



mRNA display for the *in vitro* evolution of artificial proteins and enzymes

Christopher Nicholas Rowley

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
Astbury Centre for Structural Molecular Biology

September 2016

The candidate confirms that the submitted work is his own, and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from this thesis may be published without proper acknowledgement.

© 2016 The University of Leeds and Christopher Rowley

Acknowledgements

First and foremost, I would like to thank Professors Alan Berry and Peter Stockley for their continuous support, guidance, and encouragement, and for giving me the opportunity to work on such a stimulating and demanding project. I would also like to thank Professor Adam Nelson for patient and invaluable discussions during my (ongoing) introduction to the world of organic chemistry.

Thanks also goes to all members of the Berry, Stockley, and Nelson laboratories, past and present, whose advice and support has been invaluable throughout. In particular Laura Cross, for the many insightful deliberations on the intricacies of peroxisomal import, and her input into much of the PEX5 selection work.

It would be remiss of me not to thank all the friends that I have gained during my time in the Astbury centre, for their willingness to continue discussions (scientific or otherwise) beyond the laboratory.

Last but not least, I would like to thank my family, my Mum, Dad, and Sister for their unwavering support, which has been instrumental in me being able to pursue my interests.

Abstract

Artificial proteins and enzymes have the potential to aid in the production of pharmaceuticals and to facilitate basic biomedical research. Two methods currently exist for the development of artificial proteins: rational design and *de novo* selection. Rational design requires detailed knowledge of enzyme catalysis in order to design an enzyme active site *in silico*, and then introduce this active site into a protein. However, gaps in the understanding of protein folding and structure-function relationships make this approach challenging and far from routine. In contrast, laboratory evolution approaches to isolate artificial proteins and enzymes from libraries of variants are well established.

In vitro selection techniques are powerful tools for the exploration of large areas of sequence space (up to 10^{13} unique sequences) in the search for functional proteins and enzymes. mRNA display selection methods have only recently been developed, and the application of this technique for the engineering of *de novo* enzymes has not been fully explored. This thesis describes the establishment of an mRNA display platform for the selection and evolution of novel proteins and enzymes from large, high-diversity libraries. The synthesis of novel selection substrates are described that will facilitate the application of mRNA display to the selection of Diels-Alderase enzymes. A novel application of mRNA display is described for the solution-phase selection of protein-ligand pairs using interaction-dependent reverse transcription. Further development of this research could increase the throughput of ligand discovery to complement the pace at which new macromolecular targets of interest are being discovered.

The ability to generate tailor-made enzymes that catalyse novel reactions is of considerable interest. The applications of mRNA display selection described in this thesis will help to extend the range of enzyme catalysis, and to elucidate basic mechanisms of biocatalysis and protein evolution. Moreover, such 'designer enzymes' hold promise for a huge range of applications including, but not limited to, the synthesis of chemicals, pharmaceuticals, and production of renewable fuels.

Table of Contents

Acknowledgements	i
Abstract	ii
Table of Contents	iii
List of Figures	vii
List of Tables	x
List of Abbreviations	xi
1 Introduction	1
1.1 New proteins and enzymes -----	1
1.2 Protein engineering and directed evolution-----	2
1.2.1 Genetic Diversity	3
1.2.2 Interrogating genetically diverse libraries	8
1.2.3 Rational approaches	10
1.3 <i>In vitro</i> techniques for directed evolution -----	12
1.3.1 Benefits of <i>in vitro</i> directed evolution	15
1.3.2 Ribosome display	16
1.3.3 <i>In vitro</i> compartmentalisation	17
1.3.4 DNA display	18
1.3.5 mRNA display	19
1.3.6 mRNA display scaffolds and library design.....	21
1.4 Aims and objectives -----	24
2 Materials and Methods	26
2.1 Materials -----	26
2.1.1 Bacterial strains and plasmids.....	26
2.1.2 Chemicals and reagents.....	26
2.2 General Methods -----	26
2.2.1 General experimental.....	26
2.2.2 Protein and Nucleic acid sequence alignment	26
2.2.3 pH measurements.....	27
2.2.4 Culture growth.....	27
2.2.5 Transformation.....	27
2.2.6 Flash chromatography	27
2.2.7 Mass spectrometry.....	28
2.2.8 NMR.....	28

2.3	Nucleic acid methods -----	28
2.3.1	Phenol:chloroform extraction.....	28
2.3.2	Ethanol precipitation.....	29
2.3.3	Nucleic acid quantification.....	29
2.3.4	Ligation-independent cloning	29
2.3.5	Plasmid DNA purification.....	29
2.3.6	Native gel electrophoresis	29
2.3.7	Denaturing PAGE.....	30
2.3.8	Synthetic Gene Design	30
2.3.9	PEX5 receptor library construction	30
2.4	mRNA-display methods -----	31
2.4.1	Modification of DNA templates at the 5' and 3' ends by PCR	31
2.4.2	<i>In vitro</i> transcription.....	31
2.4.3	Puromycin linker synthesis.....	32
2.4.4	Photo-crosslinking reactions	32
2.4.5	Translation <i>in vitro</i> and mRNA-protein fusion formation.....	32
2.4.6	Oligo-dT cellulose synthesis.....	33
2.4.7	Oligo-dT cellulose binding assay.....	33
2.4.8	Oligo-dT cellulose purification of mRNA-protein fusions	33
2.4.9	Preparation of peptide-coupled agarose beads	34
2.4.10	PEX5*-peptide selection.....	34
2.4.11	Interaction-dependent RT-PCR assay.....	35
2.4.12	Mock library selection by IDRT-PCR.....	36
2.5	Protein methods-----	36
2.5.1	Determination of protein concentration.....	36
2.5.2	Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)	36
2.5.3	Ni ²⁺ -NTA purification of His ₆ -tagged proteins from <i>E.coli</i>	37
2.5.4	Dialysis and concentration	38
2.5.5	Covalent coupling of DNA oligonucleotides to streptavidin	38
2.6	mRNA display substrate synthesis -----	39
2.6.1	Synthesis of 5'-acrylamide oligonucleotides	39
2.6.2	p-(methoxycarbonyl)benzyl trans-1,3-butadiene-1-carbamate.....	39
2.6.3	4-Carboxybenzyl <i>trans</i> -1,3-butadiene-1-carbamate.....	40
2.6.4	N-Boc-norbiotinamine	40
2.6.5	Norbiotinamine trifluoroacetate.....	41

2.6.6	4-carboxybenzyl trans-1,3-butadiene-1-carbamate norbiotinamide	41
3	Generating RNA-protein fusions for <i>in vitro</i> selection.....	43
3.1	Synthesis of 3'-puromycin oligonucleotides and templates	45
3.2	<i>In vitro</i> translation and mRNA-protein fusion formation.....	52
3.3	Synthesis of oligo-dT cellulose	57
3.4	Purification of mRNA-protein fusions using oligo-dT cellulose.....	59
3.5	Synthesis of tools for the selection of bond forming enzymes by mRNA display	61
3.5.1	Synthesis of a dienophile-linked reverse transcription primer	63
3.5.2	Synthesis of a biotinylated diene	65
3.6	Summary.....	66
4	Selection for orthogonal ligand-receptor pair using mRNA-display....	69
4.1	PEX5 RNA-protein fusion generation	73
4.2	Construction of the PEX5 mutant library	76
4.3	Selection for orthogonal PEX5-PTS1 interactions	80
4.4	Characterisation of selected mutant – PEX5.YY.4.3	82
4.5	Analysis of YQSYY library sequences	85
4.6	Summary.....	87
5	mRNA display with interaction-dependent reverse transcription.....	91
5.1	Interaction-dependent PCR	91
5.2	Interaction-dependent reverse transcription (IDRT) using RNA-protein fusions	94
5.2.1	Design of hybridisation sequences.....	95
5.2.2	Covalent coupling of DNA oligonucleotides to streptavidin	96
5.3	Interaction-dependent reverse transcription PCR.....	97
5.4	Mock selection for SBP peptides.....	101
5.5	Summary.....	104
6	Concluding remarks and future directions	106
6.1	Summary.....	106
6.2	Future directions	107
6.2.1	<i>In vitro</i> selection for Diels-Alderase enzymes using mRNA display	107

6.2.2	Engineering of orthogonal PEX5-peptide interactions	108
6.2.3	Interaction-dependent reverse transcription with mRNA display.....	108
6.2.4	The future of mRNA display selection	108
6.3	Concluding remarks-----	109
7	Appendix	111
7.1	Chapter 3-----	111
7.1.1	DFPase synthetic gene DNA sequence	111
7.1.2	DFPase synthetic gene amino acid sequence.....	111
7.1.3	DFPase 3'-fragment DNA sequence	111
7.1.4	Primer Sequences.....	112
7.2	Chapter 4-----	112
7.2.1	PEX5 445-728 DNA sequence [‡]	112
7.2.2	PEX5 445-728 amino acid sequence	113
7.2.3	Primer Sequences.....	113
7.2.4	PEX5* Round 0 Sequences	114
7.2.5	PEX5* Round 4 Sequences	115
7.3	Chapter 4 - supplementary figures -----	116
7.4	Chapter 5 DNA sequences -----	116
7.4.1	Streptavidin binding peptide (SBP) DNA sequence*	116
7.4.2	SBP amino acid sequence	116
7.4.3	SBP Δ DNA sequence*	116
7.4.4	SBP Δ amino acid sequence.....	116
7.4.5	Chapter 5 Primer Sequences	117
7.5	Chapter 5 - supplementary figures -----	117
	References.....	119

List of Figures

Figure 1.1. Directed evolution of proteins and enzymes in the laboratory.	3
Figure 1.2. Protein fitness landscapes.	4
Figure 1.3. A schematic representation of two types of library creation.....	7
Figure 1.4. A schematic representation of the throughput limits of many commonly used techniques for genetic library interrogation.....	9
Figure 1.5. A schematic overview of the computational enzyme design strategy.	11
Figure 1.6. Comparison of the backbone structures and side chains of RNA and proteins.....	13
Figure 1.7. Selection strategies for the isolation of functional proteins and enzymes by mRNA display.....	21
Figure 1.8. Protein scaffolds used for engineering of novel activities using mRNA- display technology.	23
Figure 3.1. A schematic overview of the experimental procedure for a single round of mRNA display selection.	43
Figure 3.2. Structures of tyrosyl-tRNA (a) and puromycin (b).....	45
Figure 3.3. Step-by-step comparison of strategies typically used to generate 3'- puromycin templates for mRNA display.	47
Figure 3.4. Denaturing PAGE of the synthesised puromycin-oligonucleotide	48
Figure 3.5. Mass spectrometry analysis of puromycin-oligonucleotide synthesised by phosphoramidite chemistry.	49
Figure 3.6. Schematic of puromycin attachment to mRNA by psoralen-mediated photo-crosslinking and subsequent mRNA-protein fusion generation.	50
Figure 3.7. Psoralen-mediated photo-crosslinking between the puromycin-oligonucleotide and a model fragment of the DFPase mRNA.	51
Figure 3.8. Schematic showing the formation of an mRNA-protein fusion on the ribosome.....	53
Figure 3.9. (a) SDS-PAGE autoradiography analysis of <i>in vitro</i> translation reactions shows RNA-DFPase fusion formation is proportional to input template concentration. (b).....	55
Figure 3.10. Optimisation of mono- and divalent salt concentration in the <i>in vitro</i> translation reaction for increased RNA-protein fusion formation.....	56
Figure 3.11. Scheme for the automated solid-phase synthesis of oligo-dT cellulose...	58
Figure 3.12. Comparative binding analysis of a dA ₁₅ oligonucleotide probe to oligo-dT matrices.....	59

Figure 3.13. SDS-PAGE-autoradiography analysis of RNA-DFPase fusion purification using synthesised oligo-dT cellulose.....	60
Figure 3.14. Demonstration of RNA-protein linkage and reverse transcription to generate cDNA-protein fusions.....	61
Figure 3.15. A schematic overview of the experimental procedure for a single round of mRNA display selection for bond-forming enzymes.....	62
Figure 3.16. The Diels-Alder reaction.....	63
Figure 3.17. Synthesis of a dienophile-linked reverse transcription primer.....	64
Figure 3.18. Reaction scheme for the synthesis of a biotinylated diene for selection of Diels-Alderase enzymes using mRNA display.	65
Figure 4.1. Schematic representation of PEX5 mediated import of folded proteins across the peroxisomal membrane.	70
Figure 4.2. X-ray crystal structure of the <i>Homo sapiens</i> PEX5 receptor-PTS1 complex (PDB: 1FCH).	71
Figure 4.3. A schematic overview of the experimental procedure for selection of orthogonal PEX5:PTS1 binding pairs by mRNA display.....	73
Figure 4.4. Optimisation of monovalent salt concentration in the <i>in vitro</i> translation reaction for increased RNA-PEX5 fusion formation.....	74
Figure 4.5. SDS-PAGE-autoradiography analysis of RNA-PEX5 fusion purification using synthesised oligo-dT cellulose.....	75
Figure 4.6. Comparative binding of RNA-PEX5 fusions on immobilised model peptide targets.	76
Figure 4.7. Relationship of mutated residues to the model PTS1 peptide YQSKL.....	77
Figure 4.8. Schematic of the PEX5* library assembly strategy.....	79
Figure 4.9. Assembly of the 900 base-pair linear PEX5* DNA library.....	79
Figure 4.10. SDS-PAGE-autoradiography analysis of the purification of RNA-PEX5* library fusions at round 0.....	81
Figure 4.11. Sequence logos displaying the amino acid distribution amongst clones generated from sequencing data after four rounds of selection.....	82
Figure 4.12. Analysis of the expression and purification of the PEX5.YY.4.3 clone (black arrow) by SDS-PAGE.....	83
Figure 4.13. Binding of the wild-type PEX5 (WT) and the selected mutant PEX5.YY.4.3 to the canonical (YQSKL) and non-canonical (YQSYY) peptides.....	84
Figure 4.14. The amino acid composition of randomised codons during selection against the YQSYY peptide.	86

Figure 5.1. (a) Solution phase identification of ligand-target pairs from libraries of small molecule ligands by interaction-dependent PCR.....	92
Figure 5.2. Streptavidin binding peptide (SBP) sequences.	94
Figure 5.3. Design of SBP mRNA display constructs to incorporate a range of hybridisation lengths.	95
Figure 5.4. (a) Schematic of the strategy used to generate DNA-tagged streptavidin.	97
Figure 5.5. Proof-of-concept interaction dependent reverse transcription.	99
Figure 5.6. Agarose gel analysis of exonuclease I optimisation in the IDRT-PCR with an 8-nucleotide hybridisation region.	100
Figure 5.7. Binding and non-binding SBP mutants can be distinguished by restriction digest.....	102
Figure 5.8. (a) IDRT selection of SBP peptides in a mock library format.....	103

List of Tables

Table 1.1. Overview of methods for the <i>in vitro</i> selection or screening of proteins.	15
Table 4.1. Percentage of each nucleotide at each position of the randomised codons in the nascent PEX5* library prior to selection.	80
Table 4.3. Amino acid distribution for randomised codons during selection against the YQSY Y peptide	85
Table 5.1. Positive and negative control SBP constructs discovered by Wilson <i>et al.</i> using mRNA display	101

List of Abbreviations

Amino acids are abbreviated according to their standard three-letter or single-letter codes. Other abbreviations are as follows:

cDNA = Complementary deoxyribonucleic acid

dH₂O = deionised water

DFP = Diisopropyl fluorophosphate

DFPase = Diisopropyl fluorophosphatase

DNA = Deoxyribonucleic acid

dNTP = Deoxyribonucleotide triphosphate

E.coli = *Escherichia coli*

EDTA = Ethylenediaminetetraacetate

IPTG = Isopropyl β-D-1-thiogalactopyranoside

L.vulgaris = *Loligo vulgaris*

mRNA = Messenger ribonucleic acid

Ni²⁺-NTA = Nickel-nitriloacetic acid

Oligo-dT = oligo-deoxythymidine

PCR = Polymerase chain reaction

RNA = Ribonucleic acid

RNase = Ribonuclease

RT = Reverse transcription/reverse transcriptase

SBP = Streptavidin binding protein

Taq = *Thermus aquaticus* DNA polymerase

TMV = Tobacco mosaic virus

Tris = Tris-(hydroxymethyl)-aminomethane

tRNA = Transfer ribonucleic acid

1 Introduction

1.1 New proteins and enzymes

Naturally evolved proteins composed of the twenty canonical amino acids are responsible for a large number of cellular processes that make life possible. The vast range of cellular functions carried out by proteins includes transport, signalling, molecular recognition, and catalysis. The staggering array of biological functions and their exquisite cellular specificity can be attributed to the intricacy and diversity of protein structure in three-dimensions. This has inspired a large field of research devoted to the creation of novel, bespoke proteins with functions defined not by natural selection but by the needs of contemporary society.

Molecular recognition is a fundamental pillar of biology, without which life as we know it could not exist^{1,2}. An archetypal example of the versatility of proteins for molecular recognition is the production of antibodies by the immune system of higher organisms. It is possible to generate antibodies against almost any target, from large, multi-domain proteins to small molecules. However, traditional antibody technology faces several limitations. They are large (~150 kDa) multimeric proteins containing numerous disulphide bonds and post-translational modifications such as glycosylation, which hinders large-scale production³, additionally many are prone to aggregation⁴. Recently, there has been an interest in the development of antibody mimetics to overcome these limitations. A number of small, thermostable, single-domain proteins have been described that serve as tools for bespoke molecular recognition⁵. Applications of tailored binding proteins are manifold, for example, developments in our understanding of the mechanistic underpinnings of disease has fuelled the need for new biomolecular recognition tools for therapeutics⁶, *in vivo* imaging⁷, and clinical diagnostics⁸. Novel biological targets of interest are constantly being discovered, and the development of the new generation of antibody mimetics depends on efficient selection in the laboratory.

Enzymes are Nature's catalysts, accelerating the chemical reactions necessary to sustain life, from the breakdown of compounds in metabolism to the construction of complex molecules that serve a wide range of functions from structure to defence^{9,10}. Consistent with their crucial biological roles, they are highly efficient catalysts, and often exhibit exquisite chemical selectivity¹¹. Combined with their ability to function in mild aqueous environments, they provide attractive alternatives to traditional organic chemical transformations¹¹.

The advantages of the use of enzymes as industrial catalysts are clear due to their high catalytic activity, stereoselectivity and ability to function in environmentally friendly conditions^{11,12}. Applications of enzymes in an industrial setting include the improved synthesis of drug compounds¹³, and breakdown of biomass for fuels¹⁴ and fine chemicals¹⁵. However, despite their incredible diversity, natural biological systems do not provide enzymes to catalyse every desired chemical reaction. This is perhaps unsurprising given that evolutionary pressures generally limit the catalytic scope of natural enzymes to reactions that give their host organism a survival advantage. A common goal in the engineering of novel enzymes is the expansion on known enzyme classes into novel applications, for example evolution of DNA polymerases that accept modified nucleotides^{16,17}. However, an area of great interest is the generation of enzymes with completely new activities, tailored to any desired application.

1.2 Protein engineering and directed evolution

Iterated mutation and natural selection over many generations during biological evolution has resulted in diverse solutions to the plethora of challenges faced by organisms in the natural world. However, these naturally evolved traits seldom overlap with the characteristics of organisms and biomolecules that are sought by humans. As a result, humans have for centuries used artificial selection to guide evolutionary processes and access more useful phenotypes, beginning with the domestication of crops¹⁸ and animals¹⁹. More recently, advances in molecular biology have facilitated the development of robust and highly effective laboratory-based evolution techniques for optimisation of individual genes and gene products at the molecular level.

Termed directed evolution, the laboratory-based process closely mimics that of biological evolution (Figure 1.1). Typically a diverse library of genes is expressed as a corresponding library of gene products in a manner that maintains an association between the genotype and phenotype. This allows screening or selection for functional variants, which are replicated for characterisation or to serve as templates for subsequent rounds of diversification and screening or selection.

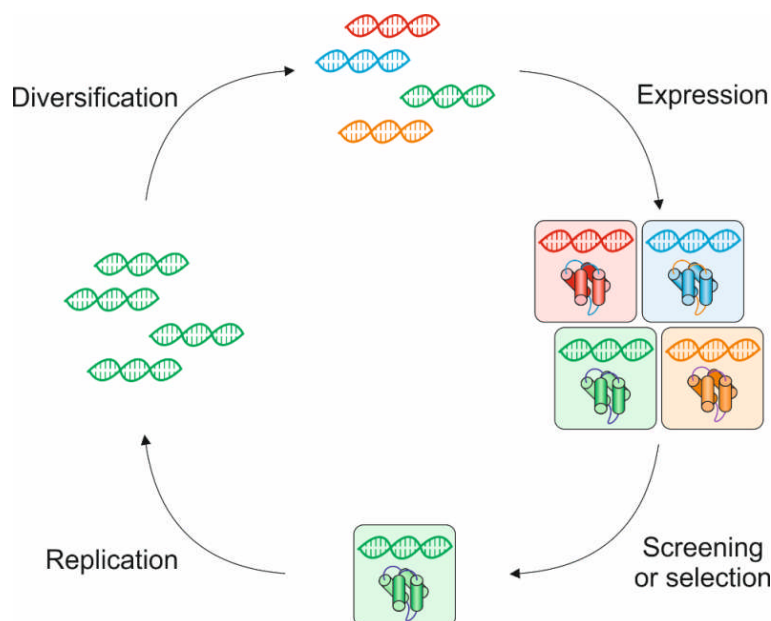


Figure 1.1. Directed evolution of proteins and enzymes in the laboratory. A diverse library of genetic sequences are expressed in a manner that preserves a link between genotype and phenotype (the function of interest). Screening or selection for the desired functionality allows isolation of active proteins or enzymes and recovery of the genetic information either for further rounds of directed evolution or characterisation.

Both natural and laboratory evolution are driven by genetic diversity, however the rate of spontaneous mutation that has fuelled natural evolution over millions of years is not practical over time scales amenable to laboratory evolution. A number of techniques have been developed for the generation of diverse populations of gene variants to accelerate the exploration of sequence space, discussed in detail below.

1.2.1 Genetic Diversity

The vastness of mutational space for a typical protein makes it impossible in practice to cover the entirety of sequence space for all but the smallest peptides. Consider that the number of mathematically possible unique sequences for a protein n amino acids in length is equal to 20^n . For a 100 residue protein this amounts to approximately 10^{130} , which is more than the total number of atoms in the known universe (approximately 10^{80}). Even a library with the mass of the Earth itself, 5.98×10^{27} g, would comprise at most 3.3×10^{47} different sequences, a miniscule fraction of such diversity²⁰. This clearly outlines one of the key barriers to the generation of lab evolved proteins with comparable function and conformation to proteins found in Nature, especially with regard to complex

functions such as catalysis. Due to this inherent caveat, most gene diversification methods perform sparse sampling of sequence space. It should be noted that only an infinitesimal fraction of possible protein sequence space is likely to have been explored by Nature over billions of years of evolution²¹. Furthermore, only those that relevant to survival were retained, with many – even those representing solutions to other interesting problems – discarded over the course of evolution.

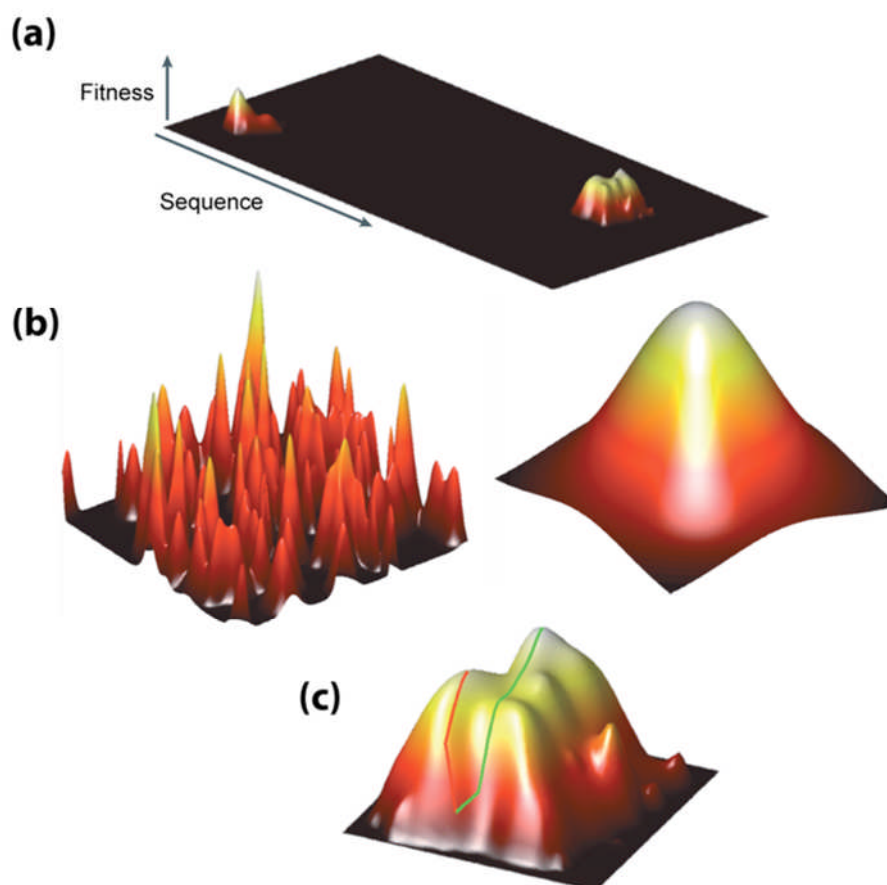


Figure 1.2. Protein fitness landscapes. **(a)** The fitness of a protein to perform a desirable trait is plotted against sequence space to give a landscape of directed evolution which is sampled by library members and surveyed by a screening or selection system. Although not known, it is assumed that most sequences are non-functional (black), and that rare functional sequences are clustered near to one-another. **(b)** Such fitness landscapes can be rugged with multiple local maxima (left), or smooth with a single global maximum. **(c)** Local optima may restrict the path taken during directed evolution (red line), however many alternative routes lead to the optimum fitness (green line). Figure adapted from Romero *et al.*²².

The idea of protein sequence space was first described in 1970, with the description of protein evolution as a walk from one functional protein to the next in the space of all possible protein sequences²³. In this description, each sequence can be ascribed a

fitness value, which in directed evolution is a measure of the ability to perform the experimentally defined function. Protein evolution toward any function can therefore be visualised as a walk along large, multidimensional protein fitness-landscapes (Figure 1.2a). Depending on the activity under selection, these landscapes can be smooth trajectories toward a single optima or more complex, with multiple local optima (Figure 1.2b). Directed evolution strategies look to navigate “uphill” in protein sequence space in a step-by-step manner, however it is possible to reach local optima that do not represent the best possible protein fitness (Figure 1.2c). These three-dimensional visualisations of protein sequence space mapped against fitness do not fully represent the multidimensionality of sequence space, and understanding the relationships between the two remains far from trivial²⁴.

In Nature, random variation at the nucleic acid level has shaped the evolution of organisms over billions of years. However, the low rates of mutation observed in natural systems are not practical for directed evolution in the laboratory. To circumvent this, many strategies have been developed to induce an elevated rate of genetic variation. Traditionally, these approaches for the introduction of random genetic diversity can be divided into recombinatory and non-recombinatory techniques, which can be used alone or in combination.

The first techniques described for the introduction of random genetic variation used chemical or physical agents to damage DNA randomly *in vivo*. These agents include alkylating²⁵ and deaminating²⁶ compounds, base analogues²⁷, and ultraviolet irradiation²⁸. These methods are sufficient to deactivate genes randomly for genome-wide genetic screening, but not commonly used in directed evolution due to biases in mutational spectrum^{25,26} and the difficulty in targeting specific genes of interest. More recently, non-chemical methods of gene randomisation have been developed that act to increase the rate of error incorporation during DNA replication. In *Escherichia coli* (*E.coli*), DNA replication by DNA polymerase III has an inherent mutation rate of 10^{-10} mutations per replicated base²⁹. Mutator strains have been developed with inactivated DNA proofreading and repair enzymes that increase the rate of mutation. The commercially available XL-1 red strain of *E.coli* permits a mutation rate of 10^{-6} per base per generation³⁰. A drawback to these mutator strains is that errors are not exclusively incorporated in the gene of interest and can also result in deleterious mutations in the host genome³⁰.

The lack of control and relatively low rate of mutation offered by *in vivo* techniques have resulted in a preference towards the use of *in vitro* methods of random mutagenesis.

Error-prone PCR (epPCR) was a seminal development in the generation of non-recombinatory genetic diversity for directed evolution (Figure 1.3a). First described by Goeddel and co-workers in 1989, the technique exploits the low fidelity of DNA polymerase enzymes under certain conditions to generate point mutations during PCR amplification of a gene of interest³¹. A combination of increasing magnesium concentration, supplementing with manganese, using skewed dNTP concentrations, and mutagenic dNTP analogues³² can increase the mutation rate to 10^{-3} per replicated base³³. In practice, the accumulation of mutations after each cycle of PCR allows tuning of the average number of mutations per gene by controlling the number of cycles. A general concern with error-prone PCR is that it results in a biased mutational spectrum. Indeed, *Thermus aquaticus* (*Taq*) DNA polymerase generates more transitions (purine/purine exchanges) than transversions (purine/pyrimidine exchanges)³⁴ and more AT to GC than GC to AT exchanges³⁵. This phenomenon can be mitigated by adjusting the concentration and ratio of the different bases³⁶ or using a mixture of *Taq* and a DNA polymerase with a complementary mutational spectrum³⁵. Furthermore, the degeneracy of the genetic code means that access to all twenty amino acids at any one codon would require the mutation of up to all three bases, an extremely unlikely event under epPCR conditions. Indeed, single-base substitutions within a codon results in an average coverage of only five different amino acid substitutions³⁷.

Recombinatory techniques, much like natural evolutionary processes, rely on the re-assortment of genetic variation in order to access beneficial combinations of mutations. There are a number of techniques that mimic these natural processes by homologous recombination of sequences of interest. One of the most widely used methods, DNA shuffling (Figure 1.3b), involves fragmentation of the gene(s) of interest with DNase followed by re-assembly by PCR, without additional primers³⁸. The resulting recombined library contains mutations from different parental DNA sequences. Recently, the decreasing cost of synthetic oligonucleotides has facilitated the increase in popularity of analogous assembly PCR protocols for genetic library generation³⁹⁻⁴¹. These methods all use overlapping oligonucleotides that prime and extend one another, yielding full-length genes containing combinations of mutation-bearing oligonucleotides.

Techniques that rely on recombination are only desirable when there is a level of diversity already present amongst the gene(s) of interest. They are therefore commonly used between rounds of directed evolution to combine mutations from ancestrally distinct clones⁴². Additionally, recombination with the wild-type DNA sequence can eliminate

neutral but non-beneficial mutations that have accumulated over rounds of evolution in a process that is analogous to back-crossing³⁸.

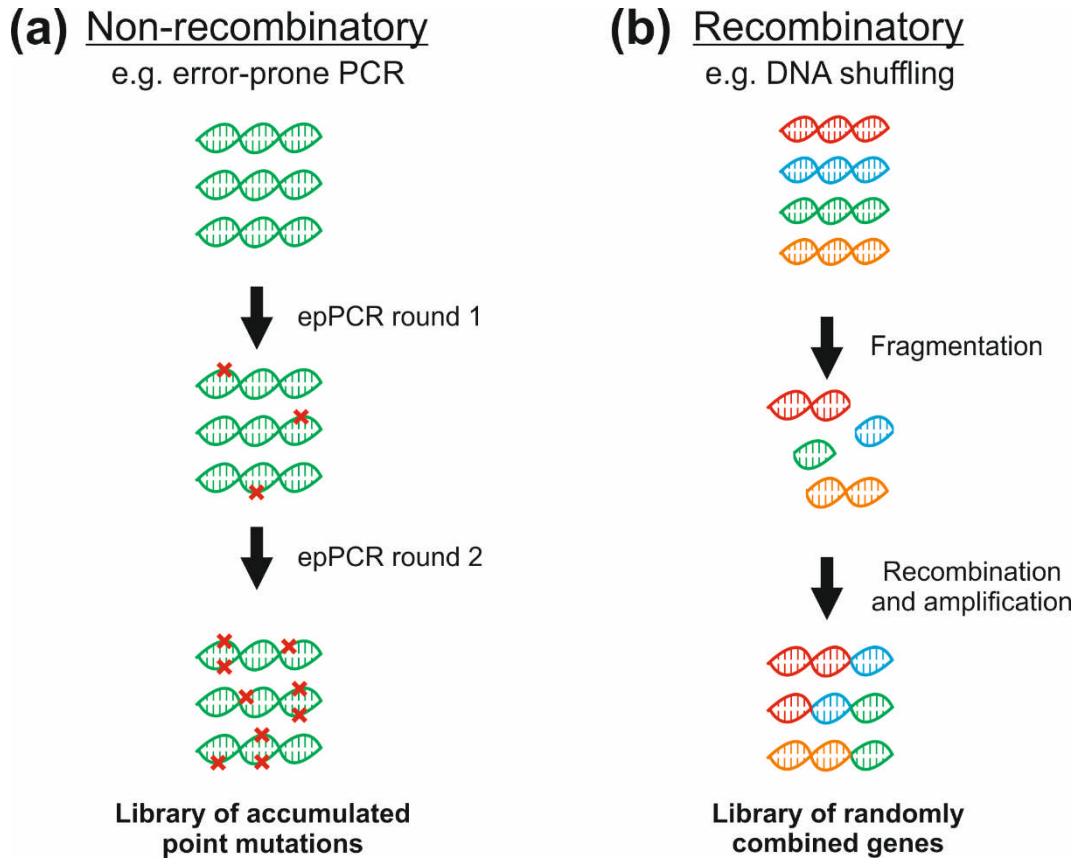


Figure 1.3. A schematic representation of two types of library creation. (a) Non-recombinatory methods (e.g. error-prone PCR). Point mutations are accumulated in the target gene over a number of rounds to generate a library of mutant genes. (b) Recombinatory methods, (e.g. DNA shuffling). A family of homologous sequences is used to create a library of mutant genes by fragmentation and recombination. A combination of both methods is often used to combine different mutations in mutant gene libraries created by error-prone PCR.

The more random the approach taken to mutant library creation, the less prior structural or mechanistic knowledge of the functionality under selection required. Larger libraries allow exploration of greater areas of sequence space and thus give a greater chance of identifying an improved variant. However, as libraries become larger the potentially problematic issue of throughput becomes more prominent. A great number of strategies and techniques have been developed to facilitate more efficient and higher throughput exploration of sequence space, these are discussed in more detail below.

1.2.2 Interrogating genetically diverse libraries

The second critical component of laboratory evolution is the ability to identify and isolate library members with desired properties. All evolutionary systems, whether in Nature or in the laboratory, require a link between genotype (nucleic acid) and phenotype (any functional trait encoded by the genotype). In natural evolution at the organism scale, genotype and phenotype are inherently coupled within each organism. However, during laboratory evolution at the level of individual genes, generating and maintaining this genotype-phenotype association can be far from trivial (this vast field has been reviewed extensively – please see references ⁴³⁻⁴⁶ for more examples).

Approaches to library interrogation in directed evolution of functional biomolecules in the laboratory can generally be divided into two categories: screening and selection (Figure 1.4). Each approach has its own inherent advantages and disadvantages. Screening typically involves assaying variants in the selection library individually, for example by measuring enzymatic activity of individual clones in cell lysates⁴⁷. Screening in this manner generates large amounts of data, including useful information such as the proportion of library members that are functional, and the range of activity between the best and worst library members. However, screening approaches often have the drawback of being relatively low-throughput and, depending on the specific protocol, being both time and labour intensive.

Laboratory selection, by contrast, mimics that of Darwinian evolution - if a new function enables an organism (or biomolecule) to survive, it is selected⁴⁸⁻⁵⁰. Experimentally this results in a binary “yes/no” outcome determined by the selection threshold. Selection in this manner enables much higher-throughput interrogation of sequence space for desired protein function. Indeed, entire libraries can be examined in a single experiment using these methods. However, as a selection experiment only provides this basic binary information, library “hits” often cannot be ranked without further characterisation of selected sequences. This is perhaps especially pertinent to selections using very large libraries, where enriched pools may still have relatively high diversity⁵¹.

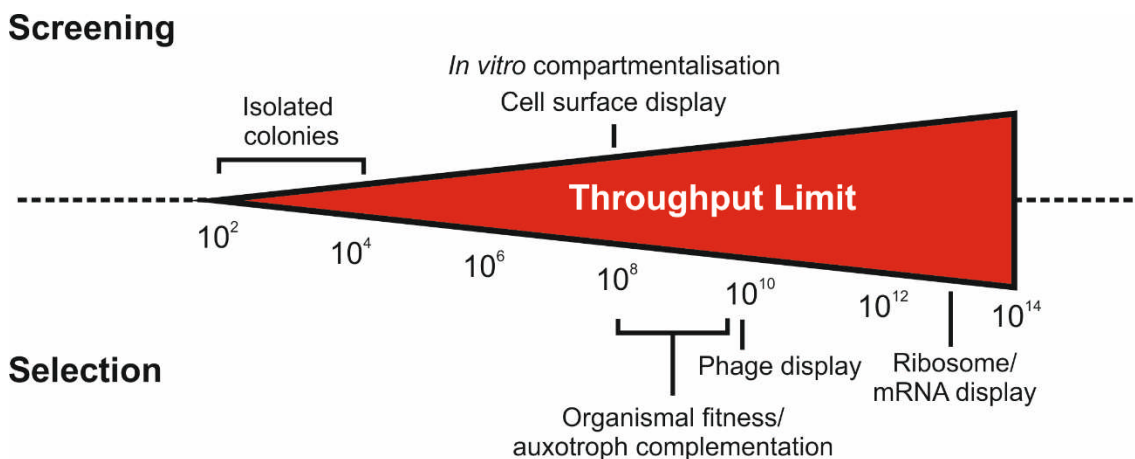


Figure 1.4. A schematic representation of the throughput limits of many commonly used techniques for genetic library interrogation.

Traditional screening and selection methods that utilise whole microorganisms such as *E.coli* or yeast cells to maintain a genotype/phenotype linkage have been, and continue to be, invaluable components of the protein engineer's toolbox. However, maintaining the genotype/phenotype link *in vivo* imposes a limit to the size of the libraries that can be interrogated. This upper-limit is dictated by the efficiency of the transformation of the genetic library into the host microorganism, which even for the most readily transformable organism *E.coli* is approximately 10^{10} sequences⁵². A number of elegant strategies have been developed to circumvent the upper-limit on achievable library size.

The occurrence of functional proteins in random sequence space has been estimated to be about 1 in 1×10^{11} sequences⁵³, far exceeding the number accessible by methods that rely on an *in vivo* translation step. To overcome this potential drawback, a number of high-throughput selection technologies have been developed to isolate novel peptide and protein ligands. However, though hugely successful in the modification of existing enzymes and discovery of novel peptide and protein ligands, the amount of sequence space available for practical exploration using these directed evolution methods has, until recently, precluded the generation of *de novo* enzymes. Complementary strategies, that reduce laboratory selection efforts by screening candidates *in silico*, are becoming powerful tools for engineering novel proteins and enzymes. These rational design approaches are discussed in detail below.

1.2.3 Rational approaches

Over six decades ago, Linus Pauling proposed the conceptual requirements for enzyme design – that enzyme catalysis may be achieved by lowering the activation energy of a reaction via the stabilisation of the transition state⁵⁴. However, the vastness of protein sequence space, in combination with an incomplete understanding of fundamental structure-function relationships in proteins, has hindered the realisation of rational protein design.

Initial attempts to exploit this idea for the generation of artificial biocatalysts utilised catalytic antibody technology^{55,56}. Catalytic antibodies represent some of the earliest ventures into engineered biocatalysis, the first reports of catalysis by antibodies date back more than 25 years^{57,58}, and have been generated to catalyse a wide-range of chemical transformations⁵⁵. Obviating the need to design and build entire proteins from scratch, these techniques utilise transition state analogues of a reaction in combination with clonal selection in the immune system to elicit antibodies with binding pockets tailored for catalytic activity and selectivity. Catalytic antibodies, whilst often exhibiting high selectivity, are capable of only modest rate accelerations in comparison to natural enzymes⁵⁵.

Relying on the same principles of stabilising the transition state as antibody catalysis, computational design has recently blossomed as a powerful approach to generate *de novo* enzymes *in silico*⁵⁹. Enzymes that catalyse a range of chemical reactions including Kemp elimination⁶⁰, retro-aldol⁶¹, and Diels-Alder⁶² reactions have been successfully generated using these methods. The general strategy (Figure 1.5) begins with the *in silico* stabilisation of the known or predicted transition state in three-dimensional atomic models of minimal active sites (theozymes). Then, possible theozyme combinations are matched to known protein scaffolds that could accommodate the proposed active site. These models then undergo further rounds of refinement to better stabilise the transition state, followed by ranking and selection for expression and characterisation for the designed activity⁶³. In total, this process can generate enzymes from a theoretical pool of more than 10^{19} unique active site configurations, significantly more than could be reasonably examined in the laboratory⁶².

A notable example is the computational design of enzymes that catalyse an intermolecular Diels-Alder reaction⁶² – widely used to form carbon-carbon bonds in organic chemistry but not usually found in Nature. The Diels-Alder reaction is a [4+2]

cycloaddition in which a diene and dienophile react in a concerted manner to generate a cyclohexene containing product, with the generation of two new carbon-carbon bonds and up to four new stereocenters⁶⁴.

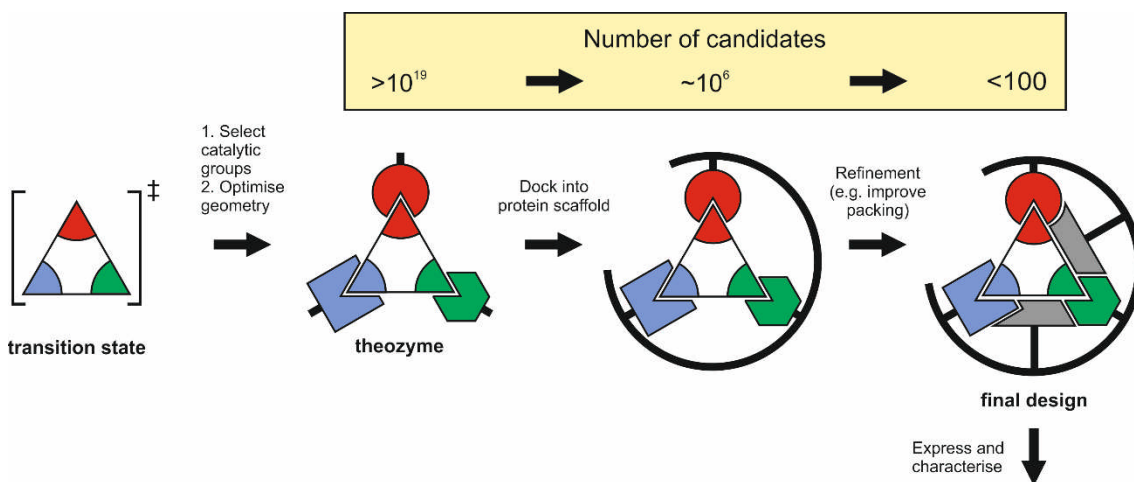


Figure 1.5. A schematic overview of the computational enzyme design strategy. A reaction is chosen for which a transition state is calculated and modelled with potential catalytic groups *in silico* to generate the theozyme. This is then computationally docked into a suitable scaffold from known protein structures in the Protein Data Bank (PDB). Active sites are then refined by computational mutation of surrounding residues and ranked. The most promising enzymes *in silico* are then expressed and characterised experimentally. Using this strategy, more than 10^{19} possible active site configurations can be reduced many orders of magnitude to less than 100 candidates for experimental characterisation⁶³.

A detailed mechanistic knowledge of the Diels-Alder reaction was utilised to stabilise the transition state *in silico*. The computational design process narrowed down the vast number of potential theozyme combinations to just 84 designs that were expressed for experimental validation. Of these, 50 were soluble proteins, and only two were found to have measurable Diels-Alderase activity⁶². The best Diels-Alderase enzyme, DA_20_10, was improved more than 18-fold in a novel manner by utilising the power of game-driven crowdsourcing to modify a loop structure proximal to the active site of the enzyme resulting in the enzyme CE6⁶⁵. The catalytic proficiency ($(k_{\text{cat}}/K_{\text{M}})/k_{\text{uncat}}$) was improved 10^4 -fold compared to the original computational design using directed evolution to yield the most proficient known Diels-Alderase enzyme⁶⁶.

The *de novo* computational design of enzymes is a promising, yet nascent field, and as such, the catalytic efficiencies achieved to date remain orders of magnitude lower than most natural enzymes. However, computationally designed enzymes appear to be

generally highly evolvable. For example, Rothlisberger *et al.* used sixteen rounds of random mutagenesis and shuffling to improve a computationally designed Kemp eliminase more than 2000-fold⁶⁷. In another example, directed evolution of a computationally designed retro-aldolase increased the specific activity by more than 4,400-fold but resulted in complete remodelling of the designed active site⁶⁸. Further evolution using droplet-based microfluidic screening resulted in an enzyme that catalyses a $>10^9$ rate enhancement – which rivals that of natural class I aldolases⁶⁹. These studies demonstrate that there remains a great deal to be learned about optimal rational protein design strategies, and that directed evolution is a powerful tool to increase the typically low activities of designed enzymes to levels comparable with natural enzymes.

1.3 *In vitro* techniques for directed evolution

As discussed previously, in order to perform selections directly to populations of molecules, genotype and phenotype must be linked in some manner. The earliest examples of *in vitro* genetic selection exploited the ability of RNA to encode both the genotype (in this case a nucleic acid sequence that can be copied) and phenotype (a functional trait that varies according to sequence). This was first demonstrated in the 1960s in work by Sol Spiegelman, who applied Darwinian selection to the RNA bacteriophage Q β in a cell-free system⁷⁰. Spiegelman exploited the fact that the Q β viral genome can be copied *in vitro* by the Q β replicase enzyme, allowing hundreds of generations of genome replications to be performed quickly. The inherently high propensity of the Q β replicase to insert mutations generated the required genetic variation at each generation. With replication speed as the phenotype under selection, the Q β genome adapted by deleting regions not required for recognition by the Q β replicase, thus shortening replication time⁷⁰.

These pioneering *in vitro* selection experiments, though elegant, were restricted (i) by the fact that replication speed was the only phenotype amenable to selection, and (ii) by the limited mutation rate set by the Q β replicase enzyme. It took until the 1990s, and the advent of chemical DNA synthesis, for the impact of Spiegelman's research to translate into further phenotypic selections (for example binding affinity) on populations of RNA molecules. Chemical synthesis meant it was, for the first time, possible to generate vast sequence diversity by incorporating entire regions of completely random sequence into DNA strands. Concomitant developments in molecular biology, such as the invention of the polymerase chain reaction (PCR)⁷¹ and the isolation of reverse transcriptase, also

made it possible to replicate almost any nucleic acid sequence *in vitro*. With these advancements came the development of techniques for the use of *in vitro* selection and directed evolution to interrogate sequence space for novel functional RNA sequences⁷²⁻⁷⁴.

The abundance of documented *in vitro* laboratory selections for functional RNA molecules, both for binding affinity⁷⁵⁻⁷⁹ and catalysis⁸⁰⁻⁸⁶, demonstrate that RNA is capable of performing many useful and complex tasks. However, during the course of evolution, almost all known biological systems adopted proteins as the preferred means of performing many essential cellular functions.

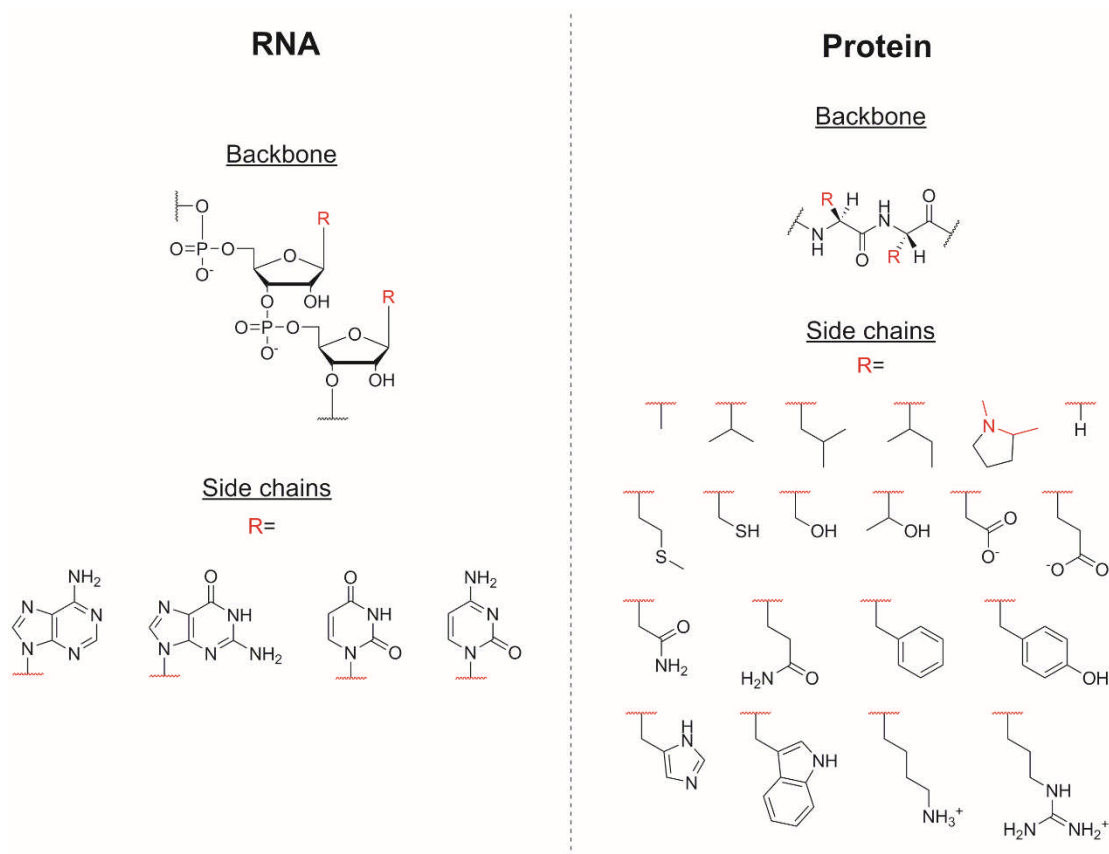


Figure 1.6. Comparison of the backbone structures and side chains of RNA and proteins.

This can be rationalised in a number of ways – from a mechanistic perspective, nucleic acids lack the diverse functional groups of amino acid side chains and the additional chemistries that they confer; the imidazole of histidine, carboxylate of aspartate and glutamate, the primary amine of lysine, and the sulfhydryl of cysteine (Figure 1.6). This comparatively low side-chain diversity, combined with more limited conformational

precision as a result of a highly charged backbone, means RNA lacks characteristics considered important in proteins for complex functions such as catalysis. For these reasons, and because proteins are much more extensively used in diagnostic, therapeutic, and industrial applications, there is great interest in the development of methods for the *in vitro* selection and directed evolution of proteins.

The greatest barrier to the development of effective methods for protein evolution is the recovery of genetic information encoding a protein sequence after translation. Traditional methods typically utilise an *in vivo* step that allows genetic information to be physically segregated into individual organisms, which can be individually isolated and assayed, for example by picking individual colonies on agar plates or sorted using FACS instrumentation. These methods typically allow the sampling of thousands to millions (10^3 - 10^8) of unique variants, which is usually enough to find mutants with improved enzyme activity or altered substrate specificity but not to isolate completely *de novo* activities.

In vitro directed evolution offers a means to engineer proteins by exploring very large libraries containing trillions (10^{12} - 10^{13}) of variants that far exceed that accessible by traditional *in vivo* methods. This was made possible by the development of cell-free translation systems that permit the expression of proteins outside of cells. In lieu of a cellular compartment, several innovative strategies have been developed to maintain the crucial genotype-phenotype link when a library of protein variants is expressed *in vitro*. These strategies generally rely on either direct physical conjugation or artificial compartmentalisation to maintain the genotype-phenotype link, and are summarised in Table 1.1.

All *in vitro* directed evolution methods follow the same general strategy. The initial DNA library is transcribed and translated, either sequentially or in a one-pot reaction. Next, the genotype-phenotype link is established, either via a physical linkage (ribosome display, mRNA display, and DNA display), or compartmentalisation (IVC, IVC-based DNA and microbead display) (Table 1.1). Active variants and can then be isolated by screening or selection. Finally, the genotype of active variants is recovered and either analysed by sequencing or subjected to additional rounds of evolution.


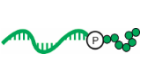


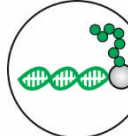
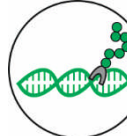
	Physical Linkage			Compartmentalisation		
	Ribosome display	mRNA display	DNA display	<i>In vitro</i> compartmentalisation	IVC & microbead display	IVC & DNA display
Genotype-phenotype linkage						
	Non-covalent mRNA-ribosome complex	Covalent mRNA-protein fusion via puromycin	Covalent or non-covalent complex of DNA-protein	Spatial segregation	Compartmentalised complex of DNA-microbead-protein	Compartmentalised covalent or non-covalent complex of DNA-protein
Reported throughput	$\sim 10^{13}$	$\sim 10^{13}$	$\sim 10^{12}$	$\sim 10^9$ (selection) 10^6 - 10^8 (screening)	$\sim 10^9$ (selection) 10^6 - 10^8 (screening)	10^8 - 10^9
Examples	Proof of concept selection for sialyltransferase ⁸⁷ , β -lactamase ⁸⁸ , and dihydrofolate reductase ⁸⁹ enzymes.	Selection for <i>de novo</i> RNA ligase ⁴²	Proof of concept binding selections ^{90,91}	Selection for DNA modifying enzymes ^{92,93} . Proof of concept screening for β -galactosidase ⁹⁴	Screening for phosphotriesterase ⁹⁵ . Proof of concept selection of biotin ligase ⁹⁶	Selection of heterodimeric antibodies ⁹⁷

Table 1.1. Overview of methods for the *in vitro* selection or screening of proteins.

1.3.1 Benefits of *in vitro* directed evolution

A key step towards the generation of *de novo* enzymes by laboratory evolution has been the development of *in vitro* evolution and selection technologies such as ribosome display, mRNA/cDNA display, and *in vitro* compartmentalisation. These entirely *in vitro* selection techniques are powerful tools in the search for novel functional proteins and peptides from diverse libraries.

The ability to conduct laboratory evolution *in vitro* has a number of benefits; (i) library construction and selection is independent of any potential effects on cell viability, meaning proteins that are toxic to host cells *in vivo* can be evolved; (ii) by eliminating the need for a transformation step, libraries of vastly greater size can be generated and iteratively screened; (iii) selections can be carried out under abiological conditions, allowing tighter control over selection conditions and stringency and permitting the use of non-canonical amino acids, extremes of pH or temperature, and other desired non-physiological conditions, and (iv) enriched library DNA can be directly manipulated after each round of selection. Approaches such as these, which are centred on the application of a selection pressure at the functional level are invaluable in the search for novel

proteins and peptides, especially in the absence of explicit knowledge of structures or mechanisms of activity. In the past decade *in vitro* display technologies have proved indispensable in the discovery of protein ligands^{49,53,98,99} as well as for *in vivo* interaction analysis^{100,101}.

By uncoupling the maximum library size from transformation into a host cell, *in vitro* evolution methods greatly increase the potential number of unique variants that can be accessed in a single experiment compared to *in vivo* approaches. Indeed, the largest reported *in vitro* protein libraries contain 10^{14} DNA sequences¹⁰². By comparison, the highest-throughput technique with a transformation step – phage display – allows the generation of up to 10^{10} unique variants from a single transformation⁵⁰. Phage display libraries of up to 10^{12} variants have been reported by pooling a large number of separate transformations⁵², however this scale-up may not be feasible for many laboratories. Typical library sizes for other *in vivo* selections are significantly lower –between 10^6 and 10^8 variants. *In vitro* approaches allow the generation of libraries with sequence complexity 10,000-fold that of phage display¹⁰³ and *in vitro* compartmentalisation¹⁰⁴, 10^6 -fold that of yeast display or yeast two- and three-hybrid systems^{105,106}, and approximately 10^9 -fold over plate-based screening approaches¹⁰⁷. The ability to survey greater areas of sequence space makes *in vitro* methods particularly well suited for isolating rare functional sequences, for example, those responsible for very high-affinity binding or enzyme catalysis.

A major advantage of *in vitro* selection for *de novo* enzymes is the independence of any prior knowledge of structures and reaction mechanisms – a prerequisite for rational design strategies. Furthermore, in contrast to techniques which contain an obligate *in vivo* step, the complexity of *in vitro* libraries can be extended beyond the 20 naturally encoded amino acids, raising the prospect of developing completely novel protein and enzyme activities beyond the scope of Nature¹⁰⁸.

1.3.2 Ribosome display

Ribosome display was first described in 1973 as a method for the purification of specific mRNA sequences based on immunoprecipitation of their encoding proteins¹⁰⁹, and subsequently developed to select and evolve peptides and proteins *in vitro*^{110,111}. In ribosome display, phenotype-genotype linkage is maintained via a ternary complex of translated protein, its encoding mRNA, and a stalled ribosome (Table 1.1). The non-covalent ternary complex in ribosome display is stabilised by high magnesium concentrations and low temperatures following translation. In order to maintain this

linkage, subsequent selection steps must also be performed at low temperatures and in the presence of high magnesium concentrations.

Ribosome display has primarily been used for the selection of binding peptides and proteins¹¹², although several model selections for enzymatic activity have been reported – with enrichment of 10 to 100-fold per round of selection⁸⁷⁻⁸⁹. The general strategy for enzymatic selection in these examples is via selection for binding to an immobilised substrate or substrate analogue. Whilst these binding strategies can successfully isolate enzymes with known properties (e.g. from metagenomic libraries), they are not well suited to discovering new activity or improving activity.

Direct selection for product formation using ribosome display has been described for the isolation of T4 DNA ligase enzymes¹¹³. Ribosome-displayed enzymes capable of ligating a DNA adaptor to the 3'-end of their encoding mRNA were enriched via PCR amplification using adaptor-specific primers. Using this strategy, active enzymes were enriched 40-fold over inactive sequences. This presents a general approach for the selection of other catalysts using ribosome display via attachment of alternative substrates to the 3'-end of the mRNA.

1.3.3 *In vitro* compartmentalisation

In vitro compartmentalisation (IVC) methods mimic Nature's strategy for linkage of phenotype and genotype by enclosing proteins and their encoding nucleic acids within water-in-oil droplet compartments¹⁰⁴ (Table 1.1). Unlike display technologies, which are inherently reliant on selection by immobilisation, IVC methods allow both selection and screening of protein and enzyme libraries for desired functionality. Several examples of model enzyme selections have been described using IVC⁹²⁻⁹⁴, as well as the directed evolution of an existing enzyme for increased activity⁹⁵.

Compartmentalisation of individual library members is achieved by dilution of an aqueous solution of DNA and a coupled transcription/translation system into a mixture of mineral oil and surfactants¹¹⁴, so that the average droplet contains no more than a single gene. The ability to generate microdroplets with volumes as low as 10 femtolitres means that single DNA molecules can be present at low-nanomolar concentrations inside the compartments. This allows efficient *in vitro* transcription and translation inside the droplet. IVC-based selections are primarily limited to nucleic acid modifying enzymes, where the DNA acts as both the genotype and the substrate. In one of the first demonstrations of enzyme evolution using IVC, the activity of M.HaeIII methyltransferase was improved for a non-native substrate⁹².

IVC approaches also allow screening of protein and enzyme libraries, increasing the diversity of selection strategies beyond nucleic acid modifying enzymes, but also reducing the library throughput compared to selection strategies. Screening approaches generally rely on fluorescence activated cell sorting (FACS) or microfluidic-based droplet sorting to separate active and inactive enzymes by detection of fluorescent product formation.

Microfluidic approaches are a promising route for IVC-based protein engineering for a number of reasons. Due to the modular and customisable nature of individual components, screening strategies can be readily tailored to the activity of choice. This includes the ability to fuse droplets together to deliver new reagents, and perform PCR on droplets without breaking of the emulsion. Furthermore, the ability to generate highly monodisperse droplets enables on-chip kinetic analysis of generated enzymes¹¹⁵. However, unlike commercially available FACS instruments, the assembly of microfluidics devices requires significant expertise.

1.3.4 DNA display

DNA display strategies establish a physical linkage between DNA and the encoded protein (Table 1.1). Several DNA display methods have been developed, including plasmid display^{116,117}, CIS display^{90,91}, SNAP display¹¹⁸, and a range of bead-display approaches^{95,97,119}. The phenotype-genotype linkage can be direct, in the case of plasmid display and CIS display approaches that rely on fusion with DNA binding proteins such as the *lac* repressor or RepA^{90,91,116}. These proteins bind to specific DNA sequences within the encoding gene, generating either a covalent⁹¹ or non-covalent^{90,116} complex. Strategies have also been developed that generate a link via fusion with proteins that bind small molecules attached to the encoding gene. For example, STABLE (streptavidin-biotin linkage in emulsions) utilises biotinylated DNA and streptavidin fusion proteins¹²⁰. The attachment of small molecule suicide inhibitors has also been demonstrated to generate a covalent link between DNA and encoded protein (for example, SNAP display^{118,121} and covalent DNA display^{122,123}). However, a caveat to many of these strategies is that they rely on the encapsulation of genes inside droplets, as specific phenotype-genotype linkage cannot be achieved in bulk solvent.

Alternatively, the link can be indirect, for example via the capture of both DNA and translated protein onto the same microbead inside a droplet^{95,119}. This strategy has been used to improve the catalytic efficiency of a phosphotriesterase enzyme over 60-fold via FACS based screening⁹⁵. Bead display techniques can provide an advantage over direct

DNA display methods as the number of proteins under selection can be increased and the level of display on the bead, and thus selection stringency, can be controlled.

The bead display strategy has also been demonstrated in a selection format, thus having the potential to interrogate larger libraries⁹⁶. In this example, an active biotin ligase enzyme was enriched from a background of inactive genes. Following product formation and immobilisation, the genotype was recovered by incubation of the beads with product-specific antibodies conjugated to a gene-specific PCR primer. Re-emulsification within microdroplets, followed by droplet PCR, resulted in 20-fold enrichment of biotin ligase genes.

1.3.5 mRNA display

mRNA display (Table 1.1), originally developed and reported independently by two groups in 1997, relies on covalent attachment of proteins to their encoding mRNA via the aminonucleoside antibiotic puromycin^{124,125}. This covalent linkage of phenotype to genotype is achieved by *in vitro* translation of 3'-puromycin mRNA templates. During translation, puromycin – an aminoacyl-tRNA mimic – enters the ribosome, which catalyses its covalent attachment to the nascent peptide chain. Since its inception, optimisation of protocols has resulted in the ability to perform selection experiments on libraries that contain up to 10^{13} molecules^{98,102,126}. mRNA display methods have been applied to evolutionary protein engineering for the generation of novel ligands and enzymes^{42,51}, as well as to the dissection of complex biological processes including protein-protein, DNA-protein, RNA-protein and drug-protein interactions^{100,127,128}.

Advances in mRNA display methodology led to the first example of the laboratory selection and evolution of a *de novo* enzyme activity from a randomised library based on a non-catalytic scaffold⁴². Active RNA ligase enzymes were isolated from a library of 4×10^{12} unique proteins based on the zinc-finger domain of the human retinoid-X-receptor with two randomised loops⁵¹. By attaching a substrate molecule to the mRNA-protein fusion complex by reverse transcription, Seelig *et al.* were then able to select for RNA ligase activity by the addition of the second substrate molecule carrying a selectable anchor group. cDNA encoding active protein was subsequently isolated by immobilisation on a solid support due to bond formation between substrates.

In vitro selections for *de novo* protein function come with their own intrinsic limitations. The sequence space available for exploration is vast, even in short peptides, which makes finding complex functionality a daunting technical challenge. However, the techniques facilitate selection from libraries as large as 10^{13} , increasing the likelihood of

discovering rare functional proteins in the starting library. Enzymatic selection using display strategies often requires covalent attachment of the substrate to the displayed protein. This means selection is performed with a high local substrate concentration, effectively removing selection pressure for high substrate affinity. In addition, selection pressure is directed towards the formation of a single product molecule, potentially precluding the generation of multiple turnover enzymes. Though it should be noted that the RNA ligase enzyme generated by Seelig *et al.* did exhibit multiple turnover⁴². In future, combinatorial methods may lead to a solution, for example, *in vitro* display techniques may be used initially to isolate an enzyme with low level novel catalytic activity from a large library of sequences ($>10^{13}$). This enzyme could then be used as a starting point for further directed evolution using complementary techniques (e.g. IVC) with a shift in selection pressure for multiple turnover and increased substrate affinity.

Typically, mRNA display has been used for the selection of binding peptides and proteins using affinity panning strategies (Figure 1.7a). The utility of mRNA display was recently expanded for selection of bond formation reactions^{42,129} (Figure 1.7b), where $A + B \rightarrow A-B$. Substrate A is attached to the cDNA-displayed protein complex via reverse transcription with a substrate-containing oligonucleotide primer. Then, substrate B carrying a selectable anchor group is added, allowing for the immobilisation of sequences that encode active enzymes. This general selection scheme can be readily modified for the isolation of *de novo* enzymes that catalyse bond breaking reactions as well as other transformation reactions. For bond breaking reactions (Figure 1.7c), the substrate to be cleaved (A-B) and an immobilised anchor group is incorporated by reverse transcription. Thus, sequences encoding active enzymes cleave off their anchor group and remain in solution whereas inactive sequences are removed by immobilisation. In the case of enzymes catalysing other chemical transformations (Figure 1.7d), the substrate is again incorporated via the reverse transcription primer, and functional sequences can be isolated by an agent (e.g. an antibody) that specifically binds to the product.

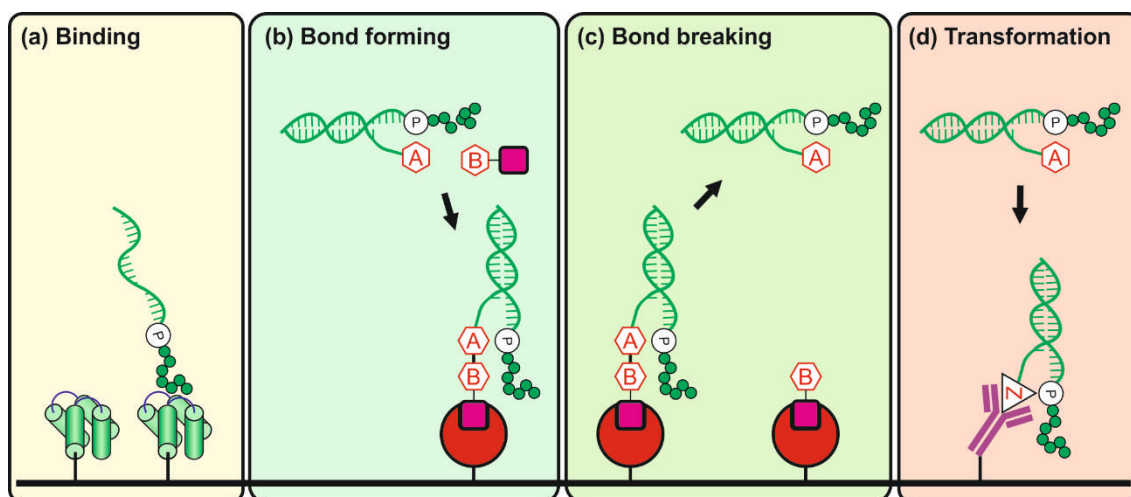


Figure 1.7. Selection strategies for the isolation of functional proteins and enzymes by mRNA display. **(a)** Selection for binding by affinity panning against an immobilised target. **(b)** Bond formation reactions. Substrate [A] is attached to the cDNA-displayed protein complex via the reverse transcription primer. Substrate [B] carries a selectable anchor group (magenta), allowing for the immobilisation of cDNAs that encode active enzymes. **(c)** Bond breaking reactions. The cDNA is modified with the anchor group via the substrate to be cleaved [A–B]. Sequences encoding active enzymes cleave off their anchor group and remain in solution whereas inactive sequences are removed by immobilisation. **(d)** Transformation reactions. An immobilised agent (e.g. antibody) that specifically binds to the product (Z) is used to isolate genes encoding active enzymes.

1.3.6 mRNA display scaffolds and library design

Key considerations before undertaking any display experiment are the nature of both the protein scaffold and the mutant library for use in selection. Various protein scaffolds have been explored for use in mRNA display, ranging from exploration of completely random sequence, to libraries based on well characterised protein folds.

The interrogation of completely random sequence libraries, though a formidable task for all but the shortest peptides, has provided some success for the selection of binders^{53,130}. Completely random sequences allow the potential exploration of all sequence space, unconstrained by the conformation of a scaffold or starting protein. However, the number of selection cycles in a typical *in vitro* directed evolution experiment pales in comparison to the evolutionary course of natural proteins. Notable successes include mRNA display selection of novel ATP ($K_d = 100 \text{ nM}$)⁵³ and streptavidin ($K_d = 5 \text{ nM}$)¹³⁰ binding peptides from libraries of more than 6×10^{12} linear peptides possessing at least 80 contiguous random amino acids. From the results of these selections, the authors estimated functional proteins to be present in random sequence at a rate of roughly 1 in 10^{11} , emphasising the need for large library sizes when performing such experiments.

Although linear peptides with strong binding characteristics can be discovered from large libraries, it is known that the affinity of linear peptides can often be improved by constraining them, for example, in loops within existing structural scaffolds¹³¹. These scaffolds, often based on natural binding proteins, contain stable structural elements such as α -helices or β -sheets linked by randomised loop regions that facilitate binding. Suitable protein scaffolds for display typically possess a number of key characteristics. The first is that the protein must be monomeric, otherwise multivalency and hetero-oligomerisation could interfere in the selection step. Second, libraries are often constructed using highly stable proteins, often derived from thermophilic organisms. This helps to abrogate the destabilising effect on the library of inserting multiple randomised residues and/or loop insertions.

The human tenth fibronectin type III (¹⁰FnIII) domain has been developed and implemented as a scaffold for mRNA display selection of binding proteins (Figure 1.8a). ¹⁰FnIII, a 94 residue domain of the multifunctional, extracellular matrix glycoprotein fibronectin, forms a structure with striking similarities to antibody complementarity determining regions. This feature of the ¹⁰FnIII domain has been successfully exploited in the design and generation of several antibody-mimetic libraries that bind a range of specific ligands with high affinity^{132,133}. Libraries for mRNA display experiments were constructed by randomising 21 residues in three loops. In one example, subsequent selection against tumour necrosis factor alpha (TNF- α) yielded a range of mutants that bound the target with 1-24 nM affinity, which was further improved to 20 pM by affinity maturation¹³².

Another such library, in which two loops of the DNA binding domain of the human retinoid-X-receptor (RXR α) were randomised, was initially used for the selection and evolution of novel ATP binding proteins by mRNA display (Figure 1.8b)⁵¹. Notably the same library was subsequently used to select for *de novo* enzymes with RNA ligase activity⁴². Further rounds of evolution at elevated temperature resulted in a thermostable enzyme with a T_m of 72 °C¹³⁴. Subsequent structural characterisation revealed no discernible secondary structure and high conformational flexibility¹³⁵, features not usually associated with thermostability. This remains the only reported instance of a *de novo* enzyme selected in the laboratory, and although the ability of this scaffold to harbour other distinct enzymatic activities remains unexplored, the zinc finger scaffold is an intriguing candidate.

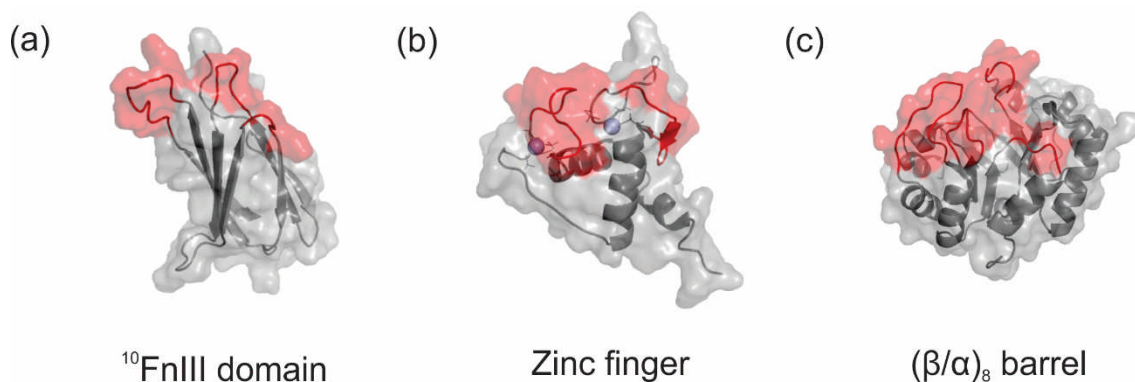


Figure 1.8. Protein scaffolds used for engineering of novel activities using mRNA-display technology. **(a)** Human ¹⁰FnIII domain (PDB ID: 1TTG) and **(b)** the zinc finger domain from human retinoid-x-receptor (PDB ID: 1RXR) have been used to engineer novel binding proteins and enzymes by mRNA-display. Potential mRNA display scaffolds based on **(c)** TIM barrel domains (PDB ID: 1IGS). Regions previously randomised in protein engineering experiments are highlighted in red. Figure generated in PyMol¹³⁶.

An attractive scaffold for the *in vitro* selection of *de novo* enzymes by mRNA display is the TIM (β/α)₈ barrel fold (Figure 1.8c). Characteristically composed of an interior barrel of eight β-sheets surrounded by eight α-helices, the TIM barrel is among the most common of protein folds. It is found in a variety of enzymes with diverse activities across five of the six enzyme classes, including some that are diffusion rate limited¹³⁷⁻¹³⁹. TIM barrel enzymes typically possess distinct structural and catalytic features, with the core of the barrel providing a structural foundation and the eight loops connecting the α-helices and β-sheets on one side of the barrel responsible for substrate binding and catalysis. This is a desirable characteristic from a protein engineering perspective, as modification of functional regions should be less likely to have a profound effect on structural stability¹⁴⁰. The fact that Nature has adopted the TIM barrel fold for many unrelated enzyme activities over the course of evolution is testament to its versatility as an enzyme scaffold. The first high diversity library to be described with a TIM barrel scaffold was based on the indole-3-glycerol phosphate synthase (IGPS) from the hyperthermophilic archaeon *Sulfolobus solfataricus*¹⁰². In this example, 49 residues were inserted in place of eight loops that connect β-sheets and α-helices on one face of the barrel¹⁰². However, despite construction and evaluation of this mRNA display library over 15 years ago – it has not yielded any examples of functional proteins or enzymes.

More recently, a more complex library assembly strategy has been described based on glycerophosphodiester phosphodiesterase (GDPD) from the hyperthermophile *T. maritima*¹⁴¹. The authors randomised 32 residues in seven of the eight loops, followed by mRNA display selection for folded library members using selection based on protease

susceptibility. There are, as yet, no published examples of successful selection of functional proteins from this library. However, the enrichment of folded library members during the library assembly process may allow greater exploration of functional protein sequence space. The versatility, stability, and modularity of the TIM barrel fold make it an attractive candidate scaffold for selection of *de novo* enzyme activities using *in vitro* techniques such as mRNA display. Nonetheless, it remains unknown whether existing enzyme structures are optimised for harbouring activities not typically found in Nature.

1.4 Aims and objectives

The primary aim of this work was to develop protocols for the generation and purification of RNA-protein fusions, and the subsequent selection of functional proteins and enzymes from large high-diversity libraries using mRNA display. Chapter 3 describes the establishment of mRNA display techniques, including the generation and purification of RNA-protein fusions, as well as the development of tools for the *in vitro* selection of novel carbon-carbon bond forming enzymes using mRNA display. Chapter 4 describes the investigation of a key peptide-receptor interaction in the import of folded proteins into peroxisomes. A high-diversity library based on the soluble peroxisomal import receptor (PEX5) was constructed and mRNA display selection was used to interrogate this library for orthogonal peptide-receptor interactions. Chapter 5 describes the development of solution-phase selection techniques using RNA-protein fusions and DNA-tagged targets via interaction-dependent reverse transcription (IDRT). This includes optimisation of amplification of nucleic acid encoding the streptavidin binding peptide (SBP) specifically in the presence of DNA-tagged streptavidin, and proof-of-concept selection from a large background of non-functional input template. The work discussed in this thesis can be divided into several individual objectives:

Aim 1: Establish the mRNA display technology

- Design and synthesise a puromycin containing oligonucleotide for attachment to the 3'- end of template RNA.
- Demonstrate covalent attachment of the puromycin oligonucleotide to the 3'- ends of RNA transcripts via psoralen mediated photocrosslinking.
- Translate of RNA-puromycin conjugates *in vitro* to generate mRNA-protein fusions.
- Optimise *in vitro* translation for efficient RNA-protein fusion formation.

- Synthesise oligo-dT cellulose for the purification of RNA-protein fusions from crude translation lysate.
- Demonstrate recovery of genotype from RNA-protein fusions by reverse transcription.
- Design and synthesise substrate analogues for *in vitro* selection of novel carbon-carbon bond forming enzymes using mRNA display.

Aim 2: Select novel peptide-receptor binding interactions using mRNA display

- Design and assemble a high-diversity library based on an existing receptor protein.
- Use mRNA display methods to interrogate the library for interactions with non-canonical peptide ligands.
- Express, purify, and characterise enriched sequences.

Aim 3: Develop novel solution-phase selection strategies using mRNA display

- Design and synthesise model selection constructs based on a known protein-protein interaction.
- Demonstrate the concept of interaction dependent reverse transcription RNA-protein fusions.
- Optimise the reaction conditions for improved signal:noise ratio.
- Perform model solution-phase selections and estimate the round-by-round enrichment using this technique.

2 Materials and Methods

2.1 Materials

2.1.1 Bacterial strains and plasmids

E. coli XL-10 Gold ultracompetent cells (Tet^rΔ(*mcrA*)183 Δ(*mcrCB-hsdSMR-mrr*)173 *endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Hte* [F' *proAB lac⁺ZΔM15 Tn10* (Tet^r) *Amy Cam^r*]) used for cloning, and *E. coli* strain BL21 Gold (DE3) (B F⁻ *ompT hsdS*(r_B⁻ m_B⁻) *dcm⁺ Tet^r gal λ*(DE3) *endA Hte*) used for high level protein expression were purchased from Stratagene (Cheshire, UK). Plasmids were kindly supplied by Professor A. Berry (University of Leeds, UK).

2.1.2 Chemicals and reagents

Analytical grade chemicals were used throughout. Isopropanol, phenol, chloroform, ethanol, 1-butanol, sodium acetate, and sodium chloride were supplied by Fisher Scientific Limited (Loughborough, UK). DNA size markers were from Promega (Southampton, UK). *Taq* DNA polymerase, Vent polymerase, T4 DNA ligase and the restriction enzymes were obtained from New England BioLabs (Ipswich, MA, USA). All other reagents used were molecular biology grade. Desalted oligonucleotides were synthesised by Integrated DNA technologies, and unless stated used without further purification. All other solvents and reagents were of analytical grade and used as supplied.

2.2 General Methods

2.2.1 General experimental

All non-aqueous reactions were carried out under an atmosphere of nitrogen. Water-sensitive reactions were performed in oven-dried glassware cooled under nitrogen before use. Solvents were removed under reduced pressure using either a Büchi rotary evaporator and a Vacuubrand PC2001 Vario diaphragm pump, or a Genevac HT-4 evaporation system.

2.2.2 Protein and Nucleic acid sequence alignment

Protein sequences were obtained using the online tool UniProtKB¹⁴². Multiple sequence alignment was achieved using the European Bioinformatics Institute's online tool

ClustalW Version 2.0¹⁴³. Sequences were aligned and visualised using Bioedit 7.1.3 software¹⁴⁴.

2.2.3 pH measurements

pH measurements were taken using either Fisherbrand® pH paper (Leicestershire, UK) or using a Jenway 3020 pH meter calibrated according to the manufacturer's instructions.

2.2.4 Culture growth

Unless stated otherwise, all media was supplemented with appropriate antibiotic(s) to a final concentration of 50 µg mL⁻¹. Bacterial cultures were grown in 2TY medium. 1 L of medium contained 16 g tryptone, 10 g yeast extract, and 5 g NaCl, and was made up to 1 L with distilled water before autoclaving at 121 °C for 20 min. Solid phase culture medium was made using 2TY medium with the addition of 1.5% (w/v) agar. Aseptic techniques were employed throughout. Media and heat resistant materials were sterilised using an autoclave. Heat labile materials were filtered through 0.22 µm filters (Sartorius AG, Goettingen, Germany). Glycerol stocks were created by adding 0.5 mL bacterial culture to a sterilised 1.8 mL CryoTube™ containing 0.5 mL glycerol. These were mixed well and stored at -20 °C.

2.2.5 Transformation

Plasmids were transformed into chemically competent *E.coli* cells. For each transformation 50 µL competent cells were thawed on ice in a 14 mL round bottom Falcon tube, and 1 µL of pKKDFP plasmid DNA was pipetted into each aliquot of cells. The transformation reactions were then mixed by swirling gently, and incubated on ice for 30 min. After this incubation, the cells were heat pulsed for 45 s in a 42 °C water bath, and then put on ice for a further 2 min. 0.5 mL of 2TY media (preheated to 42 °C) was added and each transformation reaction was incubated at 37 °C for 1 h with shaking at 200 r.p.m. Cells were subsequently spread onto agar plates containing 50 µg.mL⁻¹ ampicillin using sterile technique, and incubated overnight at 37 °C. Colonies were picked and sub-cultured overnight in 5 mL 2TY media followed by plasmid purification. Insert sequences were verified by DNA sequencing (Beckman Coulter Genomics, Takeley, UK).

2.2.6 Flash chromatography

Flash column chromatography was carried out using silica (35-70 µm particles) according to the method of Still, Kahn and Mitra¹⁴⁵. Thin layer chromatography was

carried out on commercially available pre-coated aluminium plates (Merck silica 2 8 8 0 Kieselgel 60F254).

2.2.7 Mass spectrometry

Mass spectrometry of small molecules was routinely performed on a Bruker HCT ultra LC-MS spectrometer using electrospray (+) and (-) ionization. Proteins and oligonucleotide samples were desalted into 50 mM ammonium acetate pH 7.4 prior to analysis by nano-ESI-MS on a quadrupole-ion mobility spectrometry-orthogonal TOF MS (Synapt, HDMS Waters UK Ltd.) operating in positive TOF 'V' mode. Samples were analysed in acetonitrile : 1 % aqueous formic acid (50:50; v/v). Data was processed using the MassLynx v4.1 software suite. Analysis of protein and oligonucleotide samples by mass spectrometry was carried out by Dr James Ault (Astbury Centre, University of Leeds, UK).

2.2.8 NMR

¹H- NMR spectra were recorded on a Bruker Avance 500 or DRX500 (500 MHz) spectrometer using an internal deuterium lock. ¹³C-NMR spectra were recorded on a Bruker Avance DPX-300 (300 MHz) spectrometer also using an internal deuterium lock. ¹³C-NMR spectra were recorded with composite pulse decoupling using the waltz 16 pulse sequence. DEPT, COSY, HMQC and HMBC pulse sequences were routinely used to aid the assignment of spectra. Chemical shifts are quoted in parts per million downfield of tetramethylsilane, and coupling constants (J) are given in Hz. NMR spectra were recorded at 300 K unless otherwise stated.

2.3 Nucleic acid methods

2.3.1 Phenol:chloroform extraction

Half the volume of phenol (pH 8.0) was added to the crude nucleic acid preparation, followed by vigorous mixing. The mixture was then centrifuged at 2000 × g for 30 s to separate the two phases. The aqueous phase was transferred into a clean Eppendorf tube. The phenol phase was then back-extracted by adding half the volume of water and repeating the extraction step. The two aqueous phases were combined and extracted three times with an equal volume of chloroform by mixing and centrifugation at 2000 × g for 30 s to separate the two phases. The aqueous phase was then transferred to a clean Eppendorf and concentrated by the addition of four times the volume of 1-butanol followed by mixing and centrifugation at 2000 × g for 30 s.

2.3.2 Ethanol precipitation

Nucleic acid was precipitated by the addition of one-tenth of the total volume of potassium acetate (2.5 M, pH 5.5) followed by mixing, and the addition of 2.5-fold the volume of cold ethanol. The solution was mixed and stored at -80 °C for 1 h. The sample was then centrifuged for 10 min at 13,000 × g and the supernatant was carefully removed. The nucleic acid pellet was then washed with 100 µL of 70% (v/v) ethanol and air dried. The pellet was then dissolved in an appropriate volume of nuclease free water.

2.3.3 Nucleic acid quantification

Nucleic acids were quantified by measuring UV absorbance at 260 nm using a Nanodrop 2000 spectrophotometer. For dsDNA, and ssRNA extinction coefficients of 50 ng.µL⁻¹ and 40 ng.µL⁻¹ were used, respectively. For oligonucleotide quantification, sequence-specific conversion factors were used as calculated via the Nanodrop software package.

2.3.4 Ligation-independent cloning

Ligation-independent cloning was carried out using the FastCloning method¹⁴⁶. Primers were designed to amplify the target insert (gene) and vector creating complimentary 5' and 3' ends of DNA fragments. The identity and purity of the PCR products was analysed by agarose gel electrophoresis. Vector and insert reactions were then mixed in a 1:1 (v/v) ratio. The reactions were digested with *DpnI* for 1 hr and subsequently transformed into XL10 Gold® Ultracompetent cells.

2.3.5 Plasmid DNA purification

Plasmid DNA was purified from 5 mL *E.coli* cultures using the Wizard™ Miniprep DNA purification system (Promega, Southampton, UK) according to the supplied protocol. Plasmid DNA concentration was determined using a Nanodrop 2000 spectrophotometer as described in 2.3.3.

2.3.6 Native gel electrophoresis

Double stranded DNA was analysed using native gel electrophoresis (either agarose gel or native PAGE) carried out according to the protocols described by Sambrook *et al.*¹⁴⁷. Agarose gels contained between 0.7 – 2.5 % (w/v) agarose dissolved in 1 × TAE buffer (40 mM Tris base, 20 mM acetic acid, 10 mM EDTA). A 1 kb ladder was used for determining the size of DNA fragments > 2 kb, a 100 bp ladder was used for determining the size of DNA fragments < 2 kb, DNA ladders were purchased from Promega (Southampton, UK). Extraction of nucleic acid from agarose gels was carried out using

a QIAquick® Gel Extraction Kit according to the manufacturer's protocols. Extraction of nucleic acid from native PAGE gels was performed using the crush-and-soak method as previously described¹⁴⁸.

2.3.7 Denaturing PAGE

RNA and ssDNA oligonucleotides were analysed using TBE-urea page electrophoresis, using the UreaGel 29:1 Denaturing Gel System (National Diagnostics). During casting 10 µL of TEMED was added for every 10 mL of gel casting solution, and swirled gently to mix. Gel polymerisation was initiated by the addition of 10 µL of freshly prepared 10% ammonium persulfate for every 10 mL of gel casting solution, followed by gentle mixing. Upon casting, gels were allowed to polymerise for one to two hours. Gels were pre-run for 15-30 minutes before loading. Samples were prepared in 1× denaturing PAGE loading dye (47.5% formamide, 0.01% SDS, 0.01% bromophenol blue, 0.005% Xylene Cyanol, 0.5mM EDTA) and heated to 90 °C for 5 min . Gels were run at a constant voltage of 280 V, in 0.5× TBE buffer (40 mM Tris-HCl, pH 8.3, 45 mM boric acid, 1 mM EDTA). Following electrophoresis plates were allowed to cool for 10-15 minutes before separation. Gels were stained for 20 min in a 10 % (w/v) solution of ethidium bromide, rinsed in dH₂O to remove residual stain, and imaged using a

2.3.8 Synthetic Gene Design

The synthetic DFPase gene was designed using the amino acid sequence obtained from the Uniprot database (accession number: Q7SIG4). A C-terminal hexahistidine tag was incorporated to facilitate purification by Ni²⁺-NTA chromatography. To allow cloning into the expression vector pKK223-3, EcoRI and PstI restriction sites were included at the 5'- and 3'-ends of the synthetic DFPase gene, respectively. The gene was codon optimised for dual expression in *E.coli* and rabbit reticulocyte lysate and the sequence was screened to ensure the absence of the following common restriction enzyme sites; BamHI (GGATCC), HindIII (AAGCTT), DpnI (GATC), Sall (GTCGAC), SmaI (CCCGGG), XmaI (CCCGGG), EcoRI (GAATTC), and PstI (CTGCAG) by Genscript (USA). The synthetic gene was cloned into a pUC57 vector by Genscript to yield the vector pUC57-DFP. Lyophilised pUC57-DFP was resuspended in nuclease free water prior to use as per the manufacturer's guidelines.

2.3.9 PEX5 receptor library construction

The high diversity library based on the PEX5 gene from *A. thaliana* were generated by the three-step process of PCR amplification with mutagenic primers, followed by seamless restriction digest and ligation based on a previously described method¹⁰². The

PEX5 gene was divided into 6 fragments (A-F) and twelve residues were randomised using primers containing randomised NNS codons and 5'- or 3'- terminal recognition sites for the type-IIIS restriction enzyme BsaI. Following PCR amplification from the pET28-AtPEX5 Δ 444 template, DNA was phenol-chloroform extracted (section 2.3.1) and ethanol precipitated (section 2.3.2). Precipitated fragment DNA was digested with BsaI to generate sticky ends, and purified using 2% agarose gel. Purified digested fragments were ligated in a step-wise manner using a 1:1 molar ratio of DNA with T4 DNA ligase at 16 °C overnight. Ligation products were purified by 2% agarose gel and PCR amplified with the corresponding external primers to generate approximately 10 copies of each full-length template for the next step of library construction.

2.4 mRNA-display methods

2.4.1 Modification of DNA templates at the 5' and 3' ends by PCR

DNA templates were modified at their 5' and 3' termini by PCR amplification using gene specific primers with constant regions required to facilitate mRNA display (see Appendix for oligonucleotide sequences). PCR was performed in the manufacturer's 1 x PCR buffer, with a final primer concentration of 500 nM, dNTP concentration of 200 μ M, a template DNA concentration of 5 nM, and final polymerase concentration of 0.02 U. μ L⁻¹. All reactions were made up to their final volume with nuclease free water. Reaction components were combined in sterile, thin-walled PCR tubes and placed in an MJ research PTC-100 thermocycler. For a typical DNA modification reaction using Q5 polymerase, the following cycling conditions were used: 98 °C, 1 min, 98 °C, 30 sec 62 °C, 30 sec 72 °C, 1 min. Aliquots of PCR reactions were visualised by agarose gel electrophoresis to confirm accurate and high yielding amplification. PCR products were phenol:chloroform extracted using the method described in section 2.3.1.

2.4.2 *In vitro* transcription

mRNA was generated by *in vitro* transcription of PCR generated DNA templates using T7 RNA polymerase according to the recommended protocols. Reactions were performed in T7 buffer (200 mM HEPES, 35 mM MgCl₂, 2 mM spermidine, pH 7.5). A typical reaction contained 50-100 nM linear DNA template, 5 mM ATP, CTP, GTP, and UTP, 40 mM DTT, 0.1 mg.mL⁻¹ BSA, 200 U.mL⁻¹ RNasin (Promega, Southampton UK), 1 U.mL⁻¹ inorganic pyrophosphatase, and 3000 U.mL⁻¹ T7 RNA polymerase.

Components were mixed thoroughly, followed by incubation at 37 °C for 4 - 6 h. After which, 2 U TURBO DNase (Ambion) was added and reaction was further incubated for 30 min. DNase was inactivated by addition of EDTA to 15 mM followed by incubation at 75 °C for 10 min. RNA was precipitated by addition of LiCl to a final concentration of 2.5 M, followed by incubation at -20 °C for at least 1 h. RNA was pelleted by centrifugation at 13,000 × *g* for 10 min, and washed with ice cold 70 % ethanol. The RNA pellet was resuspended in DEPC-treated water and stored at -20 °C.

2.4.3 Puromycin linker synthesis

The psoralen-puromycin containing linker oligonucleotide (5'-X(uagccggugc)AAAAAAAAAAAAAAAAZZACCP-3' X = psoralen C6, lower case letters = 2'-OMe RNA, Z = spacer 9, P = puromycin, A/C = DNA) was synthesised on an ABI 394 DNA synthesiser using standard phosphoramidite chemistry¹⁴⁹. Puromycin-CPG, 2'-O-Me-RNA phosphoramidites, DNA phosphoramidites, spacer phosphoramidite 9, and psoralen C6 phosphoramidite were used according to recommended protocols (Glen Research, VA, USA). Deprotection was performed in concentrated ammonium hydroxide for 8 h at 55 °C, followed by butanol precipitation. After drying, linkers were resuspended, and stored in nuclease free water at -20 °C.

2.4.4 Photo-crosslinking reactions

The puromycin linker oligonucleotide was annealed to template mRNAs in 1× XL buffer (100 mM KCl, 1 mM spermidine, 1 mM EDTA pH 8.0, and 20 mM HEPES, pH 7.5) by heating the reaction to 80 °C for 3 min, followed by cooling to 25 °C over 5 min. Reactions contained 3 μM RNA and 7.5 μM puromycin linker. The reaction was irradiated on ice for 20 min at a distance of ~ 2 cm using a 6 watt handheld UV lamp model EN-160L (Spectroline, USA) set to 365 nm. Photo-crosslinked product mixtures were ethanol precipitated (section 2.3.2) and stored in DEPC-treated water at -20 °C.

2.4.5 Translation *in vitro* and mRNA-protein fusion formation

Translation reactions were performed in nuclease treated rabbit reticulocyte lysate, (Promega, Southampton UK) for 1 h at 30 °C. Typically, reactions contained 200 nM photo-crosslinked mRNA, 25 μM each amino acid except methionine, 10 μCi.μL⁻¹ L-[³⁵S] methionine, 1 U.μL⁻¹ RNasin ribonuclease inhibitor (Promega, Southampton UK) and 40% (v/v) lysate. Magnesium acetate and potassium chloride concentrations were optimised for individual templates for highest mRNA-protein fusion yield. After translation, fusion formation was promoted by the addition of potassium chloride and magnesium chloride to final concentrations of 531 mM and 50 mM respectively, followed

by further incubation for 5 min at room temperature. Translation products were analysed by SDS-PAGE electrophoresis followed by gel drying and visualisation by autoradiography using an FLA-5100 phosphorimager.

2.4.6 Oligo-dT cellulose synthesis

Oligo-dT cellulose was synthesised based on previously described protocols¹⁵⁰ using an ABI 394 solid phase DNA synthesiser and 1 μ mole synthesis columns. Empty columns were packed with 75 mg of acid washed chromatography grade cellulose (Sigma-Aldrich, Cat. 22184) and loaded onto the ABI 394. Columns were purged with dry acetonitrile and then with dry argon to ensure that all cellulose particles were securely in the column, followed by synthesis. The standard 1 μ mole synthesis reaction on an ABI 394 DNA synthesiser was performed with two modifications. First, the coupling time was increased to 300s, and second, the capping steps were removed. Synthesis was performed for 25 cycles, after which the cellulose was transferred to a 1.5 mL Eppendorf and treated with 1.0 mL 30% ammonium hydroxide for 1 h at room temperature to remove cyanoethyl protecting groups. The cellulose was precipitated by centrifugation and the supernatant discarded, resin was then washed twice with 1.0 mL DEPC treated H₂O. The cellulose slurry was lyophilised to dryness by freeze drying.

2.4.7 Oligo-dT cellulose binding assay

To evaluate the binding capacity of the synthesised oligo-dT cellulose to dA₁₅ oligonucleotides, 5 mg lyophilised oligo-dT cellulose was added to a 1.5 mL Eppendorf tube and suspended in 100 μ L dA₁₅ binding buffer (1 M NaCl, 10 mM Tris-HCl, pH 7.4). A solution of dA₁₅ oligonucleotide in binding buffer was added, and the tube was incubated for 15 min at room temperature with agitation. The cellulose was precipitated by centrifugation and the supernatant removed. Bound dA₁₅ was quantified by measuring the UV absorbance of the supernatant at 260 nm before and after incubation with the oligo-dT cellulose resin.

2.4.8 Oligo-dT cellulose purification of mRNA-protein fusions

Purification of mRNA-protein fusion molecules was performed in batch using the oligo-dT cellulose synthesised according to Section 2.4.6. Crude *in vitro* translation mixtures were diluted 10-fold into oligo-dT binding buffer (1 M NaCl, 10 mM EDTA pH 8.0, 0.2% (w/v) Triton X-100, 20 mM Tris-HCl pH 8.0), followed by incubation for 45 min at 4 °C. The

oligo-dT cellulose slurry was precipitated by centrifugation at 500 × *g* for 1 minute, and supernatant removed. The oligo-dT cellulose resin was washed 3 times with 1 mL oligo-dT binding buffer, followed by a final wash with 1 mL oligo-dT wash buffer (300 mM KCl, 20 mM Tris-HCl, pH 8.0). mRNA-protein fusions were eluted in two volumes of oligo-dT elution buffer (2 mM Tris-HCl, pH 8.0). Purified mRNA-protein fusions were visualised by SDS-PAGE electrophoresis followed by autoradiography using an FLA-5100 phosphorimager. Radiolabelled mRNA-protein fusions were quantified by scintillation counting in Emulsifier Safe scintillation cocktail (Perkin Elmer) on a Packard 1600TR Liquid Scintillation Counter.

The number of fusions in the oligo-dT eluate could be calculated using the following equation:

$$n_{\text{prot}} = \frac{N_A \times [\text{Met}] \times \%_{\text{Met inc.}} \times \text{Vol}_{\text{TL}}}{N_{\text{Met}}}$$

Where n_{prot} = number of proteins, N_A = Avogadro constant, $[\text{Met}]$ = total concentration of methionine in translation, $\%_{\text{Met inc.}}$ = fraction of radioactivity (^{35}S -methionine) incorporated into mRNA-displayed proteins, Vol_{TL} = volume of *in vitro* translation reaction, N_{Met} = number of methionines per protein.

2.4.9 Preparation of peptide-coupled agarose beads

The relevant N-terminally biotinylated peptide (1 mg.mL⁻¹, supplied by Laura Cross, University of Leeds, UK), was incubated with streptavidin agarose (Thermo Fisher Scientific, cat no. 20347) in selection buffer (100 mM potassium phosphate buffer, pH 7.3, 0.1% tween) at 4 °C at a ratio of 50 µg peptide per 100 µL settled resin. After 1 h, the resin was collected by centrifugation at 500 × *g* for 1 min, and the supernatant removed. The resin was subsequently washed 10 times with 1 mL selection buffer per 100 µL settled resin to remove unbound peptides. Resin was stored at 4 °C prior to use in selection experiments. Fresh peptide-coupled beads were prepared prior to each selection step.

2.4.10 PEX5*-peptide selection

For round 1 of selection, the fully-assembled PEX5* library was amplified by 10 cycles of PCR using the primers PEX5 Mod For, and PEX5 Mod Rev to append the required sequences for mRNA display (T7 promoter, TMV enhancer, puromycin link site). The modification PCR product was purified using a silica spin column and DNA concentration determined by nanodrop. A total of 8.6×10^{12} molecules of the modified PEX5* library

were transcribed *in vitro* using T7 RNA polymerase (section 2.4.2) for 6 h, followed by digestion of the template DNA using DNase I. RNA was precipitated using LiCl ready for UV photocrosslinking. 1.7×10^{13} RNA molecules were included in the crosslinking reaction with the puromycin oligonucleotide, which upon completion was precipitated to yield a solution of approximately 20 μM PEX5*-puromycin RNA. This was used as template in an *in vitro* translation reaction in rabbit reticulocyte lysate (total volume 5 mL). Oligo-dT purification from the rabbit reticulocyte lysate yielded 8×10^{12} RNA-protein fusions (as determined by scintillation counting), which were reverse transcribed to generate cDNA fusions for the selection step. The reverse transcription reaction was diluted 5-fold into selection buffer (100 mM potassium phosphate buffer, pH 7.3, 0.1% tween), and pre-selected with 500 μL unmodified agarose beads for 1 h at 4 °C with rotation. The supernatant was collected, split into three equal aliquots and incubated with the corresponding peptide-coupled agarose bead (YQSEV, YQSY Y, YQSFY) for 1 h at 4 °C with rotation. After 1 h, the beads were washed 10 times with selection buffer and used directly as template in an RT-PCR reaction to regenerate the libraries (Figure 8). Fractions from the oligo-dT purification and reverse transcription were analysed by SDS-PAGE autoradiography

2.4.11 Interaction-dependent RT-PCR assay

RNA-protein fusions were generated and purified as described in sections 2.4.1-2.4.8. Oligo-dT cellulose purified RNA-SBP fusions were incubated with DNA-tagged streptavidin (50 nM) in a total volume of 20 μL for 45 min at room temperature. The mixture was then diluted two-fold with reverse transcriptase reaction buffer to a final concentration of 50 mM Tris-HCl, pH 8.3; 75 mM KCl; 3 mM MgCl_2 , 10 mM DTT, 0.1 $\text{U}\cdot\mu\text{L}^{-1}$ RNasin® Plus (Promega), and 0.5 mM each dNTP. The IDRT reaction was incubated at 37 °C for 5 min, after which reverse transcriptase enzyme (Superscript II, Invitrogen) was added to a final concentration of 0.5 $\text{U}\cdot\mu\text{L}^{-1}$. Where applicable, IDRT reactions were supplemented with exonuclease I at final concentration of 0.5-2 $\text{U}\cdot\mu\text{L}^{-1}$. Upon the addition of enzyme(s), IDRT reactions were incubated at 37 °C for a further 15 min, after which the reaction was heated to 80 °C for 20 min to inactivate the reverse transcriptase and exonuclease enzymes. cDNA was amplified by PCR in 1x Q5 polymerase buffer (25 mM TAPS-HCl (pH 9.3), 50 mM KCl, 2 mM MgCl_2 , 1 mM β -mercaptoethanol), 0.2 mM dNTPs, 0.5 μM each forward and reverse primer, 1 μL IDRT reaction, 0.02 $\text{U}\cdot\mu\text{L}^{-1}$ Q5 DNA polymerase (NEB), in a total reaction volume of 50 μL . The thermocycling conditions were as follows: initial denaturation at 98 °C for 30 s, followed by 10 s denaturation at 98 °C; 20 s annealing at 61 °C; and 20 s extension at 72 °C for

25 cycles, followed by a final extension at 1 min extension at 72 °C. PCR products were analysed via 2 % agarose gel electrophoresis. Positive and negative DNA sequences were assessed by digestion with 2 U.µL⁻¹ SacI-HF (NEB) for 1 h at 37 °C prior to electrophoresis.

2.4.12 Mock library selection by IDRT-PCR

Three 100 µL *in vitro* translation reactions were set up as per section 2.4.5, with SBP:SBPΔ template ratios of 1:10, 1:100, and 1:1000, yielding a final template concentration of 200 nM. The resulting RNA-protein fusions were purified as described (section 2.4.8), and incubated separately with DNA-tagged streptavidin (50 nM) in a total volume of 20 µL for 45 min at room temperature. The mixtures were then subjected to IDRT-PCR as described in section 2.4.11, with a 2:1 ratio of Superscript II (0.5 U.µL⁻¹) to exonuclease I (1 U.µL⁻¹). PCR products were digested with 2 U.µL⁻¹ SacI-HF (NEB) for 1 h at 37 °C, followed by analysis via 2 % agarose gel electrophoresis.

2.5 Protein methods

2.5.1 Determination of protein concentration

Protein concentration was determined spectrophotometrically at 280 nm and calculated using the Beer-Lambert law:

$$A = \epsilon cl$$

Where A = absorbance at 280 nm, ϵ is the molar extinction coefficient, c is the concentration of the protein, and l is the path length of the spectrophotometer. Molar extinction coefficients were calculated based on amino acid sequence using the ExPASy ProtParam tool¹⁵¹.

2.5.2 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-polyacrylamide gel electrophoresis was carried out according to the protocol described by Sambrook *et al*¹⁴⁷. Pre-stained molecular weight markers 11-190 kDa (P7706) were obtained from New England BioLabs (Ipswich, MA,USA). Proteins were separated in a polyacrylamide resolving gel with a polyacrylamide stacking gel.

Protein samples were mixed in a 1:1 ratio with SDS-PAGE loading buffer (50 mM Tris-HCl; pH 6.8, 2% SDS, 20% glycerol, 1% β-mercaptoethanol, 12.5 mM EDTA, 0.02 %

bromophenol blue) and boiled for 5 min before loading onto the stacking gel. The gel cassette was placed in an electrophoresis tank with SDS-PAGE running buffer (25 mM Tris-HCl pH 8.3, 192 mM glycine, 0.1% (w/v) SDS) providing conductance between the gel and both the anode and cathode. Electrophoresis was performed at a constant current of 30 mA per gel, for approximately 3 h until the dye-front reached the base of the gel. Gels were stained for 1 h in a methanol–acetic acid–water solution (5:1:1) containing 0.1% (w/v) Coomassie Brilliant Blue, and were destained in a solution of the same but with no Coomassie Brilliant Blue.

SDS-PAGE resolving gel:

3,750-7,500 μ L	30% acrylamide
3,750 μ L	1.5 M Tris-HCl pH 8.8
150 μ L	10% (w/v) SDS
3,500-7250 μ L	H ₂ O
50 μ L	25% (w/v) ammonium persulfate
5 μ L	TEMED

SDS-PAGE stacking gel:

625 μ L	30% acrylamide
625 μ L	1 M Tris-HCl pH 6.9
50 μ L	10% (w/v) SDS
3,650 μ L	H ₂ O
50 μ L	25% (w/v) ammonium persulfate
5 μ L	TEMED

2.5.3 Ni²⁺-NTA purification of His₆-tagged proteins from *E.coli*

Expression cultures were grown in 2TY media supplemented with the appropriate antibiotic. Cells were grown at 37 °C to an A₆₀₀ of 0.8. At this point the temperature was lowered to 30 °C to promote maximal expression of soluble protein, and expression was induced by the addition of 1 mM IPTG. Incubation at 30 °C was continued overnight, after which cells were harvested by centrifugation (6000 × *g*, 20 min, 4 °C). For purification the cell pellet was resuspended in 50 mL Ni²⁺-NTA resuspension buffer (10 mM Tris-HCl, 500 mM NaCl, 5 mM imidazole, pH 7.5). Cells were lysed using a cell disruptor (Constant Systems, Daventry, UK). Lysate was centrifuged at 12,000 × *g* for 45 min at 4 °C, the resulting supernatant applied to Ni²⁺-loaded Chelating Sepharose

Fast Flow™ resin, followed by incubation for 20 min at 4 °C with shaking. A total of four wash steps with Ni²⁺-NTA wash buffer (10 mM Tris-HCl, 500 mM NaCl, 50 mM imidazole, pH 7.5) were incorporated to remove unspecific bound proteins by increasing the imidazole concentration from 5 mM to 50 mM, followed by elution of protein at an imidazole concentration of 300 mM using Ni²⁺-NTA elution buffer (10 mM Tris-HCl, 500 mM NaCl, 300 mM imidazole, pH 7.5). Elution fractions were dialysed twice against 5 L dialysis buffer (10 mM Tris-HCl, pH 7.5), at 4 °C overnight and for 4 h, respectively. Dialysed protein was subsequently sterile filtered and stored at 4 °C.

2.5.4 Dialysis and concentration

Protein samples to be dialysed were placed in dialysis tubing bags (12-14 kDa molecular weight cut-off) and were dialysed against 50-100 times the volume of the sample. Equilibration was obtained after storing for 4 h with gentle stirring at 4 °C. The dialysis tubing bag was then transferred to newly made dialysis buffer and allowed to equilibrate for a further 4 h with gentle stirring at 4 °C. Alternatively, samples were buffer exchanged in a spin column format using either Zeba™ Spin Desalting Columns (Thermo Fisher Scientific) or Micro Bio-Spin P30 (Bio-Rad) according to the manufacturer's guidelines. Protein samples were concentrated using Vivaspin® (Sartorius Stedim) centrifugal concentrators according to the manufacturer's guidelines.

2.5.5 Covalent coupling of DNA oligonucleotides to streptavidin

Streptavidin (Generon) was reacted with excess sulfosuccinimidyl 4-(N-maleimidomethyl)cyclohexane-1-carboxylate (sulfo-SMCC; Thermo Fisher Scientific) in a total volume of 600 µL, consisting of 3.8 mg.mL⁻¹ protein and 0.25 mg.mL⁻¹ sulfo-SMPB in PBS (100 mM potassium phosphate buffer, pH 7.3, 150 mM NaCl). After incubation for 1 h at room temperature, the reaction was desalted by size exclusion chromatography (Micro Bio-Spin P30; Biorad) and buffer exchanged into PBE (100 mM phosphate buffer, pH 6.8, 5 mM EDTA). The derivatised STV in PBE was mixed with 100 µL of a deprotected thiol-modified oligonucleotide (0.7 µM in PBE). Following incubation at room temperature for 2 h, excess DNA was removed by ultrafiltration (Microcon 30, Millipore). Purification and fractionation was carried out by ion exchange chromatography on a HiTrap Q HP column (GE life sciences) using an AKTA prime chromatography system (GE life sciences). Buffer A (20 mM Tris-HCl, pH 6.3, 0.3 M NaCl); sample injection; 4 ml buffer A, followed by a linear salt gradient of 22.6 mM.mL⁻¹ using buffer B (20 mM Tris-HCl, pH 6.3, 1 M NaCl); flow rate, 0.2 mL.min⁻¹. The fractions

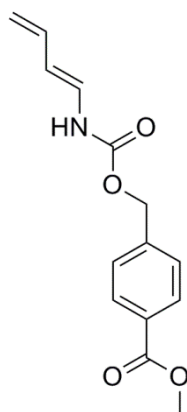
corresponding to DNA-streptavidin conjugates were collected, concentrated by ultrafiltration and stored in 20 mM Tris-HCl pH 7.3, at 4 °C.

2.6 mRNA display substrate synthesis

2.6.1 Synthesis of 5'-acrylamide oligonucleotides

For a typical labelling reaction, acrylic acid N-hydroxysuccinimide ester (Sigma Aldrich, A8060) was prepared as a 10 mg/mL stock solution in DMF. In the reaction 45 μ L of the NHS ester was added to a solution of 10 nmol of amino-terminated oligonucleotide (Integrated DNA Technologies) in 100 μ L of 0.2 M phosphate buffer pH 8.0. After 4 h at 25° C, unreacted NHS ester was removed by gel filtration (Micro Bio-Spin P30; Biorad), buffer exchanged into 20 mM Tris-HCl pH 8.0, and stored at -20 °C. Oligonucleotide labelling efficiency was analysed by ESI-MS.

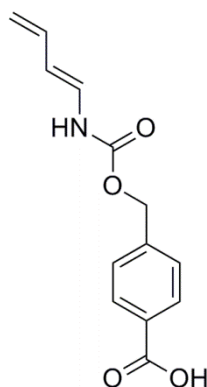
2.6.2 p-(methoxycarbonyl)benzyl trans-1,3-butadiene-1-carbamate



2,4-pentadienoic acid (2.48 g, 25.3 mmol) was dissolved in 126.5 mL DMF in a dry 250 mL round-bottomed flask equipped with a magnetic stirrer and a nitrogen inlet. Diphenylphosphoryl azide (6 mL, 27.8 mmol) and triethylamine (3.53 mL, 25.3 mmol) were added, and the reaction stirred for 10 minutes at room temperature. Methyl 4-(hydroxymethyl) benzoate (5 g, 30 mmol) was then added, and the reaction heated to 100 °C with stirring for 16 hours. Upon completion, the reaction was cooled to room temperature, washed with 250 mL H₂O, and extracted with ethyl acetate (4 x 120 mL). The combined organic layers were washed with brine, dried over MgSO₄, and evaporated *in vacuo* to yield an oily orange residue. The crude residue was subjected to

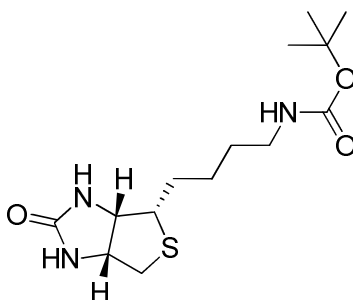
flash chromatography (SiO_2 , 9:1 petroleum ether:ethyl acetate) to afford the desired product (1.7 g, 51%) as a pale yellow solid. m/z (ES^+) 261.1. ^1H NMR and ^{13}C NMR (DMSO, 500 MHz) in agreement with previously reported data.

2.6.3 4-Carboxybenzyl *trans*-1,3-butadiene-1-carbamate



A solution of $\text{LiOH}:\text{H}_2\text{O}$ (3 equiv. 818 mg) in 10 mL of water was added to a solution of p-(methoxycarbonyl)benzyl *trans*-1,3-butadiene-1-carbamate (1 equiv., 1.7 g, 6.5 mmol) in 20 mL of THF. The mixture was stirred at room temperature for 20 hours. Upon completion, the reaction was washed with 15 mL ethyl acetate, the aqueous layer acidified to pH 4 by the addition of citric acid, and extracted with ethyl acetate (5 x 20 mL). The combined organic layers were washed with brine, dried over MgSO_4 , and evaporated *in vacuo*. This yielded the acid as a white powder (1.40 g, 87%). ^1H NMR and ^{13}C NMR (DMSO, 500 MHz) in agreement with previously reported data. m/z (ES^-) 283 (MCl^-).

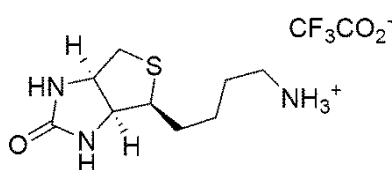
2.6.4 N-Boc-norbiotinamine



Biotin (2.5 g, 10.23 mmol) was dissolved in 36.5 mL dry tert-butanol, followed by addition of diphenyl phosphoryl azide (DPPA) (2.43 mL, 11.25 mmol), and triethylamine (1.57 mL, 11.25 mmol). The reaction was stirred for 20 h at 70 °C. The solution was concentrated *in vacuo*, followed by purification by flash chromatography (SiO_2 , 95:5

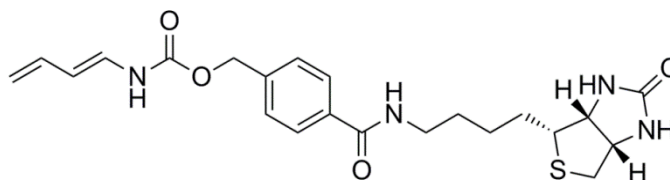
CH₂Cl₂:MeOH) to afford the desired product (2.59 g, 80%) as a white solid; ¹H NMR (500 MHz, MeOD) δ 4.49 (1H, dd, *J* = 7.6, 5.0 Hz), 4.31 (1H, dd, *J* = 7.8, 4.5 Hz), 3.23 – 3.18 (1H, m), 3.12 – 2.98 (2H, m), 2.93 (1H, dd, *J* = 12.7, 5.0 Hz), 2.71 (1H, d, *J* = 12.7 Hz), 1.75 (1H, td, *J* = 13.9, 7.6 Hz), 1.66 – 1.46 (4H, m), 1.43 (9H, s); ¹³C NMR (500 MHz, MeOD). *m/z* (ES⁺) 315.2.

2.6.5 Norbiotinamine trifluoroacetate



Trifluoroacetic acid (5 mL, 8 eq.), pre-chilled to 0 °C, was added to N-Boc-norbiotinamine (2.59 g, 8.21 mmol) followed by stirring for 1 h at 0 °C. The solution was warmed to room temperature and concentrated under reduced pressure. The crude residue was triturated in ether, yielding the product (2.6 g, 96%) as a white solid. ¹H NMR (500 MHz, MeOD) δ 4.52 (1H, dd, *J* = 7.4, 5.0 Hz), 4.33 (1H, dd, *J* = 7.5, 4.5 Hz), 3.27 – 3.15 (1H, m), 2.99 – 2.83 (3H, m), 2.73 (1H, d, *J* = 12.7 Hz), 1.85 – 1.56 (4H, m), 1.52 (2H, dd, *J* = 14.9, 7.5 Hz). ¹³C NMR (500 MHz, MeOD) δ 166.13, 63.25, 61.60, 56.75, 41.08, 40.45, 29.26, 28.41, 26.97. *m/z* (ES⁺) 329.1.

2.6.6 4-carboxybenzyl trans-1,3-butadiene-1-carbamate norbiotinamide



4-Carboxybenzyl trans-1,3-butadiene-1-carbamate (250 mg, 2 mmol) was added to dry acetonitrile followed by the addition of TBTU (320 mg, 2.5 mmol) and *N,N*-diisopropylethylamine (0.43 mL, 2.5 mmol), followed by stirring for 10 minutes at room temperature. Norbiotinamine trifluoroacetate (724 mg, 2.2 mmol) was then added

and the reaction was stirred at room temperature for 2.5 h. Upon completion, the reaction was concentrated *in vacuo*, followed by purification by flash chromatography (SiO₂, 95:5 DCM:MeOH) to afford the desired product (195 mg, 22%) as a white solid; R_f 0.32 (9:1 DCM:MeOH); ¹H NMR (500 MHz, MeOD) δ 7.82 (d, 2 H, J = 8.1 Hz), 7.46 (d, 2 H, J = 7.8 Hz), 6.69 (d, 1 H, J = 14.1 Hz), 6.28 (dt, 1 H, J = 10.2, 6.6 Hz), 5.79 (dd, 1 H, J = 13.2, 10.5 Hz), 5.20 (s, 2 H), 4.98 (dd, 1 H, J = 17.1, 1.2 Hz), 4.85 (d, 1 H, J = 10.2 Hz), 4.49 (1H, dd, J = 7.6, 5.0 Hz), 4.31 (1H, dd, J = 7.8, 4.5 Hz), 3.23 – 3.18 (1H, m), 3.12 – 2.98 (2H, m), 2.93 (1H, dd, J = 12.7, 5.0 Hz), 2.71 (1H, d, J = 12.7 Hz), 1.75 (1H, td, J = 13.9, 7.6 Hz), 1.66 – 1.46 (4H, m), 1.43 (9H, s); ¹³C NMR (500 MHz, DMSO). δ 165.69, 162.7, 139.34, 135.11, 134.15, 128.42, 127.49, 127.44, 123.69, 114.77, 65.51, 60.86, 59.11, 55.47, 29.10, 27.99, 26.02; *m/z* (ES⁺) 444.4.

3 Generating RNA-protein fusions for *in vitro* selection

In vitro selection for functional proteins using mRNA-display requires the ability to produce stable, covalent mRNA-protein fusions. This chapter describes the generation and purification of mRNA-protein fusions using the diisopropyl-fluorophosphatase (DFPase) enzyme from *L.vulgaris* as a model protein scaffold. The *L.vulgaris* DFPase enzyme is a single protein chain arranged into a six-bladed β -propeller architecture, and possesses a small pocket previously determined to be suitable for binding of Diels-Alder substrates^{62,152}. The physiological role of *L.vulgaris* DFPase has yet to be elucidated, however it has been identified as an organophosphate detoxifying enzyme that degrades several organophosphate compounds including DFP, sarin, cyclosarin, soman, and tabun¹⁵³. The DFPase enzyme has previously been used by Siegel *et al.* to generate the computationally designed Diels-Alderase DA_20_10⁶². An overview of the general experimental procedure for mRNA display selection is shown in Figure 3.1. The various steps required for a typical round of selection by mRNA display are discussed in detail below.

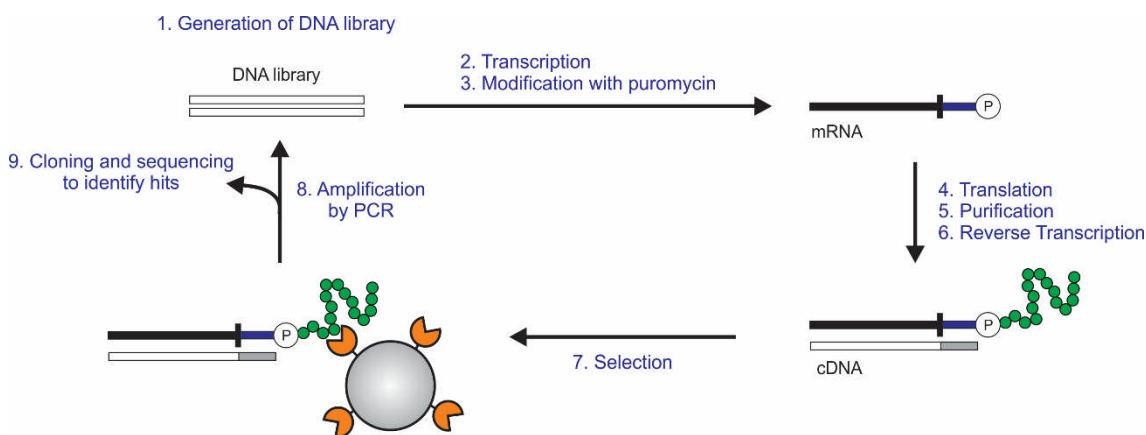


Figure 3.1. A schematic overview of the experimental procedure for a single round of mRNA display selection. A DNA library is transcribed into RNA, followed by addition of a puromycin to the 3'-end. Subsequent translation to produce mRNA-protein fusions, followed by reverse transcription enables functional selection, for example selection for binding against an immobilised target. Genetic material encoding functionally active proteins is amplified for further rounds of selection or cloning and characterisation.

A single round of mRNA display selection can be generally divided into 9 discrete *in vitro* steps. (1) Generation of a DNA library either by chemical synthesis or from natural

sequence(s). (2) Transcription of this DNA library into its corresponding mRNA. (3) Linking of puromycin to the 3'-end of the mRNA. (4) Generation of mRNA-protein fusions via *in vitro* translation of mRNA-puromycin templates. (5) Purification of mRNA-protein fusions from the translation extract. (6) Reverse transcription to generate cDNA-protein fusions. (7) Selection for desired function, for example by an immobilisation strategy. (8) Amplification by PCR to generate an enriched DNA library. (9) This enriched library can then either be used for further enrichment by iterative rounds of selection, or can be cloned and sequenced to isolate individual functional sequences. In contrast to the isolation of novel ligand binding proteins, where an affinity panning strategy can be used, a more specialised selection strategy is required for the generation of enzymes by mRNA display (discussed further in Results 3.5).

The key feature of mRNA display is the ability to form a covalent phenotype-genotype link via the aminonucleoside antibiotic puromycin (Figure 3.2). This stable link between a protein and its encoding nucleic acid allows selection at the functional level and subsequent recovery of the genetic information encoding the desired functionality. Puromycin, an aminoacyl-tRNA analogue, inhibits both eukaryotic and prokaryotic protein synthesis by entering the ribosomal A site and forming a stable amide linkage with the C-terminus of the nascent peptide chain as a result of ribosomal peptidyl transferase activity. This results in the termination of translation and premature release of the polypeptide from the ribosome¹⁵⁴. Almost 20 years ago, two groups independently surmised that attachment of puromycin at the 3'-end of an RNA transcript, followed by *in vitro* translation, would result in covalent attachment of the puromycin moiety to the protein chain – generating an mRNA-displayed protein^{124,125}. The covalent nature of the linkage between phenotype and genotype in mRNA-display permits selections in a greater range of conditions than comparable technologies such as ribosome display, where the conjugation between phenotype and genotype is noncovalent and thus more fragile¹¹¹.

The DFPase enzyme from *L. vulgaris* was identified as a suitable scaffold for establishing the mRNA display methodology based on a number of observations. A monomeric protein with an available crystal structure¹⁵², DFPase had also recently been used as a scaffold protein for the engineering of a Diels-Alderase enzyme using computational design methods^{62,65,66}. Establishing mRNA display methodology with the DFPase enzyme would therefore allow investigation of the potential for selection of enzyme activities based on this scaffold.

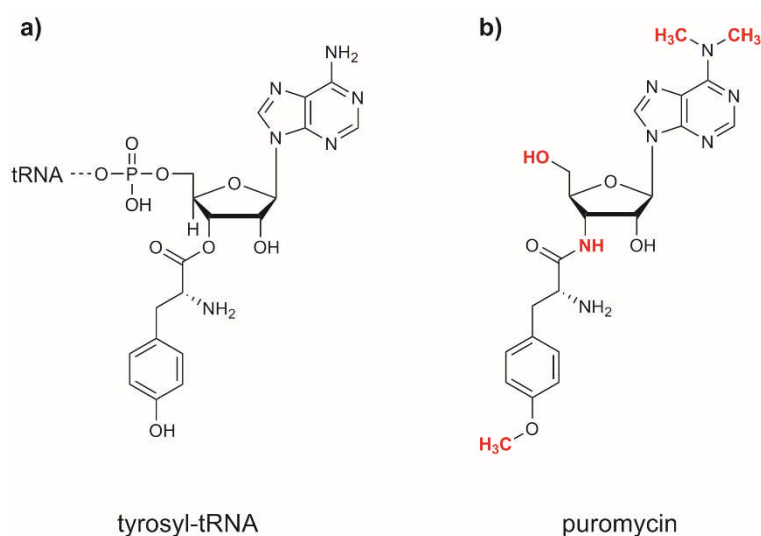


Figure 3.2. Structures of tyrosyl-tRNA **(a)** and puromycin **(b)**. Differences between the two molecules are highlighted in red.

3.1 Synthesis of 3'-puromycin oligonucleotides and templates

An essential requirement of any directed evolution experiment is the establishment of a genotype-phenotype link. Many techniques exist for the establishment of a physical linkage between genes and the proteins they encode, including but not limited to; display on phage coat proteins¹⁵⁵, eukaryotic and prokaryotic cells^{106,156}, and directly on encoding DNA¹¹⁶ and RNA¹¹¹. Many of these techniques require an *in vivo* step that drastically limits the maximum library size available for interrogation. Of the completely *in vitro* techniques, ribosome display¹¹¹ – conceptually the closest relative to mRNA display – allows access to very large libraries but is limited by the fragile, non-covalent association between phenotype and genotype. The covalent phenotype-genotype link provided by puromycin is a crucial aspect of the mRNA-display platform, and therefore the attachment of the mRNA to the puromycin-moiety is a key step.

Several strategies for this ligation have been previously described^{124,125,157,158}. Examples in the literature almost exclusively utilise a puromycin-containing oligonucleotide adaptor in combination with enzymatic or chemical ligation. Early cases of mRNA display relied on enzymatic ligation strategies (Figure 3.3a) involving incubation with T4 DNA or RNA

ligase and required an additional gel purification step, making them relatively time consuming and inefficient^{159,160}. The most well established method for the chemical ligation of puromycin is the psoralen-mediated photo-crosslinking strategy described by Kurz *et al.*^{129,158,161}. In this approach, a photoactivatable psoralen moiety is incorporated into the 5'-end of the puromycin linker, allowing covalent crosslinking of the mRNA and linker via irradiation under UV light (Figure 3.3b).

An oligonucleotide linker was designed (Figure 3.3c) for efficient UV photo-crosslinking to template RNA and subsequent RNA-protein fusion formation on the ribosome. The linker had the following features; (i) a 5'-psoralen moiety for covalent inter-strand crosslink formation with RNA templates upon irradiation with UV light ($\lambda = 365$ nm); (ii) a 10-base hybridisation region comprised of 2'-OMe RNA that is complementary to the 3'-constant region of template RNAs; (iii) a dA₁₅ region to facilitate purification of RNA-protein fusions using oligo-dT cellulose; (iv) hexaethylene glycol (TEG) spacer residues to give the linker more flexibility; (v) a puromycin moiety at the 3'-terminus.

The 10-nucleotide hybridisation region of the oligonucleotide linker was deliberately designed to form a duplex with the 3'-end of the target mRNA and contained a 5'-UA to facilitate efficient psoralen crosslinking^{162,163}. The hybridisation region was synthesised using 2'-OMe RNA, as this has been demonstrated to form a more stable duplex with RNA than either DNA or RNA oligonucleotides^{164,165}. Using 2'-OMe RNA also confers resistance to degradation of the mRNA by RNase H – an endoribonuclease that specifically hydrolyses the phosphodiester bonds of RNA hybridised to DNA¹⁶⁶ – which is present in rabbit reticulocyte lysate¹⁶⁷. The dA₁₅ region functions as a flexible linker to allow the 3'- puromycin to enter the ribosomal A-site and form a covalent linkage with the nascent peptide chain. It also serves a second purpose; facilitating affinity purification of mRNA-protein fusions from crude lysate using oligo-dT cellulose. Hexaethylene glycol spacer residues (Z, Figure 3.3c) are included given that the flexibility of linker has been previously shown to be an important factor in the efficiency of mRNA-protein fusion formation^{126,158}.

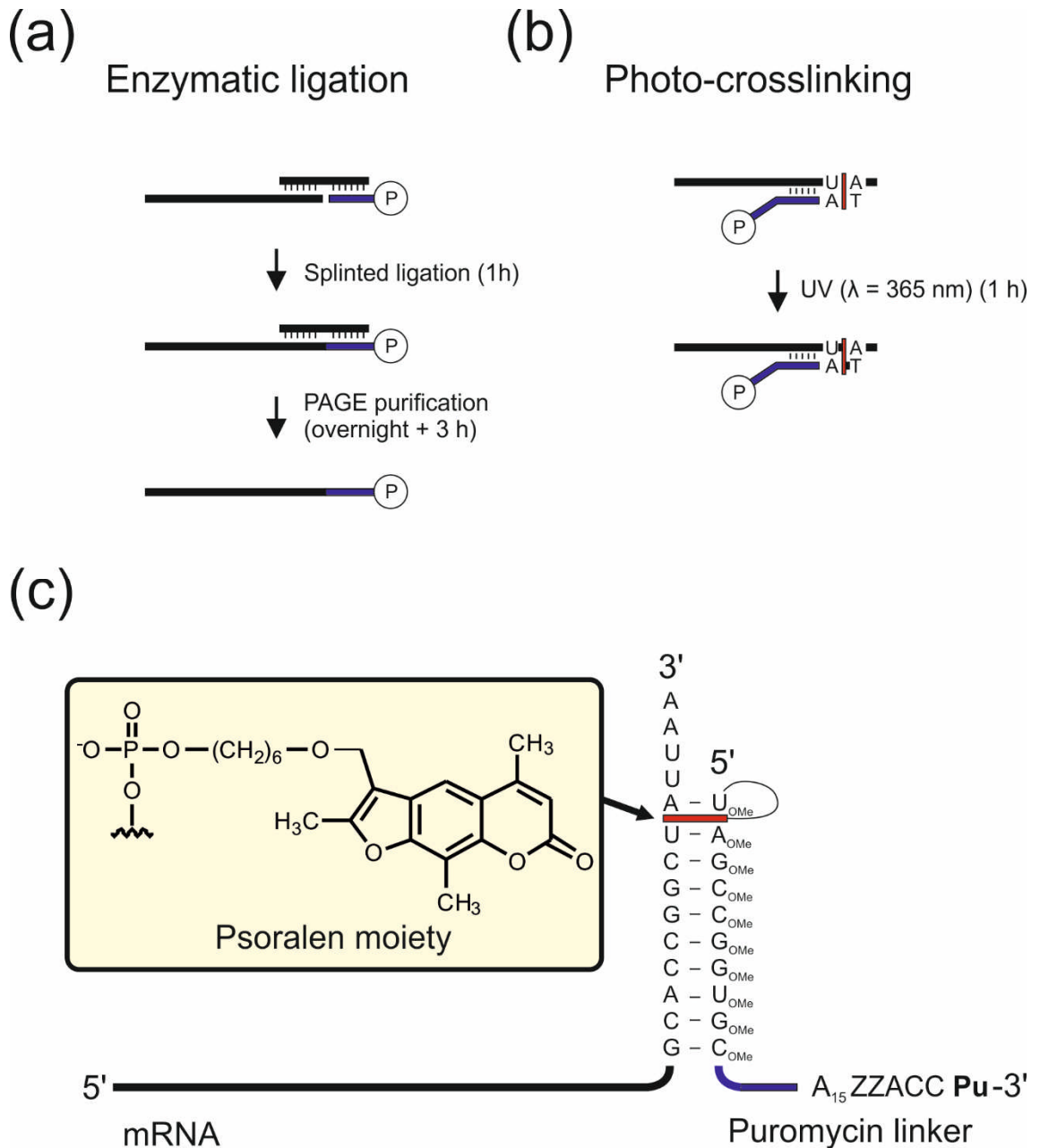


Figure 3.3. Step-by-step comparison of strategies typically used to generate 3'-puromycin templates for mRNA display. **(a)** Template directed enzymatic ligation using T4 DNA/RNA ligase requires a gel purification step and takes 1 day¹⁶⁸. **(b)** Photo-crosslinking of the puromycin to mRNA templates can be performed in 1 h and can be used in *in vitro* translation and fusion formation without further purification¹⁵⁸. **(c)** Duplex formation between the 3'-constant region of the mRNA (black) and the puromycin-containing oligonucleotide linker (blue). Z = hexaethylene glycol spacer, A/C = DNA bases, P = puromycin. The putative psoralen crosslinking site shown in red, the structure of psoralen is also shown.

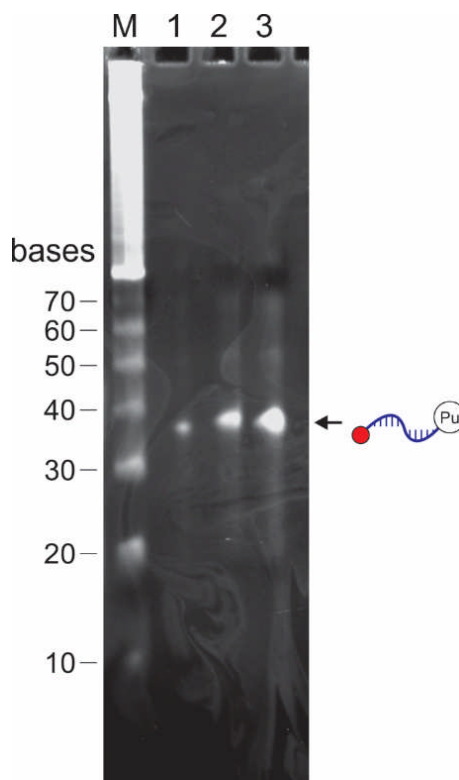


Figure 3.4. Denaturing PAGE of the synthesised puromycin-oligonucleotide (sequence shown in Figure 3.3). Lane 1, puromycin-oligonucleotide 1:100 dilution; Lane 2, puromycin-oligonucleotide 1:20 dilution ; Lane 3, puromycin-oligonucleotide 1:10 dilution; Lane M, 10 bp ladder. The puromycin-oligonucleotide is present at approximately the correct size and at an appropriate purity.

The puromycin-containing linker was synthesised using solid phase DNA synthesis using standard phosphoramidite chemistry (described in detail in Materials and Methods). Following oligonucleotide synthesis and deprotection, the length and purity of the puromycin-oligonucleotide was assessed by denaturing PAGE (Figure 3.4). The 30 base dual-modified oligonucleotide resolves on the denaturing PAGE gel at between 30 and 40 nucleotides, the reduction in mobility likely due to the additional mass of the psoralen and puromycin moieties at the 5'- and 3'-ends respectively. Mass spectrometry analysis gave a molecular mass for the synthesised oligonucleotide of 10283.76 Da (Figure 3.5) which agrees well with the calculated molecular mass of 10284.15 Da. These results demonstrated that synthesis had been successful, and that the oligonucleotide was present at sufficient purity for crosslinking to RNA.

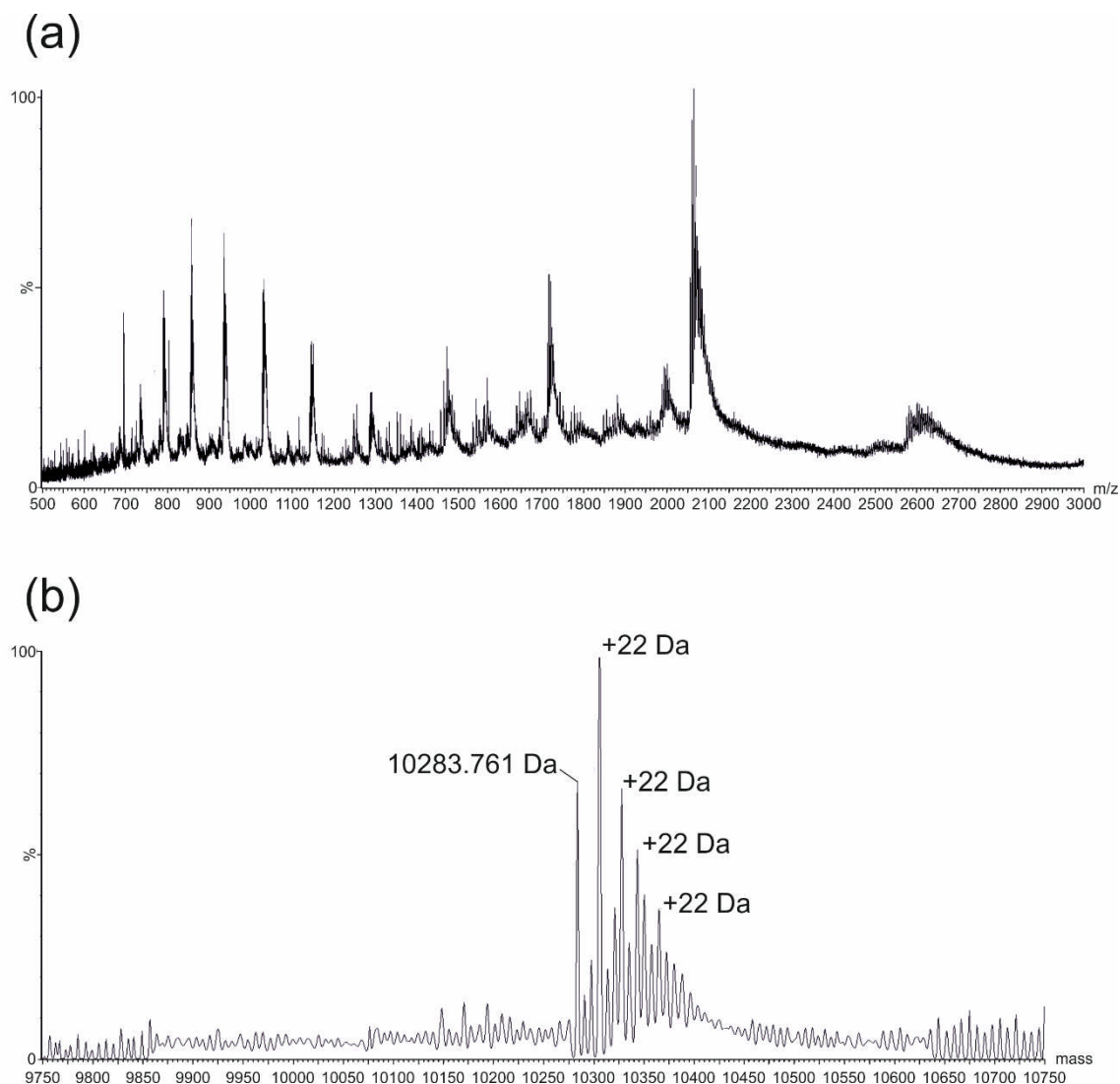


Figure 3.5. Mass spectrometry analysis of puromycin-oligonucleotide synthesised by phosphoramidite chemistry. **(a)** Negative ESI-MS m/z spectrum. **(b)** Molecular mass profile derived from **(a)**, indicating a molecular mass of 10283.76 Da which agrees well with the predicted molecular mass of 10284.15 Da. Peaks at +22n Da correspond to Na^+ adducts typical of oligonucleotide spectra.

Following characterisation, the ability of the puromycin linker oligonucleotide to covalently incorporate a puromycin moiety into the 3'-end of template RNA molecules via psoralen-mediated photo-crosslinking was tested¹⁵⁸. In this strategy (Figure 3.6), the puromycin linker is annealed to the 3'-end of RNA templates by heating followed by gradual cooling to 4 °C. Following hybridisation, the psoralen group is able to intercalate into the nucleic acid duplex and form covalent, inter-strand crosslinks with pyrimidine bases (preferentially at 5'-UA sites) upon irradiation with UV light ($\lambda = 365 \text{ nm}$)¹⁶³.

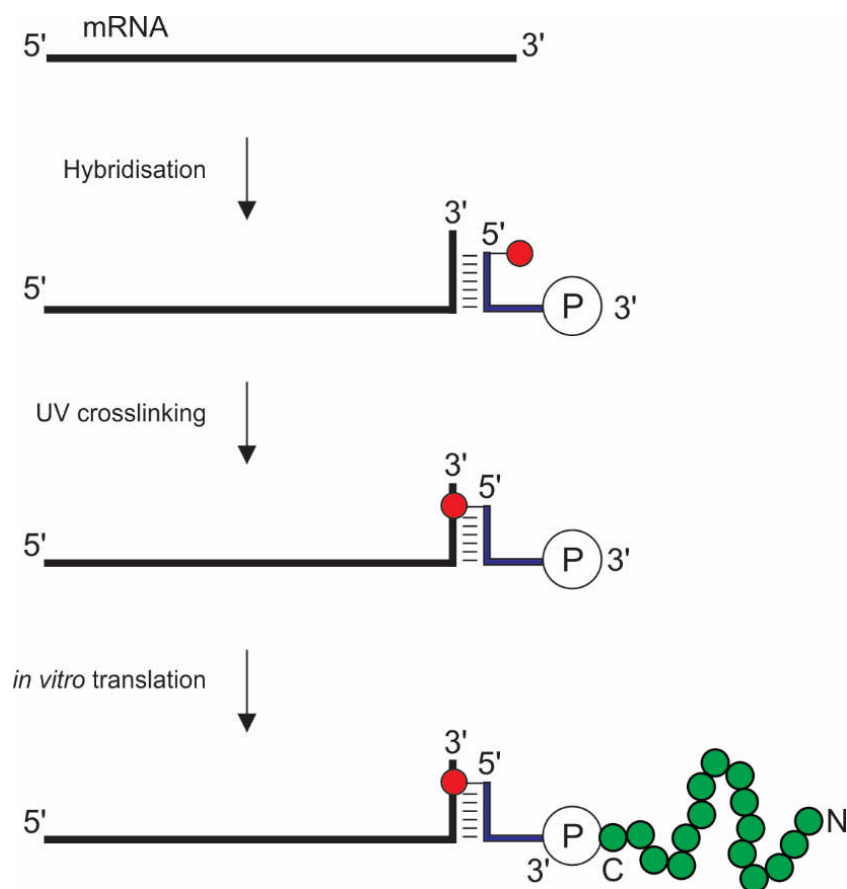


Figure 3.6. Schematic of puromycin attachment to mRNA by psoralen-mediated photo-crosslinking and subsequent mRNA-protein fusion generation. mRNA is hybridised to a DNA linker carrying psoralen (red circle) and puromycin (P) at the 5'- and 3'-termini respectively. Irradiation with UV light results in covalent crosslink formation between the linker and mRNA 3'-end. Subsequent *in vitro* translation of the photo-crosslinked RNA-puromycin template resulted in mRNA-protein fusion production.

Upon absorption of a long wavelength photon, either the furan-side or pyrone-side of molecule reacts with a pyrimidine base to form a cyclobutanyl monoadduct. Furan-side monoadducts can then absorb a second photon to cross-link the thymine base on the complementary strand of the DNA duplex^{169,170}. The product of this covalent crosslinking can then be used as a template for the generation of mRNA-protein fusions via *in vitro* translation. As discussed previously, RNA templates contained a 3'-terminal 10-nucleotide hybridisation sequence ending with 5'-UA to facilitate psoralen crosslinking¹⁶². In addition, a downstream stop codon (UAA) was included in order to induce ribosomal release from any uncrosslinked template mRNA present in the *in vitro* translation reaction, thus freeing up ribosomes for optimal RNA-protein fusion formation.

Due to the large size of the DFPase mRNA template (>1000 nucleotides), the relative size difference between mRNA and crosslinked mRNA cannot be easily distinguished using standard laboratory techniques such as gel electrophoresis. Therefore, in order to demonstrate that the photo-crosslinking step was functioning as expected, the photo-crosslinking protocol was initially performed using a 120 nucleotide 3'-RNA fragment from the DFPase gene as a model template (Figure 3.7a). This allowed the reaction to be monitored via denaturing PAGE (Figure 3.7b).

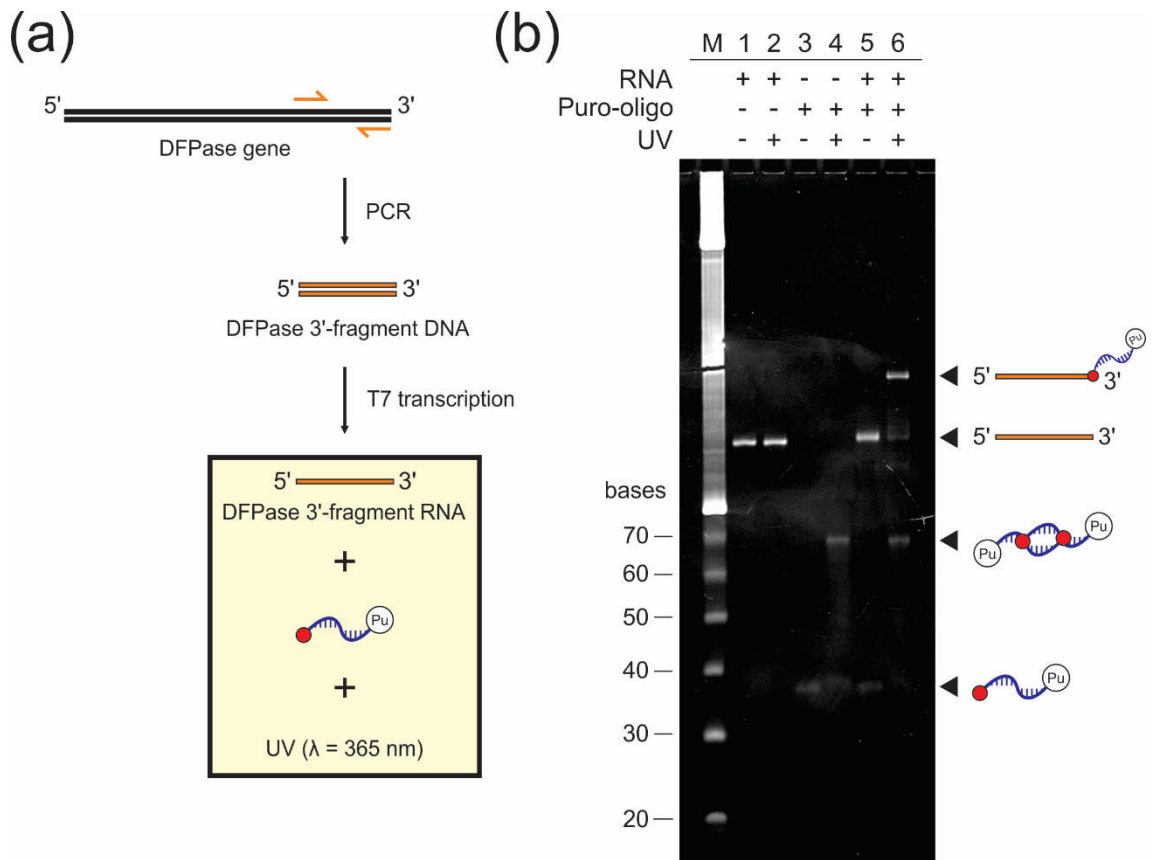


Figure 3.7. Psoralen-mediated photo-crosslinking between the puromycin-oligonucleotide and a model fragment of the DFPase mRNA. **(a)** The 120 nucleotide RNA fragment from the 3'-end of the DFPase gene was generated by PCR, followed by T7 transcription *in vitro*. The resulting RNA fragment was incubated either in the presence or absence of both the puromycin oligonucleotide and UV light ($\lambda = 365$ nm) **(b)** Denaturing PAGE analysis of photocrosslinking reactions. RNA, DFPase 3'-fragment RNA; Puro-oligo, puromycin oligonucleotide; UV, UV light ($\lambda = 365$ nm); M, 10 base pair ladder.

When the 3'-RNA fragment from the DFPase gene was incubated in the absence of the puromycin oligonucleotide, irradiation at 365 nm had no effect on the size of the RNA template (Figure 3.7, lanes 1 and 2), suggesting that the protocol does not result in degradation of RNA templates. Irradiation of the puromycin linker alone revealed that the oligonucleotide undergoes UV-induced covalent dimerisation, likely as a result of background hybridisation and psoralen intercalation (Figure 3.7, lanes 3 and 4). As expected, a clear shift in the apparent size of the RNA is observed only in the presence of both the puromycin linker and irradiation under UV light at 365 nm (Figure 3.7, lane 6). With this model RNA template, the crosslinking was highly efficient (> 75% of input RNA is covalently modified with the puromycin oligonucleotide as estimated by comparing the ratios of intensities of the bands in Figure 3.7, lane 6). This demonstrates that the synthesised puromycin linker can be rapidly and efficiently linked to RNA templates in a covalent manner using UV photo-crosslinking. Subsequently, all full-length templates for *in vitro* translation and RNA-protein fusion formation were prepared using the conditions established with this model RNA fragment.

3.2 *In vitro* translation and mRNA-protein fusion formation

Translation of puromycin-modified mRNA to generate a diverse pool of RNA-protein fusions is one of the most fundamental steps in the mRNA display process. The maximum library size that can be generated, and thus the upper limit of sequence space that can be explored, is defined primarily by both the size and efficiency of translation and the efficiency of RNA-protein fusion formation on the ribosome.

In principle, any *in vitro* translation system could be used to generate mRNA-protein fusions. Indeed, examples exist that utilise wheat germ¹⁷¹, bacterial extracts¹²⁶, mammalian cell lysates^{124,125}, and completely reconstituted systems using purified translation machinery¹⁷². A large proportion of the work performed on mRNA-display to date describes the use of rabbit reticulocyte lysate as the preferred *in vitro* translation medium. Previous investigations into the efficiency of fusion formation in different systems led to an early preference for rabbit reticulocyte lysate due to low levels of nucleic acid degradation in contrast to wheat germ and *E.coli* extracts¹⁷³ and a low dependence on mRNA 5'-capping for efficient translation¹²⁶.

The puromycin-modified RNA templates have two important features that promote efficient translation and fusion formation in rabbit reticulocyte lysate. Firstly, at the 5'-end, the presence of a tobacco mosaic virus (TMV) translation enhancer sequence in the template mRNA promotes ribosomal recruitment and results in efficient translation in eukaryotic *in vitro* translation systems¹⁷⁴. Secondly, at the 3'-end, the puromycin linker provides a covalently linked puromycin moiety to accept the nascent peptide chain during translation. The linker is also hybridised by base-pairing to the 3'-end of the RNA template, which acts to stall the ribosome and enable the puromycin end of the linker to access the ribosomal A site. The formation of RNA-protein fusions in the *in vitro* translation is shown schematically in Figure 3.8.

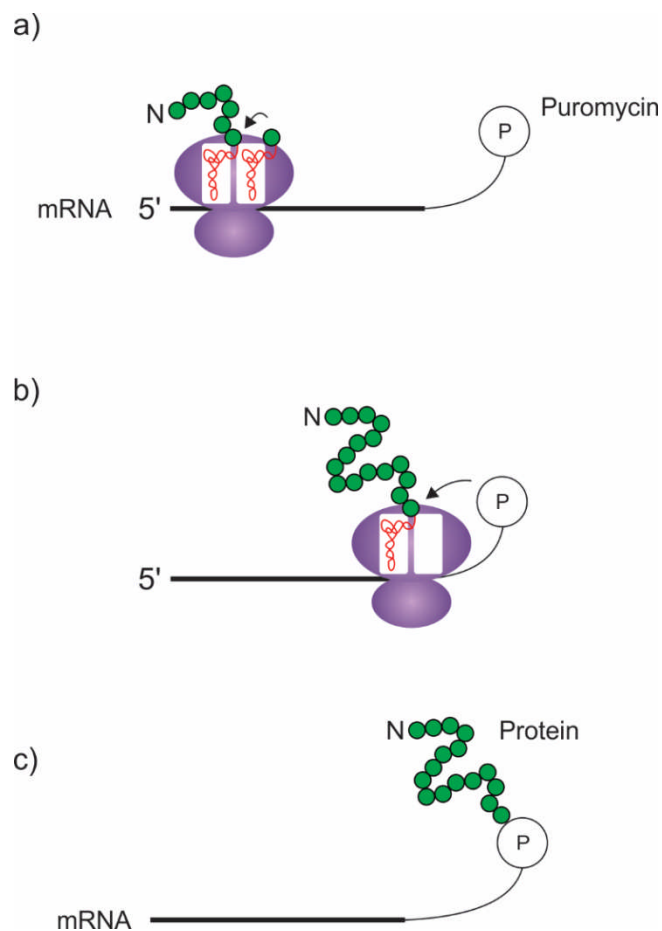


Figure 3.8. Schematic showing the formation of an mRNA-protein fusion on the ribosome. **(a)** The ribosome initiates translation of 3'-puromycin RNA templates in a 5'→3' direction. **(b)** Puromycin enters the ribosomal A-site resulting in transfer of puromycin to the C-terminus of the nascent peptide, generating a covalent RNA-protein fusion **(c)**. tRNAs (red) and amino acids (green) are shown in the P- and A- site of the ribosome.

Covalent mRNA-protein fusion formation is thought to occur in a relatively slow manner based upon two previously published observations^{125,126}. Firstly, homogeneity of fusion product length indicates that premature entry of puromycin into the ribosome is unlikely to occur during elongation, as this would result in a population of truncated fusion products heterogeneous in size¹²⁵. Furthermore, it has been observed that post-translational incubation under conditions that prevent ribosomal dissociation results in optimal fusion formation¹²⁶. This finding, together with previous data indicating that termination is a relatively slow step in translation¹⁷⁵, led to a model of RNA-protein fusion formation in which the 3'-puromycin templates are sequestered in a ternary complex with the ribosome following translation. These complexes then slowly give rise to the RNA-protein fusions as the puromycin moiety finds its way into the peptidyl transferase site of the ribosome¹²⁶. For these reasons, here, K^+ and Mg^{2+} ions were post-translationally added to *in vitro* translation reactions, followed by incubation at room temperature in order to encourage mRNA-DFPase fusion formation. *In vitro* translation of puromycin-modified DFPase mRNA was performed in rabbit reticulocyte lysate using template RNA generated as described in Results 3.1. To facilitate specific and sensitive detection of translation products, newly synthesised proteins were labelled by the addition of ³⁵[S]-methionine to *in vitro* translation reactions.

In vitro translation of *DFP* mRNA-puromycin conjugates generated a protein product commensurate with the expected apparent molecular mass of a covalent RNA-protein conjugate (~300 kDa) when resolved via SDS-PAGE (Figure 3.9a) Formation of this high molecular mass band was found to be proportional to input template concentration (Figure 3.9b).

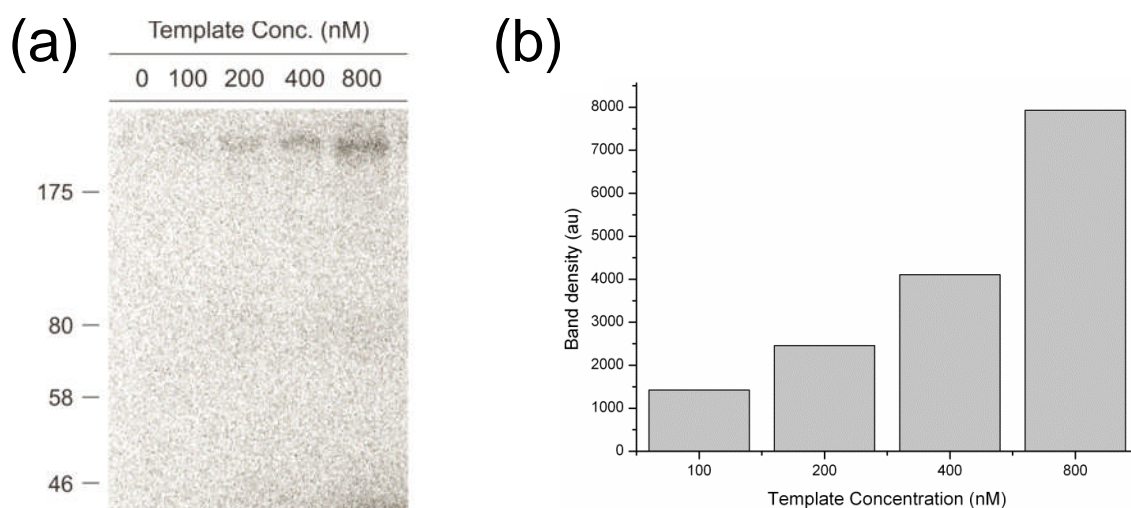


Figure 3.9. (a) SDS-PAGE autoradiography analysis of *in vitro* translation reactions shows RNA-DFPase fusion formation is proportional to input template concentration. (b) Data also expressed as radioactivity present in the RNA-protein fusion band as determined by densitometry. All translation reactions were performed at 30 °C for 1 hr using 0 – 800 nM DFPase RNA, 0.5 mM MgOAc, 100 mM KCl, and ^{35}S -Met as label. Approximate location of protein markers indicated. Densitometry was performed using ImageJ software¹⁷⁶.

To increase the efficiency of *in vitro* translation and RNA-protein fusion formation, the concentration of mono- and divalent cations must be optimised in the translation reaction. The concentration of Mg^{2+} has been previously shown to be the most critical parameter for efficiency and fidelity of *in vitro* translation in rabbit reticulocyte lysate, with a narrow optimal range of Mg^{2+} concentrations¹⁷⁷. Furthermore, RNA transcripts tend to exhibit unique optimal Mg^{2+} concentrations, therefore this parameter needs to be optimised for each template. The Mg^{2+} concentration for *in vitro* translation and fusion formation was optimised for the DFP template by varying the magnesium acetate (MgOAc) concentration in the translation reaction between 0.25 and 2.0 mM, and determining the level of RNA-protein fusion formation using SDS-PAGE, followed by autoradiography and densitometry (Figure 3.10).

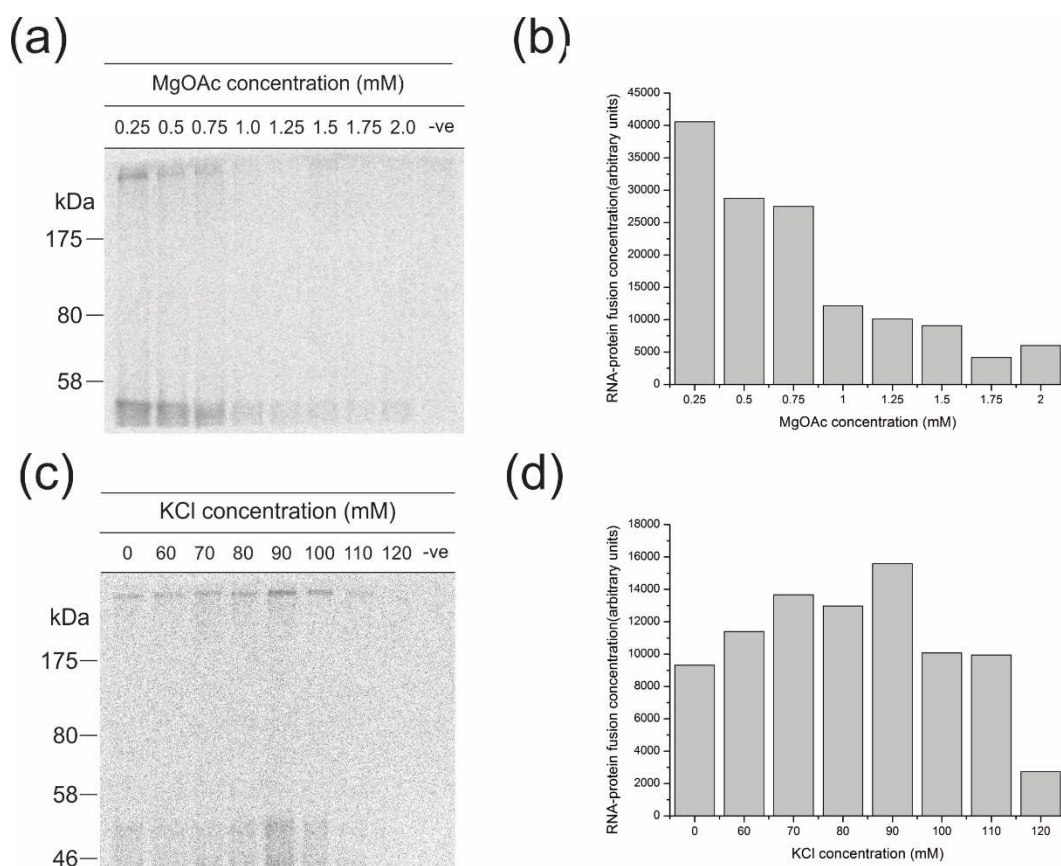


Figure 3.10. Optimisation of mono- and divalent salt concentration in the *in vitro* translation reaction for increased RNA-protein fusion formation. **(a)** Prior to translation, additional magnesium acetate (MgOAc) was added corresponding to 0 - 2.0 mM final concentration. Following translation, mixtures were assayed by SDS-PAGE followed by autoradiography. **(b)** Data also expressed as radioactivity present in the RNA-protein fusion band as determined by densitometry. Addition of 0.25 mM MgOAc results in optimal fusion formation. **(c)** Effect of additional potassium chloride (KCl) corresponding to 60 - 120 mM final concentration on fusion formation. **(d)** Data also expressed as radioactivity present in the RNA-protein fusion band as determined by densitometry. Addition of 90 mM KCl results in optimum RNA-protein fusion formation. All translation reactions were performed at 30 °C for 1 hr using 400 nM DFPase RNA, and ^{35}S -Met as label. Approximate location of protein markers indicated. Densitometry was performed using ImageJ software¹⁷⁶.

The highest yield of RNA-protein fusion formation was observed with a MgOAc concentration of 0.25 mM (Figure 3.10a). The next parameter for optimisation in the *in vitro* translation reaction is the K^+ concentration. Addition of KCl instead of KOAc has been shown to increase initiation fidelity in rabbit reticulocyte lysate¹⁷⁷ and is thought to result in more efficient translation and fusion formation¹²⁶. Furthermore, enhanced translational efficiency has been reported for uncapped RNA templates with viral 5'-UTRs when high levels (120 mM) of KCl are present¹⁷⁸. Carrying forward the optimal MgOAc concentration of 2.5 mM, the optimal KCl concentration in the reaction was then determined between 60 - 120 mM (Figure 3.10c). A final KCl concentration of 90 mM in

combination with an MgOAc concentration of 0.25 mM in the translation reaction gave the highest yield of DFPase RNA-protein fusions (Figure 3.10d).

3.3 Synthesis of oligo-dT cellulose

To circumvent interference from free RNA, protein, and components of the *in vitro* translation mixture, purification of the mRNA-protein fusions is desirable before performing selection experiments, especially when selecting for catalysis. Using rabbit reticulocyte lysate as the *in vitro* translation medium precludes purification via a genetically encoded His₆ tag using immobilised metal ion affinity chromatography due to co-purification of haemoglobin from the lysate^{179,180}. The predominant method for purification of mRNA-protein fusions from the crude lysate in the literature is via affinity capture on an oligo-dT derivatised solid support^{42,125,126,129,130,181}. This purification is facilitated by the presence of a designed poly-dA region in the puromycin linker, which is targeted for capture by oligo-dT residues on the solid-support via Watson-Crick base pairing.

Prior to commencing these studies, the commercial source of this high-capacity oligo-dT cellulose was discontinued. Similar resins are available for the purification of cellular mRNA, however these resins, both in our hands (data not shown), and in the hands of others¹⁵⁰, were insufficient for the purification of mRNA-protein fusions from crude lysates. This dearth of commercially available high-capacity oligo-dT cellulose was recognised and addressed by Sau *et al.* who described the solid-phase synthesis of oligo-dT cellulose for mRNA-protein fusion purification¹⁵⁰. This method uses standard phosphoramidite chemistry and solid-phase oligonucleotide synthesis protocols to sequentially add thymidine monomers to the surface of unmodified cellulose using an automated DNA synthesiser (Figure 3.11). During each synthesis cycle, the dT-phosphoramidite monomers are activated using 1*H*-tetrazole and coupled directly to cellulose (Figure 3.11, Step 1). The resulting phosphite linkage is oxidised to a phosphate ester using iodine and water (Figure 3.11, Step 2) followed by detritylation with trichloroacetic acid to yield a deprotected 5'-OH nucleoside (Figure 3.11., Step 3). Following completion of the synthesis protocol, protecting groups are removed via incubation with concentrated aqueous ammonia to yield the final oligo-dT cellulose matrix. The protocol differs from routine oligonucleotide synthesis in the following ways; i) cellulose is used as the solid-support rather than traditional derivatised controlled pore

glass (CPG); ii) the coupling time is increased to 300 seconds as the coupling efficiency is anticipated to be lower due to the relatively hydrophilic environment around the cellulose; iii) the capping step - which functions to reduce the occurrence of truncated products, has been removed. The rationale for the removal of this step being that shorter oligo-dT sequences should still hybridise to the poly-dA region on the RNA-protein fusion molecules.

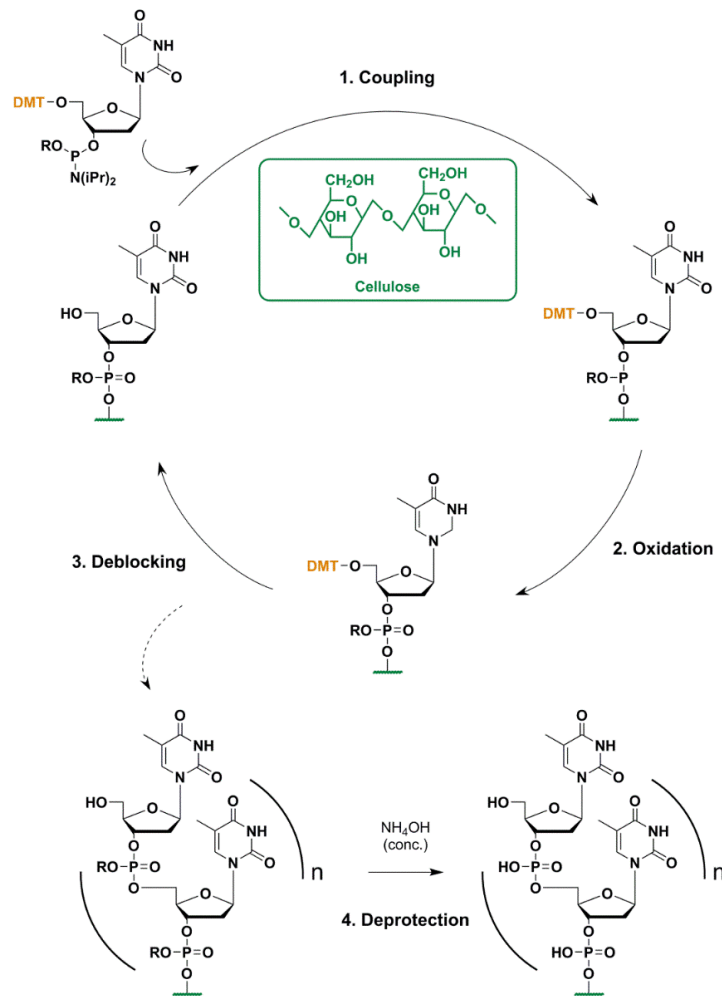


Figure 3.11. Scheme for the automated solid-phase synthesis of oligo-dT cellulose.

Oligo-dT cellulose was synthesised as shown in Figure 3.11 (described in detail in Materials and Methods). Following 25 synthesis cycles and deprotection, the binding capacity of the oligo-dT cellulose (dT25.300s) was evaluated using a poly-dA binding assay with a dA_{15} oligonucleotide (corresponding to the length of the polyA region in the puromycin linker) as a model ligand (Figure 3.12). This assay demonstrated that the newly synthesised oligo-dT cellulose was far superior to the commercially available oligo-dT cellulose (NEB S1408) for the purification of dA_{15} sequences – binding more than 2

nmol.mg⁻¹ of resin. In fact, the commercial resin bound only slightly more dA₁₅ oligonucleotide than the cellulose control in this assay, explaining its ineffectiveness for purification of mRNA-protein fusions. The performance of the synthetic oligo-dT₂₅ cellulose in the dA₁₅ oligonucleotide binding assay indicated that it may efficiently purify RNA-protein fusions from rabbit reticulocyte lysate.

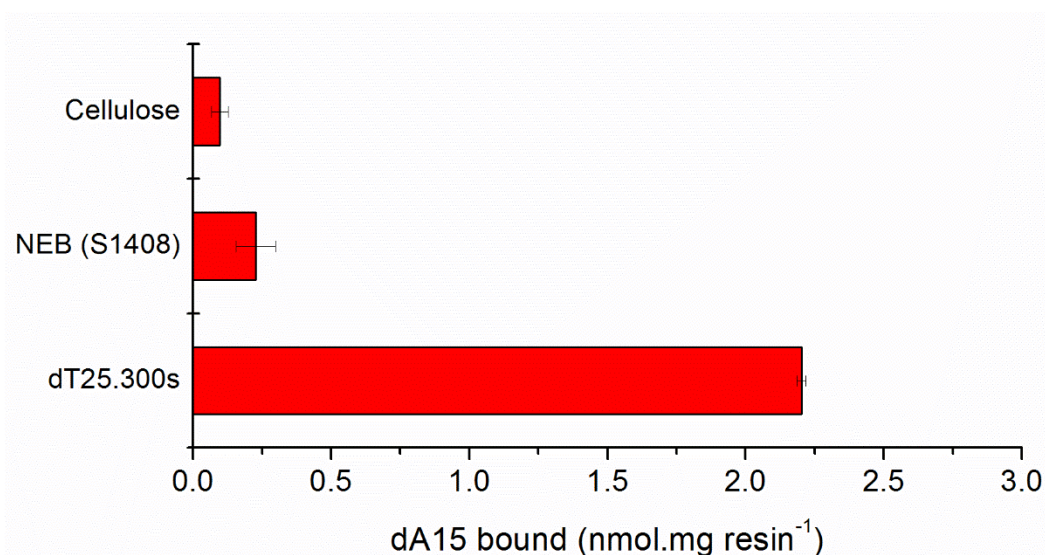


Figure 3.12. Comparative binding analysis of a dA₁₅ oligonucleotide probe to oligo-dT matrices. Each matrix was incubated with dA₁₅ oligonucleotide for 15 minutes with agitation. Bound dA₁₅ was quantified by measuring the UV absorbance of the supernatant at 260 nm before and after incubation with the oligo-dT cellulose resin. Data represents the mean \pm standard error for three experimental replicates.

3.4 Purification of mRNA-protein fusions using oligo-dT cellulose

Having demonstrated the efficacy of the synthesised oligo-dT cellulose for purification of dA₁₅ oligonucleotides, the resin was then tested for its ability to capture and purify mRNA-protein fusions from crude rabbit reticulocyte lysate. Following *in vitro* translation and fusion formation, crude lysates were applied to the oligo-dT cellulose and purification was performed as described in Materials and Methods. mRNA-protein fusions could be specifically and efficiently purified from crude lysates using the synthesised oligo-dT cellulose (Figure 3.13). The majority of mRNA-protein fusions eluted in two fractions,

after which the amount detectable by autoradiography was negligible, consistent with the expected elution behaviour based upon previous reports^{49,150}.

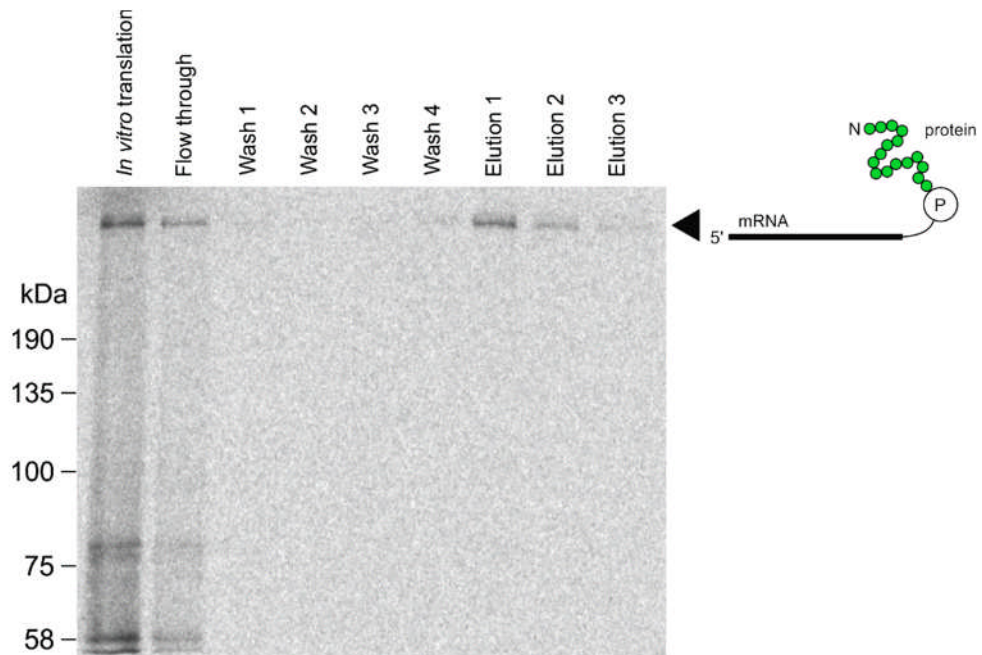


Figure 3.13. SDS-PAGE-autoradiography analysis of RNA-DFPase fusion purification using synthesised oligo-dT cellulose. Upon completion, crude *in vitro* translation reactions were incubated with 5 mg oligo-dT cellulose in oligo-dT binding buffer for 1 h at 4 °C. The resin was washed three times with binding buffer, once with wash buffer, and eluted in twice in 20 mM Tris-HCl pH 8.0. *In vitro* translation reaction was performed at 30 °C for 1 hr using 200 nM DFPase RNA template, and ³⁵S-Met as label. Approximate location of protein markers indicated.

Purified RNA-protein fusions were then reverse transcribed to form cDNA-protein fusions, this serves several purposes. The cDNA allows amplification of selected mRNA-displayed proteins by PCR, and the cDNA-RNA hetero-duplex confers resistance to digestion by nucleases. Additionally, the double stranded nature of the duplex prevents the formation of secondary and tertiary structures in the RNA that could interfere in the subsequent selection step. The *in vitro* translation, oligo-dT purification, and reverse transcription of RNA-DFPase fusions are shown in Figure 3.14. Digestion of the RNA portion of the fusions with RNase A resulted in complete removal of the high molecular mass band corresponding to the RNA-protein fusions, confirming their identity (Figure 3.14a). Genetic information can be readily recovered from cDNA-protein fusions by PCR amplification (Figure 3.14b). Amplification occurs in a reverse transcription dependent manner, indicating the lack of contaminating DNA in the RNA-protein fusion mixture that would generate unwanted background signal in the selection step.

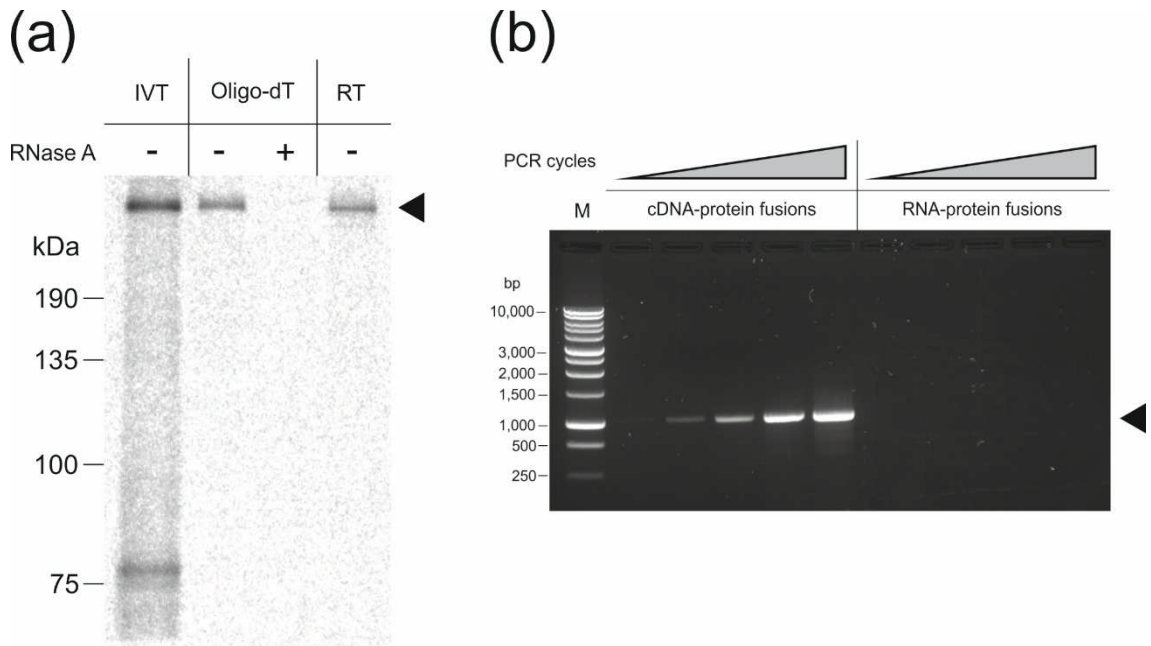


Figure 3.14. Demonstration of RNA-protein linkage and reverse transcription to generate cDNA-protein fusions. **(a)** Digestion with RNase A removes the RNA portion of the fusion resulting in loss of the low-mobility ^{35}S -labelled protein band (RNase A +). Purified RNA-protein fusions are readily reverse transcribed to form cDNA-protein fusions (RT). IVT, *in vitro* translation; oligo-dT, purified RNA-protein fusions; RT, reverse transcription. **(b)** Recovery of genetic information via PCR amplification of the cDNA portion of the fusion (cDNA-protein fusions). Amplification does not occur in the absence of first-strand synthesis via reverse transcription (RNA-protein fusions). Aliquots were taken from PCR reactions after 18 cycles and every 3 cycles thereafter. M; 1 kb DNA ladder.

3.5 Synthesis of tools for the selection of bond forming enzymes by mRNA display

Advances in mRNA display methodology led to the first example of the laboratory selection and evolution of a *de novo* enzyme activity⁴². A study by Seelig *et al.* demonstrated that enzymes that catalyse ligation of a 5'-triphosphorylated RNA oligonucleotide to the terminal 3'-hydroxyl of a second RNA, could be isolated from a library of 4×10^{12} unique proteins based upon the non-catalytic zinc-finger domain of the human retinoid-X-receptor^{42,51}. This was achieved by attaching a substrate molecule to the mRNA-protein fusion complex via reverse transcription, allowing selection for RNA ligase activity by the addition of the second substrate molecule carrying a selectable anchor group⁴². cDNA encoding active protein was subsequently isolated by immobilisation on a solid support due to bond formation between substrates. This general selection strategy can be adapted for any bond-forming reaction of interest (Figure 3.15).

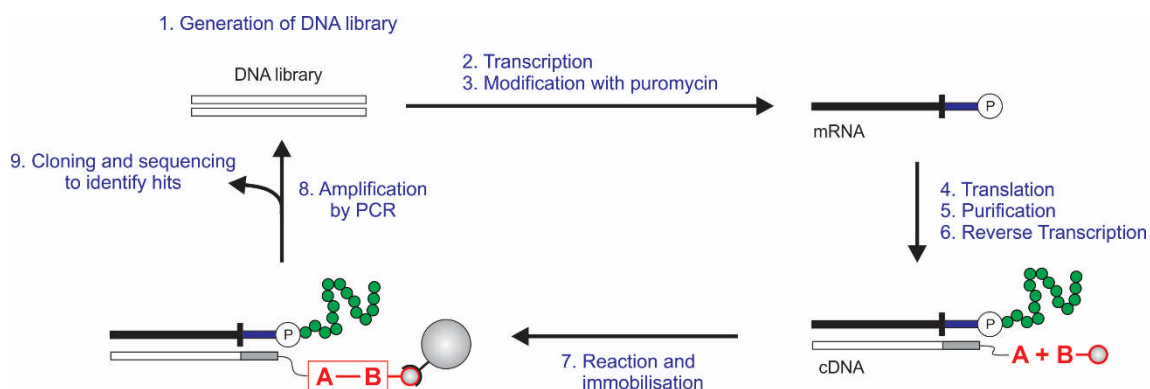


Figure 3.15. A schematic overview of the experimental procedure for a single round of mRNA display selection for bond-forming enzymes. A DNA library is transcribed into RNA, followed by addition of a puromycin to the 3'-end. Subsequent translation to produce mRNA-protein fusions, followed by reverse transcription with a substrate A-modified oligonucleotide results in a cDNA-protein-substrate complex. Addition of substrate B conjugated to a selectable anchor group (e.g. biotin) enables functional selection. Genetic material encoding functionally active proteins is amplified for further rounds of selection or cloning and characterisation.

A number of interesting bond-forming reactions are amenable to selection using mRNA display. This work focusses on development of tools for the selection of enzymes that catalyse a bimolecular Diels-Alder reaction. Pioneered over 80 years ago by Otto Diels and Kurt Alder, the Diels-Alder reaction is a specific type of [4 + 2] cycloaddition in which a diene and dienophile react in a concerted manner to generate a cyclohexene containing product^{64,182}. Characteristically between a conjugated electron-rich diene and an electron-poor dienophile, the reaction occurs via a single pericyclic transition state (Figure 3.16). The prominence of the Diels-Alder reaction in modern organic synthesis has resulted in a great deal of interest in the development of enzymes that catalyse this useful [4 + 2] cycloaddition. Advances in the understanding of biological catalysis over the last 25 years have facilitated the development of an array of engineered biomolecules able to catalyse [4+2] cycloadditions, however none rival the rate enhancements achieved by natural enzymes^{62,81,183,184}.

The specific Diels-Alder reaction in focus is that between 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate and *N,N*-dimethylacrylamide (Figure 3.16, blue and red respectively). This reaction was chosen as it has been the focus of previous work on the development of biomolecular Diels-Alder catalysts, including catalytic antibodies^{183,184} and computational designed enzymes^{62,65,66}. Key requirements for selection for bond-forming reactions using mRNA display are (i) a substrate-modified reverse

transcription primer and (ii) a substrate modified with a selectable anchor group such as biotin.

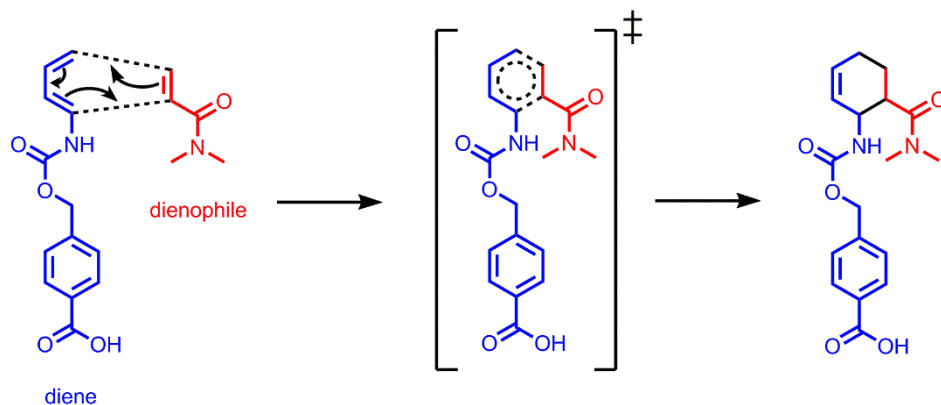


Figure 3.16. The Diels-Alder reaction. A diene (4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate, blue) and dienophile (*N,N*-dimethylacrylamide, red) undergo a pericyclic [4 + 2] cycloaddition to form a chiral cyclohexene ring-containing product. The reaction generates two new carbon-carbon bonds (black), and up to four new stereocenters.

3.5.1 Synthesis of a dienophile-linked reverse transcription primer

primer

A reverse transcription primer was designed containing a template-specific annealing region, a dT₁₈ region and two internal hexaethylene glycol spacer groups (to function as a flexible linker between the cDNA-fusion complex and the dienophile substrate) and a 5'-amino modification for the chemical conjugation of acrylic acid *N*-hydroxysuccinimide ester (Figure 3.17a).

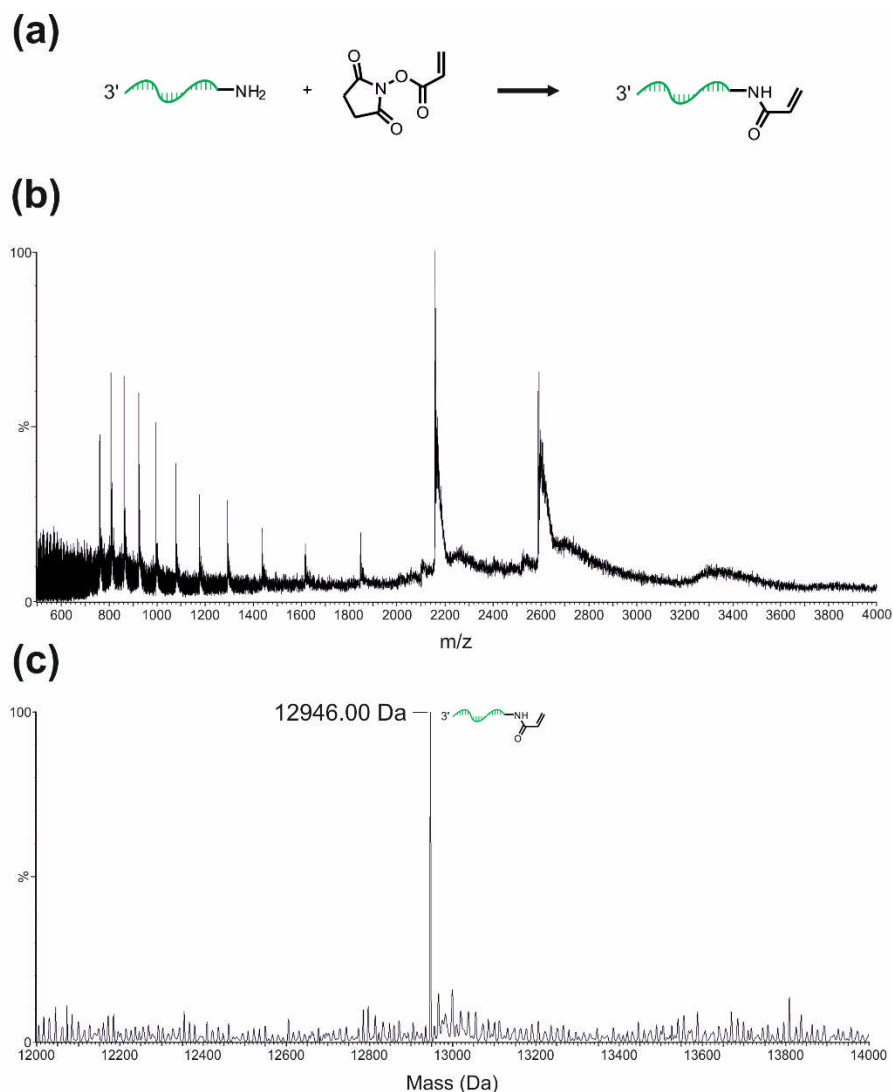


Figure 3.17. Synthesis of a dienophile-linked reverse transcription primer. **(a)** Schematic representation of the conjugation reaction between a 5'-amino oligonucleotide and acrylic acid N-hydroxysuccinimide ester. **(b)** Mass spectrometry analysis of the 5'-acrylamide oligonucleotide synthesised using peptide coupling between a 5'-amino-oligonucleotide and acrylic acid N-hydroxysuccinimide ester. **(c)** Negative ESI-MS m/z spectrum. **(c)** Molecular mass profile derived from **(b)**, indicating a molecular mass of 12946.00 Da which agrees well with the calculated molecular mass of 12,946.55 Da.

The 5'-amino oligonucleotide was conjugated to acrylic acid N-hydroxysuccinimide ester to generate the 5'-acrylamide oligonucleotide as a dienophile for mRNA display selection (Figure 3.17a). Mass spectrometry analysis of the peptide coupling reaction revealed a single species with a molecular mass of 12946.00 Da (Figure 3.17c), which agrees well with the calculated molecular mass of 12,946.55 Da for the 5'-acrylamide oligonucleotide.

3.5.2 Synthesis of a biotinylated diene

The second requirement for selection of Diels-Alderase enzymes using mRNA display is that the diene is modified with a selectable anchor group. The diene – 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate (Figure 3.18, **4**) – was synthesised via Curtius rearrangement between 2,4-pentadienoic acid (Figure 3.18, **1**) and methyl 4-(hydroxymethyl) benzoate (Figure 3.18, **2**), followed by hydrolysis of the methyl ester (Figure 3.18, **3**), using lithium hydroxide. The synthetic strategy was based upon a published method¹⁸⁵, which was modified here into a one-pot procedure using diphenylphosphoryl azide (DPPA) to circumvent isolation of the potentially explosive acyl azide intermediate^{186,187}. Biotin was chosen as a selectable anchor group, as its interaction with streptavidin represents one of the strongest known ligand-receptor interactions ($K_d = 40 \text{ pM}$)¹⁸⁸. Biotin was conjugated to 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate via a biotin derivative with a terminal amino group in place of the carboxylic acid - norbiotinamine (Figure 3.18, **7**). This was obtained from biotin (Figure 3.18, **5**) via modified Curtius rearrangement using DPPA in *t*-BuOH, followed by hydrolysis of the *tert*-butoxycarbamate intermediate (Figure 3.18, **6**)¹⁸⁹. The biotinylated 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate analogue (Figure 3.18, **8**) was obtained via standard peptide coupling of 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate and norbiotinamine, using TBTU.

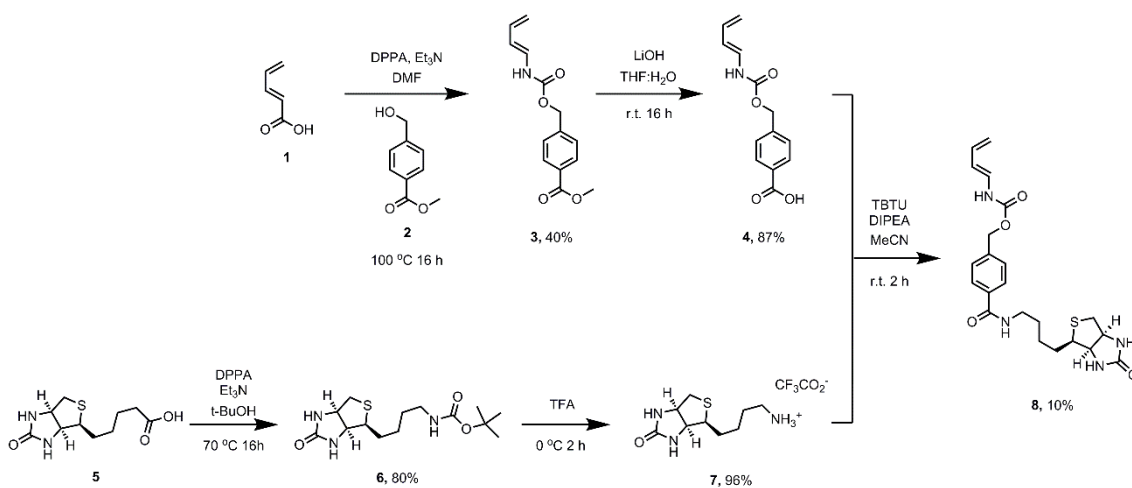


Figure 3.18. Reaction scheme for the synthesis of a biotinylated diene for selection of Diels-Alderase enzymes using mRNA display.

Selections are yet to begin using these substrate analogues in mRNA display experiments, however the synthesis of these molecules should, in theory, enable the selection of Diels-Alderase enzymes from large high-diversity libraries using mRNA display.

3.6 Summary

In this chapter, protocols have been successfully implemented and optimised for the generation of mRNA-protein fusions for *in vitro* selection by mRNA display using the DFPase gene from *L. vulgaris* as a model protein. Key steps in the mRNA display process have been validated and optimised. First, a puromycin-containing oligonucleotide linker was designed and synthesised to enable the covalent attachment of puromycin to the 3'- end of template RNA. The high efficiency of psoralen-mediated photo-crosslinking between the puromycin-oligonucleotide and the RNA template was demonstrated using a truncated 3'-fragment of the RNA template. Here a psoralen-mediated photo-crosslinking approach was chosen due to its perceived ease, versatility, and speed with respect to alternate approaches. This work confirmed the efficacy of psoralen-mediated photo-crosslinking, with an estimated crosslinking efficiency of 75% in 15 minutes of irradiation. A number of alternative linker designs have been reported, for example containing fluorescent groups for detection of RNA-protein fusions¹⁹⁰. Since embarking upon this work, alternative crosslinking chemistries have also been described that may provide faster, more efficient attachment of puromycin than psoralen based strategies. For example, a puromycin oligonucleotide could be covalently linked to mRNA templates via irradiation for only 30 s at 366 nm using a novel 3-cyanovinylcarbazole photocrosslinker¹⁹¹.

Full-length RNA-puromycin conjugates were subsequently generated and used as templates for *in vitro* translation in rabbit reticulocyte lysate, generating covalent RNA-protein fusions. Careful optimisation of mono- and divalent salt concentration in the translation reaction was shown to be crucial for the optimal yield of RNA-protein fusions. Since the original description of mRNA display, protocols have been developed that utilise cell-free translation systems from a variety of sources (e.g. *E.coli* extracts, rabbit reticulocyte lysate, and wheat germ extract). Choice of *in vitro* translation medium depends to some extent on the aims of the selection. For example, cell-free protein synthesis using recombinant elements of the *E.coli* translational machinery (PURE)¹⁹²

allows the incorporation of non-canonical amino acids for exploration of abiological chemical space¹⁹³.

The use of *in vitro* translation systems other than rabbit reticulocyte lysate may improve RNA-protein fusion yields. Indeed, a recent study found a template independent increase in RNA-protein fusion efficiency using wheat germ extracts ($\leq 95\%$ of input template) compared to rabbit reticulocyte lysate ($\leq 65\%$ of input template)¹⁹⁴. This observation might be partially explained by the greater concentration of ribosomes in wheat germ extract (approximately 4 μM or 2.5×10^{15} per mL) compared with rabbit reticulocyte lysate (approximately 0.2 μM or 1.2×10^{14} per mL)¹⁷³. For “single-turnover” display techniques such as mRNA display and ribosome display, in which ribosomal turnover is unlikely due to the absence of stop codons, ribosome concentrations may be proportional to fusion yields. Despite this, increasing the concentration of ribosomes above 0.33 μM reduced selection yields in ribosome display experiments when using a reconstituted translation system based on *E.coli* translation machinery¹⁹⁵.

In the absence of a suitable commercial alternative, high-capacity oligo-dT cellulose was synthesised and demonstrated to efficiently purify nucleic acids containing dA₁₅ sequences, including RNA-protein fusions. Isolation of RNA-protein fusions from crude translation mixtures is desirable to reduce background interference from components of the lysate but also from free proteins that may be present after translation. High levels of haemoglobin in rabbit reticulocyte lysate impede His-tag purification, due to co-purification when using immobilised metal affinity chromatography methods^{179,180}. Typically, mRNA display protocols get around this limitation by utilising oligo-dT cellulose to target the poly-dA region on the puromycin linker^{42,51,125,129,173}.

Overall, the results presented in this chapter demonstrate that mRNA-protein fusions can be readily generated using the comparatively large 36 kDa DFPase enzyme from *L.vulgaris*. Typical selections to date have been performed on peptides and small proteins (<15 kDa), with the largest previously reported mRNA-protein fusions based on the 29 kDa indole-3-glycerol phosphate synthase from *S.solfataricus*¹⁰². Proteins of this size are more likely than short peptides to fold into functional domains and adopt complex native conformations or structures such as those typically essential for catalysis. These results demonstrate the applicability of mRNA display for the study of larger proteins that perform complex functions such as receptors and enzymes. Additionally, the methodology described here is broadly applicable to accommodate almost any protein scaffold, provided it can be expressed in an *in vitro* translation system, allowing the potential investigation of multiple protein architectures for complex functionalities.

Finally, a strategy has been devised for the selection of enzymes that catalyse the bimolecular Diels-Alder reaction between 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate and *N,N*-dimethylacrylamide. A 5'-acrylamide modified reverse transcription primer was generated, allowing the dienophile to be linked to the mRNA-protein fusion complex by reverse transcription. In addition, 4-carboxybenzyl *trans*-1,3-butadiene-1-carbamate was synthesised and conjugated to biotin as a selectable anchor group. These tools will allow future *in vitro* selection for enzymes that catalyse the Diels-Alder reaction via specific immobilisation of sequences on a streptavidin-derivatised solid support.

4 Selection for orthogonal ligand-receptor pair using mRNA-display

Peroxisomes are a family of ubiquitous single-membrane bound organelles that contain enzymes involved in a variety of metabolic reactions, including several aspects of energy metabolism and the detoxification of reactive oxygen species¹⁹⁶. Peroxisomes possess a number of notable characteristics that separate them from other cellular organelles such as mitochondria and chloroplasts. Unusually, they can form *de novo* from the endoplasmic reticulum (ER)^{197,198}, in addition they do not contain DNA or ribosomes and thus lack the ability to encode their own proteins. As a result, nuclear-encoded peroxisomal matrix proteins are translated in the cytosol and imported into the peroxisome post-translationally¹⁹⁹. A remarkable feature of this import process is that proteins can be imported in fully-folded, oligomeric, and cofactor-bound states²⁰⁰⁻²⁰². The importance of peroxisomes to cellular function is demonstrated by the large group of human genetic diseases in which peroxisome biogenesis or metabolic function is impaired²⁰³.

These unique features have led to interest in the use of peroxisomes in a variety of applications that span the fields of biotechnology and synthetic biology. Peroxisomes have been proposed as storage vessels for the accumulation of heterologously expressed proteins. Indeed, the peroxisomes of methylotrophic yeasts such as *Pichia pastoris* can expand to approximately 80% of the total volume of the cell²⁰⁴, suggesting that high protein capacity is achievable. Furthermore, the expected absence of protein-modifying enzymes inside peroxisomes, for example those responsible for phosphorylation, glycosylation, and proteolysis, may obviate undesired modification of proteins that may occur in the cytosol or in the ER. Peroxisomes are also particularly attractive compartments for the storage of toxic proteins, as the peroxisomal membrane provides a barrier that prevents leakage out of the organelle²⁰⁵.

The growing number and variety of metabolic pathways identified to be hosted within peroxisomes suggests that they may also be suitable host organelles for localisation of novel engineered biosynthetic pathways, particularly those derived from intermediates of the β -oxidation pathway – the catabolic breakdown of fatty acids into acetyl-CoA – commonly found in peroxisomes^{206,207}. Compartmentalisation strategies can minimise diffusion of pathway intermediates, increase local enzyme concentrations, and provide a level of spatial control over pathway intermediates and competing pathways for optimal product formation²⁰⁸. A central tenet to the successful exploitation of peroxisomes for

these biotechnological applications is the efficient and specific targeting of heterologously expressed proteins from the cytosol to the peroxisomal lumen. However, there is currently a lack of mechanistic understanding of the processes that dictate peroxisomal import.

The vast majority of proteins destined for peroxisomal import possess a C-terminal peroxisomal targeting signal type 1 sequence (PTS1) that directs import via the cytosolic receptor, PEX5^{209,210}. The import of PTS1 containing proteins into peroxisomes is outlined in Figure 4.1. The process begins in the cytosol, where the PTS1-containing cargo protein, destined for import, is recognised by the PEX5 receptor protein, which directs docking at the translocon on the peroxisomal membrane. Translocation of receptor and cargo into the lumen then occurs via a mechanism that is yet to be fully elucidated, followed by cargo unloading and receptor recycling²¹¹.

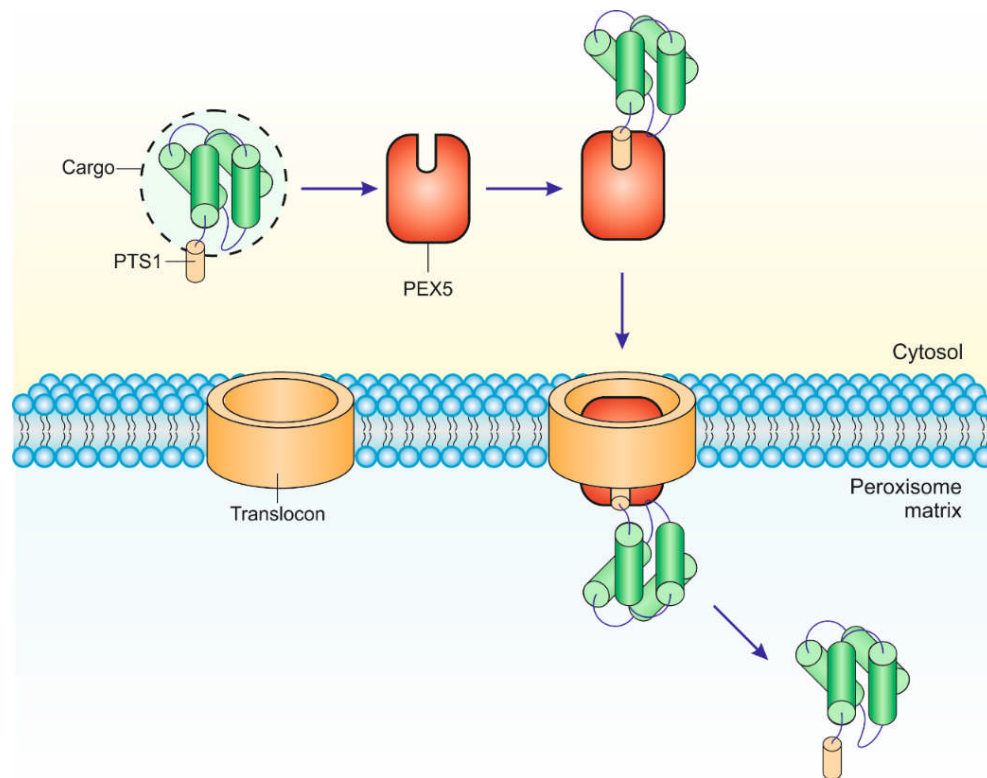


Figure 4.1. Schematic representation of PEX5 mediated import of folded proteins across the peroxisomal membrane. Cargo binding via the C-terminal PTS1 signal sequence triggers docking at the translocon. Translocation occurs via an unknown mechanism, followed by cargo unloading into the peroxisomal matrix.

The PTS1 was initially identified as the C-terminal tripeptide motif SKL²¹², however a range of tripeptides have since been shown to function as PTS1s *in vivo*. The general

consensus PTS1 sequence is [S/A/C]-[K/R/H]-[L/M]²¹³, although variants have been identified which only possess two of the three consensus residues²¹⁴. The PEX5 receptor that recognises the PTS1 signal peptide in the cytosol is a modular protein composed of a number of domains with distinct functionality. The N-terminal domain plays a role in a number of functions including docking at the peroxisomal membrane and receptor recycling. The C-terminal region of the receptor contains the domain responsible for specific recognition of the PTS1 signal sequence. X-ray crystallography studies of the C-terminal domain of the *Homo sapiens* and *Trypanosoma brucei* PEX5 receptor bound to model PTS1 peptides revealed that the binding domain is a series of structurally conserved tetratricopeptide repeat (TPR) motifs^{215,216} (Figure 4.2). These are common protein-protein interaction motifs formed from repeats of 34 amino acids, comprising two α -helices separated by a turn²¹⁷. The structure of the C-terminal region of PEX5 consists of seven TPR motifs split into two clusters separated by a hinge region (Figure 4.2). TPRs 1-3 form the first cluster (Figure 4.2, green), and TPRs 5-7 form the second cluster (Figure 4.2, cyan), separated by the putative TPR 4 that forms a continuous α -helix, thought to act as a 'hinge' region²¹⁵.

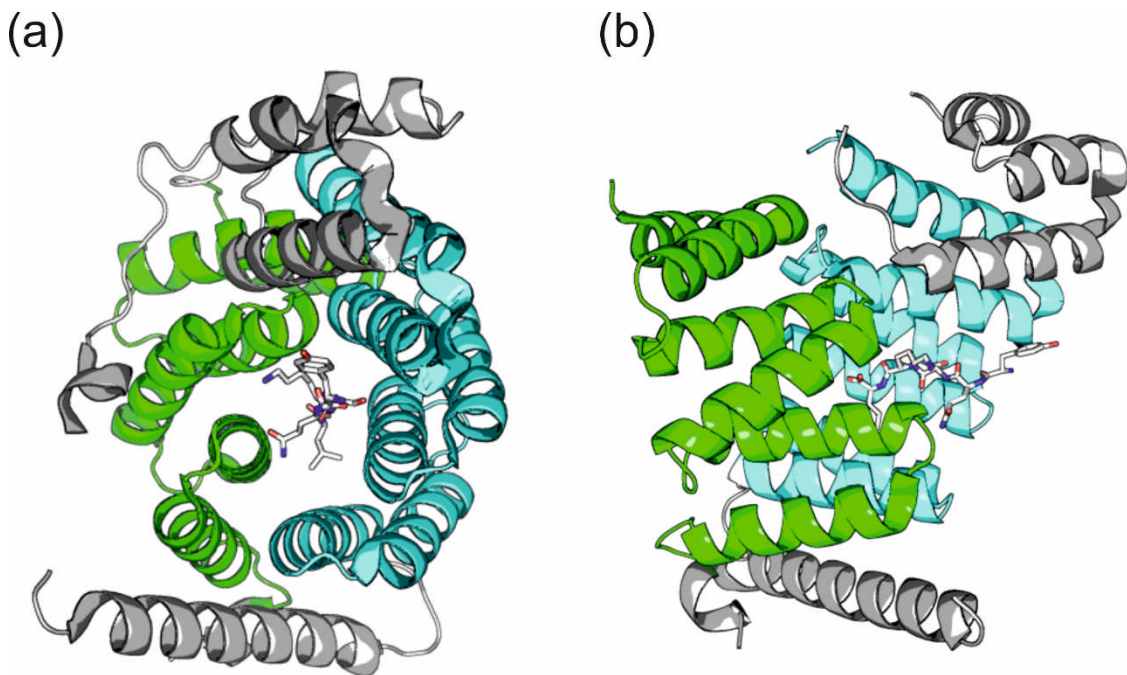


Figure 4.2. X-ray crystal structure of the *Homo sapiens* PEX5 receptor-PTS1 complex (PDB: 1FCH). **(a)** A view down the backbone of the canonical PTS1-containing peptide YQSKL (coloured by elemental composition). TPRs 1-3 are shown in green, TPRs 5-7 in cyan, all other regions are shown in grey. **(b)** A view rotated 90° from **(a)**.

Heterologously expressed proteins, and indeed whole metabolic pathways, can be targeted to the peroxisomal lumen simply by the incorporation of a canonical PTS1 motif at the C-terminus^{205,207}. A shortcoming of this strategy, however, is that accumulation of heterologous products may have a detrimental effect on normal peroxisomal function, which can have severe implications on overall viability of the host organism. One solution to this problem is the generation of synthetic organelles, allowing spatial (and temporal – in the case of inducible pathways and gene circuits) segregation of engineered and endogenous cellular processes. Conceptually, such synthetic organelles have two basic requirements; (i) the presence of a physical barrier separating the contents from the cellular environment and (ii) the ability to specifically and efficiently target heterologous proteins to the inside of the organelle. The natural diversity of peroxisomal function and the dynamic nature of their size, shape, and abundance make them intriguing candidates for the development of semi-synthetic organelles.

Preventing the accumulation of native peroxisomal proteins and enzymes in a synthetic organelle requires orthogonal import machinery, which can function completely separately to the natural PEX5-PTS1 system. This requires a novel peptide motif that does not bind to the wild-type PEX5 receptor, and a novel PEX5 receptor that binds to this novel peptide motif, but not to the canonical PTS1 sequence. A number of C-terminal tripeptides have already been identified that do not possess discernible binding to the wild-type PEX5 receptor and therefore are not imported by the natural peroxisomal import machinery²¹⁸. The aim of the work described in this chapter was to engineer a mutant PEX5 peroxisomal import receptor that binds to one of these non-canonical PTS1 sequences and thus generate a completely orthogonal import recognition system. This would provide a tool for further study of the peroxisomal import via PEX5 and serve as a starting point for the development of peroxisomes as semi-synthetic organelles *in vivo*.

There are more than fifty known plant PTS1 signals²¹⁹, with the general C-terminal tripeptide motif of small-basic-hydrophobic, making the identification of the precise molecular determinants for recognition non-trivial. Indeed, many of the residues in the PTS1 binding pocket appear to interact specifically with the peptide backbone or the C-terminal carboxylate, rather than amino acid side chains²¹⁵. This lack of mechanistic understanding, in combination with the likely dynamic nature of the PEX5 receptor, obviates the straightforward rational redesign of the PEX5 binding pocket. The sheer number of potential combinations of mutations makes the problem amenable to a selection approach such as mRNA display.

As discussed above, the PTS1 signal peptide occurs at the C-terminus of proteins destined for peroxisomal import. As the phenotype-genotype link in mRNA display is formed directly between puromycin and the C-terminus of the nascent peptide chain, the interaction cannot be studied using RNA-PTS1 peptide fusions. Instead, the C-terminal PTS1 binding domain of the soluble PEX5 receptor was chosen for RNA-protein fusion formation. The experimental strategy for selection of orthogonal PEX5:PTS1 binding pairs is shown in Figure 4.3.

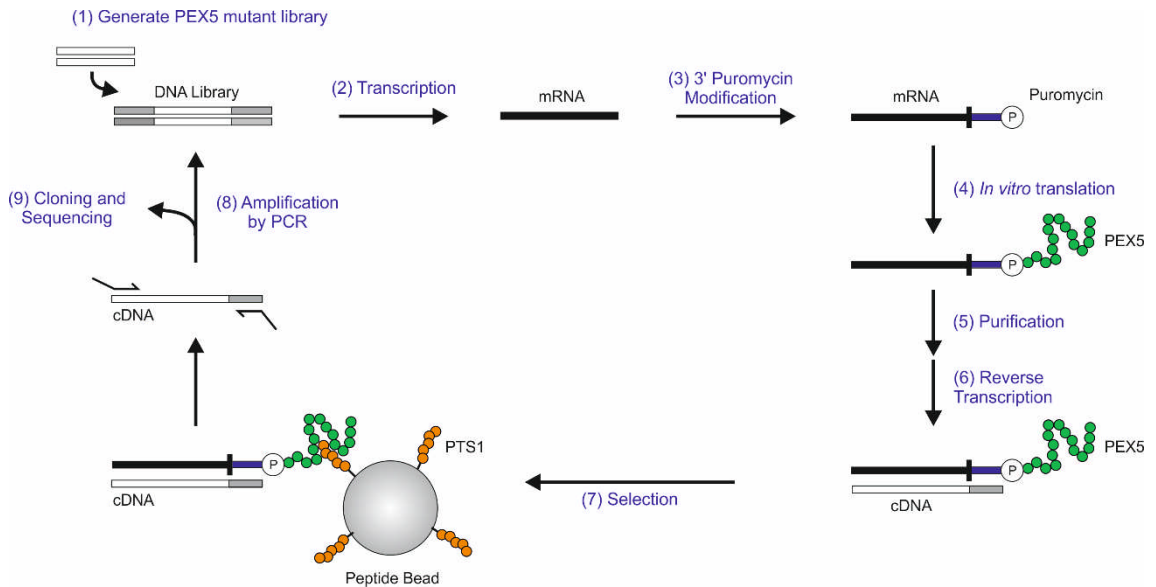


Figure 4.3. A schematic overview of the experimental procedure for selection of orthogonal PEX5:PTS1 binding pairs by mRNA display. A DNA library is transcribed into RNA, followed by addition of a puromycin to the 3'-end. Subsequent translation to produce mRNA-PEX5 fusions, followed by reverse transcription enables functional selection. In the selection step (7), mRNA-PEX5 fusions are incubated with beads displaying the unnatural PTS1 peptide. Genetic material encoding functionally active proteins is amplified for further rounds of selection or cloning and characterisation.

4.1 PEX5 RNA-protein fusion generation

In order to maximise the number of sequences interrogated in the selection step of mRNA display, the *in vitro* translation reaction was optimised for RNA-protein fusion formation using the RNA transcript encoding the PEX5 receptor. An N-terminal truncation of the *Arabidopsis thaliana* PEX5 receptor protein comprising amino acid residues 445-728 (equivalent to the *Homo sapiens* construct used to generate the three-dimensional structure of PEX5 in Figure 4.2 without the N-terminal loop region) was chosen for mRNA-display. This was based on the observation that mRNA display generally works

well when displayed proteins are less than 300 residues in length, with larger proteins typically showing lower fusion formation efficiencies⁴⁹. Previous work in our laboratory demonstrated that the PEX5 445-728 N-terminal truncation possessed binding affinity to a canonical PTS1 peptide (YQSKL) comparable to the full-length PEX5²¹⁸. To increase the efficiency of *in vitro* translation and PEX5 RNA-protein fusion formation, the concentration of mono- and divalent cations were varied in the translation reaction (Figure 4.4).

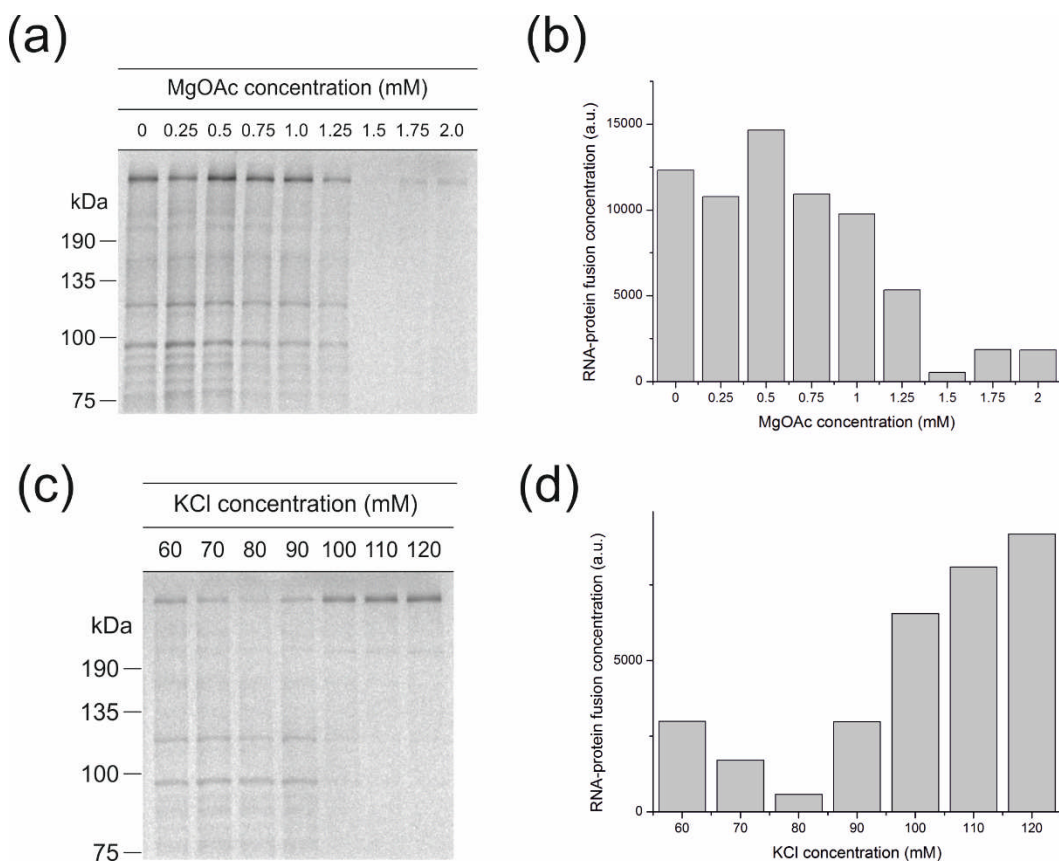


Figure 4.4. Optimisation of monovalent salt concentration in the *in vitro* translation reaction for increased RNA-PEX5 fusion formation. **(a)** Prior to translation, additional magnesium acetate (MgOAc) was added corresponding to 0 - 2.0 mM final concentration. Following translation, mixtures were assayed by SDS-PAGE followed by autoradiography. **(b)** Data expressed as radioactivity present in the RNA-protein fusion band as determined by densitometry. Addition of 0.5 mM MgOAc produces an increase in fusion formation relative to the control. **(c)** Effect of additional potassium chloride (KCl) corresponding to 60 - 120 mM final concentration on fusion formation. **(d)** Data expressed as radioactivity present in the RNA-protein fusion band as determined by densitometry. Addition of 120 mM or higher KCl results in optimum RNA-protein fusion formation. All translation reactions were performed at 30 °C for 1 hr using 200 nM PEX5 RNA, and ³⁵S-Met as label. Approximate location of protein markers indicated. Densitometry was performed using ImageJ software¹⁷⁶.

The optimum MgOAc concentration of 0.5 mM (Figure 4.4a) in combination with a KCl concentration of 120 mM (Figure 4.4c) in the translation reaction gave the highest yield

of PEX5 RNA-protein fusions. The contrast between the optimum concentrations determined here for PEX5 and those for DFPase in Chapter 3 (0.25 mM MgOAc, 90 mM KCl) demonstrates the importance of this optimisation procedure, the same conditions likely resulting in negligible RNA-DFPase fusion formation (Chapter 3, Figure 3.10). An interesting observation from the SDS-PAGE autoradiography analysis of the RNA-PEX5 fusions is that certain translation conditions result in bands of intermediate size between the full-length RNA-protein fusion and that expected of the free protein. The presence of these species may be attributable to partial degradation of the RNA-protein fusions in the *in vitro* translation reaction – a highly undesirable process when preparing libraries for selection experiments. The PEX5 RNA-protein fusions were readily purified using oligo-dT cellulose (Figure 4.5) as described in Materials and Methods, demonstrating the versatility of the synthesised dT-cellulose matrix for the general purification of RNA-protein fusions.

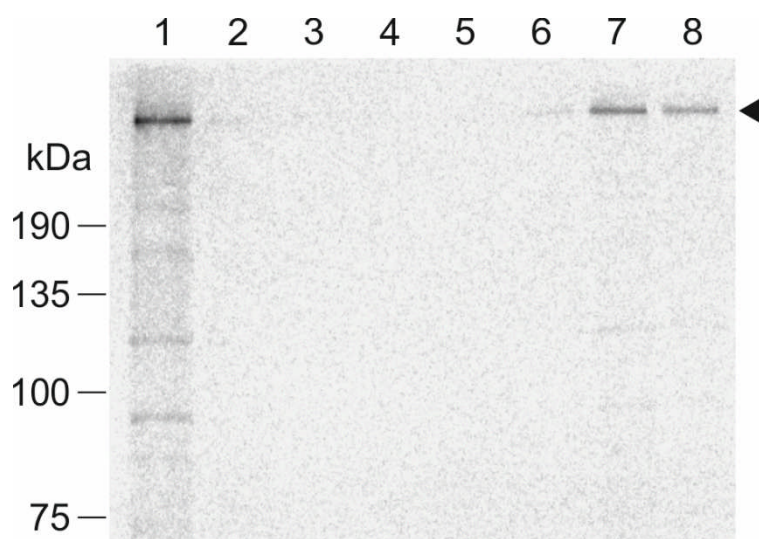


Figure 4.5. SDS-PAGE-autoradiography analysis of RNA-PEX5 fusion purification using synthesised oligo-dT cellulose. Upon completion, crude *in vitro* translation reactions were incubated with 5 mg oligo-dT cellulose in oligo-dT binding buffer for 1 h at 4 °C. The resin was washed three times with binding buffer, once with wash buffer, and eluted in twice in 20 mM Tris-HCl pH 8.0. *In vitro* translation reaction was performed at 30 °C for 1 hr using 200 nM PEX5 RNA template, and ³⁵S-Met as label. Lane 1, crude *in vitro* translation; lane 2, column flow-through; lanes 3-6, washes 1-4; lanes 7 and 8; elution fractions 1 and 2. Approximate location of protein markers indicated.

Prior to performing selection against unnatural peptides for an orthogonal interaction, the ability to distinguish between binding and non-binding peptides in the selection step was

investigated using the cognate PEX5-PTS1 interaction. Wild-type RNA-PEX5 fusions were generated and purified, followed by incubation with positive (YQSKL) and negative control (YQSEV) peptides immobilised to agarose beads. The canonical PTS1, YQSKL was chosen as a positive control, and YQSEV was selected as a negative control based on the lack of detectable binding of this peptide to *A. thaliana* PEX5 using a fluorescence anisotropy based assay²¹⁸. Following washing, the proportion of RNA-PEX5 fusions retained on the solid support was quantified by scintillation counting (Figure 4.6). A notable difference was observed in the retention of RNA-PEX5 fusions on the immobilised peptides. The canonical PTS1 pentapeptide sequence YQSKL retained approximately 23% of the RNA-PEX5 fusions, compared to less than 2% of fusions when incubated with the YQSEV peptide. This indicates that the PEX5 N-terminal truncation construct retains its ability to recognise and bind cognate PTS1 sequences when displayed on its encoding mRNA.

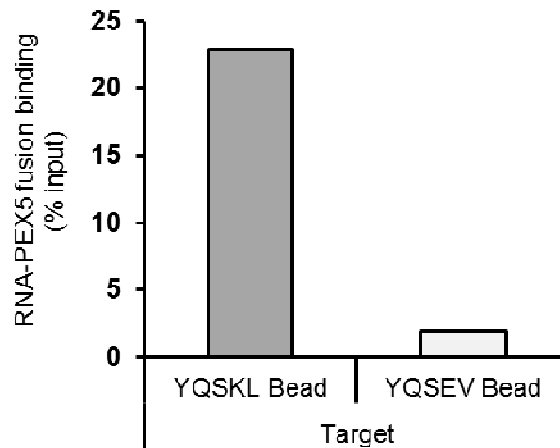


Figure 4.6. Comparative binding of RNA-PEX5 fusions on immobilised model peptide targets. The canonical PTS1 sequence YQSKL was used as a positive control, and the non-binding peptide YQSEV as a negative. Ratios of bound to unbound RNA-PEX5 fusions were determined by scintillation counting before and after incubation with the target peptide.

4.2 Construction of the PEX5 mutant library

With the aim of selecting for an orthogonal PEX5-PTS1 interaction, amino acid residues in the binding pocket were selected for randomisation based on their proximity to the bound PTS1 peptide in a crystal structure of the human PEX5 homologue (Figure 4.7, red)²¹⁵. A total of twelve positions were chosen for randomisation – resulting in a theoretical library diversity of approximately 4×10^{15} sequences. Whilst this is more than

an order of magnitude larger than that accessible to mRNA display at its practical limits, such high theoretical diversity reduces redundancy in the initial library. With little mechanistic insight into binding specificity, sparse yet wide sampling of sequence space may also further the understanding of the roles of key residues in the interaction.

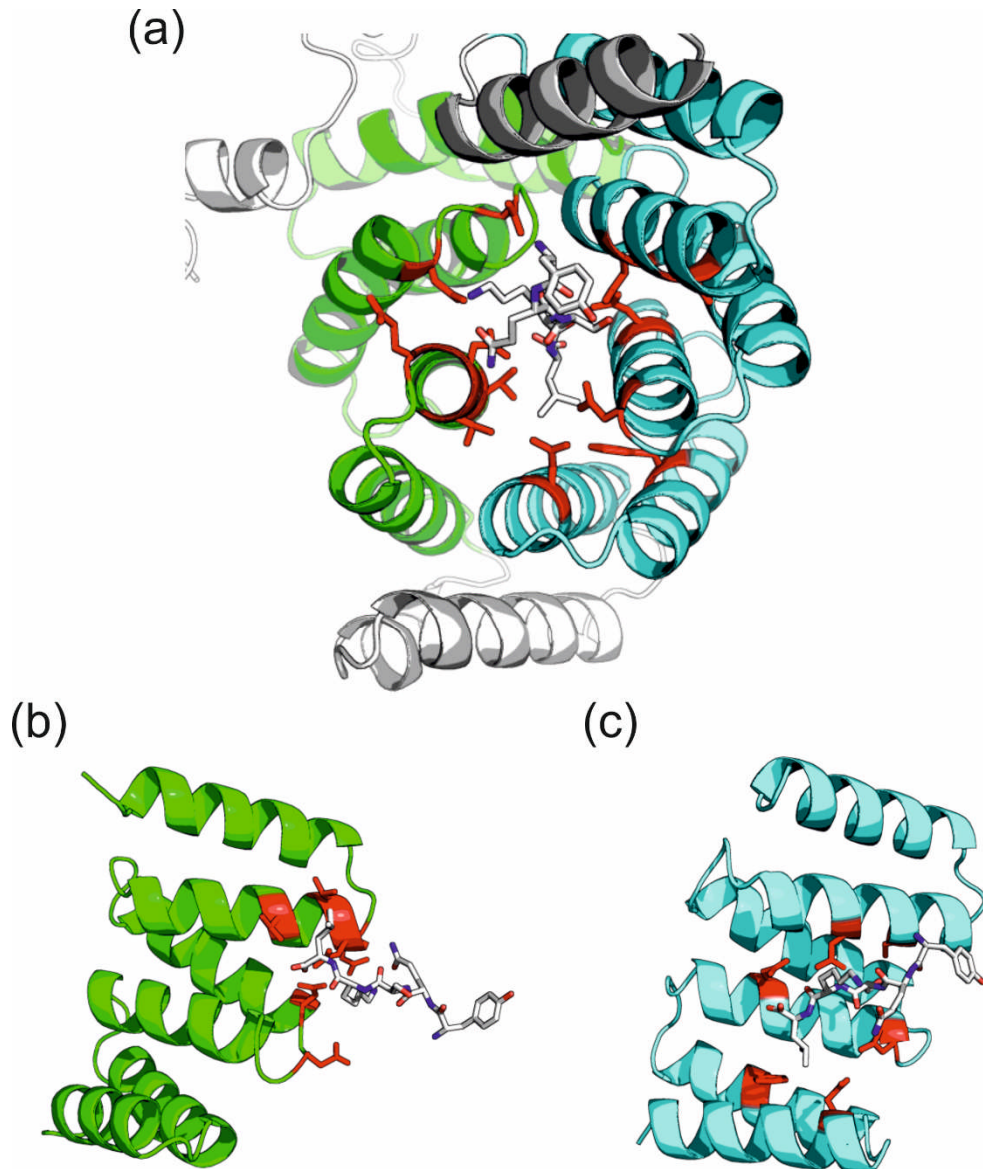


Figure 4.7. Relationship of mutated residues to the model PTS1 peptide YQSKL. **(a)** A view down the backbone of the canonical PTS1-containing peptide. **(b)** View of the interaction of TPRs 1–3 (green) with the peptide ligand. **(c)** View of the interaction of TPRs 5–7 (cyan) with the peptide ligand.

A variety of methods exist for the production of large genetic libraries for directed evolution²²⁰⁻²²³. One approach that has been successfully used to generate libraries of particularly high diversity involves the assembly of genes from chemically synthesised oligonucleotides containing randomised codons¹⁰². There are a number of considerations when assembling long DNA constructs in this manner. Deletions that occur in synthetic oligonucleotides due to imperfect coupling and capping efficiencies during solid-phase DNA synthesis cause frameshifts²²⁴. In addition, stop codons encoded in randomised regions cause premature termination of protein synthesis. Careful consideration of codon composition in randomised regions can help to reduce the appearance of stop codons, for example mixed codons such as NNK and NNS (where K = G/T and S = G/C) encode all 20 amino acids but only a single stop codon. As the library would be translated in rabbit reticulocyte lysate, randomisation was performed using NNS codons, which are more frequently used than NNK in *Oryctolagus cuniculus*¹⁸¹. The use of NNS codons also increases the theoretical proportion of sequences that are free from stop codons from 56% to 68% over NNN codons.

Due to the TPR domain architecture and the orientation of the PTS1 binding pocket, the residues chosen for randomisation spanned the entire PEX5 sequence (Figure 4.8). To circumvent this, the sequence was split into six cassettes, with randomised residues located at the beginning and/or the end to allow incorporation using PCR with randomised primers. Primers contained flanking restriction sites for the type IIS restriction enzyme BsaI, which cuts outside of its recognition sequence. This allows customisation of sticky ends, and thus scarless assembly, and makes it possible to ligate multiple fragments in a single reaction, streamlining assembly protocols. Indeed, the cassettes were designed to possess unique sticky-ends to facilitate one-pot ligation into the full-length gene. However, the one-pot assembly protocol was not efficient enough to assemble the DNA library on the scale required (data not shown). Therefore, following amplification of individual cassettes, the randomised segments were assembled sequentially using the strategy outlined in Figure 4.8. Oligonucleotide sequences for library assembly are listed in the Appendix.

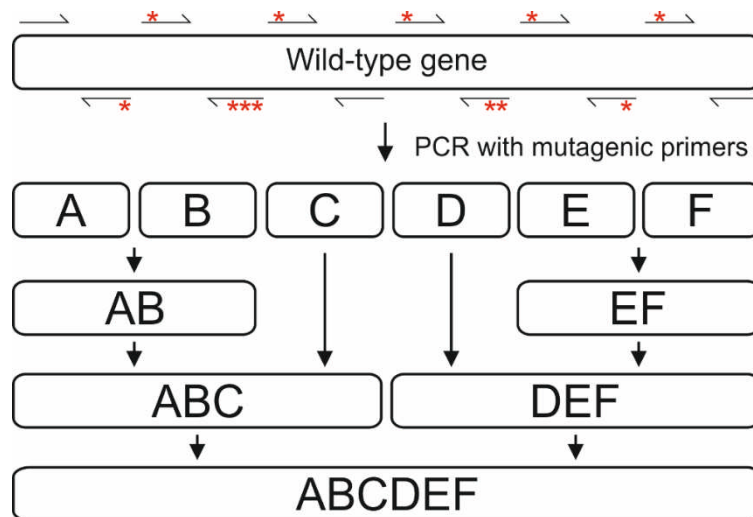


Figure 4.8. Schematic of the PEX5* library assembly strategy. Library segments were amplified by PCR using primers containing NNS codons (*). Segments were sequentially assembled by seamless restriction digestion and ligation to yield the full-length library (PEX5*).

Amplification of each cassette resulted in fragments of the expected size (Figure 4.9a). Individual fragments were digested, purified, and ligated in a step-wise manner to yield the intermediate cassettes and the full-length 900 bp PEX5* library (Figure 4.9b). A sample of the full-length linear DNA library was cloned and ten colonies were picked at random and sequenced to determine the initial library quality. All clones sequenced were identical to the wild-type PEX5 gene except at the selected codons, where successful randomisation was observed (Appendix).

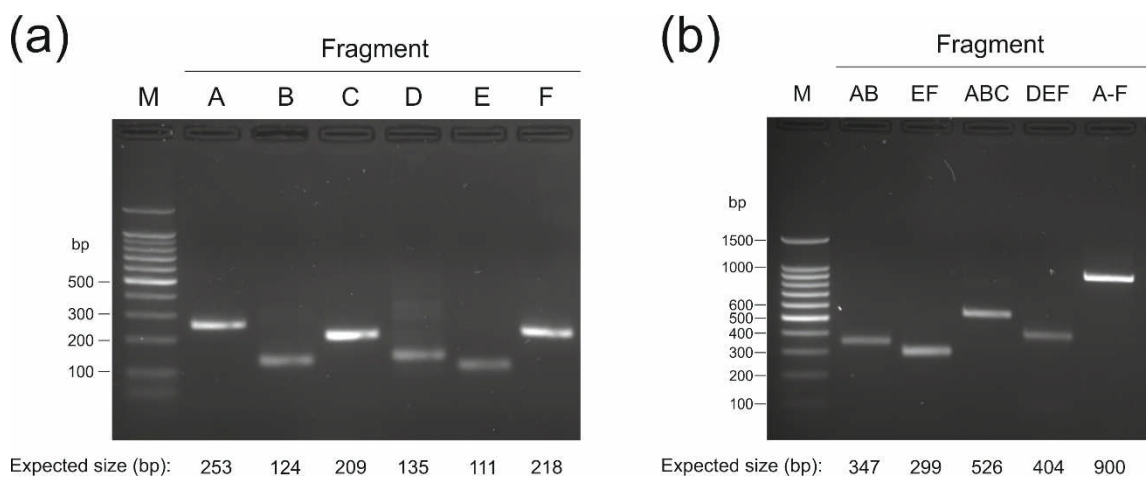


Figure 4.9. Assembly of the 900 base-pair linear PEX5* DNA library. **(a)** Amplification of library fragments A-F using PCR with mutagenic primers. **(b)** Sequential assembly of fragments into intermediate (AB, EF, ABC, DEF) and full length (A-F) constructs.

A summary of the observed nucleotide frequencies at each position for all randomised codons is shown in Table 4.1. The observed nucleotide frequencies in the library were within 10% of the intended frequency in every case, although at codon positions 1 and 2 an overall under-representation of A and T nucleotides was observed. This could be due to an inherent bias introduced during oligonucleotide synthesis given that automated mixing of phosphoramidite monomers can result in bias due to differences in delivery time to the solid support and relative coupling efficiencies during synthesis. The remainder of the linear PEX5* library was modified by PCR to append sequences for mRNA display, yielding 4×10^{13} molecules of template DNA.

Codon Position	% T	% C	% A	% G
1	15	38	23	25
2	18	35	15	33
3	0	54	0	46

Table 4.1. Percentage of each nucleotide at each position of the randomised codons in the nascent PEX5* library prior to selection. Total number of codons sequenced = 120.

4.3 Selection for orthogonal PEX5-PTS1 interactions

With the PEX5* library assembled and determined to contain the correct distribution of mutated residues, selection for orthogonal PEX5-PTS1 interaction was performed as outlined in Figure 4.3. Three candidate orthogonal PTS1 peptides – YQSEV, YQSFY, and YQSYY – were selected based on the previous observation that they do not possess significant binding activity to the wild-type PEX5 receptor when assayed using fluorescence anisotropy²¹⁸. These peptides only differ in the residues at the -1 and -2 positions relative to the C-terminus of the canonical PTS1 sequence. This is consistent with a greater thermodynamic cost to binding for the PEX5-PTS1 interaction when mutations are made at these positions, as observed by Gatto *et al.*²²⁵.

The PEX5* library was transcribed *in vitro* and a total of 1.7×10^{13} RNA molecules were modified with puromycin as described in Materials and Methods. The resulting puromycin modified RNA was used as template in a large-scale *in vitro* translation reaction. Oligo-dT purification from the rabbit reticulocyte lysate yielded 2.5×10^{12} RNA-protein fusions, which were reverse transcribed to generate cDNA fusions. SDS-PAGE analysis of the *in*

in vitro translation of the PEX5* library and purification of the resulting RNA-protein fusions revealed that the fusions migrated as two major species (Figure 4.10). This may be attributable to a decrease in protein stability experienced by a large proportion of the library due to the simultaneous randomisation of twelve residues^{226,227}.

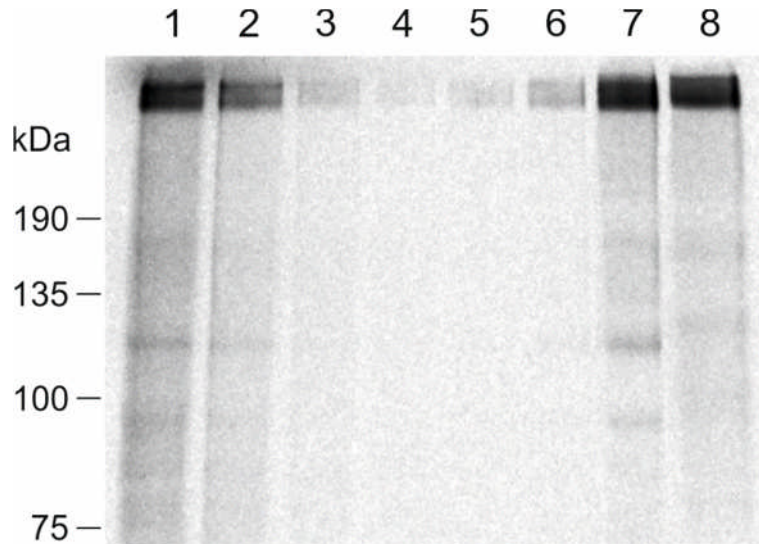


Figure 4.10. SDS-PAGE-autoradiography analysis of the purification of RNA-PEX5* library fusions at round 0 by oligo-dT cellulose chromatography. Lane 1, crude *in vitro* translation; lane 2, column flow-through; lanes 3-6, washes 1-4; lanes 7 and 8; elution fractions 1 and 2. Approximate location of protein markers indicated.

In the selection step, cDNA-PEX5 fusions were preselected against unmodified beads, split into three sub-libraries and incubated with the corresponding peptide (YQSEV, YQSYY, and YQSFY) immobilised on streptavidin-agarose beads via an N-terminal PEG-biotin moiety. Following selection, PEX5* cDNA was amplified by PCR to regenerate the libraries. In all subsequent rounds, cDNA-PEX5 fusions were preselected against YQSKL beads to remove any sequences that bind to the wild-type PTS1 sequence. The mRNA display selection procedure was performed a total of four times against the three peptide targets, after which a subset of each library was cloned and sequenced. In total, ten randomly selected clones from libraries after four rounds of selection were sequenced, the resulting amino acid distributions at each randomised position are shown in Figure 4.11. The libraries were found to have converged strongly onto a family of common sequences that appeared to be independent of the target peptide.

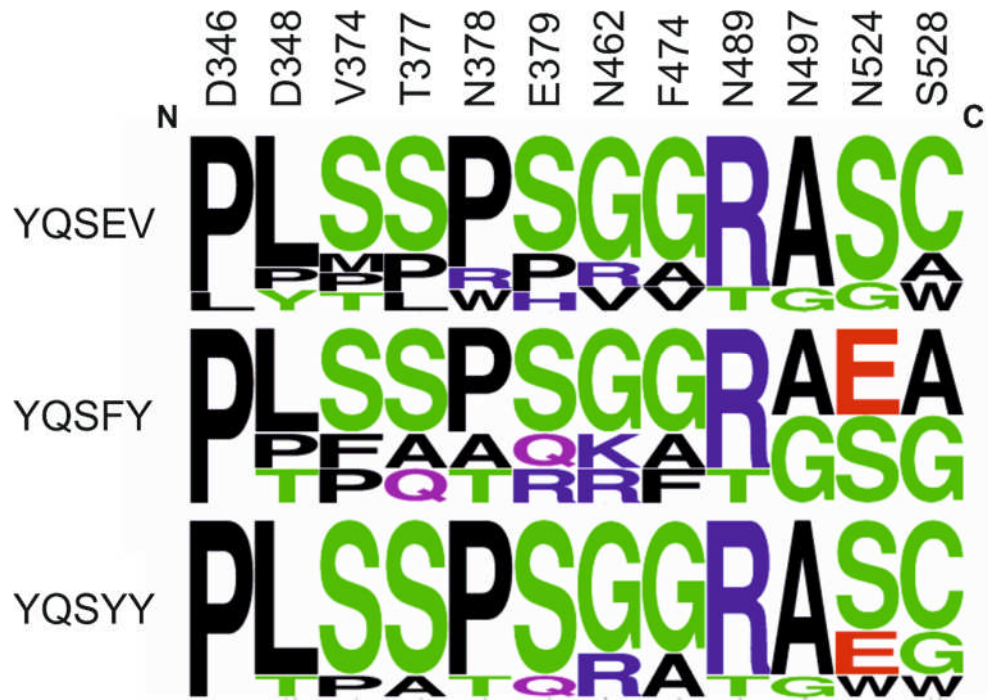


Figure 4.11. Sequence logos displaying the amino acid distribution amongst clones generated from sequencing data after four rounds of selection. Amino acids are shown in single letter format, the height of the letter is proportional to its representation in the library. The wild-type amino acid and number are shown (top). Number of clones sequenced for each library = 10. Amino acid sequences used to generate this figure can be found in the Appendix. Sequence logos were generated using Weblogo²²⁸.

The similarity of the sequenced clones from round four of the selection across all three libraries – with the most abundant amino acid shared at nine of the twelve positions – indicates a number of possibilities. The most likely cause is that the predominant selection pressure was common to all libraries, regardless of chosen target. Another possibility is that selection under these conditions resulted in enrichment of a sequence encoding a promiscuous receptor, capable of binding all three non-canonical peptides. To investigate whether this was the case, the most abundant clone after four rounds (observed in both the YQSEV and YQSY libraries) was chosen for further characterisation.

4.4 Characterisation of selected mutant – PEX5.YY.4.3

A representative clone (PEX5.YY.4.3 - D346P/D348L/V374S/T377S/N378P/E379S/N462G/F474G/N489R/N497A/N524S/S528C) was chosen from the YQSY library for further characterisation. This clone was chosen as it was the most abundant of all those sequenced from round four of the selection. Here, the PEX5.YY.4.3 mutant was cloned into a larger PEX5 construct comprising amino acids 340–728, which has previously

been used in studies of PEX5/PTS1 binding^{229,230}. The PEX5.YY.4.3 mutant was expressed in *E.coli* and purified by Co²⁺-NTA chromatography (Laura Cross, University of Leeds) (Figure 4.12). Soluble PEX5.YY.4.3 protein could be purified from the majority of other cellular proteins (Figure 4.12, Lane 10), however a high proportion of the expressed PEX5.YY.4.3 was present in the insoluble protein fraction (Figure 4.12, Lane 3). Mass spectrometry analysis of the purified PEX5 mutant gave an observed mass of 45315.37 Da (Appendix), which agreed well with the calculated mass of 45316.2 Da.

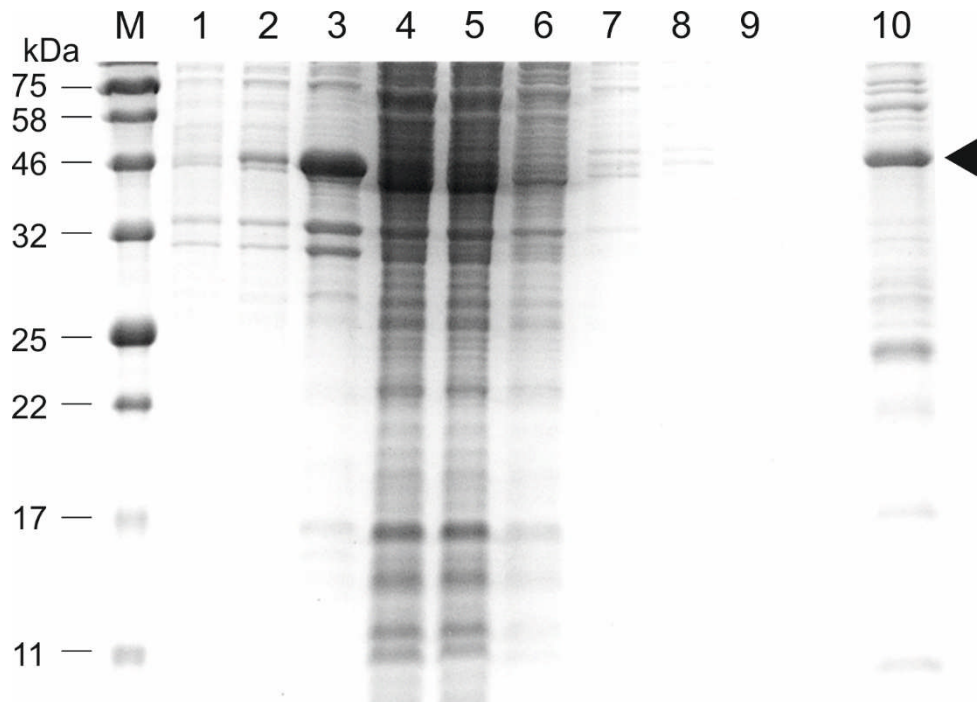


Figure 4.12. Analysis of the expression and purification of the PEX5.YY.4.3 clone (black arrow) by SDS-PAGE. Lane 1, uninduced cells; lane 2, autoinduced cells; lane 3, Insoluble fraction; lane 4, supernatant; lane 5, Co²⁺-NTA flow through; lanes 6-9, wash fractions; lane 10, elution fraction. M, prestained protein marker, masses of which are labelled in kDa.

The binding of the purified PEXYY.4.3 mutant and the wild-type PEX5 receptor to the canonical (YQSKL) and non-canonical (YQSYY) PTS1 peptides was then assayed by fluorescence anisotropy (Laura Cross, University of Leeds) (Figure 4.13). The wild-type PEX5 receptor showed typical binding activity towards the canonical PTS1 peptide (Figure 4.13, black squares), as well as much lower affinity binding to the non-canonical peptide (Figure 4.13, red circles). However, for the PEX5YY.4.3 mutant, whilst detectable binding to the canonical peptide was abolished (Figure 4.13, blue triangles), no

significant binding to the non-canonical peptide was observed (Figure 4.13, purple triangles).

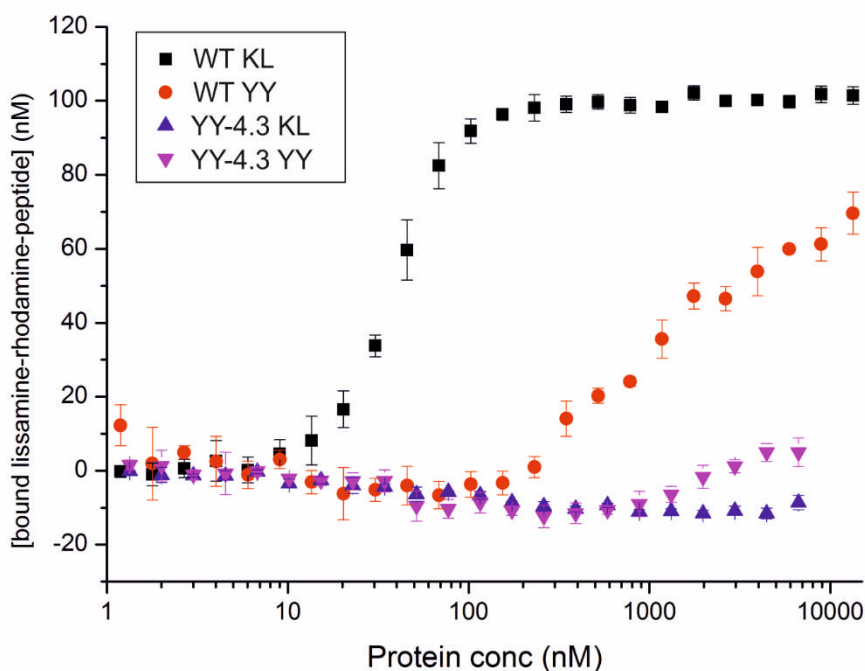


Figure 4.13. Binding of the wild-type PEX5 (WT) and the selected mutant PEX5.YY.4.3 to the canonical (YQSKL) and non-canonical (YQSY Y) peptides. Fluorescence anisotropy measurement of bound lissamine-rhodamine-peptide concentration against PEX5 concentration. A dilution series of PEX5 variants (0.3 nM – 10 μ M) were assayed against 100 nM lissamine-peptide. Each point represents the mean and standard deviation of three independent measurements. Fluorescence anisotropy was performed by Laura Cross (University of Leeds, UK).

The lack of detectable binding to the YQSY Y peptide suggests that the source of enrichment of this sequence during selection may have been artefactual and not based on the ability of these mutants to bind the target peptide. The fact that the PEX5.YY.4.3 clone was the most abundant randomly selected clone across all three libraries after four rounds of selection is suggestive of a common source of artefactual enrichment.

The relative insolubility of the PEX5.YY.4.3 clone when expressed in *E.coli* may have indicated the spurious recovery of sequences based on unfolding or aggregation of the RNA-receptor fusions in the selection step. However, it has been previously observed that the mRNA portion of the fusion improves the solubility of attached proteins, facilitating functional selection of sequences that can aggregate or are only partially soluble when expressed in isolation^{51,128}. It remained a possibility that the selected sequences were enriched based on artefactual binding to either the agarose matrix or walls of tubes. To test this theory, the PEX5.YY.4.3 sequence was expressed as an

RNA-protein fusion and assayed for binding to immobilised YQSYY in the same manner as the wild-type PEX5 in Figure 4.6. No binding to any of the selection matrix or tube could be detected using this assay (data not shown). Therefore, despite the apparently strong enrichment over four rounds of selection using mRNA display, binding affinity was almost certainly not the primary selective pressure present in the selection.

4.5 Analysis of YQSYY library sequences

In an attempt to examine the source of the observed amino acid enrichment during the selection, randomly selected clones from different stages of the selection against YQSYY were sequenced.

Amino acid		NNS ^[a]	Round 0 library ^[b]	Round 2 library ^[b]	Round 4 library ^[b]	
polar	Asn	3.1	3.3	0.9	0	
	Gln	3.1	2.5	2.3	0	
	Ser	9.4	8.3	10.6	22.9	
	Thr	6.3	7.5	7.1	4.2	
basic	Arg	9.4	12.5	10.8	8.3	
	His	3.1	2.5	4	2.1	
	Lys	3.1	3.3	1.1	0	
acidic	Asp	3.1	0	1.7	0	
	Glu	3.1	1.7	3.1	0	
aliphatic	Ala	6.3	10	8.6	8.3	
	Ile	3.1	1.7	1.7	2.1	
	Leu	9.4	9.2	7.1	8.3	
	Met	3.1	0.8	2	0	
	Val	6.3	3.3	6.6	6.3	
	stop					
aromatic	Phe	3.1	2.5	0.8	0	
	Trp	3.1	2.5	4	2.1	
	Tyr	3.1	0	1.1	0	
structural	Cys	3.1	2.5	2.2	4.2	
	Gly	6.3	10	11.6	14.6	
	Pro	6.3	14.2	12.7	16.7	
stop			3.1	1.7	0	0
Codons sequenced:			120	648	120	

Table 4.2. Amino acid distribution for randomised codons during selection against the YQSYY peptide, shown in %. ^[a]Theoretical amino acid distribution for NNS codon ^[b]Experimentally observed values from sequencing analysis of individual library clones.

A total of 54 clones were randomly selected from the PEX5.YY library after two rounds of selection and sequenced. The distribution of amino acids observed at the 888 sequenced randomised codons was analysed, allowing a round-by-round comparison across the mRNA-display selection (Table 4.2). The frequencies of sequenced codons

in the round 0, round 2, and round 4 library clones were compared to that of the theoretical amino acid distribution for an NNS library (Figure 4.14).

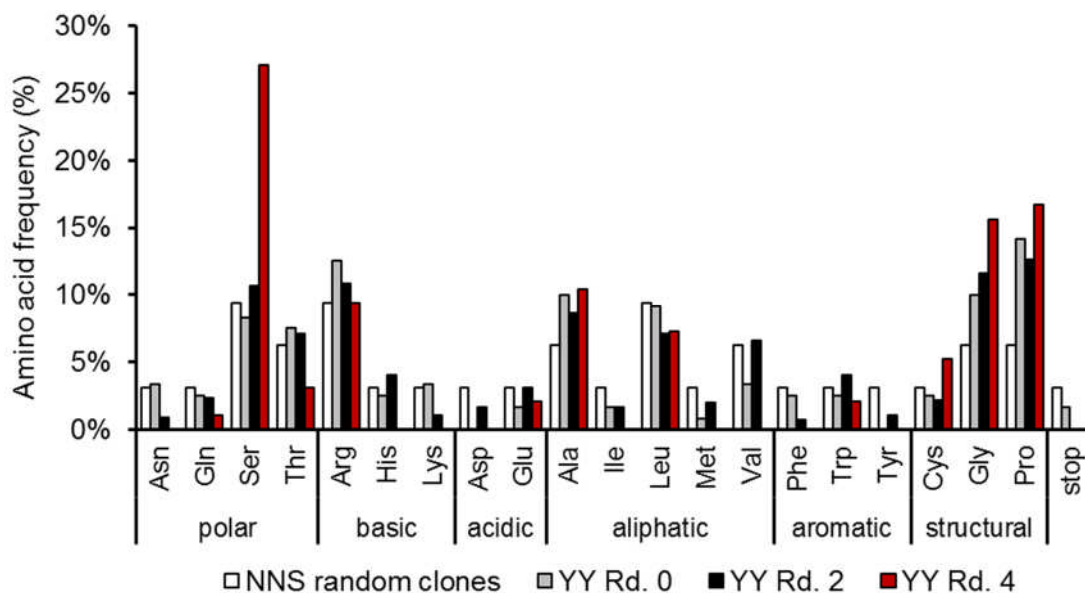


Figure 4.14. The amino acid composition of randomised codons during selection against the YQSYY peptide. Amino acids are grouped according to chemical properties. Theoretical NNS library composition (white); unselected library, randomly picked clones (grey, $n = 120$ codons); YQSYY round 2 randomly selected clones (black, $n = 648$ codons), YQSYY round 4 randomly selected clones (red, $n = 120$ codons).

Initial analysis of the nascent, unselected library (Figure 4.14, grey bars) revealed that although much of the amino acid distribution closely matched that expected from a perfectly randomised NNS codon (Figure 4.14, white bars), a disparity was observed for a small number of amino acids. For example, alanine, glycine, and proline were over-represented by >1.5 -fold in the nascent library. This is perhaps significant, with these amino acids making up a high proportion of the consensus sequences after four rounds of selection (Figure 4.11). However, serine is also significantly over-represented at randomised positions after four rounds of selection, despite being under-represented in the nascent library prior to selection (Figure 4.14).

Unfortunately, an orthogonal PEX5-PTS1 interaction was not isolated by mRNA display selection from a high-diversity library of PEX5 variants. However, the mRNA display selection procedure used led to efficient enrichment of consensus sequences based on a mechanism that has yet to be elucidated. The possible reasons for the observed artefactual enrichment are discussed in detail below.

4.6 Summary

The objective of the work presented in this chapter was to use mRNA display to select a mutant PEX5 peroxisomal import receptor capable of binding to a non-canonical PTS1 sequence and thus generate orthogonal peroxisomal import recognition apparatus. A number of important steps towards this goal were achieved. Initially, an N-terminal 33 kDa truncation of the *Arabidopsis thaliana* PEX5 receptor protein, comprising the PTS1 recognition domain, was successfully displayed on its encoding RNA. Next, the *in vitro* translation conditions were optimised for efficient display of the construct with a concomitant reduction in truncated fusion products (likely due to inhibition of ribosomal mis-initiation and/or degradation of the fusions during translation). In agreement with much of the literature describing mRNA display^{49,126,231}, the PEX5 receptor was shown to retain its PTS1 binding activity when covalently attached to its encoding RNA, demonstrating the potential for selecting novel PEX5 mutants that bind to unnatural peptide ligands.

A high diversity library was assembled using a seamless restriction-ligation strategy, utilising type-IIS restriction enzymes. This strategy has been successfully used to assemble linear double-stranded DNA libraries with high fidelity and low bias for *in vitro* selection^{102,141}. The assembly process was originally designed to allow one-pot digestion and ligation of the full-length library, owing to the designed unique four-base overhangs generated using this method. However, one-pot ligation did not prove to be efficient enough to generate large enough quantities of library DNA for high-throughput selection. Alternatively, the sequential assembly of six fragments was performed in order to generate the full-length PEX5* library. Sequencing analysis of the nascent library confirmed the high fidelity of library assembly – the correct twelve residues were randomised in all sequenced clones, and there was no evidence of accumulation of undesired point mutations during the assembly process (Appendix). The full-length 900 base pair PEX5* library was modified with 5'- and 3'- constant sequences for mRNA display, yielding a DNA library of 4×10^{13} sequences.

The PEX5* library was translated on a large-scale *in vitro* to yield a total of 2.5×10^{12} mRNA-protein fusions. These PEX5*-RNA fusions were split into three sub-libraries and selected against one of three immobilised peptide targets (YQSEV, YQSFY, and YQSY) in the first round. Four rounds of selection were then performed, after which the libraries had converged onto a family of sequences.

The most abundant sequence at four rounds – PEX5.YY.4.3 – was cloned, expressed and its binding to the canonical and non-canonical PTS1 peptides assessed by fluorescence anisotropy. Unexpectedly, the mutant receptor was found to possess no detectable binding either as an RNA-protein fusion or as a free protein. Sequence analysis of clones from before, and throughout, the selection process indicated a nucleotide bias in the library. This bias may have been non-specifically enriched round-by-round as a result of inherent PCR amplification bias.

One of the major strengths of mRNA display is the entirely *in vitro* nature of the selection process. This affords a significant increase in throughput and control over selection conditions. However, when troubleshooting a failed selection experiment, the large number of *in vitro* steps makes it difficult to determine the precise point at which the failure has occurred. Performing control experiments at each step is essential for ensuring that key steps are functioning as expected, for example the generation and purification of RNA-protein fusions in each round. Despite careful controls, it remains challenging to pinpoint the causes of non-specific changes to library populations over several rounds of selection. There are therefore a number of feasible explanations for the apparent failure of the selection experiments described in this chapter, acting either individually or in concert to enrich sequences non-specifically.

An important consideration in an mRNA display selection is the potential to unintentionally select and enrich functional RNA molecules with affinity towards the target of interest. Indeed, selection of functional RNA-protein complexes in this manner would be an intriguing prospect. In this work, the chances of this occurring was mitigated by reverse transcribing the pool of RNA-protein fusions prior to the selection step, which should obviate the selection of sequences based only on functionality at the RNA level.

The overall theoretical diversity of the PEX5* library was 20^{12} , or 4×10^{15} sequences. Consequently, the proportion of theoretical library that was displayed on its encoding RNA in the translation step in round one of selection (2.5×10^{12} RNA-protein fusions) was just 0.06% - a very small proportion of available protein sequence space. Therefore, the activity under selection may not have been present in the starting pool of RNA-protein fusions. Indeed, in any selection experiment that is an exercise in sequence space exploration, it remains possible that the functionality under selection is not present, even in the theoretical library. However, it is worth noting that RNA-ligase enzymes were selected from a pool of 4×10^{12} RNA-protein fusions based on a library with a theoretical diversity of 20^{21} (2×10^{22}) sequences⁴².

Analysis of the *in vitro* translation of the PEX5* library showed that the library migrated as two distinct RNA-protein fusion bands compared to single band for the wild-type RNA-PEX5 fusions (Figure 4.10). Proteolysis of a significant number of library members may have meant that the effective library coverage was even lower than predicted during the first selection step. Fragmentation of the protein portion of the RNA-protein fusions under selection may also have had a profound effect on the outcome of selection. Indeed, it may even explain the enrichment of the sequences observed here via selection pressure towards a truncated species with binding activity under the conditions used in the selection step. This potentially demonstrates the importance of utilising high quality libraries when performing high-throughput selections. In a recent example, successful enrichment of randomised libraries for folded proteins was demonstrated, using mRNA display in combination with protease resistance selection, generating a high quality $(\beta/\alpha)_8$ barrel library for enzyme selections¹⁴¹.

Analysis of the composition of randomised codons in the PEX5* library at round 0 indicated a bias towards the incorporation of cytosine (38% occurrence) at position 1 of the codon, and cytosine and guanosine as position 2 (35% and 33%, respectively). In the context of an NNS codon, where the third nucleotide is always either C or G, this confers a bias towards the appearance of proline (CCT, CCC, CCG, CCC) in the library. A large number of PCR cycles are required to perform this completely *in vitro* selection technique. Indeed, over four rounds of selection the PEX5.YY DNA library was subjected to more than 100 cycles of PCR. If they were present in high enough proportion in the nascent library, this may be enough to explain the enrichment of the sequences seen after four rounds of selection. The sequence that dominated all libraries after four rounds of selection (PEX5.YY.4.3) corresponded to two of fifty four clones sequenced (3.7%) after two rounds of selection against YQSYY, further supporting the notion that inherent library bias played a role in the outcome of selection. If the enrichment of artefactual sequences in the selection step can be attributed in some part to inherent library bias, modification of library assembly strategies may help to alleviate this problem in future. Using mixtures of 20 separately synthesised oligonucleotides or utilising trinucleotide phosphoramidites in the synthesis of randomised oligonucleotides²³² can result in libraries that obviate the inherent problem of bias in the genetic code.

Stringency of the selection must also be taken into account. Indeed, if the selection step was more stringent, then the apparent bias present in the PEX5* library may not have had an opportunity to dominate the library after four rounds of selection. However, high stringency can prove disadvantageous during the first rounds of selection, as rare

functional sequences can theoretically be lost from the library. High stringency selection has been successfully used to isolate binding peptides against calmodulin, where a 40 column volume wash resulted in isolation of peptides that bound the target with affinities ranging from 5 to 300 nM after two rounds of selection²³¹. Therefore, remaining DNA libraries from early rounds of selection (for example, after rounds 1 and 2), may be amenable to selection with a significantly increased stringency in an effort to isolate functional sequences. However, it is unlikely that the ease of isolating peptide binders to a target is directly comparable to that of a large, dynamic protein such as the PEX5 receptor, where the combinatorial mutations may have significant and far-reaching implications on folding and/or stability.

In summary, a high-diversity library was designed and synthesised based on the PEX5 peroxisomal import receptor from *A. thaliana*. Four rounds of affinity panning of the resulting RNA-protein fusion library for binding to three candidate orthogonal peptides (YQSY, YQSFY, YQSEV) were performed. The ambiguous outcome of the selection experiments outlines some of the challenges associated with selection for novel activity in the constraints of an existing protein fold. The prospect that the functionality under selection was not present in the initial pool of RNA-protein fusions must be considered. Indeed, more intelligent and/or more conservative library design strategies may have resulted in a higher proportion of active PEX5 receptor variants. In parallel to the experiments described in this chapter, a more focussed, site-directed mutagenesis approach in combination with a mass-spectrometry based screen proved more successful in the search for an orthogonal receptor-targeting signal interaction²¹⁸.

5 mRNA display with interaction-dependent reverse transcription

Recent advances in genome and proteome research have led to a concomitant and dramatic increase in the number of pharmaceutically relevant macromolecular targets of interest. This has led to the development of a variety of target-oriented high-throughput screening methods for the discovery of small molecule^{233,234}, nucleic acid²³⁵, peptide²³⁶, and protein²³⁷ ligands from large libraries.

When the desired functionality is binding affinity, it is possible to select directly – typically with immobilised target molecules – whereby bound library members are washed and eluted before being amplified by PCR (as described in Chapter 4). Whilst powerful, target-oriented techniques such as these suffer from fundamental limitations; they rely on immobilised targets or ligands, which may result in artifactual binding or the loss of native conformational characteristics required for authentic binding to native targets in solution (particularly proteins)²³⁸. Furthermore, the additional washing and elution steps required for solid-phase capture selections can be experimentally cumbersome and may also result in loss of rare functional library members in early rounds of selection.

Novel *in vitro* techniques for interrogating synthetic small molecule libraries in the solution phase have recently been developed using DNA-encoded libraries. The highly combinatorial and sensitive nature of these techniques has resulted in their increasing use for the discovery of chemically synthesised drug-like ligands²³⁹. Recently, this concept has been built upon in the context of *in vitro* DNA-encoded selections for chemical reactivity²⁴⁰ and binding affinity^{241,242} from large, high-complexity synthetic libraries using interaction-dependent PCR.

5.1 Interaction-dependent PCR

Interaction-dependent PCR (IDPCR) methods are being established as tools for a range of applications including the simultaneous screening of libraries of small molecule ligands and targets^{241,242}, and the detection of biomarkers in diagnostic tests^{243,244}. In IDPCR, ligands and targets are individually tagged with synthetic DNA oligonucleotides. Binding interactions are identified based on the principle that the melting temperatures (T_m) of double stranded nucleic acids increase considerably when hybridisation occurs intramolecularly as opposed to intermolecularly²⁴⁵ (Figure 5.1a). Ligand-target binding brings the encoding DNA sequences into close proximity, thus promoting DNA hybridisation. Subsequent polymerase-catalysed extension of the hybridised DNA

generates an amplifiable double-stranded DNA sequence that encodes both the ligand and the target. The difference in duplex stability between the bound and unbound members of the library ensures preferential extension, and thus PCR amplification, of DNA sequences that encode the functionally active species. This streamlines conventional selection methods that rely on “affinity panning” in which unfit library members are discarded followed by amplification of surviving sequences in a subsequent step. In contrast, functional sequences are preferentially amplified from the complex selection milieu in a single step. IDPCR is one of few currently available methods for identification of interaction partners from libraries of ligands and libraries of targets in a single, solution-phase experiment^{246,247}.

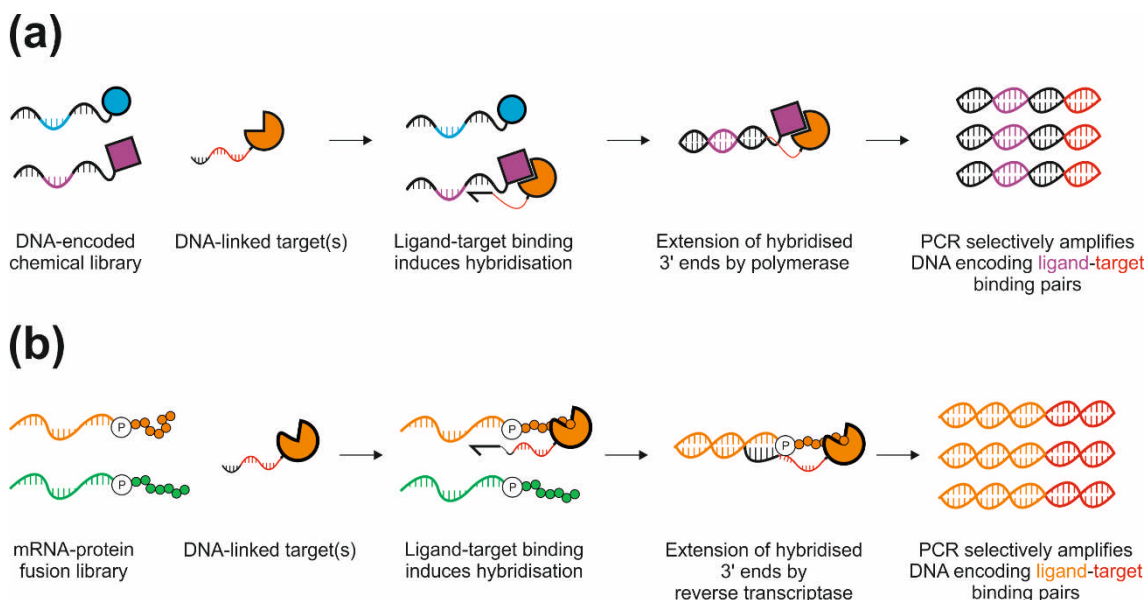


Figure 5.1. (a) Solution phase identification of ligand-target pairs from libraries of small molecule ligands by interaction-dependent PCR. A DNA-encoded small molecule library is mixed with one or more DNA-linked targets. Ligand-target binding induces hybridisation of the DNA strands, which are specifically extended by a DNA polymerase. DNA sequences encoding the identity of both the ligand (purple) and target (red) are readily amplified by PCR. (b) Solution phase identification of ligand-target pairs from mRNA display libraries by interaction-dependent reverse transcription PCR (RT-PCR). The selection scheme is analogous to that shown in (a), except that the DNA-RNA duplex is extended by a reverse transcriptase enzyme to generate a cDNA strand encoding both the ligand (orange) and target (red).

The concept of IDPCR has been demonstrated in model selections of known small molecule ligand-target pairs from mixtures of non-functional targets and nucleic acid

sequences²⁴¹. In one such study, the authors readily identified five known ligand-target pairs from a mixture of more than 67,000 possible sequence combinations by high throughput DNA sequencing, despite the fact that target affinities ranged from 40 pM to 3 μ M.

This work aims to expand the capabilities of interaction-dependent amplification methods to utilise the genetic code and Nature's translational machinery to generate a diverse library of nucleic acid-tagged ligands using mRNA display (Figure 5.1b). Subsequent incubation with one or more DNA-tagged targets would allow interaction-dependent extension of the mRNA-DNA duplex using a reverse transcriptase enzyme. Amplification and sequencing of the resulting cDNA would thus reveal the identity of ligand-target pairs from large, high-diversity protein libraries.

mRNA display lends itself to selections of this kind, due to the inherent covalent link between the protein ligand/target and its encoding nucleic acid (Figure 5.1B). Alternative *in vitro* display methods may also be compatible with the selection scheme outlined in Figure 5.1. Given that the technique requires single-stranded nucleic acid tags, DNA display methods are not compatible with interaction-dependent extension (Section 1.3.4). Ribosome display may allow functional selection in this manner, however the non-covalent ternary complex that maintains the phenotype-genotype link would likely be unstable under conditions that favour nucleic acid extension by polymerase/reverse transcriptase enzymes. Furthermore, the large size of the ribosome would potentially obviate nucleic acid hybridisation as a result of steric hindrance. In contrast, the covalent phenotype-genotype link in mRNA display is robust, able to withstand the increases in temperature required for enzymatic nucleic acid polymerisation, and is achieved via the comparatively small aminonucleoside analogue puromycin.

The combination of these two powerful selection techniques may enable streamlined solution-phase *in vitro* selection of novel functional peptide and protein biomolecules. It also raises the possibility of multiplexed one-pot selections of large randomised libraries against many targets of interest in one experiment.

5.2 Interaction-dependent reverse transcription (IDRT) using RNA-protein fusions

In order to validate IDRT-PCR using mRNA display, the relatively well characterised streptavidin binding peptide (SBP) – streptavidin interaction was chosen. The 38 amino acid SBP peptide (Figure 5.2, SBP tag) was isolated by Szostak and co-workers in 2001 via mRNA-display selection from a library of approximately 10^{13} random peptides – therefore is expected to retain functionality when covalently displayed on its encoding RNA¹³⁰. The nature of SBP binding to streptavidin has since been investigated using crystallographic methods in combination with surface plasmon resonance (SPR), resulting in a minimal 24 amino acid SBP sequence, designated SBP-tag2, (Figure 5.2, SBP tag2) with a binding affinity almost identical to that of the full-length peptide ($K_d = 1.5 \text{ nM}$)²⁴⁸. This minimal SBP sequence was chosen for mRNA-display experiments. An N-terminal methionine for translation initiation and a flexible C-terminal linker were included in the mRNA-display construct (Figure 5.2, shown in red).

SBP tag	MDEKTTGWRG GHVVEGLAGELEQLRARLEHHPQGQREP
SBP tag2	GHVVEGLAGELEQLRARLEHHPQG
mRNA display construct	MGHVVEGLAGELEQLRARLEHHPQGMGMSGSGTGY

Figure 5.2. Streptavidin binding peptide (SBP) sequences. The minimal SBP sequence is highlighted in blue. Additional amino acids added for mRNA display in this work are shown in red.

The SBP mRNA-display construct was codon optimised for expression in rabbit reticulocyte lysate and the 188 base pair linear double-stranded DNA construct (Integrated DNA Technologies) was cloned into the puc18 vector to allow bacterial propagation and stable storage of the construct. For mRNA display, the SBP gene was modified with the required 5'- and 3'- sequences by PCR, using the appropriate primers (SBP Mod For, SBP Mod Rev, see Appendix for sequences). RNA-protein fusions were generated as described in Materials and Methods. *In vitro* translation and RNA-SBP fusion formation were performed using conditions previously described by Wilson *et al.*¹³⁰, and RNA-SBP fusions were purified from crude translation reactions using oligo-dT cellulose as described in Materials and Methods.

5.2.1 Design of hybridisation sequences

A central aspect of any proximity-based extension assay is the design of the nucleic acid sequences that are expected to hybridise and prime DNA extension upon ligand-target binding. If the hybridisation region is too long, the T_m of the resulting duplex is too high, resulting in a higher likelihood of non-specific priming events. This would lead to a high level of interaction-independent amplification and thus a low signal-to-noise ratio. Too short, however, and the occurrence of an interaction between the ligand and target may not be sufficient to promote hybridisation under the conditions of the assay. Previous work in this area has favoured the use of hybridisation regions of 6-9 nucleotides in length, with the optimum length varying depending on the specific DNA polymerase used in the extension reaction^{241,242,244}.

As all previously published work in this field has featured the DNA-templated extension of DNA oligonucleotides, the RNA-templated extension of DNA in IDRT reactions may require further optimisation. Indeed, the thermodynamic stability of DNA-RNA hybrids, whilst always lower than that of the homologous RNA duplex, can be either more or less stable than the corresponding DNA duplex, depending on base-composition^{156,249,250}. For this reason, the length of complementarity in the hybridisation region was varied between 6, 8, and 10 nucleotides and assayed for specific interaction dependent reverse transcription.

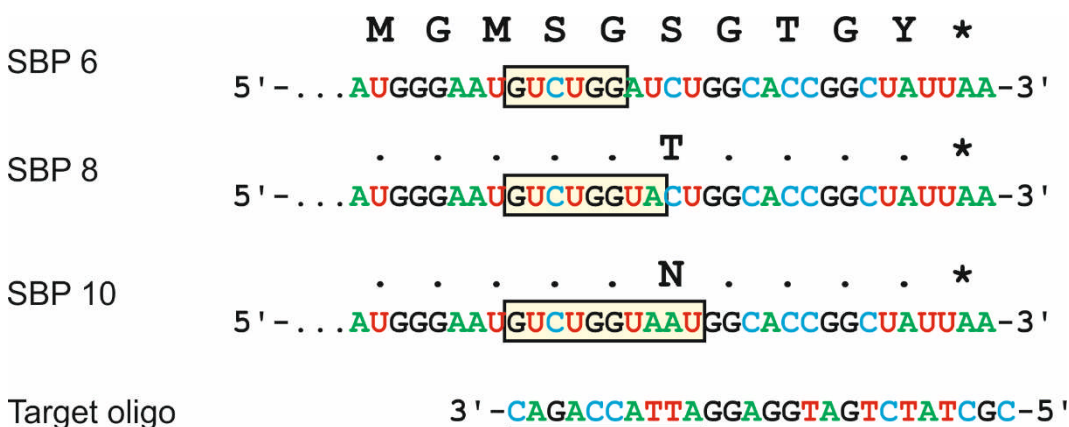


Figure 5.3. Design of SBP mRNA display constructs to incorporate a range of hybridisation lengths. Point mutations were incorporated in the 3'-region of mRNA templates encoding the flexible C-terminal linker to generate templates with a 6, 8, and 10 base region (yellow boxes) complementary to the 3'-end of the target oligonucleotide (underlined). The resulting amino acid sequence of the C-terminal linker is shown above each sequence.

This variation was introduced at the RNA transcript level during mRNA display via PCR amplification using different primers (Figure 5.3), rather than by synthesis and purification of different streptavidin-oligonucleotide conjugates, which is both time consuming and resource intensive. Point mutations were incorporated in the 3'-region of mRNA templates encoding the flexible C-terminal linker in order to generate templates with a 6- (SBP 6) , 8- (SBP 8), and 10-base (SBP 10) region, complementary to the 3'-end of the target oligonucleotide (Figure 5.3, underlined). These nucleotide changes result in single amino acid substitutions in the sequence of the C-terminal linker and do not effect the sequence of the core SBP peptide.

5.2.2 Covalent coupling of DNA oligonucleotides to streptavidin

Essential requirements for solution-phase selection in this manner are (i) a library of ligands and (ii) one or more targets, both conjugated to encoding nucleic acid tags. In IDRT-PCR, mRNA display provides a covalent link between ligands and their encoding nucleic acid, here in the form of RNA-SBP fusions. In order to generate a DNA-tagged target for SBP, streptavidin was covalently linked to a 5'-thiol modified oligonucleotide (Integrated DNA Technologies) via the heterobifunctional crosslinker sulfosuccinimidyl 4-(N-maleimidomethyl)-cyclohexane-1-carboxylate (sSMCC) (Figure 5.4a). The ϵ -amino groups of lysine side chains of streptavidin were first reacted with sSMCC to introduce a maleimide moiety, which was subsequently reacted with a 5'-thiol oligonucleotide. The resulting streptavidin-oligonucleotide conjugates were purified by anion exchange chromatography (Appendix), and analysed by native PAGE (Figure 5.4b). In order to assess the efficiency of the conjugation reaction, the same native PAGE gel was stained to detect both protein (Figure 5.4b, left panel), and nucleic acid (Figure 5.4b, right panel). Unreacted streptavidin (Figure 5.4b, lane 1) migrated as an apparent mixture of protein species, consistent with the molecular mass of the native tetrameric form (~60 kDa) and that of the dimer (~40 kDa), with no nucleic acid evident on the gel (Figure 5.4b, right panel, lane 1). The crude conjugation reaction (Figure 5.4b, lane 3), showed a similar pattern of protein migration as the unreacted streptavidin (Figure 5.4, left panel, lane 3), however when stained for nucleic acid, a band could be seen that migrated at approximately 200 bp (Figure 5.4, right panel, lane 3). Following purification by anion exchange chromatography (Figure 5.4b, left and right panel, lane 2), a single band was observed that migrated at ~40 kDa/200 bp was detected when stained for both protein and nucleic acid, thus corresponding to the streptavidin-oligonucleotide conjugates.

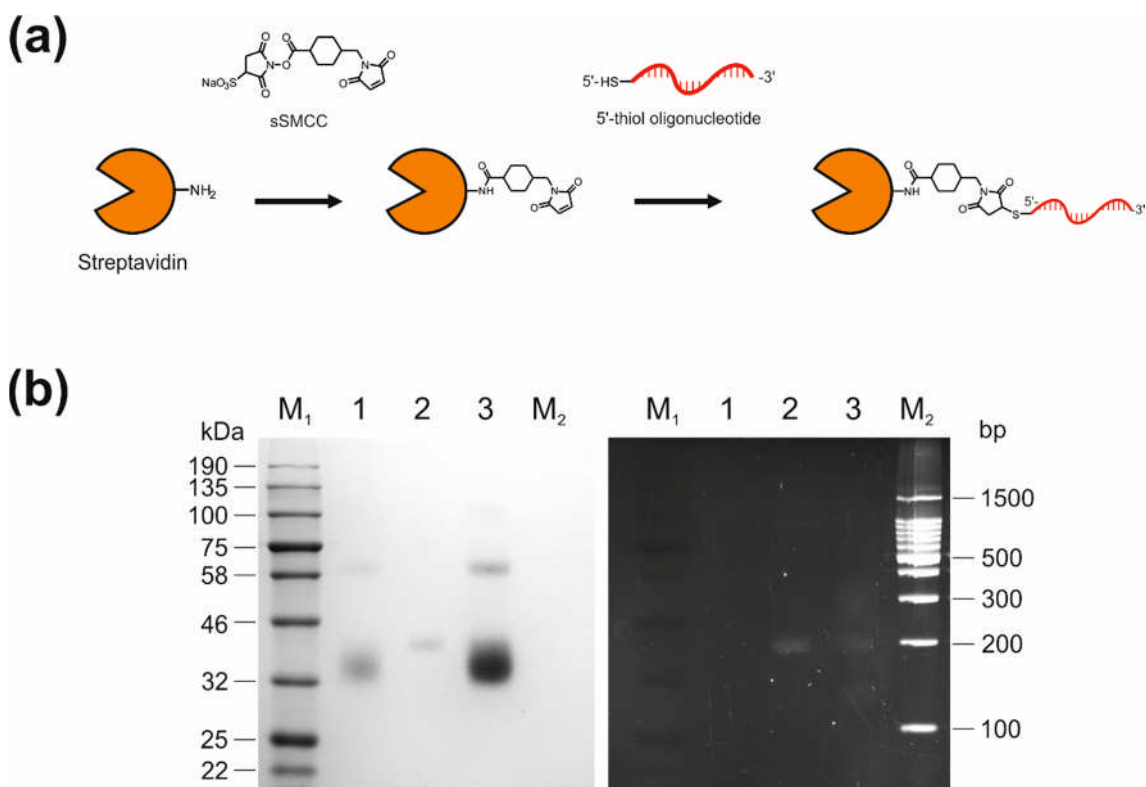


Figure 5.4. **(a)** Schematic of the strategy used to generate DNA-tagged streptavidin. First, primary amines on the surface of streptavidin were reacted with the NHS-ester group on the heterobifunctional crosslinking reagent sSMCC, followed by reaction of the maleimide with a 5'-thiolated oligonucleotide. **(b)** Native PAGE analysis of streptavidin-oligonucleotide conjugates. The gel was stained to detect both protein (left panel), and nucleic acid (right panel). M₁, prestained protein markers; lane 1, streptavidin; lane 2, pooled fractions; lane 3, crude crosslinking reaction; M₂, 100 base-pair ladder.

5.3 Interaction-dependent reverse transcription PCR

Previous work towards the development of solution-phase selection of ligand-target pairs has utilised DNA tags rather than a combination of DNA and RNA tags. Therefore, mesophilic DNA polymerases (e.g. Klenow *exo*[±], T4, T7, Pol I, and Phi29 polymerases) have been tested for use in the extension reaction, with T4 DNA polymerase resulting in the highest signal-to-noise ratios published to date^{241,242,244}.

A number of enzymes have been identified that catalyse the RNA-templated extension of DNA oligonucleotides. Of those that are commercially available, the most commonly used for molecular biology workflows are cloned and engineered variants of retroviral reverse transcriptase enzymes from Avian Myeloblastosis Virus (AMV) and Moloney Murine Leukemia Virus (M-MuLV, MMLV), which use them to make DNA copies of their

RNA genomes²⁵¹. A characteristic of wild-type versions of the AMV and MMLV reverse transcriptases that precludes their use for IDRT-PCR is their intrinsic RNase H activity. This results in the degradation of the RNA portion of a DNA/RNA duplex, which in standard reverse transcription reactions can cause truncation of cDNA products and reduction in cDNA yield²⁵².

High levels of RNase H activity in the IDRT reaction would result in digestion of the mRNA portion of the RNA-protein fusions and the dissociation of the phenotype-genotype link. Fortunately, the discovery of a point mutated variant of MMLV reverse transcriptase that possessed no detectable RNase H activity²⁵¹ and subsequent engineering efforts have resulted in a range of reverse transcriptase enzymes that possess little to no RNase H activity. The enzyme chosen for the IDRT experiments performed here was a commercially available engineered version of the MMLV reverse transcriptase (Superscript II, Invitrogen). This was based on reduced levels of RNase H activity and increased stability compared to the wild-type MMLV enzyme.

First, the ability of IDRT-PCR to specifically amplify nucleic acid in the case of a ligand-target interaction was assessed. To this end, oligo-dT cellulose purified RNA-SBP fusions with hybridisation lengths of 6, 8, and 10 nucleotides were incubated with DNA-tagged streptavidin, either in the presence or absence of a 200-fold excess of biotin, added to competitively inhibit the interaction (Figure 5.5a). The ability to generate cDNA, and the dependence of cDNA synthesis on interaction between SBP and streptavidin was tested by reverse-transcription PCR (RT-PCR), followed by agarose gel electrophoresis (Figure 5.5b).

For hybridisation regions 6, 8, and 10 nucleotides in length, cDNA was efficiently amplified from the mixtures in a reverse transcription-dependent manner (Figure 5.5b, lane 4). However, for all hybridisation lengths, amplification of SBP cDNA was observed both when the reverse transcription primer was not conjugated to streptavidin (Figure 5.5b, lane 3) and when the ligand-target interaction was inhibited by the addition of excess biotin (Figure 5.5b, lane 5). This indicates a background level of reverse transcription that occurs independently of ligand-target binding. However, amplification of SBP cDNA by RT-PCR is consistently increased in the IDRT condition, indicating that interaction between the RNA-SBP fusions and DNA-tagged streptavidin is promoting cDNA synthesis. There is no obvious correlation between the hybridisation length and the overall levels of cDNA synthesis, nor the relative levels of background amplification.

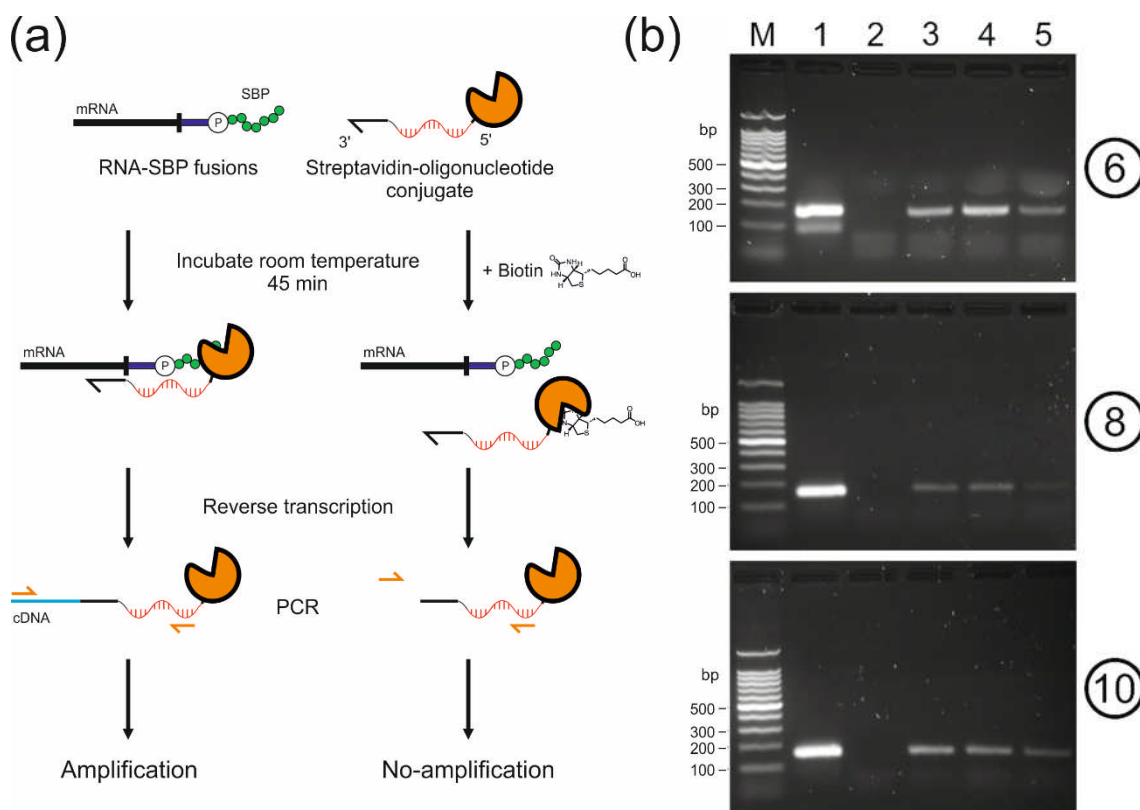


Figure 5.5. Proof-of-concept interaction dependent reverse transcription. **(a)** RNA-SBP fusions and DNA-tagged streptavidin were incubated either in the presence or absence of 10 μM biotin as a competitive inhibitor of the SBP-streptavidin interaction. The ability to specifically amplify SBP cDNA in the presence of an interaction was assayed by RT-PCR. **(b)** Agarose gel electrophoresis of RT-PCR reactions for hybridisation lengths of 6 (top panel), 8 (middle panel, and 10 (bottom panel) nucleotides. Lane 1, Positive control RT-PCR with gene specific RT primer; lane 2, negative control RT-PCR; lane 3, RT with 50 nM RT primer minus streptavidin; lane 4, RNA-SBP fusions + streptavidin-oligo IDRT-PCR; lane 5, RNA-SBP fusions + streptavidin-oligo + 10 μM biotin IDRT-PCR. M, 100 base pair ladder.

Whilst interaction between RNA-SBP fusions and DNA-tagged streptavidin appears to promote cDNA synthesis in this assay, the background levels of cDNA synthesis observed would be problematic for library-format selections. Therefore, strategies to decrease the non-specific background and increase the overall signal-noise ratio in the interaction dependent RT-PCR assay were considered.

A key finding from previous studies is that polymerases with inherent 3'-exonuclease activity result in an increased signal-to-noise ratio in these assays^{241,242,244}. This was confirmed by the discovery that this effect could be rescued when using Klenow exo⁻ by supplementing the reaction with exonuclease I²⁴⁴. This effect can be explained by the degradation of free ssDNA 3'-ends by the 3'-exonuclease, preventing the accumulation of extension products due to random proximity events during the reaction. As reverse

transcriptase enzymes are typically devoid of any 3'-exonuclease activity, exonuclease I was supplemented into the IDRT reactions at a range of concentrations.

The effect of exonuclease I in the IDRT reaction is theoretically determined by the ability of the ssDNA to hybridise to its target strand, as hybridised strands are more resistant to digestion than free DNA 3'- ends. Therefore 6, 8, and 10 nucleotide hybridisation regions were again tested. Of the hybridisation lengths tested, the signal-noise ratio in the IDRT-PCR was only improved for the 8 nucleotide construct (Figure 5.6). Inclusion of exonuclease I in IDRT-PCR with the 10 nucleotide construct did not result in an increase in signal-noise ratio (Appendix), likely due to high levels of non-specific hybridisation. In contrast, supplementation with exonuclease I abolished detectable IDRT-PCR with the 6 nucleotide construct (data not shown).

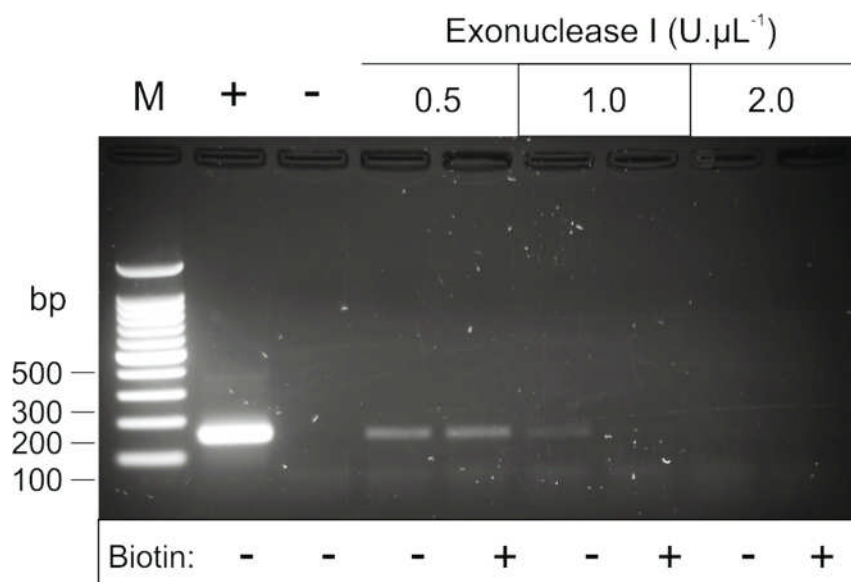


Figure 5.6. Agarose gel analysis of exonuclease I optimisation in the IDRT-PCR with an 8-nucleotide hybridisation region. Lane 1, Positive control RT-PCR with gene specific primer; lane 2, negative control RT-PCR; lanes 3-4, IDRT-PCR with 0.5 U.µL⁻¹ exonuclease I; lanes 5-6, IDRT-PCR with 1 U.µL⁻¹ exonuclease I; lanes 7-8, IDRT-PCR with 2 U.µL⁻¹ exonuclease I. M, 100 base pair ladder.

For the 8-nucleotide hybridisation construct, the addition of 0.5 U.µL⁻¹ exonuclease I did not improve the signal-noise ratio, with a significant signal observed even in the presence of biotin. However, when the exonuclease I concentration was increased to 1 U.µL⁻¹, background amplification was completely abolished in the presence of biotin. Raising the exonuclease I concentration further, to 2 U.µL⁻¹, completely abolishes IDRT-PCR

amplification. At this concentration, the 3'-exonuclease activity is likely overwhelming the cDNA synthesis activity of the reverse transcriptase enzyme. As the addition of 1 U. μ L⁻¹ exonuclease I removed detectable amplification of the SBP sequence when the SBP-streptavidin interaction was competitively inhibited by an excess of biotin, this condition was carried forward into mock selection experiments.

5.4 Mock selection for SBP peptides

Having identified conditions that allow preferential synthesis of cDNA in the event of a protein-ligand interaction, the next step was to investigate the performance of this technique in a mock library format selection. In order to do this, a negative control construct was made in which the HPQ motif, shown to be essential for binding to streptavidin¹³⁰, was mutated to HGA. This variant has been shown to retain less than 0.1% of the binding activity of the SBP peptide in a streptavidin column binding assay¹³⁰. To enable distinction between DNA sequences corresponding to binding and non-binding species, a unique SacI restriction site was introduced into the SBP Δ mutant. The inserted SacI site was silent at the amino acid level and did not result in further change to the translated amino acid sequence of the SBP Δ mutant.

Construct	Sequence	% binding activity ^[†]
SBP	MGHVVEGLAGELEQLRARLEHHPQG	100
SBP Δ	MGHVVEGLAGELEQLRARLEHGGAG	< 0.1

Table 5.1. Positive and negative control SBP constructs discovered by Wilson *et al.* using mRNA display¹³⁰. [†]binding activity of each peptide relative to the SBP peptide from the same study.

The unique restriction site in the SBP Δ DNA sequence was located so as to generate two fragments of approximately 80 base pairs upon digestion. Therefore, following digestion with SacI, binding and non-binding sequences could be visualised by gel electrophoresis (Figure 5.7). As expected, the SBP DNA was not digested by SacI (Figure 5.7, Lane 1), whereas the SBP Δ DNA migrated as two bands of approximately equal size (Figure 5.7, Lane 2). This allows simple and rapid visualisation of the ratio of SBP:SBP Δ DNA in IDRT-PCR output.

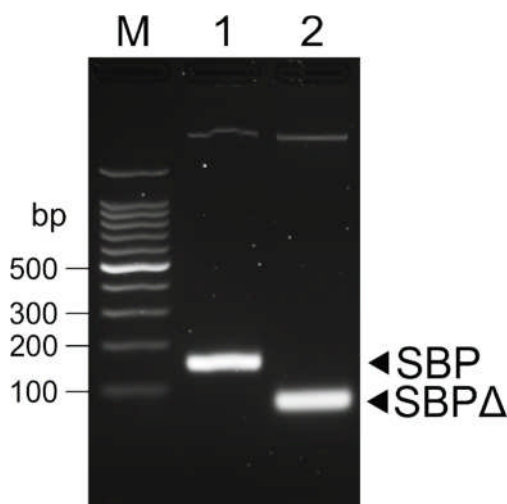


Figure 5.7. Binding and non-binding SBP mutants can be distinguished by restriction digest. SBP mutants were amplified by PCR, followed by *SacI* digestion. Lane 1, SBP; lane 2, SBP Δ ; M, 100 base-pair ladder.

In vitro selection techniques are most useful for the interrogation of large libraries of compounds or large areas of protein sequence space. For IDRT-PCR to be a useful technique for the isolation of ligand-target interactions from large libraries, it must be able to enrich functional sequences from a large excess of non-functional DNA sequences. Indeed, highly randomised protein libraries are expected to be composed of largely non-functional sequences. The ability to distinguish between SBP and SBP Δ sequences makes it possible to test the capacity of IDRT-PCR to enrich functional sequences from a large excess of non-functional sequences. Analysis of the ratio of cut:uncut DNA in the digestion of the IDRT-PCR product reflects the enrichment of the ligand (Figure 5.8a). In order to test this experimentally, puromycin-modified RNA templates were prepared for both the SBP and SBP Δ sequences, and serially diluted to achieve ratios of 1:10, 1:100, and 1:1000 SBP:SBP Δ template.

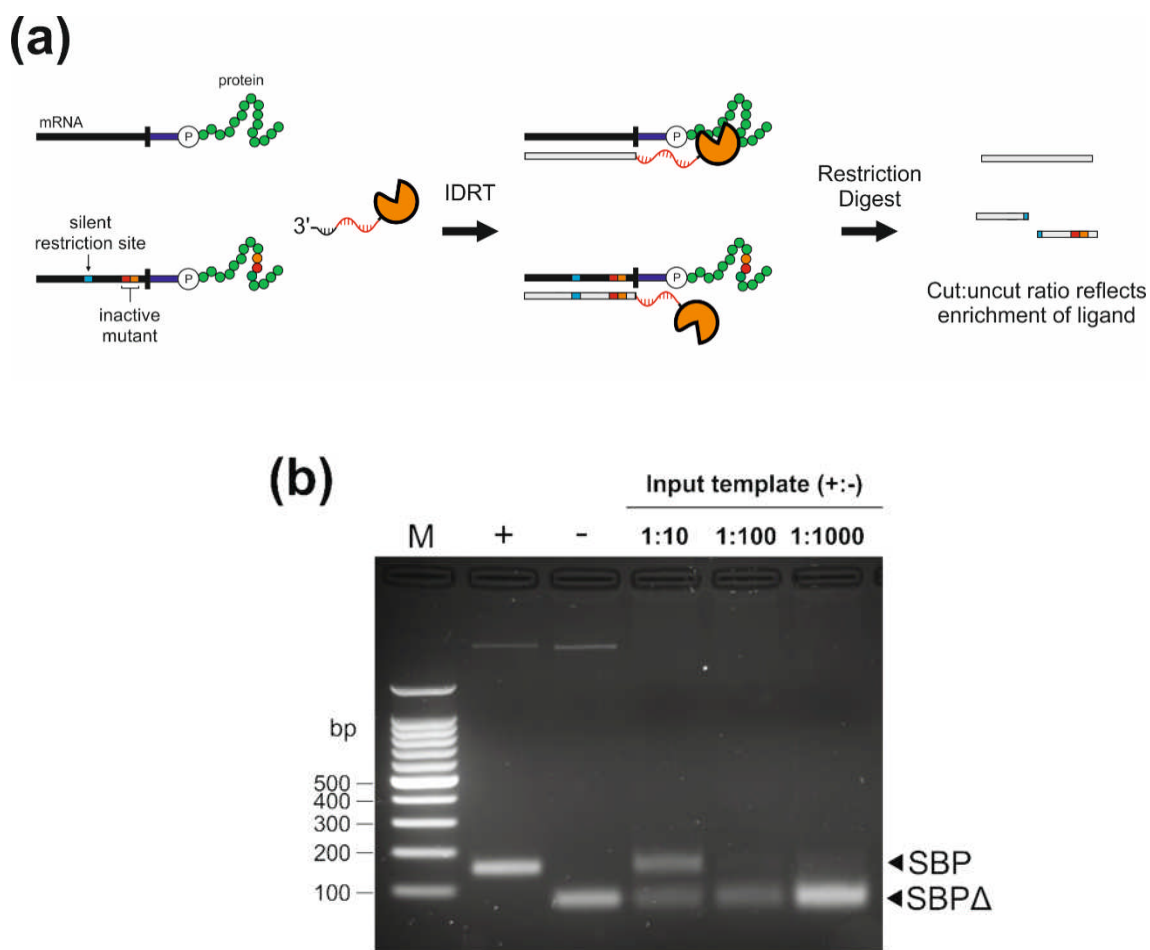


Figure 5.8. (a) IDRT selection of SBP peptides in a mock library format using an excess of non-functional SBP RNA as template in the translation reaction. **(b)** DNA generated from PCR amplification of IDRT reactions with input SBP:SBPΔ template ratios of 1:10, 1:100, and 1:1000 was restriction digested with *SacI* and analysed by gel electrophoresis. +, SBP; -, SBPΔ; M, 100 base-pair ladder.

These mixtures of RNA-puromycin templates were translated *in vitro* with a constant total template concentration of 200 nM. Following purification by oligo-dT cellulose chromatography, each mixture of RNA-protein fusions was subjected to IDRT-PCR, and the resulting PCR product was digested with *SacI* (Figure 5.8b).

Functional SBP sequences could be efficiently enriched from a 10-fold excess of non-functional SBPΔ DNA (Figure 5.8b, 1:10). There may also be a small amount of enrichment from a 100-fold excess of non-functional sequences (Figure 5.8b, 1:100), however at a 1000-fold excess the IDRT-PCR was dominated by SBPΔ sequences (Figure 5.8b, 1:1000). This indicates that an enrichment of between 10 and 100-fold is achievable using IDRT-PCR, comparable to the 10-1000 fold enrichments seen for other *in vitro* display technologies²⁵³.

5.5 Summary

DNA-tagged biomolecules are powerful tools for solution-phase functional selection, and their use has flourished due to recent advances in oligonucleotide synthesis and next-generation sequencing technologies. In this chapter, solution-phase selection has been combined with the high-throughput combinatorial power of *in vitro* display. The amalgamation of these powerful techniques should ultimately allow the solution-phase selection of libraries of ligands against libraries of targets, all within a single experiment. This will enhance efforts to discover new ligand-target interactions, to decipher target-binding specificities of protein ligands, and to detect rare functional proteins in high-diversity protein libraries.

A general requirement for the use of IDRT for ligand discovery is the generation of DNA-linked targets. In this work, streptavidin-oligonucleotide conjugates were synthesised using the heterobifunctional crosslinking reagent sSMCC, using a previously reported strategy^{254,255}. Alternative strategies exist for the preparation of a range of DNA-linked targets, most of which rely on thiol or amino modified oligonucleotides to generate a covalent link. However, the nature of the chemical crosslinking process does not allow control over the number of oligonucleotides conjugated per molecule nor their sites of attachment. It has been demonstrated that, at least for IDPCR, a ternary complex can support PCR amplification of ligands²⁴². Therefore, alternative strategies for DNA-tagging of proteins may be compatible with IDRT selection. For example, fusion with SNAP^{256,257} and HaloTag^{258,259} proteins would allow site-specific covalent attachment of oligonucleotides at a 1:1 ratio of protein:oligonucleotide. An intriguing prospect is that libraries of nucleic acid-tagged ligands and targets could be simultaneously generated *in vitro* using mRNA display, obviating the need for separate target conjugation protocols.

The mRNA display selection protocol is significantly streamlined via the incorporation of IDRT-PCR. The need to prepare immobilised targets is circumvented, and the experimentally cumbersome washing and elution steps are removed. In practical terms, unlike traditional mRNA display selection methods, the entire IDRT-PCR selection protocol can be performed in a single day. Previous studies have demonstrated the utility of IDPCR protocols in crude cell lysates²⁴² and in human serum²⁴⁴. Therefore, it is possible that IDRT-PCR could be used directly on *in vitro* translation mixtures, without purification of RNA-protein fusions. One of the greatest barriers to IDRT-PCR in crude lysate is the observed inefficiency of reverse transcription in rabbit reticulocyte lysate¹²⁶. However, reconstituted *in vitro* translation systems based on purified components^{192,193}

in combination with more robust reverse transcriptase enzymes may alleviate this problem.

The IDRT-PCR strategy described here lends itself to the multiplex one-pot interrogation of libraries for protein-ligand interactions. Diversity can be readily generated at the nucleic acid level for the *in vitro* expression of up to 10^{13} unique RNA-protein fusions. In theory this would allow the selection of a high-diversity combinatorial library against many uniquely barcoded targets of interest simultaneously. Subsequent next generation sequencing of the selected nucleic acid pool would then enable the identification of unique binding species to every target, all within a single experiment. Recent developments in the generation of immobilised peptide microarrays in next generation sequencing flow cells using ribosome²⁶⁰ and mRNA display²⁶¹ resulted from an increasing demand for high-throughput characterisation of protein-ligand interactions. IDRT-PCR could complement these developments by providing each ligand and target a solution-phase 'interaction fingerprint', encoded in the resulting nucleic acid that could then be delineated by next generation sequencing. Finally, the general strategy described here could be expanded to selections for enzyme activity. Indeed, reactivity-dependent PCR is a well-established method for the selective amplification of nucleic acid encoding ribozymes⁸⁴ and small molecules^{240,262} that undergo bond formation or bond cleavage.

In this chapter, the concept of interaction dependent reverse transcription in a library format was demonstrated for the selective amplification of DNA encoding protein-protein interaction partners. IDRT-PCR was capable of enriching DNA encoding functional SBP peptides approximately 10-fold. The extent of per-round enrichment is likely to depend on the specific interaction of interest, and could likely be improved by further optimisation to reduce non-specific amplification. IDRT-PCR has the potential to amplify rare functional sequences that may typically be lost in washing steps when selecting against immobilised targets, as RT-PCR can theoretically amplify from a single molecule of RNA template. Further development of this technique may help to accelerate the discovery of ligands that bind macromolecular targets of interest for research, diagnostic, and therapeutic applications.

6 Concluding remarks and future directions

6.1 Summary

This thesis has demonstrated some of the successes, and has highlighted some of the challenges, in the application of mRNA display to the *in vitro* selection of novel proteins and enzymes. Implementation and optimisation of mRNA-protein fusion formation, attempts to select for orthogonal PEX5-peptide interactions, and the development of solution-phase, multiplexed selection of ligand-target pairs from libraries of RNA-protein fusions and DNA-tagged targets are presented.

Chapter 1, the introduction to this thesis, presents a small cohort of the vast range of methods available to generate genetic diversity and interrogate protein sequence space in the search for novel or improved proteins and enzymes. Emerging computational design methods are introduced and compared with these laboratory-based strategies. The concept of *in vitro* selection is presented, and current methods for *in vitro* selection and evolution of proteins and enzymes are reviewed to highlight the strengths and capabilities of *in vitro* methods in comparison to *in vivo* methods. The chapter concludes with a summary of selection strategies compatible with mRNA display and of protein scaffolds successfully used in mRNA display selections.

Chapter 3 demonstrates the validation and optimisation of mRNA display techniques for the generation of libraries containing trillions of protein variants. Protocols were established and optimised for the generation and purification of mRNA-protein fusions using DFPase from *L. vulgaris* as a model enzyme. The methodology described in this chapter is broadly applicable to accommodate almost any suitable protein scaffold, provided it can be expressed in an *in vitro* translation system, allowing the potential investigation of multiple protein architectures for complex functionalities. Furthermore, a strategy was devised for the selection of enzymes that catalyse a well characterised bimolecular Diels-Alder reaction^{62,65,66,183,184,263}. The development of these tools will allow future *in vitro* selection for enzymes that catalyse the Diels-Alder reaction via specific immobilisation of sequences on a streptavidin-derivatised solid support.

Chapter 4 describes the work performed towards the selection of novel PEX5-peptide receptor-ligand pairs. A high-diversity library was designed and synthesised based on the PEX5 peroxisomal import receptor from *A. thaliana*. Four rounds of affinity panning of the resulting RNA-protein fusion library for binding to three candidate orthogonal peptides were performed. The outcome of selection outlines some of the challenges

associated with selection for novel activity in the constraints of an existing protein fold. The prospect that the functionality under selection was not present in the initial pool of RNA-protein fusions must be considered, a more intelligent and/or more conservative library design strategy may have resulted in a higher proportion of active PEX5 receptor variants. It also highlights the inherent lack of information obtained during the course of experimental selection and the prescience of the often-quoted First Law of Directed Evolution – ‘You get what you screen for’²⁶⁴.

Chapter 5 outlines work towards the development of solution-phase, multiplexed selection of ligand-target pairs from libraries of RNA-protein fusions and DNA-tagged targets. The concept of interaction-dependent reverse transcription was demonstrated using RNA-streptavidin binding protein (SBP) fusions and DNA-tagged streptavidin. The ability of this strategy to enrich functional proteins from a large excess of non-functional sequences was examined, revealing strong enrichment from a 10-fold background of non-functional sequences. Further development of this technique may help to accelerate the discovery of ligands that bind macromolecular targets of interest for research, diagnostic, and therapeutic applications.

6.2 Future directions

The results discussed in this thesis pave the way toward the detailed interrogation of large high diversity libraries for protein-ligand interactions and novel Diels-Alderase enzymes. A platform now exists for the future application of these powerful *in vitro* selection techniques for the discovery and evolution of new proteins and enzymes. There are a plethora of exciting directions for continuation of these studies:

6.2.1 *In vitro* selection for Diels-Alderase enzymes using mRNA display

Work towards future selection and evolution of Diels-Alder enzymes is presented in this thesis, with the synthesis of selection substrates presented in Chapter 3. However, prior to the commencement of selection experiments, a high-diversity library must also be prepared. This presents a significant undertaking in itself, as the premise of *in vitro* selection dictates that the desired enzyme activity is present in the starting library. This prompts contemplation over what the most appropriate protein scaffold for the selection of Diels-Alderase activity may be. Recent efforts to generate a high-diversity library for *de novo* enzyme selections highlight the importance of not only reducing stop codons and frameshifts¹⁰², but also of increasing the proportion of properly-folded proteins in the starting library¹⁴¹.

Other synthetically useful reactions that would be amenable to an mRNA display selection strategy include Friedel-Crafts reactions – a synthetically valuable family of reactions that removes a halogen from an organic molecule to create a reactive carbocation to form new carbon-carbon bonds²⁶⁵. Another example is Suzuki cross coupling - the palladium-catalysed cross coupling between an organoboronic acid and halide²⁶⁶. A Friedel-Crafts enzyme could potentially be used to form and protect the carbocation intermediate in an aqueous environment, and Suzuki cross coupling enzymes could be discovered with the addition of Pd²⁺ to the selection buffer as a potential co-factor.

6.2.2 Engineering of orthogonal PEX5-peptide interactions

The failure to select for an orthogonal PEX5-peptide interaction using mRNA display in Chapter 4 emphasises the purely exploratory nature of selection experiments. Ongoing efforts to identify the source of artefactual enrichment may shed a light on the mechanisms of background enrichment in *in vitro* selection experiments. This would inform future selection strategies to ensure the highest possible chance of discovering functional proteins. A more rational, screening-based approach to the discovery of orthogonal PEX5-peptide interactions has proved to be more successful, with candidate receptor-ligand pairs showing promise in *in vivo* peroxisomal import assays (Laura Cross, personal communication/thesis).

6.2.3 Interaction-dependent reverse transcription with mRNA display

The principle of interaction dependent reverse transcription (IDRT) PCR was demonstrated in Chapter 5 of this thesis. Prior to its use for the discovery of novel protein-ligand interactions, examination of versatility of the strategy and further optimisation would be required. The ability to select against multiple DNA-tagged targets in one experiment needs to be examined as this is one of the most attractive characteristics of solution-phase selection. More detailed optimisation of reaction conditions may provide incremental improvements in signal:noise ratio. Ultimately, this would facilitate the selection of protein-ligand pairs from libraries of RNA-protein fusions and libraries of DNA-tagged targets in a single experiment.

6.2.4 The future of mRNA display selection

mRNA display remains the only laboratory selection technique to facilitate selection of enzyme activity from a randomised, non-catalytic scaffold⁴². However, despite this significant breakthrough in *de novo* enzyme engineering, no further examples have been

described in almost a decade. This outlines the scale of the challenge of engineering enzymatic activity from scratch in the laboratory. Despite these difficulties, the potential to not only discover *de novo* enzymes but also novel protein folds¹³⁵ separates mRNA display from computational design approaches that currently rely on databases of known protein structures. Indeed, such ‘primordial’ enzymes may also provide insights into the origins and evolution of early protein catalysts. Recent work in the field has focussed on improving the quality of randomised libraries to increase the fraction of folded variants in high-diversity starting libraries¹⁴¹.

In future, *in vitro* compartmentalisation approaches (discussed in Chapter 1) could be combined with mRNA display. This would enable screening of mRNA display libraries rather than selection, thus allowing high-throughput ranking of variants (up to 10^8 per day)²⁶⁷ using automated droplet sorting. This would expand the scope of mRNA display beyond the selection of enzyme activity by immobilisation, and allow the directed evolution of properties such as substrate affinity and multiple turnover.

6.3 Concluding remarks

Artificial proteins and enzymes hold the potential to solve countless challenges in the field of biotechnology. Bespoke binding proteins are invaluable tools for a range of applications spanning therapeutics⁶, *in vivo* imaging⁷, clinical diagnostics⁸, and research²⁶⁸. In addition, new enzymes can provide cheap, chemo-, regio-, stereospecific, and environmentally friendly alternatives for a range of chemical transformations. Promising applications range from bioremediation²⁶⁹ to synthesis of drugs and commodity chemicals¹³⁻¹⁵.

The ultimate goal of the field of artificial enzyme creation is to develop a rapid and reliable process to generate bespoke catalysts that are sufficiently active, selective, and stable to be a viable commercial product¹¹. Despite the advances in understanding of protein structure-function relationships and laboratory evolution techniques, this remains a formidable undertaking. It is becoming clear that – whilst a great deal of progress has been made towards this goal in the last five years^{134,270} – for the foreseeable future the generation of such enzymes *de novo* using a single approach may not be achievable. Indeed, the shortcomings of current *de novo* enzyme engineering strategies are exemplified by the vast disparity in catalytic efficiencies observed for designed enzymes ($10^0 - 10^2 \text{ M}^{-1} \text{ s}^{-1}$)^{60-62,271} compared to those seen in Nature ($10^6 - 10^9 \text{ M}^{-1} \text{ s}^{-1}$)^{55,272}. Rather, the greatest successes to date have come from the combination of *de novo* enzyme engineering (either computationally designed⁶⁰⁻⁶² or selected⁴²) and directed evolution.

In these examples, *de novo* enzymes with initially modest catalytic efficiencies have been improved vastly^{65-67,134}, in one instance rivalling the catalytic efficiency of natural enzymes⁶⁹. Furthermore, lessons learned from directed evolution experiments can inform future strategies for *de novo* enzyme discovery²⁷³.

The selection of novel binding proteins and enzymes by mRNA display is an exciting prospect in the field of biotechnology. The work presented in this thesis has expanded the methods available for mRNA display selection by creating substrate analogues for the selection of Diels-Alderase enzymes and demonstrating the potential for multiplexed solution-phase selection using DNA-tagged targets. Whilst ongoing efforts to select for orthogonal PEX5-peptide interactions highlight that the engineering of poorly understood protein-protein interactions and the interrogation of large libraries *in vitro* remains a significant undertaking. The largest challenges facing the field of *in vitro* protein engineering are matching the pace at which new macromolecular targets of interest are discovered, and diversification into new enzymatic reactions. This would not only provide a source of novel therapeutics, research tools, and industrial catalysts, but would also further demonstrate the capability and value of *in vitro* strategies in the protein engineering toolbox.

7 Appendix

7.1 Chapter 3

7.1.1 DFPase synthetic gene DNA sequence

ATGGAATCCCGGTCATTGAACCGCTGTTTACCAAAGTTACCGAAGACATTCCGGGCGCAGAAG
 GCCCGGTCTTTGACAAAAACGGTGATTTCTATATTGTGGCGCCGGAAGTCGAAGTGAATGGCAA
 ACCGGCCGGTGAAATTCTGCGTATCGACCTGAAAACCGGCAAAAAGACGGTTATTTGCAAGCCG
 GAAGTCAACGGTTACGGTGGTATCCCGGCCGGTTGCCAGTGTGACCGTGACGCGAATCAACTGT
 TCGTGGCCGATATGCGCCTGGGCCTGCTGGTTGTGCAGACCGACGGCACGTTTGAAGAAATTGC
 GAAAAAGGATTCTGAAGGCCGTCGCATGCAAGGTTGCAACGACTGTGCCTTTGATTATGAAGGC
 AATCTGTGGATTACCGCACCGGCCGGTGAAGTTGCACCGGCTGATTATACGCGCAGTATGCAGG
 AAAAAATTTGGCTCCATTTACTGTTTACCACGGACGGTCAGATGATTCAAGTGGACACCGCATT
 TCAGTTCCCGAACGGCATTGCTGTTTCGTTCATATGAATGACGGTCGCCCCGTATCAACTGATTGTC
 GCAGAAACCCCGACGAAAAAGCTGTGGTCATACGATATTAAGGCCCGGCTAAGATTGAAAACA
 AAAAGGTTTGGGGCCATATTCCGGGTACCCACGAAGGCCGGTGCAGACGGTATGGATTTGACGA
 AGATAACAATCTGCTGGTTGCTAACTGGGGCAGCTCTCACATTGAAGTCTTTGGTCCGGATGGC
 GGTGAGCCGAAAAATGCGTATCCGCTGCCCGTTTAAAAACCGTGAATCTGCATTTCAAACCGC
 AGACCAAGACGATTTTTGTGACCGAACACGAAAAACAATGCGGTTTGGAAATTTGAATGGCAACG
 CAATGGTAAAAACAATACTGTGAAACCCTGAAGTTCGGCATCTTTGGTCCCATCACCATCAC
 CATCACTAATAA

7.1.2 DFPase synthetic gene amino acid sequence

MEIPVIEPLFTKVTEDIPGAEGPVFDKNGDFYIVAPEVEVNGKPAGEILRIDLKTGKKTVICKP
 EVNGYGGIPAGCQCDRDANQLFVADMRLGLLVVQTDGTFEEIAKKDSEGRMQGCNDCAFDEG
 NLWITAPAGEVAPADYTRSMQEKFGSIYCFTTDQMIQVDTAFQFPNGIAVRHMNDGRPYQLIV
 AETPTKKLWSYDIKGPAKIENKKVWGHIPGTHEGGADGMDFDENLLVANWGSSEHIEVFGPDG
 GQPKMIRCPFEKPSNLHFKPQTKTIFVTEHENNAVWKFQWQRNGKKQYCETLKFIFGSHHHH
 HH

7.1.3 DFPase 3'-fragment DNA sequence

TCTAATACGACTCACTATAGGTTTGAATGGCAGCGTAATGGTAAAAACAATACTGTGAAACGC
 TGAAATTCGGCATCTTTGGCTCGCACCACCACCACCATCATgCACC GGCTATTA

7.1.4 Primer Sequences

Name	Sequence (5'-3')
DFP Mod For	TCT AAT ACG ACT CAC TAT AGG GAC AAT TAC TAT TTA CAA TTA CAA TGG AAA TCC CGG TCA TC
DFP Mod Rev	TTA ATA GCC GGT GCC ATG ATG GTG GTG GTG GTG
DFP_3'FRAGMENT_FOR	TCT AAT ACG ACT CAC TAT AGG TTT GAA TGG CAG CGT AAT GGT
dA ₁₅	AAA AAA AAA AAA AAA
RT primer	/5AmMC12//iSp18//iSp18/ TTT TTT TTT TTT TTT TTT CCC ATG ATG GTG GTG GTG GTG
RT primer (unmodified)	TTT TTT TTT TTT TTT TTT CCC ATG ATG GTG GTG GTG GTG

/5AmMC12/ = 5' Amino Modifier C12

/iSp18/ = 18-atom hexa-ethyleneglycol spacer

7.2 Chapter 4

7.2.1 PEX5 445-728 DNA sequence[‡]

ATGGGCAGCAGCCATCATCATCATCACAGCAGCGGCCTGGTGGTCTACGTCTTCTCTGACA
TGAATCCTTATGTGGGTCAACCCTGAACCTATGAAAGAAGGGCAAGAATTGTTTCGAAAAGGACT
TCTGAGTGAAGCAGCGCTTGCTCTAGAAGCTGAGGTTATGAAAAACCCTGAGAATGCTGAAGGT
TGGAGATTACTTGGGGTCACACACGCAGAGAAC **GATGATGAT** CAACAGGCAATAGCTGCAATGA
TGCGTGACAGGAGGCTGATCCACAAATCTAGAGGTGCTTCTTGCCTTGGT **GTGAGTCATAC**
CAACGAGTTAGAGCAAGCAACTGCTTTGAAATATCTATATGGATGGCTGCGAAATCACCCAAAG
TATGGAGCAATTGCGCCTCCGGAGCTAGCGGATTCTTTGTACCATGCTGATATTGCTAGATTAT
TCAATGAAGCTTCTCAGTTGAATCCTGAGGACGCCGATGTGCATATAGTGTGGGCGTGCTCTA
CAATCTGTCGAGAGAGTTTCGATAGAGCAATCACATCC **TTC**CAAACAGCATTACAATAAAACCA
AACGATTATTCTCTGTGG **AAT**AAGCTAGGTGCAACGCAAGCC **AAC**AGTGTCCAGAGTGCTGATG
CCATATCTGCTTATCAACAGGCTCTAGATTTAAAACCAAATTATGTTTCGTGCTTGGGCA **AAC**AT
GGGAATC **AGT**TACGCAAACCAGGGGATGTACAAAGAATCAATCCCGTATTATGTCCGTGCCCTT
GCGATGAATCCTAAAGCTGATAACGCATGGCAATACTTGAGACTCTCGTTAAGTTGTGCATCAA
GGCAAGACATGATAGAAGCTTGTGAGTCAAGGAATCTCGATCTCTTGCAGAAAGAATTCCCGCT
GTGA

[‡]Codons randomised in the assembly of the PEX5* library are highlighted.

7.2.2 PEX5 445-728 amino acid sequence

MGSSHHHHHSSGLVVYVFSDMNPYVGHPEPMKEGQELFRKGLLSEAAALAEAEVMKNP
 ENAEGWRLLGVTSHAENDDDQQAI AAMMRAQEADPTNLEVLLALGVSHITNELEQATALKY
 LYGWLNRNHPKYGAIAPPELADSLYHADIARLFNEASQLNPEDADVHIVLGVLYNLSREF
 DRAITSEQTALQLKPN DYSLWNKLGATQAN SVQSADAI SAYQQALDLKPNYVRAWANMG
 ISYANQGM YKESIPYYVRALAMNPKADNAWQYLRLSLSCASRQDMI EACESRNLDLLQK
 EFPL

7.2.3 Primer Sequences

Name	Sequence (5'-3')
PEX5 Mod For	GCC TTC TAA TAC GAC TCA CTA TAG GGA CAA TTA CTA TTT ACA ATT ACA ATG GGC AGC AGC CAT CAT CAT CAT C
PEX5 Mod Rev	TTA ATA GCC GGT GCC AGA TCC AGA CAT TCC CAT CAG CGG GAA TTC TTT CTG CAA GAG ATC G
PEX5A For	ATG GGC AGC AGC CAT CAT
PEX5A Rev	AAA AAA GGT CTC ACT GTT GSN NAT CSN NGT TCT CTG CGT GTG TGA CC
PEX5B For	TTT TTT GGT CTC TAC AGG CAA TAG CTG CAA TGA TG
PEX5B Rev	AAA AAA GGT CTC AGC TCT AAS NNS NNS NNA TGA CTS NNA CCA AGC GCA AGA AGC AC
PEX5C For	TTT TTT GGT CTC TGA GCA AGC AAC TGC TTT GAA ATA TC
PEX5C Rev	AAA AAA GGT CTC AGT AGA GCA CGC CCA ACA CTA TAT G
PEX5D For	TTT TTT GGT CTC TCT CAN NSC TGT CGA GAG AGT TCG ATA GAG CAA TCA CAT CCN NSC AAA CAG CAT TAC AAC TAA AAC CAA ACG AT
PEX5D Rev	AAA AAA GGT CTC AGG CTT GCG TTG CAC CTA GCT TSN NCC ACA GAG AAT AAT CGT TTG GTT TTA GTT GTA ATG CTG TTT G
PEX5E For	TTT TTT GGT CTC TAG CCN NSA GTG TCC AGA GTG CTG ATG C
PEX5E Rev	AAA AAA GGT CTC ATG CCC AAG CAC GAA CAT AAT TTG
PEX5F For	TTT TTT GGT CTC TGG CAN NSA TGG GAA TCN NST ACG CAA ACC AGG GGA TGT AC
PEX5F Rev	TCACAGCGGGAATTCTTTCTGC

7.2.4 PEX5* Round 0 Sequences

```

10 110 120 130 140 150 160 170 180 190 200
PEX5* MGSSEHHHSSGLVVVFSDMNPFVGHPEPMKEGQELFRKGLLSEAAALAEAEVMKNPENAEKRWLLLVTEANDDDQQAIAAMRAQEAADPTNLEVLL
1  N.H
2  L.S
3  L.*
4  Q.P
5  P.L
6  P.P
7  P.N
8  L.H
9  P.P
10

110 120 130 140 150 160 170 180 190 200
PEX5* ALGVSHNELEQATALKYLWLEPHKYGAIAPPLEADSLYHADLARLFNEASQLNPEADVHIVLGVLYNLSREDFDRAITSFQALQKPNIDYSLWNK
1  P..ATQ
2  A..LAP
3  A..SVL
4  P..L*P
5  T..PSP
6  S..PSS
7  F..QAR
8  R..GI
9  F..CPA
10 N..RRR

210 220 230 240 250 260 270 280 290 300
PEX5* LGATQANSVQSADAISAYQOALDKPNYVRAWANMGIISYANQGMYKESIPYVVRALAMNPKADNAWQYLRILSLSCASRODMIEACEERNLDLQKEFPL
1  A..
2  W..
3  C..
4  G..
5  V..
6  A..
7  K..
8  G..
9  A..
10 W..

```

7.2.5 PEX5* Round 4 Sequences

Target	Clone	Amino acid sequence at randomised positions											
YQSEV	EV.4.1	L	L	P	L	P	P	R	A	R	A	G	W
	EV.4.2	P	L	S	S	P	S	G	G	R	A	S	C
	EV.4.3	P	L	S	S	P	S	R	A	R	A	G	W
	EV.4.4	P	Y	M	P	W	P	V	V	T	G	S	A
	EV.4.5	P	L	S	S	P	S	G	G	R	A	S	C
	EV.4.6	P	L	S	S	P	S	G	G	R	A	S	C
	EV.4.7	P	L	S	S	P	S	G	G	R	A	S	C
	EV.4.8	P	P	T	P	R	H	V	V	T	G	S	A
	EV.4.9	P	L	S	S	P	S	G	G	R	A	S	C
	EV.4.10	L	L	P	L	P	P	R	A	R	A	G	W
YQSFY	FY.4.1	P	L	S	S	P	S	G	G	R	A	E	G
	FY.4.2	P	P	F	Q	A	R	R	A	R	A	E	G
	FY.4.3	P	L	S	S	P	S	G	G	R	A	E	G
	FY.4.4	P	L	S	S	P	S	G	G	R	A	E	G
	FY.4.5	P	T	P	A	T	Q	K	F	T	G	S	A
	FY.4.6	P	L	S	S	P	S	G	G	R	A	E	G
	FY.4.7	P	L	S	S	P	S	G	G	R	G	S	A
	FY.4.8	P	L	S	S	P	S	G	G	R	G	S	A
	FY.4.9	P	P	F	Q	A	R	R	A	R	G	S	A
	FY.4.10	P	T	P	A	T	Q	K	F	T	G	S	A
YQSY Y	YY.4.1	P	T	P	A	T	Q	R	A	R	A	E	G
	YY.4.2	P	L	S	S	P	S	G	G	R	A	S	C
	YY.4.3	P	L	S	S	P	S	G	G	R	A	S	C
	YY.4.4	P	L	S	S	P	S	G	G	T	G	W	W
	YY.4.5	P	L	S	S	P	S	G	G	R	A	S	C
	YY.4.6	P	L	S	S	P	S	G	G	R	A	S	C
	YY.4.7	P	L	S	S	P	S	G	G	R	A	S	C
	YY.4.8	P	L	S	S	P	S	R	A	R	A	E	G
	YY.4.9	P	T	P	A	T	Q	R	A	R	A	E	G
	YY.4.10	P	L	S	S	P	S	G	G	T	G	W	W
Wild Type PEX5		D	D	V	T	N	E	N	F	N	N	N	S

7.3 Chapter 4 - supplementary figures

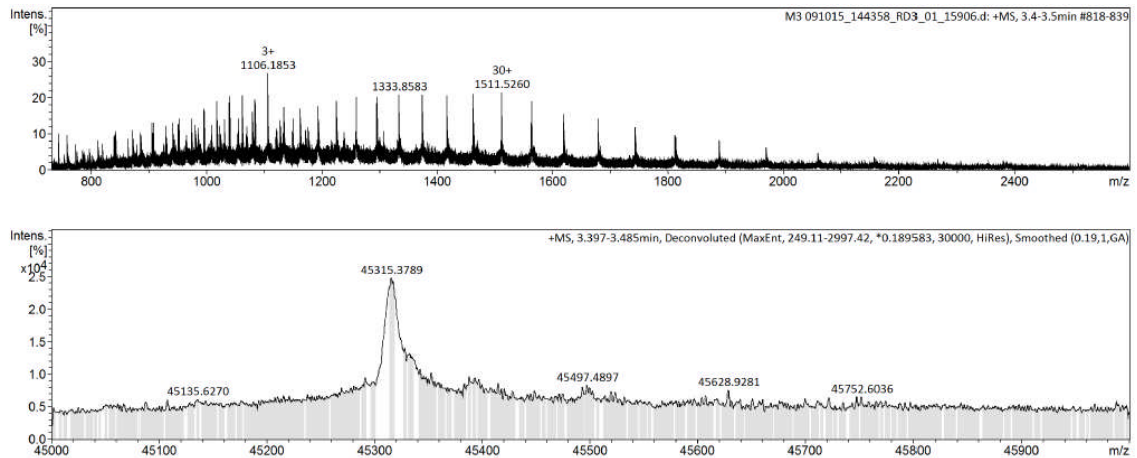


Figure 7.3.1. Mass spectrometry analysis of the purified PEX5.YY.4.3. Positive ESI-MS m/z spectrum and molecular mass profile indicates a molecular mass of 45,315.38 Da which agrees well with the calculated molecular mass of 45,316.2 Da.

7.4 Chapter 5 DNA sequences

7.4.1 Streptavidin binding peptide (SBP) DNA sequence*

GCCTTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGGCATGTTGTCTCG
 AGGGTTTGGCTGGGGAGTTGGAGCAGCTCCGGGCACGCTTGAACACCACCCCGGGTATGGG
 AATGTCTGGATCTGGCACCCGGCTATTAA

*5'- and 3'-sequence required for mRNA display is highlighted in green.

7.4.2 SBP amino acid sequence

MGHVVEGLAGELEQLRARLEHHPQGMMSGSGTGY

7.4.3 SBP Δ DNA sequence*

GCCTTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGGCATGTTGTCTCG
 AGGGTTTGGCTGGGGAGCTCGAGCAGCTCCGGGCACGCTTGAACACCACGGCGCCGGTATGGG
 AATGTCTGGATCTGGCACCCGGCTATTAA

7.4.4 SBP Δ amino acid sequence

MGHVVEGLAGELEQLRARLEHHGAGMMSGSGTGY

7.4.5 Chapter 5 Primer Sequences

Name	Sequence (5'-3')
SBP Mod For	GCC TTC TAA TAC GAC TCA CTA TAG GGA CAA TTA CTA TTT ACA ATT ACA ATG G
SBP 6 Mod Rev	TTA ATA GCC GGT GCC AGA TCC AGA CAT TCC CAT ACC
SBP 8 Mod Rev	TTA ATA GCC GGT GCC AGT ACC AGA CAT TCC CAT ACC
SBP 10 Mod Rev	TTA ATA GCC GGT GCC ATT ACC AGA CAT TCC CAT ACC
IDRT-PCR Rev	GCG ATA GAC TAC CTC CTA AT
IDRT Target oligo	/5ThioMC6-D/ /iSp18/ /iSp18/ CGC TAT CTG ATG GAG GAT TAC CAG AC

/5ThioMC6-D/ = 5' Thiol Modifier C6 S-S.

/iSp18/ = 18-atom hexa-ethyleneglycol spacer.

7.5 Chapter 5 - supplementary figures

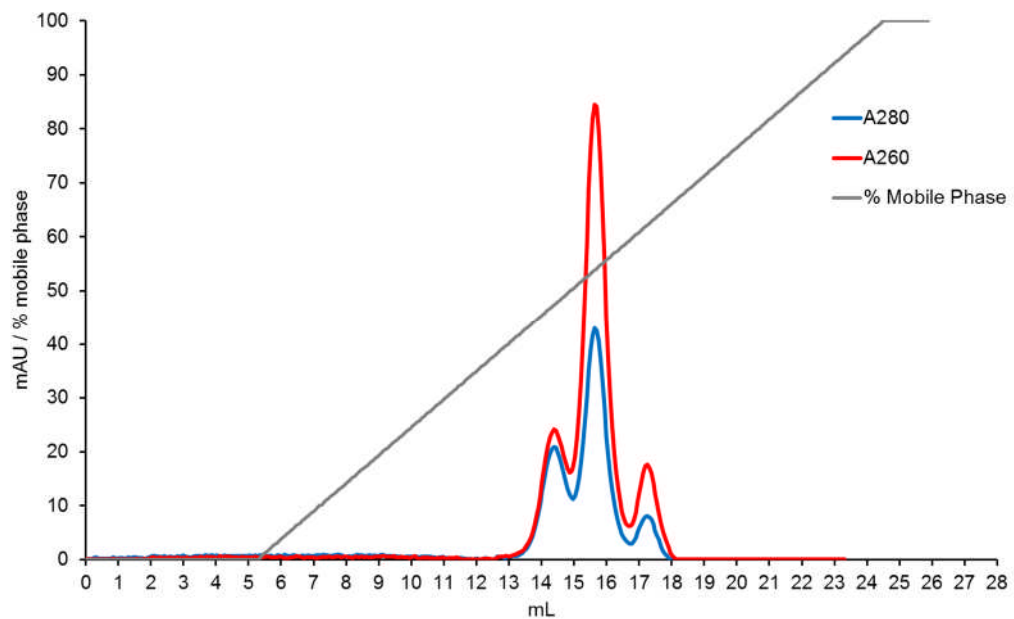


Figure 7.4.1. Elution profile of streptavidin-oligonucleotide conjugate purification by anion-exchange chromatography. The presence of multiple peaks suggests that the labelling reaction generated DNA:streptavidin conjugates of varying stoichiometry.

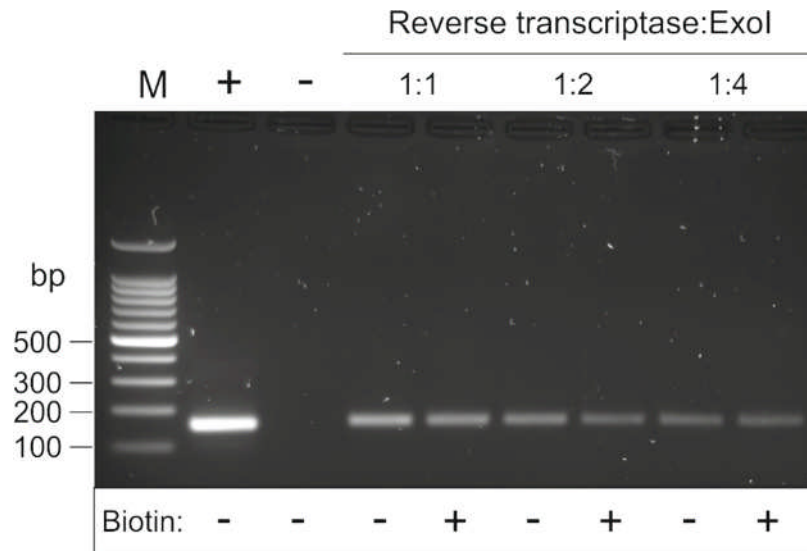


Figure 7.4.2. Agarose gel analysis of exonuclease I optimisation in the IDRT-PCR with a 10-nucleotide hybridisation region. Lane 1, Positive control RT-PCR with gene specific primer; lane 2, negative control RT-PCR; lanes 3-4, IDRT-PCR with $0.5 \text{ U.}\mu\text{L}^{-1}$ exonuclease I; lanes 5-6, IDRT-PCR with $1 \text{ U.}\mu\text{L}^{-1}$ exonuclease I; lanes 7-8, IDRT-PCR with $2 \text{ U.}\mu\text{L}^{-1}$ exonuclease I. M, 100 base pair ladder.

References

1. Nimmerjahn, F. and J.V. Ravetch. Antibody-mediated modulation of immune responses. *Immunol Rev*, 2010, **236**, 265-275.
2. Hunter, T. Signaling - 2000 and beyond. *Cell*, **100**, 113-127.
3. Chames, P., M. Van Regenmortel, E. Weiss and D. Baty. Therapeutic antibodies: successes, limitations and hopes for the future. *Brit J Pharmacol*, 2009, **157**, 220-233.
4. Vazquez-Rey, M. and D.A. Lang. Aggregates in monoclonal antibody manufacturing processes. *Biotechnol Bioeng*, 2011, **108**, 1494-1508.
5. Hosse, R.J., A. Rothe and B.E. Power. A new generation of protein display scaffolds for molecular recognition. *Protein Sci*, 2006, **15**, 14-27.
6. Buss, N.a.P.S., S.J. Henderson, M. Mcfarlane, J.M. Shenton and L. De Haan. Monoclonal antibody therapeutics: history and future. *Curr Opin Pharmacol*, 2012, **12**, 615-622.
7. Baum, R.P., V. Prasad, D. Muller, C. Schuchardt, A. Orlova, A. Wennborg, V. Tolmachev and J. Feldwisch. Molecular imaging of *HER2*-expressing malignant tumors in breast cancer patients using synthetic ^{111}In - or ^{68}Ga -labeled affibody molecules. *J Nucl Med*, 2010, **51**, 892-897.
8. Ko Ferrigno, P. Non-antibody protein-based biosensors. *Essays Biochem*, 2016, **60**, 19-25.
9. Walsh, C. Enabling the chemistry of life. *Nature*, 2001, **409**, 226-231.
10. Voet, D. and J.G. Voet. *Biochemistry*. Wiley, 2004.
11. Bornscheuer, U.T., G.W. Huisman, R.J. Kazlauskas, S. Lutz, J.C. Moore and K. Robins. Engineering the third wave of biocatalysis. *Nature*, 2012, **485**, 185-194.
12. Koeller, K.M. and C.-H. Wong. Enzymes for chemical synthesis. *Nature*, 2001, **409**, 232-240.
13. Savile, C.K., J.M. Janey, E.C. Mundorff, J.C. Moore, S. Tam, W.R. Jarvis, J.C. Colbeck, A. Krebber, F.J. Fleitz, J. Brands, P.N. Devine, G.W. Huisman and G.J. Hughes. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, 2010, **329**, 305-309.
14. Carroll, A. and C. Somerville. Cellulosic biofuels. *Annu Rev Plant Biol*, 2009, **60**, 165-182.
15. Shen, L., E. Worrell and M. Patel. Present and future development in plastics from biomass. *Biofuel Bioprod Bioref*, 2010, **4**, 25-40.
16. Pinheiro, V.B., A.I. Taylor, C. Cozens, M. Abramov, M. Renders, S. Zhang, J.C. Chaput, J. Wengel, S.Y. Peak-Chew, S.H. Mclaughlin, P. Herdewijn and P. Holliger. Synthetic genetic polymers capable of heredity and evolution. *Science*, 2012, **336**, 341-344.
17. Ramsay, N., A.S. Jemth, A. Brown, N. Crampton, P. Dear and P. Holliger. CyDNA: synthesis and replication of highly Cy-dye substituted DNA by an evolved polymerase. *J Am Chem Soc*, 2010, **132**, 5096-5104.
18. Wright, S.I., I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen and B.S. Gaut. The effects of artificial selection on the maize genome. *Science*, 2005, **308**, 1310-1314.

19. Driscoll, C.A., D.W. Macdonald and S.J. O'Brien. From wild animals to domestic pets, an evolutionary view of domestication. *Proc Natl Acad Sci USA*, 2009, **106**, 9971-9978.
20. Taylor, S.V., K.U. Walter, P. Kast and D. Hilvert. Searching sequence space for protein catalysts. *Proc Natl Acad Sci USA*, 2001, **98**, 10596-10601.
21. Mandecki, W. The game of chess and searches in protein sequence space. *Trends Biotechnol*, **16**, 200-202.
22. Romero, P.A. and F.H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol*, 2009, **10**, 866-876.
23. Smith, J.M. Natural selection and the concept of a protein space. *Nature*, 1970, **225**, 563-564.
24. Gavrillets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol*, 1997, **12**, 307-312.
25. Lai, Y.P., J. Huang, L.F. Wang, J. Li and Z.R. Wu. A new approach to random mutagenesis *in vitro*. *Biotechnol Bioeng*, 2004, **86**, 622-627.
26. Myers, R.M., L.S. Lerman and T. Maniatis. A general method for saturation mutagenesis of cloned DNA fragments. *Science*, 1985, **229**, 242-247.
27. Freese, E. The specific mutagenic effect of base analogues on phage T4. *J Mol Biol*, 1959, **1**, 87-105.
28. Bridges, B.A. and R. Woodgate. Mutagenic repair in *Escherichia coli*: products of the *recA* gene and of the *umuD* and *umuC* genes act at different steps in UV-induced mutagenesis. *Proc Natl Acad Sci USA*, 1985, **82**, 4193-4197.
29. Cox, E.C. Bacterial mutator genes and the control of spontaneous mutation. *Annu Rev Genet*, 1976, **10**, 135-156.
30. Greener, A., M. Callahan and B. Jerpseth. An efficient random mutagenesis technique using an *E. coli* mutator strain. *Mol Biotechnol*, 1997, **7**, 189-195.
31. Leung, D.W., E. Chen and D.V. Goeddel. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique*, 1989, **1**, 11-15.
32. Zaccolo, M., D.M. Williams, D.M. Brown and E. Gherardi. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J Mol Biol*, 1996, **255**, 589-603.
33. Eckert, K.A. and T.A. Kunkel. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Nucleic Acids Res*, 1990, **18**, 3739-3744.
34. Biles, B.D. and B.A. Connolly. Low-fidelity *Pyrococcus furiosus* DNA polymerase mutants useful in error-prone PCR. *Nucleic Acids Res*, 2004, **32**, e176.
35. Vanhercke, T., C. Ampe, L. Tirry and P. Denolf. Reducing mutational bias in random protein libraries. *Anal Biochem*, 2005, **339**, 9-14.
36. Cadwell, R.C. and G.F. Joyce. Randomization of genes by PCR mutagenesis. *PCR Meth Appl*, 1992, **2**, 28-33.
37. Patrick, W.M. and A.E. Firth. Strategies and computational tools for improving randomized protein libraries. *Biomol Eng*, 2005, **22**, 105-112.
38. Stemmer, W.P. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, 1994, **370**, 389-391.
39. Ness, J.E., S. Kim, A. Gottman, R. Pak, A. Krebber, T.V. Borchert, S. Govindarajan, E.C. Mundorff and J. Minshull. Synthetic shuffling expands functional protein

- diversity by allowing amino acids to recombine independently. *Nat Biotechnol*, 2002, **20**, 1251-1255.
40. Stemmer, W.P., A. Cramer, K.D. Ha, T.M. Brennan and H.L. Heyneker. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, 1995, **164**, 49-53.
 41. Zha, D., A. Eipper and M.T. Reetz. Assembly of designed oligonucleotides as an efficient method for gene recombination: a new tool in directed evolution. *Chembiochem*, 2003, **4**, 34-39.
 42. Seelig, B. and J.W. Szostak. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature*, 2007, **448**, 828-831.
 43. Xiao, H., Z. Bao and H. Zhao. High throughput screening and selection methods for directed enzyme evolution. *Ind Eng Chem Res*, 2015, **54**, 4011-4020.
 44. Leemhuis, H., V. Stein, A.D. Griffiths and F. Hollfelder. New genotype–phenotype linkages for directed evolution of functional proteins. *Curr Opin Struct Biol*, 2005, **15**, 472-478.
 45. Williams, G.J., A.S. Nelson and A. Berry. Directed evolution of enzymes for biocatalysis and the life sciences. *Cell Mol Life Sci*, 2004, **61**, 3034-3046.
 46. Arnold, F.H. and G. Georgiou. *Directed enzyme evolution: screening and selection methods*. Humana Press, 2003.
 47. Bottcher, D. and U.T. Bornscheuer. High-throughput screening of activity and enantioselectivity of esterases. *Nat Protoc*, 2006, **1**, 2340-2343.
 48. Taylor, S.V., P. Kast and D. Hilvert. Investigating and engineering enzymes by genetic selection. *Angew Chem Int Ed*, 2001, **40**, 3310-3335.
 49. Cotten, S.W., J. Zou, C.A. Valencia and R. Liu. Selection of proteins with desired properties from natural proteome libraries using mRNA display. *Nat Protoc*, 2011, **6**, 1163-1182.
 50. Kehoe, J.W. and B.K. Kay. Filamentous phage display in the new millennium. *Chem Rev*, 2005, **105**, 4056-4072.
 51. Cho, G.S. and J.W. Szostak. Directed evolution of ATP binding proteins from a zinc finger domain by using mRNA display. *Chem Biol*, 2006, **13**, 139-147.
 52. Sidhu, S.S., H.B. Lowman, B.C. Cunningham and J.A. Wells. Phage display for selection of novel binding peptides. In: S.D.E. JEREMY THORNER and N.A. JOHN, eds. *Method Enzymol*. Academic Press, 2000. 333-IN335.
 53. Keefe, A.D. and J.W. Szostak. Functional proteins from a random-sequence library. *Nature*, 2001, **410**, 715-718.
 54. Pauling, L. Chemical achievement and hope for the future. *Am Sci*, 1948, **36**, 51-58.
 55. Hilvert, D. Critical analysis of antibody catalysis. *Annu Rev Biochem*, 2000, **69**, 751-793.
 56. Jencks, W.P. *Catalysis in chemistry and enzymology*. McGraw-Hill series in advanced chemistry. New York: McGraw-Hill, 1969.
 57. Pollack, S.J., J.W. Jacobs and P.G. Schultz. Selective chemical catalysis by an antibody. *Science*, 1986, **234**, 1570-1573.
 58. Tramontano, A., K.D. Janda and R.A. Lerner. Catalytic antibodies. *Science*, 1986, **234**, 1566-1570.
 59. Nanda, V. and R.L. Koder. Designing artificial enzymes by intuition and computation. *Nat Chem*, 2010, **2**, 15-24.

60. Rothlisberger, D., O. Khersonsky, A.M. Wollacott, L. Jiang, J. Dechancie, J. Betker, J.L. Gallaher, E.A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K.N. Houk, D.S. Tawfik and D. Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 2008, **453**, 190-195.
61. Jiang, L., E.A. Althoff, F.R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J.L. Gallaher, J.L. Betker, F. Tanaka, C.F. Barbas, D. Hilvert, K.N. Houk, B.L. Stoddard and D. Baker. *De novo* computational design of retro-aldol enzymes. *Science*, 2008, **319**, 1387-1391.
62. Siegel, J.B., A. Zanghellini, H.M. Lovick, G. Kiss, A.R. Lambert, J.L. St.Clair, J.L. Gallaher, D. Hilvert, M.H. Gelb, B.L. Stoddard, K.N. Houk, F.E. Michael and D. Baker. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, 2010, **329**, 309-313.
63. Kries, H., R. Blomberg and D. Hilvert. *De novo* enzymes by computational design. *Curr Opin Chem Biol*, 2013, **17**, 221-228.
64. Diels, O. and K. Alder. Synthesis in the hydro-aromatic tier. *Liebigs Ann Chem*, 1928, **460**, 98-122.
65. Eiben, C.B., J.B. Siegel, J.B. Bale, S. Cooper, F. Khatib, B.W. Shen, F. Players, B.L. Stoddard, Z. Popovic and D. Baker. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat Biotechnol*, 2012, **30**, 190-192.
66. Preiswerk, N., T. Beck, J.D. Schulz, P. Milovnik, C. Mayer, J.B. Siegel, D. Baker and D. Hilvert. Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc Natl Acad Sci USA*, 2014, **111**, 8013-8018.
67. Khersonsky, O., G. Kiss, D. Rothlisberger, O. Dym, S. Albeck, K.N. Houk, D. Baker and D.S. Tawfik. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci USA*, 2012, **109**, 10358-10363.
68. Giger, L., S. Caner, R. Obexer, P. Kast, D. Baker, N. Ban and D. Hilvert. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat Chem Biol*, 2013, **9**, 494-498.
69. Obexer, R., A. Godina, X. Garrabou, P.R.E. Mittl, D. Baker, A.D. Griffiths and D. Hilvert. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat Chem*, 2016, Advance online publication. doi:10.1038/nchem.2596.
70. Spiegelman, S. An approach to the experimental analysis of precellular evolution. *Q Rev Biophys*, 1971, **4**, 213-253.
71. Saiki, R.K., S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich and N. Arnheim. Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 1985, **230**, 1350-1354.
72. Ellington, A.D. and J.W. Szostak. *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, 1990, **346**, 818-822.
73. Szostak, J.W. *In vitro* genetics. *Trends Biochem Sci*, 1992, **17**, 89-93.
74. Tuerk, C. and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 1990, **249**, 505-510.
75. Ciesiolka, J. and M. Yarus. Small RNA-divalent domains. *RNA*, 1996, **2**, 785-793.

76. Morris, K.N., K.B. Jensen, C.M. Julin, M. Weil and L. Gold. High affinity ligands from in vitro selection: complex targets. *Proc Natl Acad Sci USA*, 1998, **95**, 2902-2907.
77. Nieuwlandt, D., M. Wecker and L. Gold. In vitro selection of RNA ligands to substance P. *Biochemistry*, 1995, **34**, 5651-5659.
78. Pan, W., R.C. Craven, Q. Qiu, C.B. Wilson, J.W. Wills, S. Golovine and J.F. Wang. Isolation of virus-neutralizing RNAs from a large pool of random sequences. *Proc Natl Acad Sci USA*, 1995, **92**, 11509-11513.
79. Ringquist, S., T. Jones, E.E. Snyder, T. Gibson, I. Boni and L. Gold. High-affinity RNA ligands to Escherichia coli ribosomes and ribosomal protein S1: comparison of natural and unnatural binding sites. *Biochemistry*, 1995, **34**, 3640-3648.
80. Baskerville, S. and D.P. Bartel. A ribozyme that ligates RNA to protein. *Proc Natl Acad Sci USA*, 2002, **99**, 9154-9159.
81. Seelig, B., S. Keiper, F. Stuhlmann and A. Jäschke. Enantioselective ribozyme catalysis of a bimolecular cycloaddition reaction. *Angew Chem Int Ed*, 2000, **39**, 4576-4579.
82. Chapman, K.B. and J.W. Szostak. Isolation of a ribozyme with 5'-5' ligase activity. *Chem Biol*, 1995, **2**, 325-333.
83. Lehman, N. and G.F. Joyce. Evolution in vitro of an RNA enzyme with altered metal dependence. *Nature*, 1993, **361**, 182-185.
84. Bartel, D. and J. Szostak. Isolation of new ribozymes from a large pool of random sequences. *Science*, 1993, **261**, 1411-1418.
85. Green, R. and J.W. Szostak. Selection of a ribozyme that functions as a superior template in a self-copying reaction. *Science*, 1992, **258**, 1910-1915.
86. Beaudry, A.A. and G.F. Joyce. Directed evolution of an RNA enzyme. *Science*, 1992, **257**, 635-641.
87. Bieberich, E., D. Kapitonov, T. Tencomnao and R.K. Yu. Protein-ribosome-mRNA display: affinity isolation of enzyme-ribosome-mRNA complexes and cDNA cloning in a single-tube reaction. *Anal Biochem*, 2000, **287**, 294-298.
88. Amstutz, P., J.N. Pelletier, A. Guggisberg, L. Jermutus, S. Cesaro-Tadic, C. Zahnd and A. Pluckthun. In vitro selection for catalytic activity with ribosome display. *J Am Chem Soc*, 2002, **124**, 9396-9403.
89. Takahashi, F., T. Ebihara, M. Mie, Y. Yanagida, Y. Endo, E. Kobatake and M. Aizawa. Ribosome display for selection of active dihydrofolate reductase mutants using immobilized methotrexate on agarose beads. *FEBS Lett*, 2002, **514**, 106-110.
90. Odegrip, R., D. Coomber, B. Eldridge, R. Hederer, P.A. Kuhlman, C. Ullman, K. Fitzgerald and D. Mcgregor. CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc Natl Acad Sci USA*, 2004, **101**, 2806-2810.
91. Reiersen, H., I. Lobersli, G.A. Loset, E. Hvattum, B. Simonsen, J.E. Stacy, D. Mcgregor, K. Fitzgerald, M. Welschof, O.H. Brekke and O.J. Marvik. Covalent antibody display--an in vitro antibody-DNA library selection system. *Nucleic Acids Res*, 2005, **33**, e10.
92. Cohen, H.M., D.S. Tawfik and A.D. Griffiths. Altering the sequence specificity of *HaeIII* methyltransferase by directed evolution using in vitro compartmentalization. *Protein Eng Des Sel*, 2004, **17**, 3-11.

93. Doi, N., S. Kumadaki, Y. Oishi, N. Matsumura and H. Yanagawa. *In vitro* selection of restriction endonucleases by *in vitro* compartmentalization. *Nucleic Acids Res*, 2004, **32**, e95.
94. Mastrobattista, E., V. Taly, E. Chanudet, P. Treacy, B.T. Kelly and A.D. Griffiths. High-throughput screening of enzyme libraries: *in vitro* evolution of a beta-galactosidase by fluorescence-activated sorting of double emulsions. *Chem Biol*, 2005, **12**, 1291-1300.
95. Griffiths, A.D. and D.S. Tawfik. Directed evolution of an extremely fast phosphotriesterase by *in vitro* compartmentalization. *EMBO J*, 2003, **22**, 24-35.
96. Kelly, B.T. and A.D. Griffiths. Selective gene amplification. *Protein Eng Des Sel*, 2007, **20**, 577-581.
97. Sumida, T., N. Doi and H. Yanagawa. Bicistronic DNA display for *in vitro* selection of Fab fragments. *Nucleic Acids Res*, 2009, **37**, e147.
98. Barrick, J.E., T.T. Takahashi, A. Balakin and R.W. Roberts. Selection of RNA-binding peptides using mRNA-peptide fusions. *Methods*, 2001, **23**, 287-293.
99. Wilson, D.S. and J.W. Szostak. *In vitro* selection of functional nucleic acids. *Annu Rev Biochem*, 1999, **68**, 611-647.
100. Baggio, R., P. Burgstaller, S.P. Hale, A.R. Putney, M. Lane, D. Lipovsek, M.C. Wright, R.W. Roberts, R. Liu, J.W. Szostak and R.W. Wagner. Identification of epitope-like consensus motifs using mRNA display. *J Mol Recognit*, 2002, **15**, 126-134.
101. Mcpherson, M., Y. Yang, P.W. Hammond and B.L. Kreider. Drug receptor identification from multiple tissues using cellular-derived mRNA display libraries. *Chem Biol*, 2002, **9**, 691-698.
102. Cho, G., A.D. Keefe, R. Liu, D.S. Wilson and J.W. Szostak. Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J Mol Biol*, 2000, **297**, 309-319.
103. Smith, G.P. and V.A. Petrenko. Phage display. *Chem Rev*, 1997, **97**, 391-410.
104. Tawfik, D.S. and A.D. Griffiths. Man-made cell-like compartments for molecular evolution. *Nat Biotechnol*, 1998, **16**, 652-656.
105. Fields, S. and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 1989, **340**, 245-246.
106. Boder, E.T. and K.D. Wittrup. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol*, 1997, **15**, 553-557.
107. Arnold, F.H. and A.A. Volkov. Directed evolution of biocatalysts. *Curr Opin Chem Biol*, 1999, **3**, 54-59.
108. Li, S., S. Millward and R. Roberts. *In vitro* selection of mRNA display libraries containing an unnatural amino acid. *J Am Chem Soc*, 2002, **124**, 9972-9973.
109. Schechter, I. 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*, 1973, **70**, 2256-2260.
110. Hanes, J. and A. Plückthun. *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci USA*, 1997, **94**, 4937-4942.
111. Mattheakis, L.C., R.R. Bhatt and W.J. Dower. An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc Natl Acad Sci USA*, 1994, **91**, 9022-9026.

112. Plückthun, A. Ribosome Display: A Perspective. *In*: A.J. DOUTHWAITE and H.R. JACKSON, eds. *Ribosome Display and Related Technologies: Methods and Protocols*. New York, NY: Springer New York, 2012. 3-28.
113. Takahashi, F., H. Funabashi, M. Mie, Y. Endo, T. Sawasaki, M. Aizawa and E. Kobatake. Activity-based *in vitro* selection of T4 DNA ligase. *Biochem Biophys Res Commun*, 2005, **336**, 987-993.
114. Miller, O.J., K. Bernath, J.J. Agresti, G. Amitai, B.T. Kelly, E. Mastrobattista, V. Taly, S. Magdassi, D.S. Tawfik and A.D. Griffiths. Directed evolution by *in vitro* compartmentalization. *Nature methods*, 2006, **3**, 561-570.
115. Song, H. and R.F. Ismagilov. Millisecond kinetics on a microfluidic chip using nanoliters of reagents. *J Am Chem Soc*, 2003, **125**, 14613-14619.
116. Cull, M.G., J.F. Miller and P.J. Schatz. Screening for receptor ligands using large libraries of peptides linked to the C terminus of the *lac* repressor. *Proc Natl Acad Sci USA*, 1992, **89**, 1865-1869.
117. Speight, R.E., D.J. Hart, J.D. Sutherland and J.M. Blackburn. A new plasmid display technology for the *in vitro* selection of functional phenotype-genotype linked proteins. *Chem Biol*, 2001, **8**, 951-965.
118. Stein, V., I. Sielaff, K. Johnsson and F. Hollfelder. A covalent chemical genotype-phenotype linkage for *in vitro* protein evolution. *Chembiochem*, 2007, **8**, 2191-2194.
119. Diamante, L., P. Gatti-Lafranconi, Y. Schaerli and F. Hollfelder. *In vitro* affinity screening of protein and peptide binders by megavalent bead surface display. *Protein Eng Des Sel*, 2013, **26**, 713-724.
120. Doi, N. and H. Yanagawa. STABLE: protein-DNA fusion system for screening of combinatorial protein libraries *in vitro*. *FEBS Lett*, 1999, **457**, 227-230.
121. Kaltenbach, M., V. Stein and F. Hollfelder. SNAP dendrimers: multivalent protein display on dendrimer-like DNA for directed evolution. *Chembiochem*, 2011, **12**, 2208-2216.
122. Bertschinger, J., D. Grabulovski and D. Neri. Selection of single domain binding proteins by covalent DNA display. *Protein Eng Des Sel*, 2007, **20**, 57-68.
123. Bertschinger, J. and D. Neri. Covalent DNA display as a novel tool for directed evolution of proteins *in vitro*. *Protein Eng Des Sel*, 2004, **17**, 699-707.
124. Nemoto, N., E. Miyamoto-Sato, Y. Husimi and H. Yanagawa. *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*, 1997, **414**, 405-408.
125. Roberts, R.W. and J.W. Szostak. RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc Natl Acad Sci USA*, 1997, **94**, 12297-12302.
126. Liu, R., J.E. Barrick, J.W. Szostak and R.W. Roberts. Optimized synthesis of RNA-protein fusions for *in vitro* protein selection. *Method Enzymol*, 2000, **318**, 268-293.
127. Hammond, P.W., J. Alpin, C.E. Rise, M. Wright and B.L. Kreider. *In vitro* selection and characterization of Bcl-X_L-binding proteins from a mix of tissue-specific mRNA display libraries. *J Biol Chem*, 2001, **276**, 20898-20906.
128. Takahashi, T.T., R.J. Austin and R.W. Roberts. mRNA display: ligand discovery, interaction analysis and beyond. *Trends Biochem Sci*, 2003, **28**, 159-165.

129. Seelig, B. mRNA display for the selection and evolution of enzymes from *in vitro*-translated protein libraries. *Nat Protoc*, 2011, **6**, 540-552.
130. Wilson, D.S., A.D. Keefe and J.W. Szostak. The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl Acad. Sci. USA*, 2001, **98**, 3750-3755.
131. Nygren, P.A. and M. Uhlen. Scaffolds for engineering novel binding sites in proteins. *Curr Opin Struct Biol*, 1997, **7**, 463-469.
132. Xu, L., P. Aha, K. Gu, R.G. Kuimelis, M. Kurz, T. Lam, A.C. Lim, H. Liu, P.A. Lohse, L. Sun, S. Weng, R.W. Wagner and D. Lipovsek. Directed evolution of high-affinity antibody mimics using mRNA display. *Chem Biol*, 2002, **9**, 933-942.
133. Getmanova, E.V., Y. Chen, L. Bloom, J. Gokemeijer, S. Shamah, V. Warikoo, J. Wang, V. Ling and L. Sun. Antagonists to human and mouse vascular endothelial growth factor receptor 2 generated by directed protein evolution *in vitro*. *Chem Biol*, 2006, **13**, 549-556.
134. Morelli, A., J. Haugner and B. Seelig. Thermostable artificial enzyme isolated by *in vitro* selection. *PLoS One*, 2014, **9**, e112028.
135. Chao, F.A., A. Morelli, J.C. Iij, L. Churchfield, L.N. Hagmann, L. Shi, L.R. Masterson, R. Sarangi, G. Veglia and B. Seelig. Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution. *Nat Chem Biol*, 2013, **9**, 81-83.
136. *The PyMOL molecular graphics system, version 1.3, Schrödinger, LLC.* [CD-ROM].
137. Wierenga, R.K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett*, 2001, **492**, 193-198.
138. Blacklow, S.C., R.T. Raines, W.A. Lim, P.D. Zamore and J.R. Knowles. Triosephosphate isomerase catalysis is diffusion controlled. *Biochemistry*, 1988, **27**, 1158-1167.
139. Sterner, R. and B. Hocker. Catalytic versatility, stability, and evolution of the (β/α)₈-barrel enzyme fold. *Chem Rev*, 2005, **105**, 4038-4055.
140. Gerlt, J.A. and F.M. Raushel. Evolution of function in (β/α)₈-barrel enzymes. *Curr Opin Chem Biol*, 2003, **7**, 252-264.
141. Golynskiy, M.V., J.C. Haugner and B. Seelig. Highly diverse protein library based on the ubiquitous (β/α)₈ enzyme fold yields well-structured proteins through *in vitro* folding selection. *Chembiochem*, 2013, **14**, 1553-1563.
142. Consortium, T.U. UniProt: a hub for protein information. *Nucleic Acids Res*, 2015, **43**, D204-D212.
143. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. Mcgettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson and D.G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 2007, **23**, 2947-2948.
144. Hall, T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acid S*, 1999, **41**, 95-98.
145. Still, W.C., M. Kahn and A. Mitra. Rapid chromatographic technique for preparative separations with moderate resolution. *The Journal of Organic Chemistry*, 1978, **43**, 2923-2925.
146. Li, C., A. Wen, B. Shen, J. Lu, Y. Huang and Y. Chang. FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnol*, 2011, **11**, 92.
147. Sambrook, J., E.F. Fritsch and T. Maniatis. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, 1989.

148. Sambrook, J. and D.W. Russell. Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. *CSH Protoc*, 2006.
149. Beaucage, S.L. and M.H. Caruthers. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.*, 1981, **22**, 1859-1862.
150. Sau, S.P., A.C. Larsen and J.C. Chaput. Automated solid-phase synthesis of high capacity oligo-dT cellulose for affinity purification of poly-A tagged biomolecules. *Bioorg. Med. Chem. Lett.*, 2014, **24**, 5692-5694.
151. Gasteiger, E., C. Hoogland, A. Gattiker, S.E. Duvaud, M.R. Wilkins, R.D. Appel and A. Bairoch. Protein Identification and Analysis Tools on the ExPASy Server. In: J.M. WALKER, ed. *The Proteomics Protocols Handbook*. Totowa, NJ: Humana Press, 2005. 571-607.
152. Scharff, E.I., J. Koepke, G. Fritsch, C. Lücke and H. Rüterjans. Crystal structure of diisopropylfluorophosphatase from *Loligo vulgaris*. *Structure*, 2001, **9**, 493-502.
153. Blum, M.-M., F. Löhr, A. Richardt, H. Rüterjans and J.C.H. Chen. Binding of a Designed Substrate Analogue to Diisopropyl Fluorophosphatase: Implications for the Phosphotriesterase Mechanism. *J Am Chem Soc*, 2006, **128**, 12750-12757.
154. Pestka, S. Inhibitors of ribosome functions. *Annu Rev Microbiol*, 1971, **25**, 487-562.
155. Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 1985, **228**, 1315-1317.
156. Ratmeyer, L., R. Vinayak, Y.Y. Zhong, G. Zon and W.D. Wilson. Sequence specific thermodynamic and structural properties for DNA:RNA duplexes. *Biochemistry*, 1994, **33**, 5298-5304.
157. Kurz, M., P. Lohse and R.W. Wagner. *Peptide acceptor ligation methods*. Patent number: US7790421. 27/07/1999.
158. Kurz, M., K. Gu and P.A. Lohse. Psoralen photo-crosslinked mRNA–puromycin conjugates: a novel template for the rapid and facile preparation of mRNA–protein fusions. *Nucleic Acids Res.*, 2000, **28**, e83.
159. Mochizuki, Y., M. Biyani, S. Tsuji-Ueno, M. Suzuki, K. Nishigaki, Y. Husimi and N. Nemoto. One-pot preparation of mRNA/cDNA display by a novel and versatile puromycin-linker DNA. *ACS Comb Sci*, 2011, **13**, 478-485.
160. Olson, C.A., H.I. Liao, R. Sun and R.W. Roberts. mRNA display selection of a high-affinity, modification-specific phospho-IkBa-binding fibronectin. *ACS Chem Biol*, 2008, **3**, 480-485.
161. Kurz, M., K. Gu, A. Al-Gawari and P.A. Lohse. cDNA - protein fusions: covalent protein - gene conjugates for the *in vitro* selection of peptides and proteins. *Chembiochem*, 2001, **2**, 666-672.
162. Sinden, R.R. and P.J. Hagerman. Interstrand psoralen cross-links do not introduce appreciable bends in DNA. *Biochemistry*, 1984, **23**, 6299-6303.
163. Gamper, H., J. Piette and J.E. Hearst. Efficient formation of a crosslinkable HMT monoadduct at the Kpn I recognition site. *Photochem Photobiol*, 1984, **40**, 29-34.
164. Kibler-Herzog, L., G. Zon, B. Uznanski, G. Whittier and W.D. Wilson. Duplex stabilities of phosphorothioate, methylphosphonate, and RNA analogs of two DNA 14-mers. *Nucleic Acids Res*, 1991, **19**, 2979-2986.

165. Lamond, A.I. and B.S. Sproat. Antisense oligonucleotides made of 2'-O-alkylRNA: their properties and applications in RNA biochemistry. *FEBS Lett*, 1993, **325**, 123-127.
166. Cerritelli, S.M. and R.J. Crouch. Ribonuclease H: the enzymes in Eukaryotes. *FEBS J*, 2009, **276**, 1494-1505.
167. Cazenave, C., P. Frank and W. Büsen. Characterization of ribonuclease H activities present in two cell-free protein synthesizing systems, the wheat germ extract and the rabbit reticulocyte lysate. *Biochimie*, 1993, **75**, 113-122.
168. Keefe, A.D. Protein selection using mRNA display. *Curr Protoc Mol Biol*, 2001, **24**, Unit 24 25.
169. Isaacs, S.T., C.-K.J. Shen, J.E. Hearst and H. Rapoport. Synthesis and characterization of new psoralen derivatives with superior photoreactivity with DNA and RNA. *Biochemistry*, 1977, **16**, 1058-1064.
170. Eichman, B.F., B.H. Mooers, M. Alberti, J.E. Hearst and P.S. Ho. The crystal structures of psoralen cross-linked DNAs: drug-dependent formation of Holliday junctions. *J Mol Biol*, 2001, **308**, 15-26.
171. Erickson, A.H. and G. Blobel. Cell-free translation of messenger RNA in a wheat germ system. *Method Enzymol*, 1983, **96**, 38-50.
172. Josephson, K., M.C.T. Hartman and J.W. Szostak. Ribosomal synthesis of unnatural peptides. *J Am Chem Soc*, 2005, **127**, 11727-11735.
173. Roberts, R.W. Totally *in vitro* protein selection using mRNA-protein fusions and ribosome display. *Curr Opin Chem Biol*, 1999, **3**, 268-273.
174. Sleat, D.E., D.R. Gallie, R.A. Jefferson, M.W. Bevan, P.C. Turner and T.M. Wilson. Characterisation of the 5'-leader sequence of tobacco mosaic virus RNA as a general enhancer of translation *in vitro*. *Gene*, 1987, **60**, 217-225.
175. Wolin, S.L. and P. Walter. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J*, 1988, **7**, 3559-3569.
176. Schneider, C.A., W.S. Rasband and K.W. Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 2012, **9**, 671-675.
177. Beckler, G.S., D. Thompson and T. Van Oosbree. *In vitro* translation using rabbit reticulocyte lysate. *Method Mol Biol*, 1995, **37**, 215-232.
178. Jackson, R.J. Potassium salts influence the fidelity of mRNA translation initiation in rabbit reticulocyte lysates: unique features of encephalomyocarditis virus RNA translation. *Biochim Biophys Acta*, 1991, **1088**, 345-358.
179. Elmer, J., D. Harris and A.F. Palmer. Purification of hemoglobin from red blood cells using tangential flow filtration and immobilized metal ion affinity chromatography. *J Chromatogr B*, 2011, **879**, 131-138.
180. Porath, J., J.a.N. Carlsson, I. Olsson and G. Belfrage. Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature*, 1975, **258**, 598-599.
181. Takahashi, T. and R. Roberts. *In vitro* selection of protein and peptide libraries using mRNA display. In: G. MAYER, ed. *Nucleic Acid and Peptide Aptamers*. Humana Press, 2009. 293-314.
182. Huisgen, R. Cycloadditions — definition, classification, and characterization. *Angew Chem Int Ed Engl*, 1968, **7**, 321-328.
183. Cannizzaro, C.E., J.A. Ashley, K.D. Janda and K.N. Houk. Experimental determination of the absolute enantioselectivity of an antibody-catalyzed Diels-

- Alder reaction and theoretical explorations of the origins of stereoselectivity. *J Am Chem Soc*, 2003, **125**, 2489-2506.
184. Gouverneur, V.E., K.N. Houk, B. De Pascual-Teresa, B. Beno, K.D. Janda and R.A. Lerner. Control of the *exo* and *endo* pathways of the Diels-Alder reaction by antibody catalysis. *Science*, 1993, **262**, 204-208.
185. Jessup, P.J., C.B. Petty, J. Roos and L.E. Overman. 1-N-Acylamino-1,3-dienes from 2,4-Pentadienoic Acids by the Curtius Rearrangement: Benzyl trans-1,3-butadiene-1-carbamate. *In: Org Synth*. John Wiley & Sons, Inc., 2003.
186. Ninomiya, K., T. Shioiri and S. Yamada. Phosphorus in organic synthesis—VII. *Tetrahedron*, 1974, **30**, 2151-2157.
187. Shioiri, T., K. Ninomiya and S. Yamada. Diphenylphosphoryl azide. A new convenient reagent for a modified Curtius reaction and for the peptide synthesis. *J Am Chem Soc*, 1972, **94**, 6203-6205.
188. Green, N.M. Avidin and streptavidin. *Methods Enzymol*, 1990, **184**, 51-67.
189. Szalecki, W. Synthesis of norbiotinamine and its derivatives. *Bioconjugate Chem*, 1996, **7**, 271-273.
190. Miyamoto-Sato, E., H. Takashima, S. Fuse, K. Sue, M. Ishizaka, S. Tateyama, K. Horisawa, T. Sawasaki, Y. Endo and H. Yanagawa. Highly stable and efficient mRNA templates for mRNA-protein fusions and C-terminally labeled proteins. *Nucleic Acids Res*, 2003, **31**, e78.
191. Mochizuki, Y., T. Suzuki, K. Fujimoto and N. Nemoto. A versatile puromycin-linker using *cnvK* for high-throughput *in vitro* selection by cDNA display. *J Biotechnol*, 2015, **212**, 174-180.
192. Shimizu, Y., A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa and T. Ueda. Cell-free translation reconstituted with purified components. *Nat Biotechnol*, 2001, **19**, 751-755.
193. Nagumo, Y., K. Fujiwara, K. Horisawa, H. Yanagawa and N. Doi. PURE mRNA display for *in vitro* selection of single-chain antibodies. *J Biochem*, 2015.
194. Naimuddin, M. and T. Kubo. A high performance platform based on cDNA display for efficient synthesis of protein fusions and accelerated directed evolution. *ACS Comb Sci*, 2016.
195. Ohashi, H., Y. Shimizu, B.-W. Ying and T. Ueda. Efficient protein selection based on ribosome display system with purified components. *Biochem Bioph Res Co*, 2007, **352**, 270-276.
196. Van Den Bosch, H., R.B. Schutgens, R.J. Wanders and J.M. Tager. Biochemistry of peroxisomes. *Annu Rev Biochem*, 1992, **61**, 157-197.
197. Kim, P.K., R.T. Mullen, U. Schumann and J. Lippincott-Schwartz. The origin and maintenance of mammalian peroxisomes involves a *de novo* PEX16-dependent pathway from the ER. *J Cell Biol*, 2006, **173**, 521-532.
198. Tam, Y.Y.C., A. Fagarasanu, M. Fagarasanu and R.A. Rachubinski. Pex3p initiates the formation of a preperoxisomal compartment from a subdomain of the endoplasmic reticulum in *Saccharomyces cerevisiae*. *J Biol Chem*, 2005, **280**, 34933-34939.
199. Lazarow, P.B. and Y. Fujiki. Biogenesis of peroxisomes. *Annu Rev Cell Biol*, 1985, **1**, 489-530.

200. Glover, J.R., D.W. Andrews and R.A. Rachubinski. *Saccharomyces cerevisiae* peroxisomal thiolase is imported as a dimer. *Proc Natl Acad Sci USA*, 1994, **91**, 10541-10545.
201. Mcnew, J.A. and J.M. Goodman. An oligomeric protein is imported into peroxisomes *in vivo*. *J Cell Biol*, 1994, **127**, 1245-1257.
202. Walton, P.A., P.E. Hill and S. Subramani. Import of stably folded proteins into peroxisomes. *Mol Biol Cell*, 1995, **6**, 675-683.
203. Wanders, R.J.A. Metabolic functions of peroxisomes in health and disease. *Biochimie*, 2014, **98**, 36-44.
204. Gleeson, M.A. and P.E. Sudbery. The methylotrophic yeasts. *Yeast*, 1988, **4**, 1-15.
205. Wang, Y., Y. Xuan, P. Zhang, X. Jiang, Z. Ni, L. Tong, X. Zhou, L. Lin, J. Ding and Y. Zhang. Targeting expression of the catalytic domain of the kinase insert domain receptor (KDR) in the peroxisomes of *Pichia pastoris*. *FEMS Yeast Res*, 2009, **9**, 732-741.
206. Poirier, Y., V.D. Antonenkov, T. Glumoff and J.K. Hiltunen. Peroxisomal β -oxidation—A metabolic pathway with multiple functions. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 2006, **1763**, 1413-1426.
207. Sheng, J., J. Stevens and X. Feng. Pathway compartmentalization in peroxisome of *Saccharomyces cerevisiae* to produce versatile medium chain fatty alcohols. *Sci Rep*, 2016, **6**, 26884.
208. Chen, A.H. and P.A. Silver. Designing biological compartmentalization. *Trends Cell Biol*, 2012, **22**, 662-670.
209. Brocard, C., F. Kragler, M.M. Simon, T. Schuster and A. Hartig. The tetratricopeptide repeat-domain of the PAS10 protein of *Saccharomyces cerevisiae* is essential for binding the peroxisomal targeting signal-SKL. *Biochem Biophys Res Co*, 1994, **204**, 1016-1022.
210. Terlecky, S.R., W.M. Nuttley, D. Mccollum, E. Sock and S. Subramani. The *Pichia pastoris* peroxisomal protein PAS8p is the receptor for the C-terminal tripeptide peroxisomal targeting signal. *EMBO J*, 1995, **14**, 3627-3634.
211. Lanyon-Hogg, T., S.L. Warriner and A. Baker. Getting a camel through the eye of a needle: the import of folded proteins by peroxisomes. *Biol Cell*, 2010, **102**, 245-263.
212. Gould, S.J., G.A. Keller, N. Hosken, J. Wilkinson and S. Subramani. A conserved tripeptide sorts proteins to peroxisomes. *J Cell Biol*, 1989, **108**, 1657-1664.
213. Lametschwandtner, G., C. Brocard, M. Fransen, P. Van Veldhoven, J. Berger and A. Hartig. The difference in recognition of terminal tripeptides as peroxisomal targeting signal 1 between yeast and human is due to different affinities of their receptor Pex5p to the cognate signal and to residues adjacent to it. *J Biol Chem*, 1998, **273**, 33635-33643.
214. Reumann, S., L. Babujee, C. Ma, S. Wienkoop, T. Siemsen, G.E. Antonicelli, N. Rasche, F. Luder, W. Weckwerth and O. Jahn. Proteome analysis of *Arabidopsis* leaf peroxisomes reveals novel targeting peptides, metabolic pathways, and defense mechanisms. *Plant Cell*, 2007, **19**, 3170-3193.
215. Gatto, G.J., Jr., B.V. Geisbrecht, S.J. Gould and J.M. Berg. Peroxisomal targeting signal-1 recognition by the TPR domains of human PEX5. *Nat Struct Biol*, 2000, **7**, 1091-1095.

216. Sampathkumar, P., C. Roach, P.A. Michels and W.G. Hol. Structural insights into the recognition of peroxisomal targeting signal 1 by *Trypanosoma brucei* peroxin 5. *J Mol Biol*, 2008, **381**, 867-880.
217. D'andrea, L.D. and L. Regan. TPR proteins: the versatile helix. *Trends Biochem Sci*, 2003, **28**, 655-662.
218. Cross, L. *Re-design of a receptor-targeting signal interaction to create a new peroxisomal trafficking pathway*. PhD thesis, University of Leeds, 2016.
219. Chowdhary, G., A.R.A. Kataya, T. Lingner and S. Reumann. Non-canonical peroxisome targeting signals: identification of novel PTS1 tripeptides and characterization of enhancer elements by computational permutation analysis. *BMC Plant Biol*, 2012, **12**, 142-142.
220. Bloom, J.D., M.M. Meyer, P. Meinhold, C.R. Otey, D. Macmillan and F.H. Arnold. Evolving strategies for enzyme engineering. *Curr Opin Struct Biol*, 2005, **15**, 447-452.
221. Reetz, M.T. and J.D. Carballeira. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc*, 2007, **2**, 891-903.
222. Shivange, A.V., J. Marienhagen, H. Mundhada, A. Schenk and U. Schwaneberg. Advances in generating functional diversity for directed protein evolution. *Curr Opin Chem Biol*, 2009, **13**, 19-25.
223. Wang, T.W., H. Zhu, X.Y. Ma, T. Zhang, Y.S. Ma and D.Z. Wei. Mutant library construction in directed molecular evolution: casting a wider net. *Mol Biotechnol*, 2006, **34**, 55-68.
224. Hecker, K.H. and R.L. Rill. Error analysis of chemically synthesized polynucleotides. *Biotechniques*, 1998, **24**, 256-260.
225. Gatto, G.J., Jr., E.L. Maynard, A.L. Guerrero, B.V. Geisbrecht, S.J. Gould and J.M. Berg. Correlating structure and affinity for PEX5:PTS1 complexes. *Biochemistry*, 2003, **42**, 1660-1666.
226. Mumford, R.A., C.B. Pickett, M. Zimmerman and A.W. Strauss. Protease activities present in wheat germ and rabbit reticulocyte lysates. *Biochem Bioph Res Co*, 1981, **103**, 565-572.
227. Parsell, D.A. and R.T. Sauer. The structural stability of a protein is an important determinant of its proteolytic susceptibility in *Escherichia coli*. *J Biol Chem*, 1989, **264**, 7590-7595.
228. Crooks, G.E., G. Hon, J.M. Chandonia and S.E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, 2004, **14**, 1188-1190.
229. Lanyon-Hogg, T., J. Hooper, S. Gunn, S.L. Warriner and A. Baker. PEX14 binding to *Arabidopsis* PEX5 has differential effects on PTS1 and PTS2 cargo occupancy of the receptor. *FEBS Lett*, 2014, **588**, 2223-2229.
230. Skoulding, N.S., G. Chowdhary, M.J. Deus, A. Baker, S. Reumann and S.L. Warriner. Experimental validation of plant peroxisomal targeting prediction algorithms by systematic comparison of *in vivo* import efficiency and *in vitro* PTS1 binding affinity. *J Mol Biol*, 2015, **427**, 1085-1101.
231. Huang, B.C. and R. Liu. Comparison of mRNA-display-based selections using synthetic peptide and natural protein libraries. *Biochemistry*, 2007, **46**, 10102-10112.
232. Virnekas, B., L. Ge, A. Pluckthun, K.C. Schneider, G. Wellenhofer and S.E. Moroney. Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed

- oligonucleotides for random mutagenesis. *Nucleic Acids Res*, 1994, **22**, 5600-5607.
233. Inglese, J., R.L. Johnson, A. Simeonov, M. Xia, W. Zheng, C.P. Austin and D.S. Auld. High-throughput screening assays for the identification of chemical probes. *Nat. Chem. Biol.*, 2007, **3**, 466-479.
234. Zhu, Z. and J. Cuozzo. High-throughput affinity-based technologies for small-molecule drug discovery. *J. Biomol. Screening*, 2009, **14**, 1157-1164.
235. Keefe, A.D., S. Pai and A. Ellington. Aptamers as therapeutics. *Nat. Rev. Drug Discovery*, 2010, **9**, 537-550.
236. Goldflam, M. and C.G. Ullman. Recent advances toward the discovery of drug-like peptides *de novo*. *Front Chem*, 2015, **3**, 69.
237. Bruce, V.J., A.N. Ta and B.R. Mcnaughton. Minimalist antibodies and mimetics: An update and recent applications. *Chembiochem*, 2016, **17**, 1892-1899.
238. Vijayendran, R.A. and D.E. Leckband. A quantitative assessment of heterogeneity for surface-immobilized proteins. *Anal. Chem.*, 2001, **73**, 471-480.
239. Mullard, A. DNA tags help the hunt for drugs. *Nature*, 2016, **530**, 367-369.
240. Gorin, D.J., A.S. Kamlet and D.R. Liu. Reactivity-dependent PCR: direct, solution-phase *in vitro* selection for bond formation. *J. Am. Chem. Soc.*, 2009, **131**, 9189-9191.
241. Mcgregor, L.M., D.J. Gorin, C.E. Dumelin and D.R. Liu. Interaction-dependent PCR: identification of ligand-target pairs from libraries of ligands and libraries of targets in a single solution-phase experiment. *J. Am. Chem. Soc.*, 2010, **132**, 15522-15524.
242. Mcgregor, L.M., T. Jain and D.R. Liu. Identification of ligand-target pairs from combined libraries of small molecules and unpurified protein targets in cell lysates. *J. Am. Chem. Soc.*, 2014, **136**, 3264-3270.
243. Hendrickson, E.R., T.M. Truby, R.D. Joerger, W.R. Majarian and R.C. Ebersole. High sensitivity multianalyte immunoassay using covalent DNA-labeled antibodies and polymerase chain reaction. *Nucleic Acids Res.*, 1995, **23**, 522-529.
244. Lundberg, M., A. Eriksson, B. Tran, E. Assarsson and S. Fredriksson. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res.*, 2011, **39**, e102.
245. Ogawa, A. and M. Maeda. Aptazyme-based riboswitches as label-free and detector-free sensors for cofactors. *Bioorg. Med. Chem. Lett.*, 2007, **17**, 3156-3160.
246. Bowley, D.R., T.M. Jones, D.R. Burton and R.A. Lerner. Libraries against libraries for combinatorial selection of replicating antigen-antibody pairs. *Proc. Natl Acad. Sci. USA*, 2009, **106**, 1380-1385.
247. Fredriksson, S., M. Gullberg, J. Jarvius, C. Olsson, K. Pietras, S.M. Gustafsdottir, A. Ostman and U. Landegren. Protein detection using proximity-dependent DNA ligation assays. *Nat. Biotechnol.*, 2002, **20**, 473-477.
248. Barrette-Ng, I.H., S.-C. Wu, W.-M. Tjia, S.-L. Wong and K.K.S. Ng. The structure of the SBP-Tag-streptavidin complex reveals a novel helical scaffold bridging binding pockets on separate subunits. *Acta Crystallogr D*, 2013, **69**, 879-887.
249. Gyi, J.I., G.L. Conn, A.N. Lane and T. Brown. Comparison of the thermodynamic stabilities and solution conformations of DNA-RNA hybrids containing purine-

- rich and pyrimidine-rich strands with DNA and RNA duplexes. *Biochemistry*, 1996, **35**, 12538-12548.
250. Lesnik, E.A. and S.M. Freier. Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. *Biochemistry*, 1995, **34**, 10807-10815.
251. Kotewicz, M.L., C.M. Sampson, J.M. D'alessio and G.F. Gerard. Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Res*, 1988, **16**, 265-277.
252. Champoux, J.J. and S.J. Schultz. Ribonuclease H: properties, substrate specificity, and roles in retroviral reverse transcription. *FEBS J*, 2009, **276**, 1506-1516.
253. Galan, A., L. Comor, A. Horvatic, J. Kules, N. Guillemin, V. Mrljak and M. Bhide. Library-based display technologies: where do we stand? *Molecular BioSystems*, 2016.
254. Kukolka, F., M. Lovrinovic, R. Wacker and C.M. Niemeyer. Covalent coupling of DNA oligonucleotides and streptavidin. *Methods Mol Biol*, 2004, **283**, 181-196.
255. Niemeyer, C.M., T. Sano, C.L. Smith and C.R. Cantor. Oligonucleotide-directed self-assembly of proteins: semisynthetic DNA-streptavidin hybrid molecules as connectors for the generation of macroscopic arrays and the construction of supramolecular bioconjugates. *Nucleic Acids Res*, 1994, **22**, 5530-5539.
256. Mollwitz, B., E. Brunk, S. Schmitt, F. Pojer, M. Bannwarth, M. Schiltz, U. Rothlisberger and K. Johnsson. Directed evolution of the suicide protein O⁶-alkylguanine-DNA alkyltransferase for increased reactivity results in an alkylated protein with exceptional stability. *Biochemistry*, 2012, **51**, 986-994.
257. Gu, G.J., M. Friedman, C. Jost, K. Johnsson, M. Kamali-Moghaddam, A. Pluckthun, U. Landegren and O. Soderberg. Protein tag-mediated conjugation of oligonucleotides to recombinant affinity binders for proximity ligation. *N Biotechnol*, 2013, **30**, 144-152.
258. Encell, L.P., R. Friedman Ohana, K. Zimmerman, P. Otto, G. Vidugiris, M.G. Wood, G.V. Los, M.G. Mcdougall, C. Zimprich, N. Karassina, R.D. Learish, R. Hurst, J. Hartnett, S. Wheeler, P. Stecha, J. English, K. Zhao, J. Mendez, H.A. Benink, N. Murphy, D.L. Daniels, M.R. Slater, M. Urh, A. Darzins, D.H. Klaubert, R.F. Bulleit and K.V. Wood. Development of a dehalogenase-based protein fusion tag capable of rapid, selective and covalent attachment to customizable ligands. *Curr Chem Genomics*, 2012, **6**, 55-71.
259. Pakkila, H., R. Peltomaa, U. Lamminmaki and T. Soukka. Precise construction of oligonucleotide-Fab fragment conjugate for homogeneous immunoassay using HaloTag technology. *Anal Biochem*, 2015, **472**, 37-44.
260. Gu, L., C. Li, J. Aach, D.E. Hill, M. Vidal and G.M. Church. Multiplex single-molecule interaction profiling of DNA-barcoded proteins. *Nature*, 2014, **515**, 554-557.
261. Svensen, N., O.B. Peersen and S.R. Jaffrey. Peptide synthesis on a next-generation DNA sequencing platform. *ChemBiochem*, 2016, **17**, 1628-1635.
262. Kanan, M.W., M.M. Rozenman, K. Sakurai, T.M. Snyder and D.R. Liu. Reaction discovery enabled by DNA-templated synthesis and *in vitro* selection. *Nature*, 2004, **431**, 545-549.
263. Yli-Kauhaluoma, J.T., J.A. Ashley, C.-H. Lo, L. Tucker, M.M. Wolfe and K.D. Janda. Anti-metallocene antibodies: A new approach to enantioselective catalysis of the Diels-Alder reaction. *J Am Chem Soc*, 1995, **117**, 7041-7047.

264. Schmidt-Dannert, C. and F.H. Arnold. Directed evolution of industrial enzymes. *Trends Biotechnol*, 1999, **17**, 135-136.
265. Rueping, M. and B.J. Nachtsheim. A review of new developments in the Friedel-Crafts alkylation - From green chemistry to asymmetric catalysis. *Beilstein J Org Chem*, 2010, **6**, 6.
266. Han, F.S. Transition-metal-catalyzed Suzuki-Miyaura cross-coupling reactions: a remarkable advance from palladium to nickel catalysts. *Chem Soc Rev*, 2013, **42**, 5270-5298.
267. Agresti, J.J., E. Antipov, A.R. Abate, K. Ahn, A.C. Rowat, J.C. Baret, M. Marquez, A.M. Klibanov, A.D. Griffiths and D.A. Weitz. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc Natl Acad Sci USA*, 2010, **107**, 4004-4009.
268. Wojcik, J., O. Hantschel, F. Grebien, I. Kaupe, K.L. Bennett, J. Barkinge, R.B. Jones, A. Koide, G. Superti-Furga and S. Koide. A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nat Struct Mol Biol*, 2010, **17**, 519-527.
269. Yoshida, S., K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, K. Toyohara, K. Miyamoto, Y. Kimura and K. Oda. A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science*, 2016, **351**, 1196-1199.
270. Huang, P.S., S.E. Boyken and D. Baker. The coming of age of *de novo* protein design. *Nature*, 2016, **537**, 320-327.
271. Faiella, M., C. Andreozzi, R.T. De Rosales, V. Pavone, O. Maglio, F. Nastri, W.F. Degrado and A. Lombardi. An artificial di-iron oxo-protein with phenol oxidase activity. *Nat Chem Biol*, 2009, **5**, 882-884.
272. Miller, B.G. and R. Wolfenden. Catalytic proficiency: the unusual case of OMP decarboxylase. *Annu Rev Biochem*, 2002, **71**, 847-885.
273. Blomberg, R., H. Kries, D.M. Pinkas, P.R. Mittl, M.G. Grutter, H.K. Privett, S.L. Mayo and D. Hilvert. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature*, 2013, **503**, 418-421.