated by denaturing PAGE gel. The gel was analyzed using a GE Healthcare (Amersham Bioscience) Phosphorimager and ImageQuant software (Amersham Bioscience).

3.6.9 Ligase activity assay of 10C and alanine mutants.

5 μM Ligase 10C (or alanine mutant) was incubated for 6 h at room temperature in the presence of 20 μM HO-substrate, 10 μM ^{32}P-labeled PPP-substrate/splint, 24 mM HEPES (pH 7.5), 130 mM NaCl, 100 μM β-mercaptoethanol, and 120 μM $ZnCl_2$. The

ligation reactions were quenched with 20 mM EDTA and 8 M urea, heated to 95°C for 4 min, and separated by denaturing PAGE gel. The gel was analyzed using a GE Healthcare (Amersham Bioscience) Phosphorimager and ImageQuant software (Amersham Bioscience).

3.6.10 Resonance assignment.

All NMR data were collected at the University of Minnesota NMR Center. NMR spectra were acquired at 298 K on a Bruker spectrometer equipped with a cryoprobe at 700 MHz and a Varian spectrometer at 600 MHz. The samples were in a buffer of 150 mM NaCl, 20 mM HEPES and 10 mM β-mercaptoethanol, pH 7.5. Moreover, all protein samples were saturated with $ZnCl_2$ by observing changes in HSQC spectra before other NMR experiments. Triple resonance spectra such as CBCA(CO)NH and HNCACB [167-169] were used to assign peaks on ^{15}N-HSQC. All resonances in these two three-dimensional spectra and ^{15}N-HSQC were picked and fed into the PISTACHIO program (National Magnetic Resonance Facility in Madison, WI, USA) [170] to obtain preliminary assignments. Final complete assignments were done by manual checks and searches. Carbonyl groups and others side-chain carbons were assigned by HNCO and C(CO)NH-TOCSY[171]; side-chain protons were assigned by ^{15}N-NOESY-HSQC, ^{15}N-TOCSY-HSQC, and HC(CO)NH-TOCSY experiments [171] with mixing times of 150 ms, 60 ms and 12 ms, respectively.

3.6.11 Distance restraints.

All proton distance restraints were determined from the cross-peak intensities in the NOESY spectra by calibration with HN(i)Hα(i-1) distances located at the C-terminal region [172], whose helix propensity was shown by chemical shift index and $^{3}J_{\mathrm{HNH\alpha}}$ coupling values [172,173]. The cross-peaks from HN(*i*)Hα(*i*–1) distances in that region were categorized as medium NOEs, so the intensities of other cross-peaks smaller than this intensity range were defined as weak NOEs, and those larger than this range belonged to strong NOEs. The upper bounds of distance restraints of strong, medium and weak NOEs were given as 2.9 Å, 3.5 Å and 5 Å, respectively, and lower bounds were set

to 1.8 Å in all cases. Starting from unambiguously assigned NOEs at the beginning of calculation, miscalibrated NOEs were adjusted, and then ambiguously assigned NOEs were gradually added into the restraint table during iterative calculation.

3.6.12 Torsion angle restraints.

Backbone ϕ angle restraints were acquired from the HNHA experiment, and the quantitative $^{3}J_{\mathrm{HNH\alpha}}$ coupling values were calculated from the intensity ratios of cross-peaks to diagonal peaks and corrected by 3.7% to account for relaxation [174]. The correction is proportional to the rotational correlation time of the protein (3 ns), which was measured from one-dimensional TRACT experiment[175]. The ϕ angle of residue *i* with J-coupling larger than 8.5 Hz was restrained from −160° to −80°, and that with J-coupling smaller than 6 Hz was restrained from −90° to −40°. Moreover, the ψ angle restraints were derived from the ^{15}N-NOESY data. If the intensity ratio of the HN(*i*)Hα(*i*) cross-peak to the HN(*i*)Hα(*i*–1) peak is smaller than one, the ψ(*i*-1) is restrained from 20° to 220°; otherwise, the ψ(*i*–1) is restrained from 80° to −140° [176,177].

3.6.13 RDC measurement.

The stability of several alignment media for ligase 10C was tested. Using 5% neutral and negatively charged acrylamide gel was first attempted, but only weak residual dipolar couplings (absolute values <5 Hz) were obtained. Additionally, the samples precipitated in both DMPC-D7PC and DMPC-D6PC bicelle preparations. We also tested the liquid crystalline medium formed by cetylpyridinium chloride (CPCl) and 1-hexanol but had poor results in terms of sample stability. The sample was finally aligned in the other liquid crystalline medium made by the mixture of C12E5 (5% alkylpoly(ethylene glycol)) and 1-hexanol (r = 0.85) [178]. The residual dipolar couplings of amide groups were obtained by measuring the splitting difference between a decoupling HSQC peak and a TROSY peak in isotropic solution and anisotropic medium.

3.6.14 Structure calculations.

Simulating annealing protocols were performed in the XPLOR package [179]. An extended structure was first generated, and the initial temperature was set at 3,500 K, then the temperature was cooled down to 0 K with 15,000 steps. The structure with the lowest energy was used for refinement with the initial temperature of 5,000 K and 30,000 steps. The resulting structure was further refined with RDC data after optimization of the parameters Da and Rh. The angle restraints of the zinc coordination geometry were based on ideal geometries derived from X-ray data [180,181], which are in quantitative agreement with the EXAFS experiments. Distances derived from EXAFS have previously been used as restraints in NMR refinement [182,183]. Here, we report a structural ensemble of 20 conformers. The PROCHECK statistics show that 76.4% of residues are in most favored regions and 21.1% of residues are in allowed regions.

3.6.15 Zn K-edge EXAFS.

Ligase 10C protein was fully saturated with excess Zn^{2+} and then dialyzed to remove excess Zn^{2+}. The final protein sample was 1.39 mM in 15 mM Tris, pH 7.5 and 112.5 mM NaCl. Glycerol (20% v/v) was added to the protein samples to form a glass required for the EXAFS experiments. The Zn K-edge X-ray absorption spectra of ligase 10C were measured at the Stanford Synchrotron Radiation Lightsource (SSRL) on the 16-pole, 2-T Wiggler beamline 9-3 under standard ring conditions of 3 GeV and ~200-mA ring current. A Si(220) double-crystal monochromator was used for energy selection. Another optical component used for the experiment was a cylindrical Rh-coated bent focusing mirror. Spectra were collected in the fully tuned configuration of the monochromator. The solution samples were immediately frozen after preparation and stored under liquid N_2 until measurement. During data collection, the samples were maintained at a constant temperature of ~6 K using an Oxford Instruments CF 1208 liquid helium cryostat. Data were measured to k = 16 $Å^{-1}$ by using a Canberra Ge 100-element monolith detector. Internal energy calibration was accomplished by simultaneous measurement of the absorption of a Zn foil placed between two ionization chambers

situated after the sample. The first inflection point of the foil spectrum was fixed at 9,660.7 eV. Data presented here are a 15-scan average. The data were processed by fitting a second-order polynomial to the pre-edge region and subtracting this from the entire spectrum as background. A five-region spline of orders 2, 3, 3, 3 and 3 was used to model the smoothly decaying post-edge region. The data were normalized by subtracting the cubic spline and assigning the edge jump to 1.0 at 9,680 eV using the Pyspline program [184]. Theoretical EXAFS signals $\chi(k)$ were calculated by using *FEFF* (Macintosh version 8.4) [185-187]. Initial model was based on the $Zn(Cys)_2(His)_2$ active site in a zinc finger protein (1MEY)[188]. On the basis of the preliminary fits, the models were modified to accommodate a six-coordinate active site (4 Zn-N/O and 2 Zn-S(Cys)).

The theoretical models were fit to the data using EXAFSPAK [189]. The structural parameters varied during the fitting process were the bond distance (R) and the bond variance σ^2, related to the Debye-Waller factor resulting from thermal motion and static disorder of the absorbing and scattering atoms. The nonstructural parameter E_0 (the energy at which $k = 0$) was also allowed to vary but was restricted to a common value for every component in a given fit. Coordination numbers were systematically varied in the course of the fit but were fixed within a given fit.

3.6.16 Accession codes

Protein Data Bank: the accession code for ligase 10C is 2LZE. Biological Magnetic Resonance Data Bank: the accession number for ligase 10C is 18749.

3.7 Supplementary information

Table S 3.1- Summary of NMR structural statistics of 20 conformers.
The RMSD of the structural ensemble is calculated within well-structured regions (residues 17-35 and 49-69).

	Protein
NMR distance and dihedral constraints	
Distance constraints	
Total NOE	354
Intra-residue	106
Inter-residue	248
Sequential ($\|i - j\| = 1$)	162
Medium-range ($\|i - j\| < 4$)	34
Long-range ($\|i - j\| > 5$)	52
Intermolecular	0
Hydrogen bonds	0
Total dihedral angle restraints	34
ϕ	15
ψ	19
Total RDCs *Q (%)*	26 14.6
Structure statistics	
Violations (mean and s.d.)	
Distance constraints (Å)	0.1 (0.01)
Dihedral angle constraints (º)	1.3 (0.4)
Max. distance constraint violation (Å)	0.8 (0.4)
Max. dihedral angle violation (º)	5.7 (1.6)
Deviations from idealized geometry	
Bond lengths (Å)	0.008
Bond angles (º)	1.0
Impropers (º)	0.5
Average pairwise r.m.s. deviation** (Å)	
Heavy	1.4
Backbone	0.8

Table S 3.2- Thermodynamic parameters for Zn^{2+} binding determined by Isothermal Titration Calorimetry.

After completely removing the Zn^{2+} from the protein with Chelex 100 chelating ion exchange resin (BioRad), 5 μM ligase enzyme was slowly titrated with 400 μM $ZnCl_2$ solution, and the heat release was monitored using a MicroCal VP-ITC instrument (GE Healthcare). All samples contained at 150 mM NaCl, 20 mM HEPES, 10 mM β-mercaptoethanol, and pH 7.5, and the data were fitted with a sequential two-site binding model. The values represent the average of three measurements.

	Average value	**Standard deviation**
K_d1 (μM)	3.0	0.6
ΔH1 (kcal/mole)	122.9	14.8
ΔS1 (cal/mole/°)	437.7	49.7
K_d2 (μM)	92.8	8.9
ΔH2 (kcal/mole)	-123.7	13.3
ΔS2 (cal/mole/°)	-396.3	45.0

Table S 3.3- EXAFS least squares fitting results for ligase 10C.

Coordination/Path	**R(Å) [a]**	**σ^2(Å^2)[b]**	**E_0 (eV)**	***F* [c]**
4 Zn-N	2.00 (0.005)	731	-12.3	0.18
2 Zn-S	2.30 (0.003)	453		
6 Zn-C	3.00 (0.015)	1,077		
12 Zn-C-N	3.09 (0.034)	1,077 [d]		
6 Zn-C	4.16 (0.012)	169		
6 Zn-C-N	4.19 (0.013)	169 [d]		
6 Zn-C-N	4.30 (0.014)	169 [d]		

[a] The estimated standard deviations for the distances were calculated by EXAFSPAK and are given in parentheses (see also Figure S3.8 caption).

[b] The σ^2 values are multiplied by 10^5.

[c] Error is given by $\Sigma[(\chi_{obsd} - \chi_{calcd})^2 k^6]/\Sigma[(\chi_{obsd})^2 k^6]$.

[d] The σ^2 value for the Zn-C (single scattering) and Zn-C-N (multiple scattering) paths were linked to be the same value.

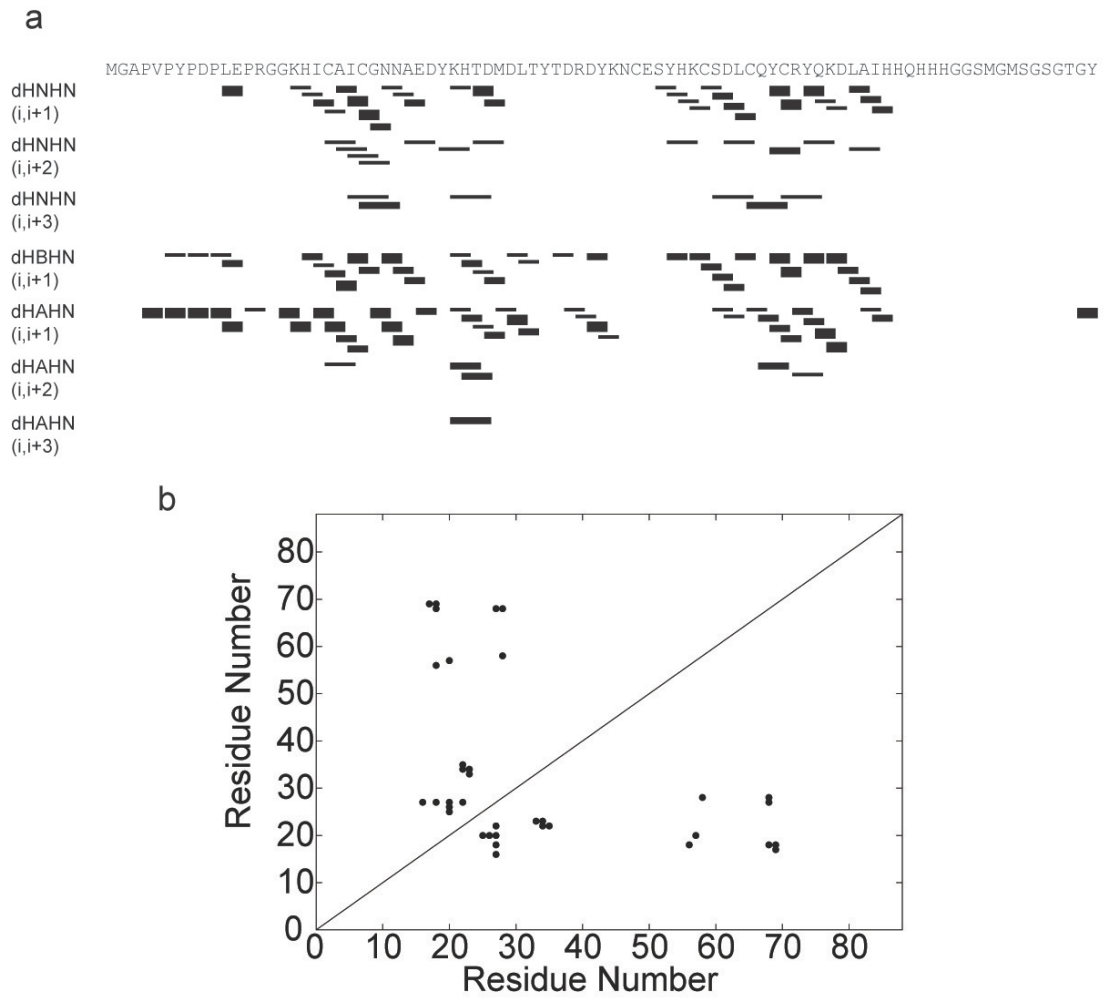

Figure S 3.1-Summary of the NOEs observed from NOESY spectra.
(a) Horizontal bars show the presence of NOE signals between residues. Bar thickness corresponds to the NOE intensity. **(b)** Long-range NOEs observed in ligase 10C.

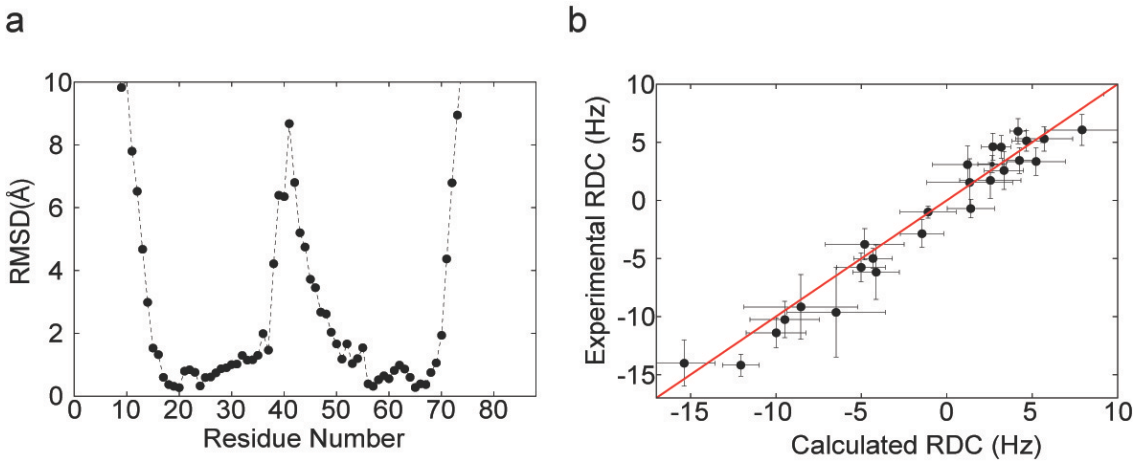

Figure S 3.2-Convergence of the structural ensemble of 20 conformers.
(a) Average backbone RMSD of the conformational ensemble. **(b)** Correlation between experimental RDC values and average back-calculated RDC values from the ensemble.

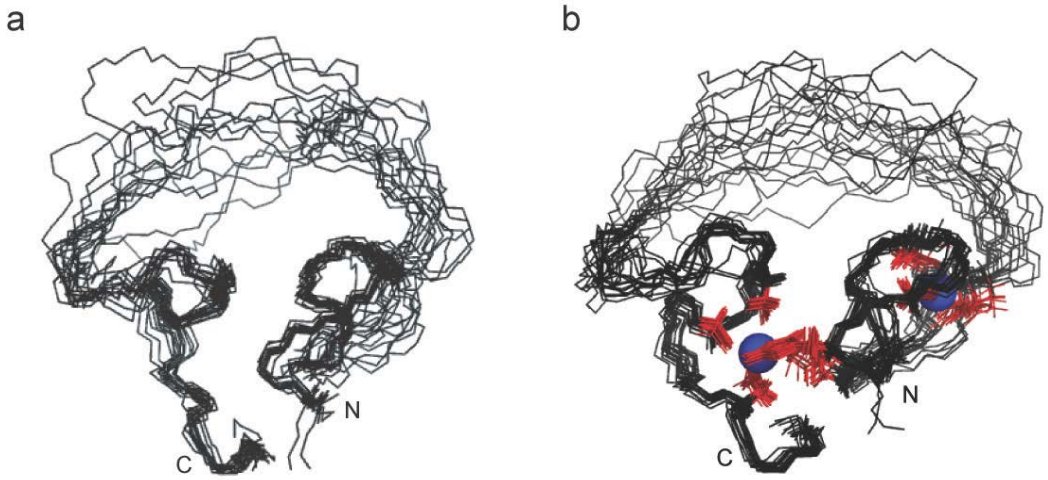

Figure S 3.3-The structural ensembles are calculated before and after incorporating Zn^{2+} ions into the coordinates.
(a) Ensemble of 20 lowest energy conformers (residues 17-69) selected from 100 structures. The NOEs, torsion angles, and RDC values were included in these calculations. Note that the two Zn^{2+} ions and coordination were not included. **(b)** Ensemble of 20 lowest energy conformers (residues 17-69) obtained including Zn^{2+} ions and coordination. The two Zn^{2+} ions are shown as blue spheres. The side chains involved in the coordination are displayed in red (H18, E28(OD2), C57, C60, D65(OD1), and C20, C23, D34(OD1)) and, additionally, both zinc binding sites each contain a single water ligand (not shown) resulting in a hexacoordinated and tetracoordinated sites, respectively. The backbone RMSD between the well-structured regions (residues 17-35 and 49-69) of the ensembles with Zn^{2+} and without Zn^{2+} is 0.52 Å.

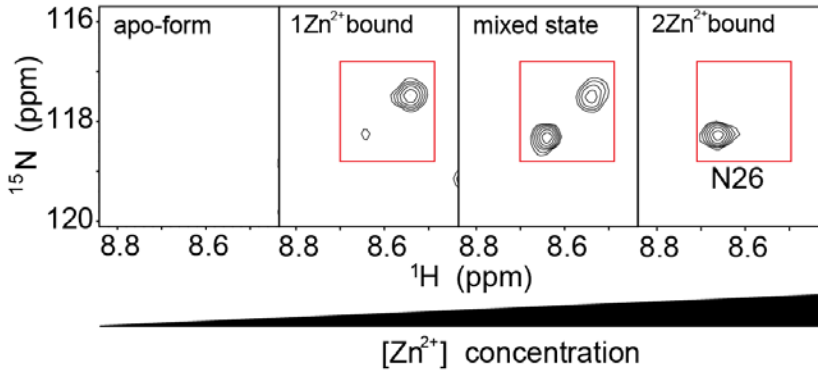

Figure S 3.4-Zn^{2+} titration into ^{15}N-labeled ligase 10C monitored by NMR.
A selected region of HSQC spectra recorded during Zn^{2+} titration is shown. Residues of partially Zn^{2+}-saturated sample displayed slow exchange on the NMR time scale between forms bound to one or two Zn^{2+}.

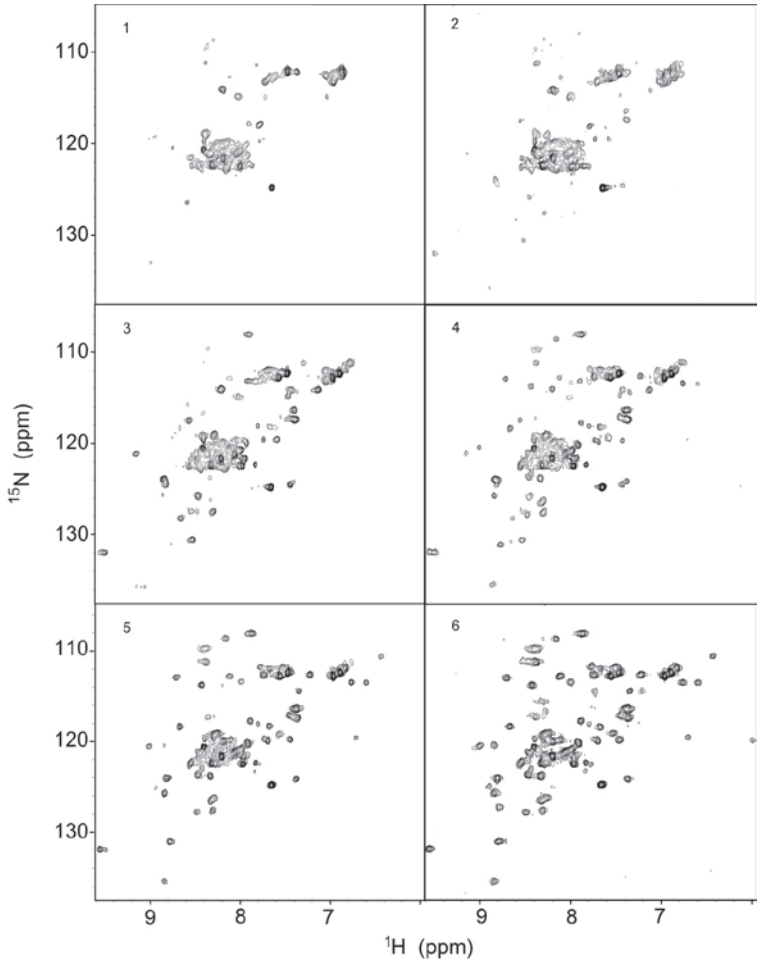

Figure S 3.5-HSQC spectra recorded upon Zn^{2+} titration.
Molar ratios of ligase 10C to zinc were: 1) 10C:Zn=1:0, 2) 10C:Zn=1:1, 3) 10C:Zn=1:2, 4) 10C:Zn=1:3, 5) 10C:Zn=1:4, 6) 10C:Zn=1:6.

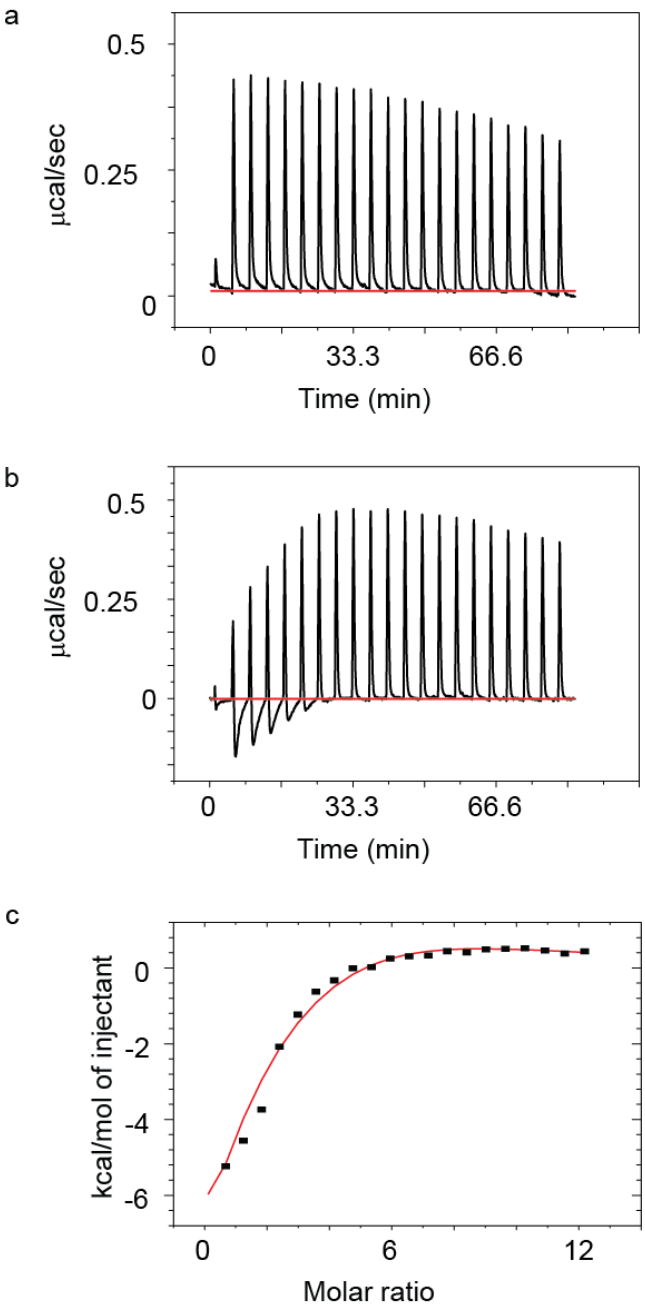

Figure S 3.6-Zn^{2+} titration into the ligase enzyme monitored by ITC.
Samples contained 5 µM ligase 10C, 150 mM NaCl, 20 mM HEPES, 10 mM β-mercaptoethanol, pH 7.5 and were measured by Isothermal Titration Calorimetry using a MicroCal VP-ITC instrument (GE Healthcare). **(a)** The graph represents the raw data for the blank titration (buffer without ligase). **(b)** The graph represents raw data for the Zn^{2+} titration of ligase 10C. **(c)** The figure shows the heat release of the Zn^{2+} titration of ligase 10C after subtracting the blank titration. The data is fit to a model of two binding sites. The data can be fitted to models with two or more Zn^{2+} binding sites, however, the fit does not improve significantly with n>2. The Zn^{2+} titration was carried out in triplicate and the errors are summarized in the Table S3.2.

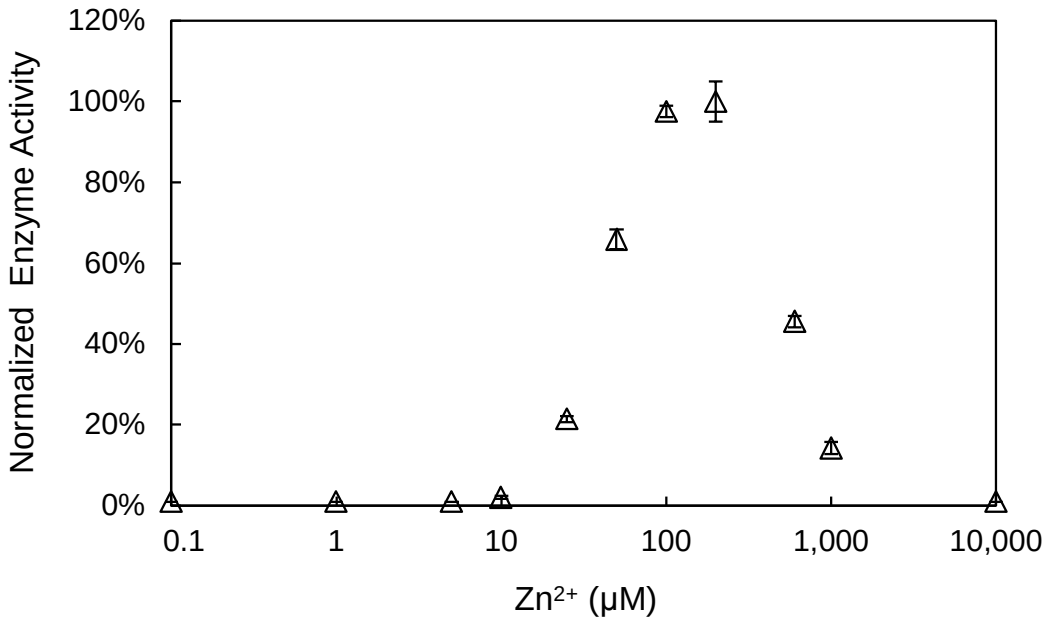

Figure S 3.7-Zn^{2+} dependence of ligase activity.
The maximum activity was observed at 145 µM Zn^{2+}. Towards lower Zn^{2+} concentrations the activity sharply drops, matching the expected behavior predicted from the dissociation constants measured by Thermocalorimetry. Towards higher Zn^{2+} concentrations, the activity also decreases but more slowly. One possible explanation is that Zn^{2+} at high concentrations might also bind to additional sites with lower affinity thereby reducing the activity. Error bars represent one standard deviation. Ligation activity for samples at 0.1, 1, 5 and 10,000 µM $ZnCl_2$ was below the detection limit of 1%.

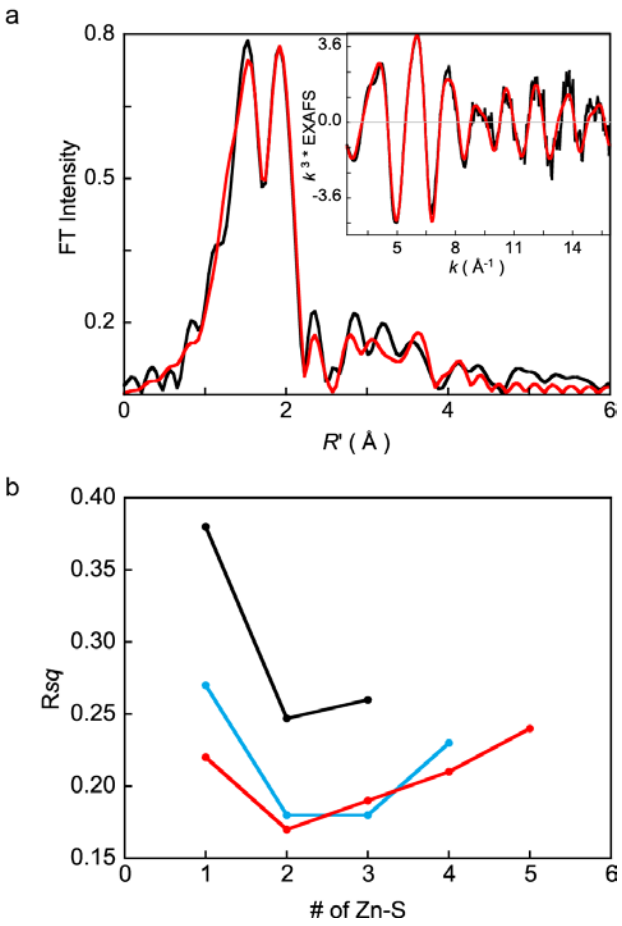

Figure S 3.8-Analysis of zinc coordination by EXAFS spectroscopy (Extended X-ray Absorption Fine Structure).

(a) The k^3 weighted Zn K-edge EXAFS (inset) and their corresponding non-phase shift corrected Fourier transforms for ligase 10C are presented. The experimental data are shown as black lines and the fit as red lines. The best-fit parameters are given in Table S3.3. While a coordination with four ligands is most commonly observed for zinc ions, a coordination geometry including six ligands has been observed in natural proteins numerous times [180,190]. The first shell coordination number was varied from four-coordinate to six-coordinate. In each case the number of Zn-S and Zn-N/O components was systematically varied to obtain the best *F* value. These fits show that the data are most consistent with a six-coordinate site with 2 Zn-S and 4 Zn-N/O components. A 1:1 occupation of the two sites modeled from NMR analysis would have resulted in best-fit with 3 Zn-N/O and 2 Zn-S coordination. However, the process of dialysis (removal of excess Zn is necessary for EXAFS experiments) may lead to stripping of some Zn from the weakly bound N-terminal site. This leads to an increase in the number of six-coordinate sites over four-coordinate sites in the protein and results in a best-fit first shell with more than 3 Zn-N/O paths. The EXAFS data are dominated by first shell Zn-N/O and Zn-S, while second and third shells are significantly weaker. The second and third shells were fit with single (Zn-C) and multiple-scattering (Zn-C-N) theoretical paths generated using a representative Zn-N(His) model. These weak features are due to a combination of single and multiple scattering from the amino acid ligands. The multiple scattering features are different from characteristic Zn-N(His)n [191] or ZnS(Cys)n [192] systems due to interference between second shell components of Cys and His ligands. Note that standard deviations in bond distances obtained from EXAFSPAK assume the use of raw, low-noise data. Although the data quality presented here are quite high, it is important to note that in the presence of several single and multiple scattering paths, the

choice of a specific path to represent an average of multiple paths will also affect the standard deviations. Typically second shell paths have errors of the order 0.05 to 0.1 Å. Furthermore these standard deviations do not reflect the fact EXAFS analysis typically underestimates bond distances (relative to crystallography). The protein samples used for EXAFS analysis were extensively dialyzed and had no extraneous source of sulfur, precluding non-protein based Zn-S ligation. A visual inspection of the FEFF fit presented here shows that the first peak (corresponding to Zn-N/O paths) in the Fourier Transform is a slightly poorer fit relative to the second peak (corresponding to Zn-S paths). In an attempt to improve the fits and to differentiate between 3 Zn-N/O and 4 Zn-N/O fits, split first shell fits were performed. Significant statistical improvement was not observed. **(b)** The R_{sq} values ($\Sigma[(\chi_{obsd} - \chi_{calcd})^2 k^6]/\Sigma[(\chi_{obsd})^2 k^6]$) for four- to six- coordinate first shell fits are presented as a function of increasing number of Zn-S ligands with concomitant decrease in the number of Zn-N/O ligands. (—) four-coordinate, (—) five-coordinate. (—) six-coordinate. The R_{sq} of the four-coordinate fit is significantly worse than that of the five- or six-coordinate fits. Note that although the best R_{sq} value is obtained with 4 Zn –N/O and 2 Zn-S ligands, the five coordinate fits with either 3 Zn-N/O and 2 Zn-S or 2 Zn-N/O and 3 Zn-S ligands also have reasonably low R_{sq} values. For the 2 Zn-N/O and 3 Zn-S fit to be correct, the two Zn sites need to have 2 Zn-S and 4 Zn-S ligands, respectively. Such a structure is ruled out by NMR data, which show that the sites do not have more than two S-based ligands. Since the first shell coordination number error can be up to 20%, it is difficult to differentiate between the 4 Zn-O/N and 3 Zn/O fits with a high level of statistical confidence. Note that both the 4 Zn-N/O and 3 Zn-N/O fits indicate that the high Zn-affinity site is six-coordinate. Six-coordinate Zn sites account for at least 11% of all Zn sites in biology based on NMR and crystallography studies [190]. EXAFS studies with cysteine ligands are typically limited to four- and five-coordinate sites [193]. Studies have been performed on six-coordinate sites, but typically with all light atom ligands [194]. In general, a comparison of total EXAFS intensity can give an insight into the coordination number but the presence of two different first shell ligands (N/O and S) modulates the EXAFS data strongly, making an accurate comparison of EXAFS data between two systems with different coordination numbers difficult [193]. In such a situation, the error in first shell coordination number determination can be greater than 20%. Since the EXAFS data are best fit with between 3 and 4 light atom ligands, the higher error indicates that the second site can be between tetra- and septa-coordinate. Since, a seven coordinate site has no biological precedence, the second site is between tetra- and hexa-coordinate.

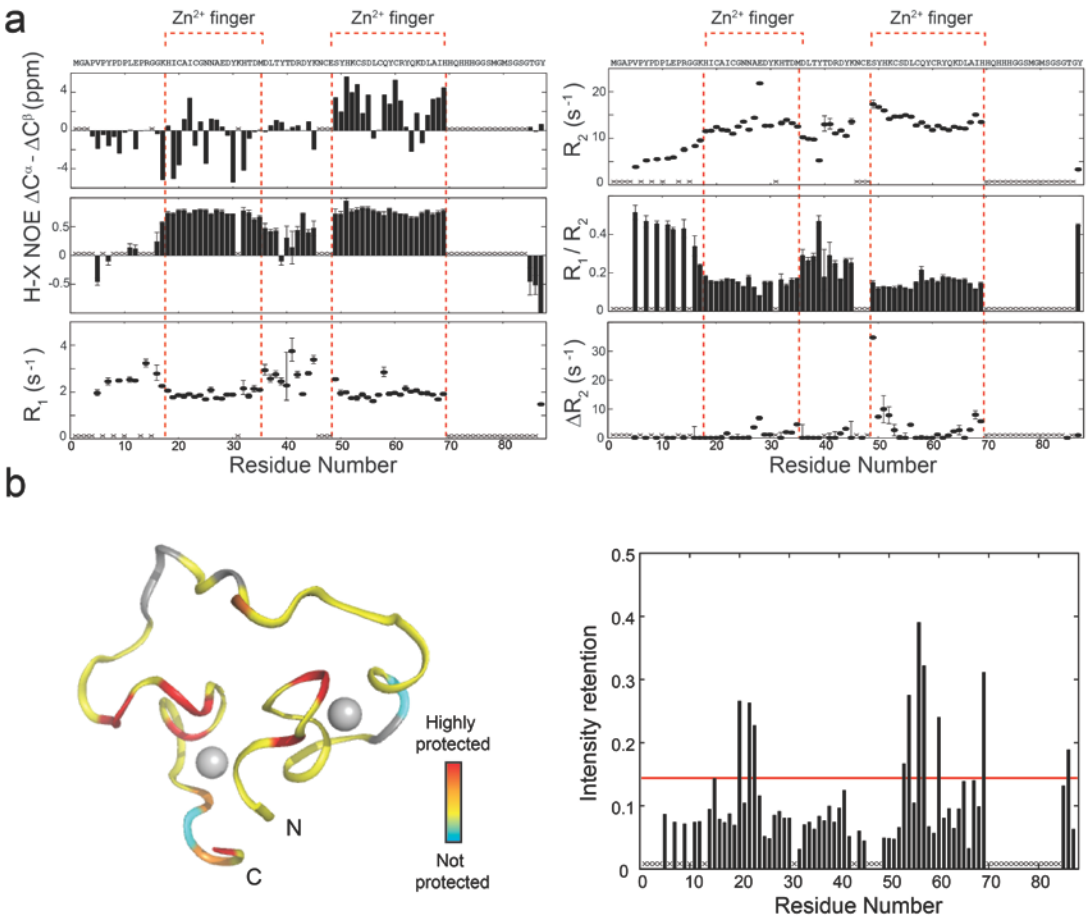

Figure S 3.9-Structure and conformational dynamics probed by NMR experiments.
(a) Chemical shift indexes (ΔCα-ΔCβ), steady-state NOE, longitudinal relaxation rates (R_1), transverse relaxation rates (R_2), and R_1/R_2 ratios as determined by NMR spectroscopy (unassigned residues are marked with an "X"). The errors are estimated by the signal-to-noise (H-X NOE), standard deviations of the fitting (R_1, R_2, and R_1/R_2), or duplicate experiments (ΔR_2). The two zinc fingers are highlighted with dashed red lines. **(b)** The decrease of peak intensities due to H/D exchange was mapped onto one NMR conformer (residues 17-69 are displayed). The intensity of the peaks was normalized to a reference HSQC spectrum of the ligase 10C in 10% D_2O. The sample was lyophilized and dissolved in the same volume of 80% D_2O. After 6 minutes, the HSQC spectrum was acquired and compared with the initial spectrum to monitor solvent exposed amide groups. The solid red line in the diagram represents the average intensity retention plus 2σ.

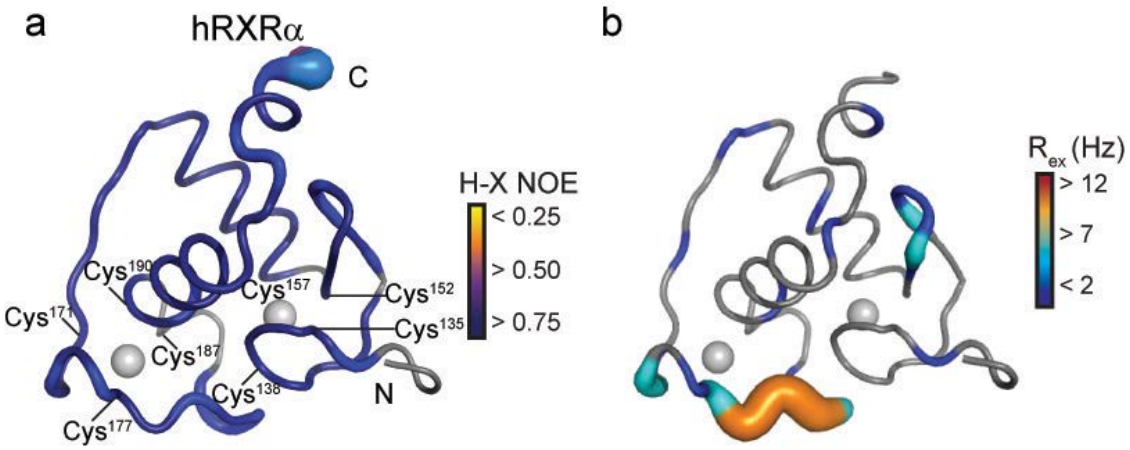

Figure S 3.10-Mapping of the conformational dynamics of the DNA binding domain (hRXRα).
(a) Heteronuclear NOEs (proxy for fast dynamics on a picosecond-nanosecond time scale) on the structure of hRXRα [159]. **(b)** Exchange rates (R_{ex}) obtained from relaxation dispersion measurements as a proxy for slow dynamics (microsecond-millisecond time scale). The color gradient and the thickness of the backbone indicate the intensity of the motions.

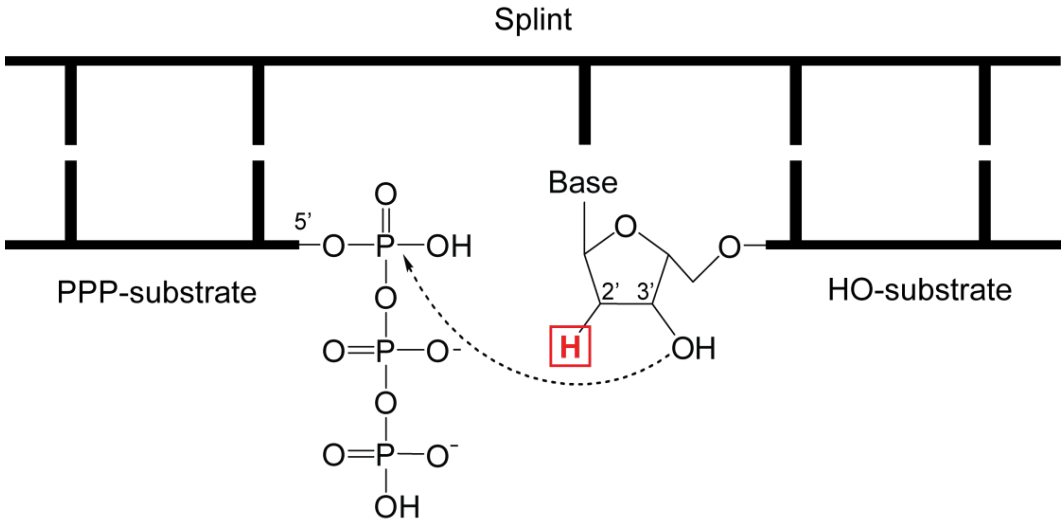

Figure S 3.11-Chemical structure of inactive ligation substrate.
Substitution of the 2'-hydroxyl group of the terminal nucleotide in the HO-substrate with a 2'-deoxy modification (red box) results in inactivation of the ligation reaction. Ligation of active substrates occurs between a 5'-triphosphorylated RNA (PPP-substrate) and the 3'-hydroxyl group of the second RNA (HO-substrate) while both RNAs are base-paired to a complementary oligonucleotide (splint). The dashed arrow symbolizes the proposed bond formation.

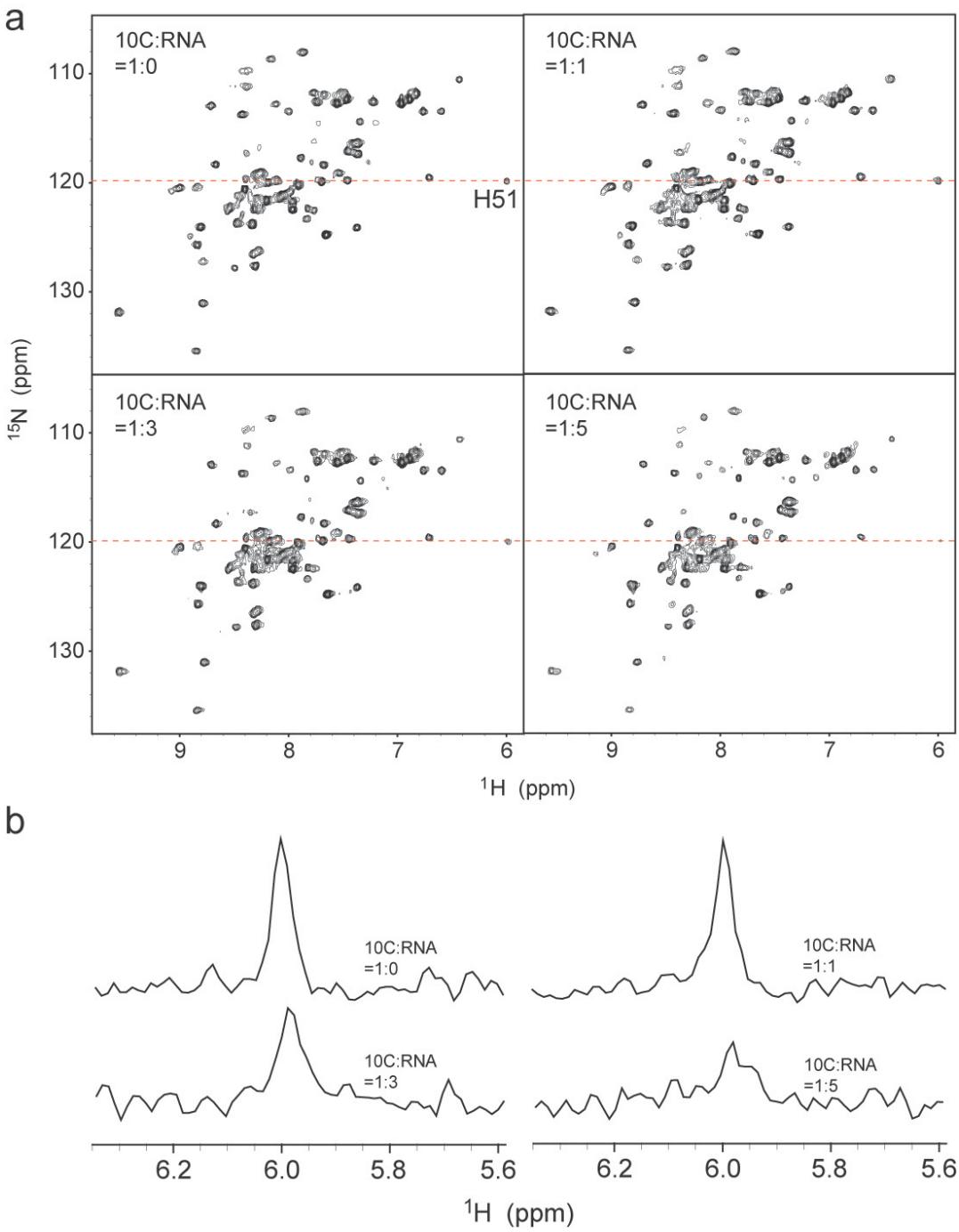

Figure S 3.12-Titration of RNA substrate into ligase 10C monitored by NMR spectroscopy.
(a) The ligase enzyme (300 μM) was titrated with the inactive RNA ligand in 150 mM NaCl, 20 mM HEPES, 10 mM β-mercaptoethanol, and pH 7.5. The HSQC spectra during the titration showed no significant changes in chemical shifts. **(b)** Slices of a selected peak (H51) in HSQC spectra during ligand titration showed significant line-broadening.

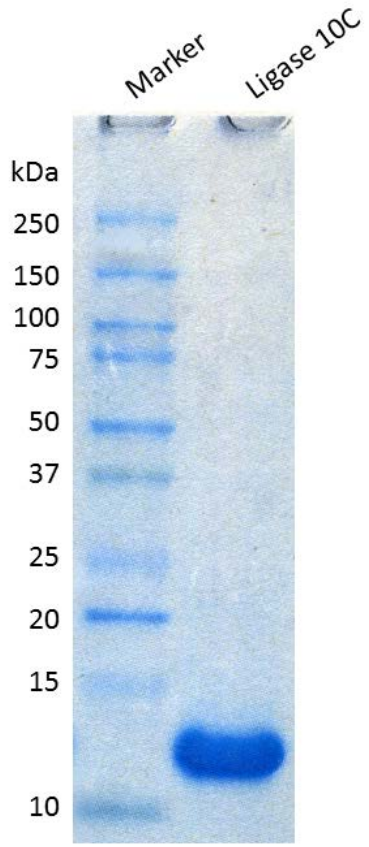

Figure S 3.13-Purity and identity of purified ligase 10C.
SDS-PAGE gel (NuPAGE 4-12% Bis-Tris gel, Invitrogen) Coomassie stained of ligase 10C purified by nickel affinity chromatography and size exclusion chromatography and 10-250 kDa ladder P7703S (New England Biolabs) used as a marker. The identity of ligase 10C was confirmed by MALDI mass spectrometry yielding a characteristic $[M+H]^+$ signal at m/z = 9,648 ± 1.4 (± s.d. from five independent measurements), which is consistent with the expected mass of the ligase 10C without the N-terminal methionine (MW = 9,648.7 Da). The purity of labeled constructs and all mutants matched that of the purified ligase 10C shown here.

Chapter 4 : Comprehensive deletion libraries mediated by *in vitro* transposition

This chapter describes unpublished experiments. Dr. Seelig and I designed the experiments and interpreted the results. I performed all the experiments. Joshua Baller and Lauren Mills from the Minnesota Supercomputing Institute (MSI) analyzed the Next Generation Sequencing (NGS) data, generated the related figures, and contributed to the interpretation of the NGS results.

4.1 Summary

We present a novel method to generate large libraries of >10,000 random deletion mutants of a given gene. This approach is based on *in vitro* transposition and routine molecular biology techniques such as PCR, restriction digestion and ligation. The method is easy to implement and can generate libraries in three to four days. This technique can be applied to the biochemical characterization of proteins and nucleic acids and *in vitro* directed evolution studies of proteins. We used the generated library as input for the directed evolution by mRNA display of the artificial ligase 10C yielding functional variants with deletions of up to 18 of a total of 96 amino acids.

4.2 Introduction

Several studies highlight the importance of deletions in protein evolution. For example, analysis of natural proteins or proteins derived from *in vitro* evolution, showed that deletions of up to 40 amino acids within loop regions can be structurally tolerated [195]. The Indel PDB database [196] provides numerous examples for deletions that are not only found in loops and unstructured regions, but also in alpha helices and beta sheets. Proteins with indels are also highly represented in essential proteins, and are very common in protein networks, where they show a high level of connectivity, suggesting important regulatory roles [197]. Finally, deletions can result in better catalytic and biophysical properties [198-200]. Given the importance of deletions, their impact on

protein properties needs to be extensively studied in order to further our knowledge of evolutionary mechanisms, and accelerate the discovery of improved variants.

Most commonly in directed evolution experiments large libraries of protein variants are created by introducing point mutations [201] or by recombination [202]. On the other hand, usually only few deletion variants are generated on the basis of structural information [199,203,204]. This approach has resulted in proteins with increased stability, however a more thorough and efficient investigation of the effects of deletions would require the generation of large libraries of deletion mutants in a combinatorial fashion. While methodologies to generate such libraries exist, they have undesired drawbacks. They are either sequence specific as they require to design primer pairs for each deletion mutant to be generated [205], or they employ nucleases such as DNaseI or exonuclease III to degrade the target gene to the desired lengths [198,206,207]. The activity of these enzymes is notoriously difficult to control, and reaction conditions, such as time or enzyme concentration, usually require extensive optimization to avoid over-digestion [198,206,207].

In order to simplify the procedures required to build libraries of deletion variants, we developed a method based on *in vitro* transposition mediated by the MuA transposase [208]. This approach overcomes the two drawbacks mentioned above, and generates a library of relatively unbiased and evenly distributed deletions. We used the MuA transposition system because it has been well characterized, and its features are ideal to develop a generally applicable method to create random deletions. First, the target site preference of MuA is low, with insertions occurring randomly along the target sequence [208-210]. Second, standard reaction conditions for MuA have been defined that can be employed for any target DNA, hence no optimization of the transposition reaction is needed. When donor (transposon) and enzyme are mixed in the presence of an excess of target for a short time, single insertions are obtained [208]. The MuA transposition system has been used for several applications such as DNA sequencing [211] and protein engineering [212], highlighting its efficiency and wide applicability.

We validated our method by creating a deletion library of the artificial RNA ligase enzyme 10C. Our goal was to identify active deletion mutants amenable to crystallization. The structure of ligase 10C comprised a well-structured core framed by highly dynamic termini. In addition, the structured core also contains a flexible internal loop region. The flexibility in those three regions likely prevented previous attempts of crystallizing ligase 10C. However, it is conceivable that some of those dynamic regions are less important for catalytic function and dispensable, hence they could be removed to promote crystallization. Yet, rational prediction of tolerable deletions has been notoriously difficult in general, and for ligase 10C in particular as the active site of this enzyme is still unknown. Therefore, we decided to employ a directed evolution approach. We used the deletion library of ligase 10C as input for an mRNA display selection to isolate active shorter variants that will eventually be screened for crystallization.

In order to build the deletion library, we first generated two separate sub-libraries of 5' and 3' terminal gene fragments by transposition and PCR. These sub-libraries were randomly recombined by restriction digestion and ligation to obtain a library comprising both terminal and internal deletions. We validated library coverage and distribution of the deletions by next generation DNA sequencing

4.3 Results

4.3.1 Library construction

We created a library of random internal and terminal deletions of variable length within a given gene. We used the ligase 10C gene as template to test the approach and to subsequently perform an mRNA display selection with the resulting library to identify functional deletion variants. In view of the mRNA display procedure, prior to starting the deletion protocol we appended a FLAG peptide sequence by PCR to allow purification of the fusions during the selection.

The first step of our protocol was a transposition reaction (Figure 4.1) which resulted in random insertions of the transposon in the gene of interest. Subsequently, two separate PCR reactions were performed to generate sub-libraries of 5' and 3' fragments

(where the 5' fragments contain 3' terminal deletions, and the 3' fragments contain 5' terminal deletions). Each PCR contained one primer that binds to the constant region at the respective terminus of the gene of interest, and a second primer that binds to the transposon. As a result, the PCR products consisted of 5' and 3' fragments of our gene with part of the transposon sequence appended (Figure 4.1, step 2). The two PCR products were then processed through a series of digestion and ligation reactions to remove the transposon sequence and randomly connect the two sub-libraries to generate the final deletion library.

We set up transposition reactions according to manufacturer instructions and included a negative control by inhibiting MuA transposase with 10 mM EDTA, and heating the reaction for 10 minutes at 75°C before adding the target gene. When the transposition reaction was amplified by PCR (Figure 4.2, lane TR) two ensembles of random length fragments consistent with the expected range of lengths of DNA were observed (957-1,218 bp for the 5' fragment sub-library; 1,247-1,514 bp for the 3' fragment sub-library). On the contrary, when the inhibited transposition reaction (IT), the transposon DNA alone (T), or the target gene ligase 10C alone were amplified, no smears were observed. These results indicate that the smears were a result of the specific amplification of the transposition product. For the 5' fragment sub-library a small amount of products at about 300 bp length was observed. Furthermore, amplification of the transposon alone (T) resulted in a band running at about 350 bp when primers to amplify the 5' fragment sub-library were used. These bands were likely the result of mis-priming, in fact optimization of the annealing temperature significantly decreased their amount (data not shown).

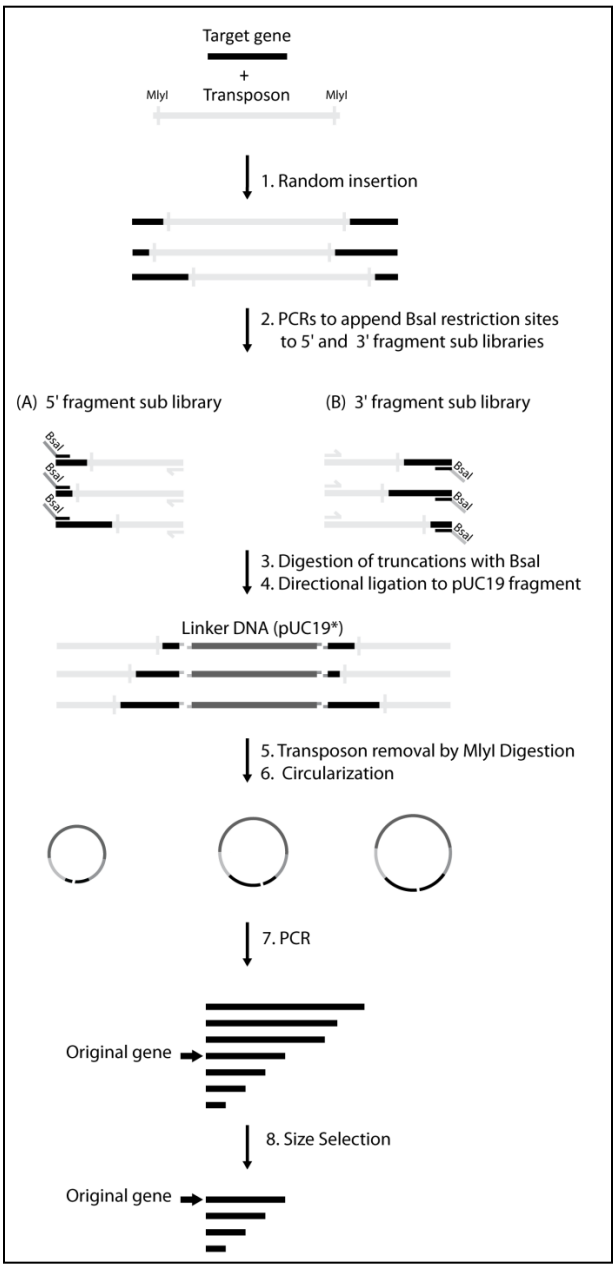

Figure 4.1-Overview of method to generate random deletions.
(1) A transposition reaction was performed to obtain random insertions of the transposon into the target gene and generate transposition products. (2) Subsequently random 5' and 3' terminal truncations of the gene were amplified in two separate PCR reactions, with primer couples in which one primer bound to the 5' or 3' constant regions of the target gene, and the other to a region on the transposon. (3) The digestion with BsaI created unique overhangs in each library. (4) Libraries were ligated to a linker DNA (pUC19 fragment *) which contained no MlyI sites. (5) The product of ligation was treated with MlyI to remove the transposon sequence. (6) Intramolecular ligation joined the 5' and 3' terminal fragments of the gene. (7) This circular library was linearized by PCR with primers complementary to the termini of the original gene. (8) The final library of deletion variants of the desired size range was isolated by gel electrophoresis.

Following PCR amplification, the two products were digested with BsaI, and ligated to a long linear linker DNA. BsaI is a type IIS restriction enzyme which cuts outside of its recognition sequence. This allowed us to design unique cut sites for each sub-library to prevent ligation of two fragments of the same kind to the linker. As linker we used a segment amplified out of the plasmid pUC19 (indicated as pUC19* in figure 4.1). This segment does not contain cut sites for the restriction endonuclease MlyI. This enzyme was used in the following step to digest the ligation product and remove the transposon sequences from the sub-libraries.

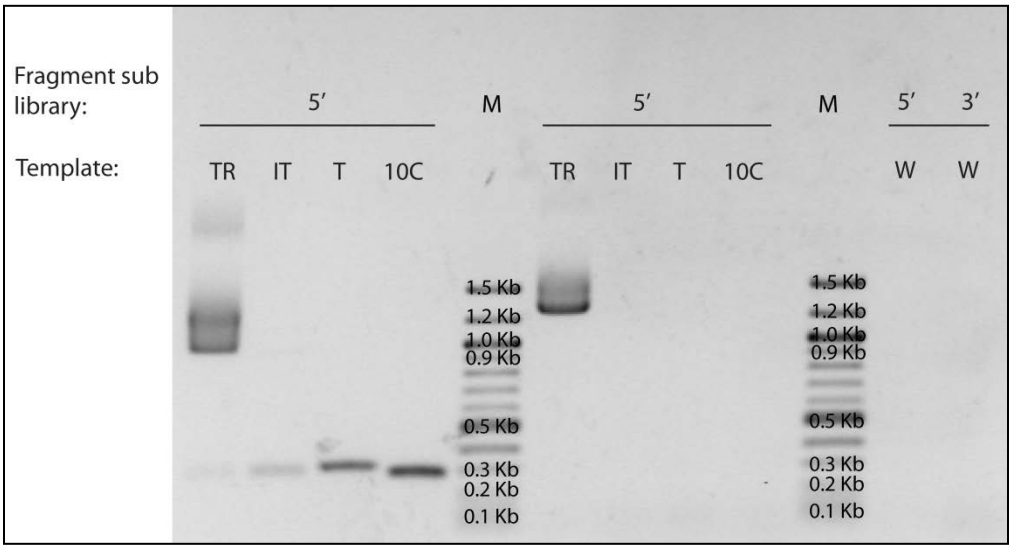

Figure 4.2-Agarose gel (1%) of PCRs with primer pairs to amplify 5' and 3' fragment sub-libraries in presence of different templates.
TR = transposition reaction; IT = inhibited transposition reaction; T = transposon; 10C = target gene coding for ligase 10C; M = Marker (100bp ladder); W = water (negative control).

After ligation (Figure S4.1) we expected a smear in the range 3,531-4,059 bp and we obtained a band around 4,000 bp. After removing the transposon by digestion with MlyI (Figure S4.2) we obtained a smear in the expected range (1,434 bp-1,962 bp) (Figure S4.2), with an overrepresentation of shorter sequences. The two transposon sequences of 904 bp and 1193 bp were successfully removed as indicated by the presence of two bands

at the expected length. The digestion product (1400-2000 bp) was purified, circularized, and used as template to amplify the final library. We gel purified the DNA in the range of interest from 53 bp to 288 bp, which represent the full length gene and deletion of the full sequence (except the regions for primer binding) respectively. As expected, we obtained a smear from about 300 bp to less than 100 bp (Figure 4.3, squared bracket). For quality control, the final deletion library was sequenced by next generation sequencing.

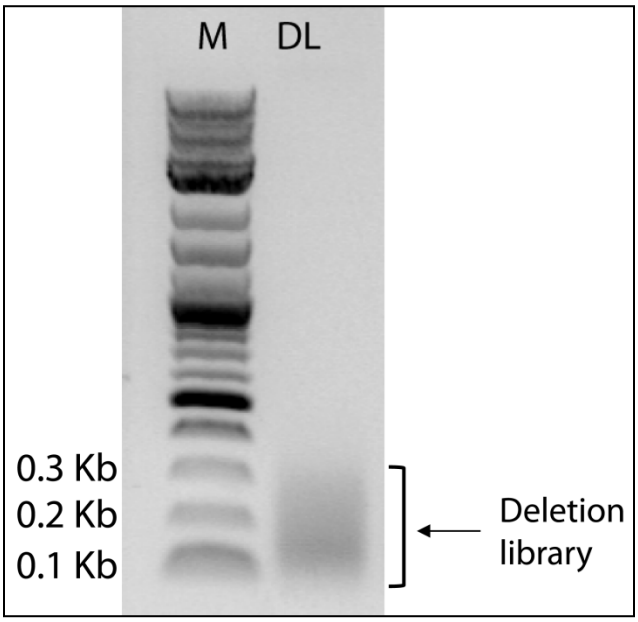

Figure 4.3-Agarose gel (1%) of deletion library after gel extraction and desalting.
M= Marker (2 log ladder); DL=Deletion library.

4.3.2 Next Generation Sequencing analysis

We conducted next generation sequencing analysis of our final deletion library to assess the efficiency of our method to generate a large number of deletion mutants randomly distributed throughout the starting gene. Sequencing yielded 10,491 unique sequences, which represented 36% coverage of the maximum theoretical complexity (27,612 sequences; calculated according to eq. 2 in section 4.6.7). In order to verify if there were any biases in the deletion distribution, the 10,491 sequences were aligned to

count the number of times each nucleotide position was deleted. The observed distribution (Figure S4.5, red line) was compared to the theoretical distribution, which was calculated for 100% coverage of the theoretical complexity (Figure S4.5, blue line). This comparison showed there is a slight bias for deletions towards the 3' end of the gene (Figure S4.5). In addition, we observed that sequences with deletions less than 20 bp in length are underrepresented by about two orders of magnitude (Figure S4.4). These two biases could be explained by the preferential amplification of shorter sequences during the PCR amplification of the 5' and 3' fragment sub-libraries (Figure 4.1, step 2), and by a potential presence of preferred transposition sites at the 3' of the target DNA. This is in agreement with previous findings which showed that although the process of transposition is random, some sites can be preferred [208-210]. Despite these biases, the sequencing data demonstrated that the method generated random deletions ranging widely in their length and in their position along the target sequence.

4.3.3 Set up of the mRNA display selection

Our goal was to isolate shorter variants of the ligase 10C, ideally catalyzing the reaction faster than the wild type and amenable to crystallization. Deletion of residues at the termini is a strategy that had been successfully employed to favor protein crystallization [203]. However, given the highly dynamic structure of the protein it would be difficult to rationally identify disordered regions that upon deletion could favor crystallization. We hence decided to take a directed evolution approach as it can test many variants at the same time. The deletion library of 10C was used as input for an mRNA display selection of shorter active variants that could then be tested for crystallization. To further increase the diversity at the beginning of the *in vitro* selection, we also prepared two additional libraries: a deletion library of ligase 10C previously mutated at random by error-prone PCR, and a double deletion library by reapplying the procedure described in this chapter to the deletion library generated previously. In order to isolate both more and less active variants we performed two selections in parallel with different selective pressures: 5 minutes and 60 minutes reaction time.

4.3.4 In vitro *selection for ligase variants*

The mRNA display procedure was performed for 6 cycles of selection and amplification as published previously [51,52], and also reported in sections 4.6.8 and 1.5.2 of this thesis. The percentage of cDNA bound to the streptavidin resin was calculated with equation 3 in section 4.6.8 to monitor selection progress.

For the '5 min selection' the percentage of bound cDNA increased from 0.03% after round 1 to 3.5% in round 5 and then decreased to 1.4% in the 6^{th} round (Figure 4.4, black bars). For the '60 min selection' we observed an increase in the percentage of cDNA bound from 0.18% to 5% in round 3. In the subsequent rounds 4-6 the percentage varied between 2.3% and 4% (round 6) (Figure 4.4, grey bars).

The PCR amplification of the eluted cDNA after round 3 of the '60 min selection' yielded a clearly discernible DNA band in the agarose gel of the same length as the original full-length gene (332 bp). Apparently, the original gene was preferentially enriched, while the smear representing shorter deletion variants was disappearing (data not shown). In order to specifically favor and therefore enrich deletion variants over the full-length gene, after rounds 3-5 of the '60 min selection', we isolated DNA shorter than the original gene (approximately 300 bp to 100 bp) by gel electrophoreses of the PCR product and used it as input for the following round. The removal of the band at 332 bp explained the drop in signal observed in rounds 4 to round 6. Despite the repeated removal by gel extraction, the band at 332 bp reappeared after every round as it was enriched during the selection step.

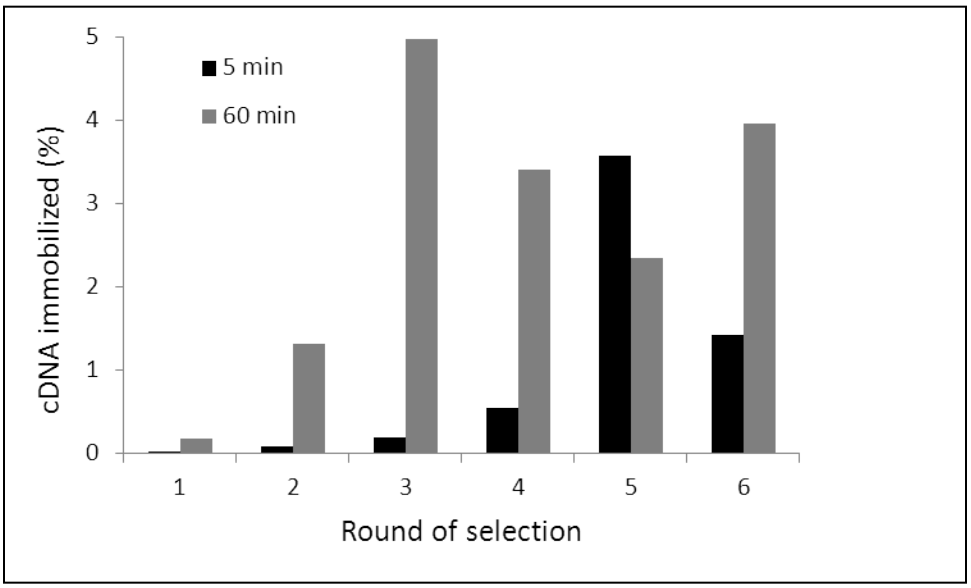

Figure 4.4-Progress of *in vitro* selections for ligases using a 5 minute (black bars) and 60 minute (grey bars) incubation time during the selection step.

The signal increase from round 5 to round 6 of the '60 min selection' can instead be explained by the enrichment of active deletion variants. An SDS PAGE gel of aliquots taken at each step of round 6 of the '60 min selection' (Figure 4.5B), showed the enrichment of shorter active variants not observed in previous rounds (lower band in the brackets in figure 4.5B). Comparison of the ligation reaction (right-most lane 'Rxn') to purified material from the preceding step (lane 'Flag II'), showed the appearance of two bands at a higher molecular weight, which represent ligation product caused by the covalent link of the 3'-OH substrate to the 5'-PPP substrate, indicating that both bands represent active clones. In contrast, the '5 min selection' was mostly enriching variants of the same length as ligase 10C (Figure 4.5A**)**, as indicated by the presence of only one band shifting up in the ligation reaction (lane 'Rxn').

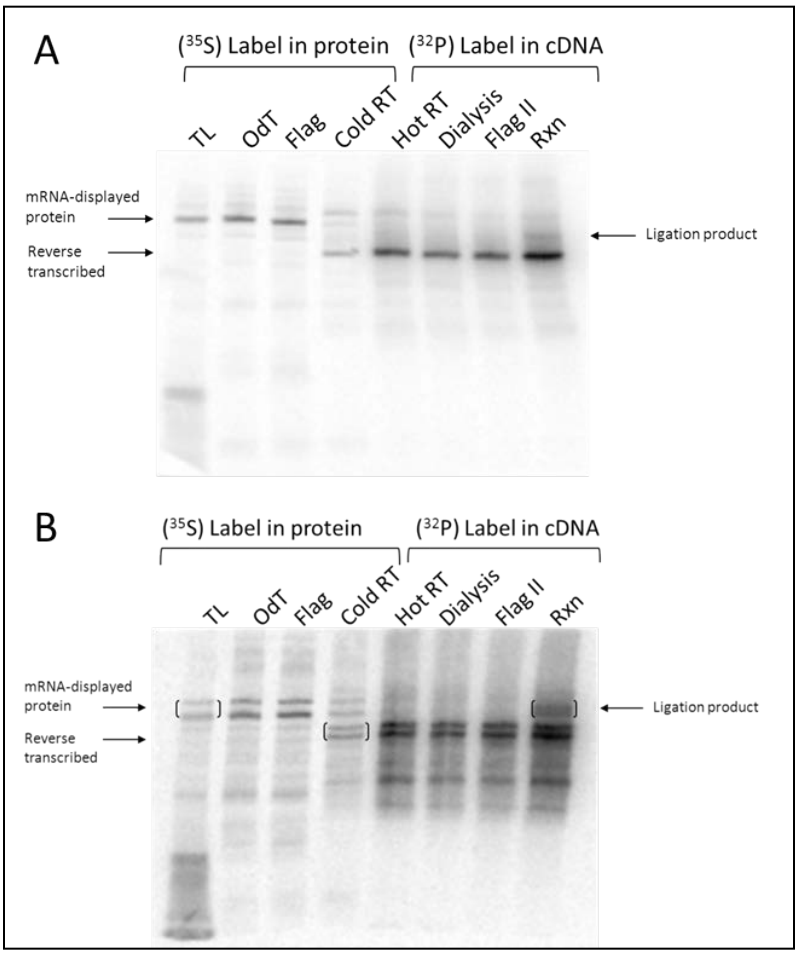

Figure 4.5-Gel electrophoresis of mRNA-displayed proteins after individual steps of round 6 of the selection (SDS PAGE).
(A) Gel for '5 min selection'. (B) Gel for '60 min selection'. TL: translation; OdT: eluate after purification by oligo-(dT) cellulose chromatography; Flag: eluate from Flag affinity purification; Cold RT: reverse transcription with no ^{32}P-α-dATP added (only ^{35}S-Met label on protein) to verify efficiency of reverse transcription; Hot RT: reverse transcription with ^{32}P-α-dATP (label on cDNA); Dialysis: sample recovered after dialysis in Flag buffer; Flag II: eluate from second Flag affinity purification; Rxn: ligation reaction with 3'-OH biotinylated substrate.

cDNA from round 6 of both selections was amplified by PCR (Figure 4.6). Consistent with the results observed by SDS PAGE gel, in round 6 of the '60 min selection' two major bands were observed at 332 bp and at about 275 bp, respectively. The '5 min selection' yielded a main band at 332 bp.

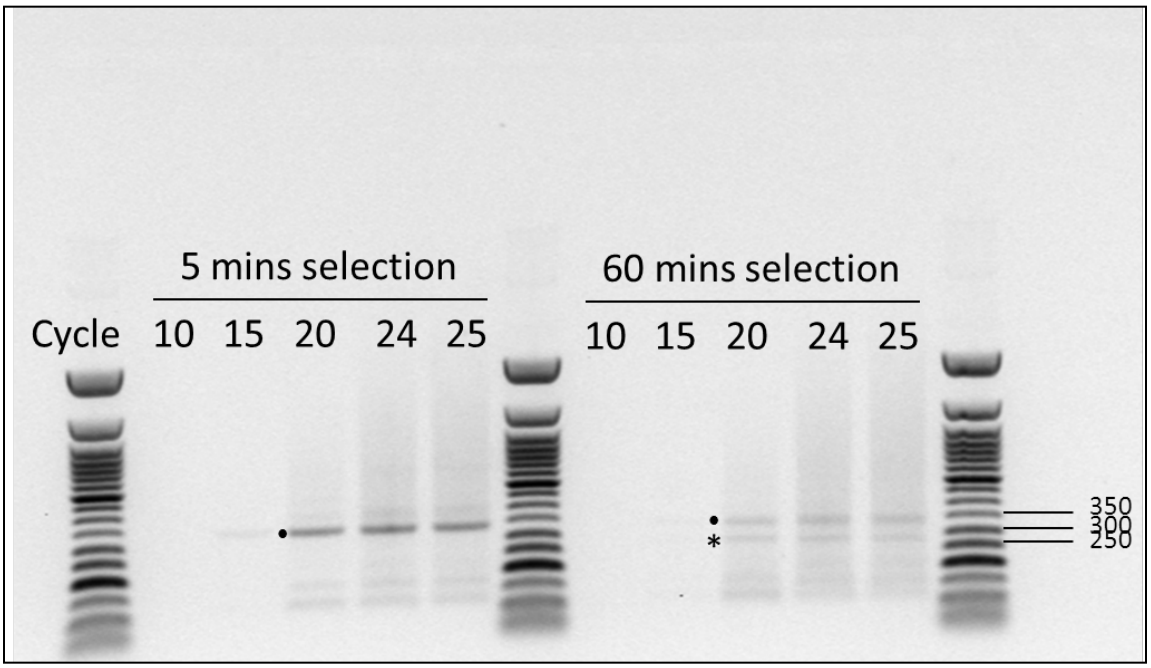

Figure 4.6-Agarose gel electrophoresis of PCR products from cDNA eluted in round 6 for both 5 min and 60 min selection.
For the '5 min selection', the presence of a major band running at the same length of ligase 10C (marked with a dot) is consistent with a similar observation in the SDS PAGE gel in figure 4.5A. In the '60 min selection', an additional major band (marked by an asterisk) at about 275 bp is present, consistent with the two bands in figure 4.5B.

4.3.5 DNA sequencing analysis of enriched clones from round 6 of selection

For both selections, cDNA after round 6 was cloned into plasmids and sequenced. We obtained 68 sequences for the '5 min selection' and 40 sequences for the '60 min selection' which were aligned to the original 10C sequence. After manual inspection of the alignments we removed any short peptides of only 20-47 amino acids in length, which were composed only of the first 13 amino acids from the N-terminus (MDYDDDDKGAPV) and of 8 amino acids at the C-terminus (MSGSGTGYStop), and of duplications of these sequences. Since these peptides were missing crucial regions such as the putative zinc and substrate binding site [127], we reasoned that these peptides were selection artifacts and highly unlikely to possess activity. The remaining 54 sequences from the '5 min selection' were either wild type ligase 10C (63%) or point-mutants thereof (37%), but no deletions variants were detected. In addition, a single

sequence with a one amino acid insertion causing a frame shift near the C-terminus was observed. Of the remaining 18 sequences from the '60 min selection', 15 sequences (83%) were either 10C wild type (7 sequences) or had point mutations (8 sequences). The other 3 sequences (17%) were deletion variants missing 13 and 18 amino acids near the N terminus. In particular, the 18 amino acids deletion corresponded to a deletion of 54 bp in the original 332 bp DNA, resulting in a band of 278 bp. This size was consistent with the lower molecular weight band observed after PCR amplification of the cDNA resulting from the 6th round of selection (Figure 4.6, band marked by an asterisk). Interestingly, sequences with mutations in the putative zinc binding sites (residues 17-35 and residues 49-59) were present in multiple copies. In particular M35V appears four times and M35T twice. Furthermore, D43Y appears four times. The double mutant M35V/D43Y is present in duplicate. In addition, one of the clones from the 5 minutes selection contains 5 point mutations, mainly distributed in the flexible termini of the protein structure.

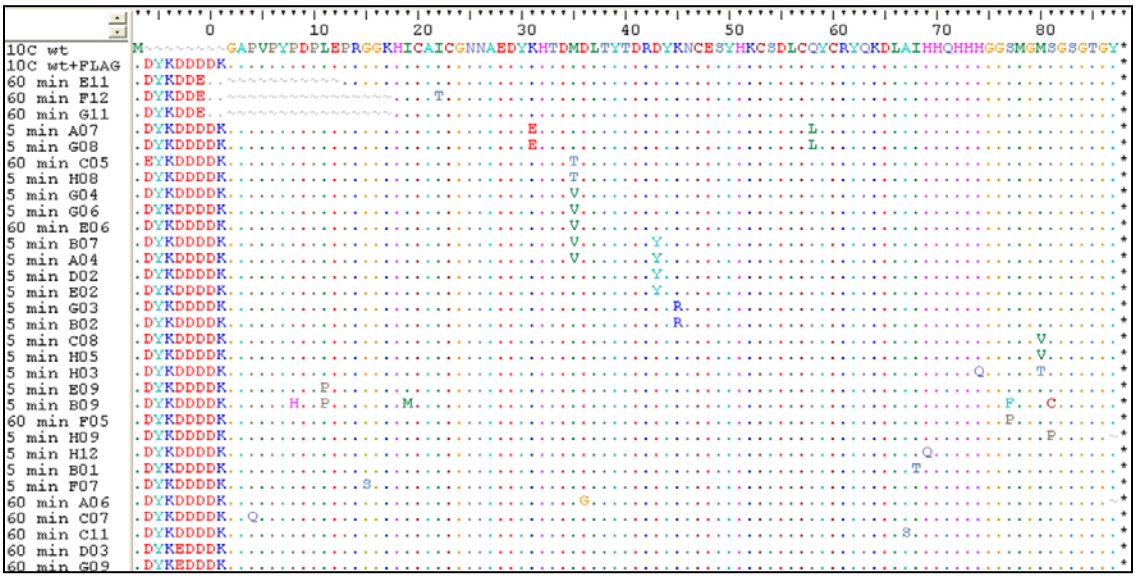

Figure 4.7-Alignment of selected variants of ligase 10C from cDNA after round 6 of the 5 min and 60 min selections.
Variants containing a 13 and 18 amino acids deletion (E11, F12, G11) were isolated during the '60 min selection'. In addition, variants with point mutations were isolated from both selections. Top sequence is reference wild type ligase 10C (10C wt). The second sequence represents wild type ligase 10C modified by addition of the FLAG peptide for purification purposes during each cycle of selection and amplification. Numbering of residues is set to match numbering used to discuss ligase 10C structure in chapter 3.

4.4 Discussion

Analysis of sequences and structures of naturally and *in vitro* evolved proteins reported that for some select proteins deletions of up to 40 amino acids can be accepted without significantly altering tertiary structure [195]. Furthermore, laboratory experiments showed that deletions can in some cases result in increased thermostability of proteins [199], higher transduction efficiency of viral particles utilized for gene therapy [198], and higher activity in ribozyme variants obtained by *in vitro* selection [200]. These reports demonstrated that deletions are not only tolerated, but can potentially result in improved properties, highlighting the need to consider deletions more frequently as a valuable source of genetic variability to be incorporated in directed evolution experiments. We devised a protocol capable of generating a large number of deletion mutants in 3-4 days. This method is ideal to routinely generate deletion libraries for directed evolution experiments because little optimization was required to implement it.

We applied our procedure to a gene coding for an artificial RNA ligase [127,213], and used the resulting library of deletion variants for mRNA display selections with different selection pressures: 5 minutes and 60 minutes reaction time. In the ‘5 min selection’ we observed a decrease in percentage of bound cDNA in round 6. This result was unexpected and has to be confirmed by repeating round 6.

Sequencing results of DNA from round 6 of the ‘60 min selection’ revealed the presence of both 13 and 18 amino acids N-terminal deletions (Figure 4.7). Despite the fact that we used also a double deletion library, only single deletions were found. The results obtained here suggest that regions other than the N-terminus cannot be deleted without compromising enzyme activity. However, these results might also be due to incomplete library coverage, as evaluated by analysis of Next Generation Sequencing data. Furthermore, the presence of two out of the three flexible regions suggests that crystallization might still be difficult. Variants with point mutations comprised 37% and 44% of the sequences analyzed in the ‘5 min’ and ’60 min’ selection respectively. The most represented mutations were M35V and D43Y. While M35V is present in sequences

from both selections, D43Y is present only in sequences from the ‘5 min selection’, suggesting an important role in optimization of catalytic activity. Screening and characterizing the sequences isolated in these selections will assess the impact of these deletions and point mutations on ligase 10C activity. Furthermore, the deletion mutants will be screened for crystallization.

Our data showed that the procedure presented here can generate libraries suitable for selections of functional deletion variants. However, sequencing results showed that the library still contained wild type enzyme that can potentially interfere with the enrichment of deletion variants. Indeed, in round 3 of the 60 minutes selection we noticed that a band of similar length as the wild type enzyme (332 bp) had been enriched while the smear containing shorter variants was disappearing. Nonetheless, gel extraction of the shorter variants from round 3 to round 6 in the 60 minutes selection allowed us to isolate active deletion variants despite the dominating presence of the wild type protein in the library.

Deletion variants of a given gene have not only been used in directed evolution experiments to improve properties, but deletion analysis has also been invaluable for the biochemical characterization of proteins and ribozymes. For example deletion variants were generated to investigate the core sequence required to perform a certain function [105,200], or to remove flexible regions when crystallization proved difficult [203,204]. Usually these approaches require detailed structural knowledge in order to determine which segments could potentially be removed; however, when structures are not available, or in molecules with a high degree of disorder like ligase 10C [127], it is difficult to predict which regions can be deleted. In these cases, it is more efficient to generate a library of random deletions and identify functional constructs by high throughput assays.

Furthermore, deletion libraries could be used to map interactions between binding partners: one molecule would be subjected to the deletion procedure to generate a large library, and the other one would be used as a bait to recover shorter variants that retain binding capacity. This approach, in combination with other techniques could help in

determining interaction surfaces and structures. For example, recently [214] such a multidisciplinary approach was employed to determine the structure of the yeast protein complex Nup84, at a resolution of 1.5 nm, by combining biochemical data obtained from domain deletion mapping experiments, electron microscopy data and computer modeling. In this study deletion constructs were individually generated one by one, based on available data on the domain boundaries of the complex components [214]. Generation of a random deletion library would be a better approach if these data were not available.

Our protocol can also be employed for the creation of libraries of non-homologous recombinant sequences. In this case, multiple transposition reactions should be performed with different target DNA sequences encoding for the parental genes. Alternatively, targets could be mixed in the same transposition reaction, as long as the fragment sub-libraries are amplified with very selective primers (step 2 in Figure 4.1).

Our method showed some biases: a higher number of deletions at the 3' terminus of the target gene, and underrepresentation of longer sequences. These biases might be due to preferential insertion sites at the 3' of the target gene and preferential amplification of shorter sequences. Previous work by others, observed that during library preparation for next generation sequencing, PCR amplification altered the length distribution of the original sample depending on the PCR system used [13]. These biases could be addressed either through emulsion PCR [215,216], or with different polymerase-buffer systems [13]. Furthermore, we cannot exclude the possibility of undersampling during the transposition reaction. Reaction conditions similar to ours in a previous directed evolution experiment [212], where a 2.6 Kb plasmid was used as a target, yielded 2,000 insertion events. Since the insertion rate decreases with the size of the target gene (personal communication from the Finnzymes company), it is possible that we had much fewer than 2,000 insertions for our short 288 bp target. This could provide an additional explanation for the less than complete coverage observed. Using a more efficient transposase, such as Hyper MuA transposase, or conducting larger scale reactions, could allow for a better coverage of the library complexity.

In order to verify if this strategy has general applicability, we performed a transposition reaction and PCR amplification of the sub-libraries for two other genes, one coding for beta-lactamase and the other for glycerophosphoryl diester phosphodiesterase (GDPD) (Figure S4.3). In both cases, by using the same transposition and PCR conditions applied to ligase 10C, we were able to successfully amplify 3' and 5' fragment sub-libraries by using primers appropriate for the termini of the two target genes. In particular, during amplification of GDPD truncation libraries non-specific products were formed (around 1,000 bp for the 3' truncation library and around 100 bp for the 5' truncation library). The amount of these non-specific products decreased by optimizing annealing temperature, and in both cases the main product of the reaction was the desired smear, demonstrating that the procedure can be applicable to different genes.

4.5 Conclusions

We provided an efficient tool for generating libraries of at least 10,000 random deletion mutants. We have isolated active N-terminal deletion variants of ligase 10C and have shown that the method presented here generates libraries suitable for directed evolution experiments. Given the ease of execution and general applicability, an engineering platform of this type will be a very useful tool for both laboratory evolution and characterization of proteins and functional nucleic acids, further advancing biochemical knowledge, and creating variants with improved properties.

4.6 Materials and methods

Template generation system Kit II from Finnzymes was used to perform the transposition reactions. TOPO TA cloning Kit and pUC19 vector (2.68 Kb) were purchased from Invitrogen. Phusion High Fidelity DNA Polymerase from New England Biolabs (NEB) was used in all PCR reactions with the buffer provided by the manufacturer. dNTPs stocks and restriction enzymes were purchased from NEB; primers were purchased from Integrated DNA Technologies (IDT), and EDTA (0.5 M, pH=8)

was purchased from AccuGENE. All purification kits (QIAGEN) were used according to manufacturer instructions. DNA concentrations were measured by reading absorbance at 260 nm on a Nano Drop spectrophotometer 2000c (Thermo Scientific). Power Pac 300 (Biorad) was used as power supply for gel electrophoresis, which was conducted in 0.5X Tris-Borate EDTA buffer (TBE); 10X TBE solution from national diagnostics contained Tris Borate (0.89M) pH=8.3 and 20 mM EDTA disodium salt. DNA ladders were purchased from NEB.

4.6.1 Amplification of pUC19 fragment, ligase 10C, transposon, beta-lactamase, and glycerophosphoryl diester phosphodiesterase (GDPD).

All PCR were performed in presence of 200 µM of each deoxynucleotide, 1X reaction buffer, 0.5 µM of each primer, 0.01 U/ µL of Phusion high fidelity DNA polymerase. Reaction conditions were as follows: 3 minutes at 94°C followed by a variable number of cycles with 30 seconds at 94°C, 45 seconds at 55°C, and 1-3 minutes at 72°C depending on the product length. pUC19 fragment (pUC19* from now on) was amplified from 6 fM pUC19 vector for 26 cycles in a 1 mL PCR reaction, with an extension time of 3 minutes. The DNA obtained was gel purified and desalted, and used as template for an 8 cycles PCR in the same conditions as above, but with template concentration of 1 nM. The target gene 10C was amplified from a vector obtained through TOPO TA cloning, with an extension time of 1 minute for 20 cycles.

The transposon was amplified from the CamR[3] transposon provided with the template generation system Kit II, with 2 minutes extension time for 24 cycles. Beta-lactamase was amplified from a pBAD plasmid for 20 cycles with 1 minute extension time. GDPD was amplified from a pCA24NMAF2 plasmid [217] for 30 cycles, with 1 minute extension time. Primers are listed in Table S1, along with the size of the PCR products.

Each PCR product was gel purified and subsequently desalted using gel extraction and PCR clean-up Kits respectively (QIAGEN) according to manufacturer instruction. The transposon was subsequently digested with BglII, to create pre-cut ends for proper

transposition efficiency, and desalted prior to the transposition reaction. pUC19* was cut with BsaI, and desalted as above.

4.6.2 TOPO cloning

TOPO Cloning was performed according to manufacturer instructions, except that for Blue and white screening 30 µL of 30mg/mL 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside [X-gal] were used. Cells were plated on LB-Agar containing 36µg/mL of Kanamycin.

4.6.3 Agarose gels and electrophoresis

All gels shown in the paper were 1% agarose. Samples were mixed with Ficoll to a final concentration of 2.5% prior to loading. Electrophoresis was conducted at 120V, for a duration variable between 30 and 50 minutes. Gels were stained with ethidium bromide (final concentration 0.5 ng/µL), and visualized under UV light at 254 nm. Staining was performed either pre-run (ethidium bromide was included in the gel prior to pouring), or post-run (gel was soaked in a water solution of ethidium bromide for 20 minutes, and subsequently destained in water for 20 minutes).

4.6.4 Transposition reactions

Transposition reactions were performed using the template generation system Kit II from Finnzymes, according to the manufacturer's instructions. Reactions (20 µL) were assembled using 12 ng of 10C DNA (64 fmoles), 20 ng of transposon (24 fmoles), reaction buffer to a final concentration of 1X, 220 ng of MuA transposase. The reaction was incubated at 30°C for 1 hour, stopped by deactivating the enzyme at 75°C for 10 minutes, and either frozen or directly used as template for PCR. In the case where reaction was inhibited by addition of EDTA (final concentration 10 mM), it was heated to 75°C for 10 minutes prior to addition of the target sequence, and then incubated at 30°C for 1 hour.

4.6.5 Library construction

Two separate PCR reactions were carried out to generate 5' and 3' fragment sub-libraries of 10C. Amplification was performed for 30 cycles with 200 µM of each deoxynucleotide, 1X reaction buffer, 0.5 µM of each primer, 0.02 U/µL Phusion polymerase, 0.5 µL of transposition reaction/50 µL of PCR reaction. In order to create the 5'-fragment sub-library the following primers were used:

Fwd:GCGTACTTAGGCGATTAGCTGAGACCATGGACTACAAAGACGACGACGATAAG

which binds the 5' end of the target gene and appends a BsaI site (underlined)

Rev: CGACATGGAAGCCATCACAAACGGCATGATGAACCTGAA

which amplifies 904 base pairs of the transposon (position 904 to position 1).

In order to create the 3'-fragment sub-library the following primers were used:

Fwd: ACGGAAGATCACTTCGCAGAATAAATAAATCCTGGTGTC

which amplifies 1193 base pairs from the transposon (position 109 to position 1,302)

Rev:GCCAGTATAGATTGCAGCTAGGCGGTTGAGACCTTAATAGCCGGTGCCAGATCC which binds the 3' end of the target gene and appends a BsaI site (underlined).

The two PCR products were gel purified and desalted as above. The purified products were both digested using BsaI to generate sticky ends compatible with the ends on the pUC19 fragment. DNA concentration was 100 ng/µL for the 3' fragment sub-library, and 83 ng/µL for the 5' fragment sub-library. Each reaction contained 0.375 U/µL of restriction enzyme, 1X cut smart buffer (50 mM potassium acetate, 20 mM tris-acetate, 10 mM magnesium acetate, 100 µg/ml BSA), and was incubated at 37°C for 12 hours, stopped by heat inactivating the enzyme at 65°C for 20 minutes, and desalted as above. Each purified product was simultaneously mixed with pUC19*, previously cut with BsaI, in a ratio higher than 1:1 (at least 4.6 pmoles of each library and 4.6 pmoles of pUC19*),

with the 5' fragment sub-library at a final concentration of 67 ng/µL, the 3' fragment sub-library at a final concentration of 84 ng/µL, pUC19* at a final concentration of 74 µg/µL, 40U/ µL T4 DNA ligase, and 1X ligation buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT). The reaction was incubated at 23°C for 15 minutes. The product was gel purified and desalted as above, and digested with MlyI to remove the transposon sequences for 11 hours at 37°C in a reaction containing 15 ng/µL of ligated product, 0.3 U/µL of MlyI and 1X cut smart buffer. The desired product was gel purified and desalted. The DNA was circularized by intramolecular ligation in a 50 µL reaction containing: 100 ng DNA (2ng/µL), 40U/µL of T4 DNA ligase (Stock concentration=2,000 U/µL), and 1X cut smart buffer supplemented with 50 µM ATP. The reaction was incubated at 16°C O/N, and subsequently used as a template to amplify the final deletion library for 30 cycles with Phusion polymerase using conditions described above.

The library was amplified with the same primers used for the ligase 10C gene amplification, and the following reaction conditions were applied: 3 minutes at 94°C followed by 30 cycles with 30 seconds at 94°C, 15 seconds at 55°C, and 1 minute at 72°C. The product in the range of interest (53-288 bp) was gel purified, desalted by ethanol precipitation and sequenced.

4.6.6 Transposition and PCR amplification of fragment sub-libraries for beta-lactamase and GDPD

Transposition for beta-lactamase and GDPD was carried out with the same conditions applied to ligase 10C, with the amount of target gene adjusted accordingly (32 ng for GDPD and 35 ng for beta-lactamase). PCR of the sub-libraries was performed as described for ligase 10C, with appropriate primers for the gene termini (listed in table S1).

4.6.7 Analysis of Next Generation Sequencing data

Trimmomatic [218] was used to remove Illumina specific adapter contamination from paired 150bp long reads produced by a MiSeq. Reads that were longer than 36bp

and maintained their paired status (both R1 and R2 were still viable) were retained for downstream analysis. Paired reads were merged using usearch –fastq_mergepairs [219]. Reads to be merged were truncated at the first base with a quality score below 3 and merged reads were allowed to have up to 5 mismatches. Once merged, the quality of the final sequences was evaluated and sequences with an expected error rate greater than 0.5, a quality reflecting more than 1 miscalled base in 2 sequences, were removed.

Queries were aligned to the original full-length sequence using SSEARCH36 from the FASTA36 [220,221] package using a +1/-3 match/mismatch scoring matrix and -1000/-1000 as gap open and gap extend penalties. SSEARCH36 using the options above created high identity local alignments between the queries and the original sequences. For each query library pair, the extremely high gap penalties should result in two statistically significant alignments on either side of the deletion, by combining the two alignments the position and size of the deletion in each sequence can be determined. Combined alignments that did not include the last 10 positions of the 5' constant region and the first 10 positions of the 3' constant regions, combined alignments that had insertions in the query, combined alignments that did not incorporate the entire query sequence and query/library pairs that did not result in exactly 2 alignments to be combined were removed. The unique deletions that resulted from the alignments that met the criteria were then identified and used to count the number of times each position was deleted. The maximum theoretical number of sequences that included a deletion at a specific position was calculated using the following equation:

$$p(N-p+1) \quad \text{(eq.1)}$$

p= position deleted (bp indexed [1,N])

N = Maximum size of deleted region (in bp)

The maximum theoretical complexity was calculated using the following equation:

$$(N*N/2) \quad \text{(eq.2)}$$

N= Maximum size of deleted region

4.6.8 mRNA display selection for ligase variants

mRNA display was performed as previously done [51,52], except that wherever triton was present, its concentration was increased from 0.01% to 0.25%., and AM017 (pTTTTTTTTTTTTTTTTTTCCCAGATCC) instead of oligo BS50 was used to produce the reverse transcription primer. Briefly, the input DNA for each round was used as template for T7 transcription, the resulting RNA was purified by lithium chloride precipitation and cross-linked to the XL-PSO oligonucleotide bearing puromycin and containing an A_{18} tail [51,72]. The cross-linked RNA was ethanol precipitated and used as template for *in vitro* translation in reticulocyte rabbit lysate (Promega) at 30°C for 30 minutes. Proteins were labeled with ^{35}S-methionine for radioactive detection. Efficiency (E) of each step in a round of selection was calculated as follows:

E= (Radioactivity of fraction/total radioactivity in the step)*100 (eq. 3)

For the first round of each selection, a 1 mL translation was performed, the volume was then decreased to 0.5 mL for the following rounds. The translation reaction was purified by oligo-dT (OdT) cellulose chromatography to remove uncrosslinked RNA and unfused proteins. The eluate from the OdT purification was further purified by FLAG affinity chromatography to remove the cross-linked but unfused RNA. The FLAG affinity purified fusions were reverse transcribed at 42°C for 1 hour, with the primer bearing the 5'-triphosphate substrate, which was produced as previously reported [51]. The cDNA was labeled with ^{32}P-d-ATP for increased radioactive detection. The efficiency of reverse transcription was calculated by densitometry analysis on SDS PAGE of a sample reverse transcribed without ^{32}P-d-ATP ('Cold RT' sample). Density of both reverse transcribed and non reverse transcribed mRNA-displayed proteins was measured. Total density was obtained by summing the two values. Efficiency of reverse transcription (E_{RT}) was calculated according to the following equation.

E_{RT}=(Density of reverse transcribed fusions/total density)*100 (eq.4)

The sample was then dialyzed 3 times against 1,000 volumes of FLAG buffer (50 mM HEPES pH=7.4, 150 mM KCl, 0.25% Triton X-100), and purified again by FLAG affinity chromatography. For the first round of selection, the resulting eluate was split into two aliquots which were incubated for 5 minutes and 60 minutes in presence of a photocleavable biotinylated 3'-OH substrate and a complementary splint for the ligation step. The ligation was quenched with 10 mM EDTA and purified by streptavidin-biotin affinity chromatography as previously described, except that before incubation with the ImmunoPure immobilized streptavidin agarose, the reaction was mixed with urea dry powder to a final concentration of 8M, heated at 90°C for 3 minutes and immediately transferred on ice. After washing, beads were recovered with 300 µL phosphate buffered saline (PBS), the obtained slurry was dispensed in 100 µL aliquots in round bottom 96 well plates, and irradiated for 15 minutes at 365 nm while shaking to release the bound cDNA. The released cDNA from the 5 minutes and 60 minutes selection was separately ethanol precipitated in presence of 1 µL glycogen (stock concentration= 20 mg/mL), washed with 70% ethanol, dissolved in 100 µL doubly distilled water and amplified in a 1 mL PCR reaction using primers BS3longb'/AM016C. The two samples were separately used as input for the following rounds of selections with 5 minute and 60 minute selective pressure respectively. For each round an SDS PAGE gel (4-12% BisTris Acrylamide) of aliquots taken at the end of each step was run (1.5 hours, 120V) for quality control purposes, fixed for 10 minutes in a solution of 20% acetic acid-10% methanol, rinsed with distilled water for 10 minutes, dried at 80°C under vacuum for 20 minutes, exposed overnight on a phosphor screen (GE Healthcare), and imaged on the scanner Storm 860 (Molecular Dynamics). The procedure was repeated for 6 rounds.

4.6.9 Processing and analysis of enriched sequences from round 6 of each selection

DNA from round 6 was cloned into a pCR-blunt-II TOPO vector, transformed and plated according to manufacturer instructions (Invitrogen). Plates were sent to Beckman Coulter for colony picking and sequencing. For both the 5 minutes and 60

minutes selection, 96 colonies were sequenced. The resulting sequences were analyzed with the software CLC genomics. Open reading frames were extracted by removing the plasmid, T7 promoter and TMV enhancer sequences by using the trimming function of the Next Generation Sequencing tools. Subsequently, sequences which contained tandem duplications, insertions and out of frame sequences were discarded after manual inspection of the first alignment. Upon a second alignment, sequences with internal stop codons and short peptides containing duplications of the 5'-terminal and 3'-terminal constant regions were discarded as well. The resulting set of sequences was aligned and inspected for the presence of deletion variants.

4.7 Supplementary information

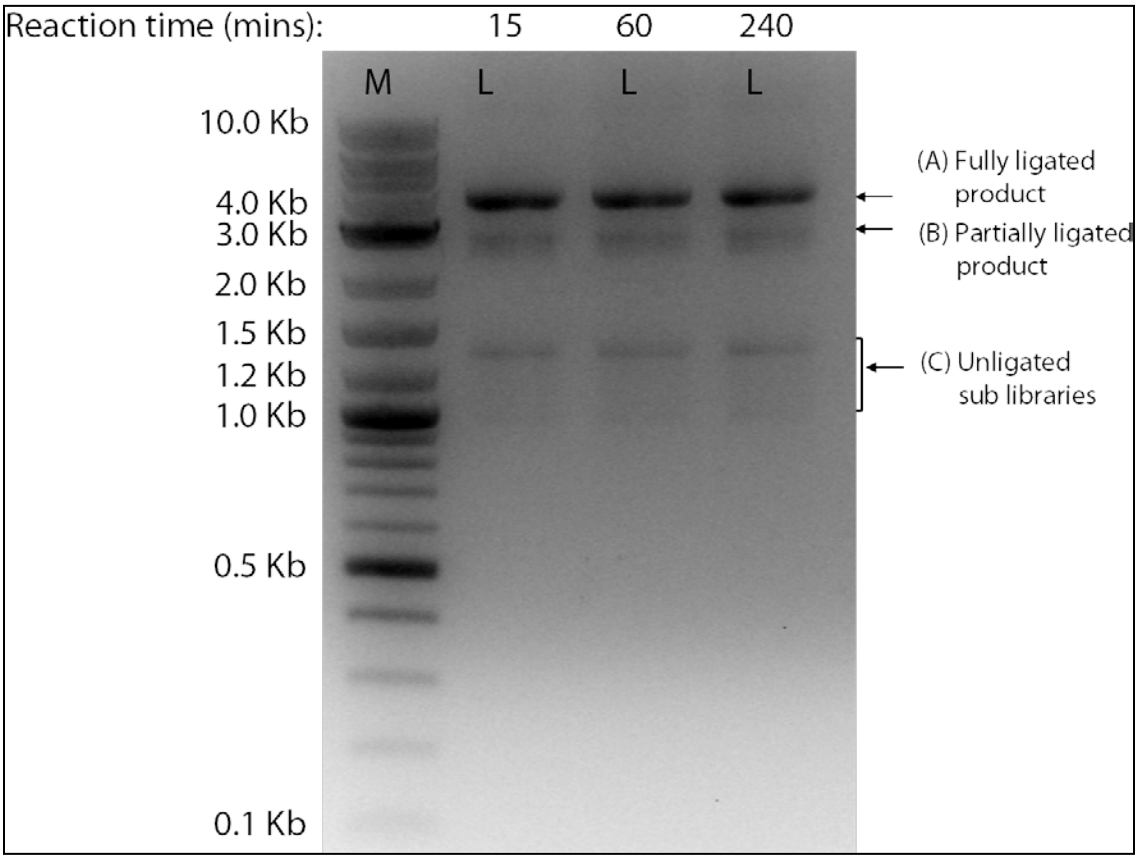

Figure S 4.1-Agarose gel (1%) of ligation of terminal truncation libraries to DNA linker.
M= Marker (2 log ladder); L=ligation; (A) Product of interest: linker with sub-libraries ligated on both termini; (B) Linker with sub-library ligated on only one terminus. (C) Unligated fragment sub-libraries Ligation conditions as reported in the methods section.

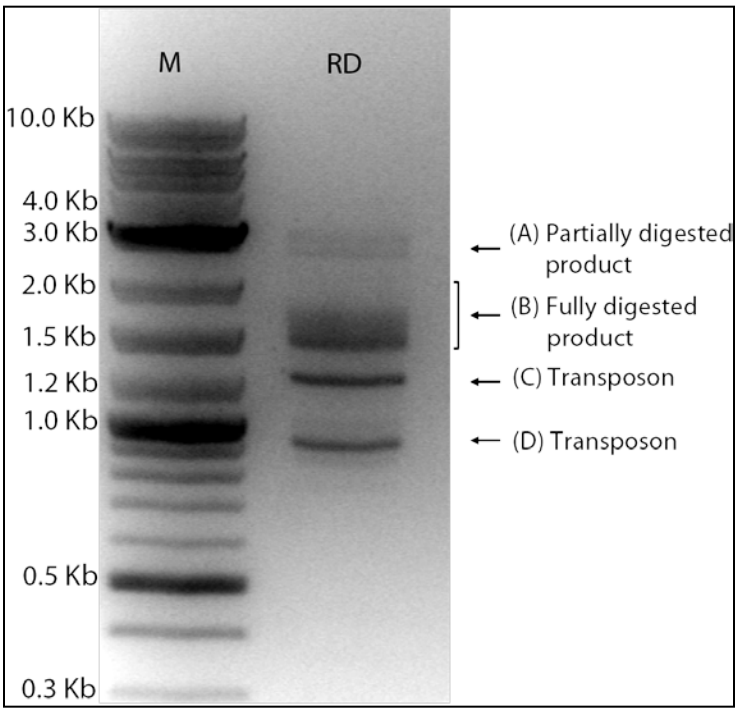

Figure S 4.2-Agarose gel (1%) of MlyI restriction digest of ligation product.
M= Marker (2 log ladder); RD= restriction digest with MlyI; (A) Partially digested product: only one transposon was removed (B) Product of interest: both transposons were removed. (C) and (D) Transposon sequences cleaved from the 3'and 5' fragment sub-libraries respectively

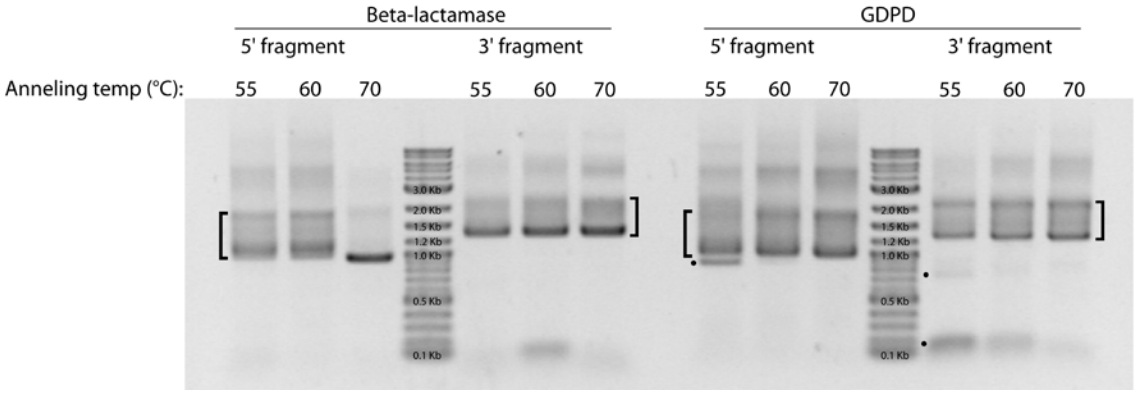

Figure S 4.3-Agarose gel (1%) of gradient PCR of beta-lactamase and GDPD fragment sub-libraries.
Squared brackets indicate the product of interest; •= undesired PCR products.

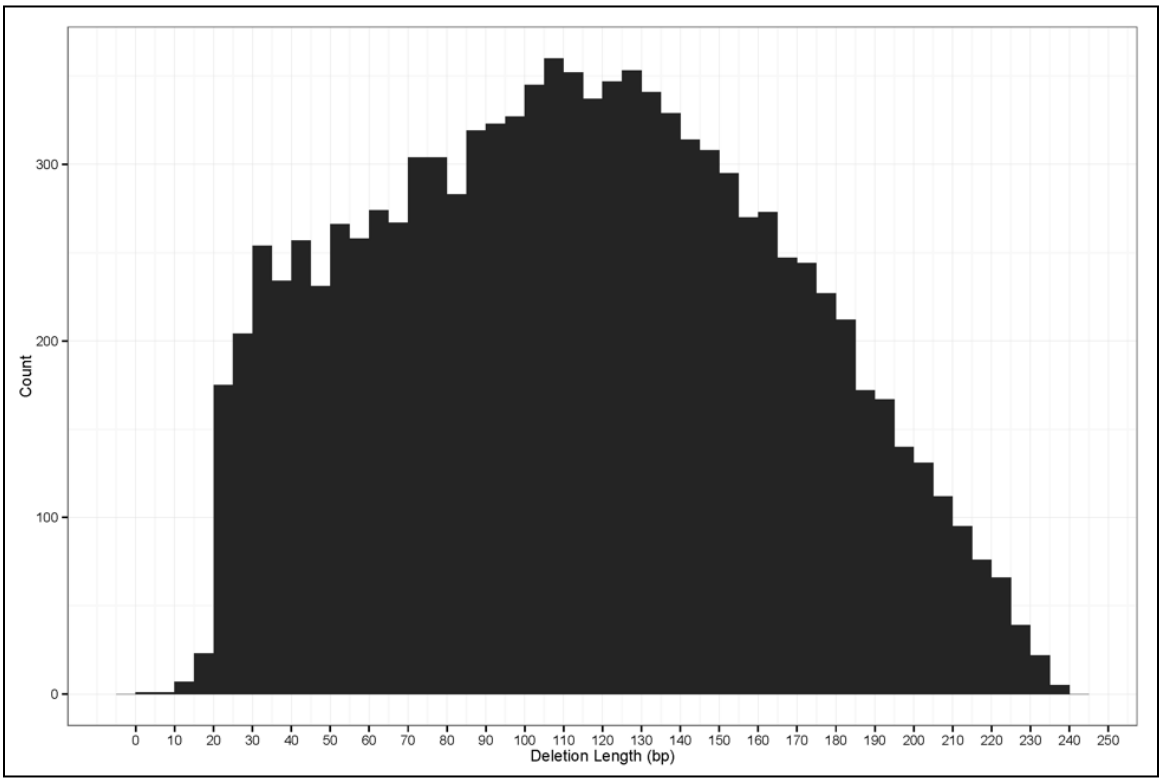

Figure S 4.4-Deletion length count for the unique sequences found in the deletion library of ligase 10C.
Sequence count is considered for 5 bp bins.

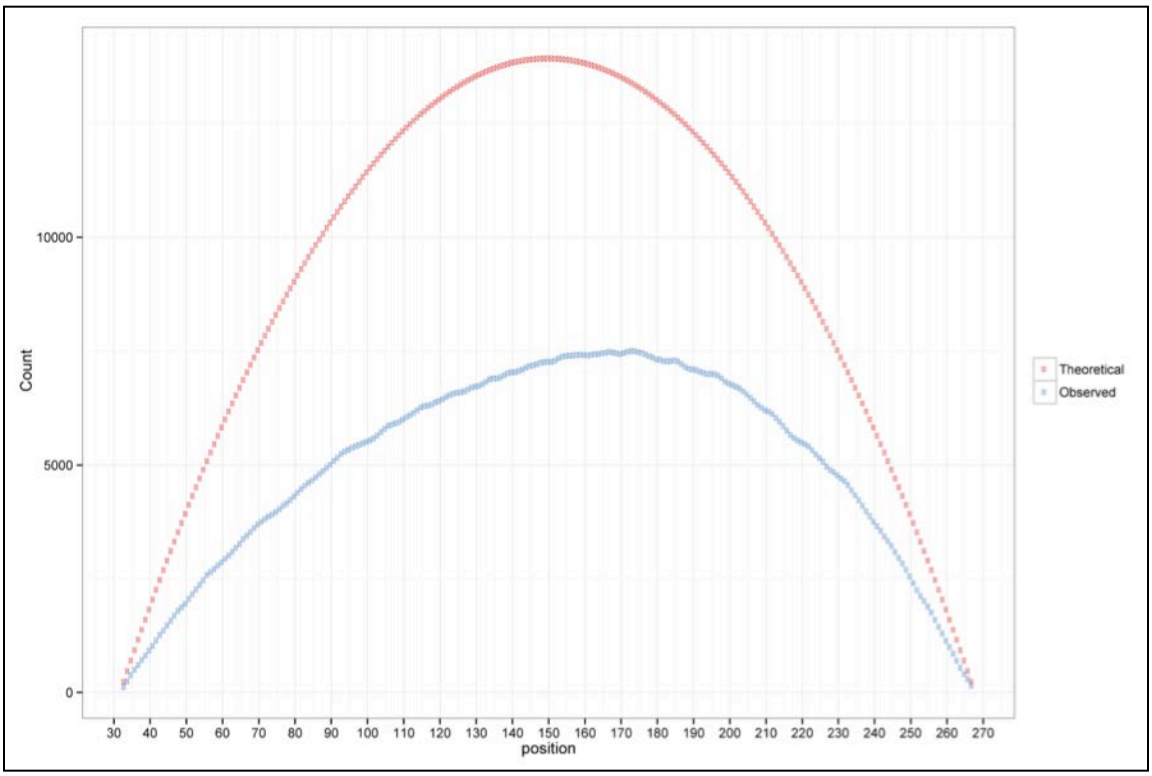

Figure S 4.5-Theoretical (red) and observed (blue) deletion counts at each potentially deleted nucleotide position.

Theoretical counts were calculated as explained in the methods section; observed counts were obtained by aligning all the unique sequences found in the library, and counting the number of instances in which a certain position was deleted.

Table S 4.1-Primers used for PCR amplification

Primers are listed in the 5' to 3' direction. The BsaI recognition site in the pUC19* reverse primer is underlined. The BglII recognition sites in the transposon primers are bold and underlined. Letters in lower case and italics in the transposon primers highlight mutations introduced to create a MlyI recognition site [212].

Amplicon	Primer pair	Amplicon length (bp)
pUC19*	Fwd: CGTGTAGATAACTACGATACG	1,404
	Rev: GCTAGGCTGAGTTGCCGCTATGAGACCCTGGCCGTCGTTTTACAACGTCG	
Ligase 10C	Fwd: ATGGACTACAAAGACGACGACGATAAG	288
	Rev: TTAATAGCCGGTGCCAGATCC	
Transposon	Fwd and Rev: GCTT**AGATCT**GA*ct*CGGCGCACGAAAAACGCGAAAG	1,320
Beta-lactamase	Fwd: ATGAGTATTCAACATTTCCG	874
	Rev: CGTTCCATGGTTATTACCAATGCTTAATCAGTGAGG	
GDPD	Fwd: AAAGAGGAGAAATTACATATGGGCAGCGATAAGATC	809
	Rev: GTTAGCGATGTACATTAATAGCCGGTGCCAGATCC	

Chapter 5 : Towards the selection of artificial RNA ligases from a library of entirely random polypeptides

This chapter describes unpublished experiments aimed at selecting an artificial RNA ligase activity from a random library by mRNA display. I performed all the experiments, and Dr. Seelig and I analyzed the results.

5.1 Summary

Early functional proteins presumably arose form random peptide sequences, which subsequently evolved for improved structure and increased performance. Several studies, both *in silico* and *in vivo* focused on testing random libraries for the presence of folded and functional proteins. While proteins with binding activities have successfully been isolated from random peptide libraries, there has been no report on the selection of protein enzymes from random sequence space. In this chapter we will present our attempts to select an RNA ligase from a random library by mRNA display. Any active variant isolated in this selection will be compared to ligase 10C. The long term goal of the lab is to characterize and compare structure and mechanism of enzymes that catalyze the same reactions, but from different origins, to identify potential evolutionary pathways.

5.2 Introduction

In extant cells, proteins perform various functions: regulation of intra and inter cellular communication, transport of molecules across the membrane and between the cellular compartments, maintenance of structural integrity, and catalysis of biochemical reactions. The folding and functionality of natural proteins depend on their primary amino acid sequence. However, extant proteins represent only a very small fraction of the vast number of all possible polypeptides that are referred to as the protein sequence space [222,223]**.** For example, it has been calculated that the ratio of natural proteins to all possible sequences, is similar to the ratio of the size of a hydrogen atom and the size of the Universe [103]. Given the size of sequence space, natural evolution can only ever

sample an extremely small subset of possible sequences for functional proteins [103]. This observation raises the question of how early functional polypeptides were selected and evolved.

It has been proposed that the earliest functional proteins originated from random polypeptides [103,224]. For example, it has been shown at the sequence level that residue distribution in natural proteins follows a random pattern based on properties such as hydrophobicity, charge, volume and secondary structure propensity, except for amino acids with a propensity for forming α-helices, which tend to cluster together [224-226].

In addition, a set of random proteins generated computationally and a set of natural proteins were compared for secondary structural content *in silico* [102]. For both sets the total content was over 60%. Furthermore, tertiary structural simulations for 6 sequences picked randomly from the set generated computationally revealed ordered structures, suggesting that a high fraction of the random polypeptides was folded.

Various studies addressed experimentally the questions of whether soluble and folded proteins could be found in random sequence space. Within libraries of random polypeptides of 95 amino acids [227] and 70 amino acids in length [228]**,** it was found that 20% of the analyzed clones are soluble in *E.coli* lysate [227] and that 20% are folded [228] by using resistance to proteolysis as folding criterion. In another study [229], random decapeptides synthesized by Merrifield synthesis, were subjected to cycles of condensation and selection for solubility under certain conditions of pH, salinity and temperature, to finally obtain a 44 residue random peptide, soluble in water with 49% α-helix, 12% β-sheet, and 39% random coil, as determined by circular dichroism. Other studies showed that random proteins can substitute domains of natural proteins without disrupting functionality [230] and that random elongation mutagenesis resulted in increased thermal stability of natural enzymes [231].

The above-mentioned studies focused on solubility and structure. However in order to consider the hypothesis of enzymes originated from random sequences plausible, there is the need for experimental evidence that random proteins can be functional and evolvable. Several articles support this idea by showing that random proteins can be

evolved for solubility [232], for DNA binding [233], and that unstructured random proteins can develop secondary structural elements within the process of selection for DNA binding [234].

The most striking result was obtained by Keefe and Szostak [105]**.** They identified novel ATP binders from a library of 6 x 10^{12} random polypeptides using the mRNA display *in vitro* selection method. The identified binders have no sequence and structural counterpart in nature, as assessed by crystallography [106]. This work estimated the frequency of functional ATP-binding proteins to be 1 in 10^{11} random sequences.

Esterase activity towards *p*-nitrophenyl acetate was previously observed by sampling a small number random polypeptides [108-110], and in a random binary patterned library capable of forming alpha helices [110]. However, the ester hydrolysis of *p*-nitrophenyl acetate has been shown many times to be particularly easy to catalyze even by non-enzyme proteins. Therefore, this reaction is considered a poorly suited benchmark to study the evolution of enzymatic activity from random sequence space [107].

Indeed all the results above showed that random proteins are capable of folding into tertiary structures and evolve to perform different functions, which is mainly binding. However the selection of an efficient catalytic activity from an entirely random library has not been reported to date. The goal of my project was to investigate whether a completely novel enzymatic activity could be selected from an unbiased random library, which neither had been evolved for substrate binding, nor contained any predetermined secondary structure patterns. This would lend support for the hypothesis that the first enzymes could have arisen from random sequences. If successful, this would also open new avenues for the selection of novel enzymatic activities for practical applications.

The long-term goal for this line of research envisions to compare the enzymes selected from the scaffold library [51] with enzymatic activities that might result from the random library. These studies will give insight into different solutions available for a catalytic problem, and how these compare to each other, and will potentially provide a simple system to investigate mechanisms of convergent evolution.

5.3 Results

In order to perform this experiment, we applied the same selection procedure used previously [51], selecting for the same activity of a 5'--triphosphate-dependent RNA ligase. However, instead of starting from the scaffold library, we used a random library previously synthesized [37], which had also been successfully used for the selection of ATP binders [105].

The random region of the mRNA display library is framed by constant regions needed to produce and purify the mRNA-displayed proteins, and to amplify the cDNA obtained at the end of each round [52]. The enhancer sequence is also used for as binding site for the forward primer. The previously selected RNA ligase [127,213] had been manipulated in our laboratory and it has the same enhancer as the random library [37,105]. In order to avoid contamination we substituted the original Tobacco Mosaic Virus sequence with an Alfalfa Mosaic Virus sequence by PCR before starting the selection procedure. Successful change was confirmed by TOPO Cloning the resulting DNA and sequencing 5 random clones (Figure S5.1).

The mRNA display procedure was carried out for 13 rounds of cycles of selection and amplification as reported previously [51,52] and as outlined in section 4.6.8. The percentage of cDNA bound to the streptavidin beads was calculated with equation 3 in section 4.6.8 to monitor the progress of the selection.

Prior to starting the selection, we performed pilot experiments to ensure that the efficiency of all the steps was comparable to results reported previously [52] . Different from the previous method, we slightly changed the streptavidin-biotin affinity chromatography procedure to increase the binding efficiency by subjecting fusions to heat denaturation for 3 minutes at 90°C in presence of 8M urea. A mock selection with a positive control of ligation showed that the denaturing treatment increased the percentage of binding to 38.3%, compared to the sample treated as in the previous protocol (25.8%), while a negative control gave a background binding of 0.08%. The positive control was

obtained by reverse transcribing purified mRNA-displayed proteins with a primer attached to a pre-ligated biotinylated product. This primer will be called 66 mer throughout this chapter.

To assess quality of the samples mRNA-displayed proteins were analyzed by SDS PAGE (Figure 5.1). The samples taken after Oligo-dT-cellulose purification and FLAG purification contained purified fusions as expected (Figure 5.1, compare 'TL' lane to 'OdT' and 'FLAG' lanes). In presence of ^{32}P-α-dATP, a secondary band at a lower molecular weight was observed. This band corresponded to cross-linked RNA not fused to protein which apparently had not completely been removed during the first FLAG purification procedure. The second FLAG purification process (Flag II) efficiently removed this band. Finally, when fusions were incubated with 8M urea at 90°C ('Rxn+urea+90°C'), two bands of lower electrophoretic mobility were observed. These bands likely represent an mRNA cDNA complex formed by re-annealing of only the constant terminal regions, and cDNA alone.

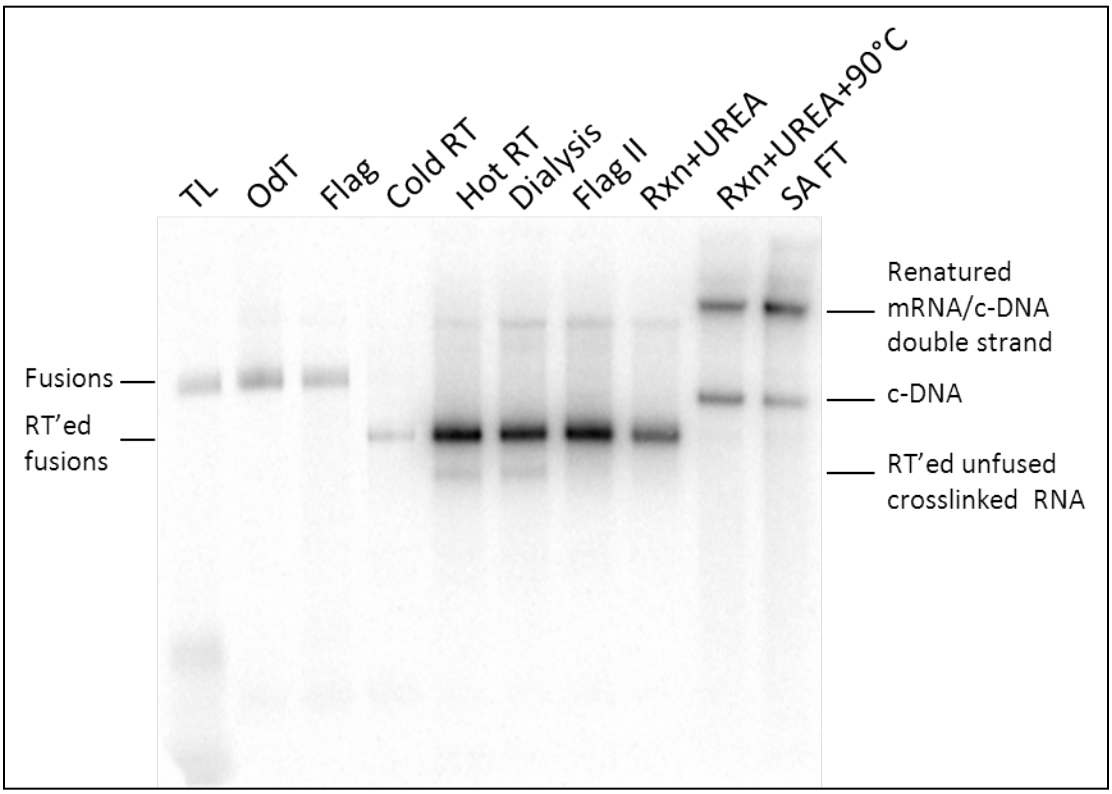

Figure 5.1-Autoradiogram of SDS PAGE gel analyzing individual steps during a typical round of selection.

TL: translation; OdT: eluate from oligo-(dT) cellulose chromatography; Flag: eluate from Flag affinity purification; Cold RT: reverse transcription without ^{32}P-α-dATP; Hot RT: reverse transcription with ^{32}P-α-dATP; Dialysis: sample recovered after dialysis in Flag buffer; Flag II: eluate from second Flag affinity purification; Rxn+UREA: quenched overnight ligation mixed with 8 M urea; Rxn+UREA+90°C: like previous sample, then heated at 90°C for 3 minutes and chilled on ice; SA FT: flow-through of streptavidin biotin affinity chromatography.

The progress of the *in vitro* selection was monitored by calculating the percentage of cDNA bound to the streptavidin resin at the end of each round. In any immobilization, a certain level of background binding is observed, which was around 0.1% for this selection. Figure 5.2 shows that no substantial increase of the signal over background was achieved after any of the rounds of selection, suggesting that no active sequences were enriched. In order to further test this assumption, cDNA isolated after round 13 was cloned by TOPO TA cloning and 96 individual colonies were sequenced. In total, 92 colonies gave good quality data and 83 sequences coded for in frame proteins. An

alignment of these 83 protein sequences (Figure S5.2 and Figure S5.3) revealed three major families: I, II and III, with family III comprising the largest number of sequences (30/83) (Figure S5.2-Figure S5.6). Six additional duplicated sequences and 17 single sequences were also present.

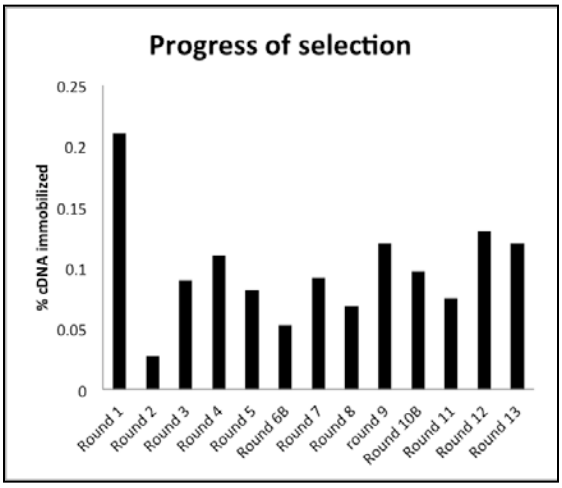

Figure 5.2-Selection progress.
The percentage of ^{32}P-labelled cDNA bound to the streptavidin Agarose resin at the end of each round monitored. No substantial increase of the signal over background was observed.

Eleven sequences, one chosen from each family and from several of the duplicate sequences were individually subjected to one round of selection. Their activity was tested by a gel shift mobility assay. Incubation of the ligation reaction with streptavidin resulted in a cDNA band with reduced mobility if the clone was active. None of the clones showed a shift, indicating that they were not active (Figure 5.3). As a positive control experiment, we reverse transcribed clones A5 and a mix of clones F2/F3/F9 with the 66 mer primer. Both A5 (gel A) and clones F2/F3/F9 (gel B) showed a shift when reverse transcribed with a positive control primer. The shift was not observed when streptavidin was not included (sample labeled as –SA), indicating that it was specifically due to binding of streptavidin to the biotinylated cDNA. An active RNA ligase (clone 13, gel B) from the previous selection [51] was used as an additional positive control for the reaction conditions. This clone in presence of streptavidin showed a shifted cDNA band,

indicating that it was active in the experimental conditions. Results were further validated by incubating clone 13 and sequence F9 from round 13 with ligation substrate and then subjecting them to the entire streptavidin biotin affinity chromatography procedure. Incubation of clone 13 with streptavidin resin resulted in 0.69% of the total radioactivity bound to the resin consistent with activity, while for F9 only 0.05% of the total radioactivity bound to the resin, and was indistinguishable from the percentage of background that bound when biotinylated substrate was not added to the reaction (0.06%).

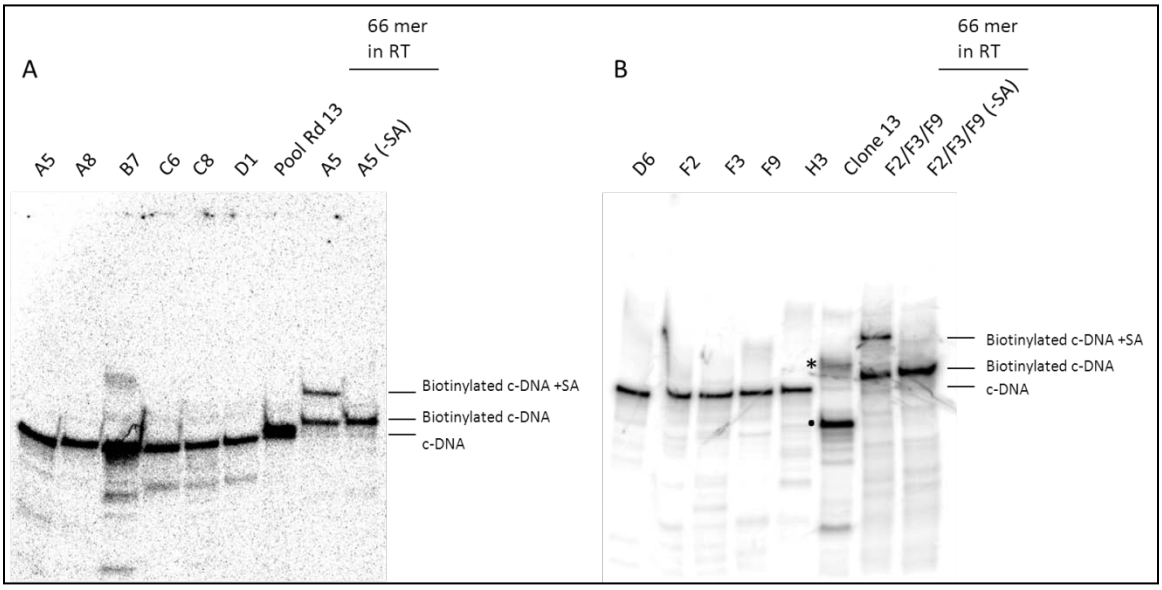

Figure 5.3-Gel shift assay to test ligase activity of clones isolated after round 13.
All samples were incubated with streptavidin otherwise indicated by (-SA). **(A)** Clones A5, A8, and B7 are unrelated sequences that were found twice in the alignment; Clones C6, C8 and D1 belonged to family II; Pool Rd 13 referred to the cDNA mixture after round 13. Clone A5 was also reverse transcribed with a positive control primer (66 mer) to verify binding of the biotinylated cDNA to streptavidin in the experimental conditions. **(B)** Clone D6 was found twice in the alignment; clones F2, F3 and F9 belong to family III, clone H3 belongs to family I. Clone 13 is an active ligase from a previous selection [51], and was used as positive control for reaction conditions; the dot indicates biotinylated cDNA, and the asterisk marks the shifted biotinylated cDNA. A mixture of clones F2/F3/F9 was also reverse transcribed with a positive control primer (66 mer) to verify binding of the biotinylated cDNA to streptavidin in the experimental conditions.

5.4 Discussion

The detailed analysis of the cDNA isolated after 13 rounds of selection indicated that no active ligase enzymes were enriched during the selection procedure. It is impossible to determine with confidence the reason for this negative result, however different potential reasons will be discussed in the following.

There is a possibility that the lack of enrichment of ligase enzymes was due to technical reasons such as biases in all but the selection step that could favor the enrichment of sequences by properties other than ligation activity. This hypothesis is supported by the fact that in round 13 three sequence families were observed, but upon screening of individual clones none of them showed activity. These sequences were clearly enriched. Those enrichment biases could be for example preferential transcription, translation, enhanced affinity to oligo(dT) or Flag resin, preferential reverse transcription or PCR amplification. Those inadvertently enriched sequences were then non-specifically carried through the streptavidin-biotin affinity chromatography step as a background. In particular, family II possesses two FLAG peptide motifs, and family I possesses a partial FLAG motif such as DYK, suggesting enrichment during the FLAG affinity chromatography step.

Another possibility is that the library does not contain active ligases. The selection of an enzymatic activity might require sampling of a larger library. Another potential reason for failing to find active enzymes is that different reaction conditions are needed to select for active RNA ligases. In order to explore this possibility, we are currently performing another selection, in presence of metal cofactors to aid folding and/or catalysis. Previous mRNA display selections have shown that metal coordination provides structural organization and supports functionality in the context of unbiased libraries [105], or in the context of randomized scaffold libraries [51], even when most of the structural information carried by the parent protein is lost and replaced by disordered regions [127].

5.5 Conclusions

Our selection did not yield active ligase variants. While multiple factors could have contributed to this, our current focus is on testing different reaction conditions. I am repeating the selection in presence of metal cofactors, which could facilitate folding and/or catalysis. Furthermore, it would be desirable to reduce the apparent bias for enhanced Flag binding by modifying the selection procedure.

5.6 Materials and methods

All chemicals were purchased from Sigma-Aldrich unless otherwise stated.

5.6.1 PCR amplification of input DNA for round 1 and subsequent rounds.

All chemicals for PCR were purchased from NEB, primers were synthesized by IDT, and Taq Polylmerase was expressed in house and purified by Ni-NTA chromatography.

DNA for round 1 was amplified by PCR in a 40 mL reaction, from a previously synthesized random library [37,105]. The template concentration was 6.4 nM, and the amplification was carried out for 5 cycles in presence of deoxynucleotides (200 μM each) Taq Polymerase (1.08 U/μL), 1X Thermopol buffer [20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$ 0.1% Triton® X-100, pH 8.8 at 25°C], and primers BS97 and BS24B (0.5 μM each). Thermal cycling comprised three steps: 94°C for 30 seconds, 55°C for 45 seconds, 72°C for 1 minute, except for first cycle where temperature was held at 94°C for 3.5 minutes. Primer BS97 replaced the TMV enhancer sequence with an AMV enhancer sequence. DNA obtained from subsequent rounds of selection was instead amplified with primers BS98B and primers BS24B'. Primer BS98B binds to the AMV enhancer sequence.

5.6.2 Purification of PCR products

DNA for the first and subsequent rounds input DNA was purified by extracting the PCR reaction with one volume of phenol, followed by extraction of the phenol phase with half a volume of water. The aqueous phase was combined, extracted with three volumes of chloroform, concentrated by mixing with butanol, and ethanol precipitated. The resulting pellet was dissolved in TE buffer (10mM Tris pH=8, 1 mM EDTA). Separation of the organic and aqueous phase during purification of the DNA for the first round was obtained by centrifugation at 3,000 rpms, for 10 minutes at 25°C, in 50 mL conical tubes, due to the large volume of the PCR reaction (40 mL). In subsequent rounds centrifugation was performed at 1,000 rpms for 30 seconds at 25°C in 1.5 mL tubes.

5.6.3 mRNA display selection for ligase enzymes

mRNA display was performed as previously [51], and as reported in section 4.6.8. except for differences reported here. Zinc Chloride and 2-mercaptoethanol were omitted from the buffers, and wherever Triton was present its concentration was increased from 0.01% to 0.25%. RNA was purified by electro elution for 2 hours at 300 V and concentrated by ethanol precipitation.

For the first round a 10 mL translation was performed. For rounds 2-6, the volume was decreased to 2 mL, and for rounds 7-13 to 1 mL. After reverse transcription the sample was dialyzed 3 times against 1,000 volumes of FLAG buffer, (50 mM Hepes pH=7.4, 150 mM KCl, 0.25% Triton X-100), and purified by a second FLAG affinity chromatography. Ligation reaction was performed for 16 hours. At the end of the first round of selection, the streptavidin agarose resin was used directly as template for PCR amplification (50 μL agarose/ mL PCR). In following rounds the cDNA was eluted by photocleavage in 300 μL phosphate buffered saline (PBS), ethanol precipitated in presence of 1 μL glycogen (stock concentration= 20 mg/mL), washed with 70% ethanol, dissolved in 100 μL doubly distilled water and amplified in a 1 mL PCR reaction using

the primers BS98B/BS24B'. SDS PAGE gel electrophoresis, gel fixation and imaging were conducted as in section 4.6.8. The procedure was repeated for 13 rounds.

5.6.4 TOPO TA cloning

TOPO TA Cloning was performed according to manufacturer procedures. For high-throughput sequencing purposes DNA from round 13 was treated as follows to increase cloning efficiency. The PCR product, previously purified by phenol/chloroform extraction and ethanol precipitation was incubated with dATP (200 μM), Thermo Pol buffer (1X), and Taq Pol (0.022 U/ μL) at 72°C for 15 minutes, purified by Phenol/Chloroform extraction and ethanol precipitation and gel purified with a commercial kit (QIAGEN) according to manufacturer guidelines. Cloning reaction was kept at 25°C for 30 minutes. TOP 10 chemically competent cells (Invitrogen) were transformed with Half of the reaction (3 μL). After 1 hour recovery in SOC media at 37°C, cells were spread on LB/Kanamycin (36 μg/ μL) plate, previously spread with 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) for blue white screening.

5.6.5 Sequencing of individual colonies

Plate obtained from TOPO TA cloning were sent to Beckman Coulter for colony picking, plasmid purification and sequencing.

5.6.6 Analysis of sequencing results

Sequencing results were analyzed with the Clone Manager, Bio edit and Clustal W software [235].

5.6.7 Ligation activity assay of mRNA-displayed proteins by gel shift

The sequences chosen for screening were PCR amplified by using primers BS98B and BS24B', and each sequence was individually mRNA-displayed as outlined above, until the dialysis step. Dialyzed fusions were incubated overnight with 3'-OH biotinylated substrate and splint. The ligation reaction was quenched with 3 volumes of loading

buffer (8 M urea, 5% glycerol, 6.5 mM EDTA), heated at 94°C for 3 minutes, transferred immediately on ice and then mixed with an excess of streptavidin. Results were analyzed by 4% denaturing UREA PAGE.

5.7 Supplementary information

```
Random library     1 tctaatacgactcactatagggtttttatttttaattttctttcaaatacttccaccatggactacaaagacgacgacgataagaagactnactgnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
Clone 1            1 tctaat-cgactc-ctatagggtttttatttttaattttctttcaaatacttccaccatggactacaaagacgacgacgataagaagacttactggttattgatgacggacgattcggctcaggggtccc
Clone 2            1 tctaatacgactcactatagggtttttatttttaattttctttcaaatacttccaccatggactacaaagacgacgacgataagaagactaactgttaccggagggagtgtgttcatgataactgggtga
Clone 3            1 ----atacgactcactatagggtttttatttttaattttctttcaaatacttccaccatggactacaaagacgacgacgataagaagactaactgtaatctggcgttctggcctgttaatggggtgaatt
Clone 5            1 tctaatacgactcactatagggtttttatttttaattttctttcaaatacttccaccatggactacaaagacgacgacgataagaagactaactgttggcataagaacgcggtgtgcaccaatagggaga
Clone 7            1 tctaatacgactcactatagggtttttatttttaattttctttcaaatacttccaccatggactacaaagacgacgacgataagaagactgactggtggttggcccagctggtgaacgagactgccaccc

Random library   131 nnnnnnnnnnnnnnnnnnnnnactgnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnactgnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
Clone 1          129 ggttcgacaggacgggggggaactgcgggattcacgcgaccgttaaggggtggttgaaggcgaacaataacctccacttcaactgccatccgacgaagattattattctcacggtctcgatgtccgagca
Clone 2          131 ttgatgcgtacgggcacacgtactgcacgcggctgaagtattcgagtaaggccattacgaagtttcagatcatggtcgagaactggagtgacgggaattatcttaatagggatccgaacgagggttcgat
Clone 3          127 attcttcgctcggtaggtgcaactgtgggacgatgggcaatatgatgtactgcccggagacgacgtacaagtcctatgagcactgcaactggcagactacctttaaggttttgtgggagtgtgctaacag
Clone 5          131 tgaacggtattatggggtgctactgggggtttttcgggatggggcctacggttcatatgaacatttttaggttctggcattactgcgctaatgacatgggtgactgcgagtatcaggccacttttcacta
Clone 7          131 tgaggaaggggattagtgctaactgtgctatggtcatgaccaggaggatctgtattaggaccggtagcagggggtcaataaactggcgcactgggtttaggaatcctaagagggtgctgtggaagcagga

Random library   261 nnnnnnnnnnnactgnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnactggtcagcgagctgccatcatcatcatcatcatatgggaatgtctggatctggcacc
Clone 1          259 gaggttcacgtactgtgacagctgttacgggcctacgatgatgaagtctatgcagtttcactggcattgtaactggtcagcgagctgccatcatcatcatcatcatatgggaatgtctggatctggcacc
Clone 2          261 taataactcgaactggaagtgcgagaccgcgatgtattctgagaagattgtggctagggcggatctgatgaactggtcagcgagctgccatcatcatcatcatcatatgggaatgtctggatctggcacc
Clone 3          257 tactcaccccaactggcacgtgtagacgcagaggctgatggttatttaggataacaagcgctactttggggactggtcagcgagc-gccatcatcatcatcatcatatgggaatgtctggatctggcacc
Clone 5          261 tgggaacgtgcactgtcattacaggaagtcggtggtgacttggtggaaggccgggaacacgccgatgtgctactggtcagcgagctgccatcatcatcatcatcatatgggaatgtctggatctggcacc
Clone 7          261 caacaaccagcactgtgtgttcgcggagatcgattgccttaacgtcacggacatgttttgcaggctgtttgactggtcagcgagctgccatcatcatcatcatcatatgggaatgtctggatctggcacc

Random library   391 ggctattaa
Clone 1          389 ggctattaa
Clone 2          391 ggctattaa
Clone 3          386 ggctattaa
Clone 5          391 ggctattaa
Clone 7          391 ggctattaa
```

Figure S 5.1- Sequences of 5 randomly chosen clones from input DNA for first round of selection, after performing TOPO TA Cloning.
As expected, all five clones contained the terminal constant regions framing the 80 amino acid random region. The four constant ACTG regions in the random region are left-over restriction and ligation sites to enable the library construction from sub-libraries of 20 random amino acids fragments [37].

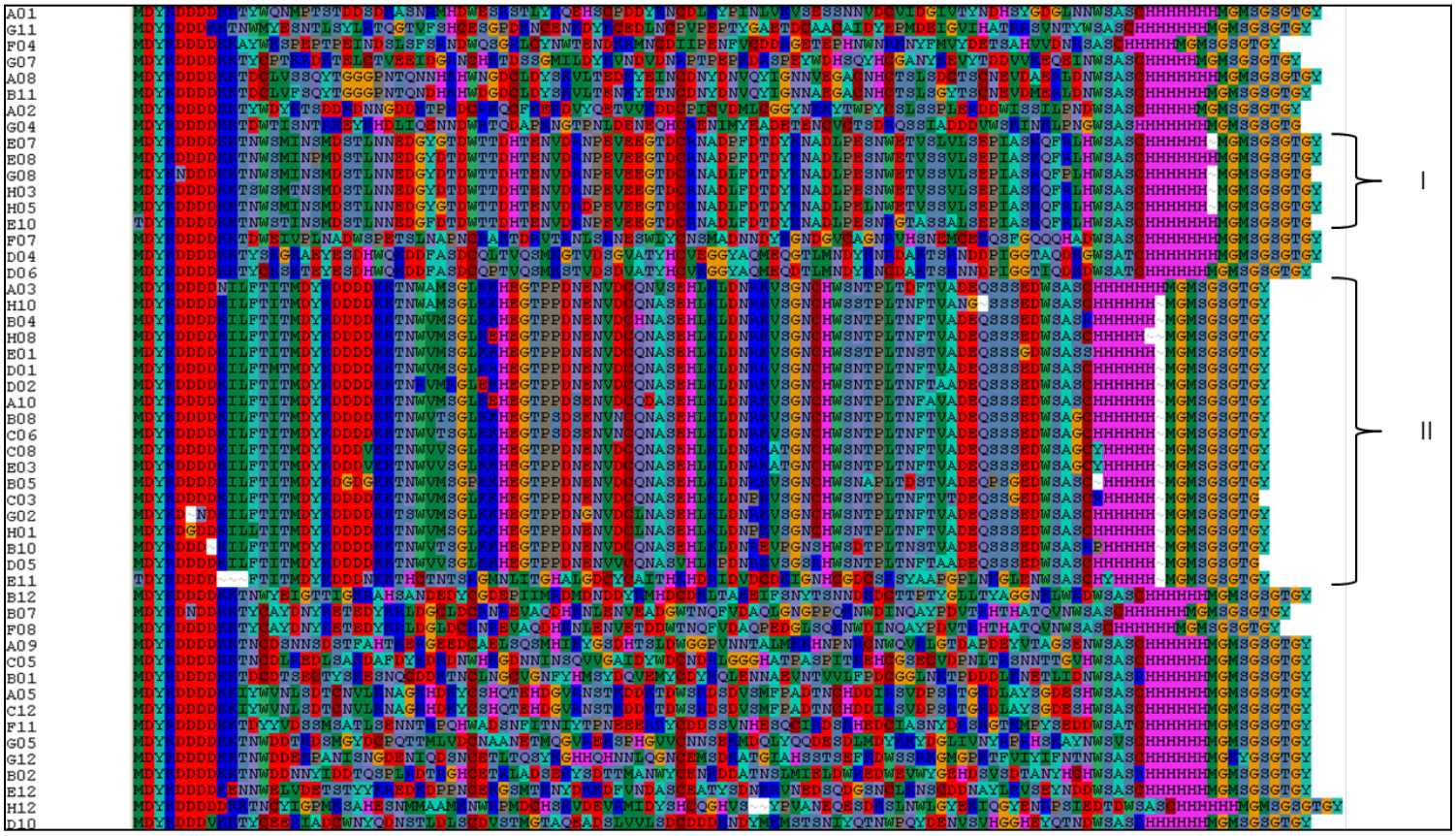

Figure S 5.2-Part I of alignment of protein sequences isolated after round 13.

This part of the alignment shows 51 out of 83 total sequences, and includes family I and II. Family I comprised 6 nearly identical sequences (7% of the total) and family II comprised 18 closely-related sequences (22% of the total). Family II contained a duplication of the FLAG peptide motif (MDYKDDDDK) starting at position 16.

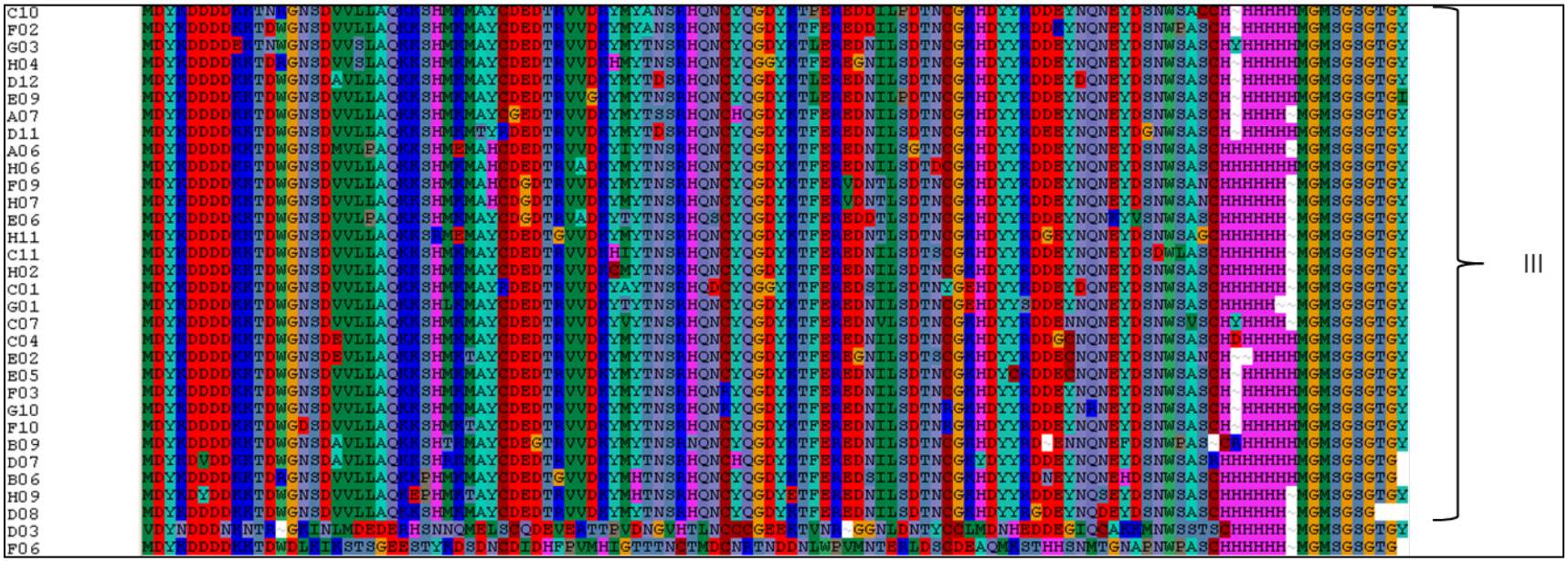

Figure S 5.3-Part II of alignment of protein sequences from round 13.
This part of the alignment shows 32 out of 83 total sequences and included family III which comprised 30 sequences (36% of the total).

Figure S 5.4-Family I from round 13.

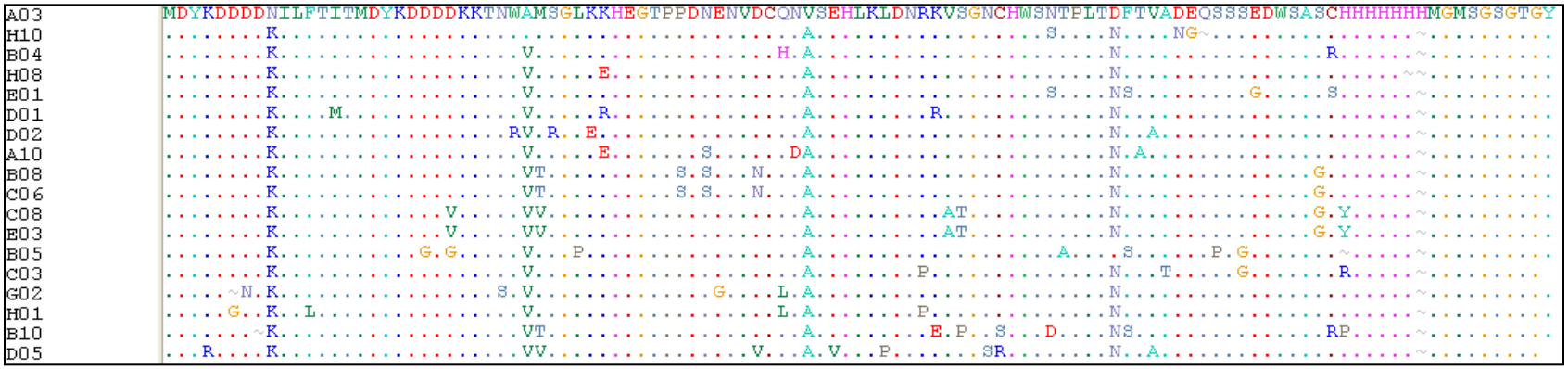

Figure S 5.5-Family II from round 13.

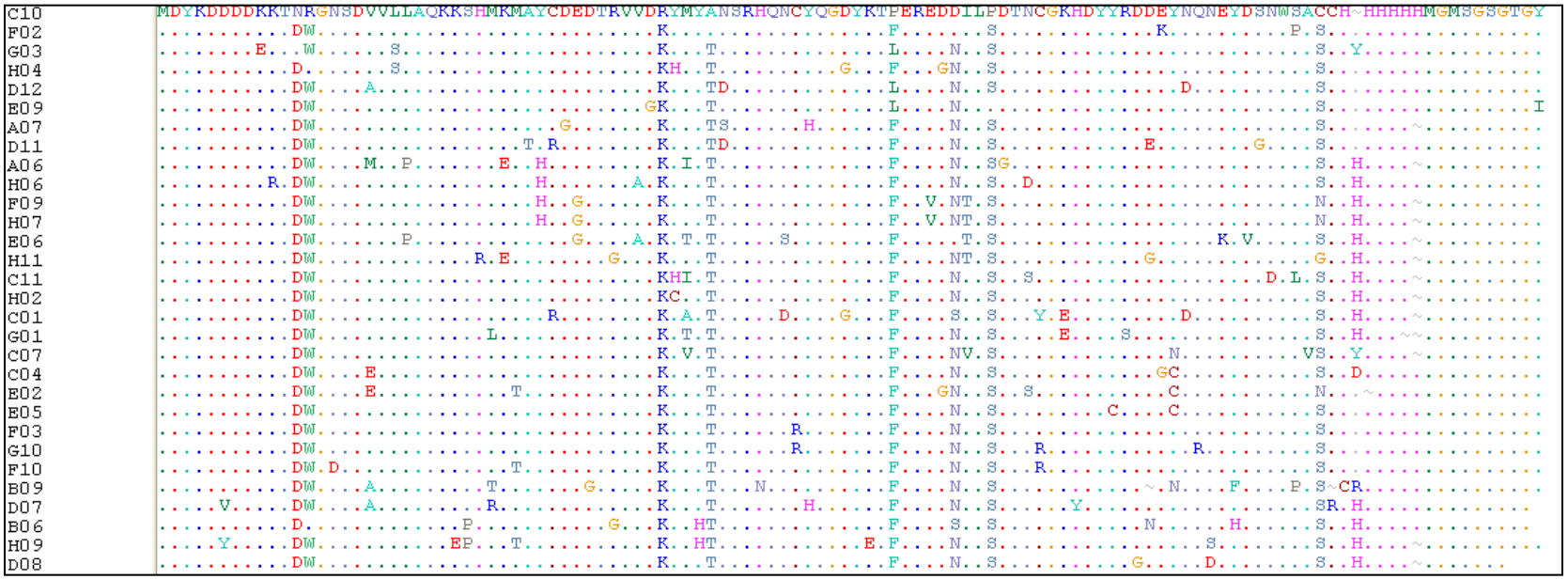

Figure S 5.6-Family III from round 13

Chapter 6 : Conclusions and future directions

Several *de novo* enzymes have been generated by computational design and *in vitro* directed evolution. While artificial enzymes hold great potential to expand the type of biocatalytic reactions available for industrial and research purposes, they are usually slower than natural enzymes. For example, ligase 10C has a turnover of less than 1 per hour. It is therefore crucial to improve the efficiency of these enzymes through engineering and evolution. In-depth studies of their structural and functional properties will enable us to generate better *de novo* enzymes in the future.

This thesis focused on the structural and biochemical characterization of the artificial RNA ligase 10C. In chapter 3, we solved the structure of ligase 10C by NMR and found that the enzyme had an unusually dynamic fold. While the protein is folded, it surprisingly lacked secondary structure elements such as α-helices and β-strands. It has been suggested that for example viral proteins are more evolvable than proteins with a tightly packed hydrophobic core, because they can accept more mutations due to their flexibility and lack of inter-residue connectivity [144]. Similarly, the flexible structure of ligase 10C could be highly evolvable, and therefore used as a model to study the optimization of primordial catalysts as ligase 10C had not been subjected to billions of years of selection and evolution like modern enzymes. Therefore, future studies in the Seelig lab will include further improvement of the catalytic activity of ligase 10C.

Ligase 10C is active at 65°C and thermostable with a melting temperature of 72°C, despite its dynamic structure. We compared the enzyme to two mesophilic ligase variants, selected at 23°C. In particular, sequence analysis of ligase 10C and the most closely related mesophilic variant ligase #7 suggested that the increased stability of ligase 10C is due to several point mutations and a 13 amino acids deletion which favors inter-residue interactions between the two termini. Furthermore, in contrast to the notion that there is a trade-off between enzyme stability at high temperatures and activity at lower temperatures, ligase 10C is more active at 23°C than at 65°C.

We have so far been unable to crystallize ligase 10C likely because it contains three flexible regions, namely at both termini and in an internal loop. We hence generated

a library of random deletions variants of ligase 10C and subjected it to six rounds of mRNA display selection to identify shorter variants, predicting that deletion of some flexible regions could facilitate crystallization. We isolated two variants with deletions of 13 and 18 amino acids near the N-terminus of the enzyme, which now need to be tested for crystallization. Yet, as two of the three flexible regions are still present in those variants, crystallization might continue to be a challenge. Through the *in vitro* selection we also identified several other variants with point mutations suggesting that these substitutions had a neutral or potentially a beneficial effect on activity. We will express the deletion mutants and the variants with point mutations individually as mRNA displayed proteins, screen them for activity, and use the most active clones for further characterization as purified free proteins.

The power of Next Generation Sequencing (NGS) data analysis could be exploited to facilitate the engineering of enzymes by directed evolution. For example, DNA from individual rounds of a selection for a Diels-Alderase ribozyme, was recently sequenced by NGS [236]. The analysis showed a good correlation between increased selective pressure and enrichment of substitutions at certain positions. These positions had been previously identified as important catalytic residues in structure-function relationship studies [236]. These results suggest that NGS data analysis could also be used to identify key target residues in ligase 10C for further directed evolution experiments. This type of approach would be particularly useful as detailed structural or biochemical information about the catalytic active site might be difficult to acquire.

We also performed a selection for RNA ligase enzymes from a completely random library, although no active enzymes have been enriched yet. Our long-term goal with this project is to characterize and compare future enzymes isolated from the random library to ligase 10C, in order to explore how different evolutionary solutions to the same catalytic problem might relate. We are now repeating this selection while using modified conditions. We included a number of divalent metal ions to the reaction with the hypothesis that these could potentially function as cofactors to aid folding and catalysis. This project is currently underway and therefore not reported here in more detail.

Ligase 10C was the first *de novo* enzyme obtained using a generalizable selection scheme for the isolation of enzymes by mRNA display However, enzymes for other bond forming reaction could be selected by following a similar strategy. The only requirement is the ability to modify substrates so that bond formation can be coupled to binding onto a solid support for purification purposes [51].

Furthermore, enzymes that catalyze bond-breaking reactions could be selected with mRNA display. In this case, the mRNA-displayed protein library would be immobilized onto a solid support via the substrate, and the cDNA released upon bond breaking would be amplified by PCR as input for the next round of selection [35]. Finally, mRNA display could be employed for the selection of enzymes which induce structural rearrangements, such as isomerases, oxidoreductases or transferases. In that case, selection can be achieved by using product-specific affinity reagents such as an antibody that binds to the reaction product but not the substrate.

The selection of *de novo* enzymes by *in vitro* directed evolution is still in its infancy, but here we have clearly shown that the method is powerful and versatile. In particular, we demonstrated an enzyme selection at high temperature, but other conditions such acidic or basic pHs, or organic solvents could be applied, as long as the RNA is not degraded. Previous experience with *de novo* enzymes obtained by computational design showed that further directed evolution can lead to 1,000 fold improvements in K_{cat}/K_m. Ligase 10C also has a relatively low activity, however additional rounds of directed evolution with methods that can directly select for multiple turnovers, such as *in vitro* compartmentalization, hold great potential to obtain improved catalysts. Despite the fact that *de novo* enzymes have so far shown only modest levels of activity, they provide a good starting point for further evolution when a suitable activity cannot be found in nature. Hence, in the future, artificial enzymes will expand the range of reactions addressed by biocatalysis.

One of the main goals for the future in the Seelig lab is to perform other selections to further demonstrate the capabilities of the method, and also to evolve ligase 10C for increased activity to demonstrate the catalytic potential of artificial enzymes. Projects to

achieve these goals are now underway. This will encourage a more extensive use of mRNA display for enzyme selection in the scientific community, and eventually lead, in combination with other technological innovations such as NGS and microfluidics, to a platform for the development of artificial catalysts with custom properties for both industrial and research purposes.

References

1. Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, et al. (2001) Industrial biocatalysis today and tomorrow. Nature 409: 258-268.
2. Rudat J, Brucher BR, Syldatk C (2012) Transaminases for the synthesis of enantiopure beta-amino acids. AMB Express 2: 11.
3. Sam M HB, Saravanan PN, Taeowan C, Hyungdon Y, (2015) Production of chiral β-amino acids using ω-transaminase from *Burkholderia graminis*. J Biotech 196: 1-8.
4. Yadav VK, Kumar A, Mann A, Aggarwal S, Kumar M, et al. (2014) Engineered reversal of drug resistance in cancer cells-metastases suppressor factors as change agents. Nucleic Acids Res 42: 764-773.
5. http://www.eufic.org/article/en/rid/modern-biotechnology-food-enzymes/.
6. Mitidieri S, Souza Martinelli AH, Schrank A, Vainstein MH (2006) Enzymatic detergent formulation containing amylase from Aspergillus niger: a comparative study with commercial detergent formulations. Bioresour Technol 97: 1217-1224.
7. Noraini MY, Ong HC, Badrul MJ, Chong WT (2014) A review on potential enzymatic reaction for biofuel production from algae. Renew Sust Energy Rev 39: 24-34.
8. Lambertz C, Garvey M, Klinger J, Heesel D, Klose H, et al. (2014) Challenges and advances in the heterologous expression of cellulolytic enzymes: a review. Biotechnol Biofuels 7: 135.
9. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, et al. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491.
10. Peake I (1989) The polymerase chain reaction. J Clin Pathol 42: 673-676.
11. Cohen SN, Chang AC, Boyer HW, Helling RB (1973) Construction of biologically functional bacterial plasmids *in vitro*. Proc Natl Acad Sci U S A 70: 3240-3244.
12. Tsiatsiani L, Heck AJ (2015) Proteomics beyond trypsin. FEBS J.
13. Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques 52: 87-94.
14. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, et al. (2012) Optimal enzymes for amplifying sequencing libraries. Nature Methods 9: 10-11.
15. van Erp PB, Bloomer G, Wilkinson R, Wiedenheft B (2015) The history and market impact of CRISPR RNA-guided nucleases. Curr Opin Virol 12: 85-90.
16. Reddy P, Ocampo A, Suzuki K, Luo J, Bacman SR, et al. (2015) Selective elimination of mitochondrial mutations in the germline by genome editing. Cell 161: 459-469.
17. Schmid RD (2003) Pocket guide to biotechnology and genetic engineering: Wiley-VCH.
18. http://penelope.uchicago.edu/~grout/encyclopaedia_romana/wine/wine.html.
19. Bornscheuer UT, Buchholz K (2005) Highlights in biocatalysis - Historical landmarks and current trends. Engi Life Sci 5: 309-323.

20. Reetz MT (2013) Biocatalysis in organic chemistry and biotechnology: past, present, and future. J Am Chem Soc 135: 12480-12496.
21. Hutchison CA, 3rd, Phillips S, Edgell MH, Gillam S, Jahnke P, et al. (1978) Mutagenesis at a specific position in a DNA sequence. J Biol Chem 253: 6551-6560.
22. Estell DA, Graycar TP, Wells JA (1985) Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. J Biol Chem 260: 6518-6521.
23. Russell AJ, Fersht AR (1987) Rational modification of enzyme catalysis by engineering surface charge. Nature 328: 496-500.
24. Matsumura M, Aiba S (1985) Screening for thermostable mutant of kanamycin nucleotidyltransferase by the use of a transformation system for a thermophile, Bacillus stearothermophilus. J Biol Chem 260: 15298-15303.
25. Bryan PN, Rollence ML, Pantoliano MW, Wood J, Finzel BC, et al. (1986) Proteases of enhanced stability: characterization of a thermostable variant of subtilisin. Proteins 1: 326-334.
26. Liao H, McKenzie T, Hageman R (1986) Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. Proc Natl Acad Sci U S A 83: 576-580.
27. Nannemann DP, Birmingham WR, Scism RA, Bachmann BO (2011) Assessing directed evolution methods for the generation of biosynthetic enzymes with potential in drug biosynthesis. Future Medicinal Chemistry 3: 803-819.
28. Labrou NE (2010) Random mutagenesis methods for *in vitro* directed enzyme evolution. Curr Protein Pept Sci 11: 91-100.
29. Arnold FH (1993) Engineering proteins for nonnatural environments. FASEB J 7: 744-749.
30. Reetz MT, Zonta A, Schimossek K, Liebeton K, Jaeger KE (1997) Creation of enantioselective biocatalysts for organic chemistry by *in vitro* evolution. Angew Chem Int Ed 36: 2830-2832.
31. Cadwell RC, Joyce GF (1992) Randomization of genes by PCR mutagenesis. PCR Methods Appl 2: 28-33.
32. Chen K, Arnold FH (1993) Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc Natl Acad Sci U S A 90: 5618-5622.
33. Turner NJ (2009) Directed evolution drives the next generation of biocatalysts. Nat Chem Biol 5: 568-574.
34. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, et al. (2012) Engineering the third wave of biocatalysis. Nature 485: 185-194.
35. Golynskiy MV, Seelig B (2010) *De novo* enzymes: from computational design to mRNA display. Trends Biotechnol 28: 340-345.
36. Cohen N, Abramov S, Dror Y, Freeman A (2001) *In vitro* enzyme evolution: the screening challenge of isolating the one in a million. Trends Biotechnol 19: 507-510.
37. Cho G, Keefe AD, Liu RH, Wilson DS, Szostak JW (2000) Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. J Mol Biol 297: 309-319.

38. Kehoe JW, Kay BK (2005) Filamentous phage display in the new millennium. Chem Rev 105: 4056-4072.
39. Sidhu SS, Lowman HB, Cunningham BC, Wells JA (2000) Phage display for selection of novel binding peptides. Methods Enzymol 328: 333-363.
40. Renesto P, Raoult D (2003) From genes to proteins - *in vitro* expression of rickettsial proteins. Ann NY Acad Sci 990: 642-652.
41. Bulter T, Alcalde M, Sieber V, Meinhold P, Schlachtbauer C, et al. (2003) Functional expression of a fungal laccase in Saccharomyces cerevisiae by directed evolution. Appl Environ Microbiol 69: 987-995.
42. Chusacultanachai S, Yuthavong Y (1994) Random mutagenesis strategies for construction of large and diverse clone libraries of mutated DNA fragments. Methods Mol Biol 270: 319-333.
43. Reetz MT, Kahakeaw D, Lohmer R (2008) Addressing the numbers problem in directed evolution. Chembiochem 9: 1797-1804.
44. Virnekas B, Ge LM, Pluckthun A, Schneider KC, Wellnhofer G, et al. (1994) Trinucleotide phosphoramidites - ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res 22: 5600-5607.
45. Janczyk M, Appel B, Springstubbe D, Fritz HJ, Muller S (2012) A new and convenient approach for the preparation of beta-cyanoethyl protected trinucleotide phosphoramidites. Org Biomol Chem 10: 1510-1513.
46. Bieberich E, Kapitonov D, Tencomnao T, Yu RK (2000) Protein-ribosome-mRNA display: affinity isolation of enzyme-ribosome-mRNA complexes and cDNA cloning in a single-tube reaction. Anal Biochem 287: 294-298.
47. Amstutz P, Pelletier JN, Guggisberg A, Jermutus L, Cesaro-Tadic S, et al. (2002) *In vitro* selection for catalytic activity with ribosome display. J Am Chem Soc. pp. 9396-9403.
48. Takahashi F, Ebihara T, Mie M, Yanagida Y, Endo Y, et al. (2002) Ribosome display for selection of active dihydrofolate reductase mutants using immobilized methotrexate on agarose beads. FEBS Lett 514: 106-110.
49. Takahashi F, Funabashi H, Mie M, Endo Y, Sawasaki T, et al. (2005) Activity-based *in vitro* selection of T4 DNA ligase. Biochem Biophys Res Commun 336: 987-993.
50. Quinn DJ, Cunningham S, Walker B, Scott CJ (2008) Activity-based selection of a proteolytic species using ribosome display. Biochem Biophys Res Commun 370: 77-81.
51. Seelig B, Szostak JW (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. Nature 448: 828-831.
52. Seelig B (2011) mRNA display for the selection and evolution of enzymes from *in vitro*-translated protein libraries. Nat Protoc 6: 540-552.
53. Odegrip R, Coomber D, Eldridge B, Hederer R, Kuhlman PA, et al. (2004) CIS display: *in vitro* selection of peptides from libraries of protein-DNA complexes. Proc Natl Acad Sci USA 101: 2806-2810.

54. Reiersen H, Lobersli I, Loset GA, Hvattum E, Simonsen B, et al. (2005) Covalent antibody display - an *in vitro* antibody-DNA library selection system. Nucleic Acids Res 33: e10.
55. Cohen HM, Tawfik DS, Griffiths AD (2004) Altering the sequence specificity of *HaeIII* methyltransferase by directed evolution using *in vitro* compartmentalization. Protein Eng Des Sel 17: 3-11.
56. Doi N, Kumadaki S, Oishi Y, Matsumura N, Yanagawa H (2004) *In vitro* selection of restriction endonucleases by *in vitro* compartmentalization. Nucleic Acids Res 32: e95.
57. Fallah-Araghi A, Baret JC, Ryckelynck M, Griffiths AD (2012) A completely *in vitro* ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. Lab Chip 12: 882-891.
58. Mastrobattista E, Taly V, Chanudet E, Treacy P, Kelly BT, et al. (2005) High-throughput screening of enzyme libraries: *in vitro* evolution of a beta-galactosidase by fluorescence-activated sorting of double emulsions. Chem Biol 12: 1291-1300.
59. Griffiths AD, Tawfik DS (2003) Directed evolution of an extremely fast phosphotriesterase by *in vitro* compartmentalization. EMBO J 22: 24-35.
60. Stapleton JA, Swartz JR (2010) Development of an *in vitro* compartmentalization screen for high-throughput directed evolution of [FeFe] hydrogenases. PLoS One 5: 1-8.
61. Kelly BT, Griffiths AD (2007) Selective gene amplification. Protein Eng Des Sel 20: 577-581.
62. Sumida T, Doi N, Yanagawa H (2009) Bicistronic DNA display for *in vitro* selection of Fab fragments. Nucleic Acids Res 37: e147.
63. Ahn JH, Kang TJ, Kim DM (2008) Tuning the expression level of recombinant proteins by modulating mRNA stability in a cell-free protein synthesis system. Biotechnol Bioeng 101: 422-427.
64. Schechter I (1973) Biologically and chemically pure mRNA coding for a mouse immunoglobulin L-chain prepared with the aid of antibodies and immobilized oligothymidine. Proc Natl Acad Sci USA 70: 2256-2260.
65. Hanes J, Plückthun A (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. Proc Natl Acad Sci USA 94: 4937-4942.
66. Mattheakis LC, Bhatt RR, Dower WJ (1994) An *in-vitro* polysome display system for identifying ligands from very large peptide libraries. Proc Natl Acad Sci USA 91: 9022-9026.
67. Lipovsek D, Pluckthun A (2004) *In-vitro* protein evolution by ribosome display and mRNA display. J Immunol Methods 290: 51-67.
68. Jestin JL, Kaminski PA (2004) Directed enzyme evolution and selections for catalysis based on product formation. J Biotechnol 113: 85-103.
69. Roberts RW, Szostak JW (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. Proc Natl Acad Sci USA 94: 12297-12302.

70. Nemoto N, Miyamoto-Sato E, Husimi Y, Yanagawa H (1997) *In vitro* virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome *in vitro*. FEBS Lett 414: 405-408.
71. Liu RH, Barrick JE, Szostak JW, Roberts RW (2000) Optimized synthesis of RNA-protein fusions for *in vitro* protein selection. Methods Enzymol 318: 268-293.
72. Kurz M, Gu K, Lohse PA (2000) Psoralen photo-crosslinked mRNA–puromycin conjugates: a novel template for the rapid and facile preparation of mRNA–protein fusions Nucleic Acids Res 28: e83.
73. Cotten SW, Zou JW, Valencia CA, Liu RH (2011) Selection of proteins with desired properties from natural proteome libraries using mRNA display. Nat Protoc 6: 1163-1182.
74. Takahashi TT, Roberts RW (2009) *In vitro* selection of protein and peptide libraries using mRNA display. Methods Mol Biol 535: 293-314.
75. Ueno S, Nemoto N (2012) cDNA display: rapid stabilization of mRNA display. Methods Mol Biol 805: 113-135.
76. Kurz M, Gu K, Al-Gawari A, Lohse PA (2001) cDNA - Protein fusions: covalent protein-gene conjugates for the *in vitro* selection of peptides and proteins. Chembiochem 2: 666-672.
77. Tawfik DS, Griffiths AD (1998) Man-made cell-like compartments for molecular evolution. Nat Biotechol 16: 652-656.
78. Miller OJ, Bernath K, Agresti JJ, Amitai G, Kelly BT, et al. (2006) Directed evolution by *in vitro* compartmentalization. Nat Methods 3: 561-570.
79. Bernath K, Hai MT, Mastrobattista E, Griffiths AD, Magdassi S, et al. (2004) *In vitro* compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. Anal Biochem 325: 151-157.
80. Ghadessy FJ, Holliger P (2004) A novel emulsion mixture for *in vitro* compartmentalization of transcription and translation in the rabbit reticulocyte system. Protein Eng Des Sel 17: 201-204.
81. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. Proc Natl Acad Sci USA 98: 4552-4557.
82. Eisenstein M (2006) Tiny droplets make a big splash. Nat Methods 3: 71.
83. Song H, Ismagilov RF (2003) Millisecond kinetics using nanoliters of reagents. J Am Chem Soc 125: 14613-14619.
84. Mazutis L, Baret JC, Treacy P, Skhiri Y, Araghi AF, et al. (2009) Multi-step microfluidic droplet processing: kinetic analysis of an *in vitro* translated enzyme. Lab Chip 9: 2902-2908.
85. Doi N, Yanagawa H (1999) STABLE: protein-DNA fusion system for screening of combinatorial protein libraries *in vitro*. FEBS Lett 457: 227-230.
86. Bertschinger J, Neri D (2004) Covalent DNA display as a novel tool for directed evolution of proteins *in vitro*. Protein Eng Des Sel 17: 699-707.
87. Bertschinger J, Grabulovski D, Neri D (2007) Selection of single domain binding proteins by covalent DNA display. Protein Eng Des Sel 20: 57-68.
88. Stein V, Sielaff I, Johnsson K, Hollfelder F (2007) A covalent chemical genotype-phenotype linkage for *in vitro* protein evolution. Chembiochem 8: 2191-2194.

89. Kaltenbach M, Stein V, Hollfelder F (2011) SNAP dendrimers: multivalent protein display on dendrimer-like DNA for directed evolution. Chembiochem 12: 2208-2216.
90. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Natl Acad Sci U S A 98: 14274-14279.
91. Kaplan J, DeGrado WF (2004) *De novo* design of catalytic proteins. Proc Natl Acad Sci U S A 101: 11566-11570.
92. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, et al. (2008) *De novo* computational design of retro-aldol enzymes. Science 319: 1387-1391.
93. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190-195.
94. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, et al. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329: 309-313.
95. Althoff EA, Wang L, Jiang L, Giger L, Lassila JK, et al. (2012) Robust design and optimization of retroaldol enzymes. Protein Sci 21: 717-726.
96. Khare SD, Kipnis Y, Greisen P, Jr., Takeuchi R, Ashani Y, et al. (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat Chem Biol 8: 294-300.
97. Richter F, Blomberg R, Khare SD, Kiss G, Kuzin AP, et al. (2012) Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. J Am Chem Soc 134: 16197-16206.
98. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D (2011) *De novo* enzyme design using Rosetta3. Plos One 6: e19230.
99. Kries H, Blomberg R, Hilvert D (2013) *De novo* enzymes by computational design. Curr Opin Chem Biol 17: 221-228.
100. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, et al. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotechnol 30: 190-192.
101. Haugner JC, Seelig B (2013) Universal labeling of 5 '-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity. Chem Comm 49: 7322-7324.
102. Minervini G, Evangelista G, Villanova L, Slanzi D, De Lucrezia D, et al. (2009) Massive non-natural proteins structure prediction using grid technologies. Bmc Bioinformatics 10 Suppl 6: S22.
103. Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, et al. (2006) Investigation of *de novo* totally random biosequences Part II On the folding frequency in a totally random library of *de novo* proteins obtained by phage display. Chemi Biodivers 3: 840-859.
104. Watters AL, Baker D (2004) Searching for folded proteins *in vitro* and in silico. Eur J Biochem 271: 1615-1622.
105. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410: 715-718.

106. Lo Surdo P, Walsh MA, Sollazzo M (2004) A novel ADP- and zinc-binding fold from function-directed *in vitro* evolution. Nat Struct Mol Biol 11: 382-383.
107. Urvoas A, Valerio-Lepiniec M, Minard P (2012) Artificial proteins from combinatorial approaches. Trends Biotechnol 30: 512-520.
108. Yamauchi A, Nakashima T, Tokuriki N, Hosokawa M, Nogami H, et al. (2002) Evolvability of random polypeptides through functional selection within a small library. Protein Eng 15: 619-626.
109. Yamauchi A, Yomo T, Tanaka F, Prijambada ID, Ohhashi S, et al. (1998) Characterization of soluble artificial proteins with random sequences. FEBS Lett 421: 147-151.
110. Wei Y, Hecht MH (2004) Enzyme-like proteins from an unselected library of designed amino acid sequences. Protein Eng Des Sel 17: 67-75.
111. Oldfield CJ, Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Ann Rev Biochem 83: 553-584.
112. Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 19: 596-604.
113. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci USA 103: 5869-5874.
114. Bommarius AS, Paye MF (2013) Stabilizing biocatalysts. Chem Soc Rev 42: 6534-6565.
115. Wijma HJ, Floor RJ, Janssen DB (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. Curr Opin Struct Biol 23: 588-594
116. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10: 866-876.
117. Eijsink VGH, Gåseidnes S, Borchert TV, van den Burg B (2005) Directed evolution of enzyme stability. Biomol Eng 22: 21-30.
118. Lane MD, Seelig B (2014) Directed evolution of novel proteins. Curr Opin Chem Biol 22: 129-126.
119. Wigley WC, Stidham RD, Smith NM, Hunt JF, Thomas PJ (2001) Protein solubility and folding monitored *in vivo* by structural complementation of a genetic marker protein. Nat Biotechnol 19: 131-136.
120. Waldo GS, Standish BM, Berendzen J, Terwilliger TC (1999) Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol 17: 691-695.
121. Sieber V, Pluckthun A, Schmid FX (1998) Selecting proteins with improved stability by a phage-based method. Nat Biotechnol 16: 955-960.
122. Martin A, Sieber V, Schmid FX (2001) *In vitro* selection of highly stabilized protein variants with optimized surface. J Mol Biol 309: 717-726.
123. Socha RD, Tokuriki N (2013) Modulating protein stability – directed evolution strategies for improved protein function. FEBS J 280: 5582-5595.
124. Golynskiy MV, Haugner III JC, Morelli A, Morrone D, Seelig B (2013) *In vitro* evolution of enzymes. Methods Mol Biol 978: 73-92.
125. Schmid FX (2011) Lessons about protein stability from *in vitro* selections. Chembiochem 12: 1501-1507.

126. Jäckel C, Bloom JD, Kast P, Arnold FH, Hilvert D (2010) Consensus protein design without phylogenetic bias. J Mol Biol 399: 541-546.
127. Chao F-A, Morelli A, Haugner JC, III, Churchfield L, Hagmann LN, et al. (2013) Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution. Nat Chem Biol 9: 81-83.
128. Haugner III JC, Seelig B (2013) Universal labeling of 5'-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity. Chem Commun 49: 7322-7324.
129. Cho GS, Szostak JW (2006) Directed evolution of ATP binding proteins from a zinc finger domain by using mRNA display. Chem Biol 13: 139-147.
130. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95-98.
131. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
132. Greenfield NJ (2006) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nat Protoc 1: 2527-2535.
133. Diaz JE, Lin C-S, Kunishiro K, Feld BK, Avrantinis SK, et al. (2011) Computational design and selections for an engineered, thermostable terpene synthase. Protein Sci 20: 1597-1606.
134. Reetz MT, Soni P, Acevedo JP, Sanchis J (2009) Creation of an amino acid network of structurally coupled residues in the directed evolution of a thermostable enzyme. Angew Chem Int Ed Engl 48: 8268-8272.
135. Palackal N, Brennan Y, Callen WN, Dupree P, Frey G, et al. (2004) An evolutionary route to xylanase process fitness. Protein Sci 13: 494-503.
136. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964-972.
137. Nashine VC, Hammes-Schiffer S, Benkovic SJ (2010) Coupled motions in enzyme catalysis. Curr Opin Chem Biol 14: 644-651.
138. Ramanathan A, Agarwal PK (2011) Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol 9: e1001193.
139. Auerbach G, Ostendorp R, Prade L, Korndorfer I, Dams T, et al. (1998) Lactate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*: the crystal structure at 2.1 A resolution reveals strategies for intrinsic protein stabilization. Structure 6: 769-781.
140. Russell RJ, Gerike U, Danson MJ, Hough DW, Taylor GL (1998) Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. Structure 6: 351-361.
141. Arnold FH, Wintrode PL, Miyazaki K, Gershenson A (2001) How enzymes adapt: lessons from directed evolution. Trends Biochem Sci 26: 100-106.

142. Macedo-Ribeiro S, Darimont B, Sterner R, Huber R (1996) Small structural changes account for the high thermostability of 1[4Fe-4S] ferredoxin from the hyperthermophilic bacterium *Thermotoga maritima*. Structure 4: 1291-1301.
143. Elias M, Wieczorek G, Rosenne S, Tawfik DS (2014) The universality of enzymatic rate-temperature dependency. Trends Biochem Sci 39: 1-7.
144. Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfik DS (2009) Do viral proteins possess unique biophysical features? Trends Biochem Sci 34: 53-59.
145. Chaput JC, Szostak JW (2004) Evolutionary optimization of a nonbiological ATP binding protein for improved folding stability. Chem Biol 11: 865-874.
146. Smith MD, Rosenow MA, Wang MT, Allen JP, Szostak JW, et al. (2007) Structural insights into the evolution of a non-biological protein: importance of surface residues in protein fold optimization. PLoS ONE 2: e467.
147. Moore MJ, Sharp PA (1992) Site-specific modification of pre-mRNA: the 2'-hydroxyl groups at the splice sites. Science 256: 992-997.
148. Chothia C (1992) Proteins - 1000 families for the molecular biologist. Nature 357: 543-544.
149. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP - a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
150. Ohno S (1971) Evolution by gene duplication: Springer-Verlag, New York, USA.
151. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300: 1701-1703.
152. James LC, Tawfik DS (2003) Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. Trends Biochem Sci 28: 361-368.
153. Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. Science 324: 203-207.
154. Bryan PN, Orban J (2010) Proteins that switch folds. Curr Opin Struct Biol 20: 482-488.
155. Cordes MHJ, Walsh NP, McKnight CJ, Sauer RT (1999) Evolution of a protein fold *in vitro*. Science 284: 325-327.
156. Mansy SS, Zhang JL, Kummerle R, Nilsson M, Chou JJ, et al. (2007) Structure and evolutionary analysis of a non-biological ATP-binding protein. J Mol Biol 371: 501-513.
157. Smith BA, Hecht MH (2011) Novel proteins: from fold to function. Curr Opin Chem Biol 15: 421-426.
158. Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, et al. (2008) Interconversion between two unrelated protein folds in the lymphotactin native state. Proc Natl Acad Sci USA 105: 5057-5062.
159. Holmbeck SMA, Foster MP, Casimiro DR, Sem DS, Dyson HJ, et al. (1998) High-resolution solution structure of the retinoid X receptor DNA-binding domain. J Mol Biol 281: 271-284.
160. Zhao Q, Chasse SA, Devarakonda S, Sierk ML, Ahvazi B, et al. (2000) Structural basis of RXR-DNA interactions. J Mol Biol 296: 509-520.

161. Maret W, Li Y (2009) Coordination dynamics of zinc in proteins. Chem Rev 109: 4682-4707.
162. van Tilborg PJ, Czisch M, Mulder FA, Folkers GE, Bonvin AM, et al. (2000) Changes in dynamical behavior of the retinoid X receptor DNA-binding domain upon binding to a 14 base-pair DNA half site. Biochemistry 39: 8747-8757.
163. Yang W, Lee JY, Nowotny M (2006) Making and breaking nucleic acids: two-Mg^{2+}-ion catalysis and substrate specificity. Mol Cell 22: 5-13.
164. Bhabha G, Lee J, Ekiert DC, Gam J, Wilson IA, et al. (2011) A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. Science 332: 234-238.
165. Baldwin AJ, Kay LE (2009) NMR spectroscopy brings invisible protein states into focus. Nat Chem Biol 5: 808-814.
166. Golynskiy MV, Seelig B (2010) *De novo* enzymes: from computational design to mRNA display. Trends Biotechnol 28: 340-345.
167. Grzesiek S, Bax A (1992) Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. J Magn Reson 96: 432-440.
168. Muhandiram DR, Kay LE (1994) Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity. J Magn Reson, Ser B 103: 203-216.
169. Wittekind M, Mueller L (1993) HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha- and beta-carbon resonances in proteins. J Magn Reson, Ser B 101: 201-205.
170. Eghbalnia HR, Bahrami A, Tonelli M, Hallenga K, Markley JL (2005) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J Am Chem Soc 127: 12528-12536.
171. Grzesiek S, Anglister J, Bax A (1993) Correlation of backbone amide and aliphatic side-chain resonances in 13C/15N-enriched proteins by isotropic mixing of 13C magnetization. J Magn Reson, Ser B 101: 114-119.
172. Wuthrich K (1986) NMR of proteins and nucleic acids. New York, USA: John Wiley and Sons.
173. Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 222: 311-333.
174. Vuister GW, Bax A (1993) Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins. J Am Chem Soc 115: 7772-7777.
175. Lee D, Hilty C, Wider G, Wuthrich K (2006) Effective rotational correlation times of proteins from NMR relaxation interference. J Magn Reson 178: 72-76.
176. Gagne SM, Tsuda S, Li MX, Chandra M, Smillie LB, et al. (1994) Quantification of the calcium-induced secondary structural changes in the regulatory domain of troponin-C. Protein Sci 3: 1961-1974.
177. Wang Y, Zhao S, Somerville RL, Jardetzky O (2001) Solution structure of the DNA-binding domain of the TyrR protein of Haemophilus influenzae. Protein Sci 10: 592-598.

178. Ruckert M, Otting G (2000) Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. J Am Chem Soc 122: 7793-7797.
179. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160: 65-73.
180. Alberts IL, Nadassy K, Wodak SJ (1998) Analysis of zinc binding sites in protein crystal structures. Protein Sci 7: 1700-1716.
181. Viles JH, Patel SU, Mitchell JB, Moody CM, Justice DE, et al. (1998) Design, synthesis and structure of a zinc finger with an artificial beta-turn. J Mol Biol 279: 973-986.
182. Ohlenschlager O, Seiboth T, Zengerling H, Briese L, Marchanka A, et al. (2006) Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. Oncogene 25: 5953-5959.
183. Banci L, Bertini I, Del Conte R, Mangani S, Meyer-Klaucke W (2003) X-Ray absorption and NMR spectroscopic studies of CopZ, a copper chaperone in Bacillus subtilis: the coordination properties of the copper ion. Biochemistry 42: 2467-2474.
184. Tenderholt A (2007) Pyspline. Stanford, USA: Stanford University.
185. Deleon JM, Rehr JJ, Zabinsky SI, Albers RC (1991) Ab initio curved-wave X-ray-absorption-fine-structure. Phys Rev B 44: 4146-4156.
186. Rehr JJ, Albers RC (2000) Theoretical approaches to x-ray absorption fine structure. Rev Mod Phys 72: 621-654.
187. Rehr JJ, Deleon JM, Zabinsky SI, Albers RC (1991) Theoretical X-ray-absorption-fine-structure Standards. J Am Chem Soc 113: 5135-5140.
188. Kim CA, Berg JM (1996) A 2.2 angstrom resolution crystal structure of a designed zinc finger protein bound to DNA. Nat Struct Biol 3: 940-945.
189. George GN (2000) EXAFSSPAK and EDG-FIT. Menlo Park, USA: Stanford Synchrotron Radiation Lightsource.
190. Patel K, Kumar A, Durani S (2007) Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. BBA-Proteins Proteom 1774: 1247-1253.
191. Kupper H, Mijovilovich A, Meyer-Klaucke W, Kroneck PMH (2004) Tissue- and age-dependent differences in the complexation of cadmium and zinc in the cadmium/zinc hyperaccumulator Thlaspi caerulescens (Ganges ecotype) revealed by X-ray absorption spectroscopy. Plant Physiol 134: 748-757.
192. Clark-Baldwin K, Tierney DL, Govindaswamy N, Gruff ES, Kim C, et al. (1998) The limitations of X-ray absorption spectroscopy for determining the structure of zinc sites in proteins. When is a tetrathiolate not a tetrathiolate? J Am Chem Soc 120: 8401-8409.
193. Penner-Hahn JE (2005) Characterization of "spectroscopically quiet" metals in biology. Coord Chem Rev 249: 161-177.
194. Bobyr E, Lassila JK, Wiersma-Koch HI, Fenn TD, Lee JJ, et al. (2012) High-Resolution Analysis of Zn2+ Coordination in the Alkaline Phosphatase Superfamily by EXAFS and X-ray Crystallography. J Mol Biol 415: 102-117.

195. Kim R, Guo JT (2010) Systematic analysis of short internal indels and their impact on protein folding. BMC Struct Biol 10: 24.
196. Hsing M, Cherkasov A (2008) Indel PDB: A database of structural insertions and deletions derived from sequence alignments of closely related proteins. Bmc Bioinformatics 9: 293.
197. Chan SK, Hsing M, Hormozdiari F, Cherkasov A (2007) Relationship between insertion/deletion (indel) frequency of proteins and essentiality. Bmc Bioinformatics 8: 227.
198. Hida K, Won SY, Di Pasquale G, Hanes J, Chiorini JA, et al. (2010) Sites in the AAV5 capsid tolerant to deletions and tandem duplications. Arch Biochem Biophys 496: 1-8.
199. Hecky J, Muller KM (2005) Structural perturbation and compensation by directed evolution at physiological temperature leads to thermostabilization of beta-lactamase. Biochemistry 44: 12640-12654.
200. Chapple KE, Bartel DP, Unrau PJ (2003) Combinatorial minimization and secondary structure determination of a nucleotide synthase ribozyme. RNA 9: 1208-1220.
201. Agresti JJ, Antipov E, Abate AR, Ahn K, Rowat AC, et al. (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution (vol 170, pg 4004, 2010). Proc Natl Acad Sci USA 107: 6550-6550.
202. Stemmer WPC (1994) DNA shuffling by random fragmentation and reassembly - *in vitro* recombination for molecular evolution. Proc Natl Acad Sci USA 91: 10747-10751.
203. Li XY, Song BA, Hu DY, Wang ZC, Zeng MJ, et al. (2012) The development and application of new crystallization method for tobacco mosaic virus coat protein. Virol J 9: 279.
204. Schwartz TU, Walczak R, Blobel G (2004) Circular permutation as a tool to reduce surface entropy triggers crystallization of the signal recognition particle receptor beta subunit. Protein Sci 13: 2814-2818.
205. Pisarchik A, Petri R, Schmidt-Dannert C (2007) Probing the structural plasticity of an archaeal primordial cobaltochelatase CbiX(S). Protein Eng Des Sel 20: 257-265.
206. Ostermeier M, Shim JH, Benkovic SJ (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. Nat Biotechnol 17: 1205-1209.
207. Sieber V, Martinez CA, Arnold FH (2001) Libraries of hybrid proteins from distantly related sequences. Nat Biotechnol 19: 456-460.
208. Haapa S, Taira S, Heikkinen E, Savilahti H (1999) An efficient and accurate integration of mini-Mu transposons *in vitro*: a general methodology for functional genetic analysis and molecular biology applications. Nucleic Acids Res 27: 2777-2784.
209. Haapa-Paananen S, Rita H, Savilahti H (2002) DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection *in vitro*. J Biol Chem 277: 2843-2851.

210. E. Poussu JJ, H. Savilahti (2005) A gene truncation strategy generating N- and C-terminal deletion variants of proteins for functional studies: mapping of the Sec1p binding domain in yeast Mso1p by a Mu *in vitro* transposition-based approach. Nucleic Acids Res 33.
211. Butterfield YSN, Marra MA, Asano JK, Chan SY, Guin R, et al. (2002) An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. Nucleic Acids Res 30: 2460-2468.
212. Jones DD (2005) Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion. Nucleic Acids Res 33: e80.
213. Morelli A, Haugner J, Seelig B (2014) Thermostable artificial enzyme isolated by *in vitro* selection. Plos One 9: 112028.
214. Fernandez-Martinez J, Phillips J, Sekedat MD, Diaz-Avalos R, Velazquez-Muriel J, et al. (2012) Structure-function mapping of a heptameric module in the nuclear pore complex. Journal of Cell Biology 196: 419-434.
215. Schutze T, Rubelt F, Repkow J, Greiner N, Erdmann VA, et al. (2011) A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. Analytical Biochemistry 410: 155-157.
216. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, et al. (2006) Amplification of complex gene libraries by emulsion PCR. Nature Methods 3: 545-550.
217. Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH (2011) *De novo* designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth. Plos One 6: e15364.
218. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.
219. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.
220. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.
221. Pearson WR (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 11: 635-650.
222. Tokuriki N, Sakamoto K, Waluyo D, Makino Y, Ogasahara K, et al. (2001) Effects of amino acid substitution on the physicochemical properties of artificial proteins with random sequences. J Biosci Bioeng 92: 167-172.
223. Prymula K, Piwowar M, Kochanczyk M, Flis L, Malawski M, et al. (2009) In silico structural study of random amino acid sequence proteins not present in nature. Chem Biodivers 6: 2311-2336.
224. White SH, Jacobs RE (1990) Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. Biophys J 57: 911-921.

225. White SH, Jacobs RE (1993) The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. J Mol Evol 36: 79-95.
226. White SH (1994) The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences. J Mol Evol 38: 383-394.
227. Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, et al. (1996) Solubility of artificial proteins with random sequences. FEBS Lett 382: 21-25.
228. Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, et al. (2006) Investigation of *de novo* totally random biosequences, Part II: On the folding frequency in a totally random library of *de novo* proteins obtained by phage display. Chem Biodivers 3: 840-859.
229. Chessari S, Thomas R, Polticelli F, Luisi PL (2006) The production of *de novo* folded proteins by a stepwise chain elongation: a model for prebiotic chemical evolution of macromolecular sequences. Chem Biodivers 3: 1202-1210.
230. Hayashi Y, Sakata H, Makino Y, Urabe I, Yomo T (2003) Can an arbitrary sequence evolve towards acquiring a biological function? J Mol Evol 56: 162-168.
231. Matsuura T, Miyai K, Trakulnaleamsai S, Yomo T, Shima Y, et al. (1999) Evolutionary molecular engineering by random elongation mutagenesis. Nat Biotechnol 17: 58-61.
232. Ito Y, Kawama T, Urabe I, Yomo T (2004) Evolution of an arbitrary sequence in solubility. J Mol Evol 58: 196-202.
233. Nakashima T, Toyota H, Urabe I, Yomo T (2007) Effective selection system for experimental evolution of random polypeptides towards DNA-binding protein. J Biosci Bioeng 103: 155-160.
234. Toyota H, Hosokawa M, Urabe I, Yomo T (2008) Emergence of polyproline II-like structure at early stages of experimental evolution from random polypeptides. Mol Biol Evol 25: 1113-1119.
235. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, et al. (2013) Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res 41: W597-600.
236. Ameta S, Winz ML, Previti C, Jaschke A (2014) Next-generation sequencing reveals how RNA catalysts evolve from random space. Nucleic Acids Res 42: 1303-1310.