In vitro evolution of artificial enzymes: method development and applications

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

John Christian Haugner III

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

Burckhard Seelig

September 2014

Acknowledgements

There are many people past and present in my thesis lab I'd like to acknowledge for all their support. First and foremost I thank my advisor, Dr. Burckhard Seelig, for guiding me in my research and challenging me to improve. I thank Dr. Misha Golynskiy for being a wonderful friend and colleague; we spent a few long years together working on those libraries and you always helped lift my spirits when experiments went wrong. I thank Aleardo Morelli who was a true comrade throughout all of our work with mRNA display and ligase 10C. I also thank Dr. Dana Morrone and Michael Lane for all the enlightening discussions we shared about enzyme selections and their input in many of my publications and presentations.

Outside my thesis lab, I thank Fa-An (Frank) Chao and the rest of the Veglia lab for their excellent work in solving the structure of ligase 10C. I thank our many collaborators described in chapter 6 who have helped us develop our application for ligase 10C.

I'd like to acknowledge the NIH Biotechnology Training Grant (T32 GM08347) for providing funding and broadening my experiences in Biotechnology. I'd also like to acknowledge the Doctoral Dissertation Fellowship for helping fund my final year of research and giving me the opportunity to share my research to a broader audience. Finally, I'd like to thank all the marvelous people of the Biotechnology Institute here at the University of Minnesota. I thank the administrators, faculty and staff for putting together an incredible program for graduate students. I also thank all of my fellow researchers within Gortner Laboratory for sharing not just materials and equipment, but a real sense of collaboration and comradery.

Dedication

For my wife, Elizabeth, and to my daughter, Evelyn, who have brought endless joy into my life.

Abstract

Artificial enzymes have the potential to aid in the production of pharmaceuticals and facilitate basic biomedical research. There are two methods for making artificial enzymes: rational design and *de novo* selection. Rational design utilizes detailed knowledge of enzyme catalysis to design an enzyme active site, and then introduces this active site into a protein. However, due to the limited understanding of protein folding and structure-function relationships this approach is still extremely challenging and far from routine. In contrast, we utilize a directed evolution approach to isolate *de novo* artificial enzymes from a large library of protein variants by *in vitro* selection. Each of the trillions of proteins in a library are tested in a single experiment to determine if any have the desired activity. The artificial enzymes are created when the library is made so a high quality library is important for success. My thesis research focuses on two goals: (1) Construct a library built on the robust $(\beta/\alpha)_8$ barrel enzyme scaffold for future enzyme selections and (2) Characterize a thermostable artificial RNA ligase and develop an application for this enzyme.

The $(\beta/\alpha)_8$ fold is used to catalyze a wide range of chemical reactions in nature. We used this fold to create a library containing $> 10^{14}$ unique proteins by replacing loops of the catalytic face with randomized codons via PCR. Small sub-libraries were subjected to a protease-based folding selection to improve library quality by enriching for folded sequences. The final folding-enriched library contained $> 10^{12}$ folded proteins representing an up to 50-fold improvement relative to a control library. These libraries will provide a valuable source of new enzymes for future *in vitro* selections.

The previously generated artificial RNA ligases join 5'-triphosphate RNA to the 3'-hydroxyl of a second RNA substrate; a reaction not observed in nature. However the enzymes were also highly dynamic, which prevented the solving of the protein structure by NMR or X-ray crystallography. A more structured enzyme, called ligase 10C, was isolated by performing the ligase selection at 65°C and its structure was solved revealing a novel primordial fold. Here, we describe the detailed biochemical characterization of ligase 10C. Using a variety of RNA substrates, we also determined how ligation rates

change with sequence composition revealing an enzyme with broad sequence specificity. We developed a method for the specific ligation and sequencing of 5'-triphosphorylated RNA. These results highlight ligase 10C as an attractive tool for the selective isolation of 5'-triphosphate RNA from a complex mixture, something which is difficult with current methods.

Table of Contents

List of Tables

List of Figures

Chapter 1:

Introduction

Sections 1.3 and 1.4 were adapted from the 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. The article is reproduced here with kind permission from Springer Science and Business Media. The article was written in collaboration with Dr. Misha Golynskiy, Aleardo Morelli and Dr. Dana Morrone with all authors having significant contributions to each section. Dr. Burckhard Seelig planned, reviewed and edited the manuscript prior to submission.

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

Figure 1.3 was reprinted from Kries, H., Blomberg, R., and Hilvert, D. (2013) *De novo* enzymes by computational design. *Current Opinion in Chemical Biology* **17**, 221-28. with permission from Elsevier. Figures 1.4, 1.5, 1.6 and 1.7 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 contains an overview of the field of artificial enzymes. The chapter begins by demonstrating the significance of artificial enzymes and showing the benefits of new enzymes to numerous fields of biotechnology and the limits of existing methods of enzyme development. A review of emerging methods for *in vitro* enzyme evolution is included to highlight the capabilities of *in vitro* evolution and its advantages over *in vivo* methods. The two general methods for making new artificial enzymes, rational design and *de novo* selection, are then introduced and compared. This leads into a summary of previous work by Dr. Burckhard Seelig describing the use of the *in vitro* method mRNA

display to isolate artificial enzymes *de novo* from a highly diverse library. The chapter ends with an in-depth look at the potential of mRNA display selections, discussing the advantages and possible pitfalls in the method.

Chapter 2 describes the design and synthesis of a new protein library based on the $(\beta/\alpha)_8$ fold. Nearly all $(\beta/\alpha)_8$ proteins found in nature are enzymes covering 5 out of 6 enzyme classes making it an attractive fold for the evolution of new enzymes. The library was constructed by randomizing the amino acids that commonly compose the active site. To preserve the protein fold, sub-libraries were constructed first and enriched for folded proteins through a protease-resistance assay, and then these sub-libraries were ligated together. The final folding-enriched library contained approximately 10^{12} unique and folded proteins.

Chapter 3 details the isolation of a thermostable variant of 5'-triphosphate dependent RNA ligase. Additional rounds of mRNA display selection were performed at 65°C with substrates modified to maintain the RNA-RNA duplex at the high temperature. The best ligase variant identified, called ligase 10C, had a melting temperature of 72°C representing a 24-35°C increase in stability relative to two closely related clones isolated from the original, room temperature selection. In addition, ligase 10C is more active at both room temperature and 65°C than the previous enzymes. While ligase 10C had been isolated in previous work, the detailed characterization was carried out as part of this thesis.

Chapter 4 describes the structure of the artificial RNA ligase 10C generated in chapter 3. Like the parent scaffold, ligase 10C binds two zinc ions which are essential for function. However, 10C adopts a unique fold not previously observed in nature and substantially different from the original protein scaffold. The region corresponding to loop 2 of the original library, which was highly conserved among all active ligases, was identified as a substrate interaction region and putative active site.

Chapter 5 contains the characterization of the sequence specificity of the artificial RNA ligase 10C. While 10C had been evolved to ligate a single RNA sequence, the ligase had comparable levels of activity on all substrates tested which showed the enzyme has broad sequence specificity. To probe the capabilities of the ligase further, the ligase

was challenged with a pool of 3 PPP-RNA substrates to determine if a specific substrate could be ligated using a specific complementary splint. Despite sharing some sequence similarity, all 3 PPP-RNAs could be ligated specifically in the mixture which suggests a potential role for isolating PPP-RNAs.

Chapter 6 demonstrates the application of the artificial RNA ligase for the sequencing of PPP-RNAs from a complex mixture. The ligase was characterized to determine the optimize ligation conditions and to indentify common chemicals that would inhibit ligation. A protocol was then developed that should enable easy sequencing of PPP-RNA utilizing Illumina's TruSeq platform.

Finally, the thesis concludes with a section of conclusions and future directions for mRNA display and artificial enzyme selections.

1.2 Significance

1.2.1 The benefits of new enzymes

Enzymes are nature's catalysts, driving chemical reactions necessary to sustain life. Enzymes are used metabolize various compounds, breaking them down into simple components and harvesting chemical energy. They can also be used to construct new complex molecules to serve roles including structure, signaling and defense. [1] Moreover, they have been shown to exhibit exquisite chemical selectivity and highly efficient catalysis in mild aqueous environments; something that few chemical transformations can boast. [2] For these reasons, enzymes hold the potential to improve many scientific fields by making existing chemical processes more efficient and by enabling new chemical transformations.

In recent years, the pharmaceutical industry has been utilizing enzymes to reduce the cost of drug synthesis by replacing inefficient steps in the chemical synthesis. Compared to organic synthesis of small molecules, enzymes and biocatalysis can offer a route to make enantiopure compounds cheaply and with fewer environmentally hazardous wastes. Enzymes such as ketone reductases and transaminases have been used extensively to create chiral products directly instead of relying on separation of enantiomers by a chiral column. [2, 3] Success stories, such as the improved synthesis of

Sitagliptin by replacing a critical step with an enzymatic catalysis, demonstrate that enzymatic catalysis can be both good science and a good business practice. [4] For these applications, the most valuable enzymes have excellent enantioselectivity, can accept a wide range of substrates and are active in partial organic solvent solutions or two-phase systems to promote substrate and product solubility. [2, 3] The pharmaceutical industry is increasingly expanding the types of enzymes used in drug synthesis as more new enzymes become available, but the field still relies heavily on chemical transformation for most synthetic steps where enzymatic transformation isn't yet feasible.

Enzymes can also be very efficient, cheap and bio-compatible catalysts for the transformation of bulk compounds. A common application is the use of hydrolases to breakdown polysaccharides and proteins into shorter fragments. [5] In the food and fermentation industry, amylases are utilized in the brewing of beer and the baking of leavened bread to promote the growth of yeast. Traditionally the amylases found in the raw flour or grain would be used to hydrolyze starch into simple sugars but the modern industry will frequently supplement with exogenous enzymes to improve the processes. [6] More recently, the rising cost of fuel and diminishing supply of petroleum has promoted new interest for the production of biofuels, plastics and other traditional petrochemicals from cellulose as an alternative and cheap source of carbon. [7, 8] As cellulose is a highly crystalline polysaccharide, many efforts have been made to develop the enzymes needed to hydrolyze it into fermentable sugars. While many such enzymes exist in nature, development is still ongoing to improve the stability and decrease costs of the enzyme production to make the process economically viable. [9, 10]

Finally, enzymes also have important research applications, allowing scientists to specifically modify and synthesize complex biomolecules in ways that would otherwise be impossible with chemical catalysts. Enzymes that act on nucleic acid substrates helped usher in the modern age of molecular biology. Seemingly simple techniques like PCR amplification of DNA became far more efficient with the development of thermostable polymerases as previously fresh enzyme had to be added each cycle. [11] Restriction enzymes and proteases allow researchers to make highly specific cuts within DNA and protein respectively. [12, 13] Enzymes such as peroxidases and luciferase play vital roles

as easy to use reporter systems for blots and enzyme assays. [14, 15] For these types of enzymes, specificity and efficiency are generally the two most important features as off-target reactions would obscure the results. Much of the commercial development in this field focuses on expanding currently known enzyme classes into new applications, such as evolving DNA polymerases to accept fluorescently labeled nucleotides for sequencing [16], but there are many additional areas of science that would benefit from new entirely enzymes.

1.2.2 Limits of established methods for enzyme development

In all these examples, there is a desire for an efficient enzyme catalyst that is well suited for the application. While many enzymes have been developed as commercially available products, this represents only a small fraction of the chemistries available in natural enzymes. [2] Moreover, even if enzymes exist in nature than can catalyze the type of reaction desired, there is no guarantee that any of them will be able to act on the desired substrates. [2, 3] While it may be desirable to have an enzyme to catalyze a specific reaction, obtaining that enzyme might involve a substantial amount of effort and time.

The development of purified enzymes for commercial applications began at the start of the 20th century which utilized essentially unmodified proteins isolated from various plant or microbial sources. [6, 17] Scientists relied primarily on finding the enzymes they needed already in nature both in terms of available chemistry and enzyme efficiency. Many different species and strains would be screened in search for the organism that produced the enzyme in large amounts and preferably with an easy purification method. [17] Enzyme evolution was possible in a limited fashion, but mutations were introduced through a mutagen which often compromised the host organism.

By the end of the 20th century, when the tools necessary to clone and manipulate genes had been developed, scientists were able to intentionally drive the evolution of sub-optimal enzymes to become more efficient for the desired application through a process called directed evolution. [18] In directed evolution, mutations are introduced specifically

into desired protein creating a library of protein variants. The variants are then each tested for activity to determine which mutations are beneficial. Evolution can then be continued by taking the best variants and introducing more mutations in an attempt to find even better versions of the enzyme. [19, 20] Today directed evolution is a well established method utilized extensively in both industry and academia to improve the properties of virtually any protein or enzyme.

Most enzyme evolution is performed either fully or partially *in vivo*: either inside cells or through the use of cells. [20] These *in vivo* methods take advantage of the intact machinery of a living cell to produce large amounts of protein from DNA transformed into the cells. *In vivo* methods are generally favored because they are cheap to use and utilize well established methods for cloning and DNA manipulation. The enzymes can be expressed inside the cytoplasm, within a specific compartment of the cell or even connected to the cell surface depending on the final application. [21, 22] However, there are limitations to these *in vivo* methods which lead scientists to develop *in vitro* platforms for enzyme evolution.

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. [23]

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. [24] 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. [25] By comparison, phage display libraries produce up to 10^{10} unique variants in a single transformation. [26] 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. [27] 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. [28] 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. [29]

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. [30] 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 2.1). 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.

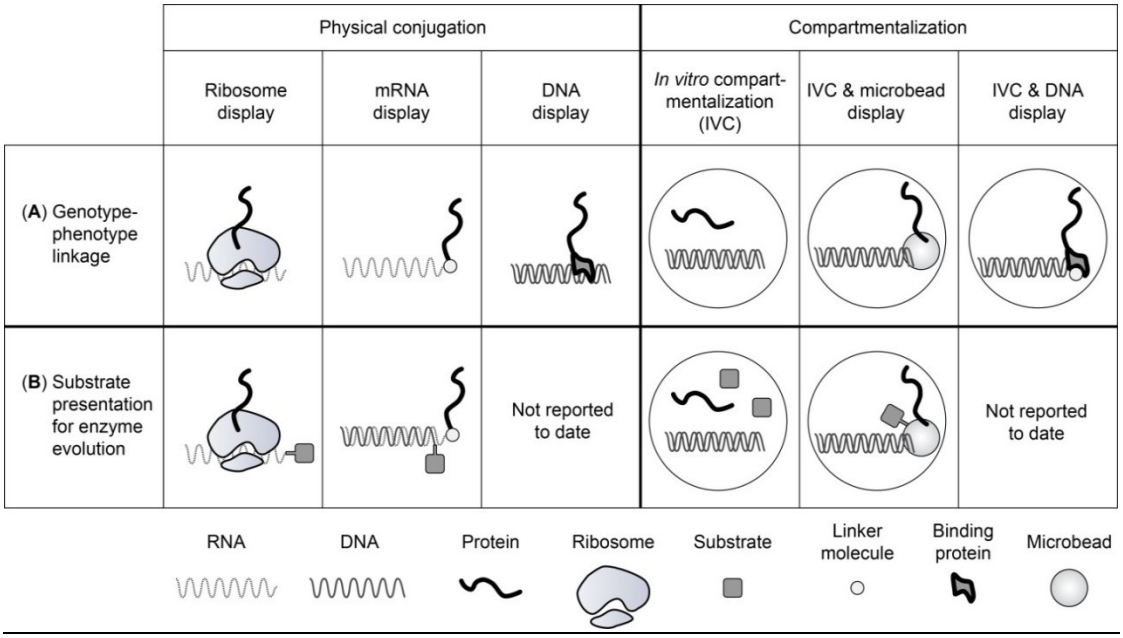


Figure 1.1 - Overview of methods for the *in vitro* selection or screening of proteins discussed in this review. (**A**) The top row shows different strategies to establish the crucial linkage between gene and protein. (**B**) The bottom row illustrates the introduction of substrate into the selection 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 2.1). 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 [32]. 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. [25] More recently, DNA synthesis via phosphoramidite trinucleotides has become commercially available. [31] 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. [32]

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 [33], beta-lactamase [34], dihydrofolate reductase [35], DNA ligase [36] and sortase [37]
mRNA display	Covalent fusions of mRNA-protein via puromycin	~10^{13}	Selection for *de novo* RNA ligase [38, 39]
DNA display	Covalent or non-covalent complex of DNA-protein	~10^{12}	Proof of concept selection of binders, but no enzymes [40, 41]
In vitro compartment-talization (IVC)	Spatial confinement	~10^{9} (selection) ~10^{6}-10^{8} (screening by FACS/microfluidics)	Selection for methyltransferase [42] and restriction nuclease [43]; Proof of concept screening for β–galactosidase [44, 45]
IVC & microbead display	Non-covalent complex of DNA-microbead-protein	~10^{9} (selection) ~10^{6}-10^{8} (screening by FACS/microfluidics)	Screening for phosphotriesterase [46]; Proof of concept screening for hydrogenase [47]; Proof of concept selection of biotin ligase [48]
IVC & DNA display	Covalent or non-covalent complex of DNA-protein	~10^{8}-10^{9}	Selection of antibodies as heterodimers, but no enzymes [49]

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 [50].

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 2.1). 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. [51] Subsequently, this method was developed further to select and evolve peptides and proteins. [52, 53] 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. [53, 54] 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 2.1). [33-35, 37] 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. [20, 55] In one example of a truly product-driven model selection, ribosome display has been employed for isolation of a T4 DNA ligase. [36] 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 2.1). [56, 57] 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. [58, 59] 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. [39, 60, 61] Through slight modifications of the mRNA display protocol, covalent fusions of protein and encoding cDNA can be generated (cDNA display). [62, 63]

There is only one published report of researchers using mRNA display to evolve an enzyme. Unlike all other examples of *in vitro* enzyme evolution, the starting library was based on a non-catalytic scaffold and a new artificial enzyme was selected *de novo.* [38] This evolution of an artificial enzyme is discussed in greater detail later in this chapter as it forms 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. [64] 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. [65] 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. [45, 64, 66] 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. [64, 67] 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. [64, 68]

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. [42] 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. [43] 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. [69] 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. [45] Recently, the same researchers reported a model enrichment of β-galactosidase using a home-made microfluidic system. [44] 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. [44, 70] 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. [44] Furthermore, reagents can be readily added to the droplets after translation, in case the translation conditions are not compatible with enzymatic assay. [71] 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 2.2). 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 2.1). [46, 47] 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. [46] 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. [46]

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. [48] 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. [47] 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 2.2). 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. [40, 41]

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 [46-48]	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 [49, 72]	Non-covalent attachment of protein to DNA, IVC is needed	Biotinylated	Streptavidin
CIS-display [40]	Non-covalent attachment of protein to DNA	RepA gene	DNA replication initiator (RepA)
Covalent DNA display [73, 74]	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 [41]	Covalent attachment of enzyme to DNA	P2A gene	Endonuclease P2A
SNAP display [75, 76]	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 2.2A). 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 2.1 and Figure 2.2B). 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 2.2C 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.

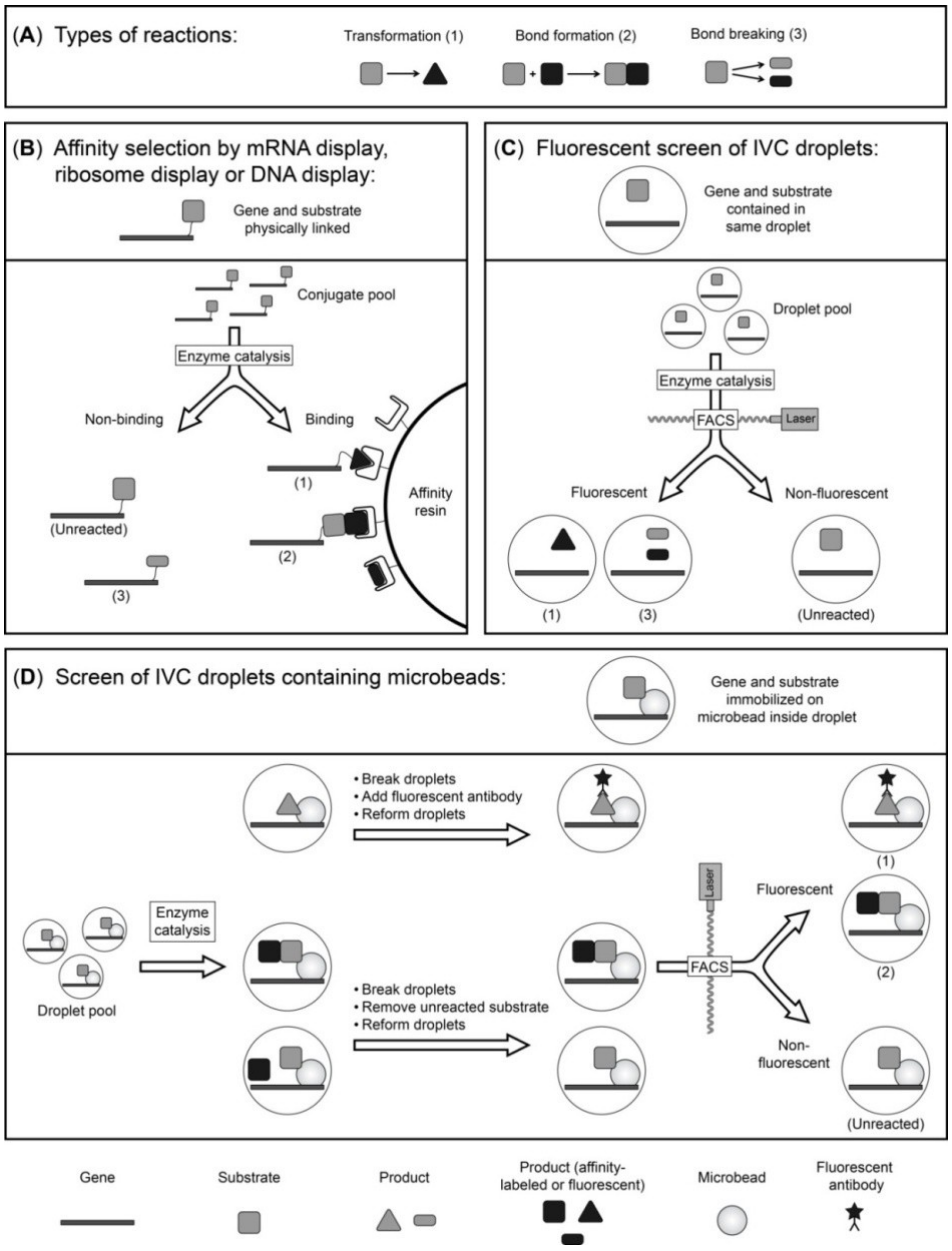


Figure 1.2 - Isolation of enzymatic activities using *in vitro* technologies. (**A**) Types of enzymatic activities that can be evolved using *in vitro* approaches. (**B**) Affinity selection of physically linked gene-substrate/product conjugates. The enzyme itself is also linked to the gene-substrate 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 General considerations for constructing artificial enzymes

A prerequisite for enzyme evolution is to first find an enzyme that can catalyze the desired reaction, even if it does so poorly. Previously, when no enzyme was available for a given reaction, enzymatic catalysis had to be abandoned as creating new enzymes *de novo* was not yet possible. Recently, a new field in biochemistry has emerged to develop the tools necessary to make artificial enzymes to meet this need. [23, 77]

Here, we define artificial enzymes as proteins that scientists have manipulated to introduce a new catalytic activity. This new activity bears no resemblance to the original

activity of the protein and is generated without first creating an intermediate. [23, 77] To date there have only been a few reports of successfully creating artificial enzymes (Table 1.3) as this field is still new. Two general strategies have been shown to create artificial enzymes: rational design and *de novo* selection. While there are significant differences between these methods, they share the same basic 3 steps to make artificial enzymes: identifying a target scaffold, manipulating the protein scaffold, testing and refinement of candidates.

Table 1.3 - Artificial enzymes created by rational design and *de novo* selection

Artificial enzyme	**Parent protein(s)**	**Method utilized**
Esterase [78, 79]	Many [a]	Rational design
Retro-aldolase [80]	Xylanase, IGPS (lyase)	Rational design
Kemp Eliminase [81]	IGPS (lyase), Aldolase, HisF (cyclase)	Rational design
Diels Alderase [82]	Phosphatase, Isomerase	Rational design
Metallo-hydrolase [83]	Many [b]	Rationally designed metal binding site [b]
5'-triphosphate dependent RNA ligase [38]	Transcription factor (non-catalytic)	*De novo* selection

[a] Numerous protein scaffolds were able to be turned into esterases including both enzymes and non-catalytic proteins

[b] While many proteins have been engineered to have metal-dependent hydrolysis activity, the catalysis is typically performed entirely by the bound metal ion with little to no participation from the protein scaffold. These are not typically considered to be "true" artificial enzymes.

1.5.1 Identifying a starting scaffold

Protein scaffolds are native or artificial proteins that will be mutated to generate the new artificial enzyme. Scaffolds often have a well characterized 3 dimensional structure which is crucial to identify where beneficial mutation could be introduced. Enzymes with active sites similar to that of the desired activity are often used as scaffolds to reduce the number of necessary mutations. Thermostable proteins are frequently used because the strong intermolecular contacts are thought to better tolerate destabilizing mutations, although mesophile proteins have been used successfully too. [77] Some researchers chose to use multiple scaffolds in their experiments to make artificial enzymes, which offers greater protein diversity but can be more time consuming and expensive to generate all the desired constructs.

1.5.2 Manipulating the protein scaffold

During scaffold manipulation, the amino acid sequence of the starting protein scaffold is changed to generate many protein variants in an attempt to create the desired artificial enzyme. The amino acid changes tend to be centered in and around the enzyme's active site to promote the interactions necessary for the desired catalysis. While there are many approaches which influence how to generate these variants, a universal feature of both methods is that most variants will not have the desired activity. Even when a fully rational approach is taken to protein design, the limited understanding of protein folding and dynamics makes it impossible to reliably predict whether a particular amino acid sequence will catalyze the desired reaction. While many groups are working to improve this predictive capability [77], others are seeking to use combinatorial approaches to create artificial enzymes without relying on such predictions. [23]

1.5.3 Testing and refinement of candidates

Once a library of artificial enzyme candidates has been made, they are then expressed and assayed for the desired activity. If no active enzymes are identified, the process starts over either with selection of a new scaffold or by choosing to introduce different mutations into the existing scaffold. However, if active enzymes are found, they will be investigated and refined to improve activity. This refinement is often necessary as the first generation artificial enzymes typically have poor activity. [23, 77] Refining these artificial enzymes is useful not only to increase the activity, but it also helps us understand how they work. This in turn promotes the further refinement of the criteria for scaffold selection and mutations.

1.6 Rational design of artificial enzymes

In rational design, scientists use computational modeling of protein folds to find a set of mutations which would introduce a new active site into a particular protein scaffold. The first step is to use the known or predicted transition state to create a theozyme, the ideal locations of amino acids to stabilize the transition state independent of a protein fold. (Figure 1.3) Second, the theozyme is compared to known protein folds

to see which folds could accommodate the proposed active site. If suitable folds are found, the active site is then modeled into the protein scaffold. Third, the models undergo additional refinement to better stabilize the active site and promote substrate binding. In total, more than 10^{19} constructs are able to be examined *in silico* but this is many orders of magnitude larger than what could be reasonably examined in the laboratory. [82] The final step is to rank the different constructs with a quality score so the top candidates can be expressed and characterized for the desired activity. [77] The calculations and refinements necessary to find artificial enzymes by this approach are not trivial and generally inaccessible to most biochemists and enzymologists.

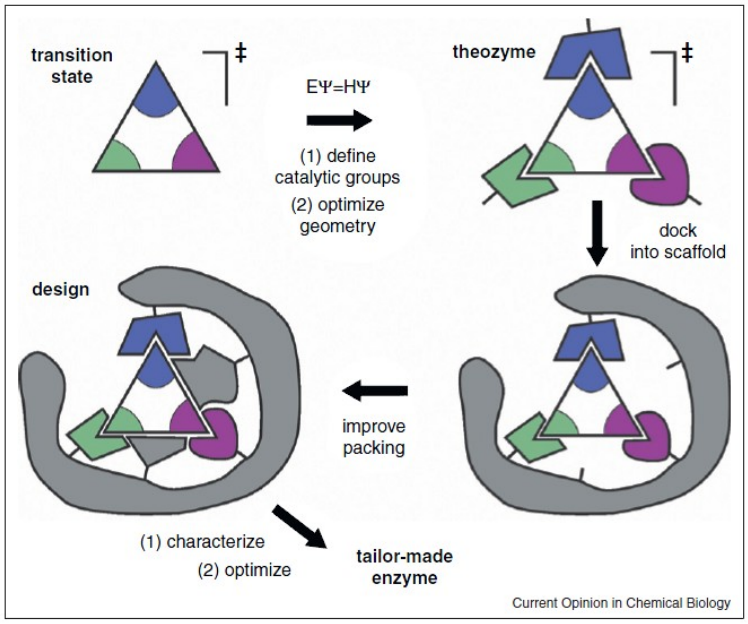


Figure 1.3 – General workflow for creating an artificial enzyme by rational design [77].

One example of rational design of an artificial enzyme is the creation of an enzyme capable of catalyzing the Diels-Alder reaction from organic synthesis. [82] The Diels-Alder reaction is a cycloaddition reaction between a diene and a dienophile forming two carbon-carbon bonds and up to four stereocenters. To design this enzyme, the authors utilized mechanistic knowledge of the Diels-Alder reaction to stabilize the transition state as well as molecular docking to promote binding of the substrate in a productive conformation. Of the vast number of scaffolds considered *in silico*, only 80 were chosen to be expressed in *E.coli* and screened for activity. Of these 80, only 2 had measureable

Diels-Alderase activity. These two hits were optimized using directed evolution further to increase catalytic activity. The best enzyme, DA_20_10 had reasonably good enantioselectivity for the desired product (>94% ee) and was capable of numerous turnovers. [82] This enzyme was eventually optimized further by additional rational design to improve the catalytic efficiency about 18 fold. Notably this improvement came from players of the online game Foldit, a game where players try to optimize protein folding utilizing intuition instead of sophisticated modeling or calculations. [84]

1.7 *De novo* selection of artificial enzymes

De novo selection of artificial enzymes uses the validated methods of directed evolution with an *in vitro* selection to isolate artificial enzymes. Directed evolution is based on the simple principle that beneficial mutation exist within protein sequence space and the experiment simply needs to sample enough mutations to find them. To evolve an existing enzyme, sampling thousands to millions (10^3-10^8) of unique variants is usually enough to find an improved enzyme. With the advent of methods for the *in vitro* enzyme evolution that can sample trillions (10^{12}-10^{13}) of variants in a single experiment, it is now possible to isolate new artificial enzymes from sequence space *de novo*.

In a 2007 publication by Dr. Burckhard Seelig, he described the first example of artificial enzymes made by *de novo* selection which catalyzed the 5'-triphosphate dependent ligation of two RNA strands forming a native 5'-3' linkage (Figure 1.4). [38] There are no natural enzymes known to catalyze this reaction, but artificial ribozymes and deoxyribozymes had been generated utilizing *de novo* selection from random pools of RNA and DNA respectively. [85-88] The 5'-triphosphate dependent ligation was used here as a model system to test whether protein enzymes could also be selected *de novo* from diverse protein libraries.

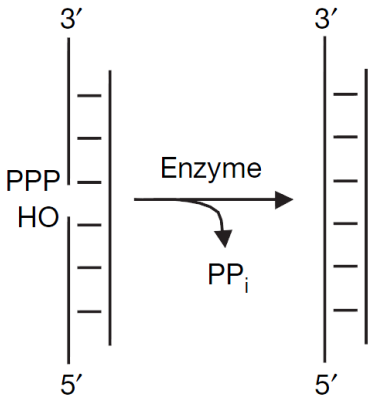


Figure 1.4 - Splinted ligation of RNA with a 5' triphosphate releasing pyrophosphate [38].

The artificial RNA ligases were generated by a selection scheme where product formation was directly linked to immobilization on a solid support (Figure 1.5). A DNA library containing $>10^{12}$ unique variants was transcribed *in vitro* to make RNA followed by modification of the 3' end of the RNA with puromycin. The RNA library was then translated *in vitro* to create the mRNA displayed fusions through the action of puromycin which covalently links each protein to its encoding mRNA. Fusions were then purified and reverse transcribed with a primer modified with 5'-triphosphate RNA substrate. After purification, a biotinylated substrate with the 3'-hydroxyl RNA was then introduced and splinted to the other substrate with a complementary oligonucleotide. If an enzyme could join these two RNA strands together, it will form a continuous covalent linkage between its cDNA and biotin allowing for it to be purified by a streptavidin column. The isolated cDNA was then PCR amplified to reintroduce the promoter to the 5' UTR which was lost during *in vitro* transcription, enabling another round of enzyme selection. Performing multiple rounds of selection enables the isolation of enzymes that might be rare in the starting library and allows for the evolution of better variants.

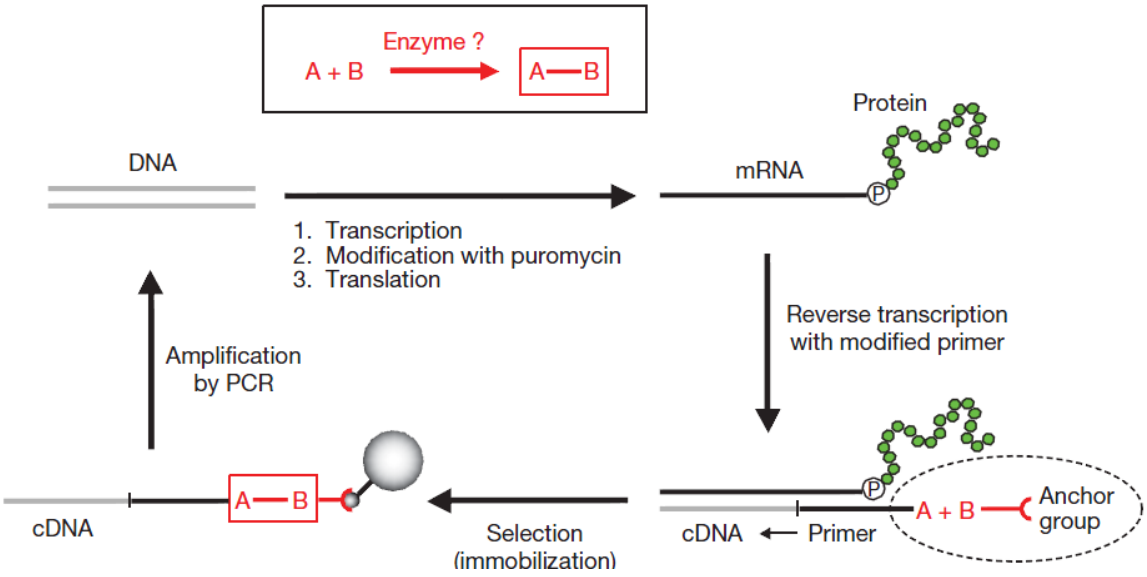


Figure 1.5 - A general scheme for the selection enzymes that catalyze bond-formation. [38] When the active enzyme forms a bond between A and B, it forms a covalent link between the anchor group and the enzyme's cDNA. The anchor group allows for the cDNA of active enzymes to be purified away from the cDNA of inactive enzymes which lack the anchor group.

The protein library used to isolate the artificial RNA ligase was based on the transcription factor hRxRα. The scaffold protein contained two zinc finger domains with a Zn^{2+} ion coordinated by four Cysteine residues each. Diversity was introduced into this protein fold by randomizing two loops of 9 and 12 amino acid positions between the two zinc finger domains, leaving the zinc coordination sites intact. [89] During the course of evolution and selection additional mutations were introduced into the constant regions of the scaffold (Figure 1.6). The artificial RNA ligases isolated had undergone significant remodeling of the protein scaffold including the loss of some of the Cys residues responsible for zinc coordination in the native scaffold. Some clones also contained large deletions of more than 10 amino acids. Nevertheless the artificial RNA ligases are dependent on zinc for activity suggesting the proteins were still capable of binding zinc

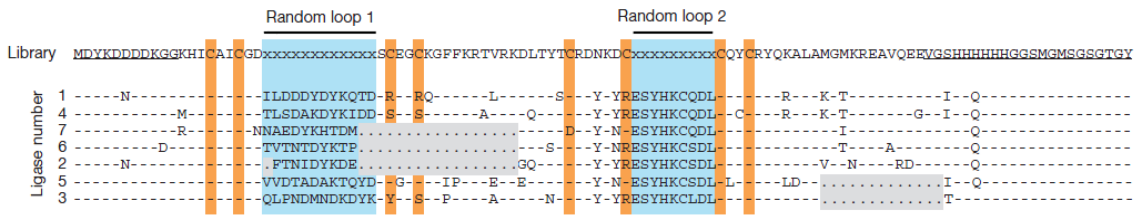


Figure 1.6 - Sequences of starting library and select artificial RNA ligases [38]. Each random loop (blue) is flanked by two Cys residues (orange) which compose the Zn^{2+} binding site of the zinc finger. For many enzymes, the scaffold has undergone substantial rearrangement as seen by the large deletions (gray dots) and loss of Cys residues responsible for Zn^{2+} coordination. Loop 2 was highly conserved among all active enzymes isolated suggesting it plays a key role in the new enzyme.

The direct selection of an artificial RNA ligase was significant milestone for the field of enzyme development. Not only did it demonstrate that directed evolution could be used to isolate artificial enzymes without using rational design, but it isolated an enzyme that could catalyze a reaction not observed in nature. It's important to note that while the selection scheme was responsible for isolating the enzymes, the enzymes came from the starting library. Therefore the quality of the starting library is a key determining factor in the success of any selection.

1.8 Outlook for *de novo* selection

Compared to rational design, the method for *de novo* selection of artificial enzymes is a more easily accessible to the general field of biochemistry. Product formation directly promotes selection of the desired enzyme so little to no knowledge of the reaction mechanism is required for success. One of the primary challenges with *de novo* selection is synthesizing the desired substrates with the appropriate modifications such that substrate and product can be separated through selective immobilization. There are many commercially available methods designed for use by non-chemists to easily and covalently link various functional groups that could help in this process. [90] Even in cases where a synthetic organic chemist might be needed to effectively synthesize the compound, organic synthesis has far fewer uncertainties than rational design of a new enzyme. Because rational design only screens a relative few constructs for the desired activity, the method could be beneficial for chemical reactions which are difficult to adapt to *de novo* selection.

The selection scheme shown in Figure 1.5 is easily amenable to other bond-formation reactions (carbon-carbon, carbon-nitrogen, carbon-oxygen, etc) as was demonstrated with the artificial RNA ligase. This process should be easily modified to select for bond breakage as well, for example to select for a protease. [23] To select for

bond-breakage, all mRNA displayed fusions would start immobilized to a solid support and only those that are able to cleave the substrate would be released from the column and amplified for the next round of selection. As any break in the link between cDNA and the immobilizing group (biotin) would be selected for, additional rounds of counter-selection may be beneficial with a linker not containing the substrate.

Not all chemical reactions easily fit into this simple bond formation/bond breakage reaction scheme, such as isomerization and oxidoreduction style reactions. Selections for these types of activity could still be possible, provided a specific way of capturing the product could be developed. For example a ketone reductase which reduces a ketone to a secondary alcohol could be captured through a second enzyme to attach the necessary anchor group. These types of chemistry might also be selected for through the use of an antibody specific to the product if one existed with a suitably high affinity.

While large libraries have a clear benefit for the evolution of new enzymes by increasing the likelihood of finding an active or improved variant, the starting scaffold and location of randomization are also important to consider. mRNA display can produce up to 10^{13} fusions which is large enough to saturate and fully cover randomization of 10 amino acid positions simultaneously. However, the successful zinc finger library randomized 21 positions with a theoretical diversity of 10^{27}, many orders of magnitude larger than what is possible to synthesize in a laboratory. Arguably though, randomizing 21 positions simultaneously allows for researchers to sample a wider spectrum of sequence space than saturation of 10 as it allows for novel discoveries in more positions.

The artificial RNA ligases selected by *in vitro* evolution accelerate the uncatalyzed background reaction at least two million-fold, but they are slow relative to natural enzymes with a turnover of about one per hour. While mRNA display is a powerful method for finding the initial rare activity, the method is primarily suited for single turnover selections. To improve enzymes to the high levels of activity found in nature, selecting or screening for multiple turnovers is necessary which might not be practical with mRNA display. However, prior to transitioning to a different evolution platform, additional rounds of mRNA display might be beneficial to select for different properties. As it is a fully *in vitro* method, it is easy to change the selection system to

select for enzymes that are active at high temperature, in the presence of organic solvents or inhibitors, or at acidic or basic pH values. If the final application for the enzyme utilizes one or more of these conditions, finding an enzyme that is active under those conditions early in the evolution would be highly advantageous to save time and effort later in the enzyme development.

Chapter 2:

Highly Diverse Protein Library Based on the Ubiquitous $(\beta/\alpha)_8$ Enzyme Fold Yields Well-Structured Proteins Through *In Vitro* Folding Selection

The following is a reprint of the article: Golynskiy, M. V., Haugner, J. C., and Seelig, B. (2013) Highly diverse protein library based on the ubiquitous $(b/a)_8$ enzyme fold yields well-structured proteins through *in vitro* folding selection. *ChemBioChem* **14,** 1553-63. The article is reprinted here with permission from John Wiley and Sons. Dr. Seelig, Dr. Golynskiy and I designed the library and analyzed the data. Dr. Golynskiy assembled the DNA fragments, performed FACS analysis and screened clones for solubility. I characterized GDPDwt and GDPDmut, designed and optimized the protease digestion selection and subjected numerous libraries to the folding selection.

Hyperlink to original publication
http://onlinelibrary.wiley.com/doi/10.1002/cbic.201300326/abstract;jsessionid=8DF048F37B6DF295C051A508718F3EC0.f04t02

2.1 Overview

Proper protein folding is a prerequisite for protein stability and enzymatic activity. While directed evolution can be a powerful tool to investigate enzymatic function and to isolate novel activities, well-designed libraries of folded proteins are essential. In vitro selection methods are particularly capable of searching for enzymatic activities in libraries of trillions of protein variants, yet quality libraries of well-folded enzymes with such high diversity are lacking. We describe the construction and detailed characterization of a folding-enriched protein library based on the ubiquitous $(\beta/\alpha)_8$ barrel fold found in five of the six enzyme classes. We introduced seven randomized loops on the catalytic face of the monomeric, thermostable $(\beta/\alpha)_8$ barrel of glycerophosphodiester phosphodiesterase (GDPD) from Thermotoga maritima. We employed an in vitro folding

selection based on protease digestion to enrich intermediate libraries containing three to four randomized loops for folded variants and then combined them to assemble the final library (10^{14} DNA sequences). The resulting library was analyzed using the in vitro protease assay and an in vivo GFP-folding assay and contains ~10^{12} soluble monomeric protein variants. We isolated six library members and demonstrated that these proteins are soluble, monomeric and show TIM barrel fold-like secondary and tertiary structure. The quality of the folding-enriched library improved up to 50-fold compared to a control library that was assembled without the folding selection. To the best of our knowledge, this work is the first example of combining the ultra-high throughput method mRNA display with a selection for folding. The resulting $(\beta/\alpha)_8$ barrel libraries provide a valuable starting point to study the unique catalytic capabilities of the $(\beta/\alpha)_8$ fold, and to isolate novel enzymes.

2.2 Introduction

Directed evolution experiments have generated numerous commercially valuable enzymes and have helped gain insight into the origins and evolution of enzymatic function. The success of any directed evolution experiment fundamentally depends on the diversity and quality of the starting library of protein variants. A protein library is considered of high quality if a substantial fraction of the library consists of well-folded, soluble and stable proteins that contain a diverse set of mutations and potential active sites for a variety of desired activities. *In vitro* selection strategies generally outperform *in vivo* or screening approaches by several orders of magnitude with regard to library diversity and are preferred for the isolation of potentially very rare mutants, e.g. novel enzymes. [1-3] However, high quality enzymatic libraries that can harness the ultra-high throughput of *in vitro* methods are currently lacking.

The ubiquitous $(\beta/\alpha)_8$ or TIM barrel fold is a promising scaffold for a general-purpose protein library that could be used for the isolation of new enzymatic activities and the understanding of the origins of enzymatic function. This versatile fold is utilized in five of the six enzymatic classes and is highly favored by natural enzymes to catalyze a wide array of different reactions, in some cases at the diffusion rate limit. [4-7] In the

$(\beta/\alpha)_8$ barrel fold, the main structural and catalytic elements are spatially separated. The barrel itself is formed by eight alternating alpha helices and beta strands and provides the structural foundation while the eight loops connecting helices and strands on one side of the barrel are responsible for substrate binding and catalysis and are known as the catalytic face of the barrel. These features are favorable for enzyme engineering since modification of functional elements is less likely to affect the structural stability of the overall scaffold. [8] In a few cases, the catalytic activities of $(\beta/\alpha)_8$ barrel enzymes were successfully swapped through protein engineering to understand how the $(\beta/\alpha)_8$ barrel fold could be recruited to perform new activities. [9-12] Although, in some other cases, desired activities could be obtained by altering the substrate specificity of existing enzymes via targeted mutagenesis, [13] the introduction of novel activities often necessitated more extensive protein remodeling. [2, 3, 14-16] In an effort to enable more divergent sequence exploration well beyond point mutations, the tolerance of $(\beta/\alpha)_8$ scaffolds to the insertion of different natural $(\beta/\alpha)_8$ loop fragments was investigated. [17-21] Furthermore, the enzymatic activity of existing $(\beta/\alpha)_8$ barrel proteins was improved or modified by a combination of rational design and directed evolution similar to proteins of other folds. [6, 22-24] In addition, rational design approaches for *de novo* enzymes repeatedly favored the $(\beta/\alpha)_8$ barrel fold over others, likely due to its ability to appropriately position catalytic and substrate binding residues. [25-27] This is particularly significant, as despite recent success in the rational re-design of enzymes, the *de novo* design of enzymes is still considered a formidable task. [26, 28-30] In summary, the combination of valuable $(\beta/\alpha)_8$ barrel protein features like catalytic versatility, efficiency, stability, structural modularity and plasticity make this fold an ideal scaffold for enzyme engineering.

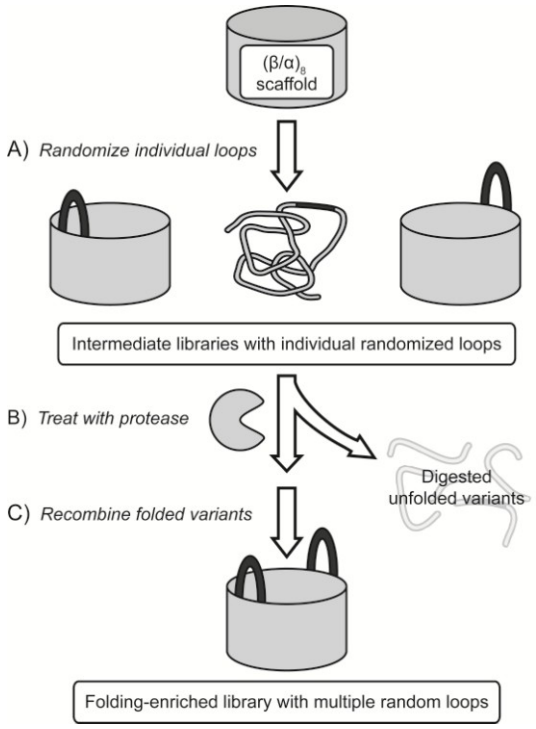


Figure 2.1 - General strategy for the stepwise construction of the folding-enriched library based on the $(\beta/\alpha)_8$ scaffold. A selection for folded proteins by protease digestion of unfolded variants is followed by recombination of folded variants to generate the final $(\beta/\alpha)_8$ library with seven randomized loops.

Herein we report the construction of a highly diverse $(\beta/\alpha)_8$ barrel library (~10^{14} unique DNA sequences) that contains seven randomized loops and is enriched for well-folded, soluble proteins. Unfortunately, the deleterious effect of mutations on stability is a major constraint in protein evolvability [31, 32] and is implicated in limiting the speed of evolution in nature. [33] Previous studies predicted that the probability of a protein to retain its structure will decline exponentially with the number of mutations. [34] An additional concern during the creation of a highly diverse protein library is the unavoidable occurrence of frameshifts and unintended stop codons caused by imperfect chemical synthesis of the respective DNA library, which can greatly reduce the number of full-length library members. [35] To generate a high quality library, we employed two complementary strategies. The first strategy removed stop codons and frameshifts from shorter library cassettes via *in vitro* selection by mRNA display. [35] The second strategy

selected for folded protein variants using protease digestion, which removed poorly folded proteins as they are more susceptible to proteolysis. [36-39] We combined these two strategies by assembling our final library *in vitro* and step-wise from intermediate libraries preselected for folded variants and the absence of frameshifts or premature stop-codons (Figure 2.1). While the selection procedures reduce the number of protein variants in the intermediate libraries, the diversity is regenerated in the final library by recombining these preselected intermediate libraries. Unlike the prior $(\beta/\alpha)_8$ library construction attempt where 49 amino acids were simultaneously inserted into all eight loops in the catalytic face of the $(\beta/\alpha)_8$ fold and likely caused unfolding of the substantial fraction of the final library, [35, 40] our conservative step-wise assembly approach aimed to significantly improve the overall library quality. In order to assess the impact of our folding selection, we additionally prepared a control library without the folding selection. The quality of the two libraries was assessed independently by orthogonal *in vitro* and *in vivo* folding assays. These libraries will be used for isolating *de novo* activities as well as for studying the origins of enzymatic function, the role of folding on the emergence of activity, and the adaptability of the omnipresent TIM barrel fold for different catalytic functions.

2.3 Results

2.3.1 Identification and characterization of a (β/α)8 scaffold protein and an unfolded control

We first sought to identify a suitable $(\beta/\alpha)_8$ scaffold candidate as a starting point for the library design. We desired a highly stable, cysteine-free, monomeric protein with a known crystal structure and chose glycerophosphodiester phosphodiesterase (GDPD) from the hyperthermophile *T. maritima* as the starting scaffold that fits all those criteria (Figure 2.2). [41] We hypothesized that the overall structure of the GDPD protein would be sufficiently stable to tolerate the replacement of loops on the catalytic face of the barrel with random sequences and even the insertion of additional amino acids. The GDPD catalytic face consists mainly of short loops and could potentially accommodate larger active sites with minimal steric clashes, similar to recent experiments that changed

TIM barrel activities. [19] To optimize our protocols for folding assessment and selection that are essential to our library assembly strategy, we prepared a destabilized GDPD construct (GDPDmut) lacking the native tertiary structure. In particular, two adjacent substitutions (G31R/V32E) were introduced to the $(\beta/\alpha)_8$ barrel to disrupt the parent GDPD structure (GDPDwt) via steric clashing and the insertion of unfavorable charge in the tightly packed core of the barrel.

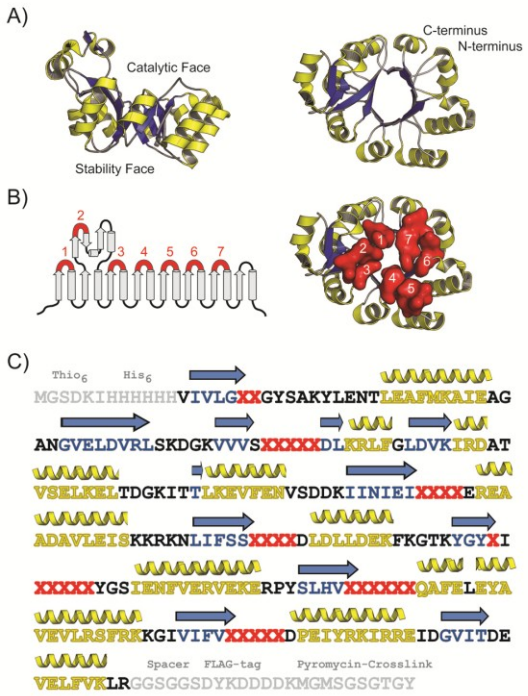


Figure 2.2 - Design of the $(\beta/\alpha)_8$ library based on the GDPD protein scaffold. A) Side view and top down view of crystal structure of the GDPD $(\beta/\alpha)_8$ scaffold that was used as a starting point for the library construction (PDB ID: 1O1Z). The □-helices and □-strands are shown in yellow and blue, respectively. **B)** Secondary structure representation and top down view of GDPD scaffold. The loops 1-7 that were randomized during library construction are numbered and shown in red. **C)** Sequence of the GDPD library. Positions randomized with the NNG/C codon are depicted as red "X". Non-native residues added to the termini of the GDPD scaffold are shown in grey (purification tags, spacers and puromycin-crosslinking region needed for mRNA display). β-strands and α-helices are colored blue and yellow, respectively.

To ascertain that the mutant construct lacks the parent $(\beta/\alpha)_8$ structure, GDPDwt and GDPDmut were expressed, purified via His_6-tag chromatography and characterized in solution. Unlike GDPDwt, GDPDmut did not express solubly but could subsequently be solubilized through purification under denaturing conditions followed by a refolding step. In contrast to the monomeric GDPDwt, GDPDmut exists almost exclusively as

oligomeric species in solution, as shown by size exclusion chromatography (Figure S3.1A). Analysis of secondary structure by far-UV circular dichroism (CD) demonstrated that both constructs possess defined, yet differing, elements of secondary structure based on the similarities at 208 nm and differences at 222 nm, wavelengths associated with α-helical structure in the far-UV CD (Figure S3.1B). In order to gain greater insight into the overall folding of the two GDPD constructs, we probed the tertiary structure via 1-anilinonaphthalene-8-sulfonic acid (ANS) fluorescence and near-UV CD (Figure S3.1C and S3.1D). Both methods showed that GDPDmut has substantially less tertiary structure and more exposed hydrophobic surface area relative to GDPDwt. After establishing that GDPDmut lacks the tertiary and quaternary structure of the parent GDPD scaffold, the two constructs were used to establish and optimize the dynamic range of the protease digestion folding selection.

2.3.2 Optimization of the folding selection by in vitro protease digestion

In order to employ the protease digestion selection to reduce the fraction of poorly folded protein variants in our $(\beta/\alpha)_8$ library, we first optimized the selection conditions to successfully discriminate between GDPDwt and GDPDmut. Selections based on protease digestion using phage and ribosome display have successfully enriched protein libraries for folded members. [36, 38] Although primarily used to improve the stability of a single protein, in one case this approach was applied to improve qualities of *de novo* libraries based on specific secondary modules. [37] Throughout our assembly protocol we utilized mRNA display, an *in vitro* selection and evolution method which employs the small molecule puromycin to covalently attach proteins to their own mRNA. [42, 43] This method had been previously used to isolate an enzyme *de novo* from a non-catalytic scaffold with two randomized loops [2, 3, 44] and is excellently suited for the long term goal of isolating enzymatic activities from the large protein libraries described here.

In pilot experiments, mRNA-displayed proteins were first treated with several proteases (data not shown) known to have preferences for hydrophobic residues, and then His_6-tag purified by immobilized metal affinity chromatography (Figure S2.2). We hypothesized that hydrophobic residues would serve as a good criterion for removing

unfolded proteins from the library as such residues are preferentially buried in the protein core and less likely to be surface-exposed in well folded proteins. [45] Chymotrypsin, which cleaves adjacent to large hydrophobic residues, showed the largest discrimination between the two control constructs in the pilot experiments and was further optimized to yield ~140-fold enrichment of GDPDwt over GDPDmut (92 ± 1.2% vs. 0.67 ± 0.12% survival). Percent survival was defined as the ratio of His_6-tag purification yields of protease-treated and untreated samples. Furthermore, mRNA-displayed fusions of GDPDmut and a GDPDmut control lacking the His_6-tag were analyzed via the same protocol to determine the level of non-specific background binding. The optimized chymotrypsin protocol was utilized for the selection and analysis of the $(\beta/\alpha)_8$ based libraries.

2.3.3 Construction of intermediate libraries with randomized loops

Intermediate libraries with several randomized loops were used as building blocks during the step-wise assembly of the final folding-enriched library (Figure S2.3). To further increase the diversity of the libraries, we also inserted one to four additional amino acids into these loops, with the exception of loop 1 (Table 2.1). We generated seven libraries with a single randomized loop each, corresponding to loops 1 through 7 on the catalytic face of the scaffold. In the next step, these libraries were used to assemble intermediate libraries with multiple randomized loops (Figure S2.3A). Specifically, fragments of the GDPD gene were PCR amplified to introduce two to six NNG/C (NNS) randomized codons at the desired loop positions, the resulting fragments were digested with restriction enzyme and ligated together to generate the full length libraries containing one or two random loops. Next, half-libraries with three or four random loops were generated by recombining PCR-amplified fragments of the libraries with one or two random loops. Loop 8 was omitted from the library assembly as its location is distant from the core of the $(\beta/\alpha)_8$ barrel and, therefore, loop 8 seemed unlikely to contribute to the formation of a potential active site with the rest of the randomized regions.

Table 2.1 - Comparison of loop length in the GDPDwt scaffold to the randomized loops used in assembling the $(\beta/\alpha)_8$ libraries

Loop #	1	2	3	4	5	6	7	8
GDPDwt loop size	2aa	3aa	1aa	1aa	5aa	2aa	4aa	1aa
Library loop size	2aa	5aa	4aa	4aa	6aa	6aa	5aa	Wild type

The introduction of multiple loops into the GDPD protein was expected to substantially destabilize the starting scaffold and reduce the fraction of folded proteins in a given library. To guide the library assembly process and decide at which step to perform either the whole folding selection or the mRNA display alone, we first analyzed the protease digestion rates of several intermediate libraries as described in the next section. The mRNA display procedure removes unintended stop codons from the library, which are introduced by the use of NNS codons for randomization, and imperfections during DNA primer synthesis. [35] The mRNA display therefore increases the quality of a library, which is beneficial for a subsequent folding selection.

2.3.4 Folding selections of the intermediate libraries by in vitro protease digestion

To evaluate the tolerance of the GDPD scaffold to amino acid insertion and randomization, several libraries containing one or two randomized loops were treated with chymotrypsin to assess the fraction of surviving library members (Table 2.2). As expected, and likely due to steric clashes between random loops from different libraries, the survival rate for libraries with two randomized loops was lower than the product of the survival rates of the two parent libraries with a single randomized loop each. The survival rates observed for the libraries containing one or two randomized loops were significantly above the GDPDmut background (Table 2.2). Therefore, to preserve some spatial context of the randomized loops, we subjected those libraries only to mRNA display to remove stop codons, and then recombined them into the two half libraries, termed “L1-4” and “L5-7”, containing randomized loops 1-4 and 5-7, respectively. Our goal was to enrich these two libraries for folded proteins until the survival rate was well above that of GDPDmut in as few rounds of selection as possible to preserve library diversity. These libraries, possessing four and three randomized loops, respectively, were therefore subjected to the folding selection (Figure S2.3A). While L5-7 exhibited 52%

survival rate, L1-4 showed a significantly lower 1.4% and the surviving variants were subjected to a second round of folding selection yielding a final survival rate of 9.2%. The increase in survival rates well above background implies that both half libraries were indeed enriched for folded sequences. Additional rounds of folding selection would decrease the diversity of enriched sequences without necessarily improving folding much further (Table 2.2).

2.3.5 Assembly of the final folding-enriched library

The stop-codon free, folding-enriched variants from the libraries L1-4 and L5-7 that survived the protease digestion selection (~10^9 and 10^{10} sequences respectively) were used to assemble the final folding-enriched library with a total of 32 randomized amino acid positions. Although combining these intermediate libraries could theoretically produce ~10^{19} unique sequences, the physical amount is limited to sub-milligram quantities of DNA that can be synthesized in the lab. Our final library contains 1.6 x 10^{14} unique DNA sequences and is at the upper limit of library sizes compatible with *in vitro* selection methods such as mRNA and ribosome display.

3.3.6 Analysis of stability of folding-enriched library and comparison to control library using the protease assay (in vitro)

In order to assess the benefits of the folding selection, a control library was prepared from the same seven single loop libraries used during the construction of the folding-enriched library (Figure S2.3B). The resulting library shared the same randomized elements and a comparable 2.9 x 10^{14} complexity as the folding-enriched library, but had not been pre-selected to maintain the parent $(\beta/\alpha)_8$ fold. A single round of mRNA display was employed to remove the stop codons and frameshifts immediately prior to the final recombination step. Rather than using the full length GDPD gene, only half-gene fragments of the L1-4 and L5-7 libraries were subjected to the round of mRNA display. By using only these fragments instead of the whole parent scaffold, we aimed to avoid a bias of the randomized loops towards the folded parent structure and allow maximum diversification. To assess the impact of the folding selection by protease

digestion, we directly compared a small fraction (~10^{10} sequences) of the control and folding-enriched libraries via our protease protocol. The folding-enriched library had a 6.6% survival rate, which is threefold higher than the control library assembled from the L1-4 and L5-7 fragments that had not been selected for folding (Table 2.2).

Table 2.2 - Results of the folding selection by *in vitro* protease digestion.

	Digested species	**% Survival [a]**
Control constructs	GDPDwt	92± 1.2
	GDPDmut	0.67± 0.12
	GDPDmut (-His6) [b]	0.4
	L3 (-His6) [b]	0.3
Analytical selections [c]	L3	28
	L4	78
	L5	80
	L3-4	10
Preparative selections [d]	L1-4 (1st round)	1.4
	L1-4 (2nd round)	9.2
	L5-7 (1st round)	52
Final libraries	Folding-enriched	6.6 ± 1.1
	Control	2.2 ± 0.3 [e]

[a] % survival is defined as fraction of mRNA-displayed species that are not digested during the chymotrypsin treatment and is calculated as the ratio (Ni-NTA purification yield of chymotrypsin treated species)/ (Ni-NTA purification yield of undigested species).
[b] Constructs lacking the His_6-tag needed for Ni-NTA purification.
[c] Small scale selections to assess tolerance of GDPDwt to the insertion of one or two loops to guide the library assembly process.
[d] Preparative selections performed to generate intermediate libraries used for the assembly of the final folding-enriched library.
[e] The % survival for the control library (loops 1-7 randomized) is higher than for library L1-4. This result is counter-intuitive and likely due to an artifact in the protease assay, potentially caused by unfolded proteins that escaped the protease digestion by aggregating (false-positives).

2.3.7 Assessment of folding of the final libraries via GFP-fused reporter assay (in vivo)

In order to confirm the efficacy of the protease digestion folding selection with an independent method, we analyzed a fraction of our libraries using a GFP-fused folding reporter system. In this system, the proteins are expressed as N-terminal fusions of GFP. The GFP fluorescence of the protein-GFP constructs is dependent on the soluble expression of the folded cargo protein and correlates with the stability to intracellular degradation. [46-48] This approach had been employed to enrich smaller protein libraries (up to 10^8) for folded variants *in vivo* and thus is an alternative to our *in vitro* folding

selection. [46] We first analyzed the several intermediate libraries that were used to construct the control library and compared them to the GDPDwt and GDPDmut controls (Figure S2.4). These intermediate libraries contained one to four randomized loops and had not been selected for folding. Since GDPDwt was shown to be solubly expressed and well behaved in solution, we were interested in the fraction of our libraries that exhibited fluorescence similar to GDPDwt-GFP fusions. In addition, we determined the mode of the GFP fluorescence as a qualitative metric for general library trends since GFP fluorescence correlates with intracellular stability. Flow cytometric analysis of *E. coli* BL21(DE3) cells expressing GDPD-GFP constructs showed a near base-line separation between GDPDwt and GDPDmut, both of which exhibit significantly higher fluorescence than cells transformed with an empty vector control plasmid. Analysis of the non-preselected libraries showed that libraries with randomized loops in the N-terminal half of the $(\beta/\alpha)_8$ barrel (libraries L1-2, L3-4 and L1-4) exhibit a lower GFP fluorescence and a lower GDPDwt-like fraction compared to the libraries with randomized loops in the C-terminal half of the $(\beta/\alpha)_8$ barrel (libraries L5, L6-7, L5-7) (Figure S2.4, Table S2.1). The folding-enriched and control libraries were analyzed in the *E. coli* BL21(DE3) Rosetta strain that provides enhanced eukaryotic protein expression since the assembly process potentially enriched for eukaryotic codons that are suboptimal for bacterial expression. We observed improvements in the folding-enriched library relative to the control library in both the mode of GFP fluorescence and the fraction of GDPDwt-like variants (Figure 2.3, Table 2.3).

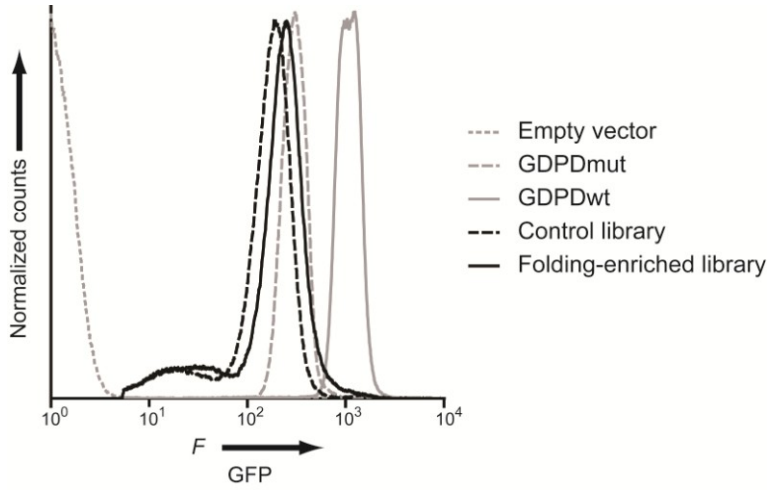


Figure 3.3 - Assessment of folding by GFP-fusion assay. Fluorescence histograms of *E. coli* BL21(DE3) Rosetta cells containing library members or control proteins fused to GFP are shown. Empty vector (gray, dotted), control library (black, dashed), folding-enriched library (black, solid), GDPDmut (gray, dashed) and GDPDwt (gray, solid). The empty vector population was gated out on the histograms of cells transformed with the GDPD constructs.

2.3.8 Isolation of well-folded members of the final libraries by cell sorting

To confirm that the soluble expression of GFP fusions is indeed closely correlated with the GDPDwt-like GFP fluorescence, control and folding-enriched libraries were sorted via fluorescence-activated cell sorting (FACS) (Figure S2.5). We subdivided the GDPDwt-like GFP fluorescence window into a low and a high GFP signal during sorting as the control library exhibited a discrete peak at high signal within this region (see gate H in Figure S2.5B). Cells with such high GFP profile could be false positives due to either insoluble aggregates or truncated proteins as noted in previous reports that used the GFP reporter system. [46]

Table 3.3 - GFP-fused *in vivo* folding assessment of the final $(\beta/\alpha)_8$ fold-based libraries. Constructs were transformed into *E. coli* BL21(DE3) Rosetta cells. Prior to analysis, data were gated to exclude cell populations that matched fluorescence and scatter profiles of cells transformed with empty vector control plasmid.

Species	Mode of GFP fluorescence [a]	% cells with GDPDwt-GFP fluorescence [b]
GDPDmut	24.6	0.01
Control library	15.4	1.4
Folding-enriched library	19.8	5.4
GDPDwt	100	98.5

[a] Values normalized to the mode of GFP fluorescence of GDPDwt-GFP.
[b] Wild type cells were gated on the forward scatter versus GFP contour plot to include ~98% of all wild type cells.

2.3.9 Analysis of soluble library-GFP fusions by Western blotting and SDS-PAGE

The four sorted populations (low and high GFP signal of each of the control and folding-enriched libraries in Figure S2.5) were re-grown in liquid culture under sorting conditions and the respective amount of soluble full-length library-GFP fusion proteins was compared by anti-GFP western blotting (Figure S2.6). A fraction of these cultures was also plated to isolate individual GFP-positive clones, express them, and analyze the soluble protein fraction of each clone by SDS-PAGE gel (data not shown). The SDS-PAGE and western blot results showed similar trends and were in good agreement with

each other (Table S2.2). The folding-enriched library populations contained a higher fraction of soluble GFP fusions in both the low and high populations compared to the control library populations. Western blot analysis also showed that the high GFP populations for both libraries contained at least ~50% false positive clones that expressed GFP alone. Based on the fraction of full length, soluble library-GFP fusions in the FACS-sorted populations, we calculated that the soluble library members comprise between 1% and 1.2% of the folding enriched library and between 0.02% and 0.033% of the control library. This corresponds to an overall 35 to 50-fold improvement in the library quality based on the fraction of soluble, monomeric and folded sequences. Therefore, the final folding-enriched $(\beta/\alpha)_8$ fold library contains about 10^{12} soluble protein variants (Table S2.2).

2.3.10 Biophysical characterization of soluble library clones

We sought to further investigate and compare the solubility of protein variants from the control and folding-enriched libraries selected at random, as well as the folding-enriched variants isolated by FACS (above). All constructs were cloned into a protein expression plasmid to express the FACS-sorted library-GFP constructs without the GFP. Only sequences from the FACS-sorted folding-enriched library produced soluble proteins (data not shown), six of which were purified for further characterization. Similar to the initial GDPDwt and GDPDmut characterization, we performed size exclusion chromatography and measured the near-UV CD and ANS fluorescence to investigate the quaternary, secondary and tertiary structure of these library variants (Figure 2.4). All of those proteins were monomeric in solution, maintained CD signatures similar to GDPDwt and ANS profiles intermediate between GDPDmut and GDPDwt.

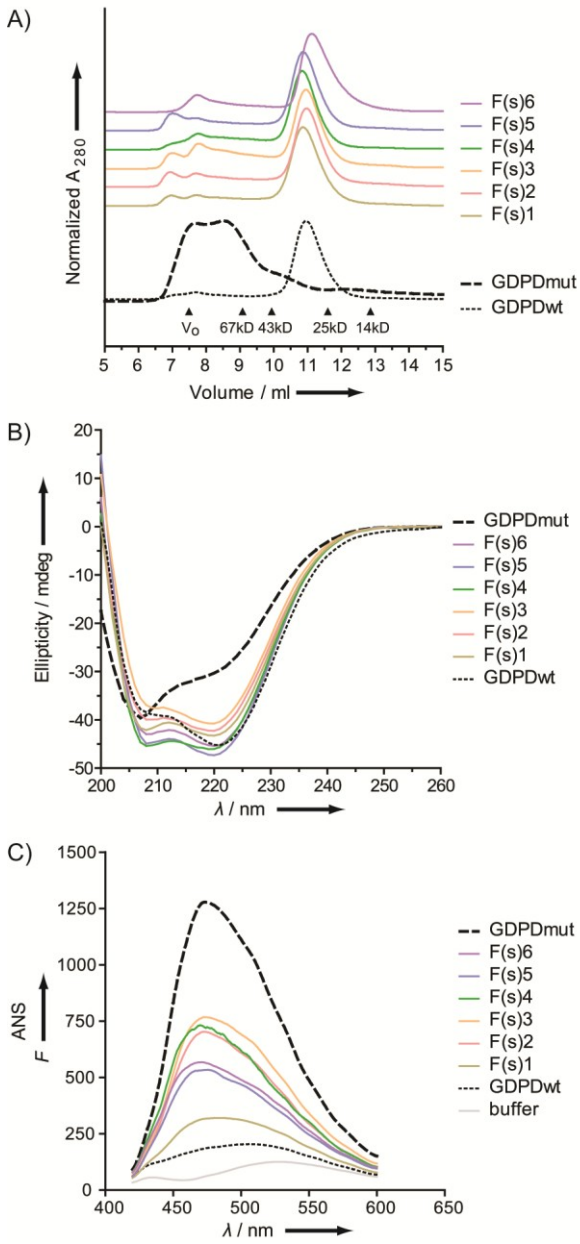


Figure 2.4 - Biophysical characterization of six soluble folding-enriched library clones from the FACS-sorted high GFP population. GDPDwt and GDPDmut data included for reference as dotted and dashed lines, respectively. **A)** Size exclusion chromatography (quaternary structure). **B)** Far-UV circular dichroism spectroscopy (secondary structure). **C)** 1-Anilinonaphthalene-8-sulfonic acid (ANS) fluorescence measurements (tertiary structure).

2.3.11 Sequence analysis of library clones

To better understand the underlying changes that occurred upon our selection for folding, we sequenced randomly chosen individual clones from the control and folding-enriched libraries, as well as the soluble folding-enriched library clones acquired by FACS sorting of the GFP-fused library. We analyzed the amino acid distribution of the 1,393 sequenced NNS codons and did not observe any stop codon, confirming that they were removed during the mRNA display step (Table 2.4). We further grouped the sequenced codons into classes of amino acids based on their properties and then compared the distributions of these classes for the control library (randomly chosen clones) and the folding-enriched library (randomly chosen clones, soluble clones) (Figure 2.5). To evaluate whether the detected distribution changes were statistically significant ($p < 0.05$), we performed pairwise t-test comparisons of the grouped codons from the folding-enriched library sequences (random and soluble clones) against the control library sequences (random clones). We observed a significant decrease in aromatic residues in the folding-enriched library relative to the control library. The soluble library clones from the folding-enriched library, isolated during FACS sorting experiment, exhibit the same decrease in aromatic residues, and, in addition, show an increase in polar residues at the expense of aliphatic residues.

Table 2.4 - Amino acid (aa) distribution for NNS codons, shown in %.

Amino acid		NNS (-stop) [a]	Control library [b]	Folding-enriched library [b]	
			Random clones	Random clones	Soluble clones
Polar	Asn	3.2	3.8	5.3	3.4
	Gln	3.2	3.2	3.8	2.8
	Ser	9.7	7.0	8.4	15.5
	Thr	6.5	7.0	5.1	6.6
Basic	Arg	9.7	10.2	10.1	11.0
	His	3.2	3.2	4.2	3.1
	Lys	3.2	5.1	5.1	3.8
Acidic	Asp	3.2	4.0	4.4	4.5
	Glu	3.2	2.2	2.7	3.8
Aliphatic	Ala	6.5	6.1	6.9	6.6
	Ile	3.2	2.9	2.7	3.1
	Leu	9.7	9.7	11.2	8.3
	Met	3.2	4.3	2.9	2.4
	Val	6.5	6.7	5.5	3.1
Aromatic	Phe	3.2	4.1	1.9	3.1
	Trp	3.2	3.0	2.9	1.4
	Tyr	3.2	2.9	1.9	2.1
Structural	Cys	3.2	1.3	1.3	1.4
	Gly	6.5	5.4	6.3	8.3
	Pro	6.5	8.0	8.0	5.9
	stop	0	Not observed	Not observed	Not observed
Codons sequenced			628	475 [c]	290 [c]

[a] Theoretical aa distribution for NNS(-stop) was calculated from the expected NNS distribution lacking a stop codon.
[b] Experimentally observed values from sequencing analysis of individual library clones.
[c] Loops containing wild type sequences were omitted from analysis.

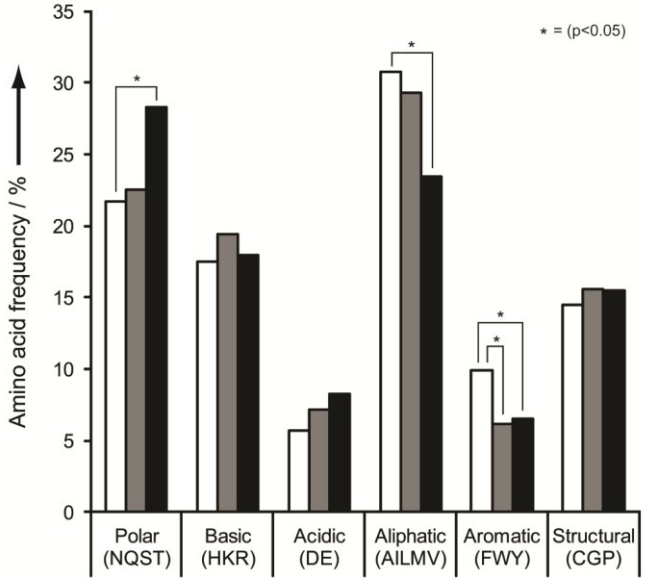


Figure 2.5 - Amino acid composition of randomized loop regions. Amino acids are grouped according to their chemical properties and the compositions were calculated from sequencing data. Control library, randomly picked clones (white); folding-enriched library, randomly picked clones (gray); folding-enriched library, soluble clones (black). Statistically significant differences, as determined by pairwise t-test, are indicated by a star.

2.4 Discussion

The objective of this study was to generate and characterize a high quality protein library based on the $(\beta/\alpha)_8$ fold by combining a step-wise assembly with an *in vitro* folding selection. We further sought to evaluate the efficacy of such an approach by comparing a representative fraction of members of the libraries using two orthogonal methods for the assessment of folding.

Our *in vitro* and *in vivo* folding assessment methods provided different metrics to measure folding stability, which are survival rates during protease digestion, the mode of fluorescence of GFP-fused library members, and the fraction of library members that behave like GDPDwt in the GFP assay. All three metrics displayed similar trends for the intermediate libraries, and showed a substantial improvement in the quality of the folding-enriched library compared to the control library, demonstrating the success of our folding selection. While those metrics were useful to characterize the libraries in bulk and assess the library construction process, they were only indirect measures for determining how much the library was enriched for soluble, well-folded protein variants that behaved

like the starting $(\beta/\alpha)_8$ scaffold. To quantify directly the fraction of those desired library variants, we cloned and expressed 20 randomly chosen proteins from both libraries in *E. coli*. We did not obtain any soluble proteins from this small sample size, indicating that the fraction of soluble variants in each library was below 5%. We therefore sorted a fraction of the GFP-fused libraries via fluorescence-activated cell sorting (FACS) and were able to isolate library members that readily expressed in bacteria, are monomeric, and exhibit behavior similar to the GDPDwt scaffold in solution. Sequencing results suggested that the improved solubility correlates, as expected, with the increased presence of polar amino acids at the expense of aliphatic residues. Furthermore, the occurrence of aromatic amino acids was reduced in the folding-enriched library compared to the control library, which might in part be a result of the selection disfavoring those residues due to chymotrypsin's preference to cleave next to aromatic amino acids. Based on the number of soluble, GDPDwt-like clones we obtained from the sorting experiment, we calculated that these sequences comprise about 1% of the folding-enriched library ($\sim 10^{12}$ variants), an increase over the control library of up to 50-fold.

The *in vitro* and *in vivo* folding methods employed in our work required the fusion of the $(\beta/\alpha)_8$ library proteins to either their own mRNA or a GFP reporter protein, which could, in principle, alter stability or solubility of the proteins. To minimize this potential issue during the *in vitro* protease digestion, the mRNA was reverse-transcribed to generate the linear mRNA-cDNA hybrid thereby preventing the mRNA from folding and affecting the digestion by obscuring protease sites. Furthermore, we assessed whether the fusion to GFP affected solubility of the library proteins by expressing soluble library-GFP constructs without the GFP fusion and analyzing them by SDS-PAGE – all proteins remained soluble in solution. Notably, during the FACS sorting experiment we encountered a substantial number of false-positive highly fluorescent cells resulting from clones that had lost their GDPD library cargo, leading to the expression of GFP alone. It has been proposed in earlier work that such false positives result from either truncated or highly aggregated and insoluble species. [46] We were able to exclude these false-positives by analyzing the soluble fraction of the expressed proteins by gel electrophoresis. The folding selection by protease digestion likely also allowed for some

false-positive protein variants to become selected. For example, we could envision certain unfolded proteins escaping the protease digest through aggregation as those proteins would be inaccessible to the protease enzyme. We counter-acted this possibility by including detergents and denaturants (Triton X-100 and SDS) in our buffers. Yet, as we cannot rule out a remaining selection bias of this kind, we deliberately chose not to further enrich the intermediate libraries L1-4 and L5-7 beyond the initial one or two rounds of selection. Finally, the biophysical characterization of individual soluble library members confirmed that our protease selection protocol successfully enriched for folded variants with a structure similar to the parental $(\beta/\alpha)_8$ scaffold.

The final folding-enriched library contains up to 32 randomized amino acid positions distributed over 7 loops. The soluble library variants isolated by FACS exhibited some variability in the location and number of loops that were randomized. Interestingly, randomization in loops 2 and 3 was disfavored in the folding-enriched library as we frequently recovered the parent GDPDwt sequence in these loops (~80% and ~40% parent sequence in randomly picked clones, respectively). All soluble clones isolated in the FACS experiments showed the parent sequence in loops 2 and 3, while containing other randomized loops. In addition, libraries that contained randomized loops 2 and/or 3 also exhibited lower protease survival rates and lower GFP-fluorescence, which is further evidence that their randomization is detrimental to the stability of the $(\beta/\alpha)_8$ barrel. We suspect that the wild type loops observed in the final library arose during the step-wise library assembly. While initially present only at very low levels in the intermediate libraries, these variants were enriched during the folding selections. However, the $\sim 10^{12}$ soluble members of the folding-enriched library have at least 3 randomized loops and at least 13 randomized amino acids. For comparison, a recent study described the switch of one $(\beta/\alpha)_8$ scaffold enzyme to an unrelated $(\beta/\alpha)_8$ activity via a single loop insertion. [19] If a new enzymatic activity can be found with the exchange of a single loop as those results suggest, our library of soluble proteins with three and more randomized loops likely has an even greater potential to contain different enzymatic activities. In addition, some of the less soluble library members may also be exploitable by *in vitro* selection methods as, for example, the mRNA display has been

shown to help keep poorly soluble proteins in solution through the attachment of a large highly-soluble RNA molecule. However, the solubility of such proteins would subsequently need to be improved through directed evolution, in contrast to the ~10^{12} already soluble library members. In summary, we demonstrated that those soluble clones have retained most of the overall structural features of the parent $(\beta/\alpha)_8$ fold despite the introduction of multiple randomized stretches of amino acids. To the best of our knowledge, this is the first report of a high quality library based on the $(\beta/\alpha)_8$ enzyme fold with such a high complexity.

Our work also allowed us to make several observations regarding the behavior of the GDPDwt $(\beta/\alpha)_8$ fold, the role of randomized loop positions and the impact of combining individual loop libraries. We observed that single, entirely randomized loop insertions into the GDPDwt resulted in libraries with 30-80% survival in the protease digestion folding selection. Interestingly, prior *in vivo* work demonstrated that single known loops inserted into an unrelated $(\beta/\alpha)_8$ barrel resulted in similar tolerances with regards to folding. [18] The authors suggested that it was the site of insertion and not the inserted sequence that had the greatest influence on the stability of the resulting protein chimera. The results we present here strongly support this notion and suggest that other $(\beta/\alpha)_8$ barrels may exhibit similar tolerances to single loop insertions, regardless of whether the loop sequence had been favored previously in nature or is entirely random. In fact, previous work suggests that random regions are beneficial in adapting known loops to the context of a new $(\beta/\alpha)_8$ barrel structure. [17] When we combined two libraries with different folding stabilities, the resulting library displayed a lower folding stability than the less stable input library, as evidenced in both the protease digestion and the GFP-fusion assay for multiple libraries. We observed a general trend where the N-terminal half of the barrel appears more vital for folding stability than the C-terminal half. This finding was inferred from the low GFP fluorescence, the high protease digestion rates, and the sequencing results for libraries containing randomized N-terminal loops. Similar positional preferences were observed in previous experiments on another $(\beta/\alpha)_8$ scaffold. [18] Although we were initially concerned that the introduction of several randomized loops into the GDPDwt scaffold would drastically unfold the $(\beta/\alpha)_8$ fold, by all our

metrics, the data indicate that this scaffold is tolerant to multiple loop insertions, particularly in the C-terminal half of the barrel. In summary, our results support the hypothesis that the core of a hyperthermophile $(\beta/\alpha)_8$ barrel fold provides sufficient stability to offset the effects of destabilizing loops of the catalytic face, and render the $(\beta/\alpha)_8$ fold an attractive scaffold in enzyme engineering by loop insertion.

2.5 Conclusion

The high quality and complexity of the libraries reported here are expected to provide an invaluable starting point for the engineering of novel enzymes and the understanding of the origins of enzymatic function in the $(\beta/\alpha)_8$ fold. By introducing randomized elements onto a stable scaffold in step-wise fashion and enriching for folded variants, we have increased the probability of finding novel enzymes with diverse activities. These initial, potentially low enzymatic activities will subsequently be evolved further under appropriate selection conditions to give rise to more efficient specialist enzymes. [40, 49, 50] Many $(\beta/\alpha)_8$ enzymes act on substrates with a phosphate group and some soluble variants of the folding-enriched library have retained the residues that compose the native phosphate binding site. This site can be used as a handle to improve substrate binding or to study the role of such handles in the evolution of enzymes. Furthermore, isolating novel activities from these libraries that are unrelated to the original GDPD function will help to elucidate whether the $(\beta/\alpha)_8$ barrel fold is predestined for certain activities, how it can be adapted to perform new functions, and what impact a library preselected for folding may have on isolation of enzymatic activity. Finally, an estimated 1% of our folding-enriched library contains sequences that are solubly expressed in *E. coli* while showing substantial diversity in the number and positioning of randomized loops. Our libraries are thus compatible with *in vitro* and *in vivo* evolution methods. Work is underway to interrogate the libraries for *de novo* enzymes using mRNA display and to study the $(\beta/\alpha)_8$ fold adaptability through bacterial selections.

2.6 Materials and methods

All chemicals were purchased from Sigma-Aldrich (St. Louis, MO) unless otherwise stated. All restriction enzymes, T4 DNA ligase and Phusion High Fidelity DNA polymerase were purchased from New England Biolabs (Ipswich, MA). All PCR reactions were performed with Phusion High Fidelity DNA polymerase. If available, high fidelity versions of the restrictions enzymes were employed. Gel extraction, PCR clean up and DNA mini-prep kits were purchased from Qiagen (Valencia, CA). Sequencing reactions were performed either by ACGT, Inc. (Wheeling, IL) or University of Minnesota BioMedical Genomics Center (St. Paul, MN).

2.6.1 Cloning and expression of GDPDwt and GDPDmut constructs:

GDPDwt gene, optimized for dual expression in rabbit reticulocyte and *E. coli*, was purchased from Genscript, USA. The construct was PCR amplified and cloned into pET28a vector (Novagen). GDPDmut was generated using standard mutagenesis protocols using pET28/GDPDwt as template. For protein expression, plasmids were transformed into BL21(DE3) Rosetta *E. coli* strains (Novagen) and grown on LB media in presence of kanamycin (34 mg/l) and chloramphenicol (34 mg/l). Overnight cultures were diluted 1:1,000 into fresh LB media and grown to OD_{600} = 1 prior to induction with IPTG (1 mM). Cells were grown an additional 4 hours at 37 °C prior to harvesting and storage at -20 °C. Frozen cell pellets were resuspended in lysis buffer (50 mM Tris-HCl pH 8.0, 50 mM NaCl) and lysed using an S-450D Digital Sonifier (Branson). Cell debris was removed by centrifugation and the His-tagged proteins were purified by affinity chromatography using Ni-NTA Superflow resin (Qiagen) under native conditions for GDPDwt and denaturing conditions for GDPDmut according to the manufacturer recommendation. Elution fractions containing GDPDmut were dialyzed to remove denaturants by first diluting 1:4 in dialysis buffer (50 mM Tris-HCl, 100 mM NaCl, pH 7.5) then dialyzing overnight in 7 kDa MWCO Snake Skin Dialysis Tubing (Pierce) in dialysis buffer. The protein purification was evaluated by SDS-PAGE on precast 4-12% gradient gels (Invitrogen

2.6.2 Circular Dichroism (CD) spectroscopy:

All CD experiments were performed on a Jasco J-815 spectropolarimeter. For far-UV experiments, ellipticity of 20 μM protein samples in 10 mM Tris-HCl pH 7.5, 20 mM NaCl was measured from 190 to 260 nm at 50 nm/min using a quartz cuvette with a 1 mm path length. Each spectrum represents the average of 10 accumulations. For near-UV experiments, ellipticity was measured the same as far-UV except a quartz cuvette with a 10 mm path length was used and wavelengths ranged from 260 to 350 nm.

2.6.3 1-Anilinonaphthalene-8-sulfonic acid (ANS) fluorescence measurements:

ANS is an environmentally sensitive dye which exhibits increased fluorescence upon interaction with hydrophobic protein surfaces and has been previously used to indirectly report on protein tertiary structure. [51] Measurements were performed on either the SpectraMax M2 or M5 plate readers (Molecular Devices) in black flat-bottom 96-well NUNC Maxisorp® plates. Samples containing protein (5 μM) and ANS (1 mM) in dialysis buffer (50 mM Tris-HCl, 100 mM NaCl, pH 7.5), were excited at 403 nm, monitoring emission at 430-600 nm in 1 nm intervals. Data was smoothed with Kaleidograph software.

2.6.4 Size exclusion chromatography:

Ni-NTA purified protein samples were loaded onto a 10 mm x 300 mm column (Tricorn) packed with Superdex 75 resin (GE Healthcare) and analyzed on the AKTA FPLC system (GE Healthcare) in a buffer containing 50 mM Tris-HCl, 100 mM NaCl, pH 7.5. Column was calibrated using Amersham low molecular weight calibration kit (GE Healthcare).

2.6.5 Library assembly:

All loop libraries were assembled via a three step process of PCR amplification, restriction digest and ligation (Figure S2.3). All PCR reactions employed a constant primer at the 5’ and 3’ termini and internal primers containing a restriction site (Tables S2.3 and S2.4). Loop randomization and insertion was carried out at the single or double

loop library level by amplifying two fragments of GDPDwt from the pET28/GDPDwt template and introducing the randomized NNS codons via one of the primers. Assembly of half libraries and final libraries was performed using internal primers that did not introduce any randomized nucleotides. Following PCR amplification, DNA was phenol-chloroform extracted and ethanol-precipitated following standard molecular biology protocols. [52] DNA was digested with appropriate restriction enzyme (Table S2.4) and purified on 2% agarose gel. Purified digested fragments were ligated with T4 DNA ligase at 16 ^{o}C overnight. The ligation product was purified on 2% agarose gel and PCR amplified with external primers to generate ~10 copies of the full length template to be used for the next set of library construction. During the construction of folding-enriched library, the L1-2, L3-4, L5 and L6-7 libraries were subjected to a single round of mRNA display (below) to remove stop codons and frameshifts and then recombined to generate L1-4 and L5-7 libraries. These libraries were subjected to protease based selection (below) and then recombined to assemble the final folding-enriched library. During the control library assembly, half gene fragments of the L1-4 and L5-7 libraries were mRNA-displayed to minimize artifacts related to folding. In the final assembly step for both the control and folding-enriched libraries, ~10^9-10^{10} L1-4 and L5-7 DNA sequences were amplified on 20 ml scale to generate ~5x10^{14} starting sequences. Due to increased scale of the *BsaI-HF* digests and final ligation reaction, DNA purification was performed via 4.5% native PAGE gel, extracted under UV-shadowing and electroeluted on S&S Elutrap (Schleicher & Schuell).

2.6.6 mRNA display:

Creation of mRNA displayed fusions was performed similarly to previously published [43] but with the following alteration. RNA was produced from the DNA library with T7 RNA polymerase (5 nM DNA template, 200 mM HEPES, 35 mM $MgCl_2$, 2 mM spermidine, 5 mM dNTP (each), 0.1 mg/ml BSA, 40 mM DTT, 1 U/ml inorganic pyrophosphatase, 150 U/ml RNaseOUT (Invitrogen), pH 7.5) and incubated at 37 °C for 3 hours. RNA was precipitated by LiCl (1/3 equivalent of 8M LiCl) at -20 °C for at least 30 min. The RNA pellet was washed with ice cold 70% ethanol and dissolved in water.

RNA was photo crosslinked (3 μM RNA, 20 mM HEPES, 100 mM KCl, 1 mM Spermidine, 1 mM EDTA, 7.5 μM oligo, pH 7.5) with a Psoralen-puromycin oligo (5′-X(tagccggtg)AAAAAAAAAAAAAAAZZACCP-3′ X = psoralen C6, lower case letters = 2′-OMe, Z = spacer 9, P = puromycin, stretch of A's and ACC = DNA) under 365 nm light on ice for 20 min with an efficiency of approximately 50%. Crosslinked RNA was ethanol precipitated and dissolved in water. A 200 μl or 1 ml translation (200 nM crosslinked RNA, 40% nuclease treated rabbit reticulocyte lysate (Promega), 25 μM amino acid mix), 25 nM ^{35}S-methionine with additional KCl and $Mg(OAc)_2$ to a final concentration of 120 mM and 0.6 mM respectively) was incubated at 30 °C for 10 min followed by high-salt incubation (550 mM KCl, 50 mM $MgCl_2$) for 5 min at RT. The translation mixture was diluted ten-fold into oligo(dT) binding buffer (20 mM Tris-HCl, 10 mM EDTA, 1 M NaCl, 0.2% Triton X-100, pH 8) and incubated with oligo(dT) cellulose (GE Healthcare, 40 mg) with rotation for 15 minutes at 4 °C. The oligo(dT) cellulose was washed on a chromatography column (Bio-Rad) with more oligo(dT) binding buffer, oligo(dT) wash buffer (20 mM Tris-HCl, 0.3 M NaCl, pH 8) and eluted with elution buffer (2 mM Tris-HCl, pH 8). The eluent was spin filtered through a 0.45 μm filter (Millipore) to remove any additional oligo(dT) cellulose and mixed with 10X phosphate buffered saline (PBS) with 0.1% Triton X-100. The mixture was added to Anti-Flag M2-Agarose Affinity Gel (25 μl, equilibrated according to the manufacturer's instructions) and incubated with rotation for at least 1 h at 4 °C. Flag resin was washed on a chromatography column (Bio-Rad) with PBS w/ 0.01% Triton X-100 followed by Flag wash buffer (50 mM HEPES, 150 mM NaCl, 0.01% Triton X-100, pH 7.4) where the final wash was performed in batch in a microcentrifuge tube. Elution was performed by incubating Flag resin with Flag peptide (56 μM in Flag wash buffer) for 10 min at 4 °C with rotation and filtered through a 0.45 μm filter (Millipore) to remove any additional Flag resin. Eluent was diluted with Flag elution buffer until mRNA displayed fusions reached $3x10^8$ fusions/μl followed by reverse transcription with Superscript II ($1.5x10^8$ fusions/μl, 50 nM RT-primer (5'-TTTTTTTTTTTTTTTN CCAGATCCAGACATTCCCAT-3'), 50 mM Tris-HCl, 3 mM $MgCl_2$, 10 mM 2-mercaptoethanol, 0.5 mM dCTP, dGTP, TTP, 5 μM dATP, 100 U/ml RNaseOUT

(Invitrogen), 500 U/ml Superscript II (Invitrogen), pH 8.3). A 10 µl sample was removed to serve as a non-radiolabeled control prior to the addition of α-^{32}P-dATP (Perkin Elmer,16 µM final concentration) to the reverse transcription. Both tubes were incubated at 42 °C for 30 min and the control was stored at -20 °C. The reverse transcription was treated with calf intestinal alkaline phosphatase (Amersham, 40 U/ml) at 37 °C for 10 min. Reverse transcribed fusions were then dialyzed in a 20K MWCO Slide-A-Lyzer (Pierce) 3-4 times against dialysis buffer (50 mM Tris-HCl, 100 mM NaCl, pH 7.5) until all unincorporated ^{32}P had been removed.

2.6.7 In vitro folding selection by protease digestion:

The dialyzed fusions were subjected to our folding selection. Triton X-100 and sodium dodecyl sulfate were added to 0.1% and 0.05% (w/v) respectively. Fusions were incubated with Chymotrypsin (Princeton Separations, 6 µg/ml) at 30 °C for 5 min, the digest was stopped by the sequential addition of phenylmethylsulfonyl fluoride (2 mM) and KCl (final concentration of 5 mM) and incubated on ice for 10 minutes. The potassium dodecyl sulfate precipitate was removed via Ultrafree-MC 0.45 µm Spinfilter (Millipore) at 4 °C followed by addition of 3 volumes of Ni-NTA binding buffer (100 mM Phosphate, 10 mM Tris-HCl, 250 mM NaCl, 6 M guanidinium hydrochloride (Amresco), 0.1% Triton X-100, pH 8). The mixture was added to 1 volume Ni-NTA agarose (Qiagen) pre-equilibrated to Ni-NTA binding buffer and incubated with rotation for at least 1 hour at 4 °C. The Ni-NTA agarose was washed on a chromatography column (Bio-Rad) with more Ni-NTA binding buffer followed by a gradient of increasing amounts of Ni-NTA native wash buffer (10 mM Tris-HCl, 250 mM NaCl, 0.01% Triton X-100, pH 8) followed by elution by Ni-NTA elution buffer (50 mM Tris-HCl, 50 mM NaCl, 500 mM imidazole, 0.01% Triton X-100, pH 8). The eluent was concentrated to a third its original volume using a SpeedVac concentrator, ethanol precipitated and dissolved in 10 mM Tris-HCl pH 8 by heating to 80 °C. cDNA was amplified by PCR with Phusion polymerase and primers to add a 5'-UTR (untranslated region). Yields from each purification step were determined via scintillation or Cerenkov counting on the Beckman LS6500 multipurpose scintillation counter.

2.6.7 GFP-based folding assay:

The GFP-based folding assay is based on the pER13a reporter plasmid previously employed to isolate protein variants with improved folding and contains an out of frame GFP. [46] A fraction of the library of interest (~10^8-10^9 sequences) was PCR amplified with Phusion polymerase and cloned into pER13a plasmid using *NdeI* and *NotI* restriction sites to generate N-terminal fusions to GFP. Libraries were ligated into the digested pER13a plasmid using T4 DNA ligase. Ligation reactions were purified via spin columns (PCR clean up kit, Qiagen) prior to electroporation into electrocompetent NEB 5-alpha cells (New England Biolabs). Following 1 hour incubation at 37 ^oC, cells were plated and grown overnight on kanamycin containing agar plates. Approximately 10^4-10^5 independent colonies were washed off the plates and their plasmids were isolated (mini-prep kit, Qiagen). BL21(DE3) and BL21(DE3) Rosetta cells (Novagen) were used for GFP-fused expression of intermediate (Table S3.1) and final libraries (Table 3.3), respectively. Electrocompetent cells prepared using standard molecular biology protocols, [52] were transformed with ~10^8 DNA sequences and grown overnight at 37 ^oC in LB media (50 ml) supplemented with kanamycin (75 mg/l) and chloramphenicol (34 mg/l). Overnight culture was used to inoculate the same medium (10 ml) and cells were grown approximately to OD_{600} = 0.6, transferred to 30 ^oC for 30 min prior to addition of IPTG (0.5 mM). Growth was continued for 6 h at 30 °C. An aliquot of the cells (1.5 ml) was pelleted by centrifugation (Eppendorf 5415R, 3 min, 4,500 rpm, room temperature), washed with phosphate-buffered saline (PBS, 1 ml) and resuspended in PBS (500 µl). Flow cytometry experiments were performed at the University Flow Cytometry Resource (University of Minnesota, Twin-Cities). Samples were analyzed on FACSCalibur (BD Biosciences) using 488 nm excitation and monitoring emission by a 530/30 nm bandpass filter. FlowJo software package (TreeStar Inc) was used for data analysis. The population of cells transformed with the empty vector was gated out from all experiments before determining the GFP mode (most frequently found fluorescence value) for the remaining cells. Cell sorting experiments were performed on FACSAria (BD Biosciences). Sorting gates were defined on the side scatter vs. GFP fluorescence dot plots. The GDPDwt-like

population gate was set based on the cells transformed with GDPDwt-GFP construct while gates for low GFP and high GFP populations were set based on the cells transformed with the GFP-fused control library. Sorted cells were used to inoculate LB medium containing kanamycin and chloramphenicol and re-grown again under sorting conditions as above for Western blot analysis. An aliquot of the re-grown cells was removed prior to IPTG induction and plated on LB agar plates containing kanamycin and chloramphenicol. Individual clones picked at random from these plates were grown in liquid culture and analyzed by SDS-PAGE for soluble expression of library-GFP variants. Six clones from the high GFP population of the GFP-fused folding-enriched library (out of 20 soluble clones identified by SDS-PAGE) were subcloned into pET28a and expressed for further characterization, as above for the GDPD control constructs.

2.6.8 Western blot analysis:

Cell pellets were lysed using BugBuster protein extraction reagent (Novagen) according to manufacturer's recommendations. Insoluble fraction was pelleted and resuspended in the original volume of the BugBuster reagent. Samples were mixed with equal volume of 2X Laemmli sample buffer (BioRad), heated for 5 min at 95 ^{o}C and spun down. A fraction of all samples was removed, diluted 10-fold and run on 4-12% gradient gel (Invitrogen). Western blotting was performed according to standard protocols [52] using affinity-purified polyclonal rabbit anti-GFP primary antibody (Abcam, cat. no 290) at 1:5,000 dilution. Anti-rabbit secondary antibody labeled with the DyLight 800 infrared dye (Cell Signaling, cat. no 5151) was used at 1:20,000 dilution and visualized using the Li-Cor Odyssey infrared imaging system. Images were analyzed using Image J software package (NIH) to quantify intensities of anti-GFP stained bands.

2.7 Supplementary Information

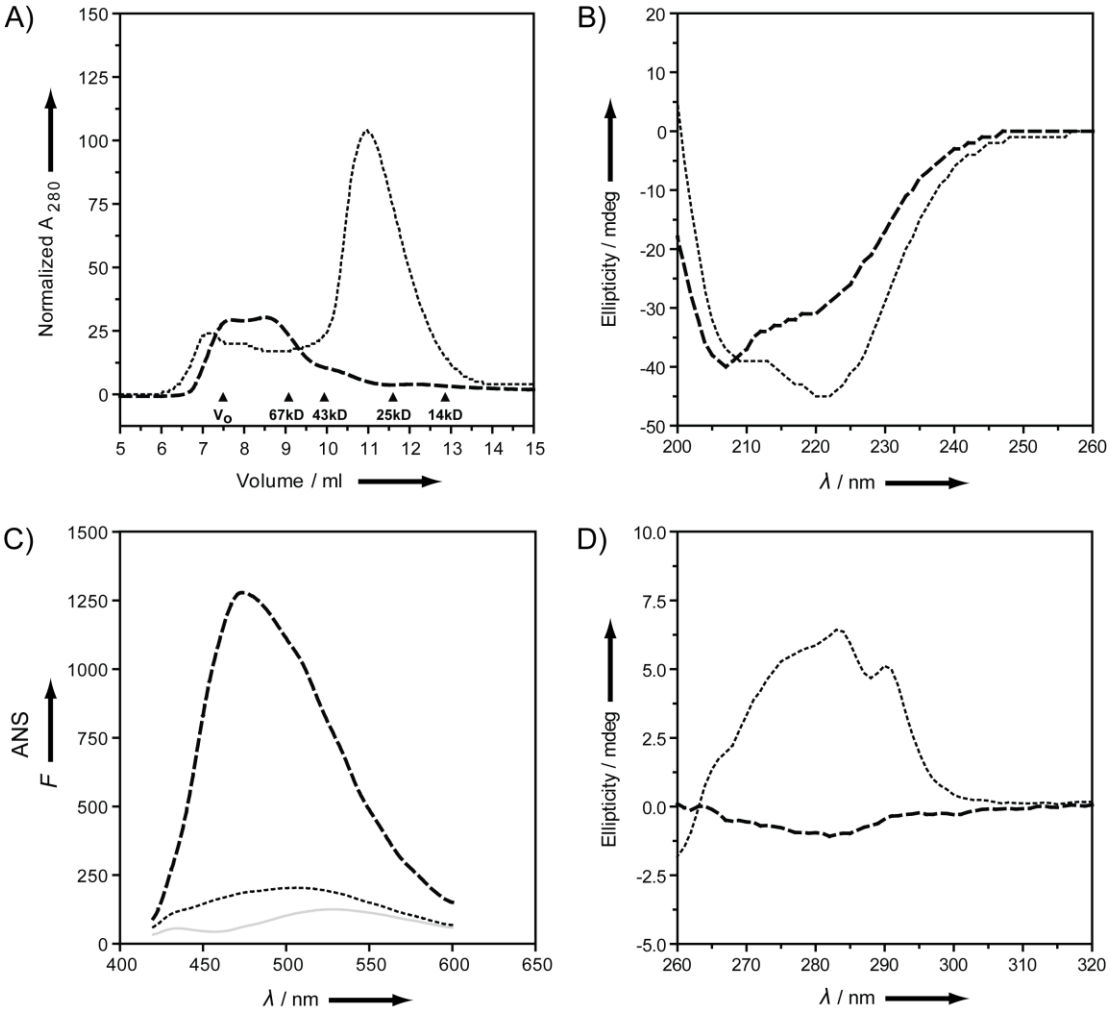


Figure S2.1 - Biophysical characterization of GDPDwt (black dotted line) and GDPDmut (black dashed line) proteins. A) Investigation of quaternary structure by size exclusion chromatography. Monomeric GDPDwt elutes at 11 ml while oligomeric GDPDmut elutes at 8 ml. **B)** Far-UV circular dichroism spectroscopy to study the secondary structure. GDPDwt exhibits characteristic alpha-helical peaks at 208 nm and 222 nm while GDPDmut has lost some secondary structure, as indicated by signal decrease in 222 nm. **C)** 1-Anilinonaphthalene-8-sulfonic acid (ANS) fluorescence measurements (tertiary structure). ANS fluorescence is known to increase upon interaction with exposed hydrophobic protein patches, suggesting increased presence of exposed hydrophobic patches in GDPDmut, associated with the loss of the native GDPDwt tertiary structure. Buffer is shown as solid light gray line. **D)** Near-UV circular dichroism spectroscopy reveals that GDPDmut has significantly less tertiary structure relative to GDPDwt.

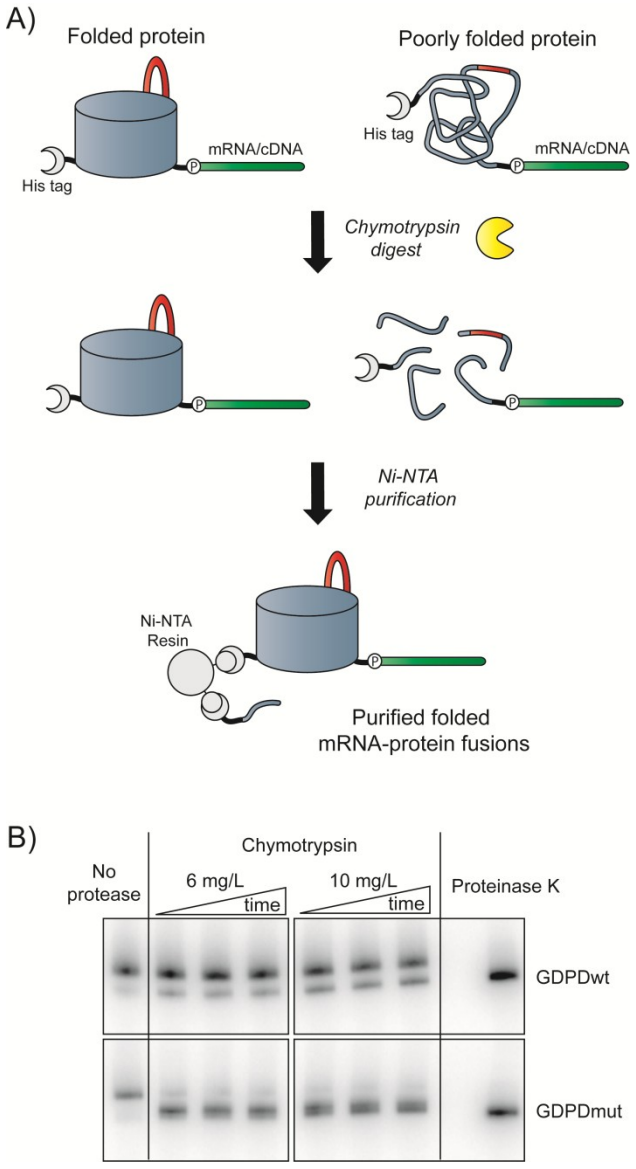


Figure S2.2 - Folding selection by *in vitro* protease digestion. **A)** Schematic of the protease-digestion based selection by mRNA display. A mixture of folded and unfolded proteins that are covalently linked to their encoding mRNA/cDNA hybrid via puromycin (P) is subjected to chymotrypsin digest and then purified under denaturing conditions via Ni-NTA affinity chromatography. Only the cDNA of the well folded proteins is immobilized on the Ni-NTA resin, and amplified by PCR for downstream applications. In contrast, cleavage of unfolded proteins severs the link between His_6-tag and cDNA, thereby preventing immobilization of the cDNA. **B)** Analysis of chymotrypsin digestion during initial optimization steps. mRNA-displayed GDPDwt and GDPDmut proteins were incubated with chymotrypsin for 10, 15 and 20 min prior to analysis of crude digest reactions by SDS-PAGE. Undigested mRNA-protein fusions and fusions treated with proteinase K (to degrade proteins non-specifically) were used as controls for 0% and 100% digestion, respectively. Under these conditions GDPDmut is preferentially digested by chymotrypsin. Final optimization steps included the additional use of detergents and an analysis of digestion based on comparison of His_6-tag purified digested and undigested samples via scintillation counting.

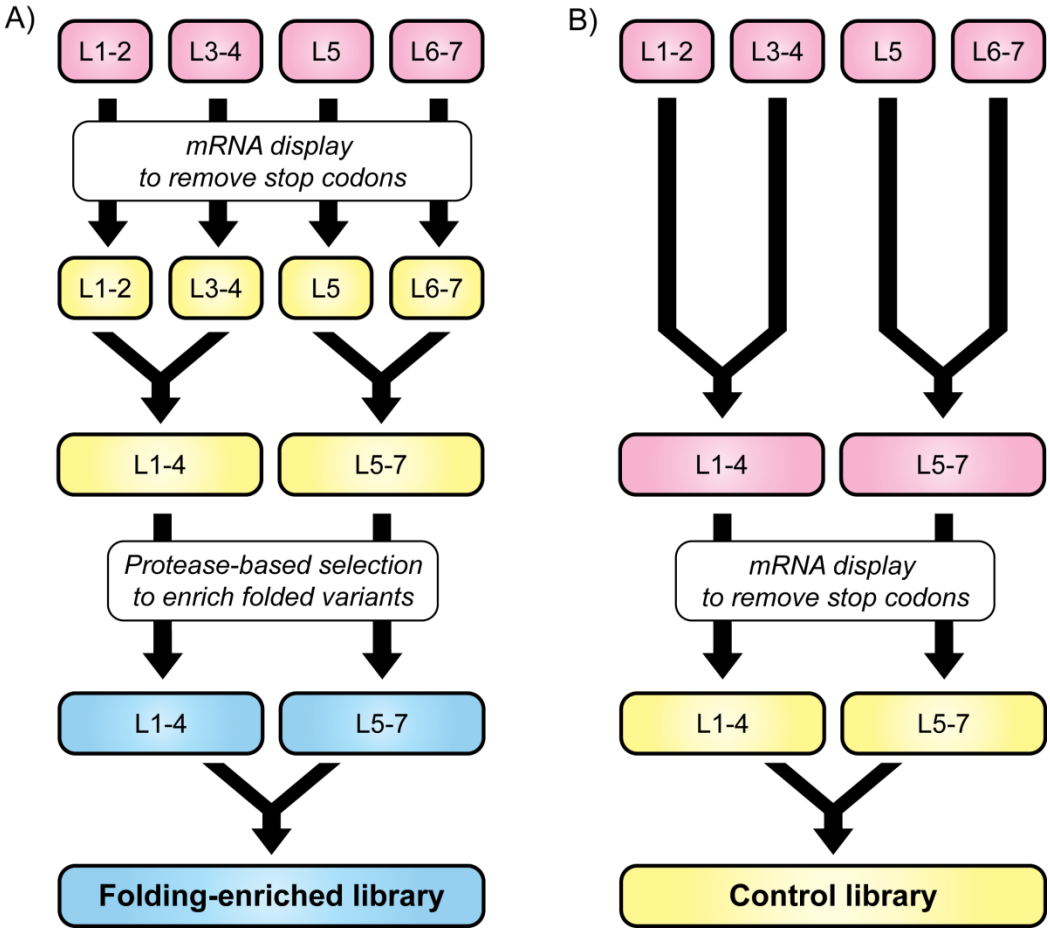


Figure S2.3 - Step-wise strategy for the construction of libraries based on the $(\beta/\alpha)_8$ fold. Each box represents an individual library with randomized loop(s) (e.g. library L1-2). Libraries containing only a single randomized loop are not shown for clarity. The libraries were mRNA-displayed to remove stop codons or, in addition, were subjected to a folding selection by *in vitro* protease digestion (shown in yellow and blue, respectively). **A)** Construction of the folding-enriched library. Full length $(\beta/\alpha)_8$ barrel libraries were used in mRNA display and folding selection. **B)** Construction of the control library. $(\beta/\alpha)_4$ half-gene fragments of the barrel libraries were used in mRNA display to prevent the native $(\beta/\alpha)_8$ fold from biasing the randomized regions in the control library.

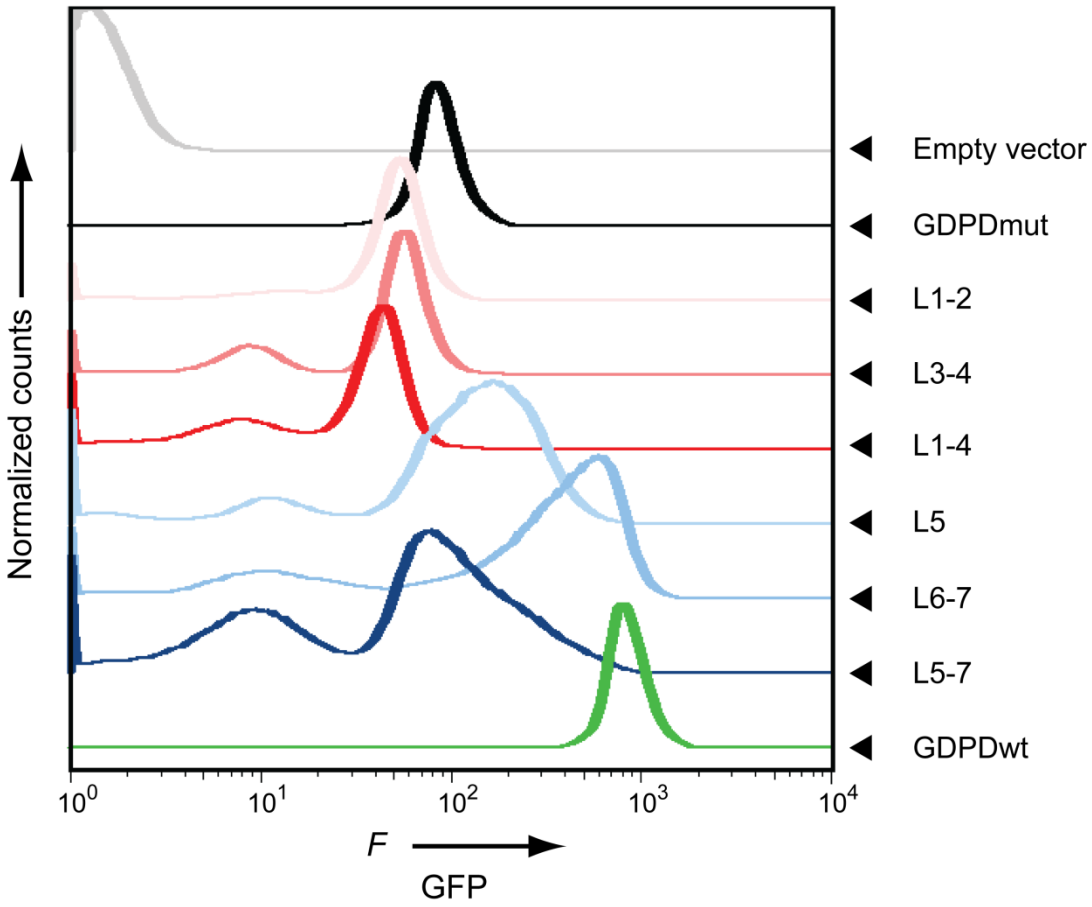


Figure S2.4 - Assessment of folding by GFP-fusion assay. Fluorescence histograms of *E. coli* BL21(DE3) cells expressing the GFP-fused control constructs and intermediate libraries. The empty vector population was gated out on the histograms of cells transformed with the GDPD constructs. None of the libraries shown here had been selected yet for folding.

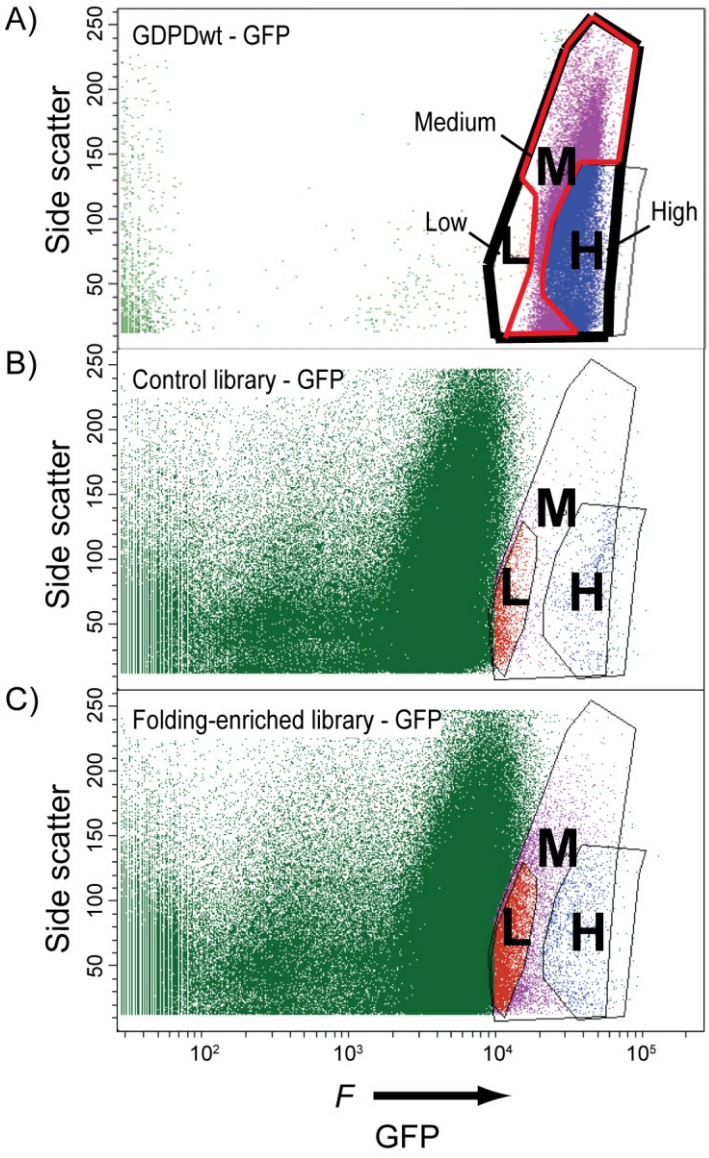


Figure S2.5 - Analysis of library populations by fluorescence-activated cell sorting experiments (FACS). Side scatter versus GFP fluorescence dot plots are shown for **A)** GDPDwt-GFP, **B)** control library and **C)** folding-enriched library. In the top panel, the GDPDwt-like population of cells is framed with a thick black line. The low (L) and high (H) populations of library-GFP cells shown in the lower two panels were collected during the sorting experiment for further analysis. The number of cells in the medium (M) population was estimated (see Table S2.2).

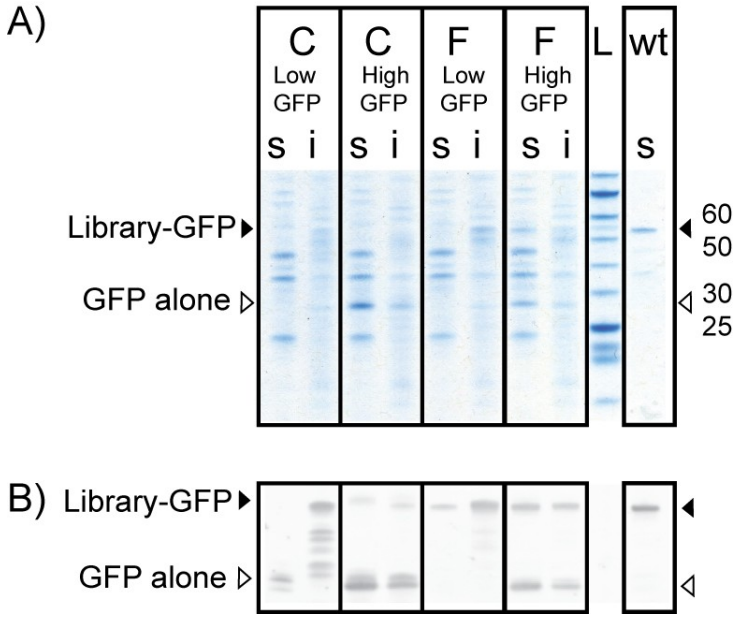


Figure S2.6 - Comparison of the soluble and insoluble fractions of the FACS sorted library populations. **A)** SDS-PAGE gel Coomassie-stained. **B)** Western blot of SDS-PAGE gel using polyclonal anti-GFP antibody. Approximate molecular weights for library-GFP fusions (57 kD, black triangle) and for GFP alone (27 kD, white triangle). Equal amounts of soluble (s) and insoluble (i) fractions were loaded for each library. The soluble fraction of over-expressed GDPDwt-GFP fusion is shown as positive control.

Table S2.1 - GFP-fused *in vivo* folding assessment of intermediate libraries. [a]

Species	Mode of GFP fluorescence [b]	% of cells with GDPDwt GFP fluorescence [c]
GDPDmut	10.4	0.02
L1-2	6.7	0.90
L3-4	7.2	0.05
L1-4	5.4	0.09
L5	21.3	9.3
L6-7	72.3	51.6
L5-7	10	6.9
GDPDwt	100	98.5

[a] Constructs were transformed into *E.coli* BL21(DE3) strain. Stop-codon containing libraries that have not been subjected to mRNA display. Prior to analysis all data were gated to exclude cell populations that matched fluorescence and scatter profiles of cells transformed with empty plasmid.
[b] Values normalized to the mode GFP fluorescence of GDPDwt-GFP.
[c] Wild type cells were gated on the forward scatter versus GFP contour plot to include ~98% of all wild type cells.

Table S2.2 - Fraction of soluble, GDPDwt-like, library-GFP fusions in the FACS-sorted populations.

		% soluble GFP fusions		
		of sorted cells		of all cells
Library population [a] [b]	% Non-empty cells [c]	By Western blot [d]	By SDS-PAGE	
Control, low GFP	0.86%	0.24%	<5% (0/18 clones)	0.002%
Control, medium GFP	0.23% [f]	0.24%-5.76% [g]	Not available	0.001%-0.013%
Control high GFP	0.31%	5.76%	<5% (0/19 clones)	0.018%
Estimated % soluble proteins in control library				0.020%-0.033% [h]
Folding-enriched, low GFP	3.18%	18.20%	29% (4/14 clones)	0.579%
Folding-enriched, mid GFP	1.68% [f]	18.20%-26.70% [g]	Not available	0.306%-0.449%
Folding-enriched, high GFP	0.54%	26.70%	24.6% (20/81 clones)	0.144%
Estimated % soluble proteins in folding-enriched library				1.029%-1.172% [h]
Enrichment of soluble proteins after folding selection [i]				35 to 50-fold

[a] Library populations fell into the GDPDwt-like profile window defined by side scatter vs. GFP fluorescence dot plot of GDPDwt-GFP construct (Figure S2.5A).
[b] Only the low and high GFP populations were sorted during the FACS experiment.
[c] % Non-empty cells defined as the ratio (# of cells analyzed - # of cells with empty-vector) / (# of cells analyzed). A total of 1.4% of the control library and 5.4% of the folding-enriched library fell into the GDPDwt-like window used for sorting and analysis.
[d] % Soluble GFP fusions of sorted cells calculated from Western blot analysis (Figure S2.6) using Image J to quantitate the intensities of all anti-GFP stained bands. Defined as the ratio (intensity of soluble GFP-fusions) / Σ(intensity of all anti-GFP stained bands).
[e] % Soluble GFP fusions of all cells calculated as (% non-empty cells x % soluble GFP fusions of sorted cells estimated from Wester blot) for the population of interest.
[f] % Non-empty cells for the unsorted medium GFP population was calculated as the difference in % non-empty cells populations (GDPDwt-like – low GFP – high GFP).
[g] Values are lower and upper estimates based on % soluble GFP fusions of sorted cells in the low and high GFP populations.
[h] Estimated % soluble proteins in a library calculated as Σ (% soluble GFP fusions of all cells) for low, medium and high GFP populations.
[i] Fold improvement defined as the ratio (estimated % soluble proteins in the folding-enriched library)/ (estimated % soluble proteins in the control library).

Table S2.3 - List of primers used during library construction.

Primer	Sequence
041B [a]	GCCTTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGCAGCGATAAGATCCACC
042 [a]	TTAATAGCCGGTGCCAGATCCAGACATTCC
018	TATGACCAAGCTTCCAGGGTGTTTTCCAGATACTTGGCGGAGTAACCSNNSNNGCCCAGCACAATCAC
019	CTGGAAAACACCCTGGAAG
047	AACAACAAGGTCTCAGGCGCTTTAAATCSNNSNNSNNSNNSNNGCTCACGACCACCTTGCC
048	AACAACAAGGTCTCGCGCCTGTTCGGTCTGGACG
028	AACAACAAGGTCTCCCTCACGTTCSNNSNNSNNSNNGATTTCGATGTTGATGATCTTG
029	AACAACAAGGTCTCGTGAGGCCGCGGACGCAGTGCTGGAGATCAGCAAAAAGCGTAAG
022	TGGAGGTCCTTGGTACCCTTGAATTTTTCATCCAGCAGGTCCAGATCSNNSNNSNNSNNGGAGCTGAAAATCAGGTTCTTAC
023	GATCTGGACCTGCTGGATG
012	AGTAGAGGTACCAAATACGGTTATNNSATCNNSNNSNNSNNSNNSTACGGTTCCATTGAAAATTTCG
015	CTTCGTCGATCAGATAACCG
003	AGTAAAGAGCTCAAAGGCCTGSNNSNNSNNSNNSNNSNNCACGTGCAGAGAGTACGGAC
017	AGGCCCTATCAGGCCTTTGAG
013	AGTAGATACGTATTTTGCGGTAGATTTCCGGATCSNNSNNSNNSNNSNNCACAAAAATCACGATGCC
016	CTGAAAGAGCTGACCGATGG
039	TTAATAGCCGGTGCCAGATCCAGACATTCCCATTTTGTCATCGTCATCCTTATAGTCGGAGCCACCGGTCTCACCCTTGAATTTTTCATCCAGC
040B	TTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGCACCATCACCATCACCATGGTCTCAAGGGTACCAAATACGGTTAT
035B	GACTGAACTGATGCCTATATGCTTGTCCGTCCAGATTGGTCTCGTGTTTGAGAACGTGTCCGATG
036B	AACTCAAGTATCGCTATGCCGGTCAATAACCGAGGTCTCAAACACTTCCTTCAGGGTGG
049	CTGCGTGGTAGCTCGTAGCACAGACTCAGCGGATACACACAGAGGTCTCAAAAATTCAAGGGTACCAAATACGG
050	GCACTCCGCTTAGATAGATAGCCAGAAGACAGACAAGGTCTCATTTTTCATCCAGCAGGTCCAG
037B	AAGTAGCATAGAGTGTCGCTCTGGATGTCAAGGTCTCAAAAGGAGCGTCCGTACTCTCTG
038B	ATGATAGCAGATGGACTTAGATTTCCGGTCAGGTGCCGAGGTCTCCCTTTTCCACGCGCTCCACG
GDPDx_001 [b]	TCTGTAAACCATGGATGGGCAGCGATAAGATCCAC
GDPDx_002 [b]	CTGTGCGCTCGAGTTAATAGCCGGTGCCAGATCC
GDPDx_003 [c]	GAAGGAGATATACATATGGGCAGCGATAAGATC
GDPDx_004 [c]	GGAGCCAGCGCGGCCGCCATAGCCGGTGCCAGATCCAG
GDPDmut_Fw [d]	GAAGCCGGCGCGAATCGTGAGGAGCTGGATGTG
GDPDmut_Rev [d]	CACATCCAGCTCCTCACGATTCGCGCCGGCTTC
030 [e]	TTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGCAGCGATAAGATCGTGATTGTGCTGGGCCATCGCGG

[a] Primers used as standard primers to amplify any full length GDPD-based template during the library construction.
[b] Primers used to amplify template DNA for insertion into the pET28a plasmid for protein expression.
[c] Primers used to amplify template DNA for insertion into the pER13 plasmid for the GFP-fusion assay.
[d] Primers used to generate pET28/GDPDmut from the pET28/GDPDwt template.
[e] Primers 030 and 042 were used to generate GDPDmut(-His_6) from the pET28/GDPDmut template.

Table S2.4 Fragments used in library assembly. [a,b]

	5'-Fragment				3'-Fragment				
Library [d]	Name	Template	FW	REV	Name	Template	FW	REV	Restriction enzyme used [c]
L1	L1_A	pET28/ GDPDwt	041B	018	L1_B	pET28/ GDPDwt	019	042	*HindIII*
L2	L2_A	pET28/ GDPDwt	041B	047	L2_B	pET28/ GDPDwt	048	042	*BsaI*
L3	L3_A	pET28/ GDPDwt	041B	028	L3_B	pET28/ GDPDwt	029	042	*BsaI*
L3(-His6)	L3_A(-His_6)	pET28/ GDPDwt	030	028	L3_B	pET28/ GDPDwt	029	042	*BsaI*
L4	L4_A	pET28/ GDPDwt	041B	022	L4_B	pET28/ GDPDwt	023	042	*KpnI*
L5	L5_A	pET28/ GDPDwt	041B	015	L5_B	pET28/ GDPDwt	012	042	*KpnI*
L6	L6_A	pET28/ GDPDwt	041B	003	L6_B	pET28/ GDPDwt	017	042	*SacI*
L7	L7_A	pET28/ GDPDwt	041B	013	L7_B	pET28/ GDPDwt	016	042	*SnaBI*
L1-2 [e]	L1-2_A	L1	041B	047	L2_B	pET28/ GDPDwt	048	042	*BsaI*
L3-4 [e]	L3-4_A	pET28/ GDPDwt	041B	028	L3-4_B	L4	029	042	*BsaI*
L6-7 [e]	L6-7_A	L6	041B	034	L6-7_B	L7	033	042	*BsaI*
L1-4 [f]	C-L1-4_A	L1-2	041B	036B	C-L1-4_B	L3-4	035B	042	*BsaI*
L5-7 [g]	C-L5-7_A	L5	041B	038B	C-L5-7_B	L6-7	037B	042	*BsaI*
L1-4 (m)	F-L1-4_A	L1-2 (m)	041B	036B	F-L1-4_B	L3-4 (m)	035B	042	*BsaI*
L5-7 (m)	F-L5-7_A	L5 (m)	041B	038B	F-L5-7_B	L6-7 (m)	037B	042	*BsaI*
Control library	C-frag_A	L1-4 frag (m)	041B	050	C-frag_B	L5-7 frag (m)	049	042	*BsaI*
Folding-enriched library	F-frag_A	L1-4 (m&p)	041B	050	F-frag_B	L5-7 (m&p)	049	042	*BsaI*

[a] Individual libraries were generated by amplifying the 5'- and 3'-fragments of the library using the specified template, and FW /REV primer pair.
[b] Libraries denoted with (m) have been subjected to mRNA display; libraries denoted with (p) have been subjected to *in vitro* folding selection by protease digestion.
[c] 5'- and 3'-fragments were digested with the indicated restriction enzyme (RE), gel-purified and ligated to produce the individual libraries. The *BsaI* restriction site, which is absent in the parent GDPDwt scaffold, was introduced into the 5-‘ and 3’ fragments by PCR amplification and was removed again during the gel purification of digested fragments.
[d] Individual libraries were gel purified after the ligation reaction. A fraction of the purified library was used as a template for the successive steps in library assembly.
[e] Libraries were subjected to mRNA display during folding-enriched library construction and then PCR amplified with 041B/042 primer pair to restore T7 promoter sequence lost during the transcription process.

[f] Library was used as template with 041B/039 primer pair to generate the control library fragment (L1-4 frag) subjected to mRNA display.
[g] Library was used as template with 040B/042 primer pair to generate the control library fragment (L5-7 frag) subjected to mRNA display.

2.7.1 DNA sequence of the GDPDwt scaffold used as template for library assembly

GCCTTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGCAGCGAT
AAGATCCACCATCACCATCACCATGTGATTGTGCTGGGCCATCGCGGTTACTCCGCCAAG
TATCTGGAAAACACCCTGGAAGCTTTCATGAAAGCGATCGAAGCCGGCGCGAATGGTGTG
GAGCTGGATGTGCGCCTGTCTAAAGACGGCAAGGTGGTCGTGAGCCATGATGAAGATTTA
AAGCGCCTGTTCGGTCTGGACGTCAAAATCCGTGACGCCACCGTGTCTGAACTGAAAGAG
CTGACCGATGGCAAAATTACCACCCTGAAGGAAGTGTTTGAGAACGTGTCCGATGACAAG
ATCATCAACATCGAAATCAAGGAACGTGAGGCCGCGGACGCAGTGCTGGAGATCAGCAAA
AAGCGTAAGAACCTGATTTTCAGCTCCTTTGATCTGGACCTGCTGGATGAAAAATTCAAG
GGTACCAAATACGGTTATCTGATCGACGAAGAGAACTACGGTTCCATTGAAAATTTCGTG
GAGCGCGTGGAAAAGGAGCGTCCGTACTCTCTGCACGTGCCCTATCAGGCCTTTGAGCTC
GAATATGCGGTGGAGGTGCTGCGCTCCTTCCGTAAAAAGGGCATCGTGATTTTTGTGTGG
ACCCTGAATGATCCGGAAATCTACCGCAAAATACGTAGAGAGATCGATGGTGTGATTACC
GACGAAGTGGAGCTGTTTGTGAAACTGCGTGGCGGCAGCGGTGGCTCCGACTATAAGGAT
GACGATGACAAAATGGGAATGTCTGGATCTGGCACCGGCTATTAA

Color code:

T7 transcription promoter / TMV translation enhancer

Thio6/His_6

GDPDwt

$(GGS)_2$ spacer

FLAG tag

Puromycin-crosslinking region

Primers 041B and 042 were used as standard primers to amplify any full length GDPD-based template during the library construction. Primers GDPDx_001 and GDPDx_002 were used to amplify template DNA for insertion into the pET28a plasmid for protein expression. Primers GDPDx_003 and GDPDx_004 were used to amplify template DNA for insertion into the pER13 plasmid for the GFP-fusion assay. Primers GDPDmut_Fw and GDPDmut_Rev were used to generate pET28/GDPDmut from the pET28/GDPDwt template. Primers 030 and 042 were used to generate GDPDmut(-His_6) from the pET28/GDPDmut template.

2.7.2 Sequence alignment of the six soluble F(s) clones characterized in this chapter (Figure 2.4).

Alignment was performed using Clustal Omega server, Clustal O(1.1.0)

Loop residues randomized during library construction are highlighted in green.

* = column contains identical amino acid

: = column contains different but highly conserved amino acids

```
GDPDwt      MGSDKIHHHHHHVIVLGHRGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)1       MGSDKIHHHHHHVIVLGSRGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)2       MGSDKIHHHHHHVIVLGHRGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)3       MGSDKIHHHHHHVIVLGRLGYSAKYLENTLEAFMRAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)4       MGSDKIHHHHHHVIVLGNGGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)5       MGSDKIHHHHHHVIVLGRVGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)6       MGSDKIHHHHHHVIVLGRLGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
            **********************  ***************:**********************************

GDPDwt      LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)1       LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKTINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)2       LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)3       LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)4       LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)5       LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)6       LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
            ********************************** ********************************

GDPDwt      LDEKFKGTKYGYIIDE-ENYGSIENFVERVEKERPYSLHVP----YQAFELEYAVEVLRSFRKKGIVIF
F(s)1       LDEKFKGTKYGYLIDE-ENYGSIENFVERVEKERPYSLHVTPTLLSQAFELEYAVEVLRSFRKKGIVIF
F(s)2       LDEKFKGTKYGYKISLWASYGSIENFVERVEKERPYSLHVSSTKDAQAFELEYAVEVLRSFRKKGIVIF
F(s)3       LDEKFKGTKYGYKIGRGGGYGSIENFVERVEKERPYSLHVYSGSPLQAFELEYAVEVLRSFRKKGIVIF
F(s)4       LDEKFKGTKYGYIISLKDTYGSIENFVERVEKERPYSLHVQRASFKQAFELEYAVEVLRSLRKKGIVIF
F(s)5       LDEKFKGTKYGYGIAEGLVYGSIENFVERVEKERPYSLHVELEFMIQAFELEYAVEVLRSFRKKGIVIF
F(s)6       LDEKFKGTKYGYLIDE-ENYGSIENFVERVEKERPYSLHVAVGRVLQAFELEYAVEVLRSFRKKGIVIF
            ************ *     *********************      **************:********

GDPDwt      VWT-LNDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)1       VKNNVCDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)2       VPCLRCDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)3       VASSTHDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)4       VAPDLPDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)5       VRADMSDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)6       VTSVTRDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
            *     ************************************************
```

Chapter 3:

Thermostable artificial enzyme isolated by *in vitro* selection

The following is a reprint of the manuscript: Morelli, A., Haugner III, J. C., and Seelig, B. (2014) Thermostable artificial enzyme isolated by *in vitro* selection. *PLOS ONE*. The article is currently in minor revisions at this open access journal. Seelig designed the high-temperature substrate and performed the selection. I expressed the artificial ligases and characterized their activity. Morelli performed all circular dichroism measurements.

3.1 Overview

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.

3.2 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 [1]. 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 [2]. Conversely, improved thermal stability correlates with mutational robustness and evolvability [3].

Methods to increase the thermodynamic stability of proteins include rational design, consensus-based design, directed evolution, and commonly some combination of these approaches [4]. Rational design introduces mutations predicted to enable additional stabilizing interactions [5]. 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 [5]. 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 [6, 7]. High throughput screens are often performed *in vivo*, utilizing colorimetric [8] or fluorescent [9] reporters to measure levels

of soluble expression as readout for stability or *in vitro* using protease resistance and phage display [10, 11]. Proteins can also be assayed directly for thermostability and activity from lysates or extracts, but these methods have a relatively low throughput [4, 12]. 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 [13, 14]. These individual methods aimed to generate more stable protein variants are frequently combined for best results [15].

We previously reported the *in vitro* selection of *de novo* RNA ligase enzymes that catalyze a reaction not observed in nature [16]. 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 [16]. 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, including ligase 10C. 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 [17, 18]. 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 [13]. 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 [16] 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 was 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 and an application for this enzyme have been described elsewhere [19, 20]. 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.

3.3 Results

3.3.1 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 [16] 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 [16] (Figure 3.1). During the selection at 23 °C, this splint base-paired to eight nucleotides of each substrate (Figure 3.1B). 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 3.1C). 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 S3.1).

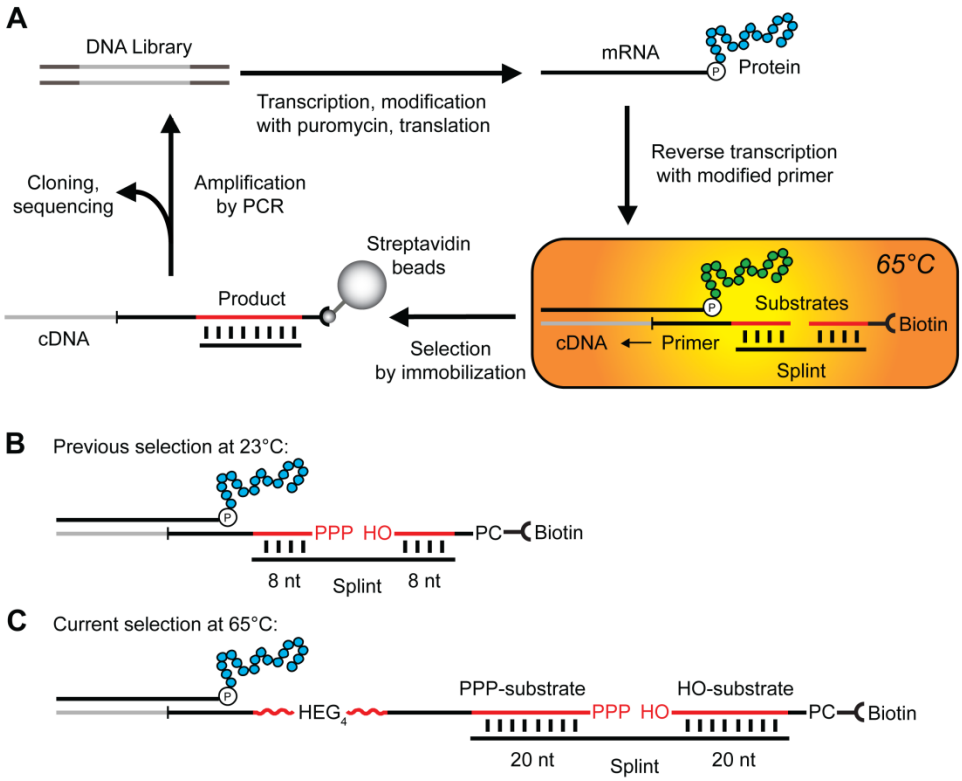


Figure 3.1 - *In vitro* selection of artificial ligase enzymes with increased stability. (**A**) Schematic of the isolation of ligase enzymes. The DNA library encodes the library of proteins that resulted from the original selection of ligase enzymes at 23 °C [16, 21]. The DNA is transcribed into RNA, modified with puromycin at the 3'-end and translated *in vitro* yielding a library of mRNA-displayed proteins [21]. 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 5'-triphosphate RNA (PPP-substrate) and 3'-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 [16, 21], 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 3.1A). 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 3.1C). 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 3.1C) resulted in a ligase activity of about 50% compared to a ligation using the shorter RT primer and the 16 nt splint (Figure 3.1B).

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.

3.3.2 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 3.1A). 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 3.2. 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.

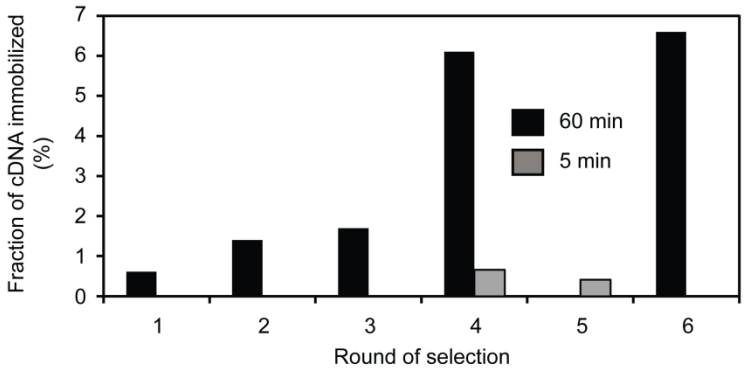


Figure 3.2 - Progress of selection for ligases at 65 °C. The fraction of ^{32}P-labelled cDNA that bound to streptavidin agarose after each round of selection is shown. The reaction time was either 60 min or 5 min as indicated by black or gray bars, respectively.

3.3.3 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 S3.2). One representative clone from each family was cloned and expressed in *E. coli* to examine soluble expression (Figure S3.3). 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 3.3). 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 [16], but these proteins were poorly soluble when expressed without an maltose-binding protein fusion and therefore unsuited for a direct comparison.

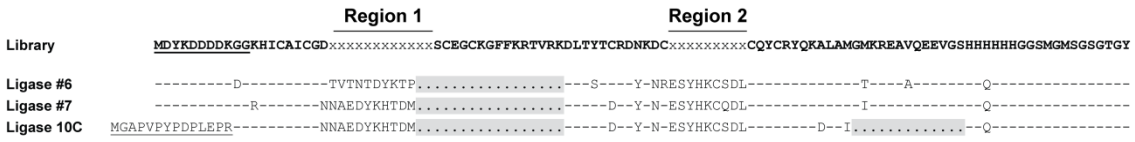


Figure 3.3 - Sequence alignment of the library used as input for original ligase selection (on top) [40] with ligases #6, #7 [16] and 10C that were selected at 23 °C and at 65 °C, respectively [41, 42]. The amino acids in regions 1 and 2 of the original library were randomized prior to the selection 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.

3.3.4 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 3.4). 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 S3.1). 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.

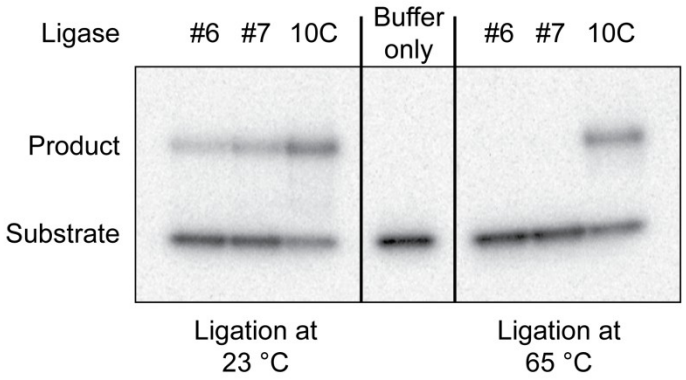


Figure 3.4 Activity of ligase enzymes assayed at different temperatures. Ligases #6 and #7 had been selected previously at 23 °C [16, 21] and ligase 10C was selected at 65 °C. In this assay, the ^{32}P-labeled PPP-substrate-65, HO-substrate-65 and 40 nt splint were incubated with the individual enzymes for 16 h and activity monitored by gel-shift.

3.3.5 Characterization of Thermal Stability by Circular Dichroism (CD)

In order to assess if the unique enzymatic activity of ligase 10C at 65 ºC 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 S3.4). 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 [23], 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 3.5). As determined from the curves, the enzymes showed very different melting temperatures. Ligase 10C had the highest melting temperature (T_m = 72 ºC), which was 35 degrees higher than the T_m of ligase #6 (37 ºC), and 24 degrees higher than the T_m of ligase #7 (48 ºC). The high melting temperature of 72 ºC for ligase 10C was in agreement with its retained enzymatic activity at 65 ºC as the enzyme has not undergone unfolding yet. In contrast, ligases #6

and #7 were fully denatured at 65 ºC, and, therefore, their complete lack of enzymatic activity at 65 ºC could be explained by their unfolding.

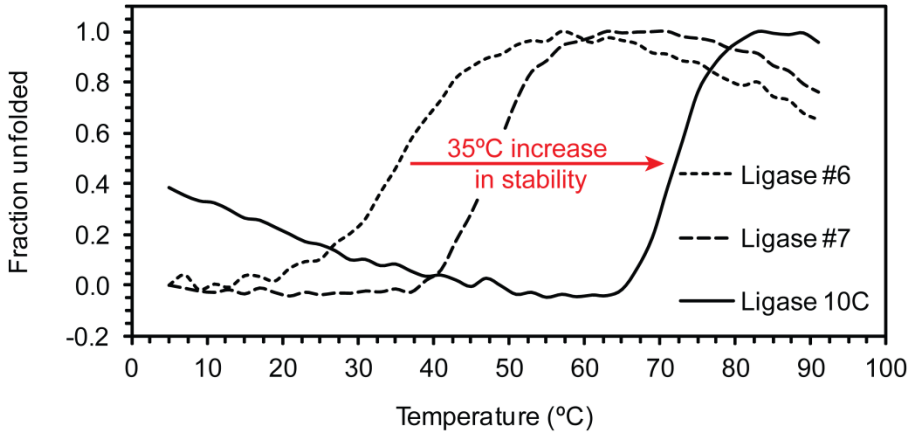


Figure 3.5 - Thermal unfolding curves of ligases #6, #7 and 10C. Thermal unfolding was monitored by circular dichroism at 222 nm. For each measurement 10 accumulations were acquired.

3.4 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 [5]. The T_M increase by 35 degrees reported here favorably compares with those rare examples of 'record-setting stabilizations' [4, 24-26]. 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 improved thermostability of ligase 10C is likely due to improved intermolecular contacts within the protein compared to the mesophilic ligases. However, it isn't clear from an examination of the primary sequences how the changes to 10C promote this stabilization. To examine this further, we mapped the differences between ligase #7 and 10C onto the previously solved structure of 10C (Figure 3.6). 10C and #7 were compared as they are very similar yet have large differences in thermostability. All differences between these two ligases are found in or near the structured region responsible for zinc coordination. In 10C there are two residues at the C-terminus, Ile68 and His69, which we previously established [19] make long range contacts with several residues at the N-terminus: Lys17, His18, Ala27 and Glu28. Notably, His18 is one of the zinc coordinating residues in 10C and mutating this position to Ala results in an insoluble protein [19]. In ligase #7, the residue corresponding to Ile68 is mutated to Methionine. In addition, between the residues corresponding to Ile68 and His69 in 10C, ligase #7 contains an additional 13 amino acids, which might displace His69 and its contacts to the N-terminus. Presumably these mutations can compromise these long range contacts and decrease the stability of #7 at high temperature. 10C also contains two additional mutations (S54 and D65) which may influence protein stability, but it's not immediately obvious how crucial they are to the overall improvement.

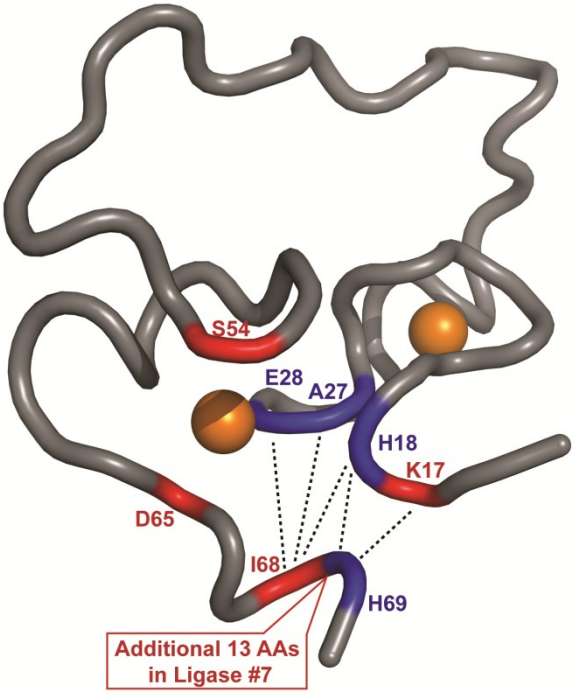


Figure 3.6 - Solved structure of ligase 10C highlighting differences between 10C and #7. We mapped the differences between ligase #7 and 10C onto the previously published structure of 10C [19] omitting the flexible termini. Mutations are shown in red, residues potentially perturbed by the mutations are shown in blue and long range NOEs are shown by dashed black lines.

Of the two ligase families we isolated in our heat-stabilization selection, the family corresponding to 10H was not solubly expressed in *E. coli*. During the mesophile selection, 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 clones like 10H and #1 might perhaps be expected as soluble expression in *E.coli* was never directly selected for in our experiments because mRNA display uses an *in vitro* eukaryotic translation system. Additionally, the fused RNA is suspected to increase protein solubility which might contribute to this result. Generally this is a favorable property of mRNA display because it can be used to characterize poorly soluble proteins. It's possible that ligase 10H could have been solubilized through MBP fusion like ligase #1, but that would have complicated structural studies. Ultimately our selection was

successful because we isolated a soluble, thermostable enzyme which we had not been able to identify previously in our library pool.

The structure of the artificial enzyme, RNA ligase 10C, does not match with any known protein folds. Considering the T_M of 72 °C for this protein, it is particularly surprising to discover the lack of secondary structural motifs combined with highly dynamic regions. While it is increasingly appreciated that catalytic activity of enzymes can require conformational flexibility [27-29], thermal stability is usually associated with tight packing and rigidity. Generally, thermophilic enzymes possess well packed hydrophobic cores [30], few exposed surface loops [31] and additional stabilizing interactions such as salt bridges [32] and a high number of hydrogen bonds [33]. 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 [34]. The structure of the ligase 10C [19] 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 appear -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 structures found in viral proteins, structural characteristic very similar to 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 [35]. 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 can be considered a model protein for primordial enzymes. For these reasons, properties of this enzyme like its evolutionary potential will be interesting to study, yet 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 [16]. 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 [20].

3.5 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 [17, 18]. 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 [36]. 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. Previous reports on mRNA display include the improvement of folding and stability of proteins by selecting for resistance to protease degradation [37], or by selecting in the presence of increasing amounts of the denaturant guanidine hydrochloride [38, 39]. 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.

3.6 Materials and Methods

3.6.1 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 [16, 21]. 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) were purchased from Dharmacon 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 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 [22] and purified by denaturing PAGE. All oligonucleotides were dissolved in ultra-pure water and concentrations determined by UV absorbance.

3.6.2 Selection of RNA Ligases at 65 °C

The mRNA display selection was performed as previously published [16], 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 [16] 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 [16]) 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 [16], and the photocleaved DNA was amplified by PCR and used as input for the following round. For round 5, the photocleaved DNA from round 4 was used that resulted from the 5 min incubation.

3.6.3 Expression & Purification of RNA Ligases

RNA ligases were expressed and purified as previously described [19].

3.6.4 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 [19]) 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.

3.6.5 Determination of Observed Rate Constants (k_{obs})

5 μM enzyme (purified by Ni-NTA affinity and size exclusion chromatography [19]) 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) giving a value in per hour (h^{-1}). The reported values are an average of 3 independent replicates ± the standard deviation. Total conversion was < 10% for all cases.

3.6.6 Circular Dichroism and Thermal Denaturation

Ligase enzymes (purified by Ni-NTA affinity and size exclusion chromatography [19]) 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.

3.7 Supporting Information

Table S3.1 - 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] Only the first 4 timepoints were used with ligase 10C as the final point had begun to plateau and would have skewed the analysis. All 5 timepoints were used for ligases 6 and 7.

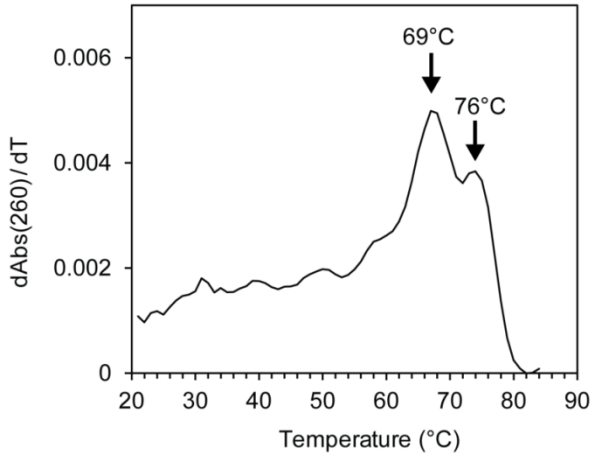


Figure S3.1 - Thermal denaturation of substrate and splint oligonucleotides used in the selection and activity assays at 65 °C. The first derivative of the melting curve for the 40 nt splint in the presence of both PPP-substrate-65 and HO-substrate-65 RNA oligonucleotides is presented. The concentration 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.

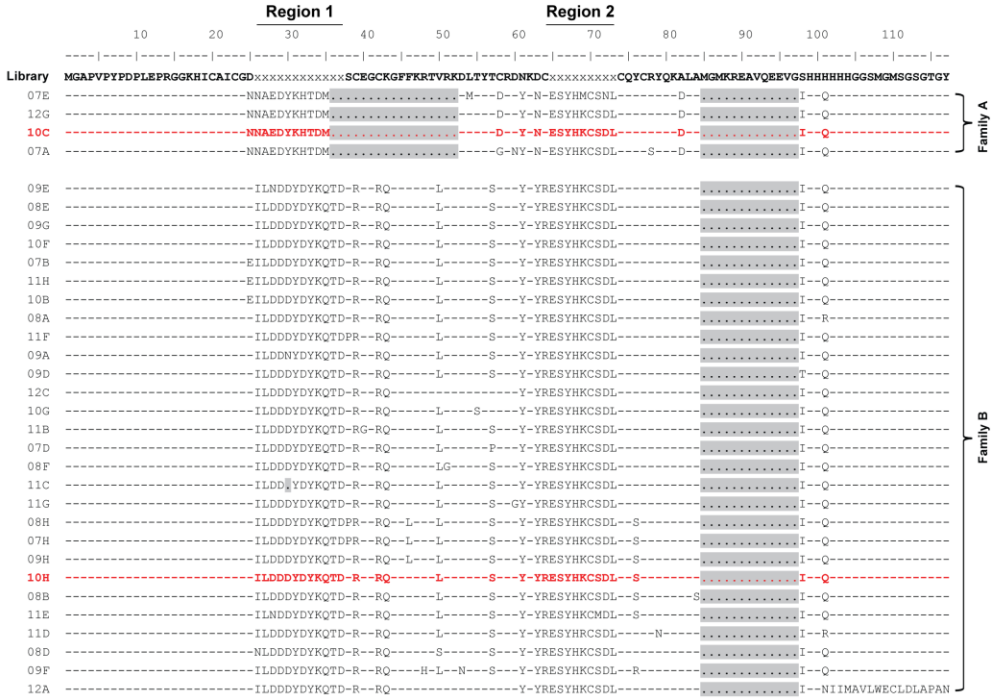


Figure S3.2 - Clones identified from round 6 of the *in vitro* selection at 65 ºC. Two protein families (A, B) were identified and a representative clone from each family was chosen for further characterization (10C and 10H, shown in red).

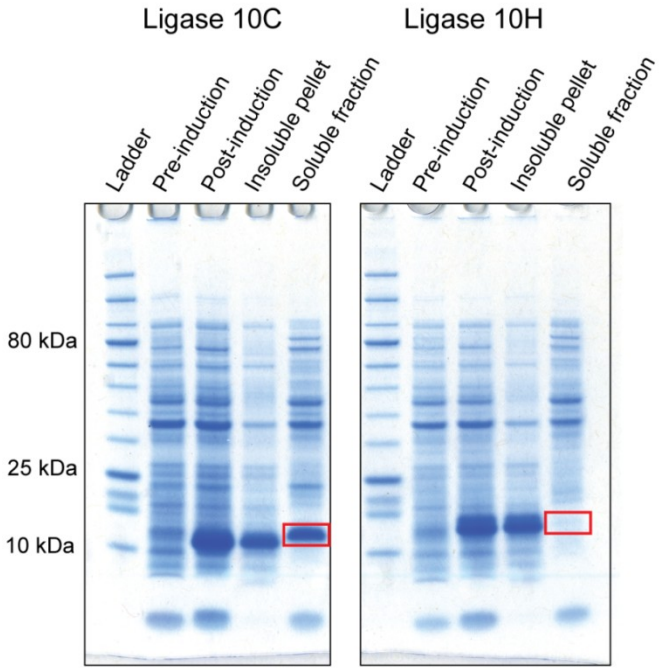


Figure S3.3 - Protein expression in *E. coli* of representative ligases selected at 65 °C. A Coomassie-stained SDS-PAGE gel shows samples of whole cells pre- and post-induction and the insoluble and soluble fractions after cell lysis and centrifugation. Red boxes in the lane "Soluble fraction" indicate the presence or absence of soluble ligases 10C and 10H, respectively.

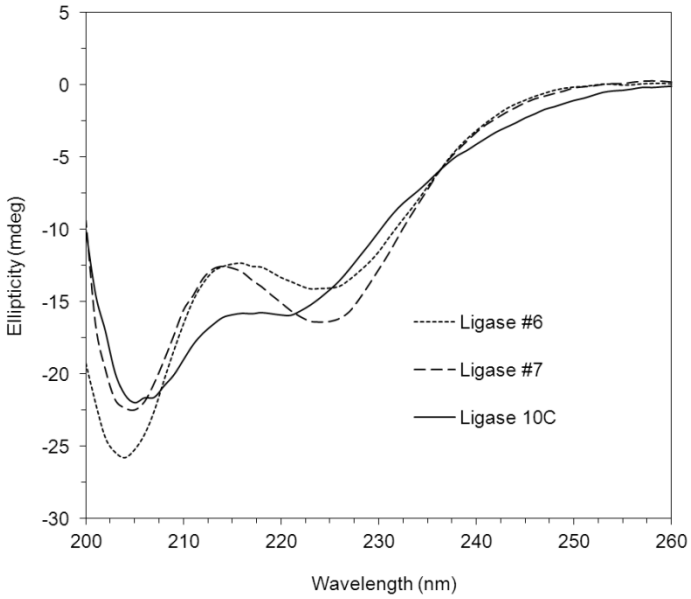


Figure S3.4 - Circular dichroism spectra of ligases #6, #7 and 10C at 25 °C.

3.7.1 Sequences of ligases selected at 65 °C

10C
MGAPVPYPDPLEPRGGKHICAICGNNAEDYKHTDMDLTYTDRDYKNCESYHKCSDLCQYCRYQKDLAIHHQ
HHHGGSMGMSGSGTGY

10H
MGAPVPYPDPLEPRGGKHICAICGDILDDDYDYKQTDSREGRQGFFKRTLRKDLTYSCRDYKYRESYHKCS
DLCQSCRYQKALAIHHQHHHGGSMGMSGSGTGY

Sequences of ligases selected previously at 23 °C for comparison
#6
MDYKDDDDKDGKHICAICGDTVTNTDYKTPDLTSTCRDYKNRESYHKCSDLCQYCRYQKALAMGTKREAAQ
EEVGSHHQHHHGGSMGMSGSGTGY

#7
MDYKDDDDKGGRHICAICGNNAEDYKHTDMDLTYTDRDYKNCESYHKCQDLCQYCRYQKALAMGIKREAVQ
EEVGSHHQHHHGGSMGMSGSGTGY

3.7.2 Oligonucleotide sequences

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)

BS76

5'-TCGTCATCGAATCGTACA AAACCAGCTAGTGAATC

BS99

5'-TCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGAGCACCAGTCCCTTACCCT
GATCCGCTGGAACCGCGTGGCGGAAAGCACATCTGC

BS24RXR2

5'-TTAATAGCCGGTGCCAGATCCAGACATTCCCATAGAACCGCCATGATGATG

Chapter 4:

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 with help from Masterson and Shi. Churchfield and I prepared ligase samples for NMR and characterized mutants by activity and CD. Morelli characterized the zinc binding and available Cys residues. Sarangi characterized the zinc coordination by EXAFS.

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

4.1 Overview

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.

4.2 Introduction

The known structures of naturally occurring proteins can be assigned to an apparently finite number of different fold families [1, 2]. 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 [3, 4]. However, the origin of those biological folds remains subject to debate [5, 6]. 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 [7-13].

Recently, we created artificial RNA ligase enzymes by *in vitro* evolution [14, 15]. 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) [16] (Figure 4.1). Two adjacent loops of this protein were randomized to generate a combinatorial library of mutants as input for the selection and evolution process [17]. 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 [14].

4.3 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 [16]. 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 4.1b). 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.

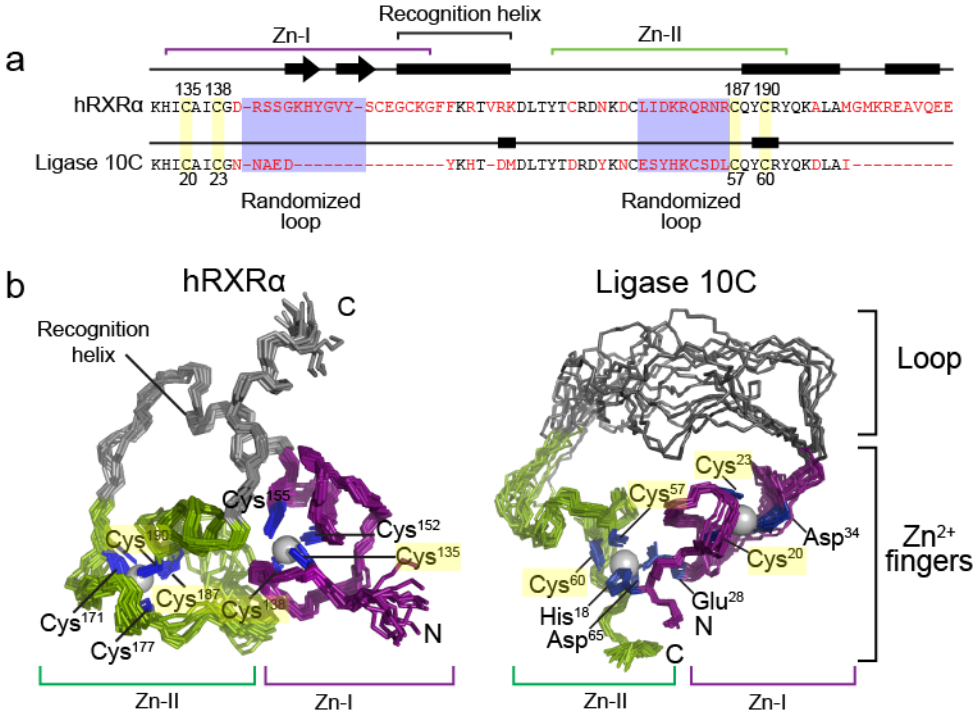


Figure 4.1 - Changes in primary sequence and three-dimensional structure upon directed evolution of the hRXRα scaffold to the ligase enzyme 10C. (a) Comparison of the primary sequences of hRXRα[16] (residues 132-208) and the artificially evolved ligase (residues 17-68). The two zinc finger regions are 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[16] 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, see also Figure 5.1A) while all other ligands differ in the two structures. In contrast to hRXRα, zinc finger Zn-II in ligase 10C comprises residues of both N- and C-terminal sequence, imposing a cyclic structure to the enzyme. Residues involved in zinc coordination are labeled and shown in blue. Note that the new ligase lost both helical domains of the DNA binding domain (grey in left structure), replacing the recognition helix with a long unstructured loop (grey in right structure).

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 4.1b). This new three-dimensional 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 [18] 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 S4.1 and Table S4.1). 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, while 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 S4.2 and S4.3).

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 S4.4 and S4.5). 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 S4.6 and Table S4.2). 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 S4.7). 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 [19]. 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 S4.2 and S4.3). EXAFS data corroborated these results, showing that both Zn sites coordinated with two S(Cys) ligands 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 S4.8).

The directed evolution process that yielded the artificial enzyme was based only on product formation without structural constraints [14]. As a result, the ligase enzyme evolved into a new structure with substantially increased conformational dynamics compared to the original DNA-binding scaffold [20]. In fact, ligase 10C showed an overall increase in structural plasticity and malleability. While the two zinc fingers exhibited heteronuclear NOEs similar to the original scaffold, the region where the two helical domains were deleted displayed much higher flexibility, with heteronuclear NOEs below 0.5. These data indicate augmentation of conformational dynamics in the picosecond to nanosecond time scale supported by longitudinal (T_1) and transverse (T_2) nuclear spin-relaxation measurements as well as hydrogen/deuterium exchange data (Figures 4.2a, S4.9 and S4.10a).

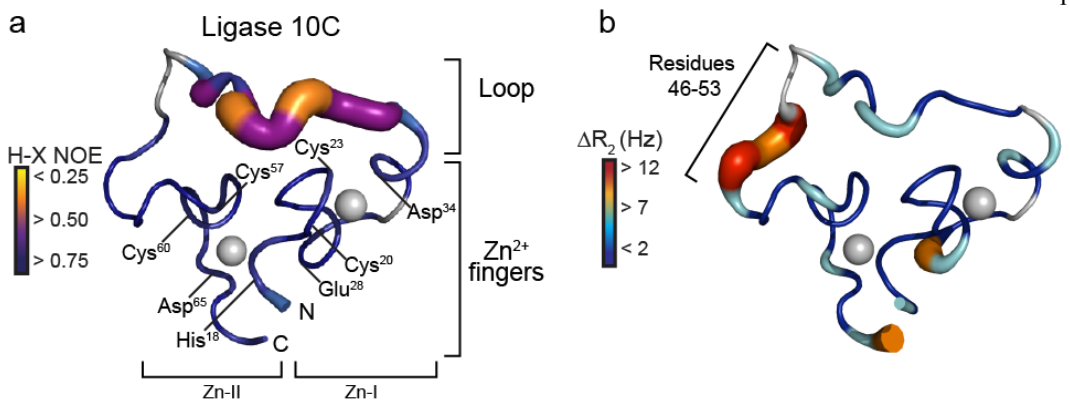


Figure 4.2 - Conformational dynamics of the ligase enzyme 10C. (a) Mapping of the heteronuclear NOEs (proxy for fast dynamics on a picosecond-nanosecond time scale) on ligase 10C. Residues involved in zinc coordination are labeled. **(b)** Mapping of the exchange rates (R_{ex}) obtained from relaxation dispersion measurements as a proxy for slow dynamics (microsecond-millisecond time scale). The color gradient and thickness of the backbone indicate that the ligase fast dynamics is located mostly in the unstructured loop, while 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.

A distinct signature of the hRXRα structure is slow (microsecond-millisecond) conformational dynamics[20] (Figure S4.10b), 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 4.2b).

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 S4.11). Chemical shift perturbation mapping of the ligase structure (Figures 4.3a and S4.12) 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 4.2b). Notably, most alanine mutations in this region decreased or completely obliterated the enzyme's activity (Figures 4.3b and 5.3c). 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 4.3d). 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 [21]. 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.

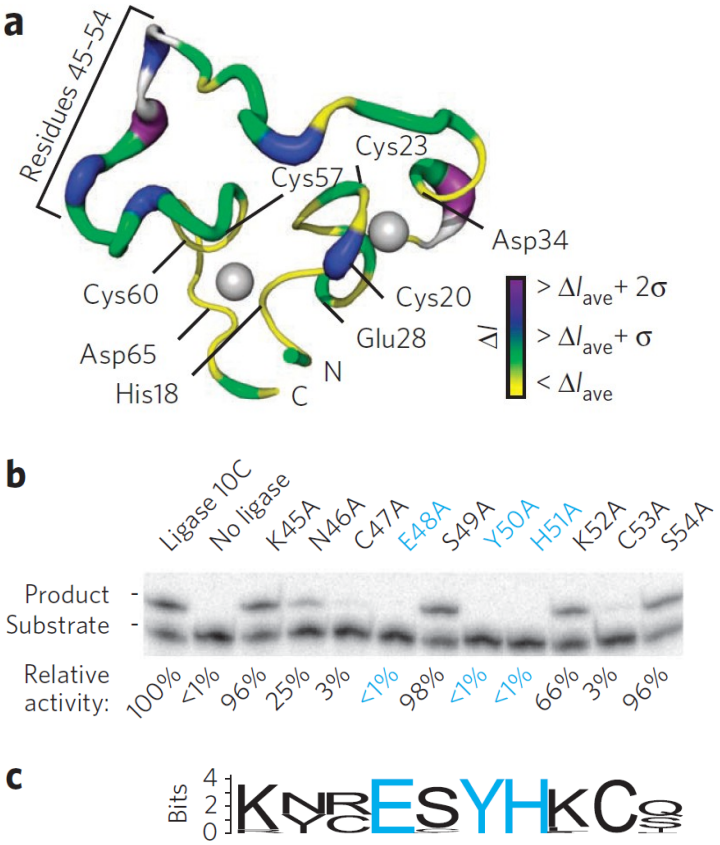


Figure 4.3 - Substrate binding surface of ligase 10C probed by NMR and alanine scanning. (a) Mapping of NMRchemical shift perturbations (intensity changes, Δ*I*) for ligase 10C shows regions affected by formation of the complex with RNAsubstrate. The most perturbed regions are indicated by thicker lines 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 [14]. 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 (C53) or had one alternative amino acid (C47). None of the other residues in this region were conserved.

4.4 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 [16]. 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 [22-24]. 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 [17]. 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. While flexibility has been suggested to increase the probability of developing new functions [5], it also reduces overall protein stability; a trade-off which enzymes must balance during evolution.

4.5 Conclusions

This report describes 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? While 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 [25].

4.6 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.

4.6.1 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.

4.6.2 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 S4.13).

4.6.3 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.

4.6.4 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 (Cys, Leu, Lys or Tyr), 600 mg of the remaining 19

unlabeled amino acids and, except for labeling ^{15}N Cys, one of the following additional amino acid supplements: 900 mg Gln, Asn and Arg when labeling ^{15}N Lys; 900 mg Val, and Ile when labeling ^{15}N Leu; and 900 mg Phe, Trp, Ala, Ser, Gly, and Cys when labeling ^{15}N Tyr. 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.

4.6.5 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
```

4.6.6 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 an OD_{600} of 0.8-1.0 at 37˚C. Cultures 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.

4.6.7 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, 20 mM TrisHCl at pH 7.5; pre-treated 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 w/ ESI PC3 Peltier cooled spray chamber, Department of Earth Sciences at the University of Minnesota).

4.6.8 Ligase activity assay for zinc dependence.

Ligase 10C was purified as reported previously [14]. Zinc was removed from ligase 10C by treatment with ion exchange resin (Chelex 100, Bio-Rad). 5 μM Ligase 10C was incubated with 20 μM HO-substrate, 10 μM ^{32}P-labeled PPP-substrate/splint, 20 mM HEPES (pH 7.5), 150 mM NaCl, 500 μM β-mercaptoethanol, and the indicated concentrations of $ZnCl_2$ for 6 h at room temperature. The ligation reactions were quenched with 20 mM EDTA/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).

4.6.9 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).

4.6.10 Resonance assignment.

All NMR spectra were acquired at 298 K on a Bruker spectrometer equipped with cryoprobe at 700 MHz and Varian spectrometer at 600 MHz. The samples were in buffer of 150 mM NaCl, 20 mM HEPES, 10 mM β-mercaptoethanol, and pH 7.5. Moreover, all protein samples were saturated with $ZnCl_2$ by observing changes in HSQC spectra prior to other NMR experiments. Triple resonance spectra such as CBCA(CO)NH, HNCACB [26-28] were used to assign peaks on ^{15}N-HSQC. All resonances in these two 3D spectra and ^{15}N-HSQC were picked and fed into the PISTACHIO program (National Magnetic Resonance Facility in Madison, WI, USA) [29] 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[30]; side-chain protons were assigned by ^{15}N-NOESY-HSQC, ^{15}N-TOCSY-HSQC, and HC(CO)NH-TOCSY experiments [30] with 150 ms mixing time, 60 ms mixing time, and 12 ms mixing time, respectively.

4.6.11 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 [31], whose helix propensity was shown by chemical shift index and $^{3}J_{\mathrm{HNH\alpha}}$ coupling values [31, 32]. 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, 3.5, 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, mis-calibrated NOEs were adjusted and then ambiguously assigned NOEs were gradually added into the restraint table during iterative calculation.

4.6.12 Torsion angle restraints.

Backbone Phi angle restraints were acquired from the HNHA experiment, and the quantitative $^{3}J_{HNH\alpha}$ coupling values were calculated from the intensity ratios of cross peaks to diagonal peaks and corrected by 3.7% to account for relaxation [33]. The correction is proportional to the rotational correlation time of the protein (3 ns), which was measured from 1-dimensional TRACT experiment[34]. The Phi 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 Psi 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 Psi(i-1) is restrained from 20° to 220°; otherwise, the Psi(i-1) is restrained from 80° to -140° [35, 36].

4.6.13 RDC measurement.

The stability of several alignment media for ligase 10C was tested. 5% neutral and negative-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 CPCl (cetylpyridinium chloride) 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% alkyl-poly(ethylene glycol)) and 1-hexanol (r=0.85)[37]. 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.

4.6.14 Structure calculations.

Simulating annealing protocols were performed in the XPLOR package [38]. 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 [39, 40], which are in quantitative agreement with the EXAFS experiments. Distances derived from EXAFS have previously been used as restraints in NMR refinement [41, 42]. 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.

4.6.15 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. 20% v/v) glycerol was added to the protein samples in order 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. Other optical components used for the experiments were 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 [43]. Theoretical EXAFS signals $\chi(k)$ were calculated by using *FEFF* (Macintosh version 8.4) [44-46]. Initial model was based on the $Zn(Cys)_2(His)_2$ active site in a zinc finger

protein (1MEY)[47]. Based on 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 [48]. 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 non-structural 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.

4.6.16 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.

4.7 Supplementary Information

Table S4.1 - 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 S4.2 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 S4.3 - 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 S5.8 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.

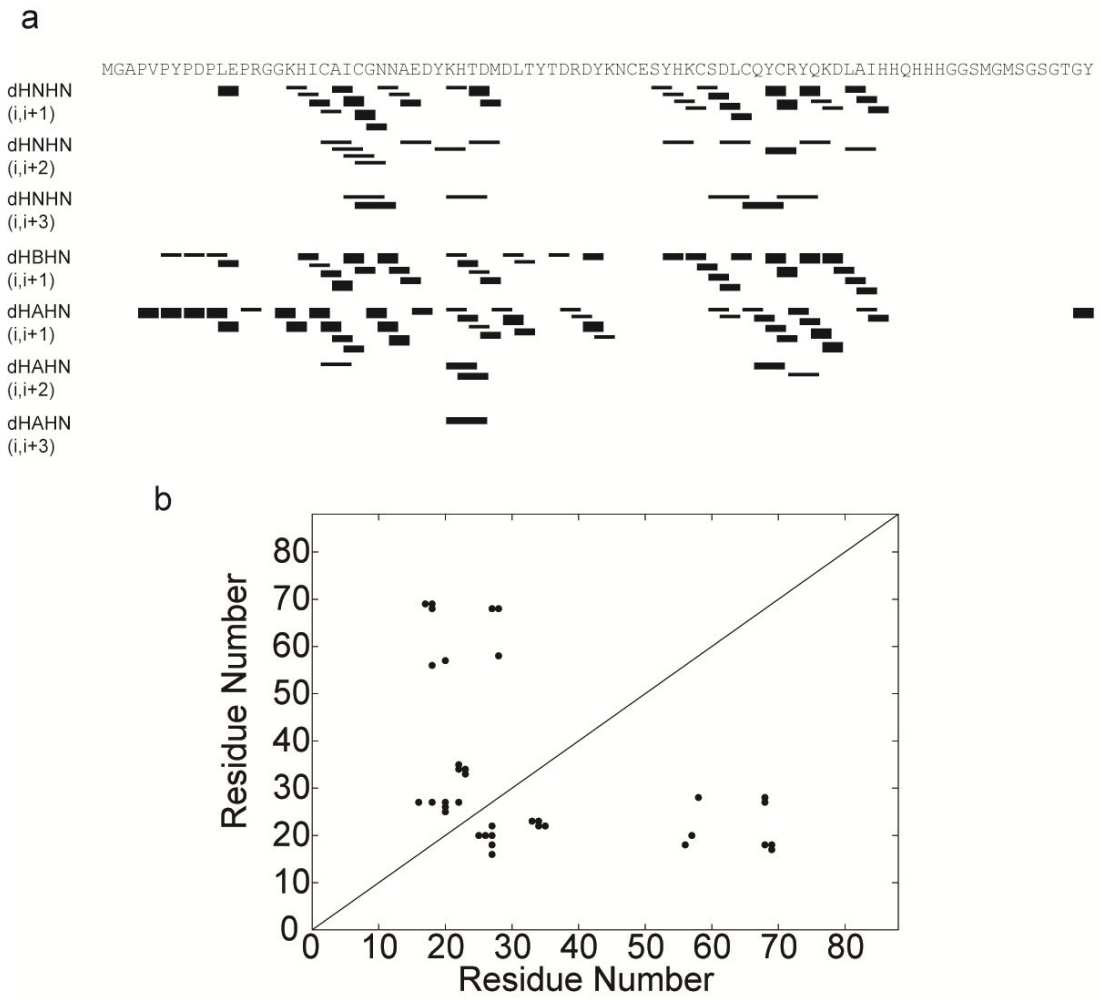


Figure S4.1 - Summary of the NOEs observed from NOESY spectra. (a) Horizontal bars show the presence of NOE signals between residues. Bar thickness corresponds to the NOE intensity. **(b)** Long-range NOEs observed in ligase 10C.

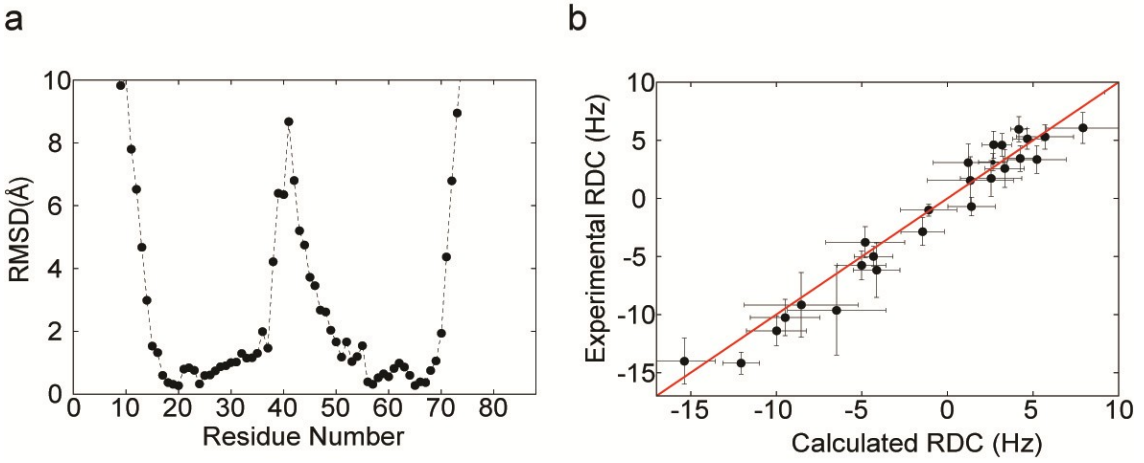


Figure S4.2 - Convergence of the structural ensemble of 20 conformers. (a) Average backbone RMSD of the conformational ensemble. **(b)** Correlation between experimental RDC values and average back-calculated RDC values from the ensemble.

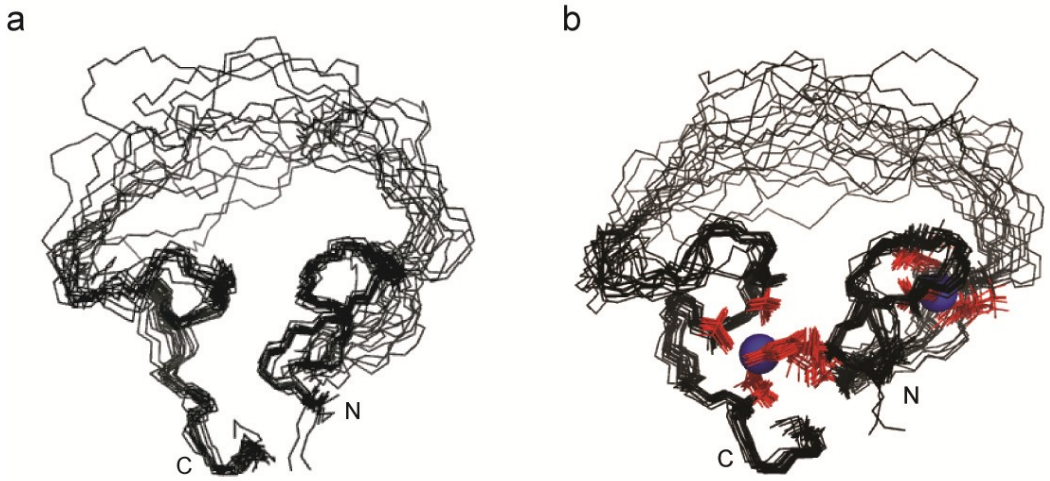


Figure S4.3 - The structural ensembles are calculated before and after incorporating Zn^{2+} ions into the coordinates. (a) Ensemble of 20 lowest energy conformers (residues 17-69) selected from 100 structures. The NOEs, torsion angles, and RDC values were included in these calculations. Note that the 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.

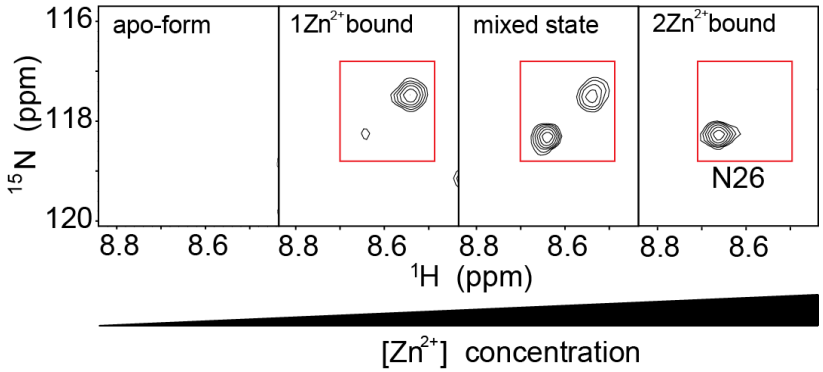


Figure S4.4 - Zn^{2+} titration into ^{15}N-labeled ligase 10C monitored by NMR. A selected region of HSQC spectra recorded during Zn^{2+} titration is shown. Residues of partially Zn^{2+}-saturated sample displayed slow exchange on the NMR time scale between forms bound to one or two Zn^{2+}.

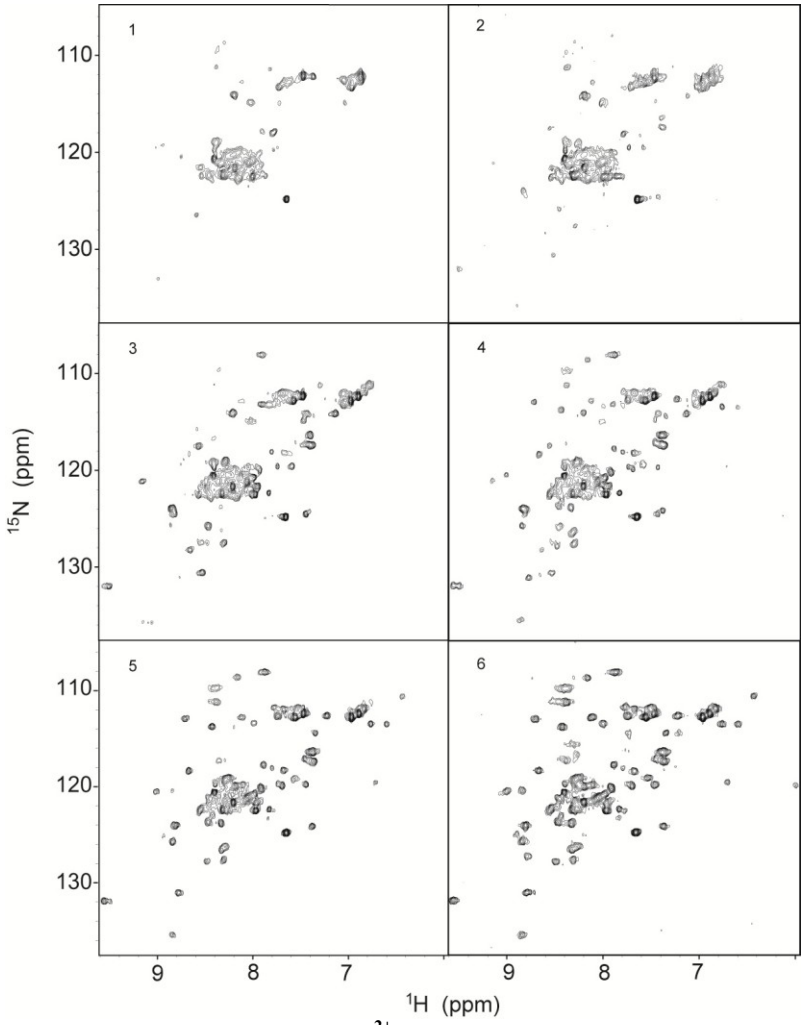


Figure S4.5 - HSQC spectra recorded upon Zn^{2+} titration. Molar ratios of ligase 10C to zinc were: 1) 10C:Zn=1:0, 2) 10C:Zn=1:1, 3) 10C:Zn=1:2, 4) 10C:Zn=1:3, 5) 10C:Zn=1:4, 6) 10C:Zn=1:6.

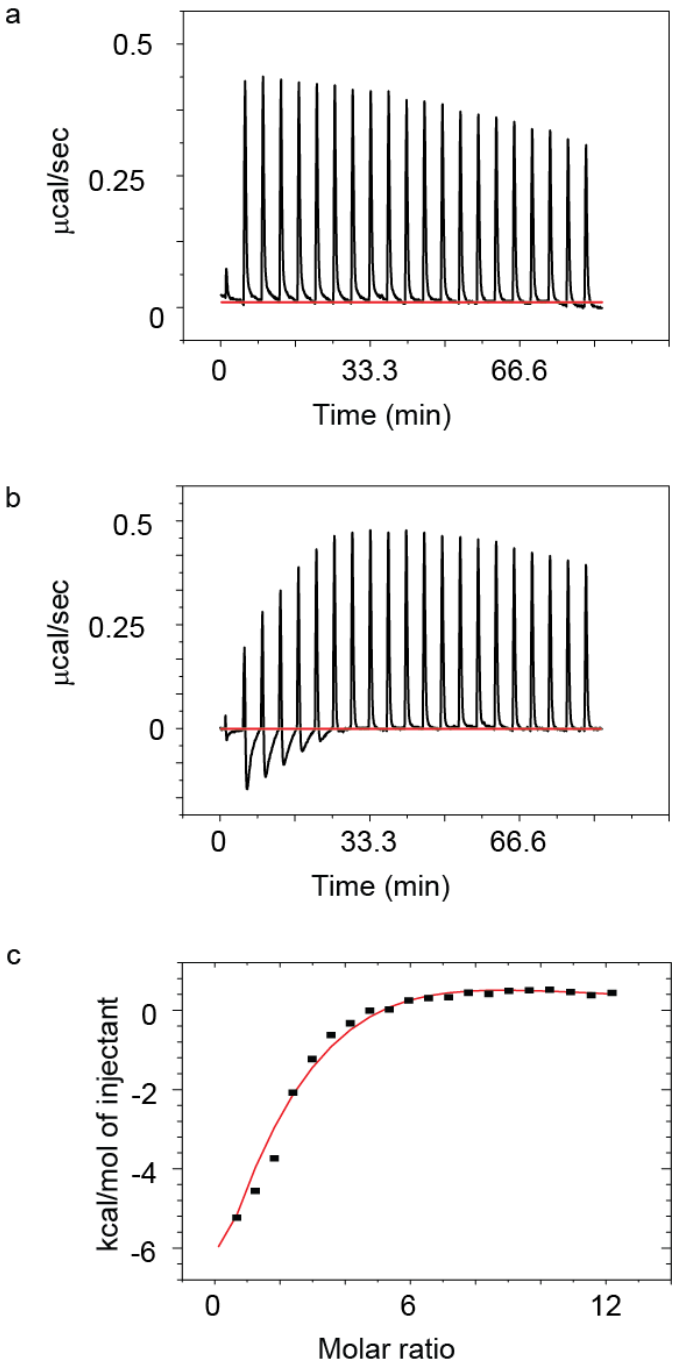


Figure S4.6 - Zn^{2+} titration into the ligase enzyme monitored by ITC. Samples contained 5 µM ligase 10C, 150 mM NaCl, 20 mM HEPES, 10 mM β-mercaptoethanol, pH 7.5 and were measured by Isothermal Titration Calorimetry using a MicroCal VP-ITC instrument (GE Healthcare). **(a)** The graph represents 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 S4.2.

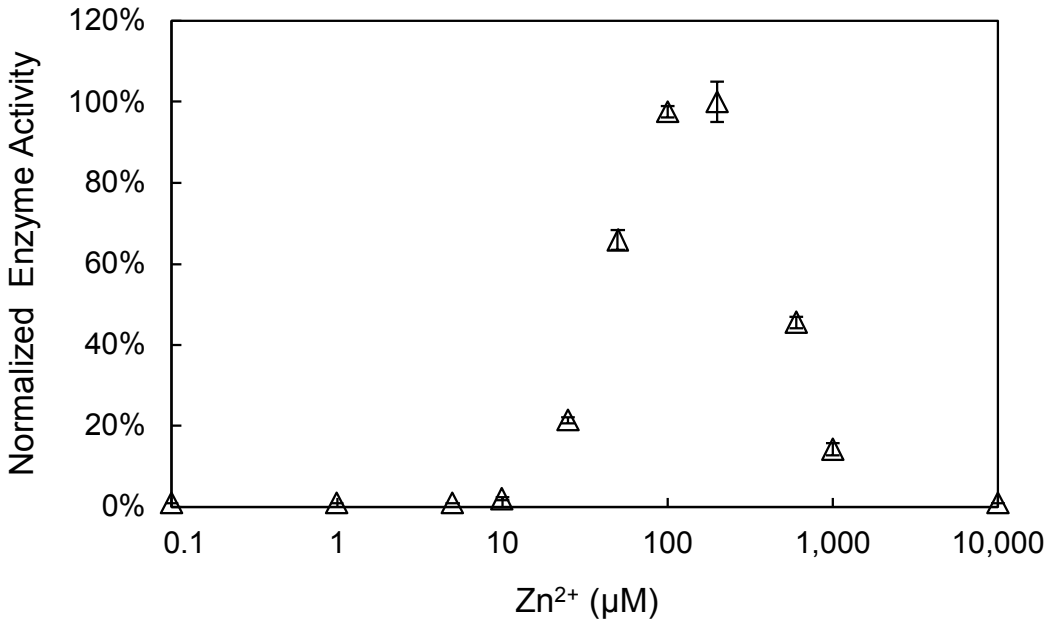


Figure S4.7 - Zn^{2+} dependence of ligase activity. The maximum activity was observed at 145 µM Zn^{2+}. Towards lower Zn^{2+} concentrations the activity sharply drops, matching the expected behavior predicted from the dissociation constants measured by Thermocalorimetry. Towards higher Zn^{2+} concentrations, 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%.

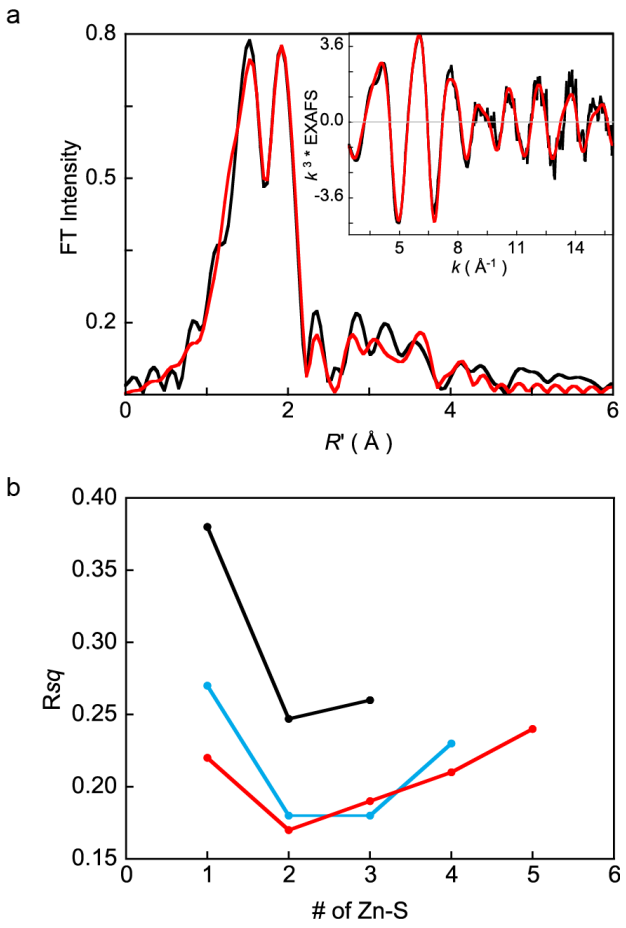


Figure S4.8 - Analysis of zinc coordination by EXAFS spectroscopy (Extended X-ray Absorption Fine Structure). (a) The k^3 weighted Zn K-edge EXAFS (inset) and their corresponding non-phase shift corrected Fourier transforms for ligase 10C are presented. The experimental data are shown as black lines and the fit as red lines. The best-fit parameters are given in Table S5.3**.** 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[39, 49]. 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[50] or ZnS(Cys)n[51] 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[49]. EXAFS studies with cysteine ligands are typically limited to four- and five-coordinate sites[52]. Studies have been performed on six-coordinate sites, but typically with all light atom ligands[53]. 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[52]. 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.

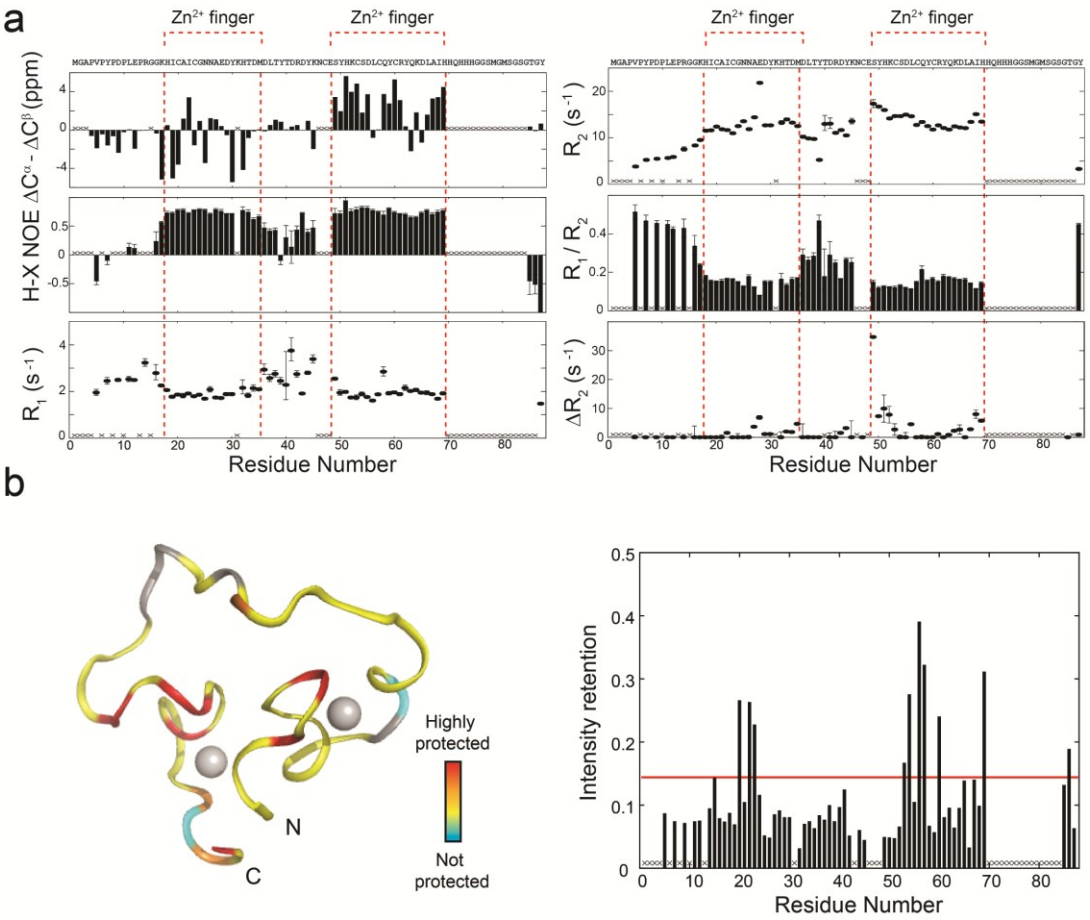


Figure S4.9 - Structure and conformational dynamics probed by NMR experiments. (a) Chemical shift indexes (ΔCα-ΔCβ), steady-state NOE, longitudinal relaxation rates (R_1), transverse relaxation rates (R_2), and R_1/R_2 ratios as determined by NMR spectroscopy (unassigned residues are marked with an "X"). 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.

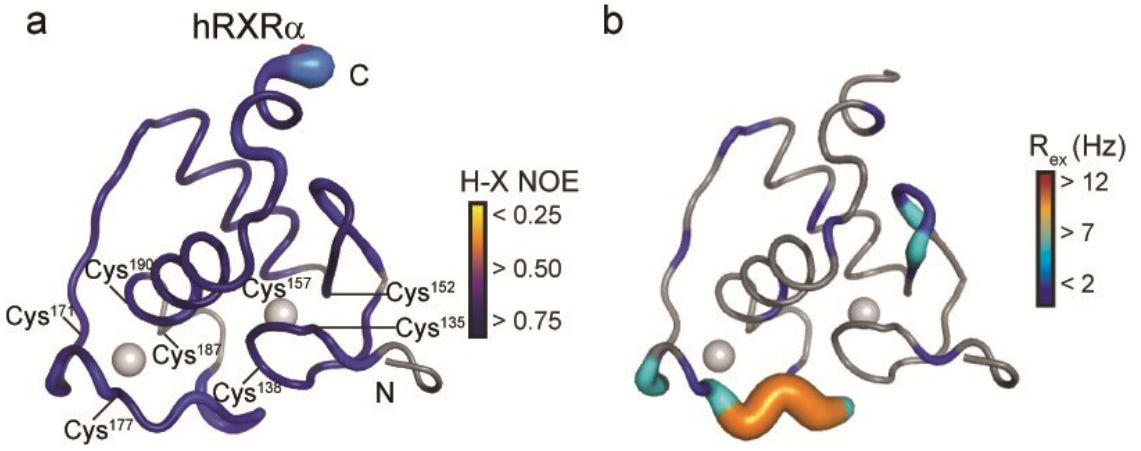


Figure S4.10 - Mapping of the conformational dynamics of the DNA binding domain (hRXRα) [16]. **(a)** Heteronuclear NOEs (proxy for fast dynamics on a picosecond-nanosecond time scale) on the structure of hRXRα[16]. **(b)** Exchange rates (R_{ex}) obtained from relaxation dispersion measurements as a proxy for slow dynamics (microsecond-millisecond time scale). The color gradient and the thickness of the backbone indicate the intensity of the motions.

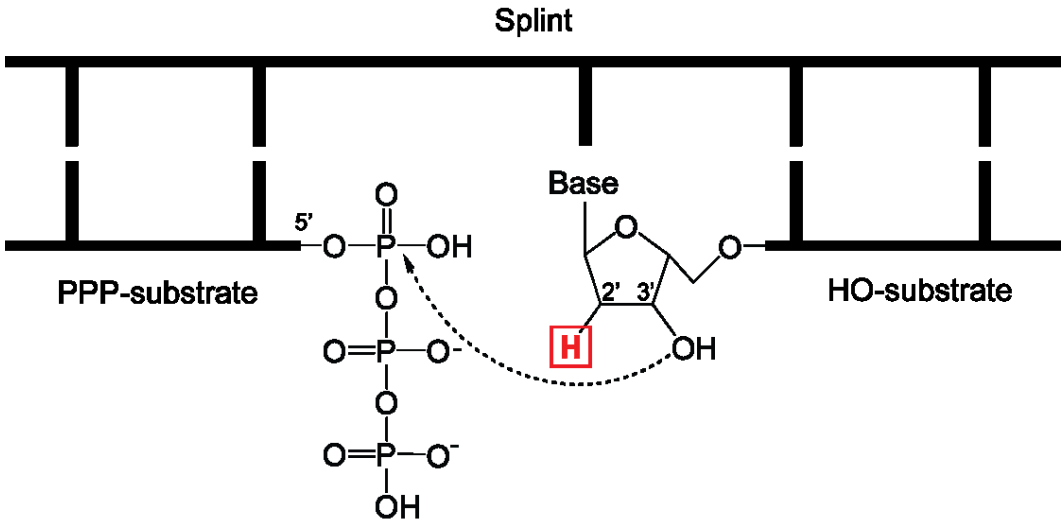


Figure S4.11 - Chemical structure of inactive ligation substrate. Substitution of the 2'-hydroxyl group of the terminal nucleotide in the HO-substrate with a 2'-deoxy modification (red box) results in inactivation of the ligation reaction. Ligation of active substrates occurs between a 5'-triphosphorylated 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.

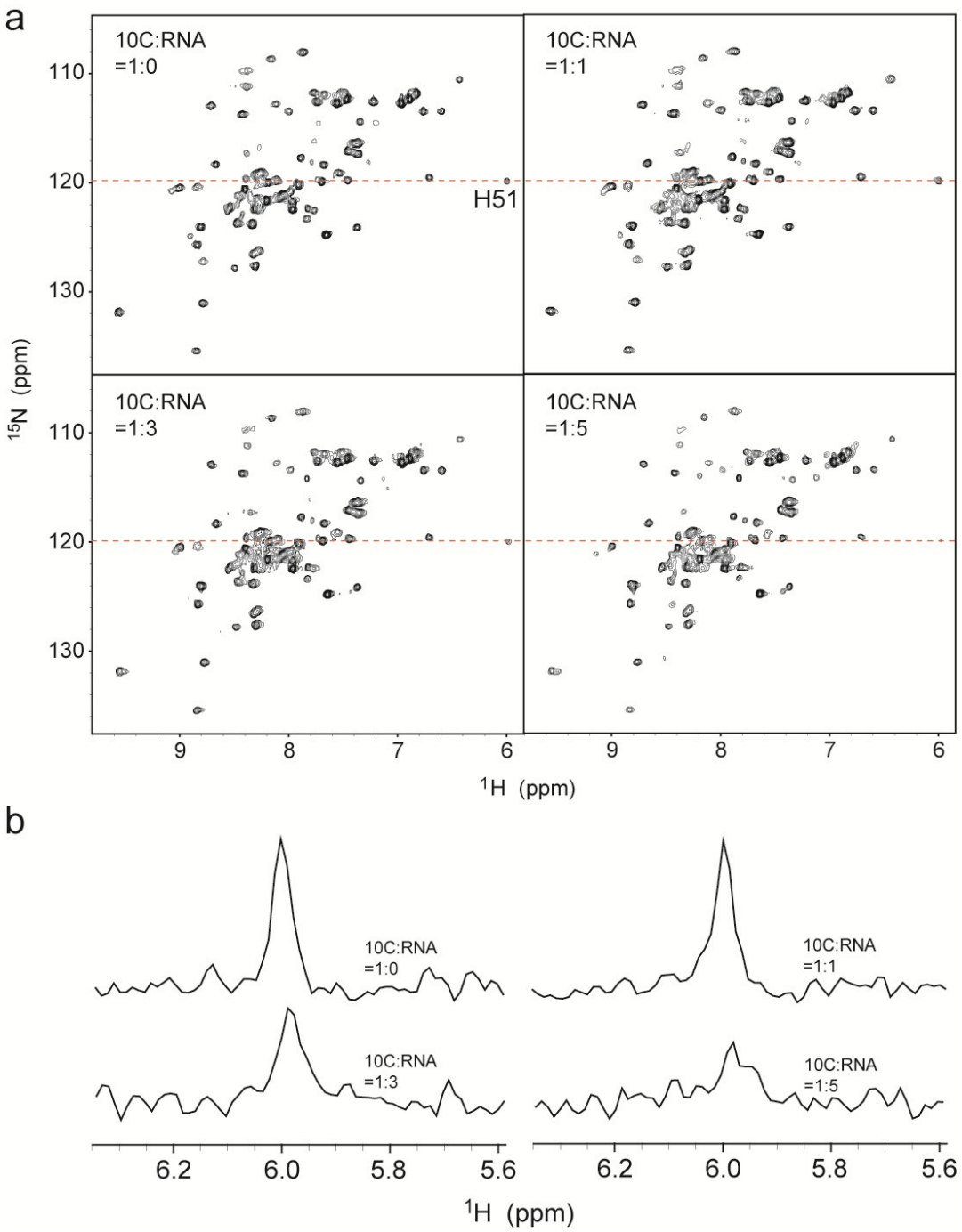


Figure S4.12 - Titration of RNA substrate into ligase 10C monitored by NMR spectroscopy. (a) The ligase enzyme (300 μM) was titrated with the inactive RNA ligand in 150 mM NaCl, 20 mM HEPES, 10 mM β-mercaptoethanol, and pH 7.5. The HSQC spectra during the titration showed no significant changes in chemical shifts. **(b)** Slices of a selected peak (H51) in HSQC spectra during ligand titration showed significant line-broadening.

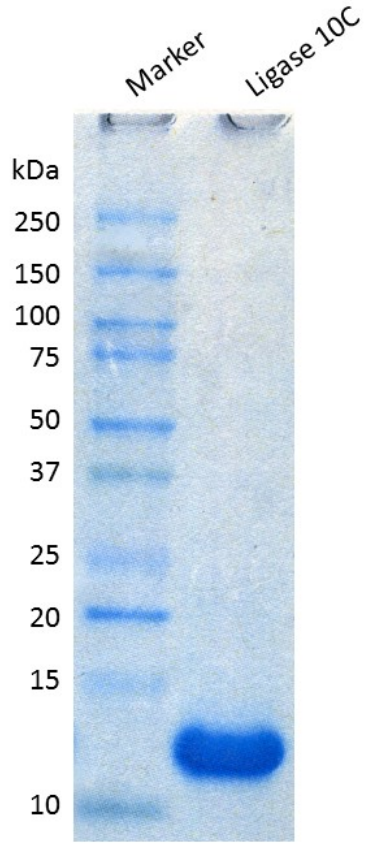


Figure S4.13 - Purity and identity of purified ligase 10C. SDS-PAGE gel (NuPAGE 4-12% Bis-Tris gel, Invitrogen) Coomassie stained of ligase 10C purified by nickel affinity chromatography and size exclusion chromatography and 10-250 kDa ladder P7703S (New England Biolabs) used as a marker. The identity 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 5:

Universal labeling of 5'-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity.

The following is a reprint of the article: Haugner III, J. C., and Seelig, B. (2013) Universal labeling of 5'-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity. *Chem. Commun.* **49,** 7322-24. The article is reproduced by permission of The Royal Society of Chemistry. Seelig and I designed all experiments and interpretated data. I performed all experiments.

Hyperlink to original publication
http://pubs.rsc.org/en/content/articlelanding/2013/cc/c3cc44454f#!divAbstract

5.1 Overview

An artificial RNA ligase specific to RNA with a 5′-triphosphate (PPP-RNA) exhibits broad sequence specificity on model substrates and secondary siRNAs with direct applications in the identification of PPP-RNAs through sequencing.

5.2 Introduction

The development of novel enzymes to manipulate nucleic acids has been a driving force in pioneering methods for molecular biology, genomics and transcriptomics. One such type of enzymes are ligases, which join together two strands of nucleic acids by forming a phosphodiester bond with a 3′-5′ linkage. Many known natural ligases couple the phosphodiester bond formation with ATP hydrolysis through a series of group transfers. [1, 2] Ligases have been used extensively in the analysis of large mixtures of RNA to add adapters of known sequence to the RNAs [3, 4], thus enabling reverse transcription, PCR amplification and ultimately their identification by microarray analysis or sequencing. [5] As all known natural ligase enzymes require a 5′-monophosphate terminus (P-RNA) for the ligation reaction to occur, RNA mixtures are

typically enzymatically modified to uniformly introduce the 5′-monophosphate in preparation for ligation and sequencing. This can prove problematic for interpreting subsequent RNA sequencing data as classes of RNA characterized by a different 5′-phosphorylation can no longer be easily distinguished in sequencing data. Identifying primary transcripts, also called immature RNA, can be particularly difficult because they have a 5′-triphosphate (PPP-RNA) and cannot be selectively purified from other classes of RNAs. While the triphosphate of eukaryotic mRNA is eventually capped, other classes of RNA retain the 5′-triphosphate as one of their only identifying features. This includes prokaryote mRNA, some viral mRNAs in eukaryotes, [6] transcripts from mitochondria [7] and chloroplasts, [8] and secondary siRNAs. [9, 10] To identify these RNAs from cell isolates, methods exist which enrich for PPP-RNAs prior to sequencing, but they commonly identify false positives and require laborious independent verification with a complementary method. [11] Therefore, an alternative, simplified method for identifying PPP-RNAs with improved accuracy would be highly desirable.

The splinted, 5′-triphosphate-dependent ligation of two RNA strands has repeatedly been used as a model reaction to generate novel catalysts. Both artificial ribozymes [12-14] and deoxyribozymes [15, 16] have been generated that catalyze this reaction. However, these catalytic nucleic acids often have strict sequence requirements near the ligation site, limiting the choice of PPP-RNA sequences that can be ligated. Furthermore, these nucleic acid enzymes act as both catalyst and complementary splint, which requires a new catalyst to be designed and synthesized specifically for each new substrate sequence. Previously, we reported the *de novo* selection and evolution of triphosphate dependent artificial RNA ligases. [17, 18] These protein enzymes ligate a 5′-triphosphate RNA substrate to the 3′-hydroxyl of a second RNA exclusively forming a 3′-5′ linkage while both substrates are aligned by a complementary oligonucleotide splint (Figure 5.1). The artificial ligases are zinc-dependent metalloproteins of ~10 kDa and were derived from the two zinc finger protein hRXRα. [17] We solved the structure of RNA ligase 10C, a highly active variant, revealing a novel protein fold with highly unusual structural features. [19]

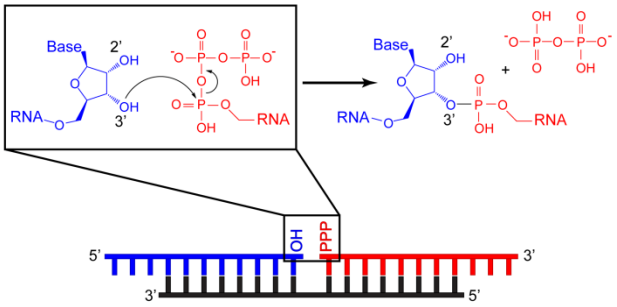


Figure 5.1 - Artificial RNA ligase catalyzing the formation of a 3′-5′-phosphodiester bond between a 5′-triphosphate RNA (red) and 3′-hydroxyl RNA (blue) substrate. The proposed mechanism of the reaction is shown. A complementary splint DNA oligonucleotide (black) anneals to the RNA and is necessary for ligation.

To determine if this enzyme could be used for the general identification of PPP-RNAs, we investigated the sequence specificity of ligase 10C. This artificial enzyme had been generated to ligate a single substrate pair, but had never been challenged to evolve an ability to ligate alternative RNA sequences. To determine the sequence specificity of 10C, we varied the nucleotides adjacent to the ligation site. We also tested a substrate pair of entirely unrelated sequence to assess the impact of large scale changes in sequence. Finally, to extend 10C for practical applications, we investigated whether this ligase can perform selective ligation in a mixture of secondary siRNAs. These experiments demonstrate RNA ligase 10C exhibits broad sequence specificity and therefore should be a valuable tool for the whole-cell isolation and identification of PPP-RNA.

5.3 Results

RNA ligase 10C was assayed for the ligation of chemically synthesized RNA oligonucleotides to PPP-RNA substrates synthesized and ^{32}P-labeled by *in vitro* transcription. Assay conditions were chosen to mimic conditions used in commercial RNA isolation and sequencing kits. [20, 21] Under these conditions, 10C was able to ligate the original substrates used during its evolution as well as an unrelated sequence, even though the nucleotides flanking the ligation site were completely different (Figure 5.2). Ligation rates were similar for these two substrates with the k_{obs} of 0.45 ± 0.01 h^{-1}

for the original substrates and 0.24 ± 0.01 h^{-1} for the substrate pair with unrelated sequence. We expanded upon these experiments by testing different nucleotides at the 3′-hydroxyl and 5′-triphosphate sides of the ligation site independently (Table S5.1). As we were limited to the transcription by T7 RNA polymerase, we only created PPP-RNA substrates beginning with a G or A. [22] For all substrates tested, the k_{obs} were similar within a 10-fold range (Figure S5.1). By comparison, the most commonly used natural enzyme T4 DNA ligase has been shown to have an approximately 3-fold difference in k_{cat} in the ligation of nicked DNA depending on the sequence surrounding the ligation site. [23] If the ligation using 10C is allowed to proceed overnight as is commonly done with commercial ligases, the differences in yield for different substrates are reduced to < 3-fold. Although the ligase 10C was originally selected to ligate only one specific pair of substrate sequences, the enzyme does not require any specific nucleotide sequence near the ligation site.

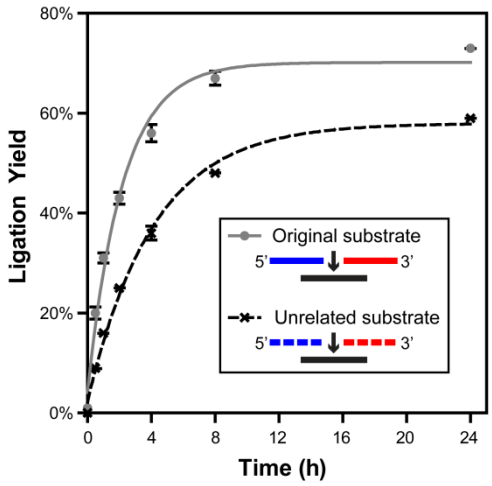


Figure 5.2 - Ligation of two pairs of RNA substrates by artificial ligase 10C. The arrow represents the site of ligation. The data were fit to first-order kinetics $Y = Y_f (1\text{-}e^{-kt})$ where Y_f = final yield and $k = k_{obs}$. The solid lines show the progress of the reaction with the substrates that were originally used to generate ligase 10C (5′ CUAACGUUCGC↓GGAGACUCUUU). Dashed lines indicate the ligation of a second substrate pair of unrelated sequence (5′ GCAUGUCAGCA↓AGGCCUAUCAA) that has the same GC content. Data are the mean of 3 replicates ± SD.

The ligation rate of ligase 10C is comparable to previously reported deoxyribozymes [15] which catalyze the same reaction under similar reaction conditions, but is lower than the typical rate observed for natural ligase protein enzymes. [24] Nevertheless, the current activity of 10C is sufficient for immediate applications, and

engineering efforts to improve the catalytic efficiency of 10C are underway. In contrast to previously generated nucleic acid enzymes, [12-15] ligase 10C can ligate all different RNA sequences without the need to tailor-make a modified catalyst for each substrate because the complementary splint is not part of, or attached to, the catalyst.

To directly demonstrate the utility of 10C, we tested the ligase with two secondary short interfering RNAs (siRNAs). siRNAs are involved in the regulation of gene expression through the RNAi pathway. Secondary siRNAs are a unique class of siRNA that were identified in *C. elegans* and have a 5′-triphosphate as they are synthesized by an RNA-dependent RNA polymerase to amplify the RNAi response. [9, 10, 25] We chose two previously published siRNAs (S2, S3 in Figure 5.3a) [9] which we transcribed with T7 polymerase. Those siRNAs lack secondary structure that could interfere in base pairing to the complementary DNA splint. We also included the original PPP-RNA substrate (S1) used to evolve ligase 10C as it shares a similar 5′-sequence with the secondary siRNAs. As expected, ligase 10C was capable of ligating 5′-triphosphorylated versions of all three substrates to an adapter sequence 8 nucleotides in length. While the 8 nt adapter was chosen to facilitate separation of RNAs in the gel, longer adapters of constant sequence can be used to enable subsequent RNA sequencing. Substrates of the same sequence but with a 5′-monophosphate instead of the 5′-triphosphate were not ligated (Figure 5.3b).

We then combined the three 5′-triphosphorylated substrates in a single reaction and tested if individual splints could specifically promote the ligation of their complementary substrates in a mixture of similar sequences. No cross-reactivity between splints was detected, despite sequence identity of the substrates in three of the four nucleotides adjacent to the ligation site (Figure 5.3b). The discrimination between substrates with a complementary splint and those with imperfect complementarity is >100-fold, using the limit of detection of the assay as a conservative estimate. Calculated melting temperatures (T_m) of the splint/substrate combinations indicate that only the matching sequence pairs will hybridize at the temperature and ionic strength of the ligation mixture (Table S5.2). By using splints with relatively short overhangs (e.g. 8 base pairs), the hybridization stringency increases dramatically as even single

mismatches can reduce the T_m by more than 10 °C. Therefore, the ligation of sequencing adapters to biologically relevant RNAs by ligase 10C can be controlled through the use of a splint which specifically base pairs to the desired PPP-RNA substrate.

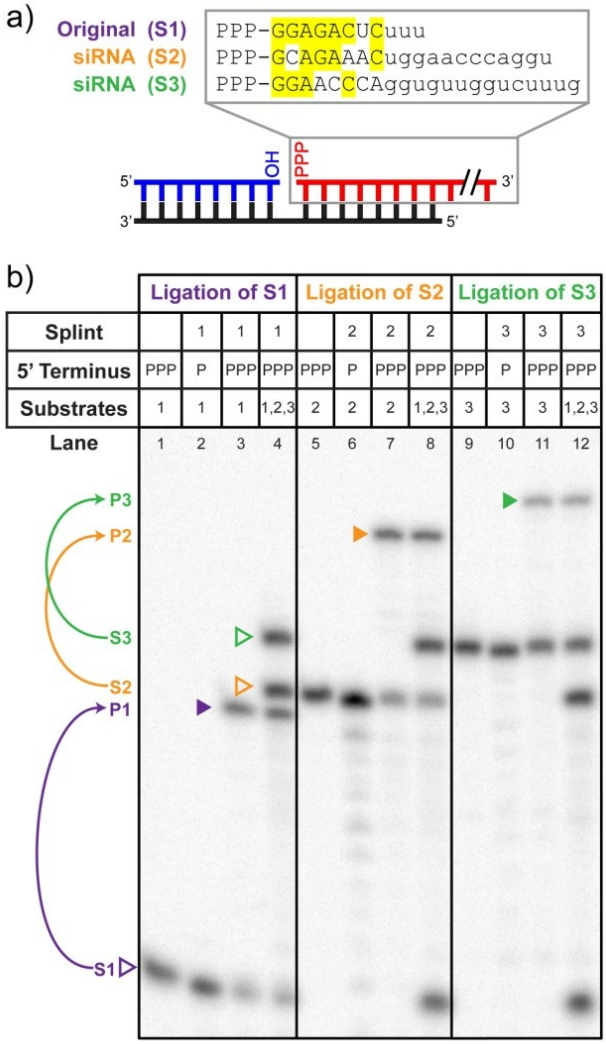


Figure 5.3 - Application of the artificial RNA ligase enzyme to selectively ligate secondary siRNA. (a) Alignment of the original ligation substrate (S1) and secondary siRNA substrates (S2) and (S3). Sequence similarities are highlighted in yellow. For each of the three substrates, a matching complementary splint oligodeoxynucleotide (black) was used to ligate them to a 3′-hydroxy-substrate "adapter". The adapter is identical in all three cases (blue). The first eight nucleotides of each substrate base-pair to its respective splint and are capitalized. **(b)** Denaturing PAGE gel depicting ligation experiments with substrates S1 to S3 to form products P1 to P3, demonstrating the necessity of a triphosphate (PPP) at the 5′-terminus (lanes 3, 7, 11). No ligation is observed for substrates with a 5′-monophosphate (P) (lanes 2, 6, 10). Lanes 4, 8 and 12 show the selective ligation of a single substrate in a mixture of substrates with highly similar sequences, which is dependent on the presence of the correct complementary splint. The substrate bands are marked by empty triangles and the product bands by filled triangles.

5.4 Discussion

The artificial RNA ligase enzyme 10C has a potential application in the sequencing of PPP-RNAs (Figure S5.2). Deep RNA sequencing (RNA-seq) is replacing traditional microarray technology for large scale RNA analysis in part because of its increased dynamic range and lower detection limit. RNA-seq has been used to characterize specific RNA subpopulations, including those which are of low cellular abundance. [5] One of the primary challenges in the analysis of specific RNA classes is the minimization of false positive results due to sample contamination with other classes of RNA or cleavage products. [5] For PPP-RNAs, ligase 10C offers a simple and direct route for the ligation of adapters without false positives as the triphosphate provides the energy needed to drive bond formation. We envision that, in addition to sequencing specific sequences, total PPP-RNA could be analyzed by using oligonucleotide splints with a short degenerate sequence overhang. This would allow for ligation of any RNA sequence and has been used previously for RNA-seq in a similar fashion. [20]

5.5 Conclusions

In summary, these results highlight the value of the artificial RNA ligase 10C for the direct isolation and identification of virtually any PPP-RNA. Efforts are underway to utilize ligase 10C in RNA sequencing efforts to evaluate the efficacy of 10C to identify a broad range of PPP-RNAs from biological samples. Finally, this exemplary application of the artificial RNA ligase highlights the potential of *in vitro* selection and evolution methods for the creation of useful artificial biocatalysts. [26, 27]

5.6 Materials and Methods

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

5.6.1 Expression & Purification of RNA Ligase 10C:

RNA Ligase 10C was expressed and purified as previously published. [19]

5.6.2 Preparation of Oligonucleotides:

The α-^{32}P-labeled PPP-RNA substrates were prepared by *in vitro* transcription using T7 RNA polymerase as previously published. [19] The inactive α-^{32}P-labeled P-RNA substrates were prepared by treating PPP-RNA with 5′ RNA Polyphosphatase from Epicentre followed by phenol/chloroform extraction or PAGE gel purification to remove the Polyphosphatase. The RNA-OH substrates were purchased from Dharmacon and prepared according to the manufacturer's protocol. DNA splints were purchased from Integrated DNA Technologies. All oligonucleotides were dissolved in ultra-pure water and concentrations determined by UV absorbance.

5.6.3 Ligation Assay:

1 μM PPP-RNA, 3 μM RNA-OH and 6 μM DNA splint (Table S5.1) were combined in a buffer containing 20 mM HEPES pH 7.5, 100 mM NaCl, 100 μM $ZnCl_2$. The oligonucleotides were annealed by heating the solution to 60 °C for 3 minutes and allowing it to cool at room temperature for 10 min. A stock of 50 μM RNA Ligase 10C in buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 100 μM $ZnCl_2$ and 0.5 mM β-mercaptoethanol was added to the oligonucleotide mix to a final concentration of 10 μM enzyme. The ligation reactions were incubated at room temperature for the indicated times, and quenched with two volumes of 20 mM EDTA in 8 M urea, 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).

5.7 Supporting Information

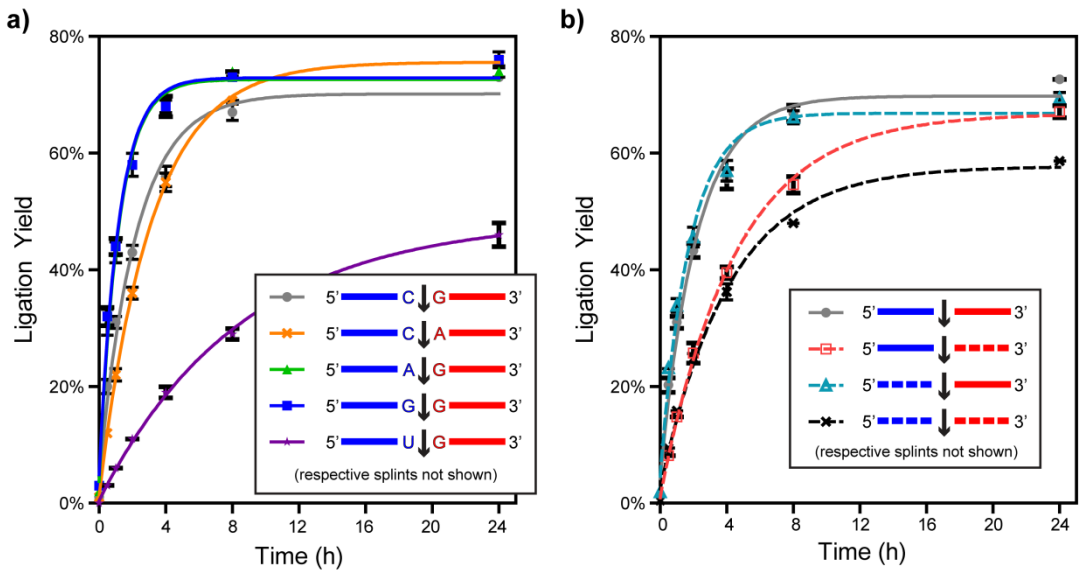


Figure S5.1 - Probing the sequence specificity of RNA ligase 10C with various substrate combinations. The PPP-RNA substrates and the RNA-OH substrates are shown in red and blue, respectively. The data are the mean of 3 replicates ± SD and were fit to first-order kinetics $Y=Y_f(1-e^{-kt})$ with Y_f=final yield and $k=k_{obs}$. The arrow represents the site of ligation and the complementary splint required for the reaction is omitted for clarity. **(a)** Single nucleotide changes next to the ligation site. The C↓G substrates are the original two sequences that were used in the selection and evolution of RNA ligase 10C. Rates of ligation (k_{obs}) ranged from 0.12 ± 0.01 h^{-1} for U↓G to 0.85 ±0.05 h^{-1} for A↓G. While not all theoretically possible substrate nucleotide combinations were tested, the combination U↓A is likely to have the lowest reaction rate. This rate can be estimated by multiplying the k_{obs} (U↓G) by the fraction of k_{obs} (C↓A) / k_{obs} (C↓G) which yields an approximate value of 0.085 h^{-1}, assuming that the reaction rate contributions from nucleotides on both sides of the ligation site are independent. **(b)** Change of whole sequence of ligation substrates. Progress of ligation is shown for combinations of the original substrate sequences used during the selection and evolution of ligase 10C (solid gray line) and substrate sequences that are completely unrelated, but have the same length and GC content (dashed lines). For detailed sequence information see **Table S5.1**.

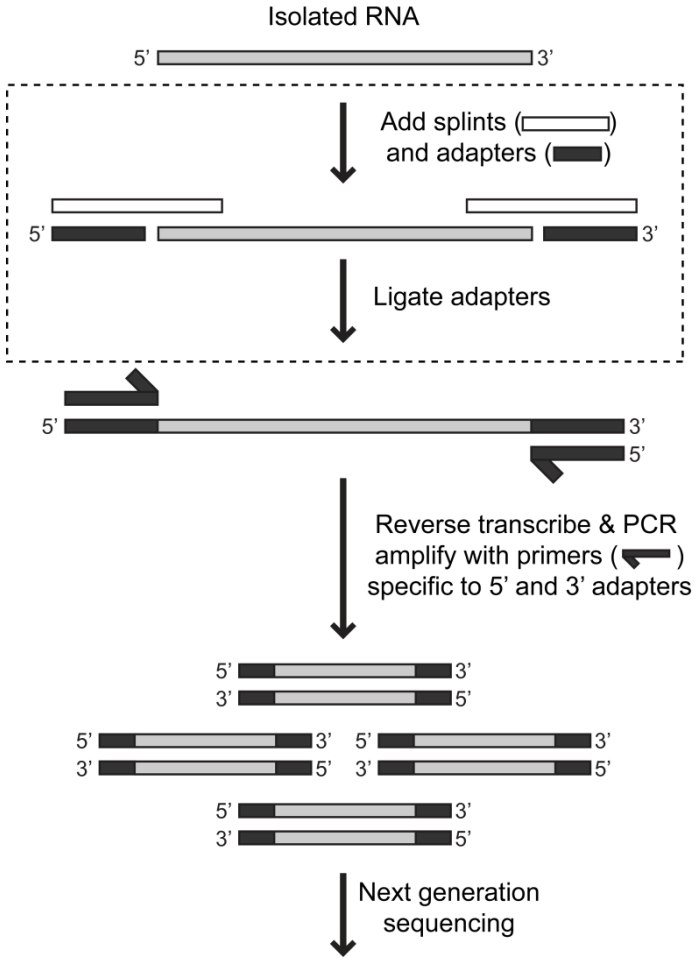
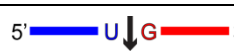
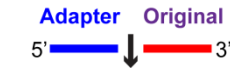


Figure S5.2 - General method for the modification of RNA samples necessary for next generation sequencing. RNA is typically not sequenced directly, but first converted to DNA through reverse transcription. Adapters add the needed terminal constant regions that facilitate the annealing of primers and reveal the orientation of the original RNA sequence. The artificial RNA ligase enzyme can be used during the "Ligate adapters" step to ligate the adapter to those "Isolated RNA" molecules that have a 5′-triphosphate group utilizing degenerate splints (dashed box). Ligase 10C can also ligate the second adapter to the 3′-terminus of the "Isolated RNA" if an adapter with a 5′-triphosphate is used.

Table S5.1 - Oligonucleotide substrate combinations used in the sequence specificity and application experiments.

Substrate combinations	RNA-OH	PPP-RNA [a]	DNA Splint
Variation at ligation site [b]			
 [c]	5′–CUAACGUUCG**C**	5′–**G**GAGACUCUUU	5′–GAGTCTCCGCGAACGT
	5′–CUAACGUUCG**C**	5′–**A**GAGACUCUUU	5′–GAGTCTCTGCGAACGT
	5′–CUAACGUUCG**A**	5′–**G**GAGACUCUUU	5′–GAGTCTCCTCGAACGT
	5′–CUAACGUUCG**G**	5′–**G**GAGACUCUUU	5′–GAGTCTCCCCGAACGT
	5′–CUAACGUUCG**U**	5′–**G**GAGACUCUUU	5′–GAGTCTCCACGAACGT
Variation of whole sequence			
 c	5′–CUAACGUUCGC	5′–GGAGACUCUUU	5′–GAGTCTCCGCGAACGT
	5′–GCAUGUCAGCA	5′–AGGCCUAUCAA	5′-ATAGGCCTTGCTGACA
	5′–CUAACGUUCGC	5′–AGGCCUAUCAA	5′–ATAGGCCTGCGAACGT
	5′–GCAUGUCAGCA	5′–GGAGACUCUUU	5′–GAGTCTCCTGCTGACA
Secondary siRNA			
	5′-ACGUUCGA	5′–GGAGACUCUUU	5′-GAGTCTCCTCGAACGT
	5′-ACGUUCGA	5′-GCAGAAACUGGAACCCAGGU	5′-GTTTCTGCTCGAACGT
	5′-ACGUUCGA	5′-GGAACCCAGGUGUUGGUCUUUG	5′-TGGGTTCCTCGAACGT

[a] All oligonucleotides in this column carry a 5′-terminal triphosphate, which is not shown here to simplify the table.
[b] Nucleotides that were varied are underlined and shown in bold.
[c] This combination of substrate sequences is identical to the original substrates used in the selection and evolution of the RNA ligase enzymes.

Table S5.2 - Melting temperatures (Tm) calculated for the hybridization of each PPP-RNA substrate with each of the three different splints used in the application experiment. The Tm values for fully complementary PPP-RNA/splint combinations are close to room temperature or above (shown in bold). All combinations that would result in mismatches yielded calculated Tm values that were substantially below room temperature (shown in parentheses and italics). In practical terms, this means that the fully complementary combinations are stable under the reaction conditions used in the application experiment, whereas the mismatched combinations are unlikely to hybridize.

PPP-RNA	Tm (by Pasteur [a])			Tm (by Stratagene [b])		
	Splint-1	Splint-2	Splint-3	Splint-1	Splint-2	Splint-3
Original (S1)	**30.1 °C**	-	-	**22.8 °C**	*(-19.9 °C)*	*(-27.3 °C)*
siRNA (S2)	-	**18.3 °C**	-	*(-14.8 °C)*	**17.6 °C**	*(-63.0 °C)*
siRNA (S3)	-	-	**30.2 °C**	*(-27.3 °C)*	*(-57.4 °C)*	**22.8 °C**

[a] Tm values were calculated using the Melting 4.1f calculator hosted by Mobyle@Pasteur (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::melting) with adjustments made for the RNA/DNA duplex, [Na+] and [oligonucleotide].

[b]Tm values were calculated using Stratagene's QuikChange® Primer Design Program (Tm = 81.5 + 0.41(%GC) - (675/N) - % mismatch, where N = total number of bases. https://www.genomics.agilent.com/CollectionSubpage.aspx?PageType=Tool&SubPageType=ToolQCPD&PageID=15) Note that this calculator assumes a DNA/DNA duplex. The calculator was used because no suitable RNA/DNA calculator was available that considers mismatches. While the real Tm values for RNA/DNA duplexes are higher than those calculated for the respective DNA/DNA duplexes, the general trend between different sequences is expected to be similar.

Chapter 6:

Development of the application of artificial ligase 10C for the next generation sequencing of 5'-triphosphate RNA

The following is a collection of unpublished experiments to develop the application of the artificial RNA ligase for the sequencing of PPP-RNAs from a complex mixture. I performed all experiments unless otherwise stated.

6.1 Overview

In the previous chapters we described an RNA ligase that was specific for RNA substrates with a 5'-triphosphate (PPP-RNA). The enzyme was made through *in vitro* evolution and catalyzes the splinted ligation of RNA with a 5'-triphosphate to a second RNA forming a native 5'-3' linkage and releasing pyrophosphate. In this chapter, we describe the use of this ligase for the profiling of PPP-RNA though RNA sequencing. Using degenerate splints, our ligase adds constant adaptor sequences to both termini of PPP-RNA allowing them to be reverse transcribed and PCR amplified. We optimized the ligation conditions to increase yields as well as identified numerous inhibitors to avoid. Finally we have adapted our protocol to be compatible with the Illumina TruSeq™ platform for RNA sequencing.

6.2 Introduction

Unprimed RNA synthesis produces RNA with a 5'-triphosphate (PPP-RNA). While the 5' end can be cleaved during the maturation of ribosomal RNAs or capped in the case of eukaryotic mRNA, several types of RNA maintain the 5'-triphosphate as their key distinguishing feature. Notable examples of PPP-RNA include mRNA in prokaryotes, secondary siRNAs in *C. elegans*, [1, 2] mRNA synthesized in the mitochondria [3] or chloroplast [4] and some viral mRNAs. [5] Currently, research on PPP-RNAs is hampered as these RNAs cannot be selectively purified from the RNA mixture and the distinguishing triphosphate is not visible after sequencing. In order to

identify PPP-RNAs a few protocols have been developed to enrich PPP-RNA within a sequencing sample, but all of these methods are susceptible to the identification of false positives. [6]

Current protocols to sequence and identify PPP-RNA rely on a specific phosphatase such as Tobacco Acid Phosphatase [7] or RNA 5' polyphosphatase that converts triphosphate to monophosphate. [6] The 5'-tagRACE method utilizes two separate adapters, the first acts as a decoy while the second one is used for sequencing. First, the short decoy adapter is ligated to all RNA in the sample with a 5'-monophosphate (P-RNA) removing them from the pool. Subsequently, the RNA sample is treated with a phosphatase to convert the 5'-triphosphate to a monophosphate followed by ligation of the sequencing adaptor. [7] An alternative RACE method utilizes Terminator 5'-phosphate-dependent exonuclease to degrade all 5'-monophosphate RNA prior to the phosphatase treatment. [6] Both protocols ultimately rely on a standard RNA ligase, which can only act on P-RNA substrates, to join the sequencing adaptors to the originally 5'-triphosphorylated RNA after first eliminating all endogenous P-RNAs from the sample. While these protocols certainly enrich PPP-RNAs in the sample, some P-RNA sequences will inevitably pass through the procedure resulting in false positive results.

Our lab has developed artificial RNA ligases that are dependent on a 5'-triphosphate for activity. [8] One particular variant called "Ligase 10C" was shown to have broad sequence specificity and has potential for the large scale identification of PPP-RNAs from complex mixtures. [9] Unlike natural ligases which use a cofactor such as ATP, 10C uses the chemical energy of the triphosphate with pyrophosphate as the leaving group to join two strands of RNA together in a mechanism that closely resembles an RNA polymerase. [10] Because of this energetic requirement, ligation of sequencing adaptors by ligase 10C should be highly selective for PPP-RNA over P-RNA. In this chapter, we describe our efforts to adapt RNA ligase 10C for the sequencing of isolated PPP-RNAs. We propose a process with a single ligation step where ligase 10C adds sequencing adaptors to both ends of PPP-RNA prior to reverse transcription and PCR amplification (Figure 6.1). While the exact adaptors used can easily be varied based on

the application or sequencing platform, we chose to base our adaptors on the Illumina TruSeq™ platform for RNA identification as it is the most prominent platforms used in RNA biology today. Our first step was to optimize the ligation conditions and to identify compounds that might be present in the RNA sample that could inhibit the ligation. We also sought to measure and if necessary, reduce the RNase activity of the ligase preparation to improve the detection sensitivity especially for low abundance RNAs. Finally, we applied our sequencing protocol to an idealized substrate to verify the efficacy of the process. We are currently collaborating with several research groups that are interested in different classes of PPP-RNA to test our kit with original RNA samples isolated from cells.

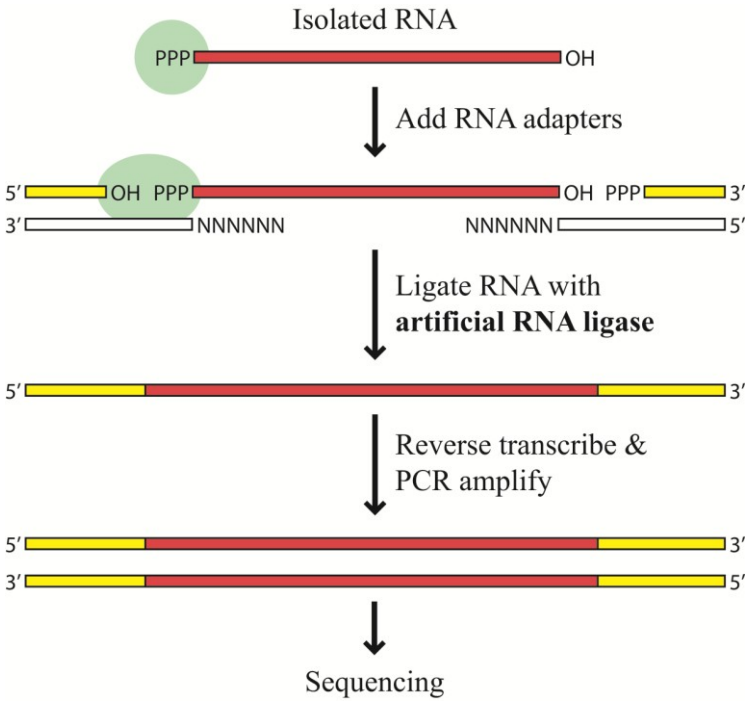


Figure 6.1 General scheme for the addition of sequencing adaptors specifically to of PPP-RNA using the artificial RNA ligase 10C. Ligase 10C ligates both adaptors in a single step although the 3' adaptor could be added by other means.

6.3 Results

To optimize ligation yields, we varied ligation conditions one variable at a time to determine the range of conditions where ligase 10C had >90% of maximum activity (Table 6.1). Ligase 10C is a 10 kDa metalloprotein that binds two zinc ions. [10] The

enzyme is dependent on zinc for structure and activity so zinc needs to be supplemented in the reaction buffer. When measured, ligase 10C has an apparent optimal concentration of 250 μM $ZnCl_2$ which was significantly different from the optimum of a related enzyme (ligase #4, 100 μM $ZnCl_2$) characterized in a previous publication. [10] However, when we initially determined the zinc concentration for maximal ligation activity of ligase 10C, we did so in the presence of 2 mM β-mercaptoethanol (βME) compared to ~0.2 mM used with ligase #4. βME is a reducing agent commonly used in protein biochemistry to maintain cysteines in their reduced state and had been included in the storage buffer of ligase 10C to improve the stability of the zinc-binding enzyme. But we found that the reducing agent was competing with ligase 10C for binding of zinc. When βME concentrations were reduced 20-fold, the zinc concentration for maximal activity was found to be approximately 120 μM (Figure S6.1). [8]

Table 6.1 - Optimum ligation conditions

Variable	**> 90% activity**	**Optimum[a]**
pH	6.8-7.6	7.3
[NaCl]	80-150 mM	120 mM
[$ZnCl_2$]	90-230 μM	143 ± 4.4 μM
βME[b]	< 100 μM	-

[a] pH value is the average of two experiments, [NaCl] is a single experiment and [$ZnCl_2$] is the average of three experiments
[b] βME or β-mercaptoethanol was initially considered a vital component of the storage buffer but further investigation revealed it acted as an inhibitor

As part of the optimization process, we also screened and identified inhibitors of ligase 10C (Table 6.2). Several inhibitors we identified act by competing with ligase 10C for zinc binding such as EDTA, Bis-Tris buffer and the reducing agents (βME, DTT & TCEP). While reducing agents and Bis-Tris buffer are not typically thought of as chelators of divalent cations, their ability to bind some metals has been previously reported. [11, 12] Ligase 10C also appears to be inhibited by magnesium (IC_{50} = 1 mM) which is noteworthy because natural enzymes that catalyze similar reactions are dependent on magnesium for activity. Ligase 10C is also inhibited by pyrophosphate,

which is one of the products of the ligation reaction. While pyrophosphate also has the potential to inhibit 10C by competing with the enzyme for zinc binding, its primary mode of inhibition is likely through a direct interaction with the enzyme itself. There are two lines of evidence that support this conclusion. First, the IC_{50} value is substantially lower than the IC_{50} of EDTA, although zinc has a higher affinity for EDTA ($K_d = 10^{-16}$) than for pyrophosphate ($K_{sp} = 10^{-8}$). Second, when the zinc concentration was increased by ~4-fold in the ligation mixture, the IC_{50} value for pyrophosphate was essentially unchanged (Figure S6.2). Theoretically this implies that ligation yields might be increased by the addition of a pyrophosphatase to breakdown pyrophosphate, but in practice the final concentration of pyrophosphate is typically less than the IC_{50} value. One unexpected inhibitor identified in our experiments was glycogen which is a commonly used as a co-precipitant for nucleic acids to minimize losses during ethanol precipitation. While the mechanism of this inhibition is unclear, omitting glycogen from precipitations is a simple precaution and RNA recoveries for this application without glycogen are regularly > 90% showing that it is unnecessary in sample preparation.

Table 6.2 - Summary of known inhibitors of ligase 10C

Compound	**IC_{50}[a]**	**Proposed mechanism[b]**
Pyrophosphate	10 μM	product inhibition
βME	2 mM	zinc binding
DTT	< 5 mM	zinc binding
TCEP	2 mM	zinc binding
Glycogen	10 μg/mL	n.d.
Bis-Tris Buffer	< 10 mM	zinc binding
Mg^{2+}	1 mM	n.d.
EDTA	50 μM	zinc binding
SUPERase In™[c]	100 U/mL	n.d.

[a] IC_{50} is the concentration of a compound which inhibits the activity of the enzyme by 50%.
[b] Proposed mechanism by which the compound might inhibit the enzyme.
[c] Only one concentration of SUPERase In™ was tested which happened to inhibit the enzyme by 50%.

In all ligation experiments no significant substrate or product degradation was observed with ligase 10C purified as previously described. [9-10] However, in these experiments we used a relatively high concentration of RNA (low micromolar) and short substrates (less than 25 nt). These parameters are different from a typical RNA sample that was isolated from a cell and used for RNA sequencing. Therefore, we needed to assess the RNase susceptibility under more realistic conditions. To better assess RNase activity we incubated samples of ligase 10C with nanomolar concentrations of a long (802 nt) RNA substrate. Under these conditions we could now clearly detect RNase activity in our ligase samples. To reduce the RNase activity, we first supplemented the mixture with the RNase inhibitor SUPERase In. However, even with SUPERase In present, substantial amounts of RNase activity was observed. Moreover while the concentration of SUPERase In wasn't sufficient to inhibit RNase activity, it decreased ligase activity by 50%. (Data not shown) Other commercial RNase inhibitors such as RNaseOUT and RNAsin were briefly considered, but they are both dependent on DTT, an identified inhibitor, and therefore cannot be used with ligase 10C. The solution instead came from an additional, orthogonal round of purification. Originally, the ligase protein was purified by Ni-NTA and size exclusion chromatography. Adding a third ion exchange purification reduced the RNase activity dramatically. (Figure 6.2)

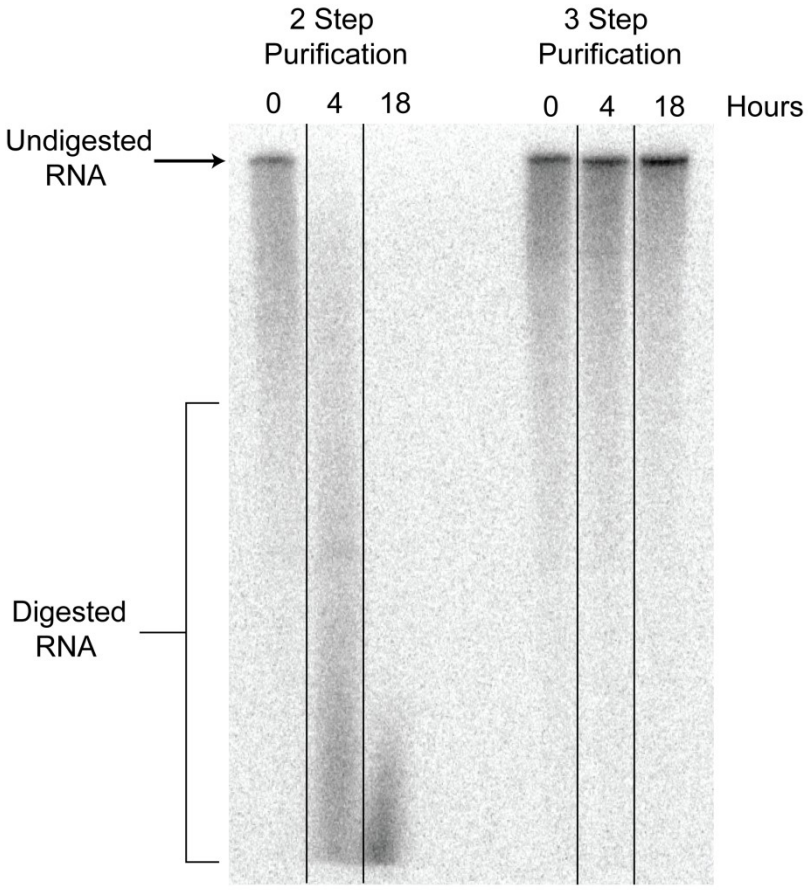


Figure 6.2 - Reducing RNase activity of the ligase 10C preparation by removing RNase contaminations in a 3 step purification process. A ^{32}P labeled RNA substrate was incubated with samples of ligase 10C for the times indicated and analyzed by PAGE to visualize RNase activity. Using the 2 step purification protocol, the RNA was completely degraded after 18 hours. Using the 3 step protocol there is no detectable amounts of degradation after 18 hours.

In our proposed scheme for sequencing PPP-RNA (Figure 6.1), the first step after the isolation of RNA is to ligate sequencing adaptors to each end of the sample RNA using ligase 10C. The 5' adaptor is a chemically synthesized RNA oligo while the 3' adaptor was made using *in vitro* transcription so that it would contain a 5'-triphosphate. Both adaptors ligate with similar efficiencies to 21 nt PPP-RNA model substrate both individually and in combination (Figure 6.3). The 21 nt substrate was chosen to mimic the size of fragmented mRNA from bacteria and secondary siRNAs from *C. elegans* which are two proposed applications for this method.

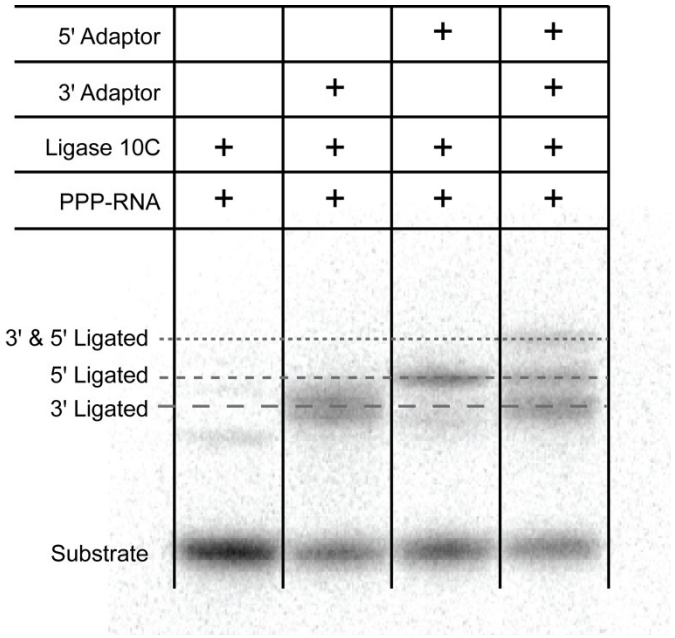


Figure 6.3 - Ligation of sequencing adaptors to a 21 nt model substrate. Urea PAGE of the ^{32}P labeled PPP-RNA substrate incubated with ligase 10C in the presence and absence of the sequencing adaptors and their corresponding complementary splints. Without any adaptors or splints present, the PPP-RNA will undergo a small amount of self-ligation. In the presence of adaptors and complementary splints, the desired product dominates. Ligation of the 3' adaptor to the 21 nt substrate consistently produces a double band on the gel for an unknown reason. Fully ligated product is observed only in the presence of both adaptors.

To demonstrate that only PPP-RNA can be amplified during PCR after ligation, we tested the protocol with an equivalent P-RNA substrate to ensure it couldn't be amplified (Figure 6.4). The PPP-RNA sample with all necessary adapters, splints and ligase yields the expected single band at approximately 140 bp. This band was excised from the gel and sequenced which confirmed that it contained the 21 nt PPP-RNA substrate ligated to the sequencing adapters. For the 5'-monophosphate substrate RNA (absence of a 5'-triphosphate), no fully ligated band was detected after the PCR although there are other lower bands present in the sample. If the total number of cycles is reduced from 25 to 10 as would be done in the sequencing experiment, only the fully ligated PPP-RNA and not the P-RNA sample produces any bands (data not shown). To investigate the identity of these additional bands, several ligation components were omitted one at a time in the case of the P-RNA substrate, while the 3' adaptor and splint were present in all lanes. A band at approximately 120 bp appears to be ligated adaptor dimer as it forms only when the 5' adaptor and ligase are present. The adaptor dimer band is also more

intense when the 5’ splint is omitted as expected. The lower band correlates with the addition of the 5' splint and forms both in the presence and absence of ligase which implies it is not a ligated product. In summary, these data show that ligase 10C selectively ligates sequencing adapters to PPP-RNA.

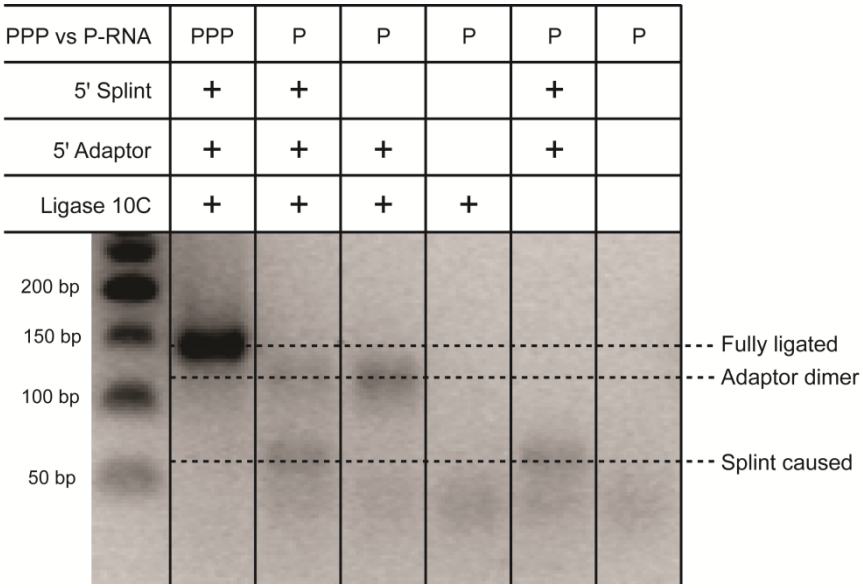


Figure 6.4 - PCR amplification after the adaptor ligation procedure performed on PPP-RNA vs. P-RNA substrates. Ligations were performed as indicated above with the 3' adaptor and splint present in all cases prior to reverse transcription, 25 cycles of PCR amplification and analysis by agarose gel. Fully ligated product was only observed with the PPP-RNA substrate as expected. In the absence of PPP-RNA, additional minor bands were observed that correspond to a ligated adaptor dimer as well as an unknown band associated with the 5' splint.

In order to test or artificial ligase 10C for sequencing applications with actual RNA samples isolated from different kinds of cells, we initiated several collaborations with RNA biologists that are interested in different classes of PPP-RNA.

6.4 Discussion

While we present a single method for the sequencing of RNA with ligase 10C, this is not the only available path for sequencing PPP-RNA with ligase 10C. In particular, only the 5' adaptor needs to be ligated by 10C for the method to work. The 3' adaptor can be introduced by 10C as shown here or by any other means desired. For example, the 3' adaptor could be added through a separate ligation step or it could be introduced through

a reverse transcription primer, avoiding the second ligation step entirely. Our experiments showed that a byproduct of our current scheme is the potential formation of adaptor dimers during the ligation step albeit at a low rate. This could be eliminated entirely through the alternate method if desired by the end user. Another alternate method for conducting this sequencing protocol would be to use a modified 5' adaptor labeled with biotin or click chemistry linkers to remove non-ligated RNAs from the sample mixture prior to reverse transcription and PCR. While the non-ligated RNAs should be eliminated through a lack of enrichment during the PCR step, this step could be made even more stringent through the additional purification.

Currently, ligase 10C is dependent on a complementary splint to ligate two RNA strands. This can provide remarkable selectivity for isolating a single desired sequence out of a mixture, but it can be difficult to ligate RNA which forms stable secondary structures near the ligation site. Ligase 10C is a thermostable enzyme with a melting temperature of 72°C so presumably the ligation could be performed at high temperature to decrease or eliminate secondary structure, but it is not possible to perform such a ligation with the degenerate splints necessary to capture all PPP-RNAs in a sample. Therefore a non-splinted 5'-triphosphate dependent thermostable ligase could be very useful for improving coverage of sequences in the RNA pool that might otherwise be difficult to detect because it forms secondary structure at lower temperature.

6.5 Conclusions

The research presented here demonstrates the capability of the artificial RNA ligase 10C to be used in the sequencing of PPP-RNA. Ligase 10C was well characterized to help avoid potential pitfalls in RNA sample preparation. While the ligase appeared to be incompatible with commercial RNase inhibitors, the triple-purified enzyme showed no RNase activity under realistic sequencing application conditions. RNA ligations for RNA-sequencing in commercial kits are also performed without RNase inhibitors as both the RNA sample and ligase are expected to be RNase free. If desired, ligase 10C can be readily inactivated through the addition of zinc chelators. Our collaborations have shown

that our kit can isolate PPP-RNAs isolated from cells, but further work is needed to increase the robustness of the protocol.

6.6 Materials & Methods

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

6.6.1 Expression & Purification of RNA Ligase 10C

RNA ligase 10C was expressed and purified as previously published. [10] For the RNA degradation experiments the protocol was modified as follows. Ni-NTA and size exclusion (Superdex G75) purified protein samples were loaded onto a MonoQ 5/50 GL column (GE Healthcare) and analyzed on the AKTA FPLC system (GE Healthcare). Buffer A: 20 mM Tris-HCl pH 8, 250 μM β-mercaptoethanol (βME) and 50 μM $ZnCl_2$, Buffer B: 20 mM Tris-HCl pH 8, 1 M NaCl, 250 μM βME and 50 μM $ZnCl_2$. Ligase 10C was bound to the column in 100% buffer A, then eluted by a gradient of 0-30% buffer B over 10 column volumes. Elution fractions were pooled and dialyzed against 2 changes of 2 L ligase storage buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 500 μM βME, 100 μM $ZnCl_2$).

6.6.2 Preparation of Oligonucleotides

The α-^{32}P-labeled PPP-RNA substrates were prepared by *in vitro* transcription using T7 RNA polymerase as previously published. [10] Non radioactive PPP-RNA substrates and 3' RNA adaptors were prepared in the same way as the ^{32}P-labeled counterparts but without the use of α-^{32}P-UTP and with equal concentrations of non-radioactive NTP in the transcription. Non radioactive P-RNA substrates and 5' RNA adaptors were purchased from Dharmacon and prepared according to the manufacturer's protocol. DNA splints were purchased from Integrated DNA Technologies. All oligonucleotides were dissolved in ultra-pure water and concentrations determined by UV absorbance prior to use in the ligation assays. See Table S6.1 for the sequences of all oligonucleotides used.

6.6.3 Ligation Assay

For the optimization of ligation conditions and identification of inhibitors 10 μM PPP-RNA, 20 μM RNA-OH and 30 μM DNA splint were combined in a buffer containing 20 mM HEPES pH 7.5, 100 mM NaCl, 100 μM $ZnCl_2$. A stock of 25 μM RNA ligase 10C in buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 100 μM $ZnCl_2$ and 0.5 mM βME was added to the oligonucleotide mix to a final concentration of 5 μM enzyme and ligation proceeded at room temperature. Potential inhibitors were added prior to the addition of ligase at the indicated concentrations.

For the ligation of sequencing adaptors 500 nM PPP-RNA, 600 nM RNA adaptors and 1.2 μM DNA splints were combined in a buffer containing 20 mM HEPES pH 7.5, 100 mM NaCl, 100 μM $ZnCl_2$. A stock of 50 μM RNA ligase 10C in buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 100 μM $ZnCl_2$ and 0.5 mM β-mercaptoethanol was added to the oligonucleotide mix to a final concentration of 10 μM enzyme and the ligation proceeded at 16°C.

The ligation reactions were quenched with two volumes of 20 mM EDTA in 8 M urea, 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).

6.6.4 Measurement of RNase activity

The α-^{32}P-labeled 802 nt RNA substrate was prepared by *in vitro* transcription using T7 RNA polymerase as previously published. [10] To test for RNase activity in preparations of the ligase enzyme, 50 nM substrate was combined with 10 μM ligase 10C in RNA ligation buffer and incubated at room temperature until stopped by the addition two volumes of 20 mM EDTA in 8 M urea. Samples were 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).

Sequence of 802 nt RNA substrate:

```
5' - GGGACAAUUA CUAUUUACAA UUACAAUGGG CAGCGAUAAG AUCCACCAUC ACCAUCACCA
     UGUGAUUGUG CUGGGCCAUC GCGGUUACUC CGCCAAGUAU CUGGAAAACA CCCUGGAAGC
     UUUCAUGAAA GCGAUCGAAG CCGGCGCGAA UCGUGAGGAG CUGGAUGUGC GCCUGUCUAA
     AGACGGCAAG GUGGUCGUGA GCCAUGAUGA AGAUUUAAAG CGCCUGUUCG GUCUGGACGU
     CAAAAUCCGU GACGCCACCG UGUCUGAACU GAAAGAGCUG ACCGAUGGCA AAAUUACCAC
     CCUGAAGGAA GUGUUUGAGA ACGUGUCCGA UGACAAGAUC AUCAACAUCG AAAUCAAGGA
     ACGUGAGGCC GCGGACGCAG UGCUGGAGAU CAGCAAAAAG CGUAAGAACC UGAUUUUCAG
     CUCCUUUGAU CUGGACCUGC UGGAUGAAAA AUUCAAGGGU ACCAAAUACG GUUAUCUGAU
     CGACGAAGAG AACUACGGUU CCAUUGAAAA UUUCGUGGAG CGCGUGGAAA AGGAGCGUCC
     GUACUCUCUG CACGUGCCCU AUCAGGCCUU UGAGCUCGAA UAUGCGGUGG AGGUGCUGCG
     CUCCUUCCGU AAAAAGGGCA UCGUGAUUUU UGUGUGGACC CUGAAUGAUC CGGAAAUCUA
     CCGCAAAAUA CGUAGAGAGA UCGAUGGUGU GAUUACCGAC GAAGUGGAGC UGUUUGUGAA
     ACUGCGUGGC GGCAGCGGUG GCUCCGACUA UAAGGAUGAC GAUGACAAAA UGGGAAUGUC
     UGGAUCUGGC ACCGGCUAUU AA
```

6.6.5 Reverse Transcription and PCR amplification of ligated PPP-RNA

Sequencing adaptors were ligated to PPP-RNA as described in section 6.6.3 in a 10 µl reaction volume. Reaction was halted by adding 0.5 µl of 4 mM EDTA, 6 µl of reaction mixture were removed and mixed with 1 µl of 7 µM reverse transcription primer. The mixture was heated to 70°C for 2 min then placed on ice. Sample was reverse transcribed through the addition of 2 µl 5x 1st strand synthesis buffer (Invitrogen), 0.5 µl of 12.5 mM dNTP (NEB), 1 µl of 100 mM DTT (Invitrogen), 1 µl RNaseOUT (Invitrogen) and 1 µl Superscript II (Invitrogen) followed by incubating at 50°C for 1 hour. PCR was conducted by adding 22.5 µl ultrapure water, 10 µl 5x phusion buffer (NEB), 2 µl of 12.5 µM primer (forward and reverse), 1 µl of 12.5 mM dNTP (NEB) and 0.5 µl Phusion polymerase (NEB). The PCR program was as follows: the sample was heated to 98°C for 30 seconds, 5 to 30 cycles of 98°C for 10 seconds, 60°C for 30 seconds and 72°C for 15 seconds, with a final extension step of 72°C for 2 min. 5 µl samples of the PCR mix were analyzed by 2% agarose gel containing ethidium bromide.

6.7 Supplementary Materials

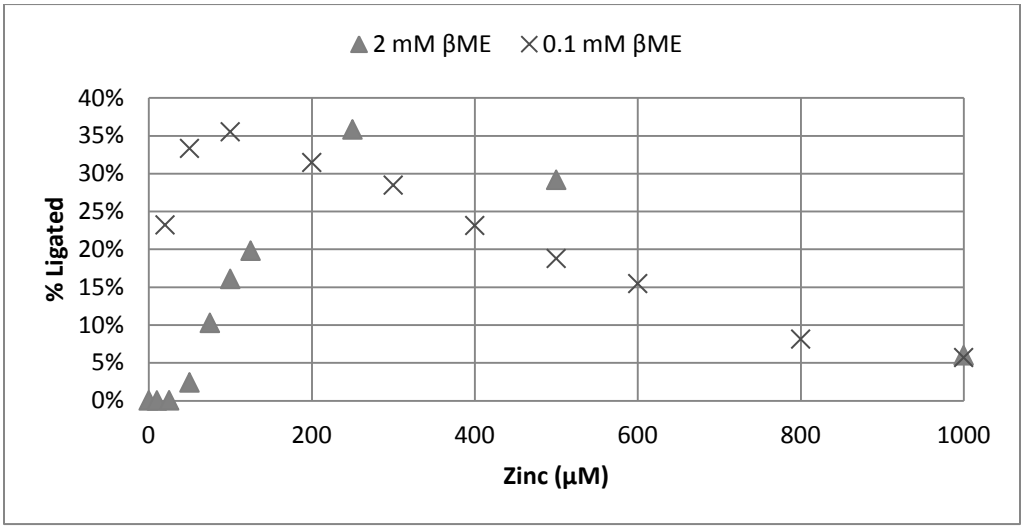


Figure S6.1 Dependence of ligation activity on zinc concentration at different concentrations of βME. Ligase 10C was originally stored in 10 mM βME which resulted in a final concentration of 2 mM in the ligation assay. At this concentration βME competes with the ligase for the binding of zinc as seen in the shifted zinc optimum compared to a low βME sample (0.1 mM).

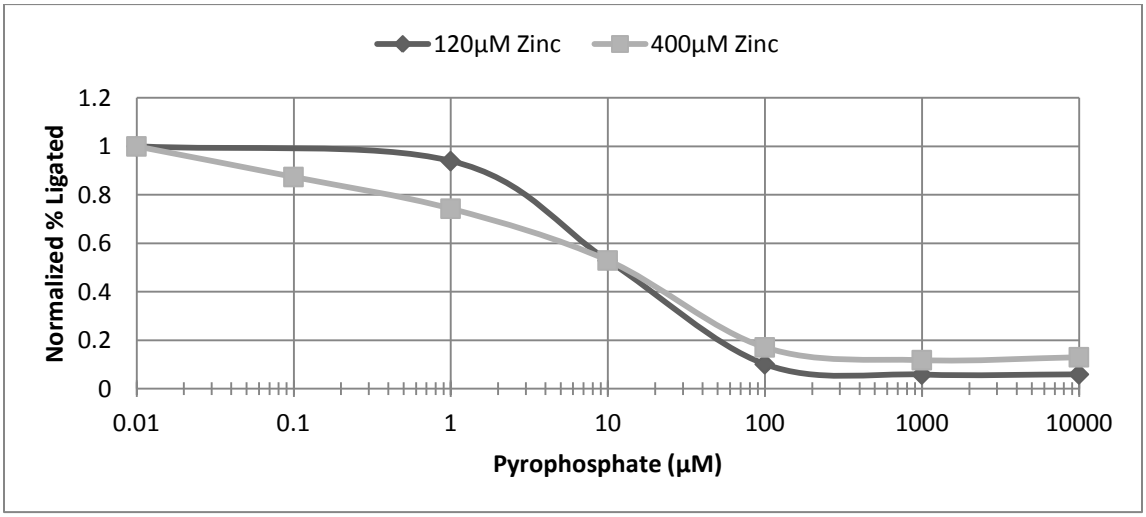


Figure S6.2 Inhibition of ligase 10C by pyrophosphate at different concentrations of zinc. The pyrophosphate inhibition curve was measured at 120 μM and 400 μM Zn^{2+} to determine if pyrophosphate was inhibiting ligase 10C through competition with zinc binding. When the curves are normalized to account for the reduced activity of ligase 10C at 400 μM zinc the 50% inhibition concentration remains the same.

Table S6.1 Oligonucleotides used in this chapter

Name		**Sequence**
Substrates for optimization of ligation yields and inhibition screen	PPP-RNA	5' – PPP-GGAGACUCUUU
	RNA-OH	5' – CUAACGUUCGA
	DNA Splint	5' – GAGTCTCCTCGAACGT
Ligation of PPP-RNA for sequencing	21 nt PPP-RNA	5' – PPP-GGAGACTCTTCTAACGTTCGG
	5' Adaptor	5' – GUUCAGAGUUCUACAGUCCGACGAUC
	5' Adaptor splint	5' – GTCTCCGATCGTCGGACTGTAGAACTCTGAAC
	3' Adaptor	5' – GGAUGGAAUUCUCGGGUGCCAAGG
	3' Adaptor splint	5' – CCTTGGCACCCGAGAATTCCATCCCCGAAC
Primers for sequencing	RT primer	5' – GCCTTGGCACCCGAGAATTCCA
	PCR fwd. primer	5' – AATGATACGGCGACCACCGAGATCTACACGTT CAGAGTTCTACAGTCCGA
	PCR rev. primer	5' – CAAGCAGAAGACGGCATACGAGATCGTGATGT GACTGGAGTTCCTTGGCACCCGAGAATTCCA

Conclusions and Future Directions

Artificial enzymes hold the potential to solve innumerable challenges in the field of biotechnology. New enzymes can make chemical transformations specific, cheap and environmentally friendly with applications in drug synthesis and bulk synthesis of commodity chemicals. [1-3] New enzymes can also enable new forms of labeling of biomolecules with applications ranging from research to medicine.[1, 4-6] The ultimate goal of the field of artificial enzyme design is to develop a rapid reliable process to generate custom catalysts that are sufficiently active, selective and stable to be a viable commercial product. Generating such an enzyme *de novo* using a single approach may not be possible in the near future. In contrast, optimizing existing enzymes is now a well-established procedure due to advances in directed evolution which have improved success rates and decreased costs. Therefore, artificial enzymes generated by rational design or *de novo* selection that may initially only exhibit low levels of activity could be evolved further with more standard directed evolution.

An important question for the *de novo* selection of artificial enzymes is how to best construct the most promising libraries of protein variants. Chapter 2 described the construction of a $(\beta/\alpha)_8$ barrel library that was enriched for folded proteins because folding is generally considered to be a prerequisite for enzyme activity. We also created a "control" library that was based on the same scaffold protein, but had not undergone any folding selection. Our results show clearly that the control library contains about 50 times fewer folded proteins, but as the control library took only a fraction of the time to develop it might still represent the better choice for developing protein libraries in the future. The folding-selected and control libraries should be used in parallel in multiple enzyme selections to compare the number of enzymes identified, their stability and level of activity. In addition to comparing these two $(\beta/\alpha)_8$ barrel libraries, it could also be useful to explore alternative library assembly strategies. To make the folding-enriched library we ultimately chose to subject the L1-4 sub-library (the first half of the barrel) to 2 rounds of folding selection prior to recombining it with the other half. This was done because the first round of selection of L1-4 had a survival rate that was only about 2-fold

greater than background. We also considered performing a folding selection on the L1-2 and L3-4 libraries first, recombining them, and then selecting the L1-4 library for folding, but the amount of additional work that would have been needed made us decide against constructing another library to test both approaches. Ultimately, we decided that selecting for folding of multiple loops together in a single experiment was more important as this examines the loops in the context of one another. However, it would be interesting to determine how such a change in the library construction protocol would affect overall stability and loop composition.

Ligase 10C is the most active and best characterized 5'-triphosphate-dependent RNA ligase we have isolated to date. The 3D-structure described in chapter 4 shows that ligase 10C adopts a highly dynamic fold that has not been found in nature. As ligase 10C has not been subjected to millions of years of Darwinian evolution, this unusual protein structure might serve as a model for primordial enzymes. The ligase is also several orders of magnitude slower than the average natural enzyme, [7] with a turnover of about one per hour. Evolving ligase 10C to have activity comparable to native enzymes represents an unprecedented opportunity to monitor the evolution from a very early state. In addition to studying the mechanisms of enzyme evolution, further evolution could help explain some features of the unusual protein fold. We demonstrated in chapter 3 that ligase 10C is a 10 kDa protein that exhibits impressive thermostability with a T_m of 72°C and has activity at 65°C. Thermostability and flexibility are not typically observed in a single protein structure so examining this further may add to our understanding of protein folding mechanisms.

A key validation of developing artificial enzymes is to demonstrate utility of the catalyst. The work described in chapters 5 and 6 confirmed that ligase 10C is an attractive enzyme for the detection and sequence analysis of RNA with a 5'-triphosphate. This interest largely stems from the broad sequence specificity exhibited by ligase 10C which should enable the linking of any RNA with a 5'-triphosphate. This property of ligase 10C was not explicitly selected for in our experiments as only a single set of selection substrates was used, but our overall selection strategy may have promoted this result. In our selection, both enzyme and substrate are physically linked providing a high local

concentration of both compounds. We hypothesized that displaying the substrate in this manner reduces the selective pressure to evolve high substrate binding which would correlate to reduced substrate specificity. Alternatively, the library we utilized may have only contained relatively primitive artificial ligases which were all incapable of specific substrate recognition. In either case, preserving this broad specificity will be an important goal in any future evolution experiments.

In the near future, two primary goals of the Seelig lab will be to continue to evolve ligase 10C for increased activity and begin enzyme selections with the $(\beta/\alpha)_8$ barrel libraries. For the ligation protocols described in chapters 5 and 6, molar excess of ligase 10C was added to substrate to compensate for the low enzyme activity and reduce ligation time. We chose this route because we can easily produce ligase 10C in large amounts and it worked sufficiently well in our experiments with well defined substrates, but increased activity will be valuable for improving applications in RNA deep sequencing and other fields. In addition, the Seelig lab will also be using the $(\beta/\alpha)_8$ barrel libraries to select for alternative 5'-triphosphate-dependent RNA ligases. While it is reasonable to expect that ligase 10C can be further evolved and optimized, ligases from the $(\beta/\alpha)_8$ barrel library might be more structured and a better platform for rapid evolution. We also have plans to select for Diels-Alderase enzymes from the $(\beta/\alpha)_8$ barrel library which is a valuable reaction from classic organic synthesis.

There are many additional reactions which would be attractive targets for an enzyme selection. For example, the Friedel-Crafts reactions are a synthetically valuable family of reactions that removes a halogen from an organic molecule to create a reactive carbocation to form new carbon-carbon bonds. [8] These reactions are usually performed in anhydrous solvents to generate the carbocation intermediate, but an enzyme could potentially be used instead to form and protect the carbocation in an aqueous environment. Assuming such an enzyme could be developed with good regioselectivity, it could be very valuable alternative in organic synthesis. Other synthetically useful reactions that would be amenable to our mRNA display selection strategy are Suzuki cross coupling (with the addition of Pd^{2+} to the selection buffer as a potential co-factor) [9] as well as a wide variety of S_N2 chemistry. S_N2 reactions might not initially appear to

be a good target for enzymes as nucleophilic substitution is one of the most basic reactions in organic chemistry, but enzymes could potentially achieve regiospecificity of the reaction in complex substrates where many alternative substitution sites are present.

In conclusion, the selection of new artificial enzymes by mRNA display is an exciting new technology for the field of biotechnology. My thesis research has expanded the methods available for mRNA display selection by creating a highly diverse protein library and developed an application for an artificial RNA ligase that was isolated through *de novo* selection. While I did not get to select for a new enzyme myself, the folding enriched $(\beta/\alpha)_8$ barrel library will hopefully be a valuable source of new enzymes in the future. The characterization of ligase 10C was originally intended to be a small portion of my thesis, however, after we learned that RNA biologists lacked robust tools to study RNA with a 5'-triphosphate, our priorities shifted to focus more of our resources on developing ligase 10C as a tool for this field. The next challenge for mRNA display selections is to branch out to new reactions and to isolate new enzymes that can catalyze them. The goal is not just to make new artificial enzymes, but to further demonstrate the capabilities of the method. What is needed now is to invest the time and resources to perform the necessary selections.

Bibliography

Chapter 1

1. Smith, A. L. (1997) Oxford dictionary of biochemistry and molecular biology. *Oxford University Press*.

2. Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012) Engineering the third wave of biocatalysis. *Nature* **485,** 185-94.

3. Huisman, G. W., and Collier, S. J. (2013) On the development of new biocatalytic processes for practical pharmaceutical synthesis. *Current Opinion in Chemical Biology* **17,** 284-92.

4. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J., Brands, J., Devine, P. N., Huisman, G. W., and Hughes, G. J. (2010) Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **329,** 305-09.

5. Aehle, W. (2007) Enzymes in Industry. *Wiley-VCH Verlag GmbH & Co. KGaA*.

6. Y. H. Hui, L. M.-G., Jytte Josephsen, Wai-Kit Nip, Peggy S. Stanfield (2004) Handbook of Food and Beverage Fermentation Technology, CRC Press.

7. Carroll, A., and Somerville, C. (2009) Cellulosic Biofuels. *Annual Review of Plant Biology* **60,** 165-82.

8. Shen, L., Worrell, E., and Patel, M. (2010) Present and future development in plastics from biomass. *Biofuels, Bioproducts and Biorefining* **4,** 25-40.

9. Kuhad, R. C., Gupta, R., and Singh, A. (2011) Microbial Cellulases and Their Industrial Applications. *Enzyme Research* **2011,** 10.

10. Li, D.-C., Li, A.-N., and Papageorgiou, A. C. (2011) Cellulases from Thermophilic Fungi: Recent Insights and Biotechnological Potential. *Enzyme Research* **2011,** 9.

11. Innis, M. A., Myambo, K. B., Gelfand, D. H., and Brow, M. A. (1988) DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proceedings of the National Academy of Sciences* **85,** 9436-40.

12. Bickle, T. A., and Kräger, D. H. (1993) Biology of DNA restriction. *Microbiological Reviews* **57,** 434-50.

13. Barrett, A., Rawlings, N., and Woessner, J. (2004) Handbook of proteolytic enzymes. *Elsvier Academic Press*.

14. Waidmann, M. S., Bleichrodt, F. S., Laslo, T., and Riedel, C. U. (2011) Bacterial luciferase reporters: The Swiss army knife of molecular biology. *Bioengineered* **2,** 8-16.

15. Martell, J. D., Deerinck, T. J., Sancak, Y., Poulos, T. L., Mootha, V. K., Sosinsky, G. E., Ellisman, M. H., and Ting, A. Y. (2012) Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat Biotech* **30,** 1143-48.

16. Ramsay, N., Jemth, A.-S., Brown, A., Crampton, N., Dear, P., and Holliger, P. (2010) CyDNA: Synthesis and Replication of Highly Cy-Dye Substituted DNA by an Evolved Polymerase. *J Am Chem Soc* **132,** 5096-104.

17. Underkofler, L. A., Barton, R. R., and Rennert, S. S. (1957) Production of microbial enzymes and their applications. *Appl Microbiol* **6,** 212-21.

18. Moore, J. C., and Arnold, F. H. (1996) Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nat Biotech* **14,** 458-67.

19. Jäckel, C., Kast, P., and Hilvert, D. (2008) Protein Design by Directed Evolution. *Annual Review of Biophysics* **37,** 153-73.

20. Turner, N. J. (2009) Directed evolution drives the next generation of biocatalysts. *Nat Chem Biol* **5,** 568-74.

21. Izard, J. W., and Kendall, D. A. (1994) Signal peptides: exquisitely designed transport promoters. *Molecular Microbiology* **13,** 765-73.

22. Lee, S. Y., Choi, J. H., and Xu, Z. (2003) Microbial cell-surface display. *Trends in Biotechnology* **21,** 45-52.

23. Golynskiy, M. V., and Seelig, B. (2010) *De novo* enzymes: from computational design to mRNA display. *Trends Biotechnol* **28,** 340-45.

24. Cohen, N., Abramov, S., Dror, Y., and Freeman, A. (2001) *In vitro* enzyme evolution: the screening challenge of isolating the one in a million. *Trends Biotechnol* **19,** 507-10.

25. Cho, G., Keefe, A. D., Liu, R. H., Wilson, D. S., and Szostak, J. W. (2000) Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J Mol Biol* **297,** 309-19.

26. Kehoe, J. W., and Kay, B. K. (2005) Filamentous phage display in the new millennium. *Chem Rev* **105,** 4056-72.

27. Sidhu, S. S., Lowman, H. B., Cunningham, B. C., and Wells, J. A. (2000) Phage display for selection of novel binding peptides. *Methods Enzymol* **328,** 333-63.

28. Renesto, P., and Raoult, D. (2003) From genes to proteins - *in vitro* expression of rickettsial proteins. *Ann NY Acad Sci* **990,** 642-52.

29. Bulter, T., Alcalde, M., Sieber, V., Meinhold, P., Schlachtbauer, C., and Arnold, F. H. (2003) Functional expression of a fungal laccase in Saccharomyces cerevisiae by directed evolution. *Appl Environ Microbiol* **69,** 987-95.

30. Chusacultanachai S, and Yuthavong Y (1994) Random mutagenesis strategies for construction of large and diverse clone libraries of mutated DNA fragments. *Methods Mol Biol* **270,** 319-33.

31. Virnekas, B., Ge, L. M., Pluckthun, A., Schneider, K. C., Wellnhofer, G., and Moroney, S. E. (1994) Trinucleotide phosphoramidites - ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res* **22,** 5600-07.

32. Janczyk, M., Appel, B., Springstubbe, D., Fritz, H. J., and Muller, S. (2012) A new and convenient approach for the preparation of beta-cyanoethyl protected trinucleotide phosphoramidites. *Org Biomol Chem* **10,** 1510-13.

33. Bieberich, E., Kapitonov, D., Tencomnao, T., and Yu, R. K. (2000) Protein-ribosome-mRNA display: affinity isolation of enzyme-ribosome-mRNA complexes and cDNA cloning in a single-tube reaction. *Anal Biochem* **287,** 294-98.

34. Amstutz, P., Pelletier, J. N., Guggisberg, A., Jermutus, L., Cesaro-Tadic, S., Zahnd, C., and Plückthun, A. (2002) In Vitro Selection for Catalytic Activity with Ribosome Display. *J Am Chem Soc* **124,** 9396-403.

35. Takahashi, F., Ebihara, T., Mie, M., Yanagida, Y., Endo, Y., Kobatake, E., and Aizawa, M. (2002) Ribosome display for selection of active dihydrofolate reductase mutants using immobilized methotrexate on agarose beads. *FEBS Lett* **514,** 106-10.

36. Takahashi, F., Funabashi, H., Mie, M., Endo, Y., Sawasaki, T., Aizawa, M., and Kobatake, E. (2005) Activity-based *in vitro* selection of T4 DNA ligase. *Biochem Biophys Res Commun* **336,** 987-93.

37. Quinn, D. J., Cunningham, S., Walker, B., and Scott, C. J. (2008) Activity-based selection of a proteolytic species using ribosome display. *Biochem Biophys Res Commun* **370,** 77-81.

38. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828-31.

39. Seelig, B. (2011) mRNA display for the selection and evolution of enzymes from *in vitro*-translated protein libraries. *Nat Protoc* **6,** 540-52.

40. Odegrip, R., Coomber, D., Eldridge, B., Hederer, R., Kuhlman, P. A., Ullman, C., FitzGerald, K., and McGregor, D. (2004) CIS display: *in vitro* selection of peptides from libraries of protein-DNA complexes. *Proc Natl Acad Sci USA* **101,** 2806-10.

41. Reiersen, H., Lobersli, I., Loset, G. A., Hvattum, E., Simonsen, B., Stacy, J. E., McGregor, D., FitzGerald, K., Welschof, M., Brekke, O. H., and Marvik, O. J. (2005) Covalent antibody display - an in vitro antibody-DNA library selection system. *Nucleic Acids Res* **33,** e10.

42. Cohen, H. M., Tawfik, D. S., and Griffiths, A. D. (2004) Altering the sequence specificity of *HaeIII* methyltransferase by directed evolution using *in vitro* compartmentalization. *Protein Eng Des Sel* **17,** 3-11.

43. Doi, N., Kumadaki, S., Oishi, Y., Matsumura, N., and Yanagawa, H. (2004) *In vitro* selection of restriction endonucleases by *in vitro* compartmentalization. *Nucleic Acids Res* **32,** e95.

44. Fallah-Araghi, A., Baret, J. C., Ryckelynck, M., and Griffiths, A. D. (2012) A completely *in vitro* ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab Chip* **12,** 882-91.

45. Mastrobattista, E., Taly, V., Chanudet, E., Treacy, P., Kelly, B. T., and Griffiths, A. D. (2005) High-throughput screening of enzyme libraries: *in vitro* evolution of a beta-galactosidase by fluorescence-activated sorting of double emulsions. *Chem Biol* **12,** 1291-300.

46. Griffiths, A. D., and Tawfik, D. S. (2003) Directed evolution of an extremely fast phosphotriesterase by *in vitro* compartmentalization. *EMBO J* **22,** 24-35.

47. Stapleton, J. A., and Swartz, J. R. (2010) Development of an *in vitro* compartmentalization screen for high-throughput directed evolution of [FeFe] hydrogenases. *PLoS One* **5,** 1-8.

48. Kelly, B. T., and Griffiths, A. D. (2007) Selective gene amplification. *Protein Eng Des Sel* **20,** 577-81.

49. Sumida, T., Doi, N., and Yanagawa, H. (2009) Bicistronic DNA display for *in vitro* selection of Fab fragments. *Nucleic Acids Res* **37,** e147.

50. Ahn, J. H., Kang, T. J., and Kim, D. M. (2008) Tuning the expression level of recombinant proteins by modulating mRNA stability in a cell-free protein synthesis system. *Biotechnol Bioeng* **101,** 422-27.

51. 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,** 2256-60.

52. Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994) An *in-vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc Natl Acad Sci USA* **91,** 9022-26.

53. Hanes, J., and Plückthun, A. (1997) *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci USA* **94,** 4937-42.

54. Lipovsek, D., and Plückthun, A. (2004) *In vitro* protein evolution by ribosome display and mRNA display *J Immunol Methods* **290,** 51-67.

55. Jestin, J. L., and Kaminski, P. A. (2004) Directed enzyme evolution and selections for catalysis based on product formation. *J Biotechnol* **113,** 85-103.

56. Roberts, R. W., and Szostak, J. W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc Natl Acad Sci USA* **94,** 12297-302.

57. Nemoto, N., Miyamoto-Sato, E., Husimi, Y., and 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,** 405-08.

58. Kurz, M., Gu, K., and Lohse, P. A. (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.

59. Liu, R. H., Barrick, J. E., Szostak, J. W., and Roberts, R. W. (2000) Optimized synthesis of RNA-protein fusions for *in vitro* protein selection. *Methods Enzymol* **318,** 268-93.

60. Takahashi, T. T., and Roberts, R. W. (2009) *In vitro* selection of protein and peptide libraries using mRNA display. *Methods Mol Biol* **535,** 293-314.

61. Cotten, S. W., Zou, J. W., Valencia, C. A., and Liu, R. H. (2011) Selection of proteins with desired properties from natural proteome libraries using mRNA display. *Nat Protoc* **6,** 1163-82.

62. Kurz, M., Gu, K., Al-Gawari, A., and Lohse, P. A. (2001) cDNA - Protein fusions: covalent protein-gene conjugates for the *in vitro* selection of peptides and proteins. *Chembiochem* **2,** 666-72.

63. Ueno, S., and Nemoto, N. (2012) cDNA display: rapid stabilization of mRNA display. *Methods Mol Biol* **805,** 113-35.

64. Tawfik, D. S., and Griffiths, A. D. (1998) Man-made cell-like compartments for molecular evolution. *Nat Biotechol* **16,** 652-56.

65. Miller, O. J., Bernath, K., Agresti, J. J., Amitai, G., Kelly, B. T., Mastrobattista, E., Taly, V., Magdassi, S., Tawfik, D. S., and Griffiths, A. D. (2006) Directed evolution by *in vitro* compartmentalization. *Nat Methods* **3,** 561-70.

66. Bernath, K., Hai, M. T., Mastrobattista, E., Griffiths, A. D., Magdassi, S., and Tawfik, D. S. (2004) *In vitro* compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. *Anal Biochem* **325,** 151-57.

67. Ghadessy, F. J., and 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,** 201-04.

68. Ghadessy, F. J., Ong, J. L., and Holliger, P. (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci USA* **98,** 4552-57.

69. Eisenstein, M. (2006) Tiny droplets make a big splash. *Nat Methods* **3,** 71.

70. Song, H., and Ismagilov, R. F. (2003) Millisecond kinetics using nanoliters of reagents. *J Am Chem Soc* **125,** 14613-19.

71. Mazutis, L., J.C., B., Treacy, P., Skhiri, Y., Fallah Araghi, A., Ryckelynck, M., Taly, V., and Griffiths, A. D. (2009) Multi-step microfluidic droplet processing: kinetic analysis of an *in vitro* translated enzyme *Lab Chip* **9,** 2902-08.

72. Doi, N., and Yanagawa, H. (1999) STABLE: protein-DNA fusion system for screening of combinatorial protein libraries *in vitro*. *FEBS Lett* **457,** 227-30.

73. Bertschinger, J., and Neri, D. (2004) Covalent DNA display as a novel tool for directed evolution of proteins *in vitro*. *Prot Eng Des Sel* **17,** 699-707.

74. Bertschinger, J., Grabulovski, D., and Neri, D. (2007) Selection of single domain binding proteins by covalent DNA display. *Prot Eng Des Sel* **20,** 57-68.

75. Stein, V., Sielaff, I., Johnsson, K., and Hollfelder, F. (2007) A covalent chemical genotype-phenotype linkage for *in vitro* protein evolution. *Chembiochem* **8,** 2191-94.

76. Kaltenbach, M., Stein, V., and Hollfelder, F. (2011) SNAP dendrimers: multivalent protein display on dendrimer-like DNA for directed evolution. *Chembiochem* **12,** 2208-16.

77. Kries, H., Blomberg, R., and Hilvert, D. (2013) De novo enzymes by computational design. *Curr Opin in Chem Bio* **17,** 221-28.

78. Bolon, D. N., and Mayo, S. L. (2001) Enzyme-like proteins by computational design. *Proc Natl Acad Sci* **98,** 14274-79.

79. Richter, F., Blomberg, R., Khare, S. D., Kiss, G., Kuzin, A. P., Smith, A. J. T., Gallaher, J., Pianowski, Z., Helgeson, R. C., Grjasnow, A., Xiao, R., Seetharaman, J., Su, M., Vorobiev, S., Lew, S., Forouhar, F., Kornhaber, G. J., Hunt, J. F., Montelione, G. T., Tong, L., Houk, K. N., Hilvert, D., and Baker, D. (2012) Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.***134,** 16197-206.

80. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Räthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De Novo Computational Design of Retro-Aldol Enzymes. *Science* **319,** 1387-91.

81. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design. *Nature* **453,** 190-95.

82. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010) Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329,** 309-13.

83. Lu, Y., Yeung, N., Sieracki, N., and Marshall, N. M. (2009) Design of functional metalloproteins. *Nature* **460,** 855-62.

84. Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Players, F., Stoddard, B. L., Popovic, Z., and Baker, D. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotech* **30,** 190-92.

85. Bartel, D. P., and Szostak, J. W. (1993) Isolation of New Ribozymes from a Large Pool of Random Sequences. *Science* **261,** 1411-18.

86. Purtha, W. E., Coppins, R. L., Smalley, M. K., and Silverman, S. K. (2005) General Deoxyribozyme-Catalyzed Synthesis of Native 3' RNA Linkages. *J. Am. Chem. Soc.Society* **127,** 13124-25.

87. Ikawa, Y., Tsuda, K., Matsumura, S., and Inoue, T. (2004) De novo synthesis and development of an RNA enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 13750-55.

88. Jaeger, L., Wright, M. C., and Joyce, G. F. (1999) A complex ligase ribozyme evolved *in vitro* from a group I ribozyme domain. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 14712-17.

89. Cho, G. S., and Szostak, J. W. (2006) Directed evolution of ATP binding proteins from a zinc finger domain by using mRNA display. *Chem Biol* **13,** 139-47.

90. Hermanson, G. T. (2013) *in* "Bioconjugate Techniques (Third edition)", pp. 299-339, Academic Press, Boston.

Chapter 2

1. 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.

2. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828-31.

3. 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.

4. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321,** 741-65.

5. Wierenga, R. K. (2001) The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492,** 193-98.

6. Sterner, R., and Höcker, B. (2005) Catalytic versatility, stability, and evolution of the $(\beta\alpha)_8$-barrel enzyme fold. *Chemical Reviews* **105,** 4038-55.

7. Blacklow, S. C., Raines, R. T., Lim, W. A., Zamore, P. D., and Knowles, J. R. (1988) Triosephosphate isomerase catalysis is diffusion controlled. *Biochemistry* **27,** 1158-67.

8. Gerlt, J. A., and Raushel, F. M. (2003) Evolution of function in $(\beta/\alpha)_8$-barrel enzymes. *Curr. Opin. Chem. Biol.* **7,** 252-64.

9. Höcker, B., Claren, J., and Sterner, R. (2004) Mimicking enzyme evolution by generating new $(\beta\alpha)_8$-barrels from $(\beta\alpha)_4$-half-barrels. *Proc. Nat. Acad. Sci.* **101,** 16448-5453.

10. Leopoldseder, S., Claren, J., Jürgens, C., and Sterner, R. (2004) Interconverting the catalytic activities of $(\beta\alpha)_8$-barrel enzymes from different metabolic pathways: sequence requirements and molecular analysis. *J. Mol. Biol.* **337,** 871-79.

11. Claren, J., Malisi, C., Höcker, B., and Sterner, R. (2009) Establishing wild-type levels of catalytic activity on natural and artificial $(\beta/\alpha)_8$-barrel protein scaffolds. *Proc. Nat. Acad. Sci.* **106,** 3704-09.

12. Evran, S., Telefoncu, A., and Sterner, R. (2012) Directed evolution of $(\beta\alpha)_8$-barrel enzymes: establishing phosphoribosylanthranilate isomerisation activity on the scaffold of the tryptophan synthase α-subunit. *Prot. Eng. Des. Sel.* **25,** 285-93.

13. Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012) Engineering the third wave of biocatalysis. *Nature* **485,** 185-94.

14. Park, H.-S., Nam, S.-H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J., and Kim, H.-S. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. *Science* **311,** 535-8.

15. Tawfik, D. S. (2006) Loop grafting and the origins of enzyme species. *Science* **311,** 475-76.

16. Heinis, C., and Johnsson, K. (2010) Using peptide loop insertion mutagenesis for the evolution of proteins. *Meth. Mol. Biol.* **634,** 217-32.

17. Ochoa-Leyva, A., Barona-Gómez, F., Saab-Rincón, G., Verdel-Aranda, K., Sánchez, F., and Soberón, X. (2011) Exploring the structure-function loop adaptability of a $(\beta/\alpha)_8$ -barrel enzyme through loop swapping and hinge variability. *J. Mol. Biol.* **411,** 143-57.

18. Ochoa-Leyva, A., Soberón, X., Sánchez, F., Argüello, M., Montero-Morán, G., and Saab-Rincón, G. (2009) Protein design through systematic catalytic loop exchange in the $(\beta/\alpha)_8$ fold. *J. Mol. Biol.* **387,** 949-64.

19. Saab-Rincón, G., Olvera, L., Olvera, M., Rudiño-Piñera, E., Benites, E., Soberón, X., and Morett, E. (2012) Evolutionary walk between $(\beta/\alpha)_8$ barrels: catalytic migration from triosephosphate isomerase to thiamin phosphate synthase. *J. Mol. Biol.* **416,** 255-70.

20. Ma, H., and Penning, T. M. (1999) Conversion of mammalian 3α-hydroxysteroid dehydrogenase to 20α-hydroxysteroid dehydrogenase using loop chimeras: changing specificity from androgens to progestins. *Proc. Nat. Acad. Sci.* **96,** 11161-66.

21. Campbell, E., Chuang, S., and Banta, S. (2013) Modular exchange of substrate-binding loops alters both substrate and cofactor specificity in a member of the aldo-keto reductase superfamily. *Prot. Eng. Des. Sel.* **26,** 181-86.

22. Griffiths, A. D., and Tawfik, D. S. (2003) Directed evolution of an extremely fast phosphotriesterase by *in vitro* compartmentalization. *EMBO J.* **22,** 24-35.

23. Vick, J. E., Schmidt, D. M. Z., and Gerlt, J. A. (2005) Evolutionary potential of $(\beta/\alpha)_8$-barrels: *in vitro* enhancement of a "new" reaction in the enolase superfamily. *Biochemistry* **44,** 11722-29.

24. Tsai, P.-C., Fox, N., Bigley, A. N., Harvey, S. P., Barondeau, D. P., and Raushel, F. M. (2012) Enzymes for the homeland defense: optimizing phosphotriesterase for the hydrolysis of organophosphate nerve agents. *Biochemistry* **51,** 6463-75.

25. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) *De novo* computational design of retro-aldol enzymes. *Science* **319,** 1387-91.

26. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design. *Nature* **453,** 190-95.

27. Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N., and Mayo, S. L. (2012) Iterative approach to computational enzyme design. *Proc. Nat. Acad. Sci.* **109,** 3790-95.

28. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329,** 309-13.

29. Baker, D. (2010) An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19,** 1817-19.

30. Golynskiy, M. V., and Seelig, B. (2010) *De novo* enzymes: from computational design to mRNA display. *Trends Biotechnol.* **28,** 340-45.

31. Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19,** 596-604.

32. Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc. Nat. Acad. Sci.* **103,** 5869-74.

33. Zeldovich, K. B., Chen, P., and Shakhnovich, E. I. (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Nat. Acad. Sci.* **104,** 16152-57.

34. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C., and Arnold, F. H. (2005) Thermodynamic prediction of protein neutrality. *Proc. Nat. Acad. Sci.* **102,** 606-11.

35. Cho, G., Keefe, A. D., Liu, R., Wilson, D. S., and Szostak, J. W. (2000) Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J. Mol. Biol.* **297,** 309-19.

36. Sieber, V., Pluckthun, A., and Schmid, F. X. (1998) Selecting proteins with improved stability by a phage-based method. *Nat. Biotech.* **16,** 955-60.

37. Matsuura, T., and Plückthun, A. (2004) Strategies for selection from protein libraries composed of *de novo* designed secondary structure modules. *Origins Life Evol. B.* **34,** 151-57.

38. Matsuura, T., and Plückthun, A. (2003) Selection based on the folding properties of proteins with ribosome display. *FEBS Lett.* **539,** 24-28.

39. Schmid, F.-X. (2012) Lessons about protein stability from *in vitro* selections. *ChemBioChem* **12,** 1501-07.

40. Khersonsky, O., and Tawfik, D. S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79,** 471-505.

41. Santelli, E., Schwarzenbacher, R., McMullan, D., Biorac, T., Brinen, L. S., Canaves, J. M., Cambell, J., Dai, X., Deacon, A. M., Elsliger, M.-A., Eshagi, S., Floyd, R., Godzik, A., Grittini, C., Grzechnik, S. K., Jaroszewski, L., Karlak, C., Klock, H. E., Koesema, E., Kovarik, J. S., Kreusch, A., Kuhn, P., Lesley, S. a., McPhillips, T. M., Miller, M. D., Morse, A., Moy, K., Ouyang, J., Page, R., Quijano, K., Rezezadeh, F., Robb, A., Sims, E., Spraggon, G., Stevens, R. C., van den Bedem, H., Velasquez, J., Vincent, J., von Delft, F., Wang, X., West, B., Wolf, G., Xu, Q., Hodgson, K. O., Wooley, J., and Wilson, I. a. (2004) Crystal structure of a glycerophosphodiester phosphodiesterase (GDPD) from *Thermotoga maritima* (TM1621) at 1.60 A resolution. *Proteins* **56,** 167-70.

42. Roberts, R. W., and Szostak, J. W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Nat. Acad. Sci.* **94,** 12297-302.

43. Seelig, B. (2011) mRNA display for the selection and evolution of enzymes from *in vitro*-translated protein libraries. *Nat. Protocols* **6,** 540-52.

44. Cho, G. S., and Szostak, J. W. (2006) Directed evolution of ATP binding proteins from a zinc finger domain by using mRNA display. *Chem. Biol.* **13,** 139-47.

45. Moelbert, S., Emberly, E., and Tang, C. (2004) Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci.* **13,** 752-62.

46. Seitz, T., Thoma, R., Schoch, G. A., Stihle, M., Benz, J., D'Arcy, B., Wiget, A., Ruf, A., Hennig, M., and Sterner, R. (2010) Enhancing the stability and solubility of the glucocorticoid receptor ligand-binding domain by high-throughput library screening. *J. Mol. Biol.* **403,** 562-77.

47. Graziano, J. J., Liu, W., Perera, R., Geierstanger, B. H., Lesley, S. A., and Schultz, P. G. (2008) Selecting folded proteins from a library of secondary structural elements. *J. Am. Chem. Soc.* **130,** 176-85.

48. Pédelacq, J.-D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C.-Y., Rho, B.-S., Park, M. S., Terwilliger, T. C., and Waldo, G. S. (2002) Engineering soluble proteins for structural genomics. *Nat. Biotech.* **20,** 927-32.

49. Amar, D., Berger, I., Amara, N., Tafa, G., Meijler, M. M., and Aharoni, A. (2012) The transition of human estrogen sulfotransferase from generalist to specialist using directed enzyme evolution. *J. Mol. Biol.* **416,** 21-32.

50. Copley, S. D. (2012) Toward a systems biology perspective on enzyme evolution. *J. Biol. Chem.* **287,** 3-10.

51. Stryer, L. (1965) The interaction of a napthalene dye with apomyoglobin and apohemoglobin - a fluorescent probe of non-polar binding sites. *J. Mol. Biol.* **13,** 482-95.

52. Sambrook, J., and Russell, D. W. (2001) Molecular cloning: a laboratory manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Chapter 3

1. Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012) Engineering the third wave of biocatalysis. *Nature* **485,** 185-94.

2. Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19,** 596-604.

3. Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 5869-74.

4. Bommarius, A. S., and Paye, M. F. (2013) Stabilizing biocatalysts. *Chem Soc Rev* **42,** 6534-65.

5. Wijma, H. J., Floor, R. J., and Janssen, D. B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*

6. Romero, P. A., and Arnold, F. H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10,** 866-76.

7. Eijsink, V. G. H., Gåseidnes, S., Borchert, T. V., and van den Burg, B. (2005) Directed evolution of enzyme stability. *Biomol. Eng.* **22,** 21-30.

8. Wigley, W. C., Stidham, R. D., Smith, N. M., Hunt, J. F., and Thomas, P. J. (2001) Protein solubility and folding monitored *in vivo* by structural complementation of a genetic marker protein. *Nat. Biotechnol.* **19,** 131-36.

9. Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999) Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17,** 691-95.

10. Sieber, V., Pluckthun, A., and Schmid, F. X. (1998) Selecting proteins with improved stability by a phage-based method. *Nat. Biotechnol.* **16,** 955-60.

11. Martin, A., Sieber, V., and Schmid, F. X. (2001) In-vitro Selection of Highly Stabilized Protein Variants with Optimized Surface. *J. Mol. Biol.* **309,** 717-26.

12. Socha, R. D., and Tokuriki, N. (2013) Modulating protein stability – directed evolution strategies for improved protein function. *FEBS J.* **280,** 5582-95.

13. Golynskiy, M. V., Haugner III, J. C., Morelli, A., Morrone, D., and Seelig, B. (2013) *In vitro* evolution of enzymes. *Methods Mol. Biol.* **978,** 73-92.

14. Schmid, F. X. (2011) Lessons about Protein Stability from *in vitro* Selections. *Chembiochem* **12,** 1501-07.

15. Jäckel, C., Bloom, J. D., Kast, P., Arnold, F. H., and Hilvert, D. (2010) Consensus protein design without phylogenetic bias. *J. Mol. Biol.* **399,** 541-46.

16. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828-31.

17. Roberts, R. W., and Szostak, J. W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 12297-302.

18. Nemoto, N., MiyamotoSato, E., Husimi, Y., and 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,** 405-08.

19. Chao, F.-A., Morelli, A., Haugner, J. C., III, 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.

20. Haugner III, J. C., and Seelig, B. (2013) Universal labeling of 5'-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity. *Chem. Commun.* **49,** 7322-24.

21. Seelig, B. (2011) mRNA display for the selection and evolution of enzymes from *in vitro*-translated protein libraries. *Nat Protocols* **6,** 540-52.

22. Moore, M. J., and Sharp, P. A. (1992) Site-specific modification of pre-mRNA: the 2'-hydroxyl groups at the splice sites. *Science* **256,** 992-97.

23. Greenfield, N. J. (2006) Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat Protoc* **1,** 2527-35.

24. Diaz, J. E., Lin, C.-S., Kunishiro, K., Feld, B. K., Avrantinis, S. K., Bronson, J., Greaves, J., Saven, J. G., and Weiss, G. A. (2011) Computational design and selections for an engineered, thermostable terpene synthase. *Protein Sci.* **20,** 1597-606.

25. Reetz, M. T., Soni, P., Acevedo, J. P., and 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,** 8268-72.

26. Palackal, N., Brennan, Y., Callen, W. N., Dupree, P., Frey, G., Goubet, F., Hazlewood, G. P., Healey, S., Kang, Y. E., Kretz, K. A., Lee, E., Tan, X. Q., Tomlinson, G. L., Verruto, J., Wong, V. W. K., Mathur, E. J., Short, J. M., Robertson, D. E., and Steer, B. A. (2004) An evolutionary route to xylanase process fitness. *Protein Sci.* **13,** 494-503.

27. Henzler-Wildman, K., and Kern, D. (2007) Dynamic personalities of proteins. *Nature* **450,** 964-72.

28. Nashine, V. C., Hammes-Schiffer, S., and Benkovic, S. J. (2010) Coupled motions in enzyme catalysis. *Curr Opin Chem Biol* **14,** 644-51.

29. Ramanathan, A., and Agarwal, P. K. (2011) Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. *PLoS Biol.* **9**.

30. Auerbach, G., Ostendorp, R., Prade, L., Korndorfer, I., Dams, T., Huber, R., and Jaenicke, R. (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,** 769-81.

31. Russell, R. J., Gerike, U., Danson, M. J., Hough, D. W., and Taylor, G. L. (1998) Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. *Structure* **6,** 351-61.

32. Arnold, F. H., Wintrode, P. L., Miyazaki, K., and Gershenson, A. (2001) How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **26,** 100-06.

33. Macedo-Ribeiro, S., Darimont, B., Sterner, R., and Huber, R. (1996) Small structural changes account for the high thermostability of 1[4Fe-4S] ferredoxin from the hyperthermophilic bacterium *Thermotoga maritima*. *Structure* **4,** 1291-301.

34. Elias, M., Wieczorek, G., Rosenne, S., and Tawfik, D. S. (2014) The universality of enzymatic rate-temperature dependency. *Trends Biochem. Sci.* **39,** 1-7.

35. Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., and Tawfik, D. S. (2009) Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* **34,** 53-9.

36. Golynskiy, M. V., and Seelig, B. (2010) *De novo* enzymes: from computational design to mRNA display. *Trends Biotechnol.* **28,** 340-45.

37. Golynskiy, M. V., Haugner, J. C., and Seelig, B. (2013) Highly diverse protein library based on the ubiquitous $(b/a)_8$ enzyme fold yields well-structured proteins through *in vitro* folding selection. *ChemBioChem* **14,** 1553-63.

38. Chaput, J. C., and Szostak, J. W. (2004) Evolutionary optimization of a nonbiological ATP binding protein for improved folding stability. *Chem. Biol.* **11,** 865-74.

39. Smith, M. D., Rosenow, M. A., Wang, M. T., Allen, J. P., Szostak, J. W., and Chaput, J. C. (2007) Structural insights into the evolution of a non-biological protein: importance of surface residues in protein fold optimization. *PLoS ONE* **2**.

40. Cho, G. S., and Szostak, J. W. (2006) Directed evolution of ATP binding proteins from a zinc finger domain by using mRNA display. *Chem. Biol.* **13,** 139-47.

41. Hall, T. A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41,** 95-98.

42. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (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,** 4673-80.

Chapter 4

1. Chothia, C. (1992) Proteins - 1000 families for the molecular biologist. *Nature* **357,** 543-44.

2. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP - a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536-40.

3. Ohno, S. (1971) Evolution by gene duplication, Springer-Verlag, New York, USA.

4. Chothia, C., Gough, J., Vogel, C., and Teichmann, S. A. (2003) Evolution of the protein repertoire. *Science* **300,** 1701-03.

5. James, L. C., and Tawfik, D. S. (2003) Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28,** 361-68.

6. Tokuriki, N., and Tawfik, D. S. (2009) Protein Dynamism and Evolvability. *Science* **324,** 203-07.

7. Cordes, M. H. J., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (1999) Evolution of a Protein Fold in Vitro. *Science* **284,** 325-27.

8. Kaplan, J., and DeGrado, W. F. (2004) De novo design of catalytic proteins. *Proc. Natl. Acad. Sci. USA* **101,** 11566-70.

9. Tuinstra, R. L., Peterson, F. C., Kutlesa, S., Elgin, E. S., Kron, M. A., and Volkman, B. F. (2008) Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl. Acad. Sci. USA* **105,** 5057-62.

10. Bryan, P. N., and Orban, J. (2010) Proteins that switch folds. *Curr. Opin. Struct. Biol.* **20,** 482-88.

11. Smith, B. A., and Hecht, M. H. (2011) Novel proteins: from fold to function. *Curr. Opin. Chem. Biol.* **15,** 421-26.

12. Keefe, A. D., and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature* **410,** 715-18.

13. Mansy, S. S., Zhang, J. L., Kummerle, R., Nilsson, M., Chou, J. J., Szostak, J. W., and Chaput, J. C. (2007) Structure and evolutionary analysis of a non-biological ATP-binding protein. *J. Mol. Biol.* **371,** 501-13.

14. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828-31.

15. Seelig, B. (2011) mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries. *Nat. Protoc.* **6,** 540-52.

16. Holmbeck, S. M. A., Foster, M. P., Casimiro, D. R., Sem, D. S., Dyson, H. J., and Wright, P. E. (1998) High-resolution solution structure of the retinoid X receptor DNA-binding domain. *J. Mol. Biol.* **281,** 271-84.

17. Cho, G. S., and Szostak, J. W. (2006) Directed Evolution of ATP Binding Proteins from a Zinc Finger Domain by Using mRNA Display. *Chem. Biol.* **13,** 139-47.

18. Zhao, Q., Chasse, S. A., Devarakonda, S., Sierk, M. L., Ahvazi, B., and Rastinejad, F. (2000) Structural basis of RXR-DNA interactions. *J. Mol. Biol.* **296,** 509-20.

19. Maret, W., and Li, Y. (2009) Coordination dynamics of zinc in proteins. *Chem. Rev.* **109,** 4682-707.

20. van Tilborg, P. J., Czisch, M., Mulder, F. A., Folkers, G. E., Bonvin, A. M., Nair, M., Boelens, R., and Kaptein, R. (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,** 8747-57.

21. Yang, W., Lee, J. Y., and Nowotny, M. (2006) Making and breaking nucleic acids: Two-Mg2+-ion catalysis and substrate specificity. *Mol. Cell* **22,** 5-13.

22. Bhabha, G., Lee, J., Ekiert, D. C., Gam, J., Wilson, I. A., Dyson, H. J., Benkovic, S. J., and Wright, P. E. (2011) A Dynamic Knockout Reveals That Conformational Fluctuations Influence the Chemical Step of Enzyme Catalysis. *Science* **332,** 234-38.

23. Baldwin, A. J., and Kay, L. E. (2009) NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* **5,** 808-14.

24. Henzler-Wildman, K., and Kern, D. (2007) Dynamic personalities of proteins. *Nature* **450,** 964-72.

25. Golynskiy, M. V., and Seelig, B. (2010) De novo enzymes: from computational design to mRNA display. *Trends Biotechnol.* **28,** 340-45.

26. Grzesiek, S., and Bax, A. (1992) Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J. Magn. Reson.* **96,** 432-40.

27. Muhandiram, D. R., and Kay, L. E. (1994) Gradient-Enhanced Triple-Resonance Three-Dimensional NMR Experiments with Improved Sensitivity. *J. Magn. Reson., Ser. B* **103,** 203-16.

28. Wittekind, M., and 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,** 201-05.

29. Eghbalnia, H. R., Bahrami, A., Tonelli, M., Hallenga, K., and Markley, J. L. (2005) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. *J. Am. Chem. Soc.* **127,** 12528-36.

30. Grzesiek, S., Anglister, J., and 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,** 114-19.

31. Wuthrich, K. (1986) NMR of proteins and nucleic acids, John Wiley and Sons, New York, USA.

32. Wishart, D. S., Sykes, B. D., and Richards, F. M. (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **222,** 311-33.

33. Vuister, G. W., and 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,** 7772-77.

34. Lee, D., Hilty, C., Wider, G., and Wuthrich, K. (2006) Effective rotational correlation times of proteins from NMR relaxation interference. *J. Magn. Reson.* **178,** 72-76.

35. Gagne, S. M., Tsuda, S., Li, M. X., Chandra, M., Smillie, L. B., and Sykes, B. D. (1994) Quantification of the calcium-induced secondary structural changes in the regulatory domain of troponin-C. *Protein Sci.* **3,** 1961-74.

36. Wang, Y., Zhao, S., Somerville, R. L., and Jardetzky, O. (2001) Solution structure of the DNA-binding domain of the TyrR protein of Haemophilus influenzae. *Protein Sci.* **10,** 592-98.

37. Ruckert, M., and Otting, G. (2000) Alignment of Biological Macromolecules in Novel Nonionic Liquid Crystalline Media for NMR Experiments. *J. Am. Chem. Soc.* **122,** 7793-97.

38. Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160,** 65-73.

39. Alberts, I. L., Nadassy, K., and Wodak, S. J. (1998) Analysis of zinc binding sites in protein crystal structures. *Protein Sci.* **7,** 1700-16.

40. Viles, J. H., Patel, S. U., Mitchell, J. B., Moody, C. M., Justice, D. E., Uppenbrink, J., Doyle, P. M., Harris, C. J., Sadler, P. J., and Thornton, J. M. (1998) Design, synthesis and structure of a zinc finger with an artificial beta-turn. *J. Mol. Biol.* **279,** 973-86.

41. Ohlenschlager, O., Seiboth, T., Zengerling, H., Briese, L., Marchanka, A., Ramachandran, R., Baum, M., Korbas, M., Meyer-Klaucke, W., Durst, M., and Gorlach, M. (2006) Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. *Oncogene* **25,** 5953-59.

42. Banci, L., Bertini, I., Del Conte, R., Mangani, S., and 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,** 2467-74.

43. Tenderholt, A. (2007) Pyspline, Stanford University, Stanford, USA.

44. Deleon, J. M., Rehr, J. J., Zabinsky, S. I., and Albers, R. C. (1991) ABINITIO CURVED-WAVE X-RAY-ABSORPTION FINE-STRUCTURE. *Phys. Rev. B* **44,** 4146-56.

45. Rehr, J. J., and Albers, R. C. (2000) Theoretical approaches to x-ray absorption fine structure. *Rev. Mod. Phys.* **72,** 621-54.

46. Rehr, J. J., Deleon, J. M., Zabinsky, S. I., and Albers, R. C. (1991) THEORETICAL X-RAY ABSORPTION FINE-STRUCTURE STANDARDS. *J. Am. Chem. Soc.* **113,** 5135-40.

47. Kim, C. A., and Berg, J. M. (1996) A 2.2 angstrom resolution crystal structure of a designed zinc finger protein bound to DNA. *Nat. Struct. Biol.* **3,** 940-45.

48. George, G. N. (2000) EXAFSSPAK and EDG-FIT, Stanford Synchrotron Radiation Lightsource, Menlo Park, USA.

49. Patel, K., Kumar, A., and Durani, S. (2007) Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. *BBA-Proteins Proteom.* **1774,** 1247-53.

50. Kupper, H., Mijovilovich, A., Meyer-Klaucke, W., and Kroneck, P. M. H. (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,** 748-57.

51. Clark-Baldwin, K., Tierney, D. L., Govindaswamy, N., Gruff, E. S., Kim, C., Berg, J., Koch, S. A., and Penner-Hahn, J. E. (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,** 8401-09.

52. Penner-Hahn, J. E. (2005) Characterization of "spectroscopically quiet" metals in biology. *Coord. Chem. Rev.* **249,** 161-77.

53. Bobyr, E., Lassila, J. K., Wiersma-Koch, H. I., Fenn, T. D., Lee, J. J., Nikolic-Hughes, I., Hodgson, K. O., Rees, D. C., Hedman, B., and Herschlag, D. (2012) High-Resolution Analysis of Zn2+ Coordination in the Alkaline Phosphatase Superfamily by EXAFS and X-ray Crystallography. *J. Mol. Biol.* **415,** 102-17.

Chapter 5

1. Lehman, I. R. (1974) DNA Ligase: Structure, Mechanism, and Function. *Science* **186,** 790-97.

2. Shuman, S. (2009) DNA Ligases: Progress and Prospects. *J. Biol. Chem.* **284,** 17365-69.

3. Moore, M. J., and Sharp, P. A. (1992) Site-Specific Modification of Pre-mRNA: The 2'-Hydroxyl Groups at the Splice Sites. *Science* **256,** 992-97.

4. Moore, M. J., and Query, C. C. (2000) *in* "Methods Enzymol." Vol. 317, pp. 109-23, Academic Press.

5. Ozsolak, F., and Milos, P. M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87-98.

6. Hornung, V., Ellegast, J., Kim, S., Brzózka, K., Jung, A., Kato, H., Poeck, H., Akira, S., Conzelmann, K.-K., Schlee, M., Endres, S., and Hartmann, G. (2006) 5'-Triphosphate RNA Is the Ligand for RIG-I. *Science* **314,** 994-97.

7. Clayton, D. A. (1984) Transcription of the mammalian mitochondrial genome. *Annu. Rev. Biochem.* **53,** 573-94.

8. Sugiura, M. (1992) The chloroplast genome. *Plant Mol. Biol.* **19,** 149-68.

9. Sijen, T., Steiner, F. A., Thijssen, K. L., and Plasterk, R. H. A. (2007) Secondary siRNAs Result from Unprimed RNA Synthesis and Form a Distinct Class. *Science* **315,** 244-47.

10. Pak, J., and Fire, A. (2007) Distinct Populations of Primary and Secondary Effectors During RNAi in C. elegans. *Science* **315,** 241-44.

11. Qiu, Y., Cho, B.-K., Park, Y. S., Lovley, D., Palsson, B. Ø., and Zengler, K. (2010) Structural and operational complexity of the Geobacter sulfurreducens genome. *Genome Res.* **20,** 1304-11.

12. Bartel, D. P., and Szostak, J. W. (1993) Isolation of New Ribozymes from a Large Pool of Random Sequences. *Science* **261,** 1411-18.

13. Jaeger, L., Wright, M. C., and Joyce, G. F. (1999) A complex ligase ribozyme evolved in vitro from a group I ribozyme domain. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 14712-17.

14. Ikawa, Y., Tsuda, K., Matsumura, S., and Inoue, T. (2004) De novo synthesis and development of an RNA enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 13750-55.

15. Purtha, W. E., Coppins, R. L., Smalley, M. K., and Silverman, S. K. (2005) General Deoxyribozyme-Catalyzed Synthesis of Native 3'-5' RNA Linkages. *J. Am. Chem. Soc.* **127,** 13124-25.

16. Flynn-Charlebois, A., Wang, Y. M., Prior, T. K., Rashid, I., Hoadley, K. A., Coppins, R. L., Wolf, A. C., and Silverman, S. K. (2003) Deoxyribozymes with 2 '-5 ' RNA ligase activity. *J. Am. Chem. Soc.* **125,** 2444-54.

17. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828-31.

18. Seelig, B. (2011) mRNA display for the selection and evolution of enzymes from *in vitro*-translated protein libraries. *Nat. Protocols* **6,** 540-52.

19. Chao, F.-A., Morelli, A., Haugner, J. C., III, 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.

20. Kuersten, S. (2010) (Pat., U., Ed.), US20100279305A1, US.

21. An excess of enzyme was used to reduce ligation times.

22. 5'-Triphosphate RNA is commonly synthesized using T7 polymerase, which strongly disfavors the incorporation of 5′-C or U.

23. Lohman, G. J. S., Chen, L., and Evans, T. C. (2011) Kinetic Characterization of Single Strand Break Ligation in Duplex DNA by T4 DNA Ligase. *J. Biol. Chem.* **286,** 44187-96.

24. While k_{obs} of ligase 10C is approximately 3 orders of magnitude lower than k_{cat} of T4 DNA ligase, ligations are completed after overnight incubation.

25. Aoki, K., Moriguchi, H., Yoshioka, T., Okawa, K., and Tabara, H. (2007) In vitro analyses of the production and activity of secondary small interfering RNAs in C. elegans. *EMBO J* **26,** 5007-19.

26. Golynskiy, M. V., and Seelig, B. (2010) *De novo* enzymes: from computational design to mRNA display. *Trends Biotechnol.* **28,** 340-45.

27. 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.

Chapter 6

1. Pak, J., and Fire, A. (2007) Distinct Populations of Primary and Secondary Effectors During RNAi in C. elegans. *Science* **315,** 241-44.

2. Sijen, T., Steiner, F. A., Thijssen, K. L., and Plasterk, R. H. A. (2007) Secondary siRNAs Result from Unprimed RNA Synthesis and Form a Distinct Class. *Science* **315,** 244-47.

3. Clayton, D. A. (1984) Transcription of the mammalian mitrochondrial genome. *Annu. Rev. Biochem.* **53,** 573-94.

4. Sugiura, M. (1992) The chloroplast genome. *Plant Mol. Biol.* **19,** 149-68.

5. Hornung, V., Ellegast, J., Kim, S., Brzózka, K., Jung, A., Kato, H., Poeck, H., Akira, S., Conzelmann, K.-K., Schlee, M., Endres, S., and Hartmann, G. (2006) 5'-Triphosphate RNA Is the Ligand for RIG-I. *Science* **314,** 994-97.

6. Qiu, Y., Cho, B.-K., Park, Y. S., Lovley, D., Palsson, B. Ø., and Zengler, K. (2010) Structural and operational complexity of the Geobacter sulfurreducens genome. *Genome Res.* **20,** 1304-11.

7. Fouquier d'He´ rouel, A., Wessner, F. o., Halpern, D., Ly-Vu, J., Kennedy, S. P., Serror, P., Aurell, E., and Repoila, F. (2011) A simple and efficient method to search for selected primary transcripts: non-coding and antisense RNAs in the human pathogen Enterococcus faecalis. *Nucleic Acids Research* **39,** e46.

8. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828-31.

9. Haugner III, J. C., and Seelig, B. (2013) Universal labeling of 5'-triphosphate RNAs by artificial RNA ligase enzyme with broad substrate specificity. *Chemical Communications* **49,** 7322-24.

10. Chao, F.-A., Morelli, A., Haugner, J. C., III, 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.

11. Langeland, B. r. T., Morris, D. L., and McKinley-McKee, J. S. (1999) Metal binding properties of thiols; complexes with horse liver alcohol dehydrogenase. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **123,** 155-62.

12. Scheller, K. H., Abel, T. H. J., Polanyi, P. E., Wenk, P. K., Fischer, B. E., and Sigel, H. (1980) Metal Ion/Buffer Interactions. *European Journal of Biochemistry* **107,** 455-66.

Conclusions and Future Directions

1. Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012) Engineering the third wave of biocatalysis. *Nature* **485,** 185-94.

2. Huisman, G. W., and Collier, S. J. (2013) On the development of new biocatalytic processes for practical pharmaceutical synthesis. *Curr Opin in Chem Bioy* **17,** 284-92.

3. Aehle, W. (2007) Enzymes in Industry. *Wiley-VCH Verlag GmbH & Co. KGaA*.

4. Bickle, T. A., and Krager, D. H. (1993) Biology of DNA restriction. *Microbio Revi* **57,** 434-50.

5. Barrett, A., Rawlings, N., and Woessner, J. (2004) Handbook of proteolytic enzymes. *Elsvier Academic Press*.

6. Waidmann, M. S., Bleichrodt, F. S., Laslo, T., and Riedel, C. U. (2011) Bacterial luciferase reporters: The Swiss army knife of molecular biology. *Bioengineered* **2,** 8-16.

7. Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S., and Milo, R. (2011) The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50,** 4402-10.

8. Rueping, M., and Nachtsheim, B. J. (2010) A review of new developments in the Friedel-Crafts alkylation: from green chemistry to asymmetric catalysis. *Beilstein J of Org Chem* **6,** 6.

9. Han, F.-S. (2013) Transition-metal-catalyzed Suzuki-Miyaura cross-coupling reactions: a remarkable advance from palladium to nickel catalysts. *Chem Society Reviews* **42,** 5270-98.