UCL

Directed evolution of bioactive compounds: oxa(thia)zole-containing post-translationally modified peptides

Author:
Pedro A. G. Tizei

Supervisors:
Dr. Vitor B. Pinheiro
Prof. Charles M. Marson

This Thesis was submitted for the degree of PhD in Structural and Molecular Biology.

Research Department of Structural and Molecular Biology

April 2018

Declaration of Authorship

I, Pedro Augusto Galvao Tizei, declare that this thesis titled, 'Directed evolution of bioactive compounds: oxa(thia)zole-containing post-translationally modified peptides' and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

The results presented in Chapter 6 have been submitted for peer-reviewed publication and are currently available as a preprint in bioRxiv: "InDel assembly: a novel framework for engineering protein loops through length and compositional variation." (Pedro A. G. Tizei, Emma Harris, Marleen Renders and Vitor B. Pinheiro, 2017, doi: 10.1101/127829)

Signed: ________________________________

Date: ________________________________

"The kind of control you're attempting is not possible. If there's one thing the history of evolution has taught us, it is that life will not be contained. Life breaks free, expands to new territories, crashes through barriers, painfully, maybe even dangerously [...]

I'm simply saying that life [pause] finds a way".

Dr. Ian Malcom (played by Jeff Goldblum) - Jurassic Park

Abstract

PhD in Structural and Molecular Biology

Directed evolution of bioactive compounds: oxa(thia)zole-containing post-translationally modified peptides

by Pedro A. G. Tizei

Thiazole/Oxazole Modified Microcins (TOMMs) are a diverse class of post-translationally modified peptides including many bioactive compounds; a potential new source for drug discovery. Despite a limited understanding of the TOMM synthase heterotrimeric complex biosynthetic mechanism, a variable degree of substrate plasticity is present in the family. This makes them attractive targets for developing novel oxazole- and thiazole-containing compounds from synthetic peptides.

Available annotation on complex members suggests the presence of different biochemical activities among homologous proteins, precluding the use of established prediction methods for identification of functional residues. A novel algorithm was developed (Normalised Shannon Entropy, NoSE) for functional prediction from sequence alignments containing mixed functions. NoSE was applied, along with established conservation- and coevolution-based metrics, to detect functional residues in the well-characterised bacterial Solute Binding Protein family, which could be validated against the extensively reported characterisation.

The strategy was applied for functional residue prediction in the TOMM synthase complex and candidate functional residues were mutated in McbC dehydrogenase of Escherichia coli. Mutants were assessed using a bacterial growth inhibition bioassay and six out of sixteen mutations reduced TOMM production, demonstrating the value of employing a prediction strategy to improve characterization of proteins. Attempts at establishing an in vitro assay for TOMM biosynthesis were unsuccessful due to difficulties in protein expression and purification, as well as inconsistent assay results.

Finally, a framework for directed evolution of length-variable proteins was developed, with the aim of engineering synthetic TOMM products. A method was developed for assembly of high-quality libraries at a low cost, along with a workflow for enriched motif detection in selection experiments. The approach was validated by isolating seven novel variants of the β-lactamase TEM-1 active on a non-cognate substrate.

Together, the developed methods represent a foundation for establishing TOMM biosynthesis as a platform for discovery of novel bioactive compounds.

Impact Statement

The steady development of powerful biotechnological tools for engineering living organisms represent a promising source of new candidate molecules for discovery of drugs for clinical use. Among the compounds that can be accessed by biological systems are the Thiazole/Oxazole Modified Microcins (TOMMs), a family of molecules produced by diverse microorganisms, with known molecules displaying antimicrobial and anticancer activities, both highly-relevant for clinical applications. Synthesis of these molecules by traditional organic chemistry methods is difficult and their biological production route starts from a genetically-encoded peptide, making it attractive for engineering with the tools of molecular biology.

Despite the potential application of synthetic TOMM molecules for compound discovery, this class of molecules has not yet been widely engineered due to limited understanding of the functioning of the enzyme complex involved in TOMM biosynthesis. This was the motivation behind the development a strategy to improve the characterisation of these enzymes, exploiting publicly-available sequence data from a wide range of microorganisms to select sites in the proteins likely to contribute to their function in TOMM biosynthesis. Beyond improving characterisation of the processes involved in TOMM biosynthesis, the prediction strategy developed here is generalisable and can be applied to many other poorly-understood families of proteins, accelerating the identification of functional residues.

Improved understanding of TOMM biosynthetic mechanisms will also be of use in future efforts to engineer the enzyme complex with the aim of establishing a generic platform for production of synthetic TOMMs. As demonstrated here in the attempts to reconstitute purified TOMM synthase complexes *in vitro*, the task was not trivial, with multiple expression and detection issues left unsolved.

The currently known enzyme complexes are not robust enough for production of repertoires of novel synthetic molecules. New engineered — or as-yet unknown natural — enzymes are needed with robust activity outside their native organisms and with sufficient tolerance towards modified substrate peptides.

In the absence of an adequate set of enzymes for development of novel TOMMs, a framework was developed here to enable efficient engineering of TOMM substrate peptides and any other protein (or region of a protein) that can have the length of its sequence changed and still function. The previously-available methods for the production of genetic diversity with variable sequence lengths were either inefficient or too costly, putting this type of diversity beyond the reach of most research groups. InDel Assembly — the method developed here — can produce sequences with targeted diversity in both length and composition through a simple experimental protocol, enabling easier exploration of length variation as a parameter in protein engineering. In addition to being used for TOMM substrates, this framework can be adapted to engineer a wide range of biological systems such as antibodies, enzymes and entire metabolic pathways.

A generalised strategy for the production of synthetic TOMMs is still not within reach, but the tools developed as part of this thesis represent useful contributions to many distinct topics within biological research, including the development of novel molecules for clinical and any other applications.

Acknowledgements

First of all, I'd like to deeply thank my primary supervisor, Dr. Vitor Pinheiro, for having me as the first member of his group and for the support, guidance, and encouragement helped me see the work through to the end. The experiences in Prof. Charles Marson's lab were invaluable in motivating me to learn more about chemistry than I did in all my pre-PhD years. His comments during Committee meetings, along with Prof. John Ward's, were of great assistance in directing the work until this final stage. I also acknowledge the CAPES foundation for providing the funding that allowed me to spend these very rewarding four years at UCL.

A very informative meeting in early Summer with Prof. Christine Orengo and Dr. Sayoni Das was essential in directing the research that became part of Chapter 3. I'm also indebted to Saira Maldonado-Puga, whose poster at the Protein Society meeting gave me the inspiration to select the target family explored in that Chapter.

Before moving onto the remaining members of the Pinheiro group, I'd like to thank J. Ewan Coates for sharing his experiences with the difficult TOMM enzymes and providing insightful input from a very different perspective. I wish him the best of luck in his current endeavours.

Li Cheah, Alberto Aparicio de Narvaez, Emma Harris, Charlie Henderson provided invaluable assistance in carrying out parts of the work described here and ongoing McbC characterisation. Friendly discussions — both in and out of the lab — with Antje Krüger, Yan Kay Ho, Chris Cozens, Eszter Csibra and Marleen Renders were always perfect for coming up with ideas and also helping me through the tough moments when things weren't working so well. I'd also like to thank Warren Hazelton, Leticia Torres, Ana Riesco, Warren Hazelton, Hugo Villanueva, Hugo Sinclair and all the others with whom I shared time in Darwin G04/G09.

All the moments with friends outside the Darwin Building were also an integral part of the PhD experience. The Rogue Hikers were there for many a weekend walking through the UK countryside, getting to know parts of the country through its mud, mountains, weather, sheep, trains, and pubs. All the early mornings (and some great evenings, too!) with Danny Bent, Joon Wong, Spike Reid, Dave Finch, Donna Marsh, Laura Plant, Chris Millar, Jake Otto, Tom McGillycuddy, Anna McNuff, and everybody else in the Project Awesome and November Project London crews helped to keep me running in more than the physical sense. The exercise, amazing friendship, and a good bit of silliness were always a great distraction from everything else.

The place of honour goes to my family: Rita, Augusto, and Luiz. Though the physical distances between us are very large, I could always count on your unwavering support at

all moments and I will always remember everything I learned from you about life, before and after moving to the UK. And finally, my girlfriend Maria Dermit, whose constant love and support gave meaning to all this time and pushed me through difficult moments, by always putting everything into perspective.

Contents

List of Figures

List of Tables

Abbreviations

ADP	Adenosine Diphosphate
AMP	Adenosine Monophosphate
ATP	Adenosine Triphosphate
BL21(DE3)	*E. coli* strain BL21(DE3)
CFU	Colony-Forming Unit
CMI	Cumulative Mutual Information
DCA	Direct Coupling Analysis
DSB	Double-Stranded Break
dsDNA	double-stranded DNA
DTT	Dithiothreitol
ER2566	*E. coli* strain NEB ER2566
FACS	Fluorescence-Assisted Cell Sorting
FMN	Flavin MonoNucleotide
HDR	Homology-Directed Repair
JSD	Jensen-Shannon Divergence
LC/MS	Liquid Chromatography-coupled to Mass Spectrometry
MALDI-TOF	Matrix-Assisted Laser Desorption Ionization Time of Flight
MBP	Maltose-Binding Protein
MI	Mutual Information
MIC	Minimal Inhibitory Concentration
MSA	Multiple Sequence Alignment
Mcb17	Microcin B17
mRNA	Messenger RNA
MS	Mass Spectrometry
NEB 10B	*E. coli* strain NEB 10-B
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End Joining
NoSE	Normalised Shannon Entropy
OD600	Optical Density of a cell suspension at a wavelength of 600 nm
PCR	Polymerase Chain Reaction
PNK	PolyNucleotide Kinase

RiPP	Ribosomally synthesised and Post-translationally modified Peptides
SDM	Site-Directed Mutagenesis
ssDNA	single-stranded DNA
T7 Express	*E. coli* strain T7 Express
T7 Express LysY/Iq	*E. coli* strain T7 Express LysY/Iq
TOMM	Thiazole/Oxazole Modified Microcin
WT	Wild-type

Chapter 1

Introduction

The biosynthetic repertoire available in the Earth's biosphere by far outstrips the capacity for traditional isolation and characterisation of natural products for clinically- or industrially-relevant bioactive compounds. However, some of the biosynthetic pathways for naturally-ocurring products can be exploited to efficiently generate and select novel compounds, using recent molecular biology techniques.

One poorly-sampled region of natural product space that can be explored in this manner is comprised of the Thiazole/Oxazole-Modified Microcins (TOMMs). These are post-translationally modified peptides containing heterocyclic thiazole and/or oxazole moieties, identified in bacterial and archaeal species. Their genomically-encoded substrate peptides are modified by sets of enzymes known to tolerate altered substrates, which makes the biosynthetic pathway amenable to engineering with directed evolution techniques towards the production of new biological activities.

1.1 Thiazole/Oxazole-Modified Microcins

1.1.1 Nomenclature and structure

TOMMs are part of a much larger category of natural products, the Ribosomally synthesised and Post-translationally modified Peptides (RiPPs). This class of compounds was the subject of a large collaborative review (Arnison et al., 2013) with the aim of standardising the nomenclature and classifying the diverse known products into smaller groups with shared biosynthetic pathways and structural elements.

In this thesis, TOMMs will be defined as ribosomal peptide-derived molecules containing heterocycles generated by post-translational modification of serine, threonine or cysteine residues. This definition encompasses three of the classes proposed in the RiPP nomenclature: the Linear Azol(in)e-containing Peptides (LAPs), the cyanobactins and the thiopeptides (Arnison et al., 2013). Despite the large structural and functional divergences between all the TOMMs (See Fig. 1.1 for examples), their heterocyclic modifications share a homologous biosynthetic enzyme system, which will be the main focus of this thesis. The flat heterocycles installed in the peptides constrain the conformational flexibility of the backbone and this constraint is crucial for making TOMMs into very specific ligands for biologically-relevant structures.

In addition to the heterocycles, TOMM substrate peptides can undergo several other modifications leading to their final, active forms. Cyanobactins and thiopeptides have all or most of their peptide backbone converted into a macrocycle, while the LAPs remain in a linear form. The side chains of non-heterocyclised residues can also receive other modifications such as methylation, dehydration, and prenylation (Wipf and Uto, 1999; Kalyon et al., 2011). These additional modifications are specific to each molecule and are performed by enzymes that are usually not involved in the heterocyclisation reactions.

1.1.2 Naturally-ocurring TOMMs

TOMM products and their homologous biosynthetic machinery have been identified in several major groups of prokaryotes: Proteobacteria (*E. coli* (Yorgey et al., 1993)), Firmicutes (*Bacillus* (Scholz et al., 2011), *Clostridium* (Gonzalez et al., 2010a), *Streptococcus* (Lee et al., 2008)), Actinobacteria (*Streptomyces* (Kelly et al., 2009)), Cyanobacteria (*Prochloron*

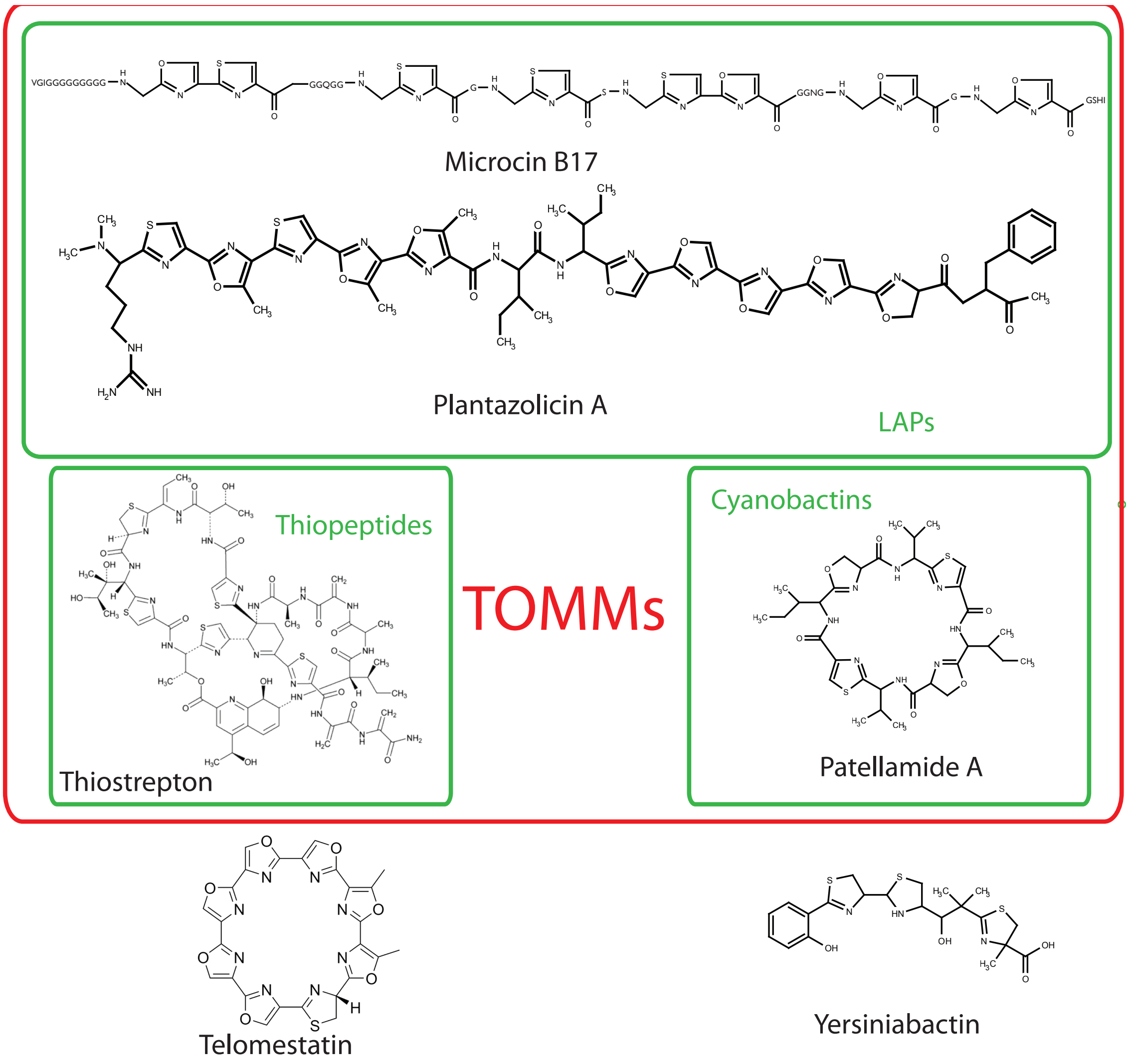


FIGURE 1.1: TOMMs and other thiazole containing natural products. Inside the green boxes are representative molecules in the classes of RiPPs (*sensu* Arnison et al 2013), that are included within TOMMs as defined in this thesis (red box). Outside the red box are azole-containing compounds not defined as TOMMs. Yersiniabactin is synthesised by a non-ribosomal peptide synthase system and the biosynthetic route to telomestatin is unknown.

(Schmidt et al., 2005)) and Archaea (*Pyrococcus* (Lee et al., 2008)). The widespread occurrence and maintenance of these compounds suggests they must provide the producing organisms with some selective advantage, however the natural biological function of many natural TOMMs has not yet been identified.

All the TOMMs which have a known function for their producing organism are involved in ecological interactions at some level. Streptolysin S was likely the first TOMM to be isolated (Weld, 1934; Czarnetzky et al., 1938), though its chemical identity and biosynthesis were only elucidated over 70 years later (Lee et al., 2008), and it is a haemolytic toxin against the vertebrate hosts of *Streptococcus pyogenes.* Clostridiolysin S, from *Clostridium botulinum*, is also a TOMM virulence factor with a similar activity to streptomycin S (Gonzalez et al., 2010a).

Anti-microbial activity is the other major natural function that has been identified for TOMMs. The first compound in this class to have its biosynthetic pathway described was microcin B17 from *E. coli*, an antimicrobial compound that acts by inhibiting DNA gyrase of closely-related bacteria (Li et al., 1996). More recently, *Bacillus amyloliquefaciens* was shown to produce a TOMM antibiotic with strong activity only against other species in the same genus (Scholz et al., 2011). Activity against phylogenetically close species can be indicative of a role in competition for similar ecological niches.

In contrast, the thiopeptide antibiotic GE37468 from *Streptomyces sp.* ATCC 55365, a member of the phylum Actinobacteria, has strong activity against Methicillin-Resistant *Staphylococcus aureus* (MRSA) which is in the phylum Firmicutes (Young and Walsh, 2011). *Streptomyces* are typically soil-associated bacteria, while MRSA and other *Staphylococcus* species are found in vertebrate skin and mucous membranes, with some being opportunistic pathogens. Since these bacteria are not likely to compete for the same resources, this activity suggests that TOMMS are involved in ecological interactions that are still largely uncharacterised.

However, most other TOMMs have no known function for the producing organism yet have been studied in detail for their clinically-relevant activities. This is the case for the patellamides and trunkamides, macrocyclic modified peptides with anti-tumour activities. These molecules are produced by *Prochloron sp.* uncultivated cyanobacterial symbionts of the colonial ascidian *Lissoclinum patella* (Schmidt et al., 2005). The cytotoxicity of some of the patellamides suggest a possible function as a deterrent against predation by other animals, but this has never been demonstrated. Experiments with extracts obtained from other species of ascidians showed strong anti-diatom activity for some of the extracts, which suggests that some compound contained in these animal's tissues could be responsible for preventing fouling of the colony's surface by diatoms (Koplovitz et al., 2011). It is possible to speculate that a cytotoxic TOMM synthesised by a cyanobacterial symbiont would be responsible for this function.

1.1.3 TOMM Biosynthesis

All biosynthetic pathways for TOMM synthesis characterised to date are encoded in genomic clusters containing the genes needed to produce the final molecule and ensure its correct function, including the sequence for the substrate peptide itself (reviewed in Arnison et al. (2013)). The core enzymatic complex needed to install the heterocycles in the peptides is composed of three components, an ATP-dependent cyclodehydratase activity, a docking domain and an FMN-dependent dehydrogenase activity (biosynthetic steps in Fig. 1.2). *In vitro*, these components have been shown to associate into a heterotrimeric TOMM synthase complex (Li et al., 1996).

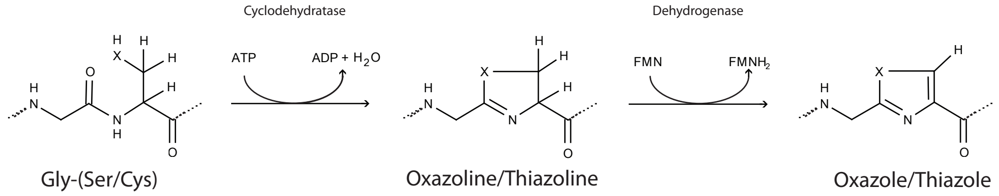


FIGURE 1.2: Biosynthetic steps catalysed by the TOMM synthase complex to generate azolines and azoles by post-translational modification of peptides. Heterocycles are installed by the cyclodehydratase activity and the resulting azolines can be oxidised by the FMN-dependent dehydrogenase activity into azoles.

The cyclodehydratase and the docking domain together catalyse the heterocyclisation of Cys/Ser/Thr side chains into reduced thiazoline or (methyl-)oxazoline moieties. In most bacteria, these two components are produced as two separate proteins, however in cyanobacteria these two domains are usually fused in a single bifunctional polypeptide (See subsection 1.3.2).

The dehydrogenase activity oxidises some (or all) of the reduced heterocycles to generate thiazole or (methyl-)oxazole rings. In most bacteria, monofunctional proteins catalyse this reaction. Some cyanobacteria have this activity as part of a bifunctional protein, together with a protease that is involved in post-heterocyclisation TOMM maturation (See subsection 1.3.3).

TOMM substrate peptides are composed of an N-terminal leader sequence, followed by a core region which contains the residues to be modified and in some cases a C-terminal recognition sequence (Fig. 1.3). The leader sequence is needed for recognition by the synthase complex and also contains sites for further modifications and a protease cleavage site to remove it from the modified peptide. At the C-terminus, the recognition sequence directs macrocyclisation activity in cyanobacterial products.

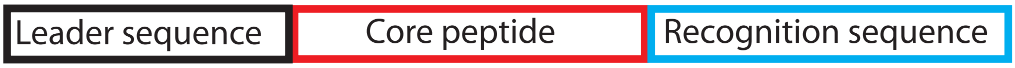


FIGURE 1.3: Generalised structure of a TOMM substrate peptide. In black is the leader sequence needed to direct TOMM synthase activity, in red is the region containing the heterocyclisable residues and in blue is an optional region containing signals to direct other post-translational modification enzymes (nomenclature from Arnison et al. (2013)).

In addition to the substrate peptide (A) and the genes encoding the heterocyclase complex (B, C and D), TOMM biosynthesis clusters usually code for one or more proteins that are needed for further modifications or to ensure proper TOMM function. Among the additional proteins that are needed by these pathways are transporters to export the products, other post-translational modification enzymes (prenylases or dehydratases, for instance) and also proteins that are needed for immunity to the final TOMM product.

1.1.4 Other -azole containing natural products

The TOMM synthase complex is not the only known pathway by which thiazole and (methyl)-oxazoles can be introduced into natural products. *Yersinia pestis* produces yersiniabactin a heterocycle-containing compound that is very similar to peptides, but is produced by a non-ribosomal peptide synthetase/polyketide synthase hybrid system from small-molecule precursors, without requiring a substrate peptide for modification (Miller et al., 2002). This compound is a siderophore with a high affinity for iron, which makes it an important virulence factor to allow *Y. pestis* colonisation of its hosts (Miller et al., 2006).

There are also other heterocycle-containing molecules which still have not had their biosynthetic mechanism elucidated, but are likely to be produced by a homologue of the TOMM biosynthetic machinery. One such molecule is telomestatin (Fig. 1.1), which was isolated from the marine bacterium *Streptomyces annulatus* and is the most specific natural product inhibitor of telomerase (Shin-ya et al., 2001). This activity makes it a potential candidate for drug development as a treatment for telomerase positive tumours (Tauchi et al., 2003). Telomestatin is a macrocyclic molecule consisting solely of five oxazoles, two methyloxazoles and one thiazoline. The total synthesis of this molecule has been done by more than one group but the final yields obtained were very low (<2%), which makes further development of the compound very costly (Doi et al., 2006; Marson and Saadi, 2006; Linder et al., 2011). The structure of telomestatin is very similar to some of the cyanobactins, which suggests a possible homologous biosynthetic pathway and that it could be synthesised *in vitro* by heterologous expression of a suitable TOMM synthase complex and substrate peptide combination.

1.2 TOMM Product Engineering

1.2.1 *In vitro* Production

Many of the TOMM-producing species are non-model organisms that are not genetically tractable or, in some cases, even culturable in the laboratory. Therefore, in order to better understand the biosynthetic pathways for future engineering, *in vitro* production and assay systems needed to be developed.

The first successful report of *in vitro* reconstitution of a TOMM biosynthetic pathway was for the *E. coli* microcin B17, by over-expression of cyclodehydratase (McbD), docking (McbB), dehydrogenase (McbC) and substrate (McbA) proteins (Li et al., 1996). This system enabled the early characterisation of the TOMM synthase complex and development of methods for activity detection. Other synthase pathways that have since been studied *in vitro* are the ones that synthesise patellamides in *Prochloron sp.* (Schmidt et al., 2005), streptolysin S in *S. pyogenes* (Lee et al., 2008), plantazolicin in *B. amyloliquefaciens* (Scholz et al., 2011) and BalhA in *Bacillus sp.* Al-Hakam (Melby et al., 2012). In addition, the trimeric synthase complex from the archaeon *P. furiosus* was heterologously expressed and shown to have activity on the *S. pyogenes* substrate peptide, converting it into the active form of streptolysin S (Lee et al., 2008).

Cell-free systems allow for easier introduction of non-proteinogenic amino acids into the substrate peptide, as well as avoiding product toxicity issues which can hinder *in vivo* TOMM production. One such system was developed for *in vitro* translation and heterocyclisation of substrate peptides, using the patellamide heterocyclase patD and increasing the chemical space accessible to engineered TOMMs (Goto et al., 2014) by allowing the introduction of non-proteinogenic amino acids into synthetic TOMMs.

1.2.2 Production of modified TOMMs by genetic engineering

The heterologously expressed TOMM synthase complexes were an important step in facilitating efforts to generate new TOMM products, while also allowing exploration of the limits of synthase activity (See Section 1.3 for characterisation of TOMM synthase enzymes). After the initial description of *E. coli* microcin B17 biosynthesis, site-directed mutagenesis was used to generate substrate peptide variants that were heterocyclised and screened for antibiotic activity (Sinha Roy et al., 1999). Some of the mutants led to new compounds but all the new TOMMs that were successfully produced in these experiments had lower antibiotic activity than the wild-type microcin. However, mutations at different heterocyclisation sites did not have equal effects. There was a significantly stronger decrease in activity when the C-terminal bisheterocyclic moiety was mutated than when similar mutations were introduced into the N-terminal bisheterocycle (See Fig. 1.1 for MicrocinB17 structure) (Sinha Roy et al., 1999), which suggests different roles for each region in the interaction with the DNA gyrase target.

Substitutions of cyclised residues for non-cyclisable ones or S-C/C-S mutations (replacing a thiazole with an oxazole and vice-versa) have been shown to generally reduce the antibacterial activity of variants against *E. coli* and the inhibitory effect on DNA gyrase (Sinha Roy et al., 1999; Zamble et al., 2001). Interestingly, Sinha Roy et al. (1999) also detected in extracts from producing cultures a minor form of the Microcin B17 product in which Ser26 is heterocyclised — bringing the total number of heterocycles in the molecule to nine (compared with the eight heterocycles in the accepted structure of Microcin B17) — and this compound was shown to have higher antimicrobial activity against *E. coli* (Zamble et al., 2001).

Terminal truncations of this TOMM have also been produced, both by *in vivo* biosynthesis of truncated precursor genes and by organic synthesis of variants(Collin et al., 2013;

Thompson et al., 2014; Shkundina et al., 2014). Most of these variants also reduced overall activity, but a synthetic variant without the N-terminal "tail" containing a stretch of nine glycines was a stronger inhibitor of DNA gyrase than Microcin B17 (Thompson et al., 2014). Despite this increased activity against the known *in vivo* target, this compound had lower antimicrobial activity against *E. coli* independently of the reduced cellular uptake of variants without this "tail" (Shkundina et al., 2014). Surprisingly, substitution of Ser26 with an oxazole (which increases the antimicrobial activity of the full-length version) in the"tail"-less variant, led to decreases in both gyrase inhibition and activity against *E. coli* (Thompson et al., 2014).

Other TOMM synthase complexes proved to be more tractable to novel TOMM production than the *E. coli* system, especially the cyanobactin synthase from *Prochloron sp.* The whole pathway was introduced into *E. coli* for *in vivo* production and saturation mutagenesis was performed at selected sites. Mutations at five out of seven sites within the trunkamide macrocycle resulted in detectable products, showing that the enzymes will tolerate substitutions to most amino acids at positions that are not crucial to the macrocyclisation process (Ruffner et al., 2014). One notable exception that is not tolerated at these five sites is cysteine, which is likely to interfere with the macrocyclisation process when it is converted into a heterocycle. The chemical space that can be accessed by this pathway was further expanded by the incorporation of non-proteinogenic amino acids into novel TOMMs (Ruffner et al., 2014). The catalytic promiscuity of the cyclodehydratase activity in *Prochloron sp.* was independently demonstrated by the cell-free translation system developed by Goto et al. (2014), which produced linear TOMMs containing up to 18 heterocyclised residues, compared to only four modified residues in the wild-type product.

While not possessing the wide substrate tolerance that was demonstrated for *Prochloron sp.* enzymes, mutagenesis was carried out for TOMMs of other organisms and the novel products had their activities compared against their wild-type forms. An engineered *Streptomyces coelicolor* strain was transformed with the production pathway for the antibiotic GE37468 from *Streptomyces sp.* ATCC 55365 and 133 variants of the substrate peptide (Young et al., 2012). Out of this repertoire of mutant peptides, 29 led to succesful production and export of a TOMM product. These were purified for detection of antimicrobial activity against *B. subtilis* and MRSA, which was retained by 12 of the mutants, including one with a 2-fold increase in activity compared to the wild-type (Young et al., 2012).

A similar experiment with Plantazolicin A production in *E. coli* also produced novel processed TOMMs, but none had improved activity over the wild-type antibiotic (Deane et al., 2013). Of special interest is the fact that the strongest negative effects in the plantazolicin were observed for mutations in the heterocyclised sites, demonstrating that these post-translational modifications are directly related to the function of this compound.

In the published literature to date, there are no reports of novel synthetic TOMMs being generated by directed evolution. However, the TOMM biosynthesis pathway meets all the requirements for directed evolution to be possible (See Section 1.5 for an overview of directed evolution). The fact that the substrate peptides are genetically encoded enables the generation of diversity by well-established molecular biology techniques and *in vivo* production of variants can exploit whole cells as a genotype-phenotype linkage. All that is needed to perform a directed evolution experiment is the development of a system in which the synthetic TOMM libraries are synthesised and selected for an activity of interest while maintaining a genotype-phenotype linkage — by physical compartmentalisation or localisation of modified peptides to the interior or surface of cells.

Directed evolution has been successfully employed to develop novel thioether-macrocyclic compounds — similar to TOMMS in being natural products derived from post-translationally modified peptides — for use as kinase inhibitors (Hayashi et al., 2012). These molecules contain only a single post-translational modification, which is a macrocyclisation between the N-terminal non-proteinogenic amino acid N-(2-chloroacetyl)-tyrosine and a cysteine residue near the C-terminus. Libraries of mRNAs coding for short peptides were translated in an *in vitro* system that resulted in the mRNA being linked to the peptide C-terminus, providing the genotype-phenotype linkage. The target kinase was immobilised on magnetic beads, which allowed the capture of positive clones while washing out the non-binders. The selected sequences could then be amplified by PCR to start a new round of selection (Hayashi et al., 2012).

1.3 TOMM biosynthetic machinery

1.3.1 Organisation and Occurrence

All characterised TOMM biosynthesis pathways are genetically encoded in clusters containing all genes needed for correct production and function of their respective compounds. Though the order of the genes within the cluster may vary, the most common organisation for TOMM clusters is depicted in Fig. 1.4A. The substrate peptide-encoding gene is usually located immediately upstream of the heterocyclisation genes. The gene coding for the dehydrogenase is followed by the two genes responsible for the cyclodehydratase activity and these two activities represent the common core of TOMM biosynthesis. Downstream of these genes are the sequences coding for auxiliary functions needed for biosynthesis, such as proteases to cleave the leader peptide, transport proteins to export products from the cell, other post-translational modification enzymes and proteins needed for self-immunity to toxic products.

Out of all studied TOMM synthase systems, the cyanobactin synthesis pathways from *Prochloron sp.* are the ones which have been more thoroughly explored, especially the genes involved in functions other than heterocyclisation. The patellamide cluster in *Prochloron sp.* is composed of seven co-directional genes, patA through patG. The first gene in this

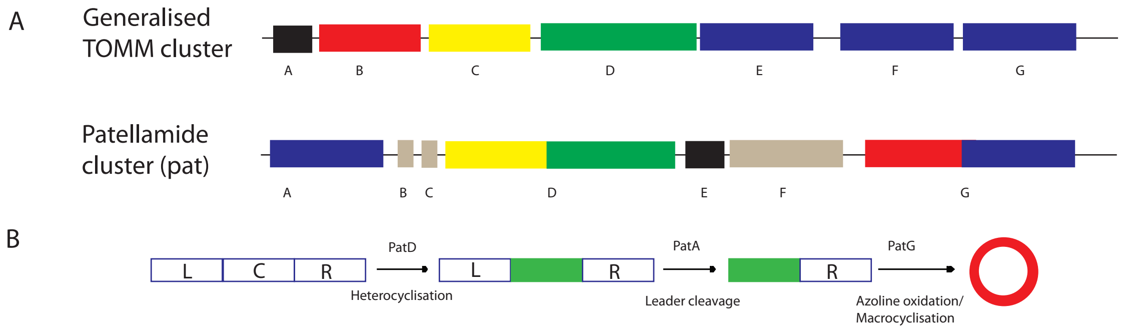


Figure 1.4: Comparison between TOMM and patellamide clusters and mechanism. A - top - Generalised TOMM synthase gene cluster with genes labeled according to Arnison et al. (2013). Components found in all such clusters are a substrate peptide (black) a dehydrogenase (red), a docking domain (yellow) and a cyclodehydratase (green). The remaining genes encode genes responsible for modifications specific to each cluster (blue). Bottom - the Patellamide A biosynthesis cluster from *Prochloron sp.*: The first gene is a protease responsible for leader peptide cleavage, followed by two short proteins of unknown function (grey), a bifunctional protein containing docking and cyclodehydratase domains, the substrate peptide, a protein of unknown function with homology to prenylases (grey) and a bifunctional protein containing dehydrogenase and macrocyclase activities. B - Steps in patellamide maturation. Substrate peptides containing leader peptides (L), core peptide region (C) and recognition sequences (R) are initially heterocyclised on the core region the PatD, generating azolines in the region (green). Then, the leader peptide is cleaved by the action of PatA, followed by oxidation and macrocyclisation by PatG, which leads to the removal of the C-terminal region past the core peptide. The final product is macrocyclic and containing azoles or azolines.

cluster codes for a protease which is responsible for cleaving the leader peptide after heterocyclisation has been carried out (Agarwal et al., 2012; Sardar et al., 2014). Immediately downstream are two short genes, patB and patC, which have no known function. PatD is a bifunctional enzyme containing the full cyclodehydratase activity in a single polypeptide, equivalent to the C and D proteins present in other species (Schmidt et al., 2005). PatE is the substrate peptide, composed of a leader peptide with recognition elements for PatA cleavage and PatD heterocyclisation, the core sequence that is modified by PatD and a C-terminal recognition sequence for PatG macrocyclisation (Sardar et al., 2014). The next gene, patF is also of unknown function, but the structure of its protein product has been determined and it is homologous to prenylating enzymes. However, the patellamides are not prenylated, the putative active site residues are mutated in this gene and no enzymatic activity could be detected from the expressed protein, so it does not have a prenylating function in this cluster (Bent et al., 2013). Finally, patG also encodes a bifunctional enzyme encompassing the TOMM dehydrogenase function and the macrocyclisation activity (Agarwal et al., 2012; Koehnke et al., 2012). There are no known transport or immunity mechanisms for the patellamide system.

The genes responsible for TOMM export and immunity in *E. coli* were discovered by deletion analyses. mcbE and mcbF code for the proteins that export Microcin B17 from cells and also prevent the TOMM from interacting with its DNA gyrase target in the producing cell (Garrido et al., 1988). The third immunity related gene, McbG is uncharacterised but its paralogue Qnr has been shown to prevent quinolone inhibition of DNA gyrase by directly binding the enzyme (Tran et al., 2005). This interaction reduces gyrase affinity for DNA, which may prevent the formation of the gyrase-drug-DNA complex that is needed for quinolone activity (Tran et al., 2005). This model of drug resistance fits the observation that McbG complements the resistance provided by the exporters McbE and McbF (Garrido et al., 1988), with only the strains that possess all three genes being highly resistant to the toxicity of their own antimicrobial compound.

1.3.2 Cyclodehydratase/Heterocyclase

Heterocycles are installed in substrate peptides by the cyclodehydratase activity, converting side chain free alcohol or thiol groups into oxazolines or thiazolines by reaction with the main chain carbonyl from the preceding residue. This reaction is ATP-dependent and results in the loss of the mass equivalent of one water molecule per heterocyclised residue (Milne et al., 1998). However, the mechanism by which this reaction occurs is still unclear and the cyclodehydratase activity has not been assigned an EC number.

There are two contradicting mechanisms which have been proposed for the TOMM cyclodehydration reaction, using two distinct enzymes as experimental models: BalhD from *Bacillus sp.* Al-Hakam and the bifunctional cyclodehydratase/docking protein TruD from

Prochloron sp.. Experiments with BalhD detected ADP formation during cyclodehydration reactions, which led the authors to propose a mechanism in which ATP directly phosphorylates the amide carbonyl in the substrate backbone, activating it to drive the reaction to heterocyclisation (Fig. 1.5 A, Dunbar et al. (2012)). Activity assays with the TruD enzymes led only to AMP formation and the mechanism proposed for this system is the adenylation of the carbonyl, releasing pyrophosphate rather than free phosphate (Fig. 1.5 B) (Koehnke et al., 2013).

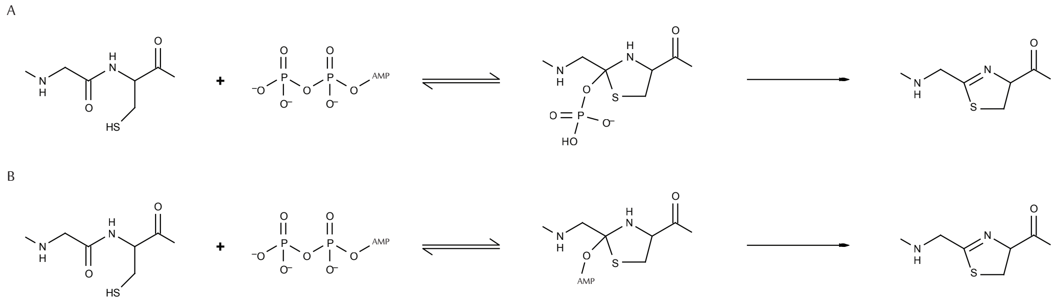


FIGURE 1.5: Simplified comparison of proposed mechanisms for the cyclodehydration reaction in TOMM synthesis. A - Phosphorylation mechanism proposed by Dunbar et al. (2012) and leading to ADP production. B - Adenylation mechanism proposed by Koehnke et al. (2013) and leading to pyrophosphate production.

A crystal structure of a TOMM D homologue from *E. coli*, YcaO, was obtained in the *apo* form and also as a complex with AMP (Dunbar et al., 2014), which could be evidence in favour of the proposed adenylation mechanism, since that model predicts an AMP complex as an intermediate in the reaction. However, this homolog does not have cyclodehydration activity and no AMP-bound TOMM cyclodehydratase structure has been obtained yet.

Apart from the proposed reaction mechanisms, there are still several poorly characterised aspects of cyclodehydratase function, such as the role of the C docking domain for the activity, how the leader peptide is recognised before cyclodehydration and the drastic differences in regioselectivity of heterocyclised residues between enzyme complexes from different species.

1.3.2.1 C docking domain

Most of the cyclodehydratases that have been studied so far are active only in the presence of a C docking protein, or are directly fused to a C domain as in the cyanobactin enzymes (Li et al., 1996; Schmidt et al., 2005; Lee et al., 2008; Scholz et al., 2011). However, the BalhD protein still possesses a reduced cyclodehydration activity in the absence of BalhC (Dunbar et al., 2012). Interestingly, in the absence of BalhC there is deregulated consumption of ATP, with more than one molecule being converted to ADP per heterocyclised residue (Dunbar et al., 2012), which suggests that the C domain is somehow involved in regulating cyclodehydratase activity.

There is also evidence of a role for the C domain in binding both the leader region of the substrate and the cyclodehydratase activity of the D domain, bringing them together for catalysis. Fluorescence polarisation experiments demonstrated that BalhD alone has low affinity for the substrate peptide, but free BalhC was shown to interact with a high affinity towards the substrate BalhA (Dunbar et al., 2014). The interaction between BalhD and BalhC is mediated by a conserved proline-rich motif PXPXP found at the C terminus of cyclodehydratases. Any mutations within this motif led to both decreased heterocyclisation activity and reduced affinity towards BalhC (Dunbar et al., 2014), further demonstrating the importance of the C domain for overall heterocyclisation activity.

1.3.2.2 Leader peptide sequence

While it is clear that the leader peptide region of the TOMM substrate has an important role in substrate recognition, some of the known TOMM cyclodehydratases can carry out heterocyclisations on peptides lacking leader sequences, albeit at lower activities and affinities. In the *Prochloron sp.* cell-free translation and heterocyclisation system developed by Goto et al. (2014), a leader-free substrate peptide was heterocyclised at half of its modifiable residues. Interestingly, expression of two peptides containing an isolated leader peptide and leader-free substrate sequence led to modification of the core peptide with a higher efficiency than when the leader-free substrate was expressed alone, which indicates that the leader can perform part of its function *in trans* (Goto et al., 2014).

Similar results were observed with leader-free substrate peptides for the *Bacillus sp.* Al-Hakam system, heterocyclisations can be observed but the affinity of the substrate towards the modifying enzymes is much lower than that observed for the wild-type peptide. (Dunbar and Mitchell, 2013). Additionally, the removal of the leader peptide negatively affected the subsequent dehydrogenation step, as a complex mixture of oxidation products was observed when the B dehydrogenase was added to reaction mixtures (Dunbar and Mitchell, 2013). While there is a clear recognition function for the leader sequence in TOMM biosynthesis, the complex effects on the dehydrogenase reaction and the trans-activating evidence point toward as-yet unexplained functions for this region in the overall modification process.

1.3.2.3 Sequence specificity in the core peptide

One aspect of the cyclodehydratases which has not yet been well characterised is the mechanistic basis for the wide variability that can be observed in modified sequence selectivity between the known TOMM synthase systems. Natural TOMM products differ widely in the number of consecutive heterocycles found in the final product; patellamides have only isolated heterocyclic moieties, microcin B17 and BalhA have bisheterocycles and plantazolicin A has five consecutive modified residues (See Fig. 1.1 for structures). The tolerance of the respective biosynthetic enzymes towards changes in the substrates were probed by mutagenesis studies, revealing differences in cyclisation site selectivity in enzymes from different species.

Both bisheterocyclic sites in the Microcin B17 product are flanked by glycines, but expression of substrate variants containing mutations at these residues demonstrated that only the N-terminal glycine is absolutely required for processing while the downstream position is more flexible (Belshaw et al., 1998; Sinha Roy et al., 1998). This was interpreted by the authors as a possible steric or conformational requirement for a small group at the position where the first heterocycle is installed, however this condition is not observed in TOMMs produced by other organisms. The BalhA2 peptide from *Bacillus sp.* Al-Hakam has heterocyclisations preceded by aspartate, lysine, proline and glycine, which shows that charge, size of side chain and conformational flexibility are not absolute requirements for heterocyclisation by this system.

The tolerance for variability at the heterocyclised sites similarly varies between species. The PatD heterocyclase has a very high substrate flexibility *in vitro*, generating linear products containing up to 18 heterocyclised residues and also variants of its wild-type product having all heterocyclisation sites mutated to cysteines, serines or threonines (Goto et al., 2014). Meanwhile, an *in vivo* screen of 40 plantazolicin A mutants at heterocyclised sites only led to 8 detectable products, 2 of which were produced at a high enough titer for antimicrobial activity assays (Deane et al., 2013). The low success rate in the plantazolicin mutants may reflect a lower tolerance for non-native sequences in the core peptide, but it could also be a failure of the *in vivo* system to process and export at a step downstream of cyclodehydration.

1.3.3 Dehydrogenase

The TOMM dehydrogenase is an FMN-dependent enzyme and is responsible for the oxidation of azolines into their respective azoles. In most organisms it is expressed a separate polypeptide, except for *Prochloron sp.*, which has this activity as part of a bifunctional protein (PatG) which also possesses macrocyclase activity. The *E. coli* Microcin B17 synthase complex is inactive in the absence of this protein (Milne et al., 1999), but other complexes have been shown to stably produce the reduced heterocyclised product in the presence of only the cyclodehydratase activity.

Attempts at expressing the BalhB protein from *Bacillus sp.* Al-Hakam only led to inactive forms of the protein, probably due to a lack of the required FMN cofactor. To circumvent this issue, the highly similar BcerB from *B. cereus* 172560W was expressed and it was capable of substituting for the BalhB protein in oxidising the BalhA heterocyclised peptides (Melby et al., 2012). To further support the validity of assembling this cross-species system, assays were performed using a dehydrogenase from an actinobacterial species and even in this cross-phylum hybrid system, the results obtained were equivalent (Melby et al., 2014).

A conserved Lys-Tyr motif located near the FMN cofactor was identified by alignment of dehydrogenase sequences to a nitroreductase with a known structure and mutants were generated to prove its catalytic importance (Melby et al., 2014). BcerB and homologues from *E. coli* and the archaeon *Sulfolobus acidocaldarius* all lost catalytic activity when both residues in the motif were mutated. However, the BcerB double mutant was still able to interact with BalhC and BalhD, further supporting the catalytic importance of these two residues for dehydrogenase activity (Melby et al., 2014).

1.4 Computational tools for protein characterisation and engineering

Despite the recent advances in the characterisation of the TOMM biosynthetic enzymes, little is still known about mechanistic aspects of enzyme function and this knowledge would be of great use in engineering TOMM production. There are a variety of computational tools to predict functional sites within protein sequences, which can then be experimentally validated and later used in efforts to improve the enzymes themselves. The choice of methods to be employed in the engineering of any specific system depends on the depth of knowledge that is already available, especially structural and mechanistic information. Here, a functional residue will be defined as any residue that, when mutated, has an impact on a known function of a protein, including catalytic activity, substrate recognition, allosteric effects, conformational changes, protein-protein interactions and structural integrity.

1.4.1 Structure-based methods

Structural data, including structures of enzyme-substrate and enzyme-inhibitor complexes, is a valuable source of information to guide protein engineering efforts. For systems in which such data is available, a diverse array of tools can be used to infer residues involved in catalysis, substrate recognition and overall protein stability.

Methods based on predicting the effect of mutations on known structures are especially relevant for reducing the experimental effort in the engineering of existing proteins. Among the strategies developed for this goal are approaches that extract rules from sets of known mutation effects (Sunyaev et al., 2001; Hurst et al., 2009), calculate the energetic impact of given mutations and search through conformation space to determine the overall impact on a protein (Lilien et al., 2005; Schymkowitz et al., 2005), and also approaches that integrate multiple sources of information to improve prediction accuracy (Karchin et al., 2005). These tools have many applications, including the engineering of protein active sites to create affinity towards non-cognate substrates (Chen et al., 2009a) and predicting the effect of a mutation on the stability of a protein (Guerois et al., 2002).

Beyond the effects of single mutations, molecular dynamics can be used to simulate interactions of every atom in a protein to provide insights into its functional mechanisms Kiss et al. (2013). However, the computational complexity involved in these simulations often requires the use of powerful clusters to produce simulations at timescales relevant for biological processes (Hospital et al., 2015) and *ab initio* predictions of protein folding are so far only possible for short proteins (Dill and MacCallum, 2012).

Alternative strategies have been developed to circumvent the need to calculate all interactions in a protein by employing knowledge from known protein structures to predict structure and function, such as the Rosetta package (Kaufmann et al., 2010). This tool has been successfully used for *de novo* design of synthetic proteins with desired structures (Lin et al., 2015) and even novel enzymatic activities that do not exist in any natural enzyme (Siegel et al., 2010).

Computational tools also exist to evaluate the ability of small molecules to interact with proteins by docking (Morris et al., 2009). This strategy enables the detection of binding pockets for small molecules (Fukunishi and Nakamura, 2011; Heo et al., 2014), information which can be used to predict the binding site of a protein that could not be crystallised in complex with a known ligand (Singh et al., 2011). In addition, these tools also enable the design of novel inhibitors for proteins of clinical relevance, by *in silico* screening of candidates and selecting for the strongest interactions (Velmurugan et al., 2014; Singh et al., 2016).

Methods such as the ones cited above can produce accurate predictions and, therefore, require a relatively small experimental validation effort to obtain the desired characterisation or engineering results (Bommarius et al., 2011). However, as mentioned, they can also require extensive structural and mechanistic information that is not available for many proteins. Of the known TOMM synthases, no structure has yet been determined of an enzyme in complex with its substrate or a substrate analog. The closest equivalent to this is the YcaO domain of unknown function, which is homologous to the cyclodehydratase proteins, and has been crystallised in complex with AMP and an ATP analog (Dunbar et al., 2014). This complex is of limited value for engineering of a cyclodehydratase, since there is no information regarding the binding site for the heterocyclised substrate or the native activity of the YcaO domain.

1.4.2 Sequence-based methods

For systems with not enough structural information to employ the strategies mentioned in Section 1.4.1, purely sequence-based methods can be an alternative to obtain candidate functional residues for characterisation and later engineering (Bommarius et al., 2011). These methods invariably require the construction of multiple sequence alignments (MSAs) containing the sequence of interest and a diverse set of homologous sequences. The constantly expanding genomic (Benson et al., 2013) and metagenomic (Mitchell et al., 2016) sequence databases are an invaluable resource in the construction of these alignments, as they cover a much greater portion of the tree of life than could be achieved a decade ago, before the availability of Next-Generation Sequencing technologies (NGS) (Metzker, 2010).

Extant functional proteins that share a common ancestor sequence represent a subset of the possible sequence space that has been explored over evolutionary time starting from the ancestral sequence. Any highly deleterious variant that arose would be eliminated by the process of natural selection and functional variants would be maintained, to be sampled in the genomes of living organisms. Homologous proteins with a common function are also likely to have closely-related structures and are classified in families (Mulder and Apweiler, 2002). Once the a sequence dataset is constructed for a chosen family, two main evolutionary signals — yielding complementary information about the sequence-function relationship in the family — can be extracted from MSAs using metrics: conservation (or its opposite, diversity) and coevolution of columns (Fig. 1.6).

1.4.2.1 Conservation detection metrics

The first aspect that can be investigated through large MSAs is the overall conservation of each column in the alignment (See Fig. 1.6 for an example of a conserved MSA column), which can be valuable in identifying residues that have been evolutionarily constrained by purifying selection, due to a required function for that residue in the function shared by the family (Echave et al., 2016). Conserved residues have been shown to play important roles in catalytic activity, protein-protein interfaces and directing the correct folding of proteins (Bharatham et al., 2011). A variety of different strategies have been proposed to

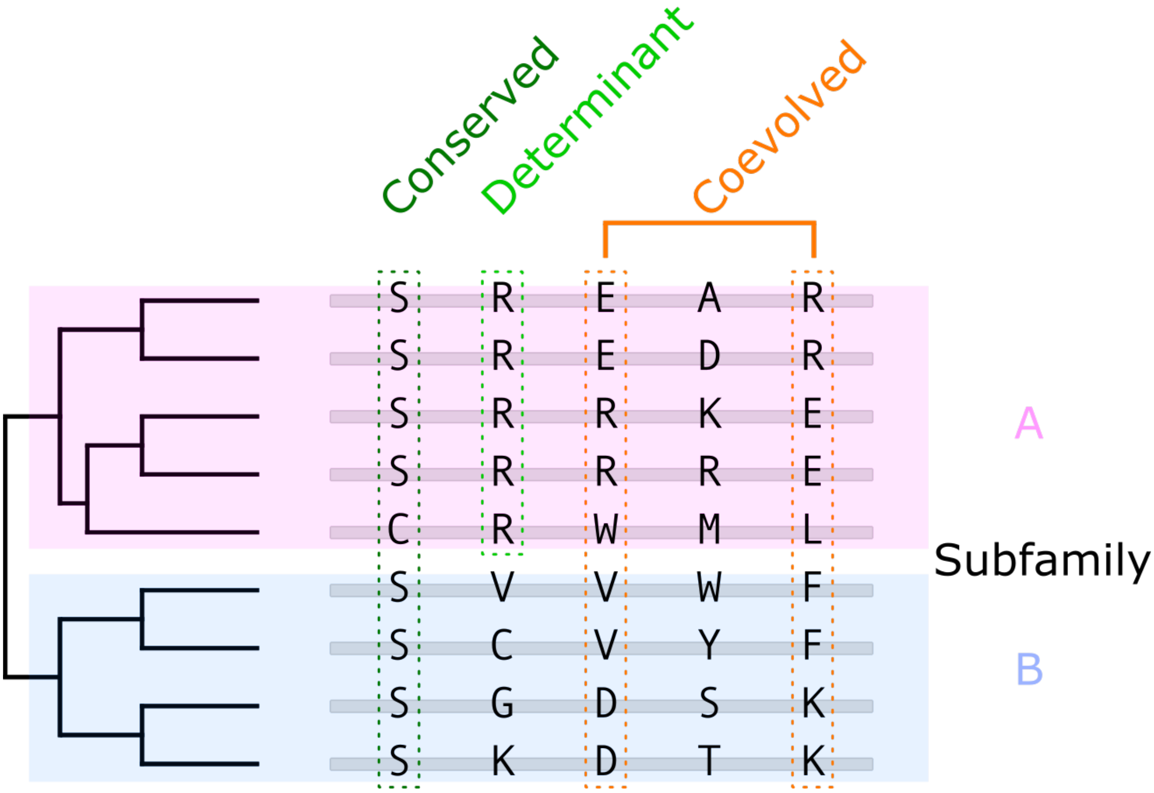


FIGURE 1.6: Evolutionary signals that can be detected in an MSA - Two homologous subfamilies are depicted, with a conserved column highlighted in dark green and a column showing signs of coevolution highlighted in orange. The column highlighted in light green for subfamily A represents a determinant residue for this subfamily, with no signs of conservation in subfamily B.

detect evolutionary conservation of an MSA column, calculating scores based on comparisons between the physicochemical characteristics of the residues, the use of metrics from information theory to estimate the degree of variability, or reconstruction of the evolutionary history of the family to infer evolutionary rates as a measure of conservation (Valdar and Thornton, 2001; Valdar, 2002).

When sequencing technologies started to become accessible and sequence databases large enough to construct alignments of homologous protein sequences appeared, comparisons based on physicochemical characteristics of amino acids were an early attempt to highlight column conservation (Valdar, 2002). One such strategy was based on Venn Diagrams showing the relations between several properties — among them, side chain size, polarity, charge, and aromaticity — and visual highlighting in MSAs or textual descriptions of the conservation within a column based on the shared characteristics identified for the observed residues (Taylor, 1986). Another method produced a numerical descriptor for conservation calculated from the number of shared physicochemical characteristics in a column, arbitrarily assigning an equal weight to every difference or similarity (Livingstone and Barton, 1993). However, the subjective nature of the textual descriptions and the arbitrary nature of the quantification led to the development and more widespread adoption of computational metrics that take into account evolutionary processes that can be inferred from alignments and, ultimately, are more efficient at detecting conserved residues (Capra and Singh, 2007).

The simplest procedures to incorporate evolutionary processes into MSA conservation measures are based on comparisons of residue frequency distributions in MSA columns to "background" distributions, which are representative of the frequency of all amino acids in sites not under selection pressure. These distributions are computed from substitutions matrices used in sequence alignment, such as BLOSUM62, that encode the probability of mutations from any residue to all other residues (Henikoff and Henikoff, 1992). Metrics based on information theory can then be used to quantify the degree of conservation within a column by measuring the divergence between the observed residue distributions in the MSA columns to the background distribution — a high divergence from a background distribution indicates a high degree of conservation. Among these are Relative Entropy (Wang and Samudrala, 2006) and Jensen-Shannon Divergence (JSD) (Capra and Singh, 2007). Capra and Singh (2007) compared the performance of these two metrics along with others and found that JSD robustly detected known conserved residues from benchmark MSAs.

Since the background distributions used for the information theory-based were computed from a large set of proteins, they will not necessarily be an accurate representation of the frequency distributions found in neutral sites of a specific protein family. For instance, transmembrane regions of proteins are known to contain higher frequencies of hydrophobic residues than cytosolic regions of the same proteins, due to the different environments to which these regions are exposed (Bordner, 2009). Therefore, methods were developed to quantify conservation by comparison to column evolutionary rates calculated for the MSA, such as Rate4Sites (Pupko et al., 2002) and ConSurf (Ashkenazy et al., 2016). However, construction of phylogenetic trees for the calculation of these evolutionary rates can be computationally difficult for large sequence sets and the computationally-simple JSD metric was shown to perform similarly to Rate4Site by Capra and Singh (2007).

In diverse protein families composed of two (or more) subfamilies that have diverged in function, residues that are required for the function of one group but not for the remaining ones can be referred to as determinant residues (Bharatham et al., 2011) and are only expected to be conserved in the first group (See Fig. 1.6 for an example of a determinant residues within an MSA). These can also be referred to as Specificity-Determining Positions and methods such as SDPPred have been proposed for their detection, by identifying conserved columns that are conserved within one functional group and divergent between groups (Kalinina et al., 2004). Since this predictor requires an MSA with an explicit subdivision between functional groups, its applicability is limited to families with sufficient annotation for the confident assignment of sequences to each group.

1.4.2.2 Coevolution detection metrics

The other evolutionary pattern that is commonly detected through MSA analysis is coevolution (See Fig. 1.6 for an example of a pair of coevolving MSA columns), yielding information that is complementary to what is obtained from conservation measures. By locating columns in the MSA that present covariation with one or more other positions, it is possible to detect residues that are free to vary but only when there are compensatory mutations at other positions in the sequence. Algorithms to detect coevolution perform pairwise comparisons of amino acid frequencies at columns and measure how the composition of one column influences the distribution of frequencies of the second column (de Juan et al., 2013). These comparisons can reveal global networks of interactions, direct or mediated through intervening residues (de Juan et al., 2013). Coevolutionary interactions often correspond to physical proximity between residues — indicating a shared function for coevolving residues, such as maintaining structural interactions or shaping the active site for diverse substrates within a family — but coevolutionary interactions between physically distant residues can also be involved in allosteric effects (Lockless and Ranganathan, 1999; Süel et al., 2003).

As was described for conservation measurements, metrics from information theory — have been employed for the detection of networks of coevolving pairs in MSAs, successfully predicting the position of catalytic sites (Marino Buslje et al., 2010). An alternative strategy, based on modelling the exploration of sequence space during protein evolution as a particle in statistical mechanics — named Statistical Coupling Analysis — has been shown to detect indirect interactions between residues, coinciding with know allosteric interactions in proteins (Lockless and Ranganathan, 1999; Süel et al., 2003).

Alternatively, methods have been devised to reduce the influence of indirect interactions on the covariation signal, with the aim of predicting physical interactions between residues. Direct Coupling Analysis (DCA) measures correlations between all possible residue pairs to measure the fraction of MI that is produced by physically-interacting residues and can predict protein-protein interactions (Weigt et al., 2009), as well as direct contacts within a single protein (Morcos et al., 2011). Further improvements to direct contact detection were achieved by a procedure named MetaPSICOV, which combines DCA with other coevolution calculations along with other alignment metrics and secondary structure prediction into a neural network, generating more accurate contact predictions than any of the metrics in isolation (Jones et al., 2015)

1.5 Directed Evolution for the engineering of proteins and other biological systems

Traditionally, two routes were recognised as viable paths for the engineering of biological systems towards phenotypes of interest: rational design — in which multiple sources of knowledge (biochemical annotation, structural information, computational predictions) are combined to generate mutations of predictable effect — and directed evolution — in which superior phenotypes are screened or selected from a diverse pool (Nixon and Firestine, 2000; Chen, 2001). However, these two extreme views are oversimplifications and neither pure option is the ideal strategy for engineering most biological systems (Lutz, 2010).

For well-characterised systems, with known functional mechanisms and structural information to guide design, the prediction and design strategies described in Section 1.4.1 have been successfully applied to produce the desired phenotypes (Siegel et al., 2010; Lin et al., 2015). However, for any system without extensive characterisation, the available prediction strategies can only direct attention towards regions of sequences likely to be involved in function, requiring further experimental data to enable precise prediction of the functional impact of mutations or the use of directed evolution approaches to reduce the number of variants to be tested.

Directed evolution is a widely-adopted strategy for the engineering of biological systems that is analogous to natural selection because they share the same pre-requisites: a heritable phenotype, genetic diversity encoding diversity in the phenotype, and a process that biases the chance of a genotype being present as a function of the encoded phenotype (Lane and Seelig, 2014; Packer and Liu, 2015; Tizei et al., 2016). The phenotype can be a single macromolecule such as a protein encoded by DNA, a functional nucleic acid that can be both phenotype and genotype, a metabolic pathway encoded by a set of genes, or even an entire cell encoded by its genome. Genetic diversity can be generated experimentally or arise spontaneously as a result of cell division, with the population of variants used in an experiment being called a library. The final element is any mechanism that can ensure an uneven partitioning of the diversity present in the population, such that the desired phenotype is enriched relative to the bulk of the population.

In natural selection, the partitioning mechanism is differential survival and reproduction of the organisms which posses the phenotypes most adapted to their environment. The offspring will carry the genotype which encodes the selected phenotype, propagating it into the next generation. In directed evolution, partitioning strategies must be designed with a connection between the desired phenotype and the genotype which encoded it — a genotype-phenotype linkage — to ensure recovery of the genotype for subsequent cycles of evolution (Tizei et al., 2016; Leemhuis et al., 2005). The nature of these mechanisms can vary greatly depending on the phenotype being selected and will be covered in Section 1.5.2.

In order for a phenotype to be carried forward in a round of directed evolution, it must be present in the diversity pool at the start of the cycle and this requirement creates an upper limit on the size of library which can be used. Biological sequence space — for nucleic acids and proteins alike — is vastly larger than what can be covered in any feasible experiment. For a relatively small protein with 200 residues in its sequence, the number of possible combinations of all 20 amino acids is 20^{200}, which vastly dwarfs the estimated number of atoms in the observable universe — around 10^{80}. This means that it is physically impossible to produce every single possible variant of such a target molecule, let alone select a library of that size. Therefore, diversity generation strategies for directed evolution must restrict their exploration of sequence space to reduce the experimental effort in selection and increase the likelihood that the desired phenotype can be isolated.

The simplest solution to this problem is to reduce the sequence length until all possible variations of the sequence can be completely covered by existing selection strategies, which range from roughly 10^8 for *in vivo* methods (Tee and Wong, 2013) up to 10^{14} for *in vitro* selection (Hoinka et al., 2015a). This can be done for nucleic acid aptamer selections, since a fully-randomised stretch of 20 nucleotides — with about 10^{12} (4^{20}) possible variants — can be covered by SELEX starting from 1 nmol (>10^{14} molecules) of a commercially-synthesised pool of oligonucleotides containing 20 N incorporations. However, not all the desired functions for aptamers can be produced by such short sequences and this restriction is even more severe for proteins with their alphabet consisting of the 20 proteinogenic amino acids — a sequence of 12 residues already has over 10^{15} possible variants, while the shortest known functional domains are at least 30 residues long (Jones et al., 1998). Since exhaustive searches of sequence space are impossible for most phenotypes of interest, rational approaches are frequently employed to restrict diversity to shorter regions within a larger sequence, reducing the number of possible variants, and increasing the probability of obtaining the desired phenotype.

In most cases, one or more known natural molecules already possessing the desired function to some degree are already known and can be used as starting points to restrict the search space for a directed evolution experiment (Arnold, 1998; Crameri et al., 1998). This approach has permitted the isolation of highly active enzymes starting from promiscuous natural enzymes with poor activity on the desired substrate (Aharoni et al., 2005) and recently led to the engineering of the first known protein able to catalyse the formation of a carbon-silicon bond (Kan et al., 2016a). Some of the computations tools mentioned in Section 1.4 have achieved a degree of success in designing entirely novel proteins with desired activities (Siegel et al., 2010), but the use of natural proteins as starting points in directed evolution experiments is still by far the most common approach.

The field of Synthetic Biology often uses a cycle comprised of Design-Build-Test-Learn steps to illustrate the process of engineering biological systems, which can be seen as analogous to the processes involved in carrying out directed evolution and will be used to guide the explanation of these processes here (Fig. 1.7) (Baldwin et al., 2015; Tizei et al., 2016).

“Design” encompasses the decisions made in targeting diversity in the genotype, with the aim of producing a library containing the desired phenotype. “Build” represents the experimental procedures to generate the library. “Test” is the partitioning step, in which variants are screened or selected to enrich for the phenotype of interest. Finally, the enriched population is analysed in the “Learn” phase to identify successful variants and direct further rounds of evolution. The description will start with the Build phase, which will be given additional emphasis due to development of a library assembly strategy for engineering of synthetic TOMM products.

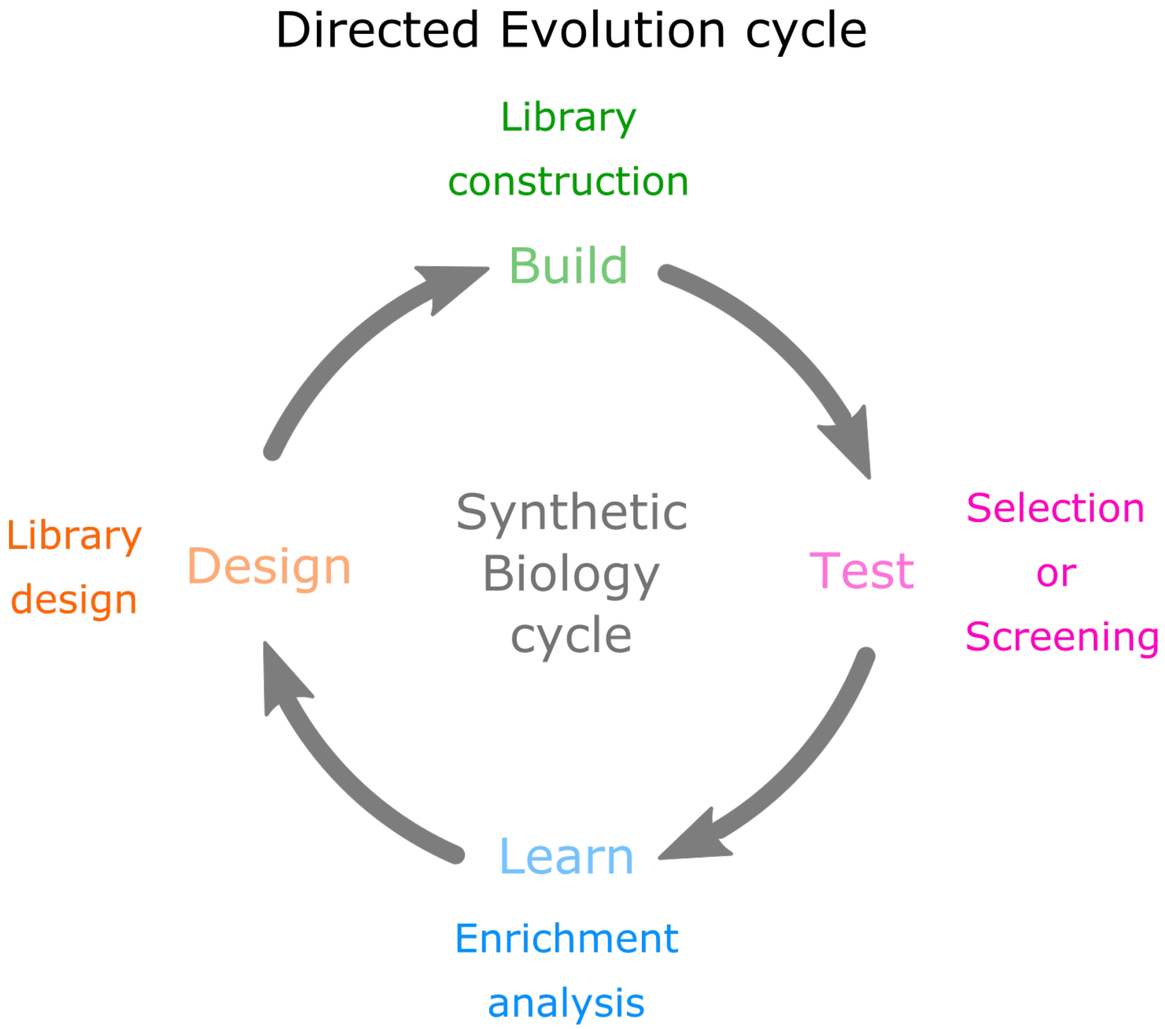


Figure 1.7: Steps involved in a round of directed evolution compared to the synthetic biology cycle for engineering biological systems.

1.5.1 Build: Diversity generation for directed evolution

As described above, a pool of variants is an integral part of directed evolution but the vast size of biological sequence space is an impediment to the complete exploration of possible sequences for most relevant targets. This conflict is resolved by employing strategies to target diversity towards subsets of the possible sequence space for a given target. Such approaches are possible due to great heterogeneity that can be observed when comparing the effects of mutations throughout a sequence. Not all possible mutations to a given biological sequence will have a positive impact on the phenotype of interest. In fact, most mutations to a given sequence are expected to have a deleterious or at least neutral effect on any given phenotype and only a small fraction are expected to contribute positively (Ohta, 1992). Prior knowledge about the relation between genotype and phenotype in the system of interest can be of great use in targeting relevant sites and, therefore, reducing the size of the search space. This knowledge can come from a variety of different sources such as structures, phylogeny (using methods such as the functional predictions in Section 1.4), functional annotation, biochemical characterisation coupled to site-directed mutagenesis, and results from previous engineering efforts for similar phenotypes.

If the available knowledge about the targeted system cannot yield any relevant insight into the genotypic contributions to the phenotype — or if the goal is to obtain more information about the system to then inform future targeted rounds of diversification — "blind" exploration of the sequence space can be employed. In this strategy, diversity is introduced without any intentional bias at any point of the chosen sequence, submitted to selection, and any positive variants obtained are carried on to further rounds of diversification and selection. Mutations that produced positive phenotypes in separate clones can be combined in a single variant, which would be expected to produce a superior phenotype if the effects of all the mutations are additive. However, this is frequently not the case and many sites will have epistatic interactions with other regions, producing greater and smaller effects than would be expected by simple addition of the effects of the single mutations.

The strategies outlined above represent extreme cases of available knowledge about the target system, with either enough information to direct mutations to a small set of specific functional residues or a complete lack of useful information that forces experimenters to apply untargeted diversification methods to the entire sequence. However, real directed evolution efforts tend to fall somewhere in between the two extremes. Even for relatively uncharacterised systems, there often is enough public information available to direct diversification to only a region of the whole sequence representing a functional domain or to predict a repertoire of candidate mutants that can be validated by selection. In the cases where these approaches do not yield reliable candidates, the results of a first round of untargeted diversification followed by selection can be used to focus deeper exploration of the sequence space on the regions around the positive hits. Whichever higher-level strategy is selected for introducing diversity into a directed evolution experiment, the repertoire of methods used for physically producing the desired variants is, in itself, extremely diverse and constantly being supplemented with novel techniques.

1.5.1.1 Overview of commonly-used diversity generation methods

The techniques which can be employed to produce the sequence diversity vary greatly in terms of experimental complexity, cost, range of selection platforms to which they can be adapted, and the degree of control over the exploration of sequence space. The choice of a method (or combination or methods) for the directed evolution of a system depends on the time and resources available for the project, the size of the sequence that is targeted, and the complexity of the desired phenotype.

The simplest way to obtain diversity for selection is by exploiting the variation that naturally arises as cells divide and — with varying degrees of fidelity — replicate their

DNA, accumulating mutations throughout the genome. This requires no input other than cells that express the phenotype to be engineered and means for the population to expand, such as sequential batch cultures or continuous growth in bacteriostats. In these conditions, mutations should be fairly spread over the entire genome, which makes this method suitable for engineering complex multi-factorial phenotypes which are not completely understood. This approach is referred to as “Evolutionary Engineering” when used to produce microbial strains with phenotypes such as resistance to challenging environmental conditions — such as oxidative stress (Cakar et al., 2005), freezing (Teunissen et al., 2002), solvent tolerance (Stanley et al., 2010) — and more efficient metabolism of novel substrates for growth and production of useful metabolites (e.g. sugars derived from hydrolysis of lignocellulosic biomass (Wisselink et al., 2007; Lee et al., 2016)).

Furthermore, the natural mutation rate can be increased by certain environmental conditions — UV (Fiedurek and Gromada, 1997) and chemical mutagens (Sandana Mala et al., 2001) — or by introducing mechanisms to increase the error rate of DNA replication — such as deleting DNA repair mechanisms (Muteeb and Sen, 2010). However, these exogenous factors to increase mutation rate must be carefully controlled, both to prevent collapse of the population due to accumulation of multiple deleterious mutations (Trindade et al., 2010) and due to the potential personal hazards involved in growing microbes under mutagenic conditions.

Populations of model organisms such as the bacterium *E. coli* and the yeast Saccharomyces cerevisiae generate point mutations at relatively low rates in standard culture conditions without any external mutagens, approximately $1.8 - 2.4 \times 10^-9$ per bp per cell division in *E. coli* (Lee et al., 2012) and 3.3×10^-9 per bp per cell division in yeast (Lynch et al., 2008). This makes *in vivo* directed evolution ineffective for engineering single proteins or phenotypes encoded by a small set of genes, since the vast majority of mutations would not be in functionally-relevant regions. This would require large culture volumes to increase the likelihood of obtaining enough mutations in the targeted regions and could still be susceptible to noisy selection results due to the relatively large number of off-target mutations that could generate unpredictable selection parasites. Therefore, more targeted approaches are often employed when the phenotype is encoded by regions much smaller than a whole microbial genome.

Much as the *in vivo* DNA replication machinery can be forced to increase its error rate, the DNA polymerases routinely used in PCR amplification can be made to have lower fidelity for diversity generation in the process known as error-prone PCR (epPCR). This is most easily done by spiking reaction buffers with Mn^{2+}, with increasing concentrations of this ion producing greater error rates which can be tuned in the range of 1-5 mutations per kb (Cirino et al. (2003) roughly, 10^6-fold higher mutation rate than obtained in vivo). However, there is a detectable bias towards A/T to G/C in the mutations generated by epPCR (Cirino et al., 2003) and the single base-pair mutations produced restrict the range of accessible amino acids in each site of a protein, since a single point mutation in any codon cannot produce codons for the other 19 amino acids. The mutational bias can be reduced using modified nucleotides (Tee and Wong, 2013) or commercial kits which claim to generate even distributions of mutations (i.e. GeneMorph II from Agilent), but the codon accessibility is an intrinsic limitation of epPCR and any other technique which produces only point mutations. Libraries produced by epPCR can also suffer from amplification bias, when some sequences become over-represented in the final library due to differential amplification efficiency caused by variations in GC content or secondary structure (Polz and Cavanaugh, 1998), or due to the stochastic nature of PCR leading to over-representation of sequences that were amplified in the initial PCR cycles (Kebschull and Zador, 2015). These effects can be mitigated by ensuring that the number of initial template molecules is greater than the final number of variants which will be present in the library used in the selection or screening processes (Firth and Patrick, 2005).

In vitro processes can also emulate the role of sexual reproduction in directed evolution, by allowing recombination between variants within a library. The process of DNA shuffling, developed by Stemmer (1994), consists of a DNA fragmentation step and reassembly by a DNA polymerase extending cross-primed fragments. This versatile technique quickly became widely adopted for its use in recombining mutations present in multiple variants obtained in rounds of selection and also for exploiting natural diversity in homologous natural sequences as starting points for directed evolution (Arnold, 1998; Crameri et al., 1998; Aharoni et al., 2005).

PCR can also be exploited to produce targeted mutations by carrying out amplifications with primers containing single mutations or degenerate oligos synthesised with defined mixtures of nucleotides at chosen positions, known as site-directed mutagenesis (Carrigan et al., 2011) (SDM). This strategy avoids the biases and single base-pair mutation disadvantages of epPCR by explicitly picking which mutations will be targeted in a library, which can be in a size range one to tens of base pairs, depending on oligo synthesis limitations, cloning strategy, and the size of the non-mutated priming region that is needed. Commercial synthesis of oligos containing degeneracies can also be exploited to allow multiple possible mutations at one site to be encoded by a single oligo. A fully degenerate codon (NNN) can encode all 20 proteinogenic amino acids in 64 combinations with three of those being stop codons (4.7% of the possible codons), but NNS or NNK (where S = C or G and K = G or T) also cover the full codon repertoire and only one possible stop codon (UAG, 3.1% of the possible codons), so the latter are more frequently used. However, the approximately equimolar amounts of each base incorporated at each degeneracy mean that amino acids encoded by multiple codons will be over-represented, so strategies such as "Small Intelligent" libraries have been developed to ensure more uniform sampling without requiring a full set of 20 oligos to produce an unbiased randomised site in a protein (Tang et al., 2012). Other degeneracy options can be made if the library design does not require full randomisation at a specific site, reducing library size and selection effort.

While PCR-based SDM allows a great deal of control over the sequence that is integrated at each site, traditional methods are limited in only being able to target a single small region of a gene at a time. If another region of a gene needs to be targeted or if a single target is larger than the maximum practical size for an oligo, multiple rounds of SDM need to be carried out sequentially, which can become very labour-intensive and time-consuming. Alternative methods have been developed to multiplex the SDM process, with varying degrees of efficiency and experimental (Seyfang and Jin, 2004; Tian et al., 2010).

With recent developments in oligo synthesis technologies, companies have started to offer custom synthesis of gene libraries. These services allow diversity to be targeted to any position within a gene and fully specified according to the desired library design. While this option can produce high-quality libraries with minimal sequence redundancy and zero experimental effort, their costs are frequently too high for routine use in most academic laboratories.

Advances in *in vivo* genome manipulation from the past decade have also proved useful in library generation, by allowing diversity to be targeted to specific genomic regions. Multiplex Automated Genome Engineering (MAGE) adapts a phage-derived recombinase to replace genomic sequences in *E. coli* with homologous single-stranded oligos containing the desired mutations (Wang et al., 2009). The endlessly-versatile CRISPR-Cas9 is another system that has been adapted to produce targeted point mutations by fusing a catalytically inactive form of the enzyme (dCas9) to the cytidine deaminase enzyme responsible for somatic hypermutation in antibody-producing cells, resulting in a fusion protein that produces point mutations around a region of interest that is targeted by single-stranded RNA (Hess et al., 2016). Using techniques such as these, it is possible to conduct directed evolution rounds fully *in vivo*, with no intermediate cloning steps and a greatly reduced library size when compared to spontaneously-occurring mutations, reducing overall experimental complexity.

Most uses of these technologies have one crucial characteristic in common, which is that the libraries produced by them are designed to contain only compositional variation or substitutions in sequences all having the same length. The term homometric library is proposed to represent these libraries, in contrast to heterometric libraries in which sequences differ in length and may or may not also contain compositional variation. There are two main reasons why substitutions are favoured over insertions and deletions (indels) for directed evolution: the methods needed to explore variation simultaneously in length and composition in the laboratory are not as versatile as the ones described above and indels have a greater chance of being deleterious, on average, than substitutions. However, by not considering length variations in library design, most protein engineering efforts are implicitly stating that the main chain length of the starting protein is the optimal one for the function of interest. In fact, for most cases there is no evidence to support this assumption and there are published examples where the opposite is true — i.e. a length variation is essential for a new or altered function.

1.5.1.2 Length diversity in natural and laboratory evolution

Indels are a very broad class of mutations, in both size and potential phenotypic effect. At the smallest end of the scale, single base-pair indels can be effectively neutral when occurring in repetitive intergenic regions with no detectable function, but they can also be highly deleterious in protein-coding regions by causing frameshifts. This is reflected in the lower frequency of indels compared to substitutions when comparing pairs of genomes of closely-related species across widely divergent groups such as primates and bacteria (Khan et al., 2015; Chen et al., 2009b). The ratio between substitutions and indels (S/I) also varies widely according to genomic region, being generally higher in protein coding regions (Khan et al., 2015; Chen et al., 2009b).

However, as the size of indels in protein coding regions increases, multiple-of-3 bp lengths have a greater likelihood of being selectively neutral or even positive, since they become indels of one or more whole codons that maintain reading frames. The effect of multiple-of-3 bp indels is still widely variable, because changes in the backbone of crucial secondary structure elements can disrupt multiple interactions, while indels within a structurally-flexible loop tend to be more tolerated and can contribute to function. These differences can also be observed in genome evolution, by comparing the frequency of full-codon length indels in sequences coding for functional protein domains to the indels of the same size in regions of DNA that code for non-regular structural parts of proteins, the latter being measurably higher due to the higher likelihood of such indels being selectively neutral or even positive (Khan et al., 2015). At even larger scales of whole protein domains or genes, this class of mutation represents an important source of novel material for selection to act upon, creating combinations of elements that did not exist before, generating redundant copies of elements that loosen purifying selection pressure upon one of the copies or even introducing new sets of genes into a species by horizontal gene transfer (Borneman et al., 2011).

Comparative sequence analysis can also yield insights into how indels in genetic sequences correspond to phenotypic differences. The phosphotriesterase (PTE) enzyme, which is thought to have evolved recently to degrade the organophosphate insecticide paraoxon, is a clear example of this. The sequence of PTE is homologous to the bacterial quorum sensing-related enzyme family called PTE-Like Lactonases (PLL) — so named for their initial description as proteins related to PTE, only later having their function characterized — and their main difference lies in a longer active-site loop in the structure of PTE (Afriat et al., 2006). Each family has strong activity on its own preferred substrate and very low residual activity on the substrate of the other family. However, deletion of the longer loop in PTE, followed by a single point mutation, produced a variant that is active on both substrates — an apparent intermediate in the evolution of PTE from PLL (Afriat-Jurnou et al., 2012).

Comparative genome studies can reconstruct evolutionary events that took place since the divergence of the species or strains under study, which generally takes place in time scales much longer than what can be recreated in a laboratory. However, relatively short laboratory evolution experiments with fast-growing microbes have also produced evidence to support the importance of indels in the evolution of novel phenotypes. One notable example is the Long Term Evolution Experiment by Richard Lenski's group that obtained *E. coli* strains capable of metabolising citrate after nearly 20 years (50,000 generations) of continuous growth. Genome sequencing of the adapted strains detected indels at frequencies greater than 10% that of point mutations (Tenaillon et al., 2016). Moreover, experiments conducted in much shorter timescales also corroborate this observation. Selection of populations of *E. coli* and *Pseudomonas* under carbon source or antibiotic selection produced the desired phenotypes in experiments ranging from 8 to 44 days and all selected strains were shown to have indels in their genomes, ranging in size from a few codons to amplifications of whole genes (Toprak et al., 2011; Wong et al., 2012; Herring et al., 2006). These examples of indels obtained under laboratory conditions were the result of spontaneous *in vivo* mutations with no deliberate interference in the natural mutation rates of the organisms. However, there are methods developed to intentionally produce diversity containing length variations and these have been successfully deployed to engineer function.

1.5.1.3 Experimental strategies to produce heterometric libraries

Though not as broad as the repertoire of techniques available for producing homometric libraries, a range of methods from the toolset of molecular biology have been adapted to produce heterometric diversity. These tools can produce indels in diverse scales and with varying degrees of control over the composition of the output sequence, also differing in terms of experimental complexity and cost. Another important limitation that must be addressed when engineering protein coding sequences is that most indels in protein coding sequences of sizes that are not multiples of three will produce deleterious frameshifts, so methods that can restrict indels to these multiples are inherently more efficient.

As was mentioned for compositional variation earlier, commercial synthesis of libraries represents the least labour-intensive path to obtain length diversity, but comes at a cost that makes it inaccessible to many academic projects. These synthetic libraries can be produced by high-throughput oligonucleotide synthesis in arrays such as the ones sold by Twist Biosciences or generated by combinatorial assembly of a smaller number of oligonucleotides using restriction endonucleases and ligases (Ashraf et al., 2013; Van den Brulle et al., 2008). Synthetic libraries allow the desired diversity to be fully specified and reduce the occurrence of redundant sequences, consequently also reducing the experimental effort needed for screening or selection of positive clones (Osuna et al., 2004).

Adaptations to chemical synthesis methods have been made to create libraries of sequences with their length as a distribution of values rather than a single length per synthesis. One such method, COBARDE, is based on synthesizing trinucleotide building blocks with orthogonal protecting groups that are added to only part of the population after each cycle and then reset, leading to failed incorporations in some cycles for a fraction of the population that produce a distribution of different lengths at the end of the synthesis (Osuna et al., 2004). Changes in reaction conditions can be exploited to tailor the length distribution of the final products. However, this synthetic strategy requires specialized equipment and organic synthesis expertise that is beyond the reach of many academic oligonucleotide synthesis facilities.

Library synthesis requires a degree of information about the targeted system to inform the diversification strategy and reduce the search space. This limitation becomes more relevant when length diversity is included, since each amino acid indel in a position creates a new set of possible variants for the targeted region that contains 20^{n-1} or 20^{n+1} new variants, which quickly grows to intractable numbers as the numbers of indels or positions is increased. Therefore, untargeted methods can be employed to probe the tolerance to indels of an entire protein coding sequence and direct deeper diversification by other methods.

Randomized Tandem Repeat Insertions (TRINS) (Kipnis et al., 2012) produces short duplications in coding sequences by fragmenting a template DNA into smaller fragments, followed by circular ligation of single-stranded fragments and a mixture of PCR and rolling-circle amplification, which produces double-stranded variants of the original sequences containing short tandem repeats. This method was validated by producing a library of TEM-1 β-lactamase variants and selecting for activity on the non-cognate substrate ceftazidime (Kipnis et al., 2012). These duplications can be of any length, so it is expected that two-thirds of them will be function-disrupting frameshifts.

The occurrence of frameshifts can be avoided by strategies that only produce fixed-length indels, such as *in vitro* transposons that insert randomly into target DNA and delete exactly 3 bp when excised by a restriction endonuclease (Jones, 2005). This produces either a deletion of a single codon — when the excision event removes an entire codon — or a deletion and a substitution — when the excision event removes part of two adjacent codons and a new codon is formed by joining the remainders of the original codons. This strategy has been used to isolate a GFP variant with faster folding kinetics, containing a deletion in an N-terminal α-helix (Arpino et al., 2014).

The strategies described until now have focused on the ones that produce indels at scales that are useful for protein engineering, since this was the focus of this work. However, the development of robust genome engineering techniques in the past decade has also produced methods that can generate large indels for the *in vivo* engineering of complex phenotypes. At the moment, CRISPR-Cas9 (Jinek et al., 2012; Chu et al., 2015) is the most used system for producing targeted indels, but other targeted nucleases such as Zinc Finger Nucleases (Bibikova et al., 2003) (ZFNs) and Transcriptional Activator-Like Effector Nucleases (Christian et al., 2010) (TALENs) can also perform the same functions. Insertions are produced by creating a double-stranded break (DSB) at a specific genomic location and introducing the desired DNA fragment containing ends which are complementary to the regions surrounding the DSB, leading to its repair by homology directed repair (HDR). Deletions can be generated by creating a pair of DSBs surrounding the targeted region, which can be repaired by non-homologous end-joining (NHEJ)(Su et al., 2016) or by HDR if a deletion cassette is provided with complementarity to both ends.

1.5.2 Test: Selection and screening strategies for directed evolution

As mentioned earlier, directed evolution requires a partitioning process by which diverse variants within a population are compared and the ones harbouring the desired phenotype can be isolated. A wide range of partitioning frameworks have been developed for directed evolution, but they can be broadly divided into two categories: screening and selection methods (Lane and Seelig, 2014; Packer and Liu, 2015). In screening, properties of each variant in the library are individually quantified and these measurements are used by experimenters to decide which variants will be subjected to further characterisation or be used as templates in a subsequent round of diversification (Packer and Liu, 2015). Conversely, selection methods probe the entire population in parallel by a physical separation mechanism or differential survival of cells carrying the phenotype of interest and experimenters do not have direct control over the individual variants that are carried to the next round (Packer and Liu, 2015). *In vivo* selection processes in directed evolution can also be seen as analogous to the complementation experiments used in genetics for discovery of gene function and some authors adopt the term “complementation” to refer to the partitioning of desired variants in directed evolution (Baker et al., 2002), but the more common “Selection” will be used here for clarity.

Screening and selection processes each have distinct advantages and disadvantages. Since variants are probed individually in a screen, the observed phenotype can be directly connected to its encoding genotype, which is impossible in a bulk selection experiment in which all genotypes are pooled for sequencing at the end of a round. Conversely, the bulk nature of selection processes allows for a much higher throughput per round than is achievable in screening strategies that require physical separation of variants. Here, a greater emphasis will be placed on selection processes, since it was the strategy used in Chapter 6.

Screening vs selection

Any selection strategy can be be visualised as the intersection of two key processes: the degree to which library variants possess the function of interest and the stringency of the method in removing variants that do not have the required function (Fig. 1.8, (Tizei et al., 2016)). A successful selection should efficiently recover functional variants, while rejecting the bulk of the diversity that does not possess the phenotype.

However, this visualisation also illustrates the failure modes that can occur in any selection process as a result of problems with the genotype-phenotype linkage (Fig. 1.8, blue and yellow regions). Functional variants that fail to be recovered by selection are considered false negative results and a high proportion of these can impair the efficiency of the entire directed evolution strategy. False negatives can occur when a functional variant is poorly expressed or the partitioning cutoff point for the desired function has high variability.

Recovery of false positives without the phenotype of interest, on the other hand, can have two distinct origins. The first, named background, is random recovery from the population bulk due to non-specific interactions in the partitioning process. For instance, in selection for affinity towards a target molecule supported on a matrix, insufficient washing can lead

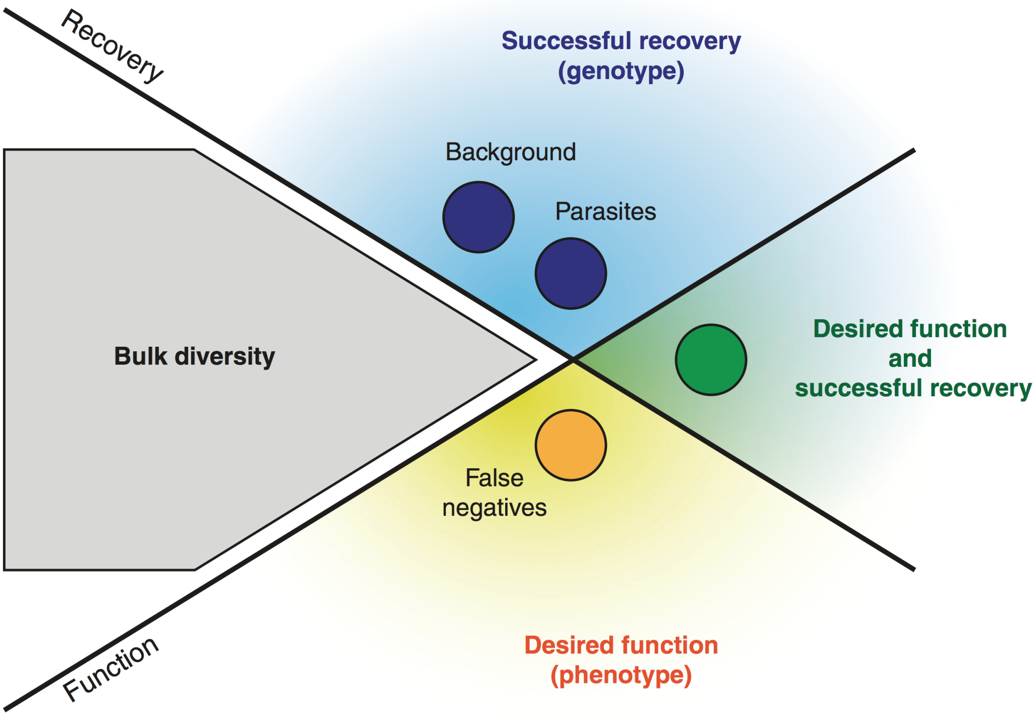


FIGURE 1.8: Possible outcomes in selection experiments. Recovered true positives are represented in green and selection failure modes are represented in relation to recovery and function. False negatives are functional variants that failed to be recovered by selection (yellow). In blue are the recovered false positives, including non-specific background variants and harmful parasites that are actively recovered by unexpected interactions with the selection mechanism. According to (Tizei et al., 2016).

to retention of non-functional molecules due to non-specific interactions with the system (Vant-Hull et al., 2000). Due to the random nature of these interactions, background negatives can be removed by further cycles of selection or changes to selection conditions.

Conversely, parasites are a class of false negative variants that are recovered due to specific interactions with the partitioning mechanism, while not harbouring the targeted phenotype. Following the example given above, a parasite would be a variant that has high affinity for a component of the matrix used to support the target molecule and, therefore, would not be eliminated by washes to remove variants interacting non-specifically. Parasites can be harmful to the selection process if recovered in high proportions relative to the true positives, leading to wasted characterisation effort in determining the phenotype of recovered variants. Possible sources of parasites must be evaluated while the selection strategy is being designed and steps taken to reduce their probability of emergence, but it is impossible to predict all possible sources of parasites and changes to the selection strategy may be needed if parasites do emerge during selection.

The wide diversity of selection and screening strategies that have been established for directed evolution can be divided into four categories, according to the nature of the genotype-linkage: *in vivo*, *ex vivo*, or *in vitro* systems. Each will be described in a separate section.

1.5.2.1 *In vivo* selection systems

In vivo selection systems are closest to natural biological evolution, using living cells as the genotype-phenotype linkage. This enables the engineering of complex phenotypes such as tolerance to environmental conditions or growth using non-natural substrates, with contributions from the entire cellular genome. Selection for growth under specific environmental conditions is the simplest strategy employed in *in vivo* directed evolution and has been used to isolate complex phenotypes. Hoesl et al. (2015) isolated a strain of *E. coli* capable of replacing L-tryptophan in its proteome with the synthetic analogue L-β-(thieno[3,2-b]pyrrolyl)alanine, using serial batch cultures in media containing the synthetic analogue. Using a turbidostat for fine control of cell growth rates and gradual changes to medium composition, Marliere et al. (2011) obtained *E. coli* that incorporate chlorouracil instead of thymine into their genetic material.

However, growth-based selections are restricted to phenotypes that are required for basic cell processes — such as the protein or nucleic acid synthesis examples mentioned above — or at least closely correlated with cell growth, such as metabolic pathways. If the desired phenotype is production of a metabolite that is not needed for cell survival, parasites may arise that are able to successfully grow under selection but do not produce the desired metabolite. A xylose-fermenting *Saccharomyces cerevisiae* strain was engineered with a metabolic pathway for fermentation of the sugar arabinose into ethanol and grown in the presence of this sugar in repeated batch cultures, leading to the isolation of a strain that was able to ferment arabinose but lost the ability to ferment xylose (Wisselink et al., 2007). This example illustrates a warning often repeated in the field of directed evolution that "You get what you select for" (Zhao and Arnold, 1997), i.e. selection conditions must be carefully selected to recover the desired phenotype.

Recently, methods for engineering biosensors for target small molecules have been established, allowing their use in logic circuits to allow for selection of phenotypes not naturally correlated with growth. Such a biosensor for progesterone was coupled to a fluorescent output, enabling the use of Fluorescence Assisted Cell Sorting (FACS) to screen for variants capable of producing higher titres of the hormone (Feng et al., 2015a). Chou and Keasling (2013) coupled an engineered biosensor for lycopene to expression of an error-prone DNA polymerase, leading to high initial mutation rates and a gradual decrease as variants capable of high lycopene yields appeared in the population. Continued development of biosensor technology will expand the range of functions which can be targeted by *in vivo* selection systems.

1.5.2.2 *Ex vivo* selection systems

Ex vivo selection — also known as surface display — platforms also employ whole cells or bacteriophage particles as their genotype-phenotype linkage, but do not exploit growth or metabolic processes for the selection mechanism. Instead, the protein to be engineered is exposed on the surface of the organism, for interaction with the chosen partitioning mechanism. The protein of interest must be fused to a protein naturally present on the outer surface of the host organism, such as a phage capsid subunit (Smith, 1985) or bacterial membrane protein (van Bloois et al., 2011), which also restricts the use of this strategy to proteins that do not disturb the export process of the fusion partner and retain their function upon fusion.

The ease of selection for high-affinity variants using solid-supported antigens led to widespread use of *ex vivo* directed evolution for the engineering of affinity reagents, enabling the development of synthetic antibodies for clinical applications (Huse et al., 1989). Although limited to few activities, enzymes displayed on the surface of phage have been engineered by capturing active variants with high-affinity transition state analogue inhibitors or “suicide substrates” that create an irreversible covalent link to the active site of active variants (Fernandez-Gacio et al., 2003). Peroxide-generating enzymes (lipases and esterases) have been selected by yeast cell surface display by coupling their activity to peroxidase-mediated production of highly-reactive tyramide radicals, which create a covalent link to cell-surface proteins for activity detection (Lipovsek et al., 2007).

1.5.2.3 *In vitro* selection systems

In *in vitro* selection strategies, cells are not used as the genotype-phenotype linkage and, consequently, other mechanisms must be designed to ensure the selected phenotypes can be linked to their encoding genotypes. The methods to establish this linkage can be broadly divided into two groups: mechanisms which physically link genotype and phenotype molecules and compartmentalisation of the population to avoid mixing of variants.

In this first category are methods for engineering functional nucleic acids, employing the same molecule as both genotype and phenotype. These have been used for the generation of affinity reagents known as aptamers (Blind and Blank, 2015) and molecules with enzymatic activity (Robertson and Joyce, 1990). Adaptations of this strategy for the directed evolution of proteins is possible by creating a physical link between between the mRNA genotype and the protein responsible for the phenotype, either via a covalent linkage to a puromycin modification on the mRNA (mRNA display, Liu et al. (2000)) or maintaining the translation complex by stalling the ribosome (ribosome display, Hanes and Pluckthun (1997)). For all of these methods, genotypes can be recovered directly by PCR if information is encoded in

DNA or by reverse transcription if encoded in RNA or synthetic nucleic acids. Since these strategies do not require intermediate steps in whole cells, faster turnaround times and higher throughputs can be achieved than with *in vivo* and *ex vivo* strategies — libraries containing up to 10^{14} variants are routinely selected for the isolation of aptamers (Blind and Blank, 2015), compared to limits of approximately 10^{8} CFU for high-efficiency *E. coli* transformation processes (You and Percival Zhang, 2012). However, RNA-based display strategies can be hindered by ubiquitous contaminations with RNA-degrading enzymes, a problem that can be mitigated by the use of nuclease-free translation extracts (Barendt et al., 2013) or conversion of mRNA to cDNA prior to selection (Yamaguchi et al., 2009).

The second category of *in vitro* selection and screening frameworks relies on individual compartmentalisation of each variant in a library, to prevent variants from coming into contact with each other due to the lack of a physical link between genotype and phenotype. Two distinct strategies are used to achieve this compartmentalisation: screening assays in individual wells of microplates or encapsulation of genotype and phenotype molecules inside a cell-like water-in-oil emulsion compartment.

In microplate-based screening strategies, the link between genotype — usually a microbial strain expressing a single library variant — and phenotype — proteins secreted into culture medium or lysates of cultures — is the ability to recover the library variant responsible for the phenotype observed in an assay well. Throughput of these systems is limited by the resources — either from human experimenters or automated liquid handling platforms — available for the project, due to the space and experimental effort requirements required for screening large numbers of microplate assays. Ambitious directed evolution experiments, such as metabolic engineering of microbial strains for biofuel production using multi-step pathways require screening of vast libraries containing over 10^{5} combinations of elements: relative copy number of genes in the pathway, promoter strength, terminators, ribosome binding sites and mRNA stability (Peralta-Yahya et al., 2012). To circumvent extensive (and costly) screens, information-rich library design strategies allow the efficient engineering of novel phenotypes, such as the recent engineering of the first carbon-silicon bond forming enzymes from sequential site-saturation mutagenesis of only three sites (60 possible variants) within a cytochrome c (Kan et al., 2016b).

Emulsion-based selection methods link their genotypes and phenotypes with the same physical separation principle as microplate screening, but using much smaller compartment volumes — with droplet diameters around the single micrometre range — containing a single cell expressing a protein variant from the library. The greatly-reduced volume of these emulsion droplets allows for higher throughputs than is possible with microplates and at a lower cost, enabling libraries of 10^9 to be selected in a total emulsion volume of 1 ml (Pinheiro et al., 2012a). Selection platforms have been developed for engineering of nucleic acid-active enzymes by capture of plasmids encoding active variants (Ghadessy and Holliger, 2007; Pinheiro et al., 2012a), allowing the development of the first synthetic genetic information storage system (Pinheiro et al., 2012a). Precise control over droplet dimensions using microfluidics devices enabled the establishment of a very high throughput screening system using fluorescent enzyme substrates and FACS, capable of screening over 10^7 variants in under 2 hours (Zinchenko et al., 2014). Though this system was only demonstrated on a model enzyme system, it was shown to capable of recovering active variants from a population of inactive variants that was 100,000-fold larger, showing the potential of such systems for enzyme engineering.

1.5.3 Learn: Analysis of directed evolution results

The results obtained from the Test step of the cycle are crucial to inform the decisions that will direct the subsequent steps of a directed evolution project. The first decision is whether to continue carrying out cycles of directed evolution and it depends on how close the phenotypes of the variants obtained in the current cycle are to the aims set for the project. If the goals have not yet been met and the decision to start a new cycle is made, the results obtained in the current cycle can be exploited to inform the subsequent Design phase.

If the project employs direct screening for a single phenotype of interest to partition the population in the Test step, the results of the screen can be used directly to decide whether the desired phenotype is present in any of the variants or whether a library will be designed for a new round of selection (Kan et al., 2016c). However, this is not a frequent scenario for many biological systems, in which multiple phenotypes are evolutionarily correlated due to being encoded by overlapping networks of residues (Pires et al., 2016). In many cases this leads to the observation of trade-offs between a phenotype of interest — such as catalytic activity — and secondary but still essential phenotypes — such as enzyme stability (Dellus-Gur et al., 2013).

Whenever such trade-offs exist, any important phenotypes correlated to the one being targeted must also be measured for the chosen variants, to prevent the isolation of undesired variants such as a highly-active enzyme that is not stable at the conditions in which it will be used (Li et al., 2016). Once a set of "hits" comprised of the most improved variants is found, these can be used to inform the design of any subsequent cycles of directed evolution that are deemed necessary. Commonly, the sequences of these variants are compared by MSAs and clustered, allowing the detection of positions or motifs that contribute to function (Breaker and Joyce, 1994).

For projects that employ selection as a partition strategy, the traditional approach is to isolate a fraction of the selected population, use Sanger sequencing to recover the genotypes and measure the phenotype of interest in these isolated variants. This sampling is needed because the number of variants recovered from selection is frequently greater than throughput of screening assays (See Section 1.5.2.3 for a discussion of screening throughput). The results of these screens can then be used to select the most improved variants for input into the design (See Section 1.5.4) step of the next cycle. This approach is commonly employed for the engineering of nucleic acid aptamers with affinity for a wide range of targets (Stoltenburg et al., 2007).

However, the development and widespread adoption of NGS within the last decade has made it possible to extract information from a much greater proportion of a selected population. If the selection strategy enriches variants proportionally to the strength of the phenotype they encode, read counts can be used as a measure of the phenotype — reducing the need for laborious screening of variants. This approach is being increasingly employed for directed evolution experiments and also as a way to increase the throughput of biochemical characterisation assays, such as binding affinities of nuclear factors for large panels of post-translational modifications of histones (Nguyen et al., 2014).

Due to the importance of phage display as a platform for the generation of antibodies, NGS was adopted to identify enriched variants in earlier rounds of selection, in which diversity is still too large to reliably isolate the best clones for Sanger sequencing. Early efforts only extracted the most frequent variants from the obtained sequences, not making attempts to identify shared motifs that could be responsible for the observed enrichment (Ravn et al., 2010). These were followed by methods that incorporated clustering of sequences for motif detection but only characterising the phenotype of a single enriched motif (Christiansen et al., 2015) or stratifying the list of identified motifs into groups of equal length (Ravn et al., 2013), overlooking any possible motifs that spanned more than a single length. Since variations in sequence length are known to be important for the function of high-affinity antibodies (Krause et al., 2011), this parameter should be incorporated into enrichment analyses to more closely reflect the natural mechanisms involved in natural antibody selection and maturation.

Employment of NGS for directed evolution has not been restricted to antibody engineering, though. Sequencing of phage display-based selections of viral proteases for activity in the presence of inhibitors enabled the detection of enriched point mutations with frequencies varying between 1% and 50% (Dickinson et al., 2014), which would have required extensive and laborious isolation of clones for similar discrimination by Sanger sequencing — and likely at a higher total cost than is possible with multiplexed sequencing of multiple populations using relatively low-throughput NGS systems such as Illumina MiSeq. Workflows for aptamer selection with analysis of enrichment in the populations have also been established, employing clustering algorithms capable of handling datasets of over 20 million sequences (Hoinka et al., 2015b) or identifying the most enriched sequences and calculating their fitness from their relative frequencies (Jimenez et al., 2013). The additional information recovered by these approaches can, then, be employed to inform library design for further rounds of directed evolution, accelerating the progress of engineering compared to traditional workflows based on characterisation of small numbers of variants per cycle.

1.5.4 Design: Integrating multiple sources of information to increase the efficiency of directed evolution rounds

The methods employed to target diversity within a sequence in a directed evolution experiment depend on the information that is available at that stage of the experiment. The possible sources of information to direct diversification are: functional annotation on the phenotype of interest from the literature, bioinformatic predictions (such as the ones described in Section 1.4), enrichment patterns observed in a prior round of directed evolution, or results from initial experiments to map positions which contribute to the phenotype of interest. The latter option can be carried out efficiently using the technique known as deep mutational scanning, in which each site within a protein is mutated to all possible residues and the phenotype of all variants is measured by an assay using NGS as its output, allowing simultaneous quantitation of the effect of all single point mutations for the sequence (Fowler and Fields, 2014).

The results obtained in previous rounds of a directed evolution experiment can be a valuable source of information for design of a diversification strategy for subsequent rounds. In these cases, the simplest option is to apply the same diversification strategy used in the previous round onto the best variant isolated from that round — if multiple positive variants were isolated, all can be diversified and pooled — to become the library for the next round of selection or screening. However, more sophisticated strategies have been developed to more efficiently explore the sequence space around selected variants.

Whenever multiple functional variants with mutations at different sites in the sequence exist, these can be combined to determine whether there is additivity in their phenotypes or any form of epistasis. While it is practical to produce combinations of a small number of mutations by traditional site-directed mutagenesis (SDM) methods, this becomes impractical for large numbers of mutations — there are 1013 possible combinations for ten separate mutations in a sequence. Therefore, techniques such as DNA shuffling were developed to quickly generate such libraries in a process that is analogous to the production of combinatorial diversity in sexual reproduction of organisms (Stemmer, 1994). Additional diversity can be generated by spontaneous mutations caused by the enzymes used in the shuffling method, but these will not be restricted to the sites of the original variants. Deeper exploration of the sequence space at these sites can be achieved by a strategy called iterative saturation mutagenesis, in which selected positions in a variant are mutated to the other nineteen amino acids, followed by selection and further saturation mutagenesis at the same positions (Reetz and Carballeira, 2007).

While the approaches mentioned above are efficient for generating libraries specifically towards engineering a single phenotype in a target system, generalised libraries can be produced as an common starting point for engineering multiple divergent phenotypes in a same system. Such libraries have been produced for antibodies (Prassler et al., 2011) and affibodies (Woldring et al., 2015, 2017) — small proteins used as affinity reagents analogously to antibodies, but based on Protein A as a scaffold — for use in isolation of binders to any target, with a reduced frequency of misfolded variants and residue compositions biased towards sequences more likely to produce more effective binders.

1.6 Thesis aim and overview

The aim of this thesis was to develop methods to enable the engineering of TOMM synthases and their products by directed evolution. Towards this aim, the following specific objectives were pursued:

1) Establishment of a sequence-based functional residue prediction strategy for the proteins of the TOMM synthase complex, including a novel metric for detection of determinant residues for TOMM biosynthesis.

2) Validation of candidate functional residues selected for *E. coli* McbC dehydrogenase, involved in biosynthesis of microcin B17.

3) Establishment of a robust *in vitro* assay system for TOMM synthase complexes, to aid in characterisation of proteins in the complex and form the basis for a directed evolution strategy for their engineering.

4) Development and validation of a framework for directed evolution of proteins that enables efficient exploration of length diversity, including methods for targeted library synthesis and detection of motifs enriched by selection in mixed-length datasets.

Chapter 2

Materials and Methods

2.1 Genetic constructs

Oligonucleotides described in this section are listed in Appendix B.

2.1.1 TOMM Synthase enzyme constructs

Amino acid sequences for BamB (CBJ61639.1), BamC (CBJ61637.1) and BamD (CBJ61638.1) were obtained from NCBI. Coding sequences for *Bacillus amyloliquefaciens* TOMM synthase complex enzymes were codon optimised for expression in *E. coli* and comercially synthesised (GeneWiz, USA).

N-terminal Maltose Binding Protein fusions of each enzyme were made by cloning into the multiple cloning site of pMAL-c2x (New England Biolabs, USA). BamB and BamC were cloned into the 5′-EcoRI-SalI-3′ sites and BamD was cloned into the 5′-EcoRI-HindIII-3′ sites.

Inserts were generated by PCR using Q5 Hot Start DNA Polymerase (New England Biolabs, USA) and primer pairs containing the appropriate restriction sites for each fusion construct (See Section 2.2.1). Primers MBP_BL_B_F and MBP_BL_B_R were used to clone BamB, MBP_BL_C_F and BL_C_R to clone BamC and MBP_BL_D_F and BL_D_R to clone BamD. PCR products and vector were digested with the appropriate restriction enzymes for each construct, followed by ligation and transformation. (See Section 2.2.6).

Expression constructs for the *E. coli* and *Bacillus sp.* Al-Hakam TOMM synthase complexes were kindly provided by Douglas A Mitchell. All constructs consist of the coding sequence of each enzyme with an N-terminal fusion to MBP and a TEV protease clevage site, in pET29b vector (Melby et al., 2012, 2014). All plasmids were transformed into cloning and expression strains of *E. coli* (see Section 2.3.3) and constructs were confirmed by sequencing before expression (see Section 2.2.9). Proteins were expressed and purified as per Sections 2.4.1 and 2.4.5.

2.1.2 Synthetic substrate constructs

Synthetic substrate constructs for the *B. amyloliquefaciens* TOMM Synthase complex were built by a combinatorial strategy, annealing and ligating short single-stranded oligonucleotides with complementary ends to generate peptide substrate variants (see construction strategy in Fig. 2.1), followed by IPCR to generate short deletions/insertions. All constructs were made in vector pTWIN1 (New England Biolabs, USA) and inserted into the 5′-NdeI-SapI-3′ sites, in frame with the C-terminal Mxe intein-Chitin Binding Domain purification tag (intein-CBD).

The peptide elements contained in the final constructs are in Table 2.1. All substrate constructs contained the *B. amyloliquefaciens* leader sequence and a streptavidin-binding tag (Strep-tag) for purification, with the core peptide region being variable. The four core peptide regions used in these constructs are named 0C, 1C, 2C and 3C, according to the number of heterocyclisable cysteine residues contained in each. The 0-cysteine construct is illustrated in Fig. 2.1 as an example.

The XCYS (X being 0 or 3) and BACA_LEADER oligos contained complementary overhangs to allow their ligation into a single double-stranded fragment.

The inserts for 0C and 3C substrate constructs were generated in two separate annealing and ligation steps, followed by PCR amplification of the ligated products. The substrate

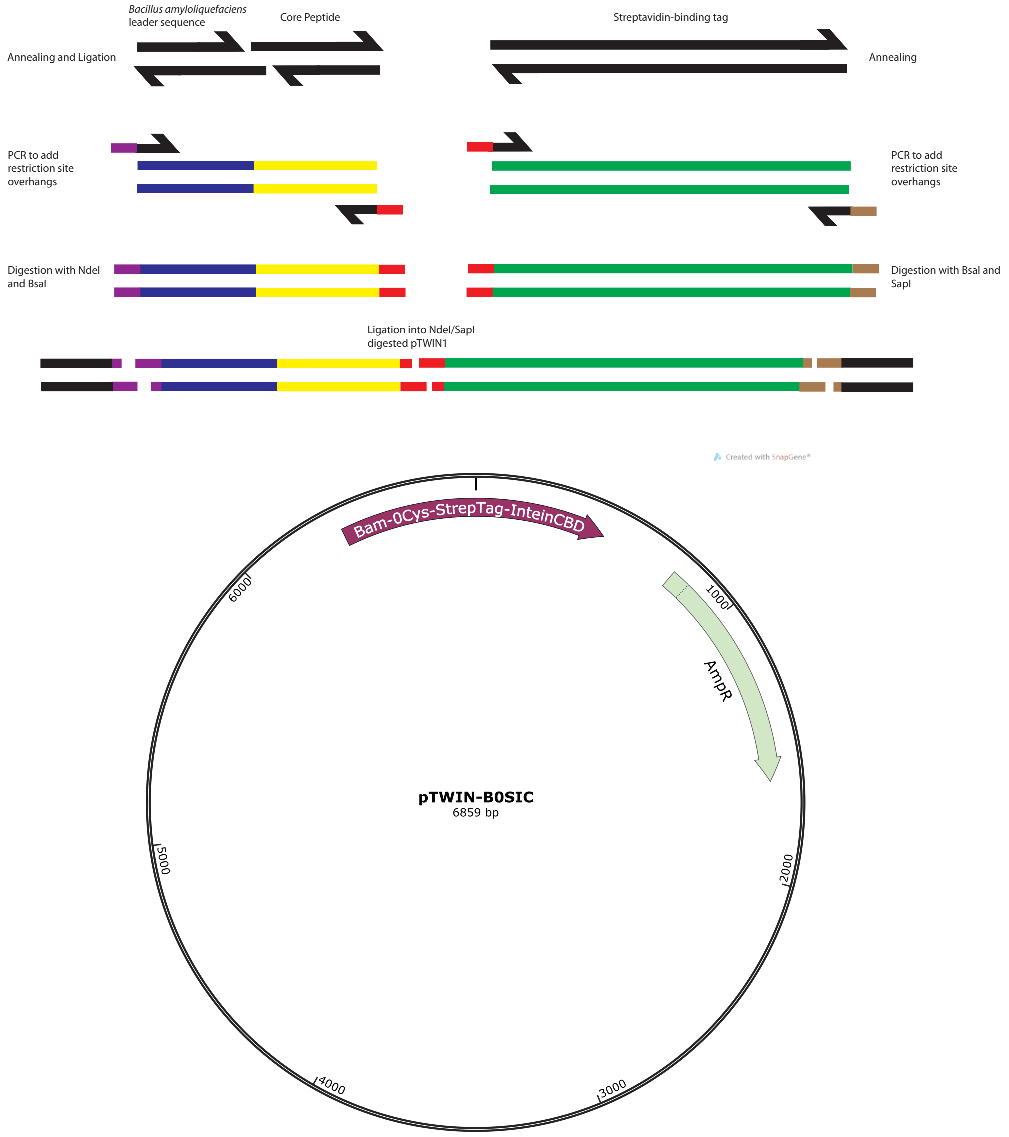

Bacillus amyloliquefaciens
leader sequence
Core Peptide
Streptavidin-binding tag
Annealing and Ligation
Annealing
PCR to add restriction site overhangs
PCR to add restriction site overhangs
Digestion with NdeI and BsaI
Digestion with BsaI and SapI
Ligation into NdeI/SapI digested pTWIN1
Created with SnapGene®
Bam-0Cys-StrepTag-InteinCBD
AmpR
1000
2000
3000
4000
5000
6000
pTWIN-B0SIC
6859 bp

Figure 2.1: Substrate construct design. Top - Strategy used for the assembly of the synthetic substrate constructs. Initially, oligonucleotides coding for the leader and core sequences were annealed and ligated, while the oligonucleotides coding for the streptavidin-binding tag were simply annealed. This step was followed by PCR amplification the two resulting double-stranded fragments with primers containing restriction enzyme sites in their overhangs. The PCR products were then digested with the appropriate restriction enzymes to enable their insertion and ligation into pTWIN1 for cloning. Bottom - map of the B0SIC construct generated by this strategy for the 0-cysteine substrate peptide.

Peptide element	Sequence
Leader	MTQIKVPTALIASVHGEGQHLFEPMAA
New-Leader	MEEVTIMTQIKVPTALIASVHGEGQHLFEPMAA
0C	YDGGK
1C	CDGGK
2C	CCDGGK
3C	CCCDGGK
PznA	CTCTTIISSSSTF
Strep-tag	MDEKTTGWRGGHVVEGLAGELEQLRARLEHHPQGQREP MMSGGCKLGS
Strep-tag-C	MDEKTTGWRGGHVVEGLAGELEQLRARLEHHPQGQREP

Table 2.1: Amino acid sequences of peptide elements contained in substrate constructs. Initially, all constructs contained Leader and Strep-tag regions, along with 0C, 1C, 2C or 3C. Deletion to generate Strep-tag-C produced BXSIC-C constructs. Leader and Strep-tag were present in all of the first round of constructs and each construct contained 0C, 1C, 2C or 3C.

single-stranded oligos XCYS_F and XCYS_R (X being 0 or 3) were annealed along with the leader oligos BACA_LEADER_F and BACA_LEADER_R (see Section 2.2.5) and the Strep-tag oligos (SB19C4_F and SB19C4_R) were annealed in a separate reaction. The annealed products containing the XCYS and BACA_LEADER oligos were ligated according to Section 2.2.6. Since the SB19C4 fragment was only composed of two complementary oligos, no ligation was necessary.

The double-stranded products of the annealing and ligation steps were then amplified by PCR with primers containing restriction enzyme recognition sites as overhangs, to enable insertion of the fragments into pTWIN1. Primers BacL-F-NdeI and XCys-R-BsaI (X being 0 or 3) were used to amplify the leader-core fragment with 5′-NdeI-BsaI-3′ restriction sites and primers Strep-F-BsaI-Cterm and Strep-R-SapI-Cterm were used to amplify the strep-tag fragment with 5′-BsaI-SapI-3′ restriction sites, according to the PCR protocols in section 2.2.1. Inserts were digested with the appropriate restriction enzymes for the respective overhangs and pTWIN1 vector was digested with NdeI and SapI, followed by simultaneous ligation of all three products and transformation according to the protocols in section 2.2.6.

Once the 0- and 3-Cysteine constructs clones were confirmed by sequencing (see Section 2.2.9), the 1- and 2-Cysteine variants were generated by PCR deletions of, respectively, 2 and 1 cysteine codons using the 3-Cysteine construct as a template. Primers XCys-IPCR-F (X being 1 or 2) and Cys-IPCR-R were used to amplify the 3-Cysteine construct with the appropriate deletions (see Section 2.2.1), followed by blunt-ended cloning as described in Section 2.2.7. These clones were also confirmed by sequencing (Section 2.2.9)

The cysteine residue present in the Strep-tag sequence was removed from all substrate constructs by an additional round of PCR deletions (see Section 2.2.1), removing the MMSGGCKLGS sequence that is not essential for streptavidin-binding function and producing Strep-tag-C (see Table 2.1). All four substrate constructs were used as templates for these PCR deletions followed by cloning (see Section 2.2.7), generating the BXSIC-C (X being 0, 1, 2 or 3) constructs which were sequenced (Section 2.2.9) and later expressed and purified for use in assay development (see Sections 2.4.3 and 2.4.6).

Expression constructs for the *E. coli* and *Bacillus sp.* Al-Hakam TOMM substrates were kindly provided by Douglas A Mitchell. All constructs consist of the coding sequence for each substrate, consisting of leader and core regions, fused at its N-terminus to MBP and a TEV protease cleavage site, in pET29b vector (Melby et al., 2012, 2014). All plasmids were transformed into cloning and expression strains of *E. coli* (see Section 2.3.3) and constructs were confirmed by sequencing before expression (see Section 2.2.9). Proteins were expressed and purified as per Sections 2.4.1 and 2.4.5.

In addition to the wild-type BalhA1 substrate from *Bacillus sp.* Al-Hakam containing 10 ciclysation sites (WT), there is a negative control construct where all the cyclisable residues were converted into alanines (NC for non-cyclisable) and also single cyclisation variants of NC where one of the alanines is converted to a cyclisable residue (NC-A40C and NC-A40T).

2.1.3 Substrate construct variants

To produce synthetic substrate constructs with the New-Leader element and the native Plantazolicin A core peptide (see Table 2.1), PCR was carried out (See Section 2.2.1) with primers BamLeadFix-0C-F and BamLeadFix-R or BamLeadFix-PznA-F and BamLeadFix-R, with B0SIC-C as template. These primers contain SapI restriction enzyme sites, to create compatible overhangs for the insertion of the fragments generated by the annealing and digestion of the oligos B1-WT-T and B1-WT-B or BP-WT-T and B1-WT-B. (See Sections 2.2.5 and 2.2.6). These inserts produced substrate peptides containing the New-Leader region and, respectively, the 1C and PznA core regions (See Table 2.1). The constructs were named B1SIC-WT and BPSIC-WT.

Once a correct clone was obtained for B1SIC-WT, this construct was used as template for PCR with primers B1-WT-CmutX-F (X being 0, 2 or 3) and B1-WT-Cmut-R. The products of these amplifications were blunt-end cloned (See Section 2.2.7) to produce the B0SIC-WT, B2SIC-WT and B3SIC-WT constructs. These substrate constructs containing the altered leader peptide were expressed and purified by the same methods used for the earlier constructs (See Sections 2.4.3 and 2.4.6).

Modified McbA substrate peptides were constructed following the SDM procedure (Section 2.2.8) to insert codons for the QMQ and KKD insertions, using oligo pairs McbA-1-46-QMQ-AarIF/McbA-1-46-AarI-R and McbA-1-46-KKD-AarIF/McbA-1-46-AarI-R. However, these were cloned by digestion with the restriction endonuclease AarI and ligation of the overhangs generated by digestion (Section 2.2.6).

2.1.4 Constructs for validation of McbC dehydrogenase mutants

Mutations were inserted into vector pCID909 containing the entire Mcb17 biosynthetic operon following the SDM procedure described in Section 2.2.8, with oligo pairs McbC-XXXX-F/R-Blunt-, where XXXX is the position of each mutation.

2.1.5 Constructs for TEM-1 directed evolution

The TEM-1 gene was cloned into pBAD30 by amplifying the gene in two parts to remove an internal BsaI site present in its sequence with a silent mutation, using oligo pairs TEM1-Nter-F/R and TEM1-Cter-F/R. The vector was amplified using the oligo pair Vec-TEM1-F/R and gene insertion was carried out through the BsaI sites added to the 5′ of each oligo. The M182T stabilising mutation and the $_{165}$YYG$_{167}$ triple mutation were created in this vector by SDM, using the oligo pairs TEM1-M182T-F/R and TEM1-YYG-F/TEM1-MutLoop-R, respectively.

Temp-Opt oligos were used for optimisation of the InDel assembly process, exploiting the internal fluorescent label for sensitive detection of assembly products. Temp-TEM1-XG oligos were the template oligos used to start assembly of the libraries for both rounds of directed evolution. Similary, Cap-Opt oligos were used as capping oligos during library design and Cap-TEM1 oligos were adopted for the TEM-1 libraries. Finally, AB-X-T/B oligo pairs form the assembly blocks for each codon used in the assembly mixtures for each step of the Indel protocol.

Oligos TEM1-Seq(1-7) were used to fully sequence the vectors of selection parasite variants PTX7S and PTX8S, by overlapping Sanger sequencing reactions.

2.2 Molecular Biology

Unless otherwise mentioned (See Appendix A for list of reagents), all enzymes used in cloning procedures were purchased from New England Biolabs, USA. dNTPs were purchased from Bioline, UK.

2.2.1 PCR

Polymerase Chain Reaction (PCR) was used to amplify DNA fragments with the high fidelity enzyme Q5 Hot Start DNA Polymerase, in Peqstar (Peqlab, Germany) and C1000 Touch (BioRad, USA) thermocyclers. Primers used for each reaction are listed in Appendix B and annealing temperatures were selected according to the Melting Temperature calculator available on the NEB website (`http://tmcalculator.neb.com/`).

Amplifications were tipically carried out in 50 µl reactions containing 1X Q5 Reaction Buffer, 200 µM dNTPs, 0.5 µM each primer, 1 ng template DNA, 0.02 U/µl Q5 Hot Start DNA Polymerase. Reaction conditions were as follows: initial denaturation at 98°C for 30 seconds; 25-30 cycles of 98°C for 10 seconds, 50-72°C for 20 seconds and 72°C for 20 seconds per kilobase of the target DNA; final extension at 72°C for 2 minutes and storage at 4°C once the reaction is completed. For amplification of DNA bound to streptavidin-coated paramagnetic beads, 1 µl resuspended bead slurry replaced the purified templated DNA in the reaction.

MyTaq HS Dna Polymerase was used for reactions where initial template was limited and Q5 polymerase did not yield robust amplification. Reactions for this enzyme contained 1x MyTaq Reaction buffer, 0.1U/µl My Taq HS DNA Polymerase, 0.4 µM each primer, 1 ng template (or less) and a total reaction volume of 50 µl. Reaction conditions were as follows: initial denaturation at 95°C for 1 minute; 25-30 cycles of 95°C for 15 seconds, 50-72°C for 20 seconds and 72°C for 30 seconds per kilobase of the target DNA; final extension at 72°C for 2 minutes and storage at 4°C once the reaction is completed.

For reactions that yielded non-specific products, the CES enhancer solution (Ralser et al., 2006) was made as a 5x stock (2.7 M betaine, 6.7 mM DTT, 6.7% (v/v)DMSO, 55 µg/ml BSA, stored at -20°C) and added to reactions at a 1x final concentration.

2.2.2 Agarose gel electrophoresis

PCR and restriction digest products were analysed in agarose gels using a horizontal electrophoresis system (AlphaLabs, UK). Gels were prepared in 0.5X TBE buffer (45 mM Tris base, 45 mM Boric acid, 0.125 mM EDTA pH 8.3) or 10 mM Lithium Acetate with varying concentrations of agarose (0.8-2% (w/v)), according to the fragment size to be analysed. SYBR Safe stain was added to gels at a final concentration of 0.5X from 10,000X commercial stock. Samples were diluted in 6X DNA loading dye (40% (v/v)Glycerol, 0.025% (w/v)Bromophenol blue) before loading onto gels. Gels were run in 1X TBE or 10 mM lithium acetate DNA bands were visualised by transillumination with blue or UV light.

2.2.3 Measurement of DNA concentration by spectrophotometry

Concentration of all DNA solutions was measured by absorbance at 260 nm and purity was evaluated by measuring ratios of absorbance at 260 nm to 280 nm to detect protein contamination and 260 nm to 230 nm to detect solvent or guanidine contamination. All measurements were conducted using the SpectroStar Nano (BMG Labtech, UK) instrument, with the LVis Plate accessory, which enables the measurement of absorbance in 1.5 µL droplets.

2.2.4 Measurement of DNA concentration by Qubit fluorometry

High-precision quantitation of DNA concentration for NGS library preparation (see Section 2.2.12) was carried out using a Qubit 3.0 fluorometer (Thermo Fisher Scientific) with a dsDNA HS assay kit. The working solution of the dye was prepared by a 200-fold dilution of the dsDNA HS reagent in the dilution buffer provided by the manufacturer. Only 0.5 ml microcentrifuge tubes recommended by the manufacturer were used to ensure accurate measurements. Standards were prepared by adding 190 µl of the working solution to tubes, followed by 10 µl of the provided concentrated standards. Samples were prepared by adding

199 µl of the working solution to 1 µl of each sample. All tubes were vortexed for 2-3 s, incubated at room temperature for 2 min and measured in the Qubit instrument.

2.2.5 Single-stranded oligonucleotide annealing

Double-stranded fragments were generated from complementary single-stranded oligonucleotides by combining the oligos to be ligated in TE (10 mM Tris-Cl pH 8.0, 1 mM EDTA) at a final concentration of 10 µM. The solution was heated to 95°C for 5 minutes, then cooled to 4°C at a rate of 0.1°C per second.

2.2.6 Cloning of fragments with complementary overhangs

Amplified PCR products were purified using the GeneJet PCR Purification Kit (Thermo Scientific) according to manufacturer's instructions. Vector and inserts were digested in separate 20 µL reactions containing 2 µL Cutsmart Buffer, 1 µg DNA and 10 U each of the enzymes need for the specific construct. Digestions of PCR products amplified from plasmid templates also contained 10 U DpnI to remove the methylated plasmid DNA and reduce false positives in the transformation. Digestions were carried out at 37°C for 2 hours and subsequently purified using the GeneJet PCR Purification Kit.

Ligations were carried out in 20 µL volume, containing 1X T4 DNA Ligase buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT) and 40 U T4 DNA Ligase. Vector and insert DNA were added to a total of approximately 100 ng per reaction, in a ratio of 1 molecule of the vector to 3 molecules of the insert. Reactions were incubated for 2 hours at 25°C.

Before transformation, ligation products needed to be purified by phenol/chloroform extraction and ethanol precipitation (see Section 2.2.10) to remove salts and other ions that interfere with transformation by electroporation.

2.2.7 Blunt-ended single fragment cloning

Purified PCR products to be used for blunt-ended cloning of a single fragment by intramolecular ligation were phosphorylated by incubation of 1 µg DNA with 10U NEB T4 Polynucleotide Kinase (PNK) and 10 U DpnI in 1x T4 DNA Ligase buffer for 1 hour at 37 °C. Phosphorylated products were purified using the GeneJet PCR Purification Kit and 20 µL ligation reactions were assembled in 1X T4 DNA Ligase Buffer with 40 U T4 DNA ligase and approximately 100 ng DNA. These reactions were incubated at 16°C for 16 hours before phenol/chloroform extraction and ethanol precipitation (see Section 2.2.10), followed by transformation (See Section 2.3.5).

2.2.8 Site-directed mutagenesis

Site-directed mutagenesis was carried out on plasmid templates containing the target sequence by amplification of the entire vector sequence with pairs of primers annealing immediately adjacent to the site where mutations were introduced. The desired mutations were added as $5'$ non-annealing sequence to only one of the oligos and PCR was done using Q5 DNA polymerase (see Section 2.2.1) for a maximum of 25 cycles to reduce the likelihood of mutations. Purified products were recircularised by blunt-ligation (see Section 2.2.7).

2.2.9 Sanger sequencing of plasmids

Aliquots of each construct to be sequenced were sent to GATC Biotech AG, Germany for Sanger sequencing, according to the service provider's instructions. Sequencing results were aligned by BLAST to the expected sequences of each construct to detect possible mutations from the cloning process. If a clone presented mutations, another colony isolated from the same transformation was submitted for sequencing.

2.2.10 Phenol/Chloroform extraction and ethanol precipitation of nucleic acids

Phenol/Chloroform/Isoamyl Alcohol (25:24:1) was added to the DNA to be purified in a 1:1 ratio, mixed by vortexing and centrifuged at 17,000 g for 5 minutes at room temperature. The aqueous phase was transferred to a new tube and precipitation was carried out with 100 mM ammonium acetate, 70% (v/v) ethanol and 10 µg glycogen azure. Ammonium acetate was used in the precipitation as it sublimates during pellet drying and glycogen azure was used to facilitate visual detection of the nucleic acid pellet by staining it blue. After vortexing, tubes were kept at -20°C for at least 1 hour, followed by centrifugation at 17,000 g for 30 minutes at room temperature. The supernatant was discarded and the pellets were washed by adding 500 µL ice-cold 70% ethanol and repeating the centrifugation step for 10 minutes. The supernatant was discarded again and pellets were air-dried at room temperature, then resuspended in dH_2O.

2.2.11 InDel Assembly

Oligos used in InDel Assembly reactions were commercially synthesised as single-stranded DNA oligos and purified by desalting (Integrated DNA Technologies). One strand of the template oligos contained a 5′ Biotin-TEG modification for capture on streptavidin beads and were purified by HPLC. Double-stranded fragments were obtained by diluting oligos in annealing buffer (10 mM Tris-HCl pH 8.0, 20 mM NaCl, 1 mM $MgCl_2$, 0.01% (v/v)Tween20), followed by freezing at -20°C for 1 hour, thawing at room temperature and heat annealing following the steps in Section 2.2.5.

Prior to annealing, the oligos corresponding to the strands providing 5′ ends for ligations were phosphorylated in 100 µl reactions containing 1x T4 DNA ligase buffer, 10 U T4 PNK and 1 nmol oligo. Reactions were incubated at 37°C for 3 h, followed by inactivation at 80°C for 20 min. Oligos were phenol-chloroform extracted, ethanol precipitated, resuspended in 90 µl annealing buffer and annealed (as above) to 1 nmol of the corresponding complementary block. Building blocks containing codons for each amino acid were mixed after annealing in the selected ratios for each assembly step.

To prepare for oligo capture, 60 µl of paramagnetic Dynabeads MyOne Streptavidin C1 (Invitrogen) beads were added to 1.5 mL microcentrifuge tubes. During all bead handling steps, any solution changes were carried out by capturing beads from slurry on a magnetic rack, removing supernatant with a pipette and resuspending in the desired solution. Fresh beads were washed twice in 500 µl BWBS (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% (v/v)Tween20) and incubated with rotation at room temperature in 500 µl BWBS for 30 min in a rotating incubator. After washing, 10 pmol biotinylated dsDNA template oligos were added to the bead slurry and incubated for 16 h at room temperature in a rotating incubator. Beads were washed once more in 500 µl BWBS, resuspended in 100 µl BWBS and the slurry was transferred to a 0.5 ml microcentrifuge tube for assembly.

All on-bead digestion steps were carried out with SapI (NEB) in 100 µl reactions (10 µl 10X CutSmart buffer, 20 U SapI, 1 µl, 0.01% (v/v)Tween20) for 2 h at 37°C with vortexing every 15-20 min to keep beads in suspension. After digestion, beads were isolated, resuspended in 100 µl BWBS and incubated at 37°C for 30 s before recapturing to start the annealing step. The supernatant containing SapI was retained and stored at 4°C for use in other assembly cycles carried out within the same day.

The mixture of building blocks selected for the assembly step was added to the washed beads, incubated at 37°C for 30s, followed by incubation at 4°C for 30 s. The supernatant containing the unbound building blocks was retained if needed for furtther steps and 100 µl ligation mixtures were added (10µl 10X T4 DNA ligase buffer, 12 µl 1,2-propanediol, 10 µl 30% (v/v)PEG-8000, 400 U T4 DNA ligase, 50 U 5′ deadenylase, 1 µl 1% (v/v)Tween20, 65 µl ddH$_2$O). Ligations were incubated at 25°C for 1 h, with vortexing every 15-20 minutes.

Ligation mixtures were retained and stored at 4°C for use in other assembly cycles carried out within the same day.

After ligation, beads were washed in BWBS with a 30 s incubation at 37°C and were then ready to be used in a new assembly cycle, starting from the digestion step. The final assembly cycle used a modified dsDNA assembly block (called a 3′ cap block) containing a priming site to amplify the assembled library from the beads. After the ligation step of the final assembly cycle with the capping oligo, beads were resuspended in 50 µl BWBS for PCR amplification and long-term storage at 4°C.

2.2.12 Preparation of libraries for Illumina sequencing

Libraries for Next-Generation Sequencing (NGS) on an Illumina MiSeq instrument were prepared by PCR with oligos containing the required 5′ and 3′ adaptor sequences, along with distinct indices to allow all pre-and post-selection libraries to be sequenced in a single lane.

Pre-selection libraries were amplified directly from the assembled bead slurries and post-selection libraries were amplified from plasmid DNA extracted from a liquid cultures into which were pooled all selected colonies of one selection round. Amplification was carried out in 50 µl reactions using Q5 polymerase and a maximum of 20 PCR cycles were used to avoid polymerase-induced mutations and reduce amplification biases. Reactions contained 1 U Q5 polymerase, 0.2 µM each of oligos XXX-MiSeqF (one oligo for each library, with varying index sequences for demultiplexing. Oligo names and sequences are in Apendix B) and TEM1-MiSeq-R, 1 ng plasmid template or 1 µl resuspended bead slurry from the assembled library, 200 µM dNTPs, 1X Q5 reaction buffer and 1X CES enhancer solution. Entire reaction products were loaded onto agarose gels for verification of product size and purity, followed by excision of bands for purification using Monarch Gel Extraction kits (NEB).

Libraries were quantified by Qubit fluorometry (see Section 2.2.4 and pooled in proportion to the desired number of reads for each sample. Sequencing was done by the UCL Genomics Facility, using an Illumina MiSeq instrument with a 150 cycle v3 kit.

2.3 Microbiology

2.3.1 *E. coli* culture

Routine *E. coli* cultures were carried out in LB medium (1%(w/v) tryptone, 0.5% (w/v)NaCl, 0.5% (w/v)yeast extract. For solid medium, 1.5% (w/v)agar was added). Additional media used for protein overexpression experiments include LBD (LB with 5% (w/v)D-glucose), TB (1.2% (w/v)tryptone, 2.4% (w/v)yeast extract, 0.4% (v/v) glycerol), 2xTY (1.6% (w/v)tryptone, 1% (w/v)yeast extract, 0.5% (w/v)NaCl). Antibiotics used as transformation controls include ampicillin (100 mg/ml stock solution in 50% (v/v)ethanol store at, 100 µg/ml working concentration), kanamycin (50 mg/ml stock solution in dH_2O, 50 µg/ml working concentration), and chloramphenicol (34 mg/ml stock solution in 100% ethanol, 34 µg/ml working concentration).

Media and glassware used for cultures were sterilised by autoclaving at 121°C for 20 minutes. All antibiotic free-media were kept at room temperature, antibiotic stocks were kept at -20°C, antibiotic-containing media were kept at 4°C.

2.3.2 Measurement of bacterial growth by spectrophotometry

Cell density of liquid *E. coli* cultures was estimated by measuring light scattering in spectrophotometer cuvettes with 1 cm path length. Absorbance at 600 nm (OD600) of culture aliquots was measured in the SpectroStar Nano instrument (BMG Labtech, Germany), compared to a culture medium blank and only measurements lower than an OD600 of 1 were considered reliable due to flattening of the response at higher values. If a culture had an absorbance value greater than OD600 of 1, both the culture and blank were diluted tenfold in dH_2O before being measured again and the obtained value was multiplied by 10 to obtain an estimate of the original culture density.

2.3.3 *E. coli* strains

E. coli NEB 10-B (New England Biolabs, USA) was used for cloning of all constructs described in this report and also for the expression of MBP-fused BamB and BamC constructs. *E. coli* T7 Express ER 2566 and *E. coli* T7 Express LysY/Iq (New England Biolabs, USA) were used for expression of constructs under control of T7 promoter or for improving protein yield. *E. coli* strain BL21(DE3) (New England Biolabs, USA) was used solely for *in vivo* TOMM biosynthesis assays. Throughout the text, *E. coli* strains will be referred to only by their strain names as shown here and these are listed in the Abbreviations table, before the start of the main text.

2.3.4 Preparation of electrocompetent cells for transformation

An overnight culture of the selected *E. coli* strain was started from a glycerol stock maintained at -80°C in 10 mL LB with no antibiotics and incubated at 37°C with shaking. After saturation, the culture was used to start a 200 mL culture in a 2 L shake flask at an initial OD600 of 0.1, incubated at 30°C with shaking. When the cell density in the culture reached 0.4 units, the culture was transferred to pre-chilled 50mL centrifuge tubes and centrifuged for 10 minutes at 3250 g at 4°C. The supernatant was discarded and the pellets were each resuspended in 50 mL filter-sterilised ice-cold 1 mM HEPES pH 7.0, then pelleted by centrifugation under the same conditions as before. Once again the supernatant was discarded, but the pellets pooled in two 50 mL tubes and each resuspended in 50 mL 1 mM HEPES. Pelleting, pooling and resuspension were repeated, resulting in a single tube with cells in 50 mL 1 mM HEPES, followed by a final pelleting step.

The final resuspension was carried out with 2 mL 1 mM HEPES containing 10% (v/v)filter-sterilised glycerol, followed by dividing the competent cell mixture into 50 µL or 100 µL aliquots and immediately flash-freezing in dry ice. Competent cell aliquots were stored at -80°C.

2.3.5 Transformation of *E. coli* by electroporation

Transformation of *E. coli* was carried out by electroporation in cuvettes with a 0.2 cm gap between electrodes (BioRad, USA). Before transformation, competent cell aliquots were removed from storage at -80°C and thawed on ice, while the cuvettes were also placed on ice. Once thawed, competent cells were added to the cuvettes, along with the DNA to be transformed. If the original DNA solution contained salts, it was first phenol/chloroform extracted and ethanol precipitated (see Section 2.2.10) to prevent arcing.

Electroporation was carried out using a Gene Pulser II electroporator (BioRad, USA) at 2.5 kV, 200 Ω and 25 µF. Immediately after electroporation, 500 µL LB was added to each cuvette with gentle mixing, followed by a 30-minute incubation at 37°C for recovery. After incubation, cells were plated in LB-agar containing the appropriate antibiotic for selection of transformants and incubated at 37°C overnight.

2.3.6 Plasmid purification from *E. coli*

Colonies obtained on transformation plates were isolated and transferred to liquid media for plasmid multiplication. Cultures were carriet out in 10 mL LB-Amp and incubated at 37°C with shaking. Before harvesting the cultures for lysis, 500 µL of each culture was transferred to a a microcentrifuge tube and sterile glycerol was added to a final concentration of 20% (v/v)for storage of clones at -80°C. The remaining culture volume was lysed and plasmids were purified using the GeneJET PCR Purification Kit. To increase final DNA concentration, culture and buffer volumes were doubled in relation to the kit manufacturer's recommendations, while the final elution volume was not altered. Once the plasmids were purified, concentration and purity were checked by spectrophotometry as described in Section 2.2.3.

2.3.7 Large-scale selection for antibiotic resistance

Assembled TEM-1 libraries were transformed by electroporation of freshly-prepared NEB 10-β cells.After recovery in liquid LB medium, cultures were plated onto LB-agar medium in 24.5 cm x 24.5 cm plates supplemented with the selected ceftazidime concentration for selection and incubated at 37°C overnight. Resistant colonies were harvested using a cell scraper, resuspended in 10 ml LB medium containing the same concentration of ceftazidime used in the solid plates, and incubated at 37°C for 2-3h. Once turbity reached 1-2 OD_{600} units, the cultures were split into three aliquots. One received an addition of glycerol to a final 20% (v/v) concentration and was frozen at -80°C for storage. A second aliquot was diluted to plate approximately 10^4 CFU on LB agar plates containing higher ceftazidime concentrations to isolate the most active TEM-1 variants from the pool. The remainder of the culture was used for plasmid extraction.

2.3.8 Broth microdilution assay for antibiotic resistance

The Minimum Inhibitory Concentration (MIC) of NEB 10-βcells carrying selected TEM-1 variants was estimated by the broth microdilution method (Balouiri et al., 2016). Strains were grown in liquid media containing varying concentrations of a small panel of antbiotics, consisting of the penems ampicillin (Amp) and carbenicillin (Cbn), the cephalosporins ceftazidime (Caz) and cefotaxime (Ctx), and the penem imipenem (Imp).

Approximately 100 CFU of a mid-log growth culture were added to 200 µl LB medium with varying concentrations of the selected antibiotics in a flat-bottom 96-well microplate (Greiner), sealed with a gas-permeable plate seal (AeraSeal, Excel Scientific), and incubated at 37°C for 16 h with shaking. Cells deposited at the bottom of wells were resuspended by mixing with a multichannel pipette and growth was estimated from OD_{600} measurements. Each plate contained a full column of no antibiotic control wells to normalise OD_{600} between independent experiments and the wells along the outer edge of the plate were not used to reduce evaporation effects. Each experiment was carried out a minimum of three times and the MIC of an antibiotic for each strain was defined as the lowest concentration of that antibiotic that fully inhibited growth of the strain.

2.3.9 Disc-diffusion assay for antibiotic resistance

Disc-diffusion assays (Balouiri et al., 2016) were carried out by placing filter paper discs (Oxoid) containing fixed amounts of antibiotic onto a lawn of the selected strain containing approximately 10^7 CFU, plated on antibiotic-free LB medium. Ampicillin discs contained 25 µg, Carbenicillin discs contained 100µg, Ceftazidime discs contained 10 µg, Cefotaxime discs contined 30 µg, and Imipenem discs contained 10 µg. Susceptibility of a strain to an antibiotic was measured as the radius of growth inhibition around the antibiotic disc, measured in four directions around the disc to reduce the effect of any assymetry. A minimum of three independent experiments were carried out for each strain and up to three discs were placed on each culture plate.

2.3.10 *In vivo* Microcin B17 synthesis assay

The agar well diffusion method (Balouiri et al., 2016) was used to detect *in vivo* production and export of Microcin B17, as a measure of the activity of McbC dehydrogenase variants. *E. coli* BL21(DE3) cells that did not carry the Microcin B17 resistance genes (sensitive strains) were grown and approximately 10^7 CFU were added to 4 ml 0.7% (w/v)agar and overlaid on antibiotic-free LB agar plates. Plates were left at room temperature for 15 min for agar solidification and four wells were created in each plated using a 9 mm diameter cork borer. Culture supernatants from overnight cultures of BL21(DE3) producing strains — carrying the Microcin B17 biosynthetic pathway in pCID909 — were obtained by centrifugation at 17,000 rcf in 1.5 ml microcentrifuge tubes for 15 min and 175 µl supernatant was added to agar wells.

Plates were incubated for 16h at 37°C with lids facing up to avoid irregularly-shaped growth inhibition zones caused by supernatant spills. One well was loaded with sterile LB medium in each plate as a negative control and to be used as a size standard to compare images taken at different zoom levels. Each plate was imaged after growth and the area of growth inhibition zones was measured using ImageJ software (Schindelin et al., 2012), allowing estimation of an overall radius and calculation of an inhibition radius by subtracting the radius of the blank well in each plate. Dilution curves of the wild-type supernatant were made to estimate a relation between observed inhibition radii and the relative amount of Microcin B17 produced by each strain.

2.4 Macromolecule Expression and Analysis

2.4.1 Induction of *Bacillus amyloliquefaciens* TOMM synthase constructs

The BamB and BamC constructs were solubly expressed using the *E. coli* NEB 10-beta clones obtained from the cloning process and the BamD plasmid was transformed into the T7 Express strain for expression, according to the the electroporation protocol in Section 2.3.5. Each of the clones was grown for 16 h in 10 mL LB-Amp, followed by innoculation at a cell density of OD600 0.1 into 200 mL fresh LB-Amp in flasks incubated at 37°C with shaking. Cell growth was monitored until the density reached OD600 values between 0.4-0.6 and IPTG was added to a final concentration of 1 mM for induction.

BamB and BamD were induced for 5 hours at 37°C and BamC was induced for 16 hours at 16°C, with continuous shaking for all constructs. Induced cells were harvested by centrifugation at 3250 g at 4°C for 10 minutes, washed in 20 mM Tris-Cl pH7.5 with 150 mM NaCl, pelleted again and stored at -20°C until lysis and purification (See Sections 2.4.4 and 2.4.5).

2.4.2 Induction of *Bacillus sp.* Al-Hakam and *E. coli* TOMM synthase constructs

Balh enzymes and substrates were induced with 1 mM IPTG for 3 hours at 37°C. All enzymes were induced in LB medium, with the addition of 50 µM riboflavin for BcerB, 50 µM ZnCl2 for BalhC. BalhA1 substrates were induced in 2xTY medium (1.6% (w/v) tryptone, 1% (w/v) yeast extract, 0.5% (w/v) NaCl).

2.4.3 Induction of pTWIN1 constructs

The pTWIN1 constructs coding for the synthetic substrate constructs were transformed into the T7 Express ER2566 (New England Biolabs, USA) strain for induction of the T7 promoter used by this vector. Each of the clones was grown for 16 h in 10 mL LB-Amp, followed by innoculation at a cell density of OD600 0.1 into 200 mL fresh LB-Amp in flasks incubated at 37°C with shaking. Cell growth was monitored until the density reached OD600 values between 0.4-0.6 and IPTG was added to a final concentration of 1 mM for induction.

All substrate constructs were induced for 5 hours at 37°C with shaking. Induced cells were harvested by centrifugation at 3250 g at 4°C for 10 minutes, washed in 20 mM Tris-Cl pH7.5 with 150 mM NaCl, pelleted again and stored at -20°C until lysis and purification (See Sections 2.4.4 and 2.4.6).

2.4.4 Lysis of induced cell pellets

Induced cell pellets were lysed in the Equilibration buffer for the column used in subsequent purification steps (See Sections 2.4.5 and 2.4.6 for buffer compositions), with the addition of 1 mM PMSF to inhibit expressed protein degradation by cellular proteases. Pellets were resuspended at a cell density equivalent to an OD600 of 60-90 and transferred to a tube appropriate for the cell suspension volume for lysis by sonication. When needed, cOmplete Mini EDTA-free Protease Inhibitor Cocktail tablets were added to lysis buffer to prevent protein degradadation during processing.

Sonication was carried out using a Soniprep 150 (MSE, UK) instrument with a microprobe accessory. The probe was immersed into the cell suspension for 9 cycles consisting of 4 seconds of sonication and 5 seconds with no sonication, at a 20 µm amplitude. All samples were kept on ice throughout sonication to prevent protein denaturation by excessive heating. This process was carried out once for samples in 1.5 mL microcentrifuge tubes, twice for samples in 15 mL tubes and three times for samples in 50 mL tubes.

After sonication, lysed samples were centrifuged at 4°C for 30 minutes at 17,000 g. The supernatant containing the soluble proteins was transferred to a new tube and the insoluble fraction was resuspended with the same volume of lysis buffer as was used for the lysis. Protein solubility was assessed by electrophoresis in polyacrylamide gels under denaturing conditions (see Section 2.4.7). Cleared lysates containing soluble protein were used for purification.

2.4.5 Purification of MBP fusions

MBP-fused enzymes were purified by affinity chromatography on MBPTrap HP (GE Healthcare, Sweden) columns with a 1 mL bed volume, manually operated with syringes and following the manufacturer's recommendations. Sample injection, column washing, and elution phases of the purification process were each collected in separate 15 mL tubes, with no further fractionation.

Before loading the cleared lysates, the column was equilibrated with 10 mL MBP equilibration buffer (50 mM Tris-Cl pH 7.5, 125 mM NaCl, 2.5% (v/v) glycerol, 0.1% (v/v)

Triton X-100, 0.5 mM TCEP). The sample was loaded onto the the column, followed by another a washing step of 20 mL MBP Wash buffer (50 mM Tris-Cl pH 7.5, 125 mM NaCl, 2.5% (v/v) glycerol, 0.5 mM TCEP). Finally, bound proteins were eluted by flushing the column with 10 volumes MBP Elution buffer (50mM Tris-Cl pH 7.5, 150 mM NaCl, 10 mM Maltose, 2.5% (v/v)glycerol, 0.5 mM TCEP).

Collected fractions were stored at -20°C. The column was regenerated with 5 volumes MBP equilibration buffer and stored at 4°C. Purity of the fractions was assessed by denaturing electrophoresis on polyacrylamide gels using the Tris-Tricine buffer system, according to Section 2.4.7

2.4.6 Purification of intein-fused synthetic substrates

Intein-CBD-fused substrates were purified by affinity chromatography on 0.8 mL bed volume Micro Bio-Spin Columns (BioRad, USA) hand-packed with 0.2 mL Chitin Resin (New England Biolabs, USA). Chromatography was conducted by gravity flow, with buffers and samples loaded by carefully pipetting directly over the resin, according to the manufacturer's recommendations. Sample injection, column washing, and elution phases of the purification process were each collected in separate 15 mL tubes, with no further fractionation.

The columns were equilibrated with 3 mL Chitin Column Buffer (Chitin CB, 20 mM Tris-Cl pH 8.5, 500 mM NaCl, 0.1% (v/v) Triton X-100) and 2 mL of each lysate was loaded, collecting the flow-through fraction. The columns were then washed with 4 mL Chitin CB to remove unbound proteins. In-column self-cleavage of the intein-CBD tag was initiated by washing the column with 0.6 mL Chitin CB with 50 mM DTT, followed by capping of the columns and overnight incubation at 4°C for the intein clevage reaction. Free peptides were eluted by loading 1.2 mL Chitin CB. The columns were regenerated by loading 0.6 mL Chitin CB with 0.3 M NaOH and capping the columns to incubate for 30 minutes at room temperature, followed by a wash with 2 mL Chitin CB with 0.3 M NaOH, 5 mL dH_2O and 1 mL Chitin CB.

The eluted peptide fractions were then reloaded onto the equilibrated columns to capture the remaining uncleaved fusion peptides. After the loaded sample was collected as flow-through, another 1 mL Chitin CB was loaded to increase the peptide yield and collected as a pool with the first flow-through fraction.

The peptides were then concentrated and buffer-exchanged by ultracentrifugation using Amicon-30K and Amicon-3K spin filters. The 30 kDa cutoff filter was used to remove any traces of the fused peptides and other proteins and the 3 kDa cutoff filter was used to concentrate the peptide and remove traces of contaminants that could interfere in downstream analyses. Purity of the concentrated material was assessed by denaturing electrophoresis on polyacrylamide gels using the Tris-Tricine buffer system, according to Section 2.4.7

2.4.7 Protein electrophoresis

All denaturing polyacrylamide gel electrophoresis experiments for protein detection (SDS-PAGE) were carried out in vertical systems (Peqlab, Germany) and gels were made with an Acrylamide:Bis-acrylamide 37.5:1 40% (w/v) solution (Alfa Aesar, USA). Two different gel and buffer compositions were used for the experiments in this report, one for proteins $>$ 30 kDa and the second for peptides $<$ 20 kDa.

For larger proteins, a two-layered (stacking and resolving gels) system was used. The stacking gel composition was 5% (w/v) Acrylamide:Bis-acrylamide, 125 mM Tris-Cl pH 6.8, 0.05% (w/v) SDS, 0.1% (w/v) APS, 0.1% (v/v) TEMED. Resolving gel composition was 8% Acrylamide:Bis-acrylamide, 375 mM Tris-Cl pH 8.8, 0.05% (w/v) SDS, 0.1% (w/v) APS, 0.1% (v/v) TEMED.

The resolving gel was poured first, followed by an overlay of 2-butanol to remove air bubbles and ensure a horizontal interface with the stacking gel. Once the first gel was polymerised, 2-butanol was washed away with dH_2O, the stacking gel was poured and the wells were cast in it.

Samples are diluted twofold in 2X SDS-PAGE sample buffer (100 mM Tris-Cl pH 6.8, 4% (w/v) SDS, 0.02% (w/v) Bromophenol Blue, 20% (v/v) glycerol, 20 mM TCEP) and incubated at 95°C for 5 minutes before loading into the wells in the stacking gel. The composition of the SDS-PAGE running buffer is 25 mM Tris, 192 mM glycine, 0.1% (w/v) SDS pH 8.3.

2.4.8 Peptide electrophoresis

For smaller peptides, a three-layered (stacking, spacer and resolving gels) system was used with a Tris-Tricine buffer system (Tricine SDS-PAGE). The stacking gel composition was 4% (w/v) Acrylamide:Bis-acrylamide, 125 mM Tris-Cl pH 6.8, 0.05% (w/v) SDS, 0.1% (w/v) APS, 0.1% (v/v) TEMED. The spacer gel composition was 10% (w/v) Acrylamide:Bis-acrylamide, 375 mM Tris-Cl pH 8.8, 0.05% (w/v) SDS, 0.1% (w/v) APS, 0.1% (v/v) TEMED. The resolving gel composition was 16% (w/v) Acrylamide:Bis-acrylamide, 375 mM Tris-Cl pH 8.8, 0.05% (w/v) SDS, 0.1% (w/v) APS, 0.1% (v/v) TEMED.

The resolving gel was poured first, followed by an overlay of 2-butanol to remove air bubbles and ensure a horizontal interface with the stacking gel. Once the first gel was polymerised, 2-butanol was washed away with dH_2O, the spacer gel was poured to a height of approximately 1 cm and overlaid with 2-butanol until polymerisation. The organic solvent was washed away, the stacking gel was poured and the wells were cast in it.

Samples for peptide gels were diluted twofold in 2X non-reducing SDS-PAGE sample buffer (100 mM Tris-Cl pH 6.8, 4% (w/v) SDS, 0.02% (w/v) Bromophenol Blue, 20% (v/v) glycerol) and incubated at 95°C for 5 minutes before loading into the wells in the stacking gel. Two different running buffers are used in this system, Tricine SDS-PAGE cathode buffer (100 mM Tricine, 100 mM Tris, 0.1% (w/v) SDS pH 8.3) and Tricine SDS-PAGE anode buffer (133 mM Tris-Cl pH 8.8).

2.4.9 Denaturing polyacrylamide gel electrophoresis for nucleic acids

Electrophoresis in denaturing polyacrylamide gels was used for detection of nucleic acid fragments smaller than 100 bp and single base-pair resolution of fragments (Urea-PAGE). Gels contained 15% polyacrylamide (19:1 acrylamide:bis-acrylamide) with 8 M urea in 1X TBE and were cast homogeneously in a single step.

An equal volume of Urea-PAGE loading solution (98% (v/v) formamide, 10 mM EDTA, 0.02% (w/v) Orange G) was added to samples, which were incubated at 95°C for 5 min before loading onto the gel. Runs were carried out at a constant current of 30 mA for 1.5-2 h. FAM-labeled oligos were detected by imaging on a Typhoon FLA 9500 scanner (GE Life Sciences).

2.4.10 Cleavage of MBP fusion tag by proteases

Factor Xa and ProTEV plus were used to remove fused MBP from TOMM synthase complex enzymes and substrates and cleavage efficiency was monitored by acrylamide gel electrophoresis. Factor Xa cleavage reactions contained 50 µg MBP fusion proteins, 20 mM Tris-Cl pH 7.5, 100 mM NaCl, 2 mM $CaCl_2$ and 0.5 µg Factor Xa protease. Reactions were incubated at at 23°C for 16 h. ProTEV plus reactions were carried out simultaneously with TOMM synthase activity reactions (Section 2.5.1).

2.4.11 Precipitation of peptides by Acetone/Trichloroacetic acid

Peptides were precipitated in microcentrifuge tubes in 80% (v/v) acetone, 10% (w/v) trichloroacetic acid, with the addition of 10 µ glycogen azure to facilitate visualization of pellets. Precipitations were incubated at -20°C for at least 16 hours, then centrifuged at 17,000 g for 1 hour. Pellets were washed twice in acetone by discarding the supernatant, adding 500 µL acetone and centrifuging at the same conditions for 15 minutes. After the second wash, the pellets were air-dried and resuspended in 50 mM TEAB pH 8.5, MBP fusion elution buffer or acetonitrile for water-insoluble peptides.

2.4.12 Protein quantitation in solution

Proteins in solution were quantitated using the Pierce 660 nm Protein Assay Reagent (Thermo Scientific, USA), according to the manufacturer's instructions. The protocol was adapted for use in the low-volume LVis accessory for the SpectroStar Nano (BMG Labtech, UK) instrument, by proportionally scaling down the recommended assay volumes to 20 µL.

2.5 TOMM Synthase activity detection

2.5.1 Activity assay conditions

TOMM Synthase activity assays were conducted according to the methods of Dunbar et al. (2012). The reaction mixtures contained 100 mM Tris-Cl pH 7.5, 125 mM NaCl, 2 mM ATP, 20 mM, MgCl2 and 10 mM DTT in a volume of 50 µl. Free Substrate peptides (100 µM) and FactorXa-cleaved synthase complex enzmyes (10 µM each) were added and incubated at 25°C for at least 16 hours. For the enzyme fusions that contain TEV protease cleavage sites, all proteins and peptides were added as MBP fusions and 5U TEV Protease was added to the reaction to release the free enzymes and substrates for reaction.

After incubation, results were analysed by precipitating the reaction products with acetone/TCA (see Section 2.4.11) and detecting migration shifts in Tricine SDS-PAGE (see Section 2.4.7), analysing by MALDI-TOF (see Section2.5.2), or thiol labeling with fluorescent dyes 2.5.3.

2.5.2 Detection of heterocyclisation by Mass Spectrometry

MALDI-TOF Mass Spectrometry was carried out in a Waters MALDI Micro MX (Waters Corporation, USA) instrument in linear negative mode for peptide detection, scanning in a m/z range of 500-10,000. Peptides in 50 mM TEAB were mixed with CHCA matrix

(See Appendix A) in equal volumes and spotted onto MALDI plates. Spots were air-dried at room temperature before loading into the instrument. Masses were normalised to Calibration Mixture 2 of the Sequazyme Peptide Mass Standards Kit (Life Technologies, USA).

LC-MS was carried out in a Waters Acquity UPLC - Single quadrupole system, using an Acquity BEH C18 column (Waters) in a gradient from 5% to 95% (v/v) acetonitrile in 5 minutes, containing 0.1% (w/v) formic acid. Samples were dissolved in acetonitrile and 20 µl were injected for each sample. Electrospray was used for ionisation and detection in the quadrupole was done in positive mode, between 150-2000 m/z and McbA peptides were detected approximately 1.5 minutes after injection. Spectra were deconvoluted using Water MassLynx software.

2.5.3 Chemical labeling of cysteine residues in peptides

Prior to chemical labeling of cysteine residues, peptides were buffer exchanged into Labeling Buffer (40 mM Tris-Cl pH 7.5, 300 mM KCl, 2 mM EDTA, 0.1% (w/v) Triton X-100). Iodoacetamide-Fluorescein (IAA-FITC) or Fluorescein-5-Maleimide (FITC-Mal) were added to the peptides in varying molar ratios and incubated for 2 hours at room temperature, while protected from light. The reaction was quenched by the addition of a 2-fold excess of L-cysteine to dye and the reaction products were loaded on Tricine SDS-PAGE gels for analysis. Fluorescent signal from the dye was detected prior to staining for proteins, using a Typhoon FLA9500 (GE Healthcare, Sweden) scanner. After recording of the fluorescent signal from the cysteine labeling, the gel was stained with InstantBlue dye and scanned a second time with the appropriate filters for protein detection.

2.6 Bioinformatics

2.6.1 Dataset construction for functional predictions

Sequence based methods were used to construct large and diverse Multiple Sequence Alignments for functional residue prediction in TOMM Synthase proteins. The methods used in this process are summarised in Fig. 2.2.

A separate functional residue prediction dataset was created for each of the three enzymes in the *E. coli* TOMM synthase complex, using publicly available sequence databases. The initial set (referred to as Validated MSA in this thesis) was constructed by performing blastp searches with the *B. amyloliquefaciens* protein sequences as queries against the NCBI nr database with a 10^{-6} e-value cutoff. This stringent cutoff value was used in the initial search to recover a number of homologues that was not too large for manual analysis in the construction of the Validated MSA. Only hits that contained genes for the remaining two genes in the TOMM synthase complex within a window of no more than 20kb were included in the initial Curated dataset. Hits retrieved by this method were aligned using the T-Coffee package (Notredame et al., 2000).

The sequences in the Curated dataset were used as queries for a new round of blastp searches with the same parameters as above (10^{-6} e-value cutoff). A computational pipeline was implemented to produce the input MSAs for functional prediction using the local pairwise alignments from the blastp output. The tools used were freely available sequence analysis software (Uclust, exonerate) and a set of perl and MATLAB scripts for conversion of file formats and construction of the final alignments (see Fig. 2.2 for an overview of the workflow and Appendix C for the scripts used in this section).

First, the script PWA-reformat.pl (See Appendix C) was used to extract the sequence IDs of each of the hits in the search, as well as all the pairwise alignments generated by the query. The sequence IDs were used to recover all the individual target sequences from the

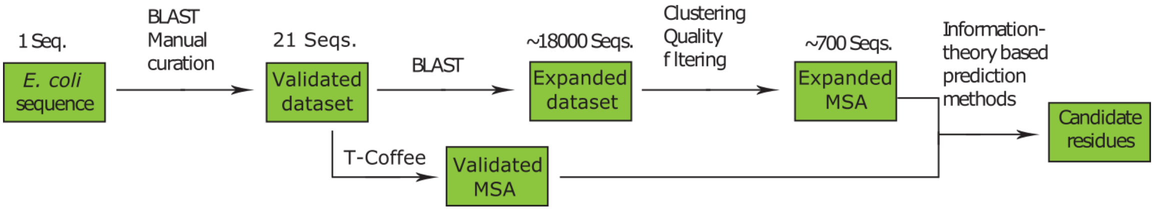


FIGURE 2.2: Analysis pipeline implemented for the *E. coli* TOMM synthase complex enzymes. The validated dataset was created by manual curation of BLAST hits from the *E. coli* query sequences, to include only sequences found within a genomic cluster that contained homologues for all three enzymes of the TOMM synthase complex. These sequences were aligned by T-Coffee to produce the Validated MSA and also used as queries for a larger BLAST search. The sequences obtained in the second BLAST search were clustered to remove duplicate sequences and very close homologues, then an Expanded MSA was created by realignment of the regions recovered from the BLAST searches and removal of columns with low alignment quality. Both MSAs were then used as input into prediction strategies.

local copy of the nr database using the fastafetch script from Exonerate (`https://www.ebi.ac.uk/~guy/exonerate/`). The pairwise alignments were used in the construction of the final MSA, after the filtering steps.

These were clustered (Edgar, 2010) in groups of >95% identity and all but one of the sequences in each of these groups was removed from the dataset, to reduce any biases due to overrepresentation of certain taxonomic groups in the sequences deposited in nr. The clustering package output was then parsed with the clusterParse.pl script (See Appendix C), extracting a representative sequence for each cluster, to be part of the final MSA. The sequence ID of each representative cluster was written to another file to be used for selection of the pairwise alignments added to the final MSA.

Each column in the Validated MSAs was also scored by the Quality values calculated by the Jalview alignment software (Waterhouse et al., 2009) and any columns with Quality values below 50 were deleted from the MSAs to focus the analysis on regions that were more highly conserved in the Curated MSAs. This score is calculated by summing the BLOSUM62 scores for each pair of mutations in an alignment column and, since substitutions between similar residues in BLOSUM62 have higher scores, columns with high Quality scores will have high proportions of similar residues. This filter by the Jalview

Quality value was used to remove columns containing large proportions of gaps, that could bias the results produced by prediction strategies that frequently ignore gaps.

The per-column quality values extracted from Jalview and BLAST reports are used as input for the script hash.pl (See Appendix C). This script detects which of the queries in the Validated dataset had the best pairwise alignment to each of the target sequences in the BLAST results and also matches the high-quality positions in the Validated MSA to their equivalent positions in the selected pairwise alignment. Each of the highest-scoring pairwise alignments is then printed into a separate FASTA file by the script pairwiseSplitFilter.pl (See Appendix C).

The MATLAB script MSAMapper.m (See Appendix C) takes as input the selected high-quality sites in each of the pairwise alignments and builds a single MSA, containing the residue found at each of these positions for all the pairwise alignments that were selected. This is done by mapping each column in the pairwise alignments back to the original Validated MSA, using the pairwise alignment start position and the residue numbering of the query sequence in each alignment. This final dataset was then exported into a FASTA format that could be submitted to functional prediction web servers and custom scripts.

2.6.2 Functional residue prediction

The Expanded MSA was submitted to functional residue prediction servers to be analysed by the Jensen-Shannon Divergence (JSD (Capra and Singh, 2007)) and Cumulative Mutual Information (CMI (Simonetti et al., 2013)) methods using the default settings. Additionally, both data sets were scored by the proposed Normalised Shannon Entropy (NoSE) method, based on the following equation (described in Chapter 3):

$$NoSE_i = P_{imaxC}(-\sum_{j=1}^{20} P_{ijE} * log_2 P_{ijE})$$

where P_{imaxC} is the frequency of the most common residue in the ith column of the Curated MSA and P_{ijE} is the frequency of residue j in the ith column of the Expanded MSA. NoSE was calculated by running the script freqscores2MSAs.m (See Appendix C) with the Validated and Expanded MSAs loaded in MATLAB.

2.6.3 Analysis of functional prediction results

The top 5% of scores for all three methods were selected for further analysis. Homology models of each enzyme in the *B. amyloliquefaciens* TOMM synthase complexes were constructed by submitting their amino acid sequences to the PHYRE2 homology modelling server, using the intensive modelling mode (Kelley and Sternberg, 2009). These models were then structurally aligned with their homologous template structures and candidate functional residues were plotted onto the aligned structures using PyMOL (The PyMOL Molecular Graphics System, Version 1.6.0 Schrödinger, LLC.)

2.6.4 Simulation of libraries produced by InDel Assembly

In silico library assemblies and selections were carried out using a set of MATLAB scripts written by Dr. Vitor Pinheiro. Script InDEL_assembly3b_CsvRead.m simulated library assembly using assembly efficiency and number of cycles as input values that can be altered directly in the script and relative frequencies of each codon for each assembly step from the fred_input.csv file. Simulated enrichment and motif detection were then carried out using scripts InDEL_selection3.m and InDel_analysis3.m

2.6.5 Treament of MiSeq data for analysis

Reads in fastq format were trimmed to remove the NNN sequence at the 5′ and all sequences past the 100th nucleotide at the 3′ end using the script fastx_trimmer from the FASTX-Toolkit (Available at: http://hannonlab.cshl.edu/fastx_toolkit/). This step was followed by a quality control step, keeping only reads with > 30 quality value over > 90% of their sequence, using the script fastq_quality_filter (FASTX-Toolkit). Reads were then demultiplexed according to the indices used for each library with the script fastx_barcode_splitter (FASTX-Toolkit) and trimmed again to remove the indices and ensure that the first nucleotide of each read was the beginning of the variable region of each library and in the +1 frame for translation. Text file NGS-analysis-Workflow-PT.rtf contains sample command lines to carry out NGS data treatment steps.

Reads were converted to FASTA format using script fastq_to_fasta (FASTX-Toolkit) to produce the input file for transeq (EMBOSS v.6.6.0, (Rice et al., 2000)). C-terminal fixed regions in loop sequences were removed by matching with the custom perl script Cter-trim.pl, producing protein sequences corresponding only to the variable region of the Ω loop. Reads were then counted using the custom perl script fasta-to-fastaCounts.pl, generating a FASTA file with each unique read as a single sequence containing its number of ocurrences in the dataset as part of the FASTA header. The custom perl script fastaCounts-toMatlab.pl was then used to produce a .csv file as input for K-mer analysis in MATLAB.

2.6.6 Detection of enriched motifs by K-mer-based analysis

NGS-derived enrichment data was analysed with script TEM1_PoissonZscores_csvOutput.m, using .csv files of sequence counts generated for each library pre- and post-selection as input. Hamming distances to selected sequences were calculated with script fastaCounts-HammD.pl. Sequence logos were created with the weblogo tool (Crooks et al., 2004), using scripts csv-to-fasta.pl and MakeLogos.sh to, respectively, reformat .csv sequence frequency files into a simple FASTA file that retained frequency information and to automate generation of logo images.

Chapter 3

Development of a strategy for sequence-based functional residue prediction in TOMM synthase proteins

3.1 Introduction

Protein sequence databases are useful tools for learning about the function of any given protein, by allowing the assignment of putative functions to uncharacterised proteins based on homology to related sequences of known function. For TOMM synthases and their homologues, these resources are unreliable, due their homologues being annotated with multiple distinct functions (see Chapter 4 for the results of TOMM synthase homology searches), such as nitroreductases among the TOMM dehydrogenase homologues and adenylases with strong similarity to TOMM docking proteins. Since many predicted functions in sequence databases are derived solely from homology-based transfer and this strategy is known to be prone to mis-annotations (Friedberg, 2006), these are not always accurate indicators of the true function of a given sequence. The lack of direct biochemical evidence for the mechanisms of TOMM biosynthetic enzymes and the unreliable database annotations make it impossible to create a "pure" dataset for functional prediction strategies, containing only sequences predicted with a high degree of confidence to encode proteins involved in TOMM biosynthesis. As the sequence similarity between homologues and known TOMM synthase proteins increases, it becomes more likely that a homologue will possess a distinct function or even both functions at the same time — though such bifunctional proteins could also have been lost along evolutionary time and the extant proteins all possess a single function.

The inability to generate clearly separated datasets containing homologous TOMM synthase proteins and related proteins with divergent functions prevented the use of tools such as SDPpred (Kalinina et al., 2004) for detecting functionally relevant residues within the groups, since these require two or more clearly delimited groups of sequences as input. Therefore, only datasets probably containing mixed functions could be produced and any attempts at functional prediction with these sequences would have to explore the evolutionary information that can be extracted from a single MSA as input. Within such a dataset, columns with high conservation and coevolution signals would be expected to contribute to all or most of the functions contained in the MSA.

Any site that is only important for the function of a subset of the larger group would be more difficult to detect due to a lack of signal from the other sequences in which that site is not functionally-relevant. To circumvent this issue, a novel metric was developed to compare conserved sites within the few validated TOMM synthase complexes to the wider diversity found in the mixed datasets, with the aim of recovering information on sites which contribute specifically towards functions related to TOMM biosynthesis.

3.1.1 Sequence-based metrics selected for functional residue prediction

Two complementary measures were selected from the literature for prediction of functional residues from these mixed-function datasets: Jensen-Shannon Divergence (JSD, Capra and Singh (2007)) and Cumulative Mutual Information (CMI, Marino Buslje et al. (2010)). JSD was designed to detect conserved columns within alignments, which are likely to be relevant to the function of all proteins in the dataset. On the other hand, CMI detects columns that show strong covariation with other sites, suggesting coevolutionary interactions maintained throughout the phylogenetic diversity contained in the input MSA.

The conservation measure JSD of a column in a large MSA is a comparison of the frequency distribution observed for that column to a background distribution, which is the set of expected frequencies from the BLOSUM62 alignment scoring matrix (Capra and Singh, 2007). Since this background distribution was computed from a large and phylogenetically diverse set of protein sequences, a large deviation from this distribution implies the existence of an evolutionary constraint maintained by purifying selection and, therefore, a possible functional importance. JSD is calculated according to the following equation:

$$JSD_i = \lambda \left(\sum_{a=1}^{20} P_i(a) * log \frac{P_i(a)}{r(a)} \right) + (1-\lambda) \left(\sum_{a=1}^{20} q(a) * log \frac{q(a)}{r(a)} \right) \tag{3.1}$$

where $p_i(a)$ is the is the frequency of amino acid a in column i, λ is a prior weight (the default value of 0.5 was used), $q(a)$ is the frequency of amino acid a in the background distribution (derived from BLOSUM62), and $r(a) = \lambda\ p_i(a) + (1-\lambda)\ q(a)$. This score can vary from 0 to 1, with 0 being the score when the distribution at the MSA column matches the background distribution perfectly and the highest values occurring when the query MSA has high frequencies for residues that are infrequent in the background distribution.

The mutual information between two frequency distributions is a measure of how dependent one distribution is on another and is calculated according to the equation:

$$MI(i,j) = \sum_{a,b=1}^{20} P(a_i, b_j) * log\ \frac{P(a_i, b_j)}{P(a_i) * P(b_j)} \tag{3.2}$$

where $(P(a_i)$ is the frequency of amino acid a occurring at position i, $P(b_j)$ is the frequency of amino acid b occurring at position j and $P(a_i, b_j)$ is the frequency of amino acids a and b occurring at positions i and j, respectively. Therefore, two columns in an MSA with a high mutual information have a high likelihood of covarying. The coevolution measure CMI for a column i within an MSA is calculated by adding all possible pairwise mutual information values between column i and the remaining positions in the MSA. A column i that has a high CMI is simultaneously covarying with several other residues within the sequence, which indicates that any mutation in that column tends be compensated by mutations at other sites to maintain the function of the high CMI residue.

To compensate for the inability to generate validated TOMM synthase complex dataset that is large enough for detection of functional residues, an alternative approach was devised to extract information from the small set of representatives of this family that have been characterised. If MSAs that can be constructed from homology searches of TOMM synthase protein sequences contain divergent functions, any residues that are strongly conserved in the known TOMM synthase proteins but are less evolutionarily constrained in the complete MSAs are expected to be relevant to TOMM synthase-specific functions. Therefore, a metric was developed to identify residues that have evolved following this pattern, as a third strategy to employ the information contained in the diversity of TOMM synthases and their homologues towards a greater understanding of the mechanisms for the production of these peptide-derived natural products. Then, a search was carried out for a suitable protein family to be used as a benchmark for the proposed metric together with the two published strategies (JSD and CMI), followed by the construction of an input dataset and validation of predictions by comparisons to published structural information and mutations with known effects.

3.2 Results and Discussion

3.2.1 Normalised Shannon Entropy

In a theoretical MSA containing only TOMM synthase complex proteins, any residue that is strictly required for function in TOMM biosynthesis would be expected to be in a fully conserved column. If the homologue of this residue is not important for the distinct functions of proteins that diverged from TOMM synthase proteins, it would be expected to contain higher diversity due to relaxation of purifying selection at that position, representing a determinant column as defined in Section 1.4.2.1. Therefore, in an extensive MSA produced by alignment of sequences of mixed function recovered from homology searches (as depicted in Fig. 1.6), this same column would be expected to contain diversity proportional to the relative amount of TOMM synthase proteins — in which the residue is conserved — and homologues with distinct functions — in which the residue is not constrained by purifying selection — in the MSA. The diversity contained within a column can be estimated by a simple metric such as Shannon Entropy(Shannon, 1948):

$$SEi = \left(- \sum_{a=1}^{20} P_i(a) * logP_i(a) \right) \tag{3.3}$$

where $P_i(a)$ is the frequency of residue a at column i of the MSA. However, this metric alone would not be expected to recover the residues that are conserved in the subset of proteins encoding for members of the TOMM synthase complex. The columns with the highest Shannon Entropy values are poorly conserved for the entire alignment, likely to be positions in the MSA which are not under selection pressure for any of the proteins contained in the dataset.

Therefore, a hybrid metric was proposed to detect conserved residues within a smaller MSA composed of validated TOMM synthase proteins — hereafter referred to as Validated MSA — and use that information to modify the SE metric by penalising high diversity columns int the larger MSA — hereafter referred to as Expanded MSA — that are not conserved in the validated set. This metric was named Normalised Shannon Entropy (NoSE) and is calculated according to the following equation:

$$NoSE_i = P_{iC} \left(- \sum_{a=1}^{20} P_{iE}(a) * log_2 P_{iE}(a) \right) \tag{3.4}$$

where P_{iC} is the frequency of the most common residue in the ith column of the Validated MSA and $P_{iE}(a)$ is the frequency of residue a in the ith column of the Expanded MSA. This score has a maximum value of approximately 4.3 ($log_2 20$), when the ith site is invariable in the Validated MSA, the Expanded MSA has all 20 aminoacids in the same frequency (5%) and the relative frequency of TOMM synthases in the larger expanded set is very low. However, this would also imply that any predictions obtained by the other two methods were made on datasets containing mostly non-TOMM sequences, which was not the case in the predictions made here, since no NoSE scores above 4 were observed in the datasets that were explored.

Other example scenarios for the effect varying degrees of conservation in both the Validated and Expanded MSAs can be found in Table 3.1. The NoSE metric penalises all scenarios except for the one where there is a high degree of conservation in the Validated MSA and a high diversity in the Expanded MSA.

P_{iC}	$P\_iE(a)$	NoSE
0.8	0.05^{a}	3.5
0.8	0.2^{b}	1.9
0.8	0.8^{c}	0.6
0.2	0.05^{a}	0.9
0.2	0.2^{b}	0.5
0.2	0.8^{c}	0.1

Table 3.1: Examples of NoSE scores for various conservation scenarios. [a] Equal frequency of 0.05 for all residues in the Expanded MSA, representing no conservation. [b] Equal frequency of 0.2 for five residues in the expanded MSA and 0 for all others, representing an intermediate level of conservation. [c] 0.8 frequency for one residue and 0.2 frequency for another, with 0 frequency for all others, representing a high level of conservation in the Expanded MSA.

However, since the sequences from the Validated MSA are also contained in the Expanded MSA, the relative size of the two alignments can also have a strong influence on the NoSE score. If the Validated MSA is exactly half the size of the Expanded MSA and has one fully conserved column, the maximum possible NoSE score for that column will be approximately 0.32, when the Expanded MSA has the 19 residues not present in the

Validated MSA in equal frequencies in its unique sequences. In this instance, the Shannon Entropy component of the Expanded MSA would only be approximately 0.64, even though the column is completely conserved outside the Validated MSA. Therefore, NoSE scores will tend to reduce as the size of the Validated MSA is increased in proportion to the entire Expanded MSA. In this work, this limitation of the method was not of large influence in the calculated scores, since in all cases the Expanded MSA was at least 20 times larger than the Validated MSA and a fully conserved column in the latter could not greatly affect the final residue frequencies of the former. Simulations with sampling of known datasets could be carried out to more thoroughly explore the effect of variations in relative size of alignments and determine what range of relative alignment sizes can produce informative predictions.

The intention for this metric was to extract additional information from the mixed-function Expanded MSAs, that should be complementary to the predictions produced by JSD and CMI, which respectively detect columns within the Expanded MSA that are conserved or possess signs of coevolution with other columns. By extracting high-diversity columns from the Expanded MSA that are conserved in the validated set, NoSE should detect residues that became fixed or at least strongly conserved during the evolution of TOMM synthases, but did not experience the same selective pressure in the set of homologous proteins that possess distinct functions.

Given the limited characterisation of the TOMM synthase complex proteins, these were not ideal first targets to characterise NoSE. Instead, a functionally-diverse family of proteins with extensive biochemical and structural characterisation was selected to validate the proposed metric in comparison to the selected coevolution- and conservation-based metrics

3.2.2 Construction of a benchmark dataset

The main requirements sought in a benchmark dataset were the existence of published structural and biochemical information on functionally-distinct but homologous protein subfamilies and a low enough level of sequence divergence between the two families to allow the construction of MSAs. These MSAs would be used as input for functional residue prediction using the JSD, CMI, and NoSE scores, to be validated by comparisons to published information on the function of the families.

Initially, the CATH database of protein superfamilies classified on the basis of structural similarity and subdivided into smaller groups likely to share similar functions named Functional Families (FunFams), was used to search for a target family (Dawson et al., 2017), along with associated functional annotation extracted from other resources, such as the Catalytic Site Atlas (CSA, Porter et al. (2004)). A particularly well-annotated superfamily — Aldolase class I (CATH Superfamily 3.20.20.70) — was selected for its abundance of catalytic site information and preconstructed alignments were downloaded for two of its FunFams. Attempts were made to merge the alignments from the two families by purely sequence-based algorithms such as MUSCLE (Edgar, 2004), but this produced alignments containing large tracts of gaps due to the high sequence divergence contained in these families defined by structural similarity. Due to time constraints, no further improvement was pursued for this superfamily and an alternative dataset was sought that could be constructed using only sequence-based alignment strategies.

The periplasmic Solute-Binding Proteins (SBP) of Gram-negative bacteria were selected as another candidate for a benchmarking family for the TOMM synthase functional predictions. These proteins act as highly-specific receptors for substrates of the diverse group of ABC importers and have been intensely studied. Structures are available for many of these proteins — some crystallised in the presence and absence of their substrate (Oh et al., 1994; Sun et al., 1998) — and their wide range of specificities has been exploited for biosensing applications, including artificial variants with altered substrate specificity (de Lorimier et al., 2002; Dattelbaum and Lakowicz, 2001; Gruenwald et al., 2012).

The Glutamine-Binding Protein (GlnBP, UniProt P0AEQ3) from *E. coli* (Sun et al., 1998) and the Lysine-, Arginine-, Ornithine-binding protein (LAO, UniProt P02911) from *Salmonella enterica* (Oh et al., 1994) were selected as the starting points for dataset construction. Crystal structures of both of these proteins are available in both apo and substrate-bound forms, providing information on substrate-binding residues as well as conformational changes that occur upon binding. A separate set of homologous sequences were constructed for each protein and these were combined into an Expanded MSA for calculation of the functional prediction scores, with the separate sets serving as Validated MSAs for calculation of the conservation component of NoSE scores.

The sequences of both proteins were used as queries for BLASTP searches of the NCBI nr database, producing tens of thousands of hits even at very stringent e-value cutoffs (10^{-20}). These numbers are too large for alignment using computationally-intensive and accurate methods such as T-Coffee, so clustering with USEARCH (http://www.drive5.com/usearch/) was used to reduce the size of the sequence sets. In addition to the reduction in computational complexity, clustering removes duplicate sequences from strain genomes and reduces the phylogenetic bias created by uneven distribution of genome sequencing efforts.

While the separate datasets were being constructed, sequences annotated as putative GlnBP were detected in the set constructed with LAO as seed and vice-versa. The low sequence divergence between GlnBP and LAO that allowed them to be easily aligned to each other also meant that homology searching by BLAST could efficiently recover members of both functional groups from either query sequence. Since this would mean that the dataset constructed for each function was "contaminated" by sequences annotated as coding for the other function, the separate datasets could not be used as a Validated MSA for the NoSE prediction which requires an alignment containing only a single function. Therefore, a single Expanded MSA was constructed from the sequences obtained by the LAO BLAST searches and a second, smaller, Validated MSA was constructed by manual curation containing only sequences annotated as GlnBP homologues from the NCBI Protein database and selecting for a wide range of phylogenetic origins. This small set contained twelve sequences, with at least one member of each of the following bacterial phyla: Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria and Firmicutes.

The larger dataset was constructed by taking sequences obtained by BLAST with the LAO as query, with an e-value cutoff of 10^{-25} and a coverage cutoff of 90% to eliminate partial alignments. Sequences with identity above 60% were clustered with USEARCH and represented by only a single centroid sequence in the final alignment, to reduce the computational effort required for alignment, producing a set of 741 sequences. The manually-curated set of twelve GlnBP was concatenated into the larger set and all sequences were aligned with T-Coffee, for a total of 753 sequences. The same Jalview Quality score cutoff of 50 was used initially to eliminate columns containing large proportions of gaps, but this still left relatively large regions (>20 contiguous columns) containing >80% gaps. Therefore, all columns with Jalview quality scores <300 were removed from this alignment to prevent the scores being skewed by the small number of sequences that contained non-gap characters at these sites.

3.2.3 Functional residue predictions in Solute-Binding Proteins

Since GlnBP from *E. coli* has had its structure determined both in the presence and absence of glutamine (PDB IDs 1ggg and 1wdn, respectively), these structures were used here to analyse the possible effects of predicted functional residues (Fig. 3.1). The overall structure of GlnBP and other SBPs is bilobed with a connecting hinge and a cavity formed between the two lobes that forms the substrate-binding site. A large inter-domain movement occurs upon substrate binding, with the ends of the lobes opposite the hinge coming closer upon substrate binding (Fig. Fig. 3.1, bound glutamine in magenta).

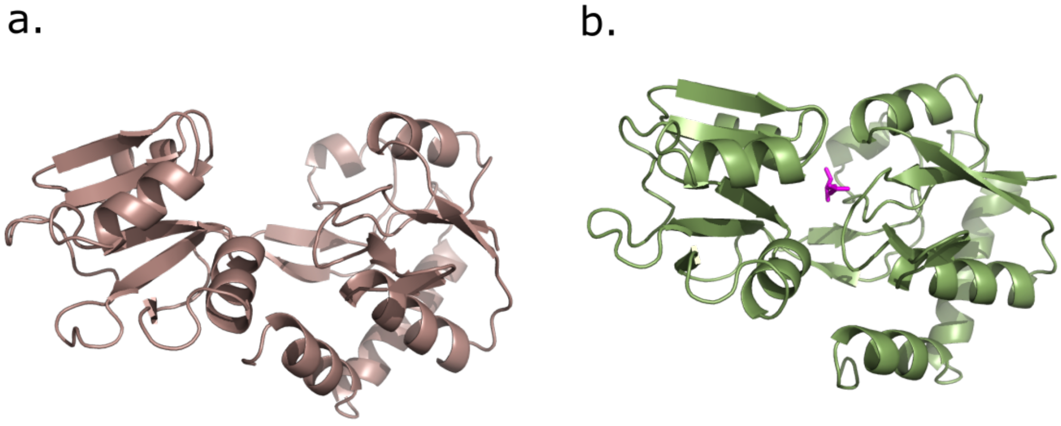


Figure 3.1: Structures of *E. coli* GlnBP. (a) GlnBP crystallised in the absence of substrate (PDB ID 1ggg). (b) GlnBP crystallised in the presence of glutamine (PDB ID 1wdn).

The completed datasets were used as input for the selected functional prediction algorigthms: JSD, CMI, and the NoSE score proposed in this work. JSD and CMI use only the Expanded MSA containing more than one functional group and NoSE uses both the large set and the manually-curated Validated MSA containing only GlnBP homologues. The top 5% of scores for each metric were selected, as an arbitrary cutoff, for further analysis. These residues were plotted onto the apo and substrate-bound forms of the GlnBP structure and compared to previously published studies that characterised mutants or studied the function of this protein by other structural techniques such as NMR or molecular dynamics (MD) simulations. As expected, the metrics selected columns within the Expanded MSA with distinct evolutionary trajectories, which can be evidenced by the lack of any overlap between the residues selected by each metric (Fig. 3.2).

The highest-scoring residues for the JSD metric in GlnBP were distributed throughout the sequence, but all were located around the large domain of the bilobed structure of the SBP proteins (Fig. 3.3). None of the top-scoring residues for JSD form direct contacts with the substrate, but F16 and G26 are part of the hydrophobic core directly underlying substrate-interacting F13 (Pistolesi and Tjandra, 2012) and could have indirect effects on

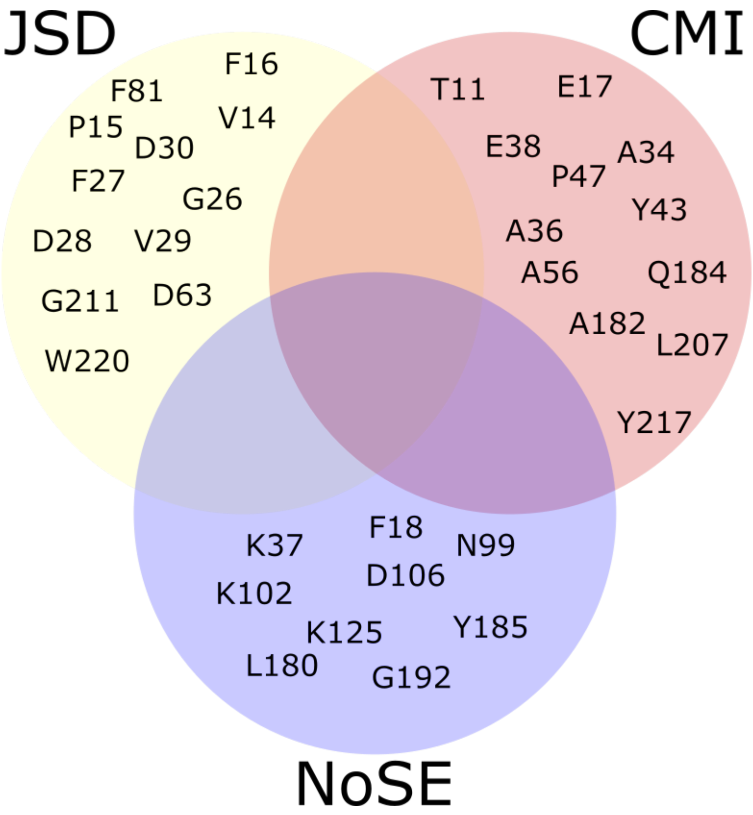


FIGURE 3.2: Venn diagram depicting the lack of overlaps between the top 5%-scoring residues in GlnBP for JSD(yellow), CMI (red), and NoSE(blue)

activity through these contacts. The mobility of V14 has been demonstrated to be significantly constrained upon substrate binding by MD simulations (Lv et al., 2017), which could also suggest an indirect role in GlnBP activity. Residue W220 has previously been assumed to be dispensable for glutamine binding and was mutated to cysteine to convert GlnBP into a biosensor by labeling with a thiol-reactive dye (de Lorimier et al., 2002). However, characterisation of a W220A mutant showed reduced affinity to glutamine, confirming a functional role in determining specificity for this residue (Gruenwald et al., 2012).

As was observed for JSD, the top-scoring residues for CMI (Fig. 3.4) were mostly located in the large domain, but two residues (A182 and Q184) from this prediction were located in the hinge region that connects the two domains. This region regulates the inter-domain movement that occurs upon substrate binding and Q184 has been shown to form part of a network of hydrogen bond contacts within this region that is though to stabilise the apo form of the protein in the transport-incompetent open state (Sun et al., 1998). Residue

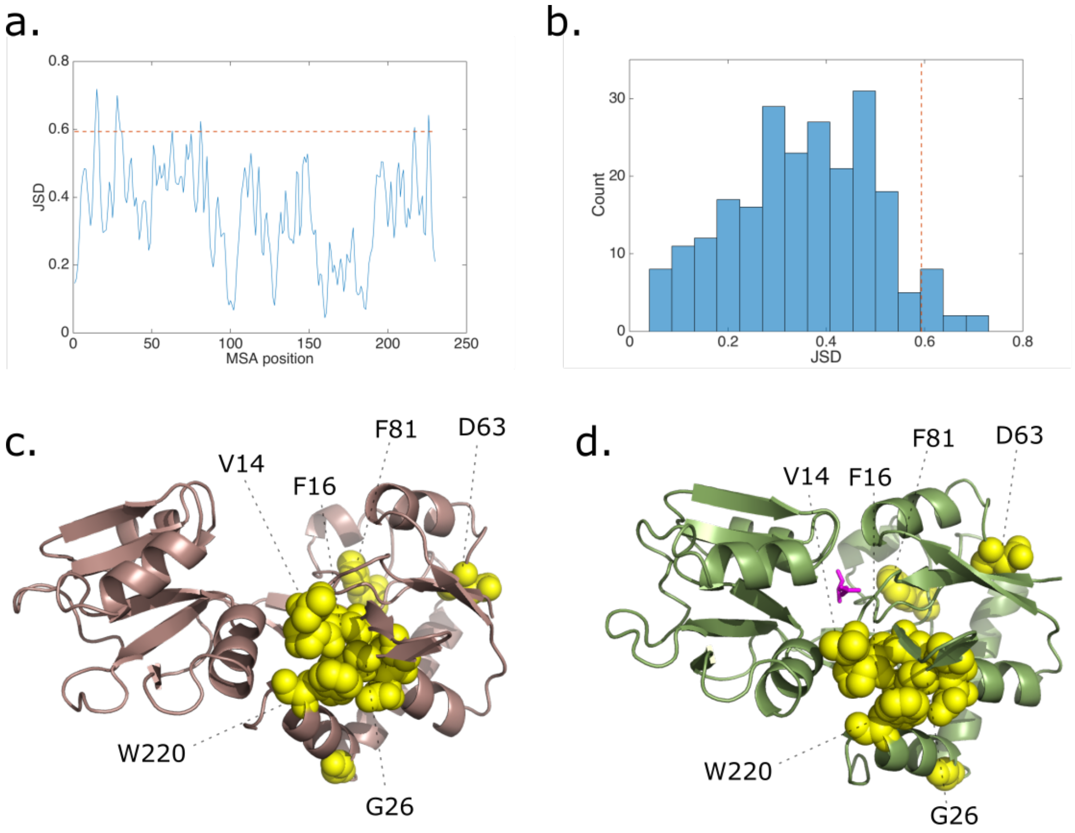


Figure 3.3: Highest scoring residues for JSD in GlnBP. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of JSD scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in yellow on the structure of the apo form of GlnBP (PDB: 1ggg). (d) Top 5% residues highlighted in yellow on the Glutamine-bound structure of GlnBP (PDB: 1wdn). The bound glutamine is highlighted in magenta.

A56 is located along the putative interaction surface with the transmembrane portion of the ABC transporter complex (Fig. 3.4 e), responsible for membrane transport of solutes.

The NoSE metric was calculated using the Validated and Expanded MSAs for GlnBP and Shannon Entropy was also calculated for the Expanded MSA so the difference between the two metrics could be visualised (Fig. 3.5 a-b). Despite the existence of a correlation between the two scores for each colum in the alignment (Fig. 3.5 c), the top 5% of scores for each was distinct, with only two columns from the top 5% of Shannon Entropy values also present in the selected columns for NoSE: Lys102 and a position that had a gap for the GlnBP sequence in the Expanded MSA and was, therefore, not included in further analyses. This comparison shows that the penalty towards non-conserved residues in the Validated MSA allowed the detection of residues with potential function in GlnBP that would not

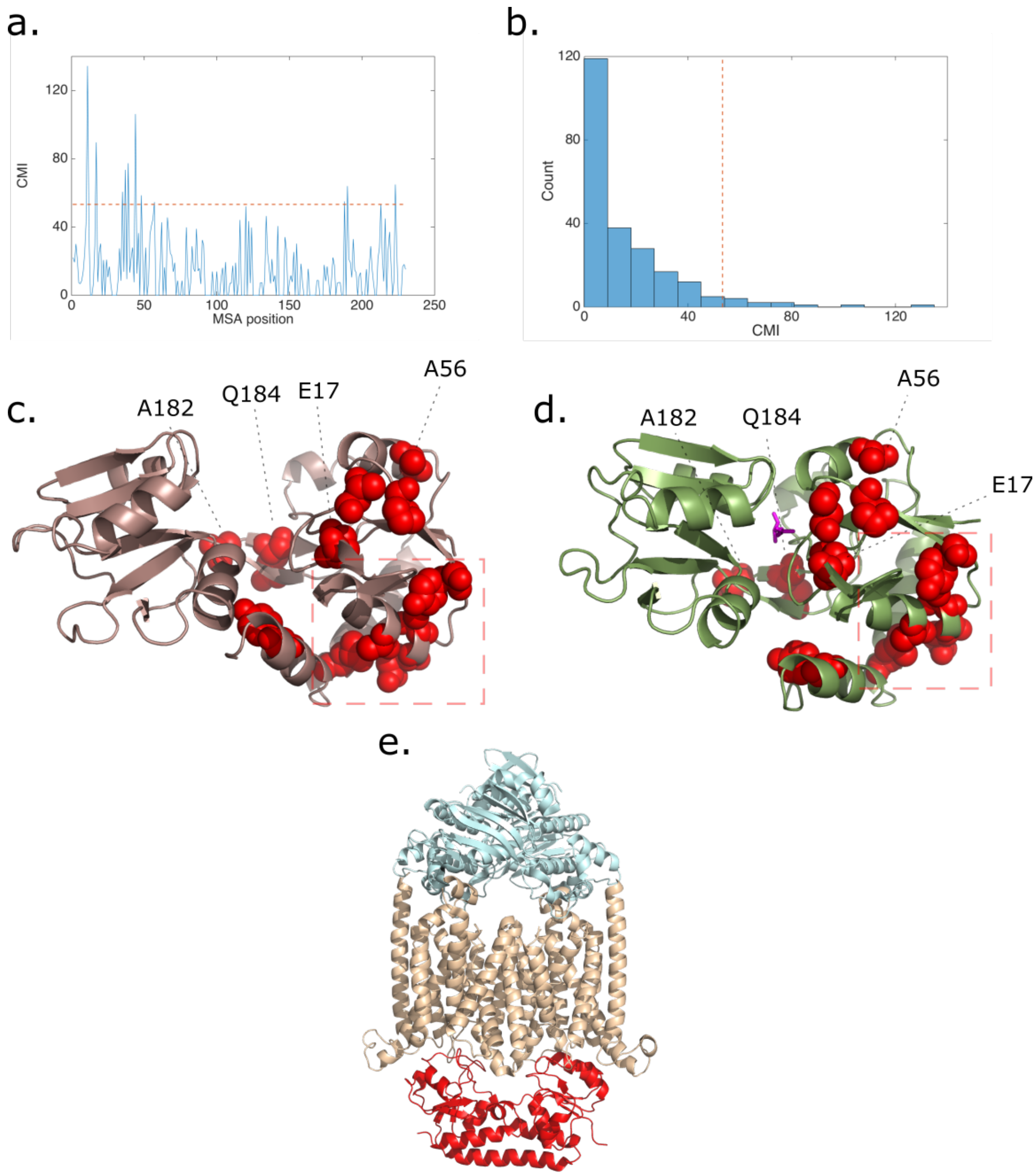


Figure 3.4: Highest scoring residues for CMI in GlnBP. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of CMI scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in red on the structure of the apo form of GlnBP (PDB: 1ggg). (d) Top 5% residues highlighted in red on the Glutamine-bound structure of GlnBP (PDB: 1wdn). (e) Structure of a full ABC transporter complex (Vitamin B12 transporter from *E. coli*, PDB: 2qi9) in the same relative orientation as the models in (c) and (d), with the solute-binding protein highlighted in red and the transmembrane and nucleotide binding subunits in gold and cyan, respectively. The bound glutamine is highlighted in magenta and the patch of residues on the surface mentioned in the text is enclosed in a red dashed box.

be selected by a strict diversity measurement in the Expanded MSA. The comparison was extended by calculating the Shannon Entropy and ScoreCons metrics for the small alignment, also showing no clear overlap between high-scoring residues in NoSE and these other metrics ((Fig. 3.5, d-e)

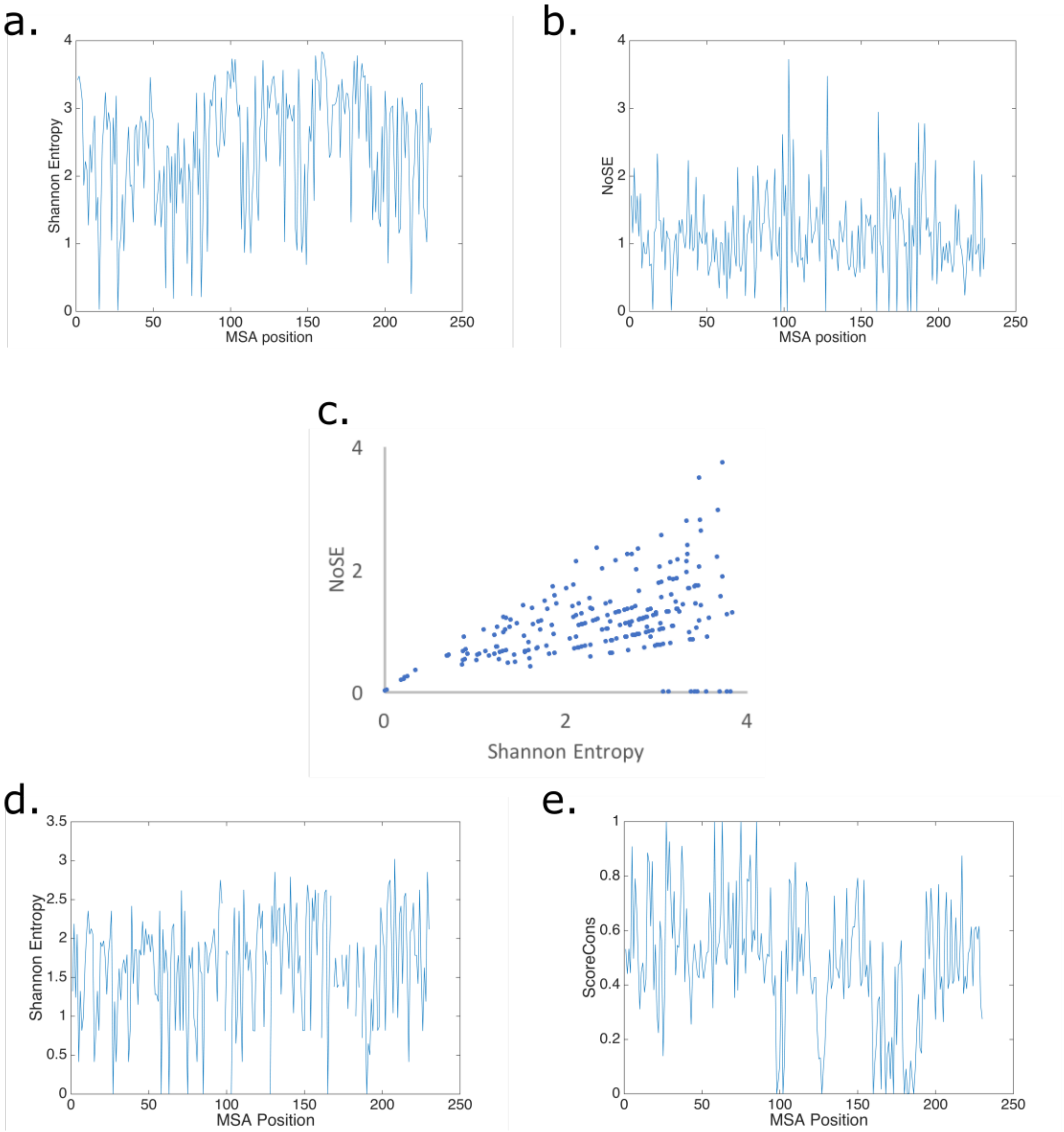


FIGURE 3.5: Comparison between NoSE and other metrics scores. (a) Shannon entropy values for each residue in the Expanded GlnBP dataset. (b) NoSE scores for each residue in the GlnBP dataset. (c) Comparison between NoSE scores and Shannon entropy values for each residue. The Pearson correlation coefficient between the two sets is 0.458. (d) Shannon Entropy values for each residue in the Validated MSA. (e) ScoreCons values for each residue in the Validated MSA.

Unlike in the published scores, among the top 5% of NoSE values for GlnBP were residues in both domains of the protein (Fig. 3.6). One of the putative functional residues

(Y185) was also in the hinge region, making hydrogen bond contacts to substrate-interacting D157, and is a candidate trigger for the ligand-induced conformational change, mediating the protein movements through its hydrogen bond contacts to the remaining hinge residues (Sun et al., 1998). This residue was also one of the sites that produced a variant with affinity towards lactate when mutated in an effort to produce biosensors for non-cognate small molecules using SBPs (Looger et al., 2003).

Similarly to residue F16 detected by JSD, F18 forms part of the hydrophobic core directly interacting with F13 from the ligand site (Pistolesi and Tjandra, 2012), suggesting a role in structuring the ligand-binding site for glutamine recognition. In addition, residues K125 and G192 in the ligand-bound structure are located along the putative interaction interface for the transmembrane subunits of the ABC transporter (Fig. 3.6 d), suggesting a possible role in transporter function.

A summary of the predicted mutations compared to the results from the literature described in the above paragraphs can be found in Table 3.2. Despite the prediction metrics not identifying residues that interact directly with glutamine, at least three residues from each prediction (out of a total of twelve, for a 25% hit rate) were found to have a known contribution in the functional mechanism of GlnBP or were structurally modelled in a position that suggests a role in ABC transporter function.

In addition to the residues discussed above and highlighted in Table 3.2 that have some direct evidence for their possible function, other residues highlighted by the three methods are positioned in the structure in ways that suggest a functional role.

For JSD, residues D63 and F81 (Fig. 3.3) are positioned on the surface of one of the lobes, in a position that could also mediate interactions with the transmembrane transporter subunit of the ABC transporter.

For CMI, residue E17 is a charged residue along the surface of the Gln-binding cavity and can have interactions with the substrate during the binding process that were not

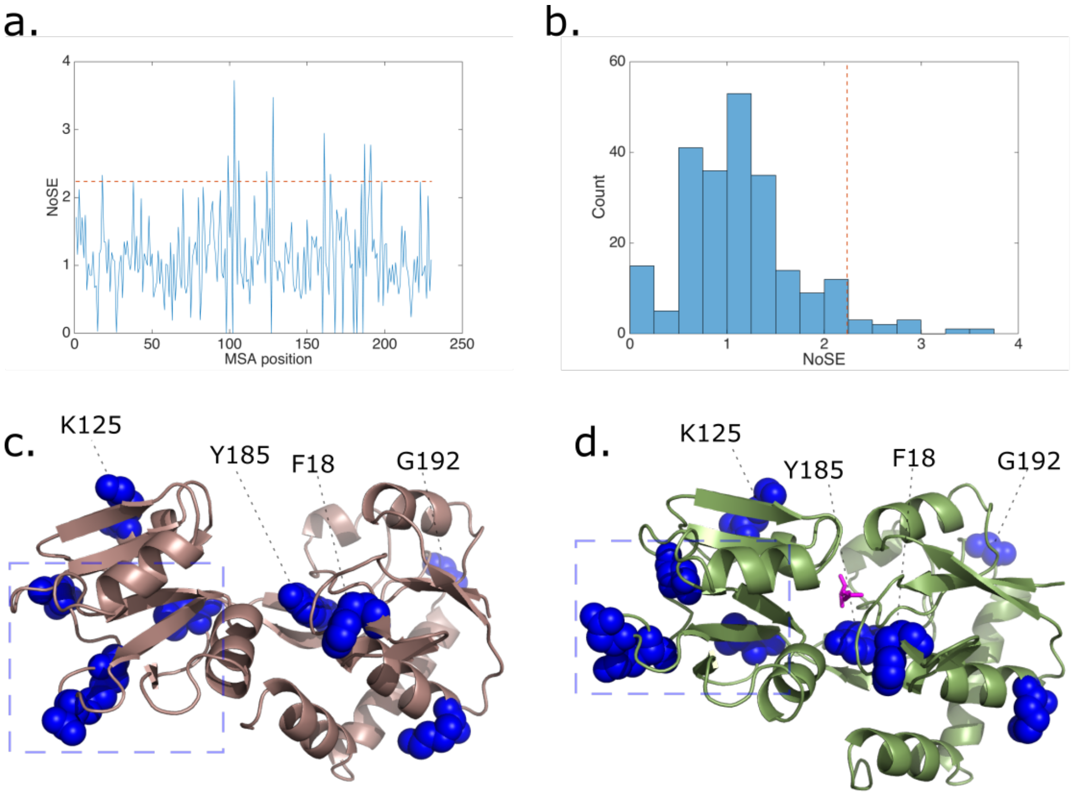


Figure 3.6: Highest scoring residues for NoSE in GlnBP. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of JSD scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in blue on the structure of the apo form of GlnBP (PDB: 1ggg). (d) Top 5% residues highlighted in blue on the Glutamine-bound structure of GlnBP (PDB: 1wdn). The bound glutamine is highlighted in magenta and the patch of residues on the surface mentioned in the text is enclosed in a blue dashed box.

captured by the crystallographic structure (Fig. 3.4). In addition, residues E34, A36, E38, Y43, and L207 were highlighted by CMI and all lay along a large flat surface away from the expected binding interface with the transmembrane subunits (Fig. 3.4, enclosed in a red dashed box). GlnBP has known protein-protein interactions with binding partners outside the glutamine ABC transporter complex, including components of ABC transporter complexes for other amino acids (Szklarczyk et al., 2015), so this patch could represent a region of the protein that coevolved to mediate interactions with diverse binding partners.

Among the residues highlighted by NoSE there was a group of residues similar to the one described above for CMI, but on the other lobe of the structure, composed of residues N99, K102, D106 and L180 (Fig. 3.6, enclosed in a blue dashed box). Since these residues score highly in NoSE, they are well-conserved within the GlnBP proteins and variable in the larger set, which can indicate a role in an interaction that only occurs in the GlnBP proteins.

Residue	**JSD**	**CMI**	**NoSE**	**Affinity**	**Hinge**	**Core**	**Dynamics**	**Interface**
V14	✓						✓	
F16	✓					✓		
F18			✓			✓		
G26	✓					✓		
A56		✓						✓
K125			✓					✓
A182		✓			✓			
Q184		✓			✓			
Y185			✓	✓	✓			
G192			✓					✓
W220			✓	✓				

Table 3.2: Summary of known function-affecting mutations and structural information for the sites selected by the three prediction metrics.

3.3 Conclusions

The results obtained from the benchmark functional predictions on GlnBP corroborate the potential in employing the NoSE metric proposed here for sequence-based detection of functional residues in diverse protein families. Using a dataset created by a simple analysis pipeline, two residues known to contribute to the specificity of GlnBP towards glutamine were detected, as well as two residues possibly involved in protein-protein interactions essential for ABC transporter function. As expected from the distinct evolutionary signals each metric was designed to detect, the residues selected by NoSE had no overlap with the highest-scoring residues for the other two metrics used here, demonstrating their ability to detect MSA columns with distinct evolutionary histories.

Among the residues detected by CMI, two distinct sets highlight the importance of studying coevolving residues to detect functional constraints in protein evolution. The A182-Q184 pair from the hinge region of GlnBP can represent a functional unit that is maintained by purifying selection, eliminating variants at one site that are not compensated by a mutation at the other site. Similarly, the patch of residues highlighted at one end of the protein not expected to participate in subunit interactions of the glutamine ABC transporter complex (Fig. 3.4 c and d, enclosed in a red dashed box) could represent a protein-protein interaction with component residues that must coevolve to produce a functional binding interface to partners that vary as the protein family evolves.

Although known functional sites in GlnBP were successfully predicted, the benchmark dataset used for these analyses was constructed using a simplified process that was chosen due to the limited timeframe available for this process before completion of the thesis. Therefore, subsequent work would necessarily include reconstruction of these benchmark datasets to ensure a wider sampling of sequences and to reduce any phylogenetic bias, following a strategy similar to what was used in the construction of datasets in Chapter 4. Additionally, the number of sequences in the Expanded MSA used for GlnBP functional predictions was reduced by aggressive clustering to enable construction of alignments on a desktop computer with limited memory, which could be avoided by carrying out alignments on dedicated servers. These improvements in the quality of the dataset should also make predictions more informative, by decreasing the noise produced in the metrics by misaligned sequences or spurious alignment due to phylogenetic biases. Another improvement that could be made to the dataset generation step is the substitution of the Jalview Quality score for another metric to estimate how reliable is each column of an alignment, possibly as a strict cutoff based on the frequency of gap characters within a column. Since gaps are ignored by the BLOSUM62 mutation scores that form the basis of the Jalview Quality metric, this value in effect is creating a bias towards more conserved columns while not directly penalising columns with high proportions of gaps.

Application of these metrics to other protein families with deeper characterisation in the literature could also be used to produce additional validation for the strategy proposed here. Protein kinases could be an interesting candidate for a second benchmark family, due to their large diversity in eukaryotes and their relevance in cancer and other diseases (Shchemelinin et al., 2006), which motivates the generation of large amounts of research on their function and effects of mutations (Lahiry et al., 2010; Patani et al., 2016). Intense efforts are also dedicated towards the development of inhibitors for clinical use, due to the difficulties in producing specific inhibitors for members of this structurally-conserved class of proteins (Arora and Scholar, 2005). Because of this difficulty, predictions of specificity-determining residues have been attempted for kinases with the aim of understanding the structural basis for specificity and off-target effects of inhibitor compounds (Caffrey et al., 2008; Bradley et al., 2017). Since these studies have already constructed datasets for functional prediction, the MSAs employed by the authors could be directly used as input for JSD, CMI and NoSE, if they are already of sufficiently high quality. The predictions produced by these analyses could be validated by literature information and any uncharacterised residues — especially ones detected by NoSE analysis targeting specific kinases — would be candidates for validation by experimental characterisation, to determine if any represent previously undetected specificity determinants and, therefore, novel targets for inhibitor development. In addition, comparisons against known benchmarks would allow a deeper characterisation of the NoSE metric itself, producing information about the trade-off between prediction performance and false discovery rate in the method, to allow the selection of thresholds for acceptance of prediction.

The NoSE metric itself could be also improved upon by changing its components that calculate diversity of columns in the Expanded MSA and conservation in the Validated MSA. As an example, the JSD metric could be used as the conservation measure to take into account the different expected frequencies of each amino acid from a neutral background distribution. This change could be beneficial because the JSD metric takes into account the frequency of each residue at neutral columns, assigning higher scores to columns that diverge strongly from these neutral frequencies. However, adopting this metric would also require larger Validated MSAs for reliable computation of frequency distributions for each column, which can be difficult for proteins such as members of the TOMM synthase complex with few validated homologues. The current components of NoSE were selected for producing predictions with such limited datasets and any variations of the calculation will need to be weighed against the available validated data.

The predictions generated here for GlnBP can contribute towards future characterisation and engineering efforts for this protein. For instance, the residues putatively involved in the interaction with the membrane subunit of the ABC transporter complex could be validated by reconstitution of the entire ABC transporter complex encoded by the genes glnH (GlnBP), GlnP (membrane subunit), and GlnQ (intracellular ATP-binding subunit) and a method for detection of activity in GlnBP variants. This could be accomplished by overexpression of the subunits and incorporation into lipid vesicles (Glavinas et al., 2008) or deletion of the genomic copy of glnH in *E. coli* followed by overexpression of mutant variants for cell-based assays. Activity detection can be carried out using radiolabeled glutamine or recovery of cell/vesicle contents followed by HPLC quantitation (Glavinas et al., 2008). Alternatively, a less cumbersome assay could be done by measureing ATP hydrolysis under different conditions, since this process is activated upon GlnBP binding to the membrane subunit and further increased by glutamine binding to GlnBP (Liu et al., 1997).

In addition to the residues highlighted in the comparisons to published functional and structural information for the GlnBP protein, other residues with no obvious functional relevance were also among the top 5% of all values for each metric. These can represent false positive predictions that in reality have no relevance to the function of the protein, but were selected by the predictions due to issues with data quality, sub-optimal choice of thresholds for accepting predictions or even chance (Yu et al., 2018).

Among these residues, some could still have an uncharacterised function that could be initially elucidated by high-throughput experiments with the aim of engineering the function of these solute binding proteins towards novel biosensors.These could be investigated by a saturation mutagenesis (using small intelligent library construction (Tang et al., 2012) to reduce screening effort) screen in microplates of all individual candidate mutations using the fluorescent assays previously employed for engineering of this protein as a biosensor (Looger et al., 2003). Any residues identified by this screen as being important for glutamine binding could be further investigated by complementary characterisation strategies and also exploited for library generation in efforts to isolate biosensors for other small molecules.

Since the assays used for engineering of GlnBP and other SBPs for altered specificity only require site-specific labeling of free cysteine residues and fluorescence intensity measurements (de Lorimier et al., 2002; Looger et al., 2003), this strategy could also be adapted into a higher-throughput screen using GlnBP libraries displayed and labeled on the surface of *E. coli*, followed by FACS. Development of a FACS-based screen would allow libraries of 10^6 or more clones to be screened in sorting runs shorter than 1 hour (Zinchenko et al., 2014), which would enable screening of all possible single (20 residues at 27 sites = 540 variants) and double mutations (540 * 540 = 291600) at the sites picked by the prediction metrics in a single experiment. A screening platform with this throughput would be a powerful tool for engineering of SBPs for recognition of non-cognate small molecules, especially when coupled to a prediction strategy that allows the construction of targeted libraries, such as the one described in this chapter.

Despite the already-acknowledged limitations of the dataset constructed for these benchmarks with GlnBP, the successful identification of residues relevant to the function of this protein by all three prediction metrics demonstrated the validity of this prediction strategy using mixed-function datasets constructed by homology searches. Therefore, the strategy could also be applied for detection of candidate functional residues in the proteins of the TOMM synthase complex.

Chapter 4

Prediction and validation of functional residues in the Microcin B17 TOMM synthase complex

4.1 Introduction

Before the functional prediction strategy described in the previous chapter could be applied towards characterisation of components of the TOMM synthase complex, two main requirements needed to be met: a suitable dataset needed to be constructed for each protein in the complex and a robust assay system was needed to detect changes in activity caused by the selected mutations. Dataset construction followed the strategy used for the GlnBP benchmark (See Section 3.2.2), with modifications to exploit the known TOMM synthases from phylogenetically distinct groups of organisms. However, experimental validation of the candidate functional residues could not be done by *in vitro* assays used in previously-published work on the TOMM synthases (Li et al., 1996; Melby et al., 2012) due to inconsistent results in attempts to establish these assays in our group (See Chapter 5). Therefore, an alternative activity detection method was needed.

The chosen strategy was a bioassay exploiting the toxic phenotype of the *E. coli* microcin B17, on sensitive strains that do not possess the immunity mechanisms encoded in the biosynthetic operon for this TOMM (Garrido et al., 1988; Tran et al., 2005). The plasmid pCID909 (kindly donated by Syngulon SA, Belgium) contains the entire microcin B17

(Mcb17) operon under the control of a constitutive promoter and turns strains carrying it into producing strains resistant to toxic effects of the DNA gyrase inhibitor (Fig. 4.1). Mcb17 is made by the TOMM synthase complex McbBCD acting on the substrate peptide McbA, while resistance to the effects of Mcb17 is conferred by the ABC exporter McbEF and the immunity protein McbG.

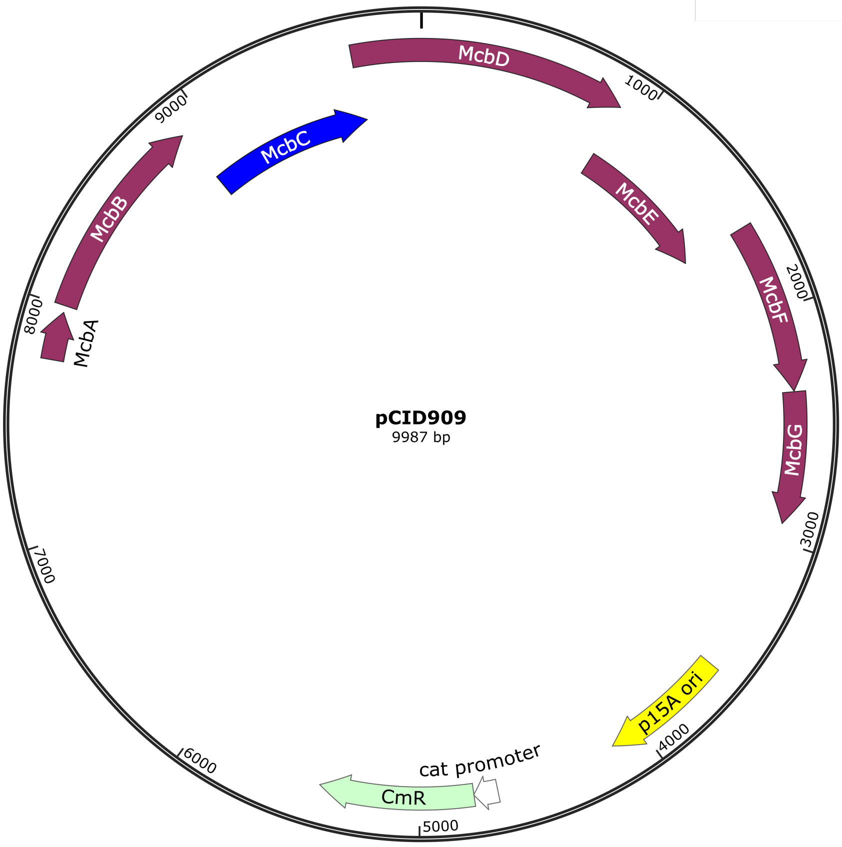


Figure 4.1: Construct for *in vivo* microcin B17 biosynthesis - All genes involved in microcin B17 biosynthesis, extracellular export and immunity are under control of the operon's uncharacterised native promoter. McbA - substrate peptide. McbB - Docking protein. McbC (in blue) - dehydrogenase. McbD - cyclodehydratase. McbE and McbF - ABC transporter subunits. McbG - Immunity protein.

Since the mature Mcb17 product is exported to the extracellular medium, production of this microcin can be detected by its toxic effect on sensitive strains of *E. coli* not carrying the pCID909 vector. The observed degree of cell death can be correlated to the amount of mature Mcb17 produced (Herrero and Moreno, 1986) and the effect of mutations on the biosynthetic components of the operon can be inferred by changes in the toxicity phenotype.

Both activities of the Mcb17 TOMM synthase complex — cyclodehydration to azolines carried out by McbD cyclodehydratase with the McbB docking protein, as well as oxidation to azoles carried out by McbC dehydrogenase — are required for production of mature Mcb17. Therefore, mutations that alter the activities of the complex will have an effect on the rate production of active Mcb17 that can be detected. Any mutation that increases the activity of an enzyme would only be expected to increase Mcb17 production if that step was limiting for the whole pathway, including McbA substrate peptide translation and export by McbEF. However, no information is available regarding the relative rates of *in vivo* cyclodehydration and oxidation, so mutations with a positive effect would not necessarily be detected.

Dataset construction and functional prediction were carried out for all three proteins in the Mcb17 TOMM synthase complex. However only the predictions for McbC dehydrogenase were experimentally validated with the bioassay as a proof-of-concept for the sequence-based functional prediction strategy. Since no structure of a homologous dehydrogenase involved in TOMM biosynthesis has been published, information obtained in these experiments could contribute to the understanding of the function of the TOMM dehydrogenases. In addition, overexpression of TOMM dehydrogenases always led to higher yields than the other proteins in the complex (See Chapter 5) and binding of the FMN cofactor represents an easily-detectable phenotype that could also be affected by mutation, making this protein into an attractive target for further characterisation attempts.

4.2 Results and Discussion

4.2.1 Dataset construction

In order to perform functional residue prediction, a large and diverse MSA was needed for each protein to be studied. As mentioned in Section 1.3, currently, there are only a few

TOMM synthase complexes that have been characterised by *in vitro* methods and structural information is scarce. Therefore, purely sequence homology-based methods were employed to construct the MSAs. The methods used in this process are summarised in Fig. 4.2.

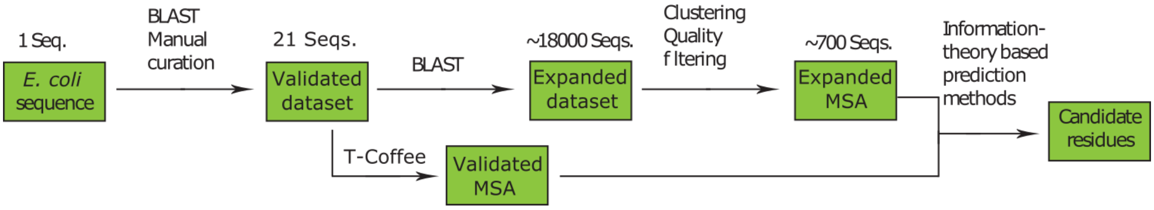


Figure 4.2: Analysis pipeline implemented for the *E. coli* TOMM synthase complex enzymes. The validated dataset was created by manual curation of BLAST hits from the *E. coli* query sequences, to include only sequences found within a genomic cluster that contained homologues for all three enzymes of the TOMM synthase complex. These sequences were aligned by T-Coffee to produce the Validated MSA and also used as queries for a larger BLAST search. The sequences obtained in the second BLAST search were clustered to remove duplicate sequences and very close homologues, then an Expanded MSA was created by realignment of the regions recovered from the BLAST searches and removal of columns with low alignment quality. Both MSAs were then used as input into prediction strategies.

A separate functional residue prediction dataset was created for each of the three enzymes in the *E. coli* TOMM synthase complex, using publicly available sequence databases. The initial set — corresponding to the Validated MSA of Chapter 3 — was constructed by performing blastp searches with the protein sequences from *E. coli*, *B. amyloliquefaciens*, *Bacillus sp.* Al-Hakam, and *Pyrococcus furiosus* as queries against the NCBI nr database. Since all of the currently characterised TOMM synthase complexes are coded by gene clusters of no more than 20kbp, only hits that contained the genes for the other two enzymes in the complex within a window of this size were included in the Validated set. Most of the hits were uncharacterised proteins, but the degree of sequence similarity (>30% identity) and the presence of the two other putative genes in close proximity were considered evidence of possible TOMM synthase activity.

This manual screen led to a set of 21 putative TOMM synthase clusters from Bacteria and Archaea, with three sequences in each. The dehydrogenase, docking protein and cyclodehydratase sequences from each cluster were joined in three FASTA files to be submitted for alignment. A variety of alignment methods were tested (T-Coffee (Notredame et al., 2000), M-Coffee (Moretti et al., 2007), MUSCLE (Edgar, 2004), Clustal Omega (Sievers et al., 2011)) and their results visually inspected to determine which had the best performance for these sequences. T-Coffee consistently produced the best alignments, with the smallest gap insertions, across all three sequences and was used to generate the Validated MSA for each dataset without any subsequent manual adjustment. In the dehydrogenase (McbC in *E. coli*) and docking protein (McbB in *E. coli* MSAs, inclusion of one or more of the sequences led to alignments containing excessive amounts of gaps — possibly due to truncations in the annotated gene sequences or indels that occurred during evolution — leading to poor alignment of homologous regions in the remaining proteins. These individual exceptions were eliminated from the sets and alignment was repeated, leading to Validated MSAs containing 18 to 21 sequences.

However, these numbers of sequences were too small for reliable use of the functional prediction methods, which recommend the use of MSAs containing at least 150 sequences (Martin et al., 2005). To increase the size of the alignments, an expanded dataset was constructed using each sequence in the Validated set as a query in a new round of blastp searches. This second round of searching with phylogenetically distant queries produced more divergent sequences, including proteins annotated as having a non-TOMM synthase function — nitroreductases for the dehydrogenase queries, adenylases for the docking protein query, and the YcaO domain of unknown function for the cyclodehydratase. In total, over 20000 hits were recovered for each of the three complex members with an e-value $<10^{-5}$, but these could not be used directly as inputs for MSA generation by global alignment due to differences in sequence length and greater sequence divergence than in the smaller set.

Therefore, a computational pipeline was implemented to produce the input MSAs for functional prediction using the local pairwise alignments from the blastp output. Sequence analysis software tools were used (USEARCH (http://www.drive5.com/usearch/), exonerate (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate)) and a set of perl and MATLAB scripts for conversion of file formats and construction of the final alignments (see Fig. 4.2 for an overview of the workflow). The first step consisted of filtering the blastp results using e-value ($<10^{-6}$) and coverage (>50%) cutoffs, to ensure that only high-quality alignments remained in the dataset. Each column in the Validated MSAs was also scored by the Quality values calculated by the Jalview alignment software (Waterhouse et al., 2009) and any columns with Quality values below 50 were ignored in the subsequent steps, to focus the analysis on regions that were not predominantly composed of gaps. Then, the sequence IDs of each of the hits were extracted and used to recover the individual target sequences from the nr database. These were clustered in groups of >95% identity and a single representative sequence was kept in the dataset for each cluster, to reduce any biases due to overrepresentation of certain taxonomic groups in the sequences deposited in nr.

After clustering, the pairwise alignments containing the selected sequences from each cluster were extracted from the blastp results. Some of these target sequences had significant alignments with more than one of the original queries, so these were compared and only the highest-scoring alignment was retained. After this step, each dataset contained at least 700 separate pairwise alignments of a filtered hit sequence to one of the original queries from the Validated MSA.

To convert the separate blastp pairwise alignments into a single MSA, the aligned residues in each pairwise alignment were mapped onto the original Validated MSA. This was done by comparing the aligned query residue to its position in the Validated MSA and adding the corresponding residue from the target sequence into the equivalent position of a new Expanded MSA. These final datasets were then exported into a FASTA format that could be submitted to functional prediction web servers and used for calculation of the

NoSE metric, encompassing a Validated MSA and an Expanded MSA for each of the three proteins in the TOMM synthase complex.

4.2.2 Functional residue predictions

All three prediction scores (JSD, CMI and NoSE) were applied to the datasets generated in the previous section and analysis of results was carried out similarly to Section 3.2.3. Residues with the top 5% values for each of the scoring methods were selected for analysis. Since none of the proteins in the Microcin B17 synthase complex have had their structures characterised, all structural comparisons were carried out with homology models built using the Phyre2 server (Kelley and Sternberg, 2009). All of the homology models described below had Phyre2 Confidence values of 100%, which indicates the probability that templates used for modeling are true homologues of the query protein, Confidence values above 90% correspond to 2-4 Å RMSD at the core of the structure, but surface loops are modelled at lower accuracy (Kelley et al., 2015)

4.2.3 Cyclodehydratase functional residue prediction

Seventeen residues were selected from the highest-scoring sites in each prediction metric for further analysis from the 332 scored sites in the McbD dataset. A small overlap was observed between the residues selected by each metric, with one site shared by JSD and CMI and another shared by CMI and NoSE (Fig. 4.3).

Homology modelling by Phyre2 produced a model (Fig. 4.4 a-b) using the cyanobacterial bifunctional TruD heterocyclase (PDB ID: 4bs9) as template, with 15% identity and 87% sequence coverage (Koehnke et al., 2013). This protein contains the cyclodehydratase ("D") and docking ("C") domains in a single polypeptide and one of its domains aligned structurally to the McbD homology model (Fig. 4.4 c-d) . Since TruD was crystallised in the absence of any ligands, the structure of the YcaO protein with the co-crystallised ATP

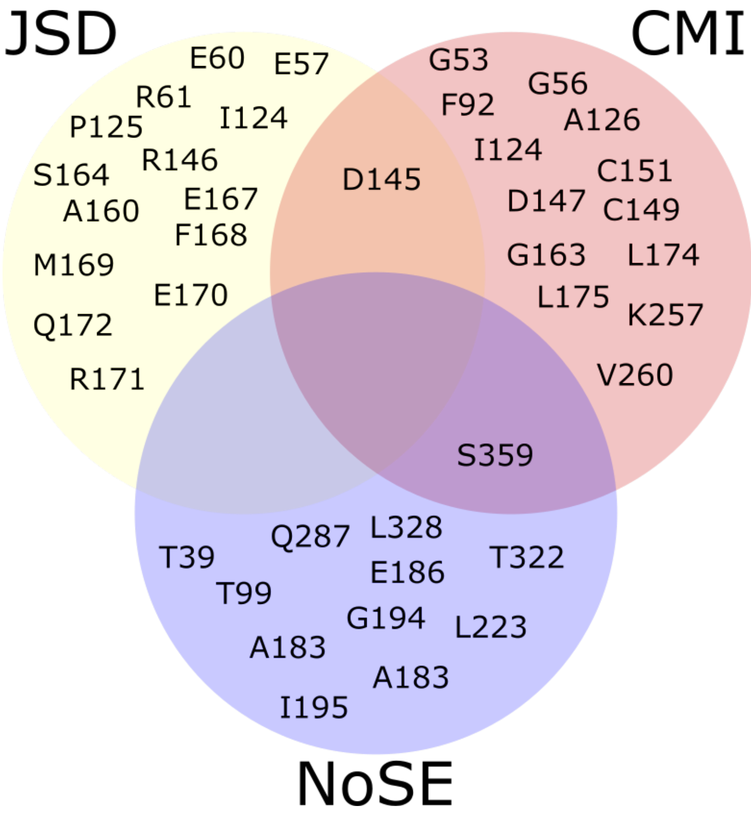


Figure 4.3: Venn diagram depicting overlaps between the top 5%-scoring residues in McbD for JSD (yellow), CMI (red), and NoSE(blue).

analogue AMPCPP (PDB ID: 4q85, Dunbar et al. (2014)) was also structurally aligned to the homology model and the position of AMPCPP in that structure was highlighted in relation to the model (Fig. 4.4, in green). This *E. coli* protein of uncharacterised function is homologous to the TOMM cyclodehydratases and names the superfamily in which they are included (CDD cl19253, Marchler-Bauer et al. (2017)). Several residues homologous to the ATP-binding site in YcaO were previously mutated and characterised in the *Bacillus sp.* Al-Hakam BalhD cyclodehydratase, confirming that the YcaO structure is representative of the TOMM cyclodehydratases (Dunbar et al., 2014). In the McbD model, the aligned AMCPP molecule is located at the end of a cleft in the structure, which can possibly represent the active site of the protein, away from the putative interaction surface with the docking protein implied by the bifunctional TruD structure (Fig. 4.4 c)

The highest-scoring residues from the JSD prediction were distributed throughout the primary sequence (Fig. 4.5a), but clustered spatially around the predicted ATP binding

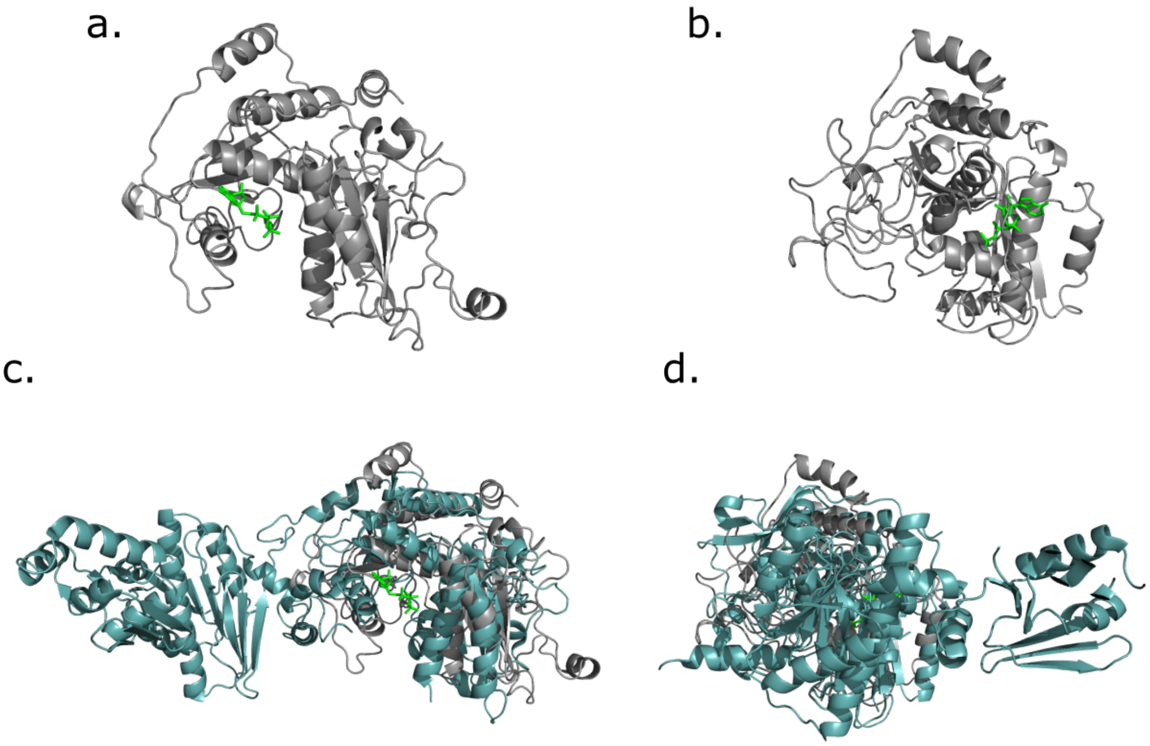


Figure 4.4: Phyre2 Homology model of McbD. (a) and (b) Views of McbD model with structurally aligned AMCPP from YcaO (PDB ID 4q85) highlighted in green. (c) and (d) Views of McbD model with the structure of TruD (PDB ID 4b29) overlaid in cyan. Views are rotated 90 °around the y axis.

site, away from the putative binding interface to the docking protein suggested by the bifunctional TruD structure (Fig. 4.5c-d). Three of these residues are in positions similar to important sites from the YcaO structure. Glutamates 57 and 170 are in similar positions to Glu75 and Glu202 in YcaO, which bind Mg^{2+} ions (Dunbar et al., 2014). Arg171 appears homologous to R203 in YcaO, coordinating the γ-phosphate of bound AMPCPP (Dunbar et al., 2014). Since the remaining residues are in contact with these known functional sites, their strong conservation detected by JSD could indicate a role in structuring the ATP binding pocket and in catalytic function.

Similarly to JSD, the highest scoring residues for CMI were dispersed throughout the protein sequence, but were physically located around the predicted position for ATP (Fig. 4.6). Gly56 is located on the distal side of this pocket compared with the other selected sites,

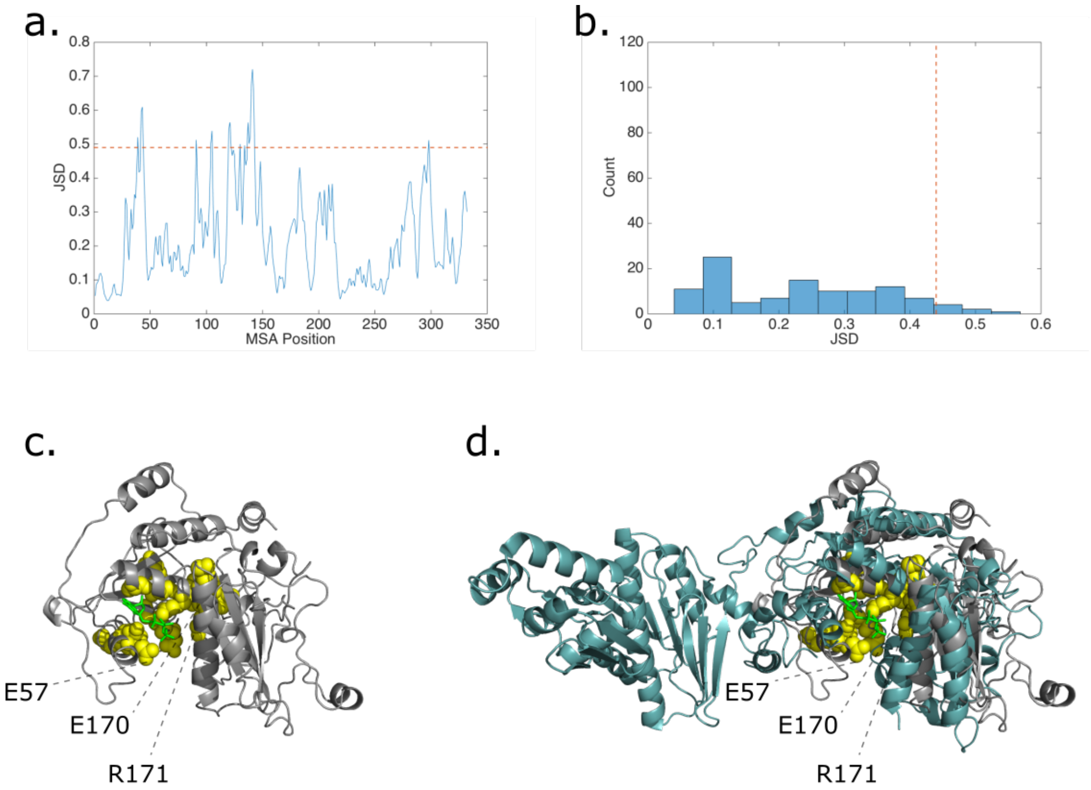


FIGURE 4.5: Highest scoring residues for JSD in McbD. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of JSD scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in yellow on the McbD homology model. (d) Top 5% residues highlighted in yellow on the homology model, after structural alignment to the TruD cyanobacterial bifunctional heterocyclase/docking protein (PDB: 4ds9), in cyan. The YcaO domain structure (PDB: 4q85) was also structurally aligned and its bound ATP analog (non-hydrolysable AMPCPP) is highlighted in green in both panels.

and forms part of the hydrophobic surface along the ATP-binding pocket of YcaO (Dunbar et al., 2014). The nearest apparent homologue to Lys257 is YcaO-Arg286, suggesting a role for this lysine in interaction with the α-phosphate of AMPCPP. In addition to the residues lining the ATP binding cavity, sites in the hydrophobic core of the domain could be important for correct folding of the protein, requiring compensatory mutations at other sites whenever they vary and producing the high CMI values observed.

Unlike the other two scores, NoSE-selected residues were dispersed throughout both the sequence and structure of McbD (Fig. 4.7). Threonines 39 and 322 are located towards the

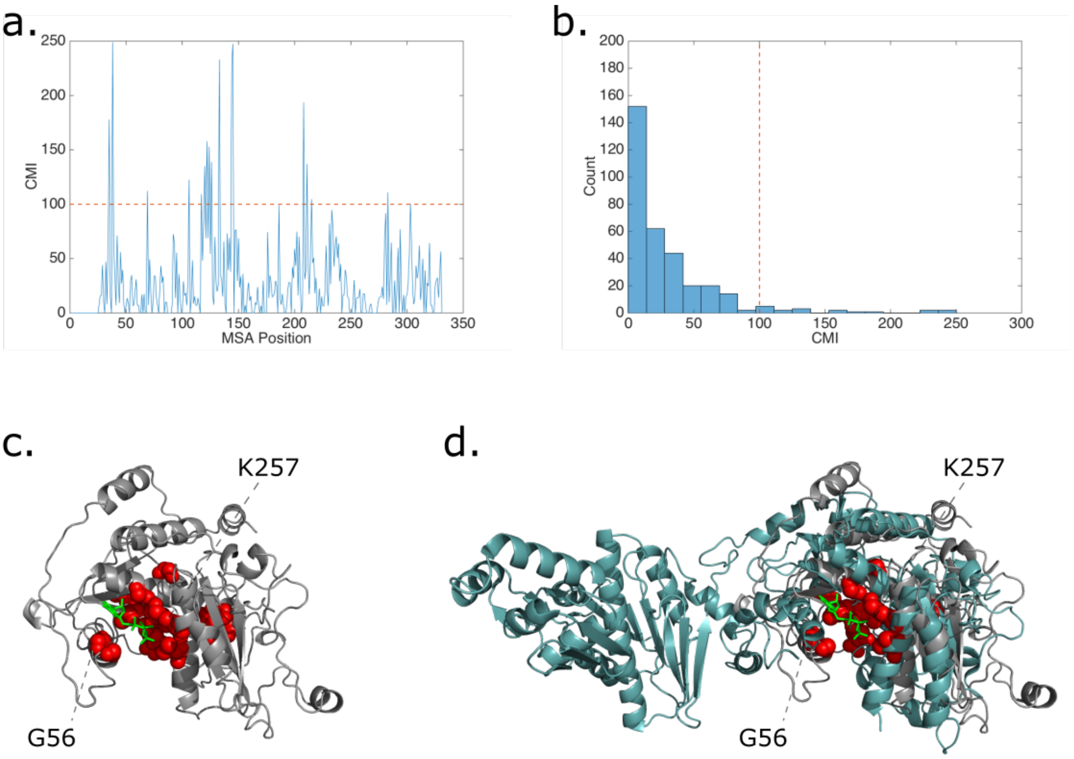


Figure 4.6: Highest scoring residues for CMI in McbD. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of CMI scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in red on the McbD homology model. (d) Top 5% residues highlighted in red on the homology model, after structural alignment to the TruD cyanobacterial bifunctional heterocyclase/docking protein (PDB: 4ds9), in cyan. The YcaO domain structure (PDB: 4q85) was also structurally aligned and its bound ATP analog (non-hydrolysable AMPCPP) is highlighted in green in both panels.

surface of the model in positions that suggest a possible function in the cyclodeydratase-docking protein interaction, which has been shown to require the C-terminal region of TOMM cyclodehydratases (Dunbar et al., 2014). The fact that none of the residues selected by NoSE map to the AMPCPP binding pocket suggests that the proteins in the expanded dataset share a conserved ATP-binding function, which would not be expected to have enough sequence divergence to score highly in NoSE. The location of the selected residues surrounding the ATP-binding site and the putative active site could indicate a more specific role in catalysis, allosteric regulation or substrate specificity.

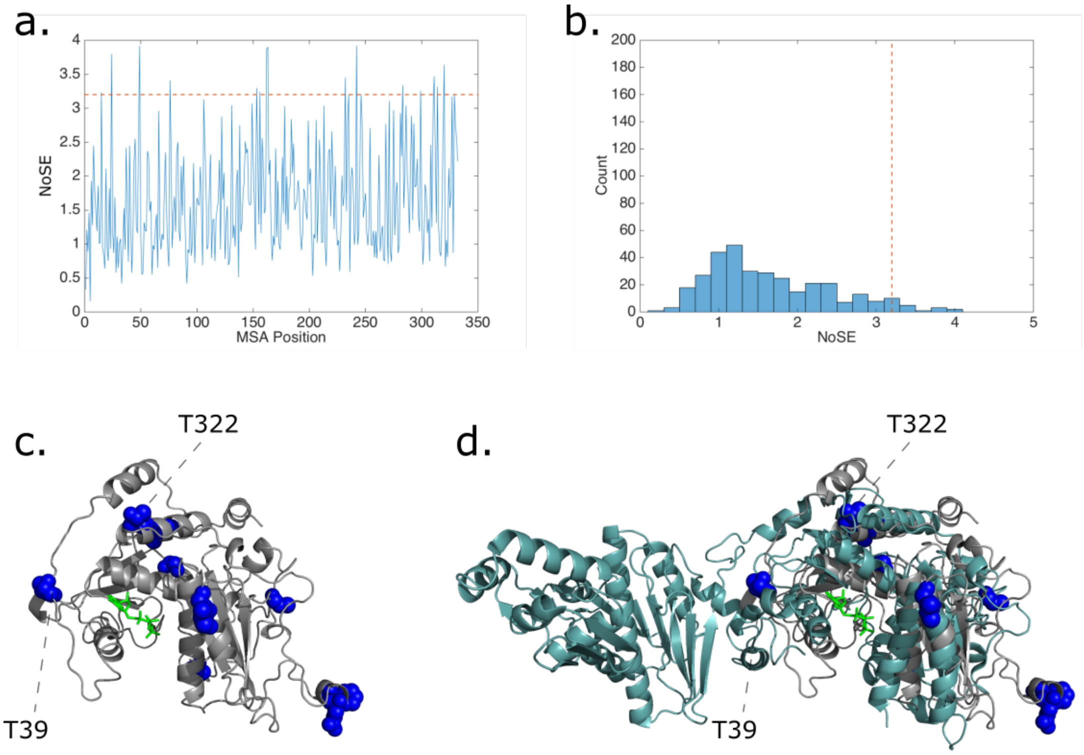


Figure 4.7: Highest scoring residues for NoSE in McbD. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of NoSE scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in blue on the McbD homology model. (d) Top 5% residues highlighted in blue on the homology model, after structural alignment to the TruD cyanobacterial bifunctional heterocyclase/docking protein (PDB: 4ds9), in cyan. The YcaO domain structure (PDB: 4q85) was also structurally aligned and its bound ATP analog (non-hydrolysable AMPCPP) is highlighted in green in both panels.

4.2.3.1 Docking protein functional residue prediction

Phyre2 failed to produce a high-confidence model for the docking protein *E. coli* McbB, so the analyses presented here were carried out using the homologous BamC from *Bacillus sp.* Al-Hakam. The homologue used for modeling was also the cyanobacterial bifunctional TruD heterocyclase (PDB ID:4bs9) (Koehnke et al., 2013), with 18% identity to BamC and the aligned region covered 92% of the query sequence (Fig. 4.8). Although no ligands for the cyclodehydratase domain were co-crystallised in this structure, the structural Zn^{2+} ion of the docking domain was present and structural alignment was used to predict its position relative to selected residues in BamC, which is near the putative binding interface with the cyclodehydratase subunit (Fig. 4.8).

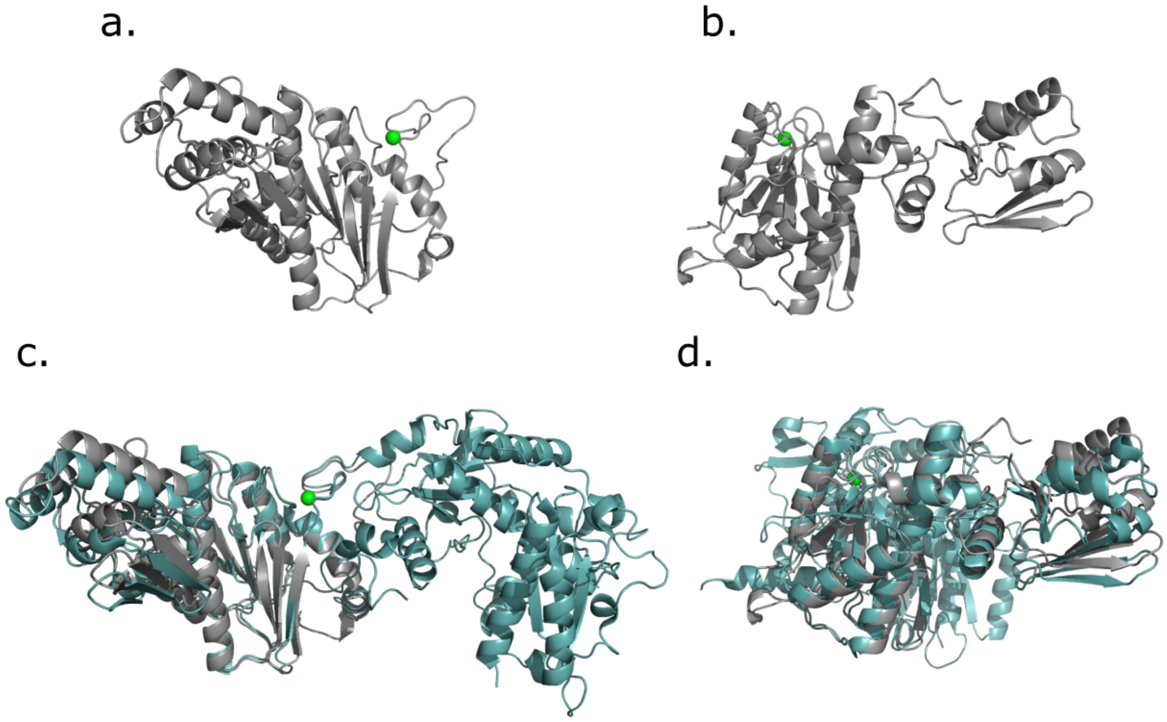


FIGURE 4.8: Phyre2 Homology model of BamC. (a) and (b) Views of BamC model. (c) and (d) Views of BamC model with the structure of TruD (PDB ID 4b29) overlaid in cyan. The zinc ion from the TruD structure is highlighted in green. Views are rotated 90 °around the y axis.

Six out of 109 scored sites were further analysed for each scoring method in the docking protein dataset. As was observed for McbD, a single high-scoring site was shared by JSD and CMI and another site was shared by CMI and NoSE (Fig. 4.9).

The residues with the highest JSD values were dispersed throughout the sequence, but the first thirty residues had distinctly lower values, indicating a low degree of sequence conservation in the N-terminal region of this family (Fig. 4.10 a). The selected residues clustered in one region of the structure and especially around the predicted Zn^{2+} binding site (Fig. 4.10 c-d, Zn^{2+} obscured by highligted residues in c). Of note are three Cys residues — 216, 308, and 311 — positioned around the predicted ion binding site, suggesting a role

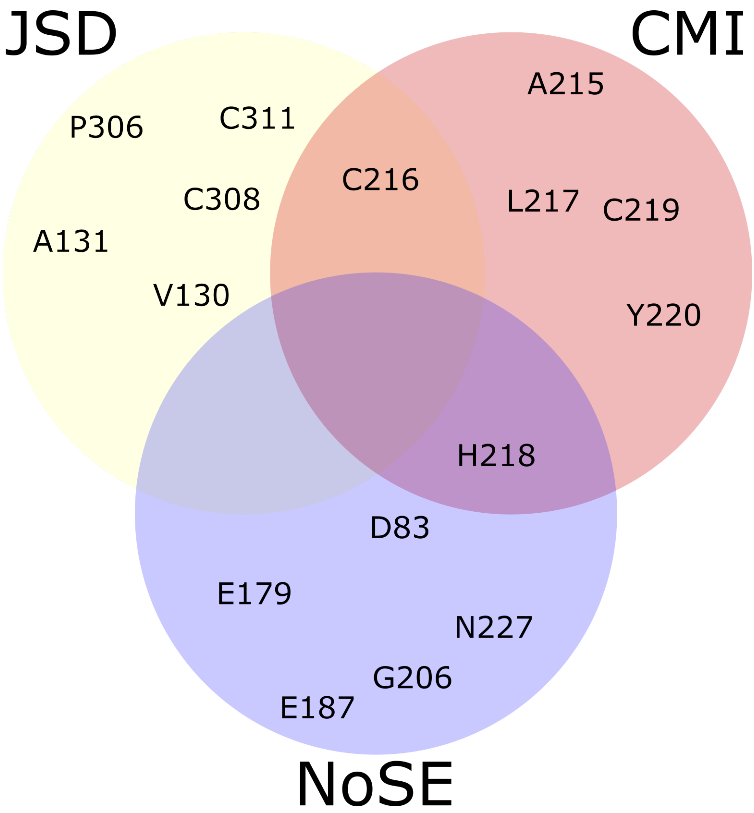


FIGURE 4.9: Venn diagram depicting overlaps between the top 5%-scoring residues in BamC for JSD (yellow), CMI (red), and NoSE (blue).

in coordination of the metal ion that is carried out by cysteines in many known structures (Pace and Weerapana, 2014).

Unlike the observed distribution along the sequence for JSD scores, all of the top 5% of CMI scores corresponded to six contiguous resides in the BamC sequence (Fig. 4.11 a). These were located near the Zn^{2+} binding site, with the positions of Cys216 (also high-scoring in JSD) and Cys219 suggesting a role in ion coordination. The imidazole ring of His218 is not oriented towards the zinc ion as is observed in structures that coordinate zinc via this amino acid (Alberts et al., 1998), suggesting that the four identified cysteines carry out this function in BamC and His218 could have an indirect function in shaping the binding site.

As was observed for McbD (Section 4.2.3), the residues with the highest NoSE scores were dispersed throughout the primary and tertiary structures of BamC (Fig. 4.12 a). His218 was also selected in this prediction, indicating a high degree of conservation within

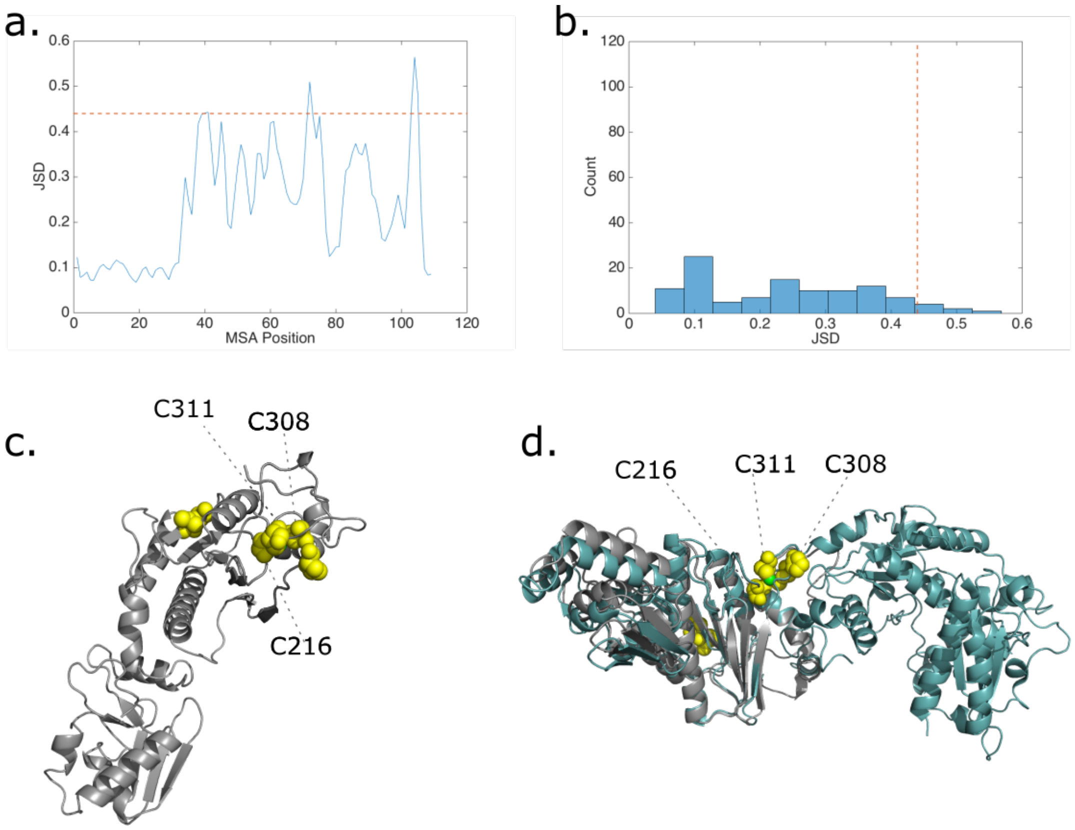


Figure 4.10: Highest scoring residues for JSD in BamC. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of JSD scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in yellow on the BamC homology model. (d) Top 5% residues highlighted in yellow on the model, after structural alignment to the TruD cyanobacterial bifunctional heterocyclase/docking protein (PDB: 4ds9), in cyan. The bound Zn^{2+} from the TruD structure is highlighted in green in both panels. The structures have been rotated 90°along the horizontal axis between panels for a clearer visualisation of highlighted residues and the putative interface with the cyclodehydratase partner.

the set of TOMM-related proteins but more relaxed evolutionary constraint in the larger dataset, leading to a high score in the CMI measure of coevolution (Fig. 4.12 c). In addition, Asn227 is adjacent to an extended loop from the cyclodehydratase domain of the TruD structure, suggesting a possible role in heterodimerisation (Fig. 4.12 d)

4.2.4 Dehydrogenase functional residue prediction

A homology model with a putative nitroreductase from *Ralstonia eutropha* as the closest structural match (PDB ID: 3hj9) was produced, with 20% identity in an alignment covering 90% of the query sequence (Fig. 4.13). This protein was crystallised as a dimer, with

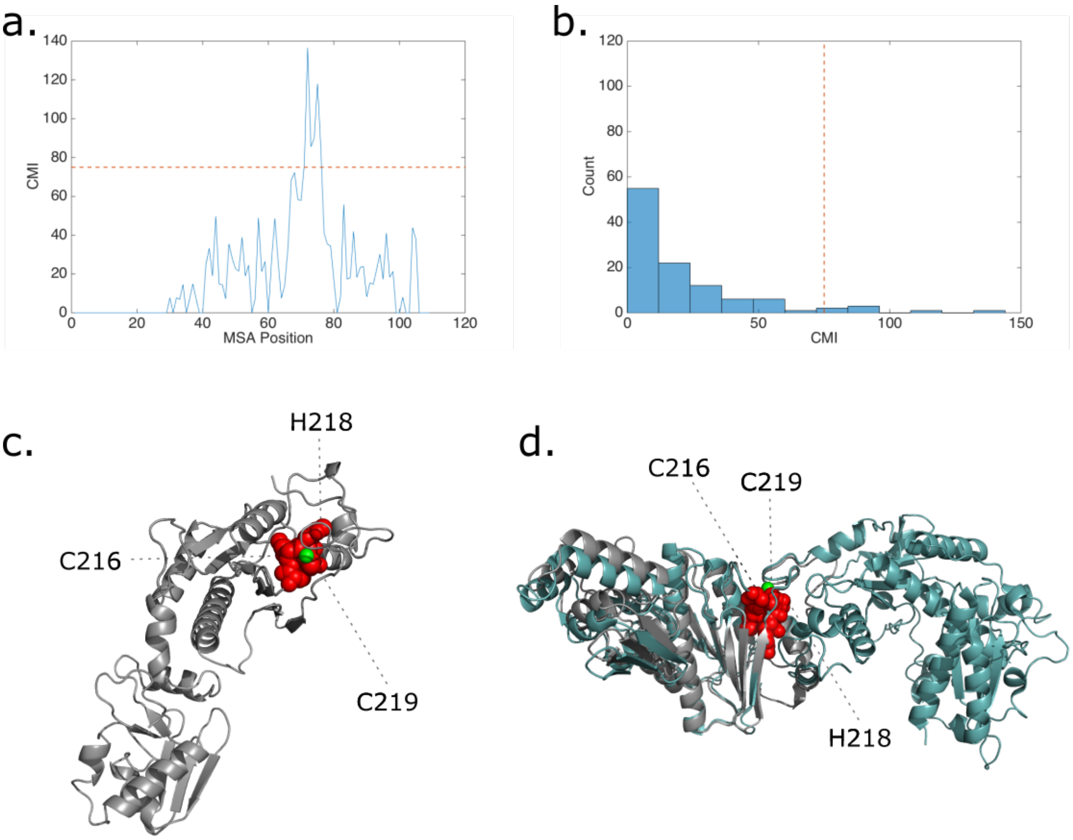


Figure 4.11: Highest scoring residues for CMI in BamC. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of CMI scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in red on the BamC homology model. (d) Top 5% residues highlighted in red on the model, after structural alignment to the TruD cyanobacterial bifunctional heterocyclase/docking protein (PDB: 4ds9), in cyan. The bound Zn^{2+} from the TruD structure is highlighted in green in both panels.The structures have been rotated 90°along the horizontal axis between panels for a clearer visualisation of highlighted residues and the putative interface with the cyclodehydratase partner.

two molecules of the FMN cofactor bound in distinct sites with contributions from both subunits. The McbC homology model was structurally aligned to one of the subunits, to identify potential roles for predicted functional residues in cofactor binding or subunit interactions. The core of the protein monomers aligned well, but there was considerable divergence between the *Ralstonia* structure and the McbC model (Fig. 4.13 c-d), which can reflect real structural divergences between the homologues or modelling errors due to the lower accuracy of the software for surface loops (Kelley et al., 2015).

All of the residues discussed in the text below were selected for mutation and validation by *in vivo* Microcin B17 biosynthesis assays, due to proximity to the FMN binding and

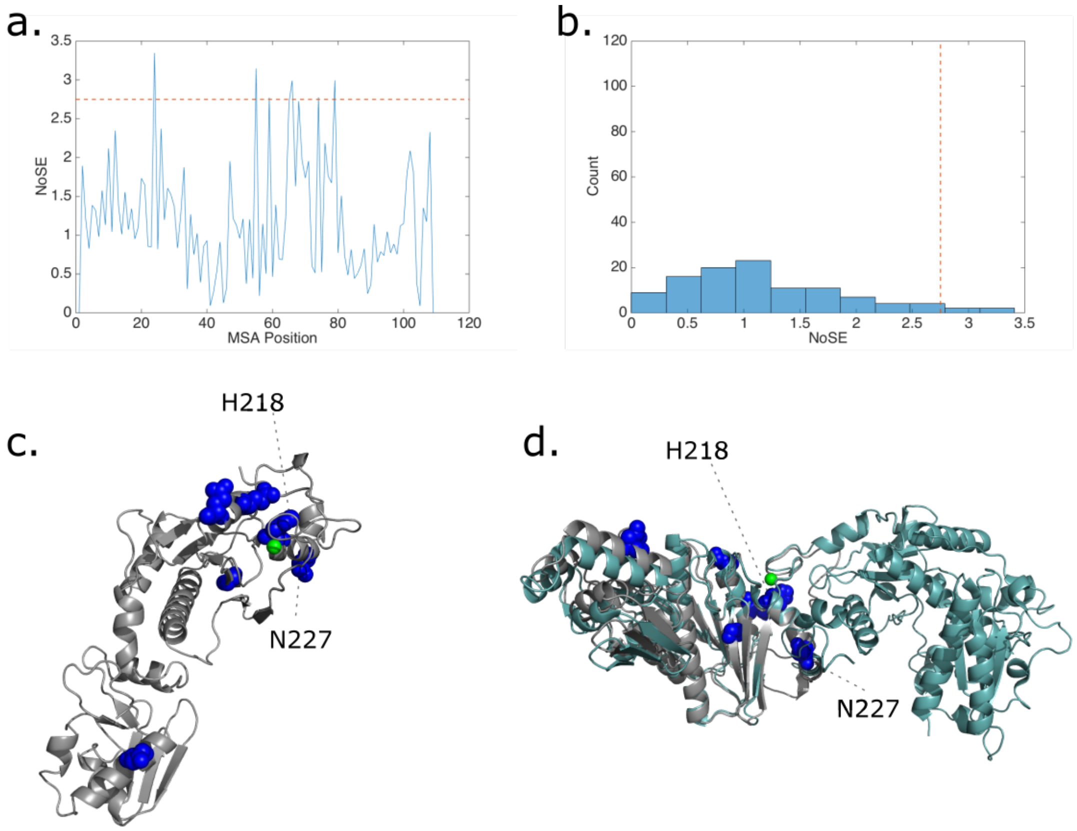


Figure 4.12: Highest scoring residues for NoSE in BamC. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of NoSE scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in blue on the BamC homology model. (d) Top 5% residues highlighted in blue on the model, after structural alignment to the TruD cyanobacterial bifunctional heterocyclase/docking protein (PDB: 4ds9), in cyan. The bound Zn^{2+} from the TruD structure is highlighted in green in both panels.The structures have been rotated 90°along the horizontal axis between panels for a clearer visualisation of highlighted residues and the putative interface with the cyclodehydratase partner.

putative active site, possible roles in protein-protein interaction, and structural integrity. Twelve out of 225 high-quality sites from the McbC dehydrogenase alignment were selected for further analysis and as candidates for validation by mutation and *in vivo* activity assays. In this dataset, five high-scoring sites were found in both the JSD and CMI datasets and a single site was present in both the NoSE and CMI sets (Fig. 4.14).

The highest-scoring residues from the JSD metric were located in the middle region of the sequence, with distinctly lower scores at the N- and C-termini of the input alignment (Fig. 4.15 a). Structurally, these predicted residues clustered mainly around one of the

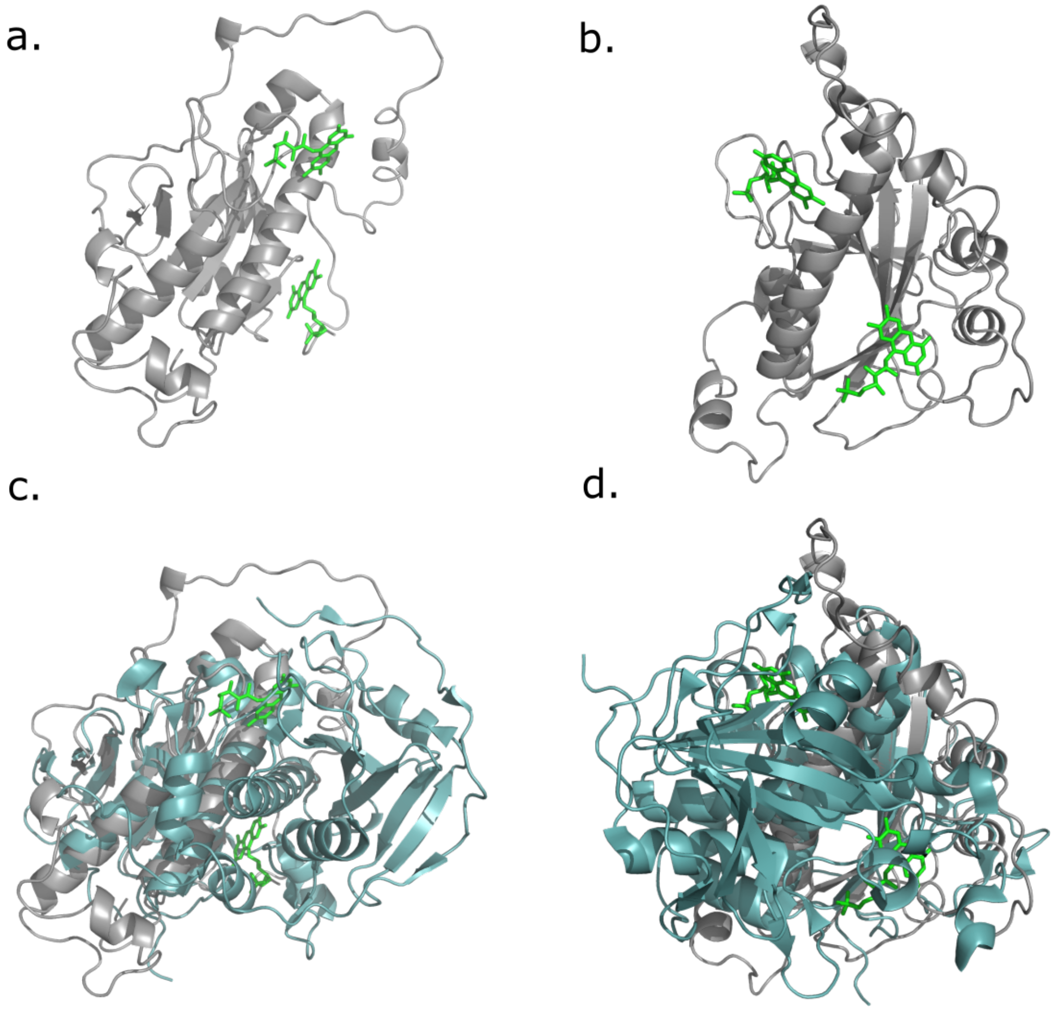


FIGURE 4.13: Phyre2 Homology model of McbC. (a) and (b) Views of BamC model. (c) and (d) Views of McbC model with the structure of *Ralstonia eutropha* nitroreductase (PDB ID:3hj9) (PDB ID 4b29) overlaid in cyan. The putative positions of the FMN cofactors are highlighted in green. Views are rotated 90 °around the y axis.

FMN half-sites and also in a hydrophobic region away from the FMN sites and the dimer interface (Fig. 4.15 c-d).

One of the residues within the top 5% JSD values — Tyr202 — has been previously demonstrated to be required for azoline oxidation activity, but not for FMN binding, by mutation of the homologous Tyr202 of BcerB dehydrogenase to Ala (Dunbar and Mitchell, 2013), so this site was selected for mutation as an expected negative control for McbC activity. The position of Gly123 suggests its main chain could form a hydrogen bond to one of the FMN hydroxyl groups and Gly215 is positioned along the long α-helix in the

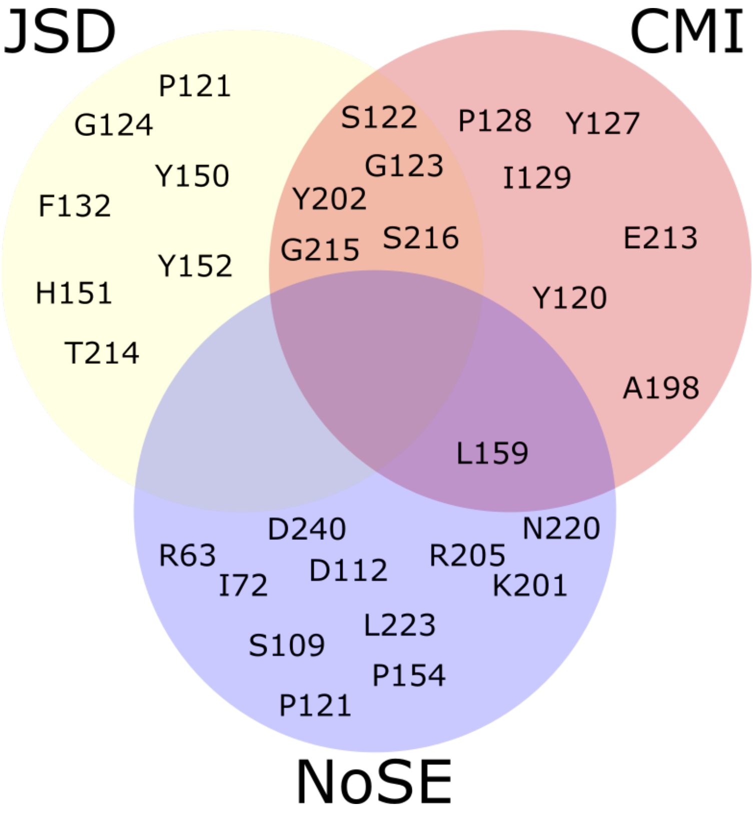


FIGURE 4.14: Venn diagram depicting overlaps between the top 5%-scoring residues in McbC for JSD (yellow), CMI (red), and NoSE(blue).

dimer interface, indicating a possible role in dimerisation. Finally, a set of three contiguous aromatic residues highlighted by JSD — Tyr150, His151, and Tyr152 — was selected to determine whether these have a shared role in maintaining structural integrity of the complex that is conserved in the larger family of homologues.

As was observed for JSD, the top 5% of CMI scores were located in the central region of the protein sequence, with markedly lower scores at the N- and C-termini of the alignment (Fig. 4.16 a). However, the residues highlighted by CMI were more clustered spatially, mostly around the FMN half-site that includes Tyr202 — present in the high-scoring residues for both JSD and CMI (Fig. 4.16 c-d). Residues Tyr120, Ser122 and Glu213 could form contacts with FMN or be part of the larger cavity which would be needed to accomodate the azoline substrate. Ala198 and Ser216 are more distant from the putative binding site and could contribute to interactions that shape the active site.

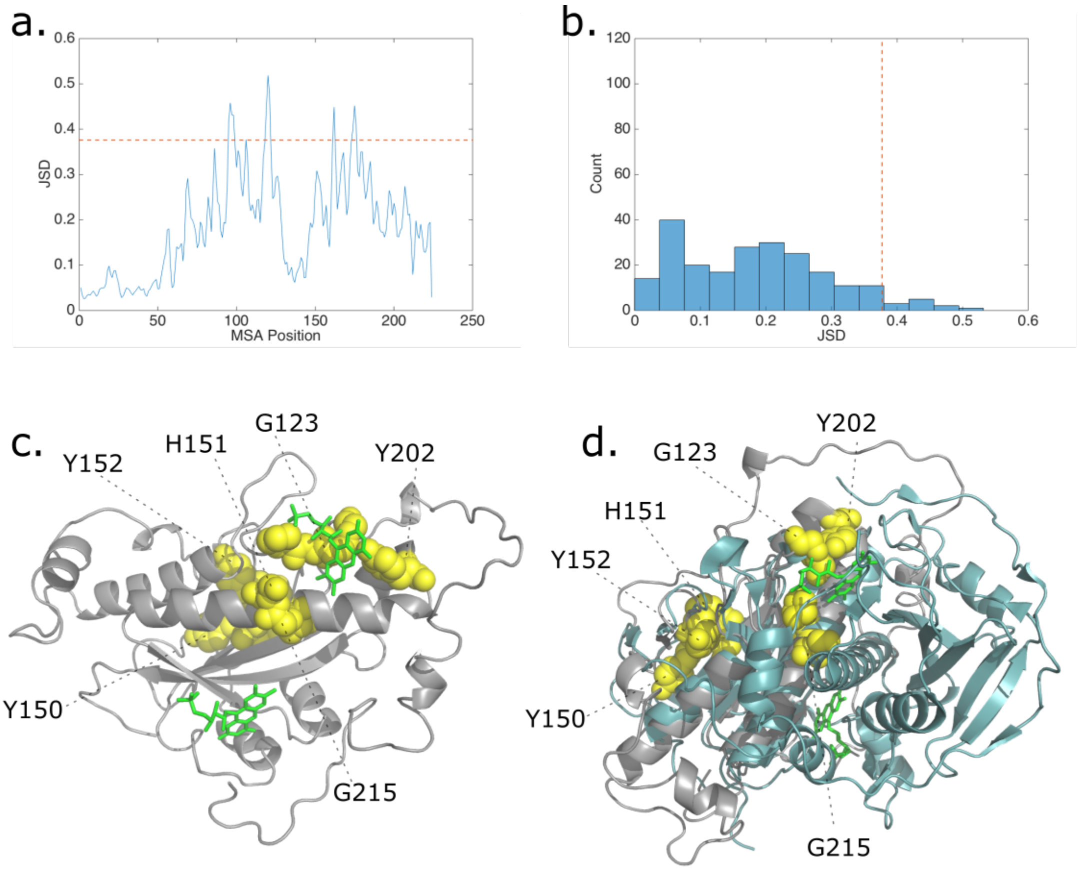


Figure 4.15: Highest scoring residues for JSD in McbC. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of JSD scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in yellow on the McbC homology model. (d) Top 5% residues highlighted in yellow on the homology model, after structural alignment to *A. variabilis* putative nitroreductase (PDB: 3eo7), with both subunits of the nitroreductase dimer displayed in cyan. The FMN cofactors from the structurally aligned nitroreductase (PDB: 3eo7) are highlighted in green in both panels and the views have been rotated for clear display of highlighted residues and dimer subunits.

The NoSE scores were more evenly distributed throughout the sequence and the structural model than the previous two scores (Fig. 4.17). Lys201 is homologous to the other known catalytically-important residue in the TOMM dehydrogenases, Lys201 of BcerB — also leading to loss of activity but maintaining FMN binding upon mutation to Ala (Dunbar and Mitchell, 2013). Leu159, Arg204 and Phe132 were selected due to their proximity to the long α-helix in the putative dimer interface. Finally, the isolated residues Ile72 and Asp240 were selected for having no obvious function based on their predicted positions, especially the the hydrophobic and apparently surface-exposed isoleucine.

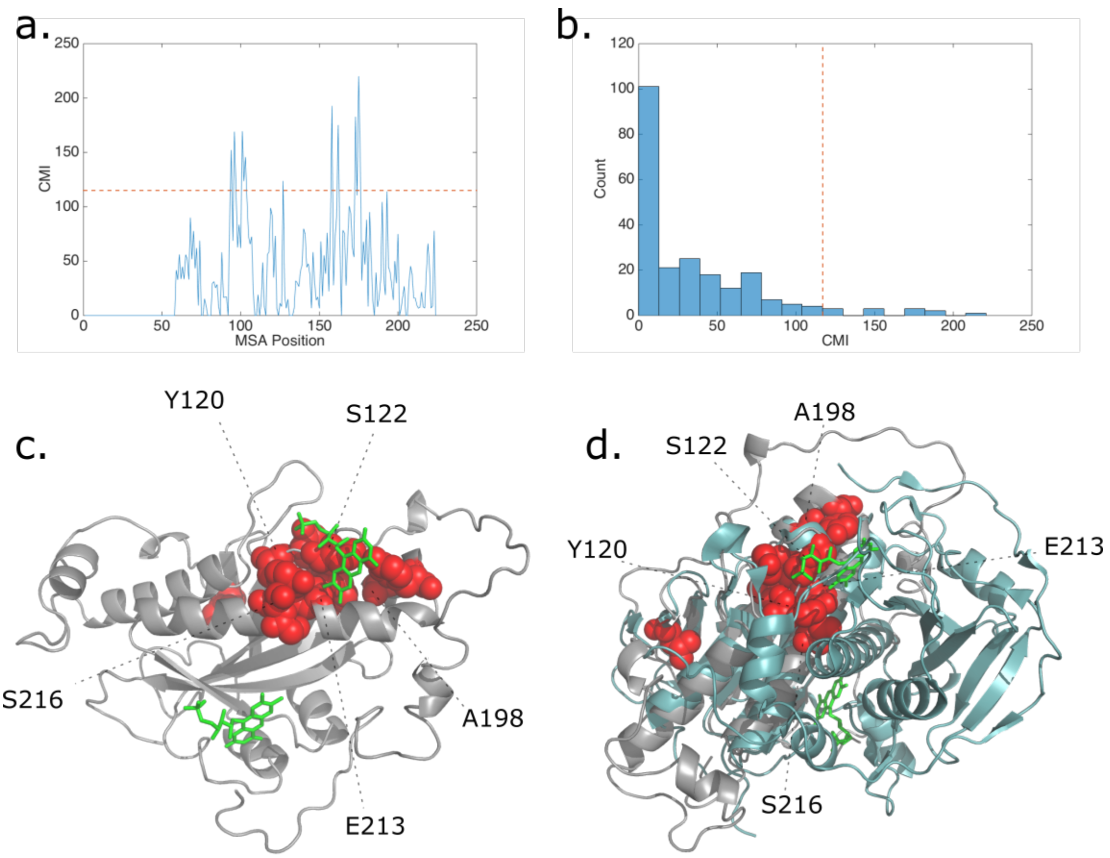


FIGURE 4.16: Highest scoring residues for CMI in McbC. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of CMI scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in red on the McbC homology model. (d) Top 5% residues highlighted in red on the homology model, after structural alignment to *A. variabilis* putative nitroreductase (PDB: 3eo7), with both subunits of the nitroreductase dimer displayed in cyan. The FMN cofactors from the structurally aligned nitroreductase (PDB: 3eo7) are highlighted in green in both panels and the views have been rotated for clear display of highlighted residues and dimer subunits.

The residues in McbC selected for mutation were: I72, Y120, S122, G123, F132, Y150, H151, Y152, L159, A198, K201, Y202, R205, E213, G125, S215, and D240. Six sites were picked for JSD and NoSE predictions and five from CMI, to produce a total of 18 variants for assaying (when the WT is included).

4.2.5 Validation of predictions by *in vivo* microcin B17 synthesis assays

The selected putative functional residues in McbC dehydrogenase were mutated to cysteine in vector pCID909 (See Section 2.2.8 for mutagenesis methods) to produce the panel of

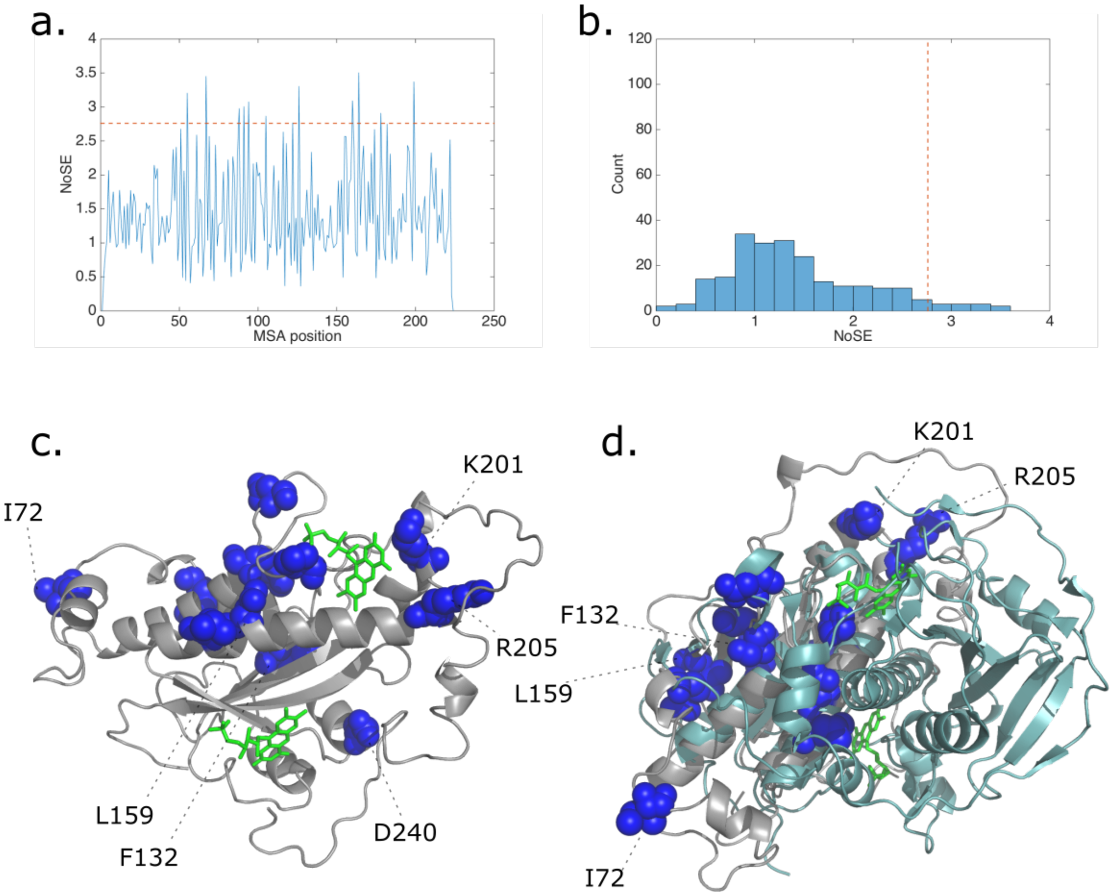


Figure 4.17: Highest scoring residues for NoSE in McbC. (a) Score for each residue in the dataset, the red dotted line is the cutoff point for the top 5% of all scores. (b) Observed distribution of NoSE scores, the red dotted line is the cutoff point for the top 5% of all scores. (c) Top 5% residues highlighted in blue on the McbC homology model. (d) Top 5% residues highlighted in blue on the homology model, after structural alignment to *A. variabilis* putative nitroreductase (PDB: 3eo7), with both subunits of the nitroreductase dimer displayed in cyan. The FMN cofactors from the structurally aligned nitroreductase (PDB: 3eo7) are highlighted in green in both panels and the views have been rotated for clear display of highlighted residues and dimer subunits. Residue Asp240 is obscured by the remainder of the structure in this view

mutants for validation. A cysteine scan was selected due to a general tolerance for cysteine in non-essential sites in proteins revealed by deep mutational scan studies (Whitehead et al., 2012) and for the possibility of labeling with thiol-reactive probes in future characterisation efforts. Construction of the Y150C variant in pCID909 failed after multiple attempts and characterisation was carried out with the remaining mutants.

4.2.5.1 Microcin B17 bioassay development

Prior to validation of the mutant panel, a set of conditions for optimal detection of Mcb17 production needed to be established. These conditions include strain choice for both production and activity detection, incubation temperatures and times, and the layout of the experiment for exposure of sensitive strains to the microcins.

Initially, a small panel of three *E. coli* strains (BL21(DE3), NEB 10-β, NEB T7 Express lysY/Iq) was tested in all nine possible combinations of producing (strain transformed with the pCID909 vector) and sensitive (no plasmid) roles. Sensitive strains were added to the plate to form a lawn, followed by spotting of producing strains onto the plate to allow formation of a colony. Detection of a zone of growth inhibition was expected around producing strain spots, where the local Mcb17 concentration became high enough for DNA gyrase inhibition to lead to cell death of the sensitive lawn.

Sensitive and producing strains were grown for 16 h at 37°C in LB medium containing chloramphenicol (producing strains) or no antibiotic (sensitive strains). Then, a lawn containing 10^7 CFU of the sensitive strains was plated onto LB agar plates with no antibiotic, 10 µl of producing strains (containing approximately 10^6 CFU) was spotted in triplicate onto the lawn, followed by overnight incubation at 37°C (Fig. 4.18). The BL21(DE3) strain appeared to be the most efficient Mcb17 producer and also most susceptible to Mcb17 toxicity when not carrying the operon. Strains T7 Express lysY/Iq and NEB 10-β produced smaller growth inhibition zones on sensitive strains and were also more resistant to the effects of the microcin.

Although growth inhibition was clearly visible in the tested plates, the cleared zones and producing strain spots were irregularly-shaped (Fig. 4.18), which would represent a source of noise during comparison of mutant strains by measurements of inhibition zones. This effect was probably caused by varying levels of moisture on each plate causing the producing strain spots to spread unevenly after pipetting, so alternative strategies were sought to avoid this issue.

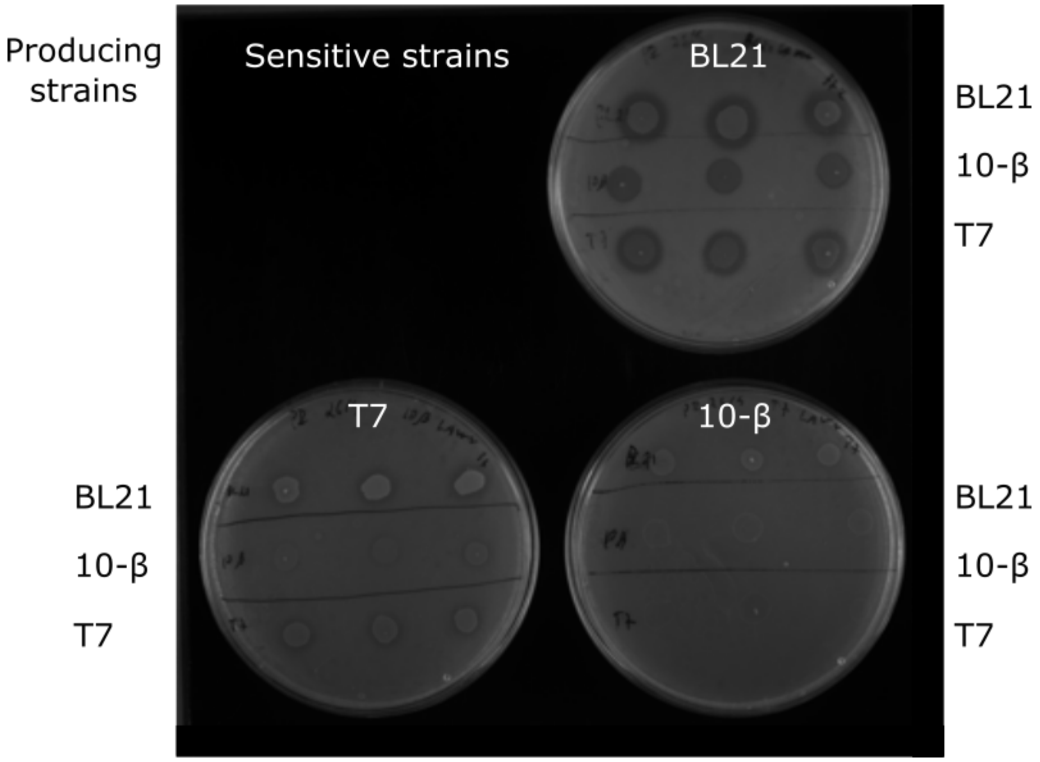


Figure 4.18: Test of producing and sensitive strains for Microcin B17 bioassay - Producing strains spotted in triplicate over sensitive strain lawns to determine microcin B17 biopthetic capability and susceptibility. BL21 - BL21(DE3); 10-β - NEB 10-β; T7 - T7 Express lysY/Iq

The same experiment was repeated with an inverted order of addition of the strains, in an attempt to create homogeneous circular inhibition zones. Producing strains were spotted onto fresh culture plates, followed by addition of the sensitive strain. Sensitive strains (10^7 CFU) were added to 3 ml 0.7% agar and spread onto plates containing dried producing strain spots. The overall size and shape of the growth inhibition zones was more consistent in this test, but producing strain spots could also be seen away from their intended positions, carried by the agar overlay as it was spread (Fig. 4.19).

The same trend for more efficient production of Mcb17 and high susceptibility by the BL21(DE3) strain was observed, so this strain was used for all subsequent experiments. Contrary to the previous experiment, the sensitive 10-β strain showed inhibition by the producing strains in this test, which could be caused by susceptibility varying during growth of this strain. Since the BL21(DE3) did not show such variation, it was also preferred for this consistency. Coverage of the plate by the sensitive strain was also more consistent with the overlay method, but the sensitive strain would need to be diluted into a higher volume of agar solution to avoid gaps in the overlay caused by premature solidification (as seen in the negative control of Fig. 4.19)

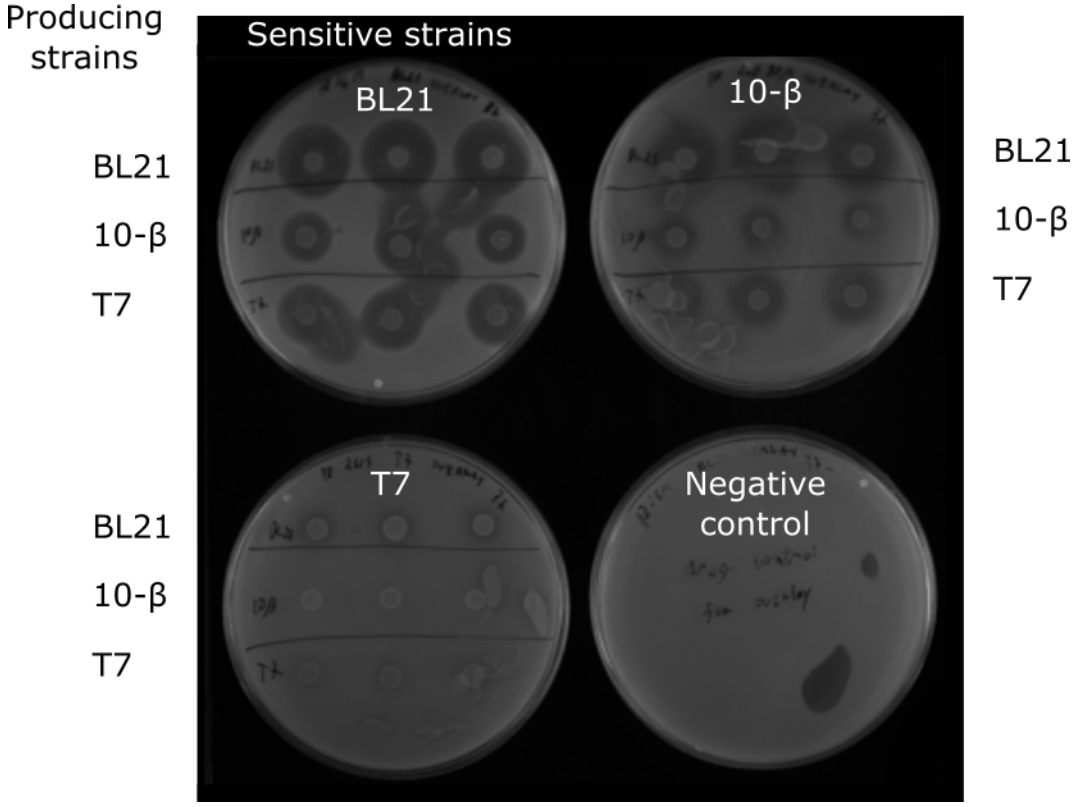


FIGURE 4.19: Use of an agar overlay of sensitive cells to detect microcin B17 production - Producing strains spotted on blank plates, dried and covered with sensitive strains diluted in agar solution. BL21 - BL21(DE3); 10-β - NEB 10-β; T7 - T7 Express lysY/Iq; Negative control - Sensitive strain overlaid on a clean plate. Growth-free zones in the negative control are due to insufficient volume of overlay suspension to cover the entire plate.

While the experimental layout was being optimised, the effect of temperature on the Mcb17 production was measured by repeating the experiment from Fig. 4.18 using strain BL21(DE3) as the sole sensitive strain and incubating plates at 37°C, 30°C, and 25°C. Incubations were carried out for up to 30 h to compensate for slower growth at lower temperatures. Since no large differences were observed in the size of the growth inhibition zones around the the producing spots (Fig. 4.20), a temperature of 37°C was used for subsequent experiments due to faster cell growth at this temperature. The lower temperatures also led to more uneven spreading of the producing spots, probably caused by slower evaporation of liquid medium at lower temperatures.

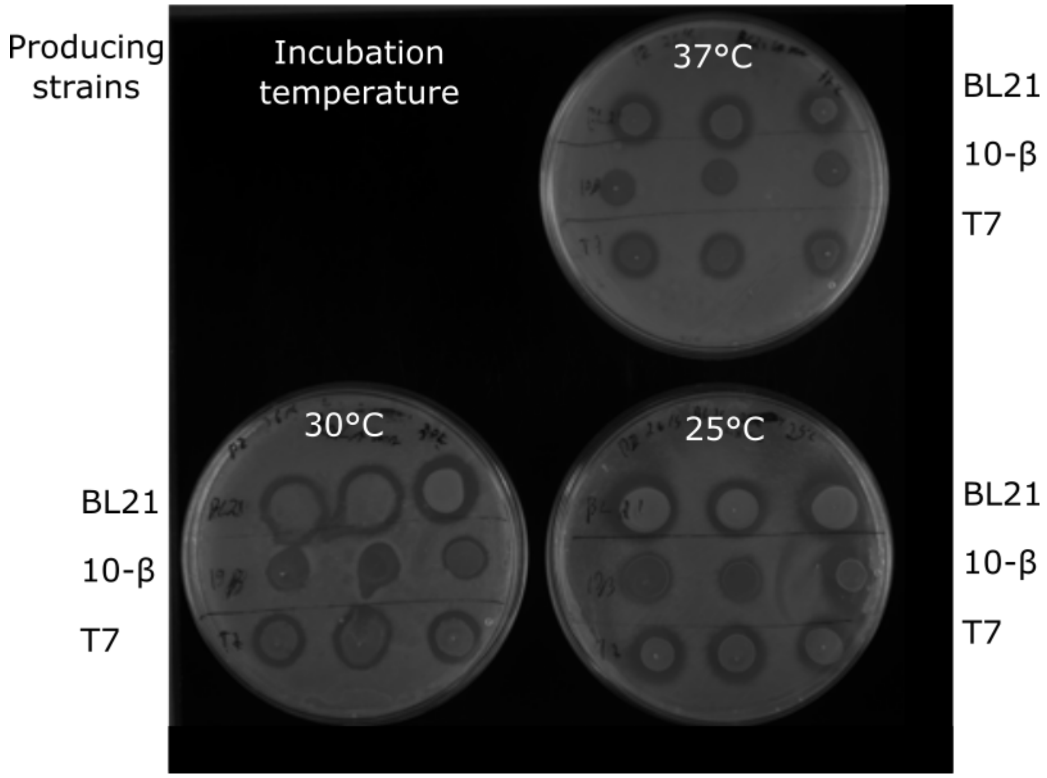


Figure 4.20: Effect of temperature on microcin B17 biosynthesis assays - Producing strains spotted over BL21(DE3) sensitive strain to determine the influence of temperature on the observed phenotype. BL21 - BL21(DE3); 10-β - NEB 10-β; T7 - T7 Express lysY/Iq

To further reduce the experimental noise caused by uneven growth inhibition zones and the formation of producing strain colonies within the growth inhibition zones, the well-diffusion assay was adopted (Balouiri et al., 2016). Briefly, an overlay of the sensitive strain is added to culture plates, wells are created using a cork borer tool, and these are filled with supernatant from overnight cultures of the producing strain (experimental details for the optimised assay conditions are in Section 2.3.10). Mcb17 contained in producing strain supernatants diffuses from the wells into the surrounding medium, creating more uniform growth inhibition zones (Fig. 4.21). The variability between biological replicates was also reduced by normalisation of cell concentrations in overnight cultures before centrifugation to remove supernatant.

Each culture plate contained four wells spaced widely enough to prevent overlap of inhibition zones and the supernatant produced by a strain carrying one McbC mutant was added to each well. To increase data collection throughput, plates were imaged and measured using ImageJ software (Schindelin et al., 2012). Each plate also contained a single blank well filled with only sterile LB medium as a normalisation standard to account for variations in zoom level of imaging equipment or plate positioning (Fig. 4.21). The areas of all inhibition zones were measured and estimated inhibition radii were calculated by approximating to a circular area and subtracting the blank well radius from each measurement. Each measurement was then normalised to the wild-type value as described in Section 2.3.10.

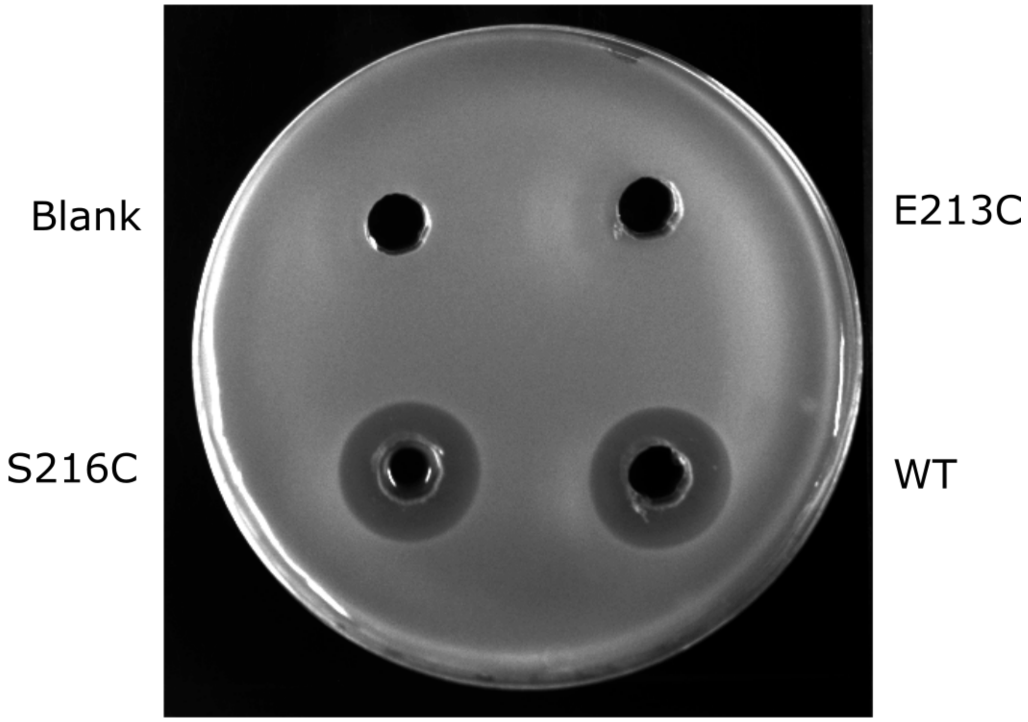


Figure 4.21: Example of a Well-diffusion assay for microcin B17 biosynthesis - Each plate was used to measure microcin B17 biosynthesis by three McbC variants, with a blank well containing LB medium for normalisation of measurements.

4.2.5.2 Activity detection for McbC variants

The optimised conditions established for the Mcb17 bioassay were used to quantitate the activity of McbC Cys mutants and validate functionally-relevant residues selected previously (Section 4.2.4). Assays were carried out with a minimum of four biological replicates per variant, with the wild-type present in every set of experiments as a normalisation control to minimise any interference from uncontrolled changes in growing conditions and sensitive strain overlays between days.

Visual inspection of plates and manual comparison of inhibition radius measurements revealed clear drops in activity in some mutants when compared to the wild-type supernatants (Fig. 4.22 a). However, some variants presented smaller trends towards reduced (i.e. Phe132) and even increased (i.e. Ala198) biosynthetic activity that would require statistical testing to verify whether the observed divergences from phenotype were significant.

The activity measured in the bioassays is the result of variable amounts of Mcb17 contained in the supernatants diffusing from wells into the surrounding medium, stopping the growth of the sensitive strain wherever the local concentration of Mcb17 is high enough to be toxic. Since the shape of the wells, height and volume of the liquid column, and height of the culture medium are constant in all measurements, this leaves the amount of active Mcb17 contained in each well and the size of the zone of growth inhibition as the only two controllable variables. The inhibition zone is created by diffusion of Mcb17 from the supernatant into the effectively two-dimensional surrounding medium, thus the inhibition radius measurement should not be linearly correlated with the concentration of active Mcb17 contained in each well.

To verify the relation between Mcb17 concentration and the size of the growth inhibition zone, dilution series of a wild-type supernatant and two mutants with apparently lower activity (S122 and E213) were assayed and the observed inhibition radii were plotted against

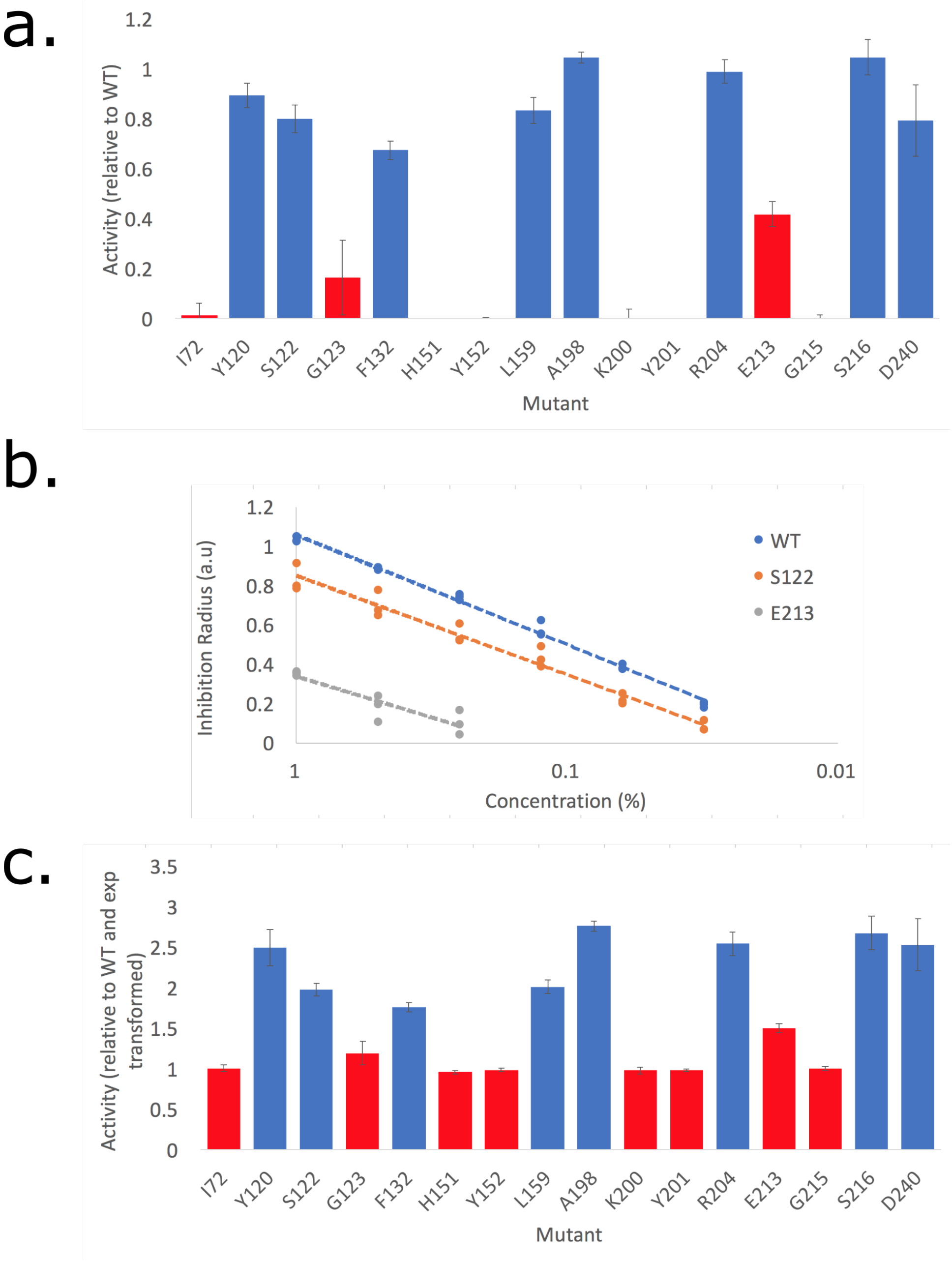


FIGURE 4.22: *In vivo* microcin B17 biosynthesis assays for McbC mutant panel. (a) Mcb17 biosynthesis activity of each McbC mutant relative to the wild-type. (b) Dilution curve of wild-type and two mutants (Ser122 and Glu213) to estimate the relation between observed inhibition radii and relative concentration of Mcb17 in supernatants. (c) Exponential normalisation of data shown in (a). Two-tailed *t*-tests with Bonferroni multiple-testing correction ($p<0.05$ was used as the significance cutoff) were carried out using the data show in (c) and mutants that differed significantly from WT are labelled in red in all panels. Wild-type in (c) had a value of $\exp(1) \approx 2.72$. Error bars represent the standard deviation of the measurements from the mean. $n \geq 4$ for all mutants, in biological replicates

their relative concentrations. As expected, no linear correlation could be observed, but a logarithmic fit was appropriate ($R^2 \geq 0.8$ for all three plots), producing an apparently linear behaviour on a plot with relative concentration on a $\log_{10}$ scale (Fig. 4.22b). Therefore, activity measurements were transformed by the exponential of the inhibition radius as an approximation of the Mcb17 concentration — and, therefore, McbC activity — before significance testing.

All mutants were compared with the wild-type measurements by two-tailed *t*-tests with Bonferroni correction for multiple testing (16 mutants). Six sites had complete activity losses, including the known Lys200-Tyr201 motif (Dunbar and Mitchell, 2013). Of the eight sites that significantly affected McbC function ($p < 0.05$ after correction), six led to complete loss of activity (Ile72, His151, Tyr152, Lys201, Tyr202, and Gly215) and the remaining two (Gly123 and Glu213) were compromised but still able to produce sufficient Mcb17 for activity to be detectable. The results for the validation of the functional predictions are summarised in Table 4.1.

Mutation	**JSD**	**CMI**	**NoSE**	**Bioassay**
I72C			✓	-
Y120C		✓		N
S122C	✓	✓		N
G123C	✓	✓		-
F132C	✓		✓	N
H151C	✓			-
L159C		✓	✓	N
A198C		✓		N
K201C			✓	-
Y202C	✓	✓		-
R205C			✓	N
E213C		✓		-
G215C	✓	✓		-
S216C	✓	✓		N
D240C			✓	N

TABLE 4.1: Effect of mutations on McbC candidate functional residues. "-" indicates reduced activity, "N" indicates no significant effect on activity, and "✓" indicates the residue was among the top 5% of scores for that prediction.

Cysteine mutations at two of the residues predicted by the homology model to contact the FMN cofactor, or form part of the active site region, led to negative impacts upon McbC activity: Gly123 selected by JSD and CMI and Glu213 selected by CMI. However, these mutations did not lead to full activity losses and two other residues selected by CMI within that same region did not significantly reduce activity upon mutation: Tyr120 and Ser122. Interestingly, both the latter mutations were conservative polar-to-polar substitutions of residues that are solvent-exposed in the homology model — especially Ser122Cys, which is only a replacement of a hydroxyl group with a thiol in the side chain. More disruptive mutations, such as introducing a charged or hydrophobic residue could be used to determine whether these positions have a functional role that was maintained when mutated to Cys or the protein was unaffected simply because they are solvent-esposed residues. These results are consistent with the hypothesis that this set of positions has a role in substrate interaction or positioning of active site residues and would, therefore, represent potential candidates for diversification if engineering of McbC for changes in substrate specificity or promiscuity were to be pursued.

The patch of buried aromatic residues detected by JSD led to complete loss of activity when His151 and Tyr152 were mutated to cysteine, suggesting a role in McbC function or maintenance of structural integrity. Attempts at producing the Tyr150Cys mutant failed, but in light of the results obtained for the two subsequent residues in the sequence it would be informative to obtain and measure the activity of this mutant. Due to their aromatic nature and position within the McbC model, it is likely that these residues are involved in maintaining the structure of the protein or in the folding process (Fukusaki et al., 2001), so more conservative mutations to other aromatic or hydrophobic amino acids could assist in elucidating their function. If conservative mutations lead to rescue of activity levels towards values close to wild-type, the activity loss observed for cysteine mutations at positions 151 and 152 could be due to a structural effect.

Only one of the mutations at sites along the putative dimerisation interface in the homology model produced a complete loss of activity: Gly215. However, the cysteine variants at positions 132 and 159 had slight non-significant trends towards activity reduction, suggesting a smaller impact on protein function which would need to be verified by mutations to other residues or a more sensitive assay method. The TOMM dehydrogenases have not been shown to form homodimers in solution, but the position of the FMN cofactors in the homology model between the two subunits and the results obtained here favour that hypothesis. Since the only evidence for the subunit composition of a TOMM synthase complex is the purification of the McbBCD complex in apparent unitary stoichiometry by Li et al. (1996), this information together with the homodimeric structural homologues detected by Phyre2 could indicate a trimer-of-dimers arrangement for the complex (2 McbB : 2 McbC : 2 McbD molecules per complex). Another hypothesis would be that either the docking protein or the cyclodehydratase acts as the binding partner for the dehydrogenase in TOMM synthase complexes. However, the fact that two dehydrogenases from phylogenetically divergent organisms (*B. amyloliquefaciens* and *E. coli*) were successfully expressed in this work with bound FMN in the absence of other TOMM synthase complex partners (Chapter 5) does not support this alternative model. *In vitro* homodimerisation of McbC could be confirmed by methods such as SEC-MALS (Hong et al., 2012) or native mass spectrometry (Heck, 2008).

Finally, the two isolated residues selected from the NoSE scores had opposite effects. A cysteine mutation at Asp240 had no detectable effect on Mcb17 biosynthesis, while the same mutation at Ile72 led to complete loss of activity. The mutation from Asp to Cys at position 240 removed the negative charge and decreased the length of the side chain, yet produced no effect on activity probably due to its position along the solvent-exposed surface of the structure, which tends to be predominantly polar in soluble proteins (Moelbert et al., 2004). Similarly, the Ile72Cys mutation would not be expected to cause a strong effect even though it is a hydrophobic-to-polar change, because this site is in an extended surface loop distant from the putative dimerisation interface in the Phyre2 homology model for McbC, which would be expected to tolerate mutations to polar residues. However, structural modelling of loops is known to be especially unreliable (Eswar et al., 2008), so the predicted location for this residue could be incorrect and the Ile72 side chain could be facing the internal part of the structure and in a close interaction with another hydrophobic residue.

4.3 Conclusions

The *in vivo* assay system established for microcin B17 biosynthesis enabled the quick validation of candidate functional residues obtained from predictions, demonstrating a significant effect on Mcb17 production for half of the mutations tested (eight out of the sixteen successfully constructed mutants). While only negative effects have been detected, it is not clear whether increases in the activity of McbC would necessarily lead to increased export of active Mcb17, which is the phenotype detected by the assay. Mutants with increased activity are only expected to produce larger zones of inhibition if the azoline oxidation step is limiting during *in vivo* production of Mcb17, but the relative velocities of each pathway step (McbA translation, heterocyclisation, and export) are unknown and likely to vary between strains and cultivation conditions. The relative velocities of each step could be intentionally changed by altering relative expression levels — under- or overexpressing individual pathway components by expression from separate higher copy number plasmids or engineering of ribosomal binding sites (Reis and Salis, 2017). If the pathway could be engineered to have oxidation by McbC as its rate-limiting step, it could be used as a bioassay capable of reliably detecting variants with higher activity than the wild-type, with the Ala198Cys mutant as a candidate and other rationally selected mutations at sites found in this work.

Another caveat of the results produced by *in vivo* biosynthesis assays is that several different phenotypes could be responsible for the differences in activity observed for the mutant panel, including phenotypes not directly relevant to the functional mechanisms of TOMM dehydrogenases. A reduction in Mcb17 synthesis in a strain carrying a variant of McbC could be explained by the targeted residue playing a role in processes such as catalytic activity, FMN cofactor binding, substrate recognition, interaction with binding partners in the TOMM synthase complex, enzyme folding, and thermostability. Therefore, additional characterisation of isolated McbC would be needed to determine which mutations are directly related to McbC catalysis and specificity. One possibility would be to overexpress and purify McbC, followed by characterisation with a simple assay such as Thermofluor (Ericsson et al., 2006) to determine whether mutations caused negative impacts on protein folding and stability, effectively representing false positive functional residues. Constructs have been prepared to overexpress the mutant panel following the optimised methods from Chapter 5, but experiments were not carried out due to time constraints.

The fact that at least half of the predicted functional residues had some impact on McbC function in the biosynthesis of Mcb17 demonstrates the power of coupling computationally-simple information theory-based prediction strategies to robust bioassays in producing data for characterisation of poorly-annotated protein families. The mutants validated by this strategy can be further characterised to provide novel insight into the mechanism of TOMM dehydrogenase activity and overall TOMM synthesis, avoiding laborious untargeted screening of variant repertoires to identify relevant residues. However, this combination can only be applied if a simple bioassay can be established for the phenotype of interest, which limits the range of biological systems that could be targeted to the ones that produce an easily-detectable phenotype such as growth, gene expression or a marker metabolite. Recent developments in directed evolution strategies to isolate biosensors for small molecules represent an alternative path towards generalisable assays and could greatly expand the range of possible molecules that could be targeted by a combination of computational predictions and bioassays (Feng et al., 2015b; Skjoedt et al., 2016).

Chapter 5

In vitro exploration of TOMM synthase complexes

Previous *in vitro* reconstitutions of TOMM synthase complexes were carried out by heterologous expression of the complex proteins — frequently as fusions to Maltose Binding Protein (MBP) to increase their solubility (Li et al., 1996; Melby et al., 2012), followed by activity assays with incubation of purified complexes and substrate peptides. The most commonly used method for detection of synthase activity is Mass Spectrometry (MS), due to the loss of 20 Da associated with the full processing — heterocyclisation followed by oxidation to azole — of each heterocycle in the substrate peptide. However, characterisation of a protein employing MS as the sole detection method in assays can become costly and labour-intensive, especially if variants of the protein are made to study the function of residues in its sequence. Alternative detection strategies employed previously include polyclonal antibodies specific for Microcin B17 (Li et al., 1996) and detection of phosphate released by ATP hydrolysis during the heterocyclisation step (Melby et al., 2012).

The aim of this section of the work was to establish a robust *in vitro* detection system for the formation of heterocycles by TOMM synthase complexes, employing MS as the main activity measurement strategy but also attempting to develop simpler methods to assay complex function. The polyclonal antibodies used by Li et al. (1996) are no longer available, but other commercially available antibodies could be employed in assays such as ELISA to quantify TOMM production. Additionally, development of a negative assay based on labeling of unreacted cysteine residues — molecules with all cysteines fully converted to thiazoles would not be labeled — was attempted, using fluorescent dyes coupled to thiol-specific reagents.

5.1 Results and Discussion

5.1.1 *In vitro* reconstitution of the *Bacillus amyloliquefaciens* plantazolicin synthase complex

This section will describe experiments carried out with the aim of heterologously producing the *B. amyloliquefaciens* TOMM synthase complex in *E. coli* and measuring its activity *in vitro*.

5.1.1.1 Expression and purification of *Bacillus amyloliquefaciens* plantazolicin synthase complex and substrates

Genes coding for the three enzymes of the TOMM synthase complex for plantazolicin biosynthesis described by Scholz et al. (2011) were synthesised (GeneWiz Inc.) and cloned by restriction-based cloning into fusion-free (pBAD30) and MBP-tagged (Fig. 5.1, in pMAL-C2x, New England Biolabs) constructs for heterologous expression in *E. coli.* Constructs were transformed into the protease-deficient *E. coli* K-12-derived strain ER2508 (NEB) for expression. After optimisation of expression conditions, soluble expression was obtained for all three enzymes fused to MBP, but not for the fusion free constructs (data not shown).

These proteins were partially purified by MBP-affinity chromatography on MBTrap columns (GE Healthcare), reaching a purity of at least 80% for all constructs (Fig. 5.2a-c), which was deemed sufficient to carry out assay development. A yellow colour was clearly visible in eluted fractions from BamB purification, indicating that the FMN cofactor needed for activity was bound. Following previous literature reports (Li et al., 1996; Melby et al.,

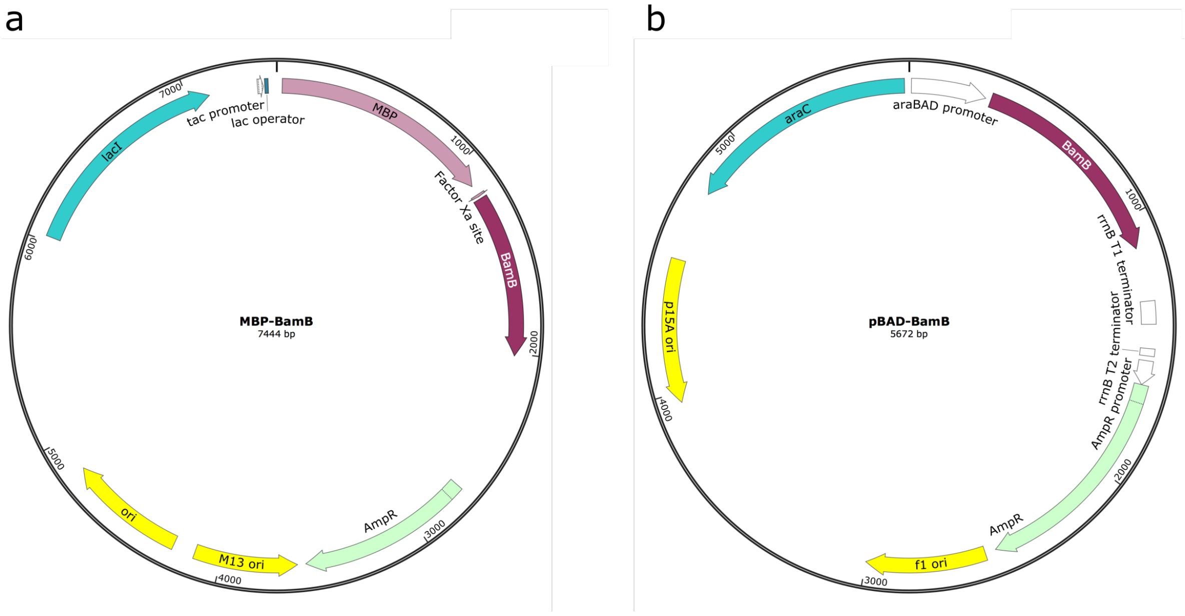


Figure 5.1: Constructs for expression and purification of *B. amyloliquefaciens* BamB, BamC and BamD. (a) pMAL-C2x-derived vector for the expression of BamB with an N-terminal MBP fusion. (b) pBAD30-derived vector for the expression of fusion-free BamB. These constructs were also made for expression of the BamC and BamD proteins

2012) that showed full activity only for fusion-free proteins, cleavage with Factor Xa was used to produce fusion-free enzymes. However, these cleavage reactions did not reach completion, even after optimisation attempts (Fig. 5.2d). Since the MBP fusion is known to interfere with TOMM synthase function (Milne et al., 1999; Melby et al., 2012), alternative constructs were designed to obtain higher cleavage efficiencies by avoiding the use of Factor Xa protease.

In order to obtain fusion-free enzyme preparations with a higher purity, a set of new constructs was designed and tested for the BamB dehydrogenase. These included three different self-cleaving intein-Chitin Binding Domain (CBD) fusions (*Mxe*, *Ssp* and *Sce* intein-CBD tags in pTWIN1 and pTYB21 vectors, New England Biolabs) as well as a *Ssp* intein-MBP fusion (pMAL-C2x) and a fusion-free protein under the control of the tightly regulated Rhamnose promoter (pD881 vector, DNA 2.0). All constructs were transformed into the protease-deficient strain ER2566, derived from BL21(DE3), for transcription from the T7 promoter. The C-terminally-tagged construct BamB-Mxe led to the highest titres of soluble BamB protein (Fig. 5.3), so constructs were made to implement this expression strategy

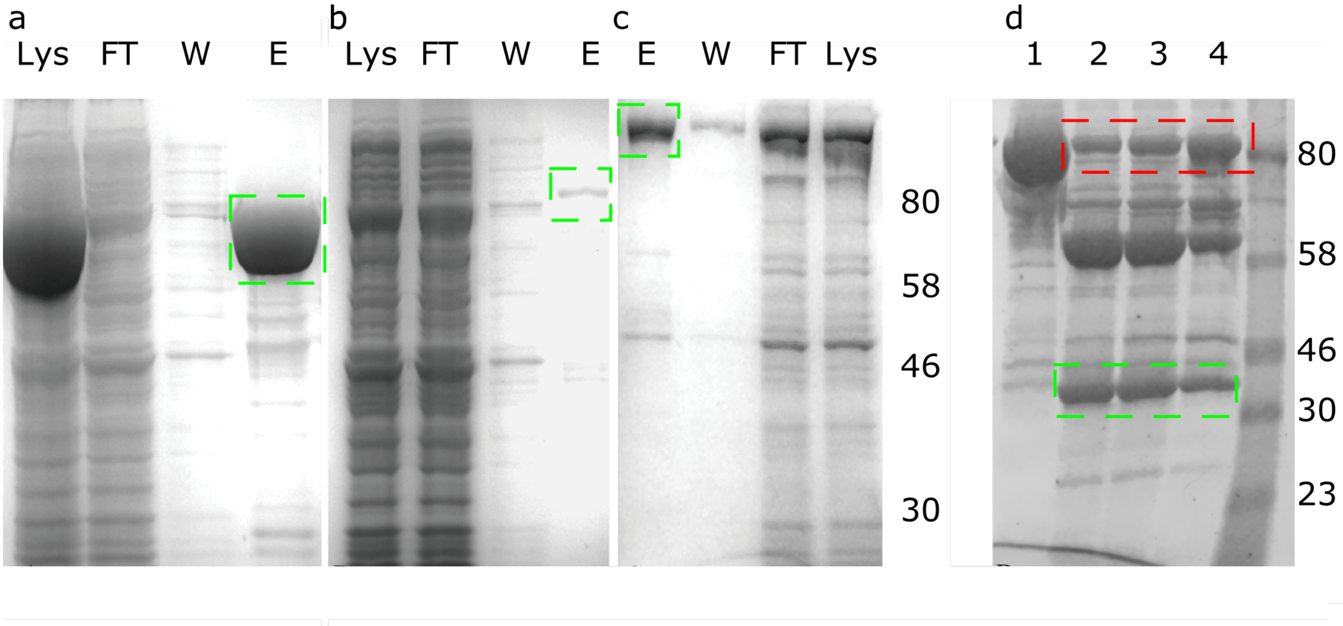


Figure 5.2: MBP-tagged enzyme expression and purification. a-c) Fractions eluted from amylose affinity chromatography column. In all images, there is significant enrichment of the band corresponding to the target MBP-enzyme fusion. LYS - Lysate injected into column. FT - Flow-through from injection. W - Material washed from the column before elution. E - equate recovered after addition of Maltose. Numbers indicate mass of ladder fragments in kDa. a) BamB purification. b) BamC purification. c) BamD purification. d) Factor Xa-mediated cleavage of the MBP tag in BamB protein, in all conditions there was still detectable uncleaved MBP-BamB. Lane 1 - no Factor Xa. Lanes 2-4 - varying incubation times with Factor Xa, according to manufacturer's instructions. Expected masses for MBP-BamB = 72.7kDa, MBP-BamC = 80.3kDa MBP-BamD = 92.8kDa, free MBP = 42.5kDa and free BamB = 30.2kDa. Bands enclosed in green boxes correspond to desired purified or fusion-free protein. Bands enclosed in red boxes are fusions to MBP.

for the other two proteins in the plantazolicin heterocyclase complex. However, no soluble expression could be obtained for BamC-Mxe and BamD-Mxe (data not shown), so assay development continued with the pMAL-based N-terminal MBP fusions with Factor Xa protease cleavage sites, using concentration by ultrafiltration to obtain enzyme stocks for activity assays.

The native plantazolicin substrate peptide has 10 heterocyclisation sites which can lead to complex mixtures of modification products in case of incomplete heterocyclisation. Therefore, shorter sequences were designed to act as positive and negative controls for enzyme activity. Synthetic core region constructs were designed with the sequence XDGGK, where X can be a non-cyclisable tyrosine (0C) or a number of consecutive cysteine residues (1C, 2C, 3C). These core regions are preceded by the plantazolicin substrate peptide leader

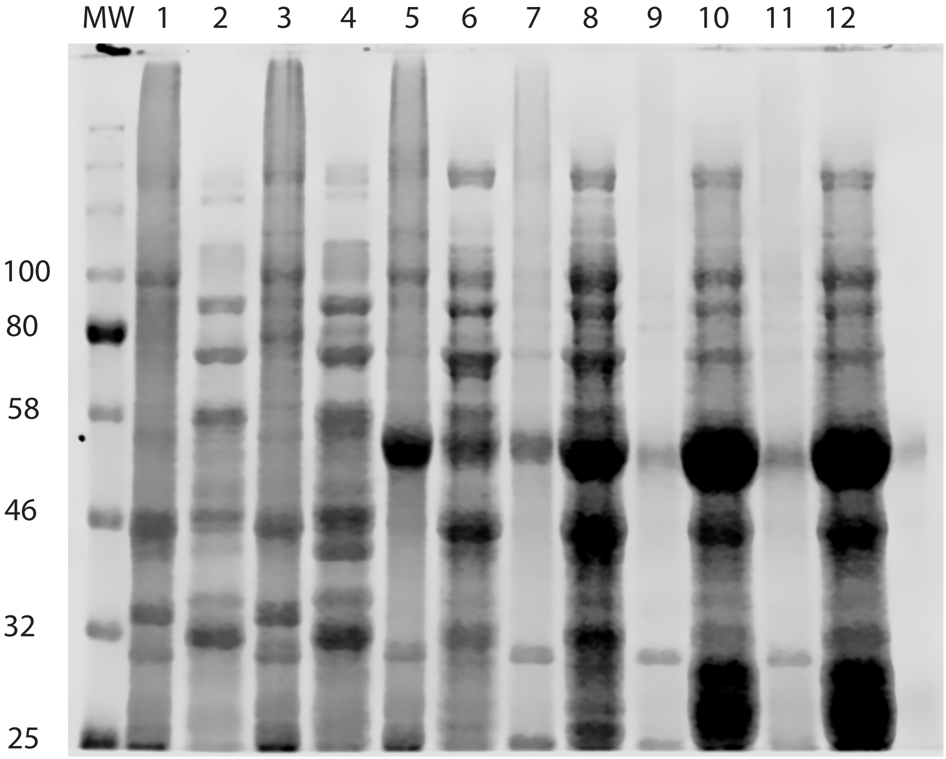


FIGURE 5.3: Solubility tests for alternative BamB constructs. MW - Molecular Weight Marker. Every pair of subsequent lanes contains, respectively, the insoluble and soluble fractions for one expression condition. Lanes 1-2 - fusion-free BamB under the control of the arabinose promoter in pD881. Lanes 3-4 - MBP-Ssp intein-BamB fusion. Lanes 5-6 - BamB-Sce intein fusion induced with 1mM IPTG. Lanes 7-8 - BamB-Sce intein fusion induced with 0.1mM IPTG. Lanes 9-10 BamB-Mxe intein fusion induced with 1mM IPTG. Lanes 11-12 BamB-Mxe intein fusion induced with 0.1mM IPTG. Expected Molecular weights: BamB = 30.2kDa, MBP-Ssp-BamB = 97kDa, BamB-Sce = 85.2.kDa, BamB-Mxe = 58.1kDa

sequence (Scholz et al., 2011). In addition to the sequences required for TOMM synthase complex recognition and activity, fusion tags were added to assist in assay development and purification. A streptavidin-binding tag (Strep-tag, Wilson et al. (2001)) was fused immediately to the C-terminal end of the artificial core sequence. This tag was used for the immobilisation of peptides on streptavidin-coated microplates, which would be required for the washing steps in the ELISA-based assays. Finally, the intein-CDB fusion tag from pTWIN1 was at the C-terminus of the fusion peptide and was used for purification on chitin resin (NEB) columns (Fig. 5.4). These constructs were named BXSIC (Bacillus leader - X cysteines - Strep tag - Intein - CBD).

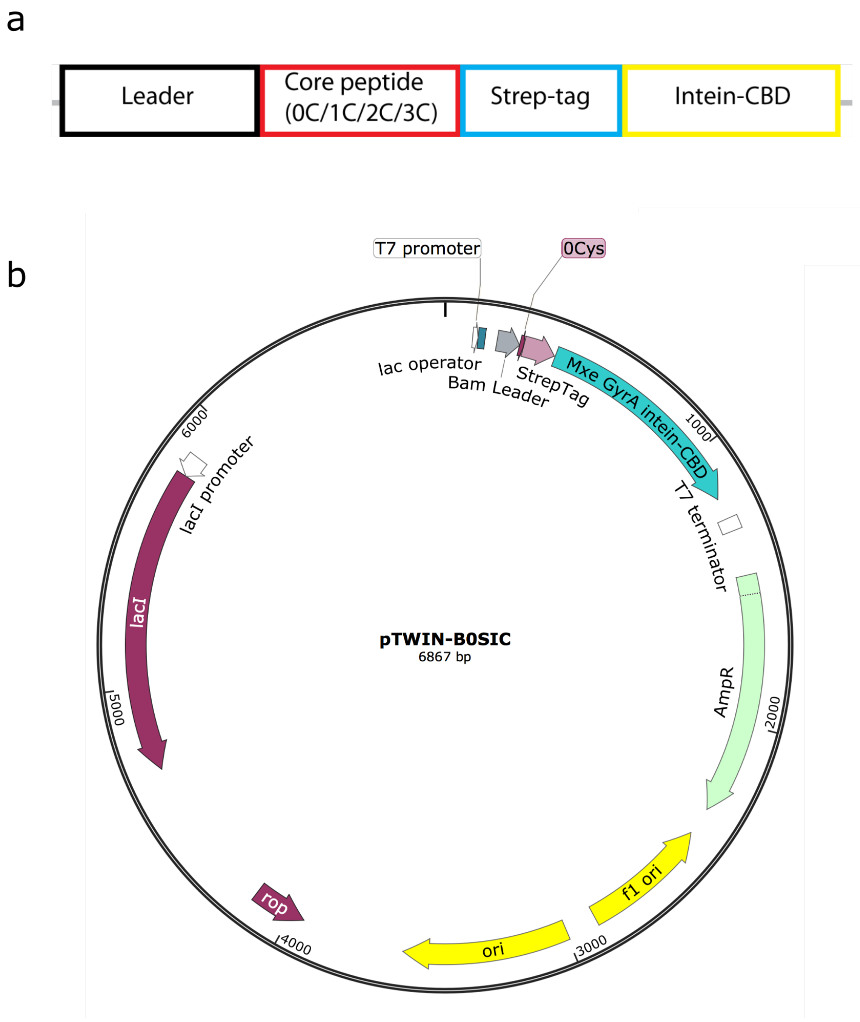


FIGURE 5.4: Synthetic *B. amyloliquefaciens* TOMM synthase substrate construct. (a) Design of the fusion protein, containing the native leader sequence, a synthetic core peptide, a streptavidin binding tag for immobilisation and an Intein-CBD tag for purification. (b) pTWIN1-derived vector for the expression of the synthetic substrate constructs. The 0C negative control substrate is shown here, but the remaining variants were made following the same design.

The 0C and 1C substrate constructs were expressed and purified with high yields by in-column cleavage, following manufacturer's instructions (Fig. 5.5 for 0C, data not shown for 1C). Significant amounts of contaminating uncleaved peptides could be detected in the eluate. Therefore, the eluate from the first run was reloaded onto the re-equilibrated column to capture the material still containing the intein CBD fusion and the flow-through fraction of this second injection was used for assay development. After the second purification step, the peptides were concentrated and buffer-exchanged using Amicon 3 kDa MW cutoff filters (Millipore).

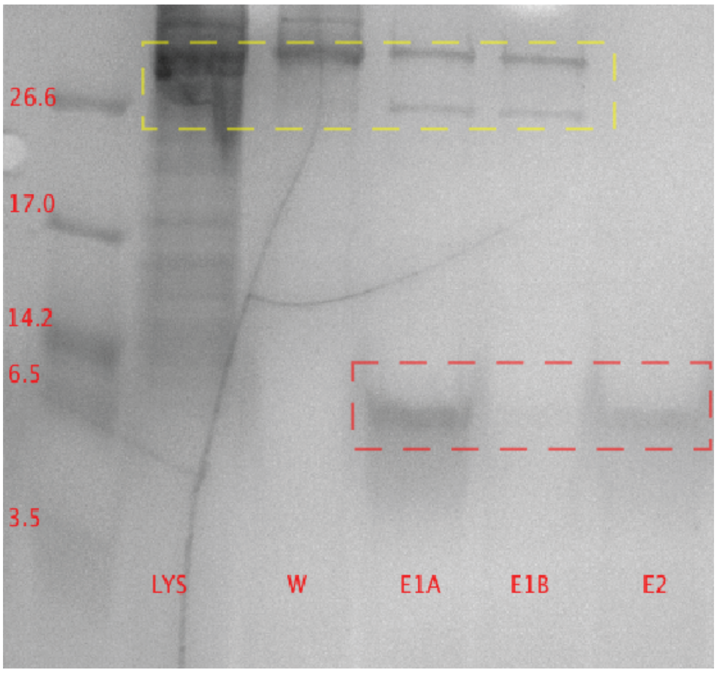


FIGURE 5.5: Single step capture and cleavage of Intein-CBD-tagged 0C substrate peptide by chitin column chromatography. LYS - lysate injected into column. W - Material washed from the column before cleavage. E1A and E1B - Eluate fractions after cleavage. E2 - Protein recovered after reloading E1A onto a reequilibrated column, showing a single band near the 6.5kDa marker band. The bands corresponding to free substrate peptides are enclosed in a red dashed box and the bands corresponding to Intein-CBD-substrate fusions and free Intein-CBD are enclosed in a yellow dashed box. Expected mass for the cleaved peptide is 6.7 kDa

5.1.1.2 Development of *in vitro* assays for the *Bacillus amyloliquefaciens* plantazolicin synthase complex

Fusion-free substrate 0C (B0S, representing B0SIC after removal of the Intein-CBD tag) peptides were analysed by MALDI-TOF (Matrix-Assisted Laser Desorption Ionization Time of Flight) MS and single peaks were detected with mass/charge ratios almost 1000 units smaller than the expected value for a singly-charged peptide (Fig 5.6). Initial cysteine labelling tests were also conducted on the 0C and the 1C peptides, with the 0C peptide acting as a negative control. These experiments resulted in strong background labelling, which could interfere with future quantitation attempts. However, there was differential labelling of both peptides, with the stronger signal coming from the 1C peptide as expected (Fig. 5.7).

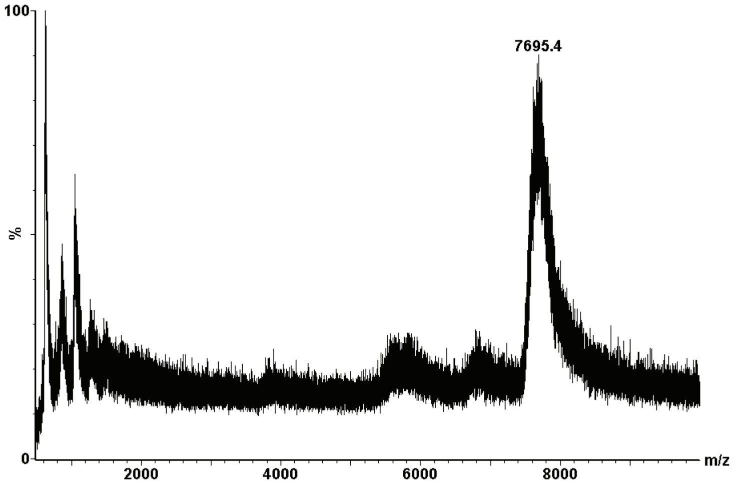


Figure 5.6: MALDI-TOF mass spectrum of purified 0C peptide, obtained in negative linear mode. Expected mass for the peptide is 6731 Da

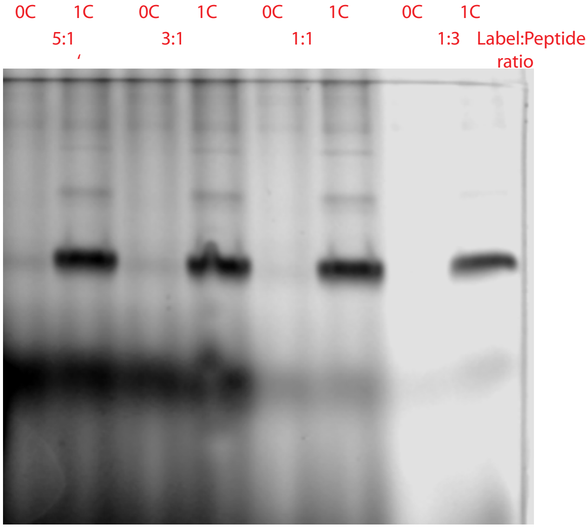


Figure 5.7: Flourescence image of gel electrophoresis after Iodoacetamide-fluorescein labelling reactions in different ratios of 0C/1C peptides to fluorescent label. As the Label:Peptide ratio decreased, there was a reduction in background signal while the specific 1C band was still strongly labeled.

Once the peptide mass was confirmed by MS, TOMM synthase activity assays were attempted by incubating Factor Xa-cleaved BamB/BamC/BamD with 0C and 1C peptides, following conditions established for the *Bacillus sp.* Al-Hakam synthase complex (Melby et al., 2012). After 72 hours, the reactions mixtures were loaded into Tricine SDS-PAGE gels to detect any changes in migration of the peptides incubated with the enzymes. The expectation was that heterocycles installed in the peptide backbone would restrict conformational flexibility during electrophoretic migration, leading to altered migration of modified peptides compared to unmodified controls.

Unexpectedly, both the 0C and 1C peptide bands shifted to a lower apparent molecular weight when incubated with the enzymes, which could be evidence of activity of the TOMM synthase even on the substrate that did not contain cysteines for modification (Fig. 5.8). However, both peptides contain two threonine residues in the C-terminally fused Strep tag, which could explain how the 0C negative control underwent modification by the enzyme complex. Other forms of evidence, such as a shift in MS peaks or changes in iodoacetamide labelling, would still be needed to confirm this result. Additionally, the shifts could be caused by changes in buffer conditions and protein loading, which were not controlled between positive and negative lanes.

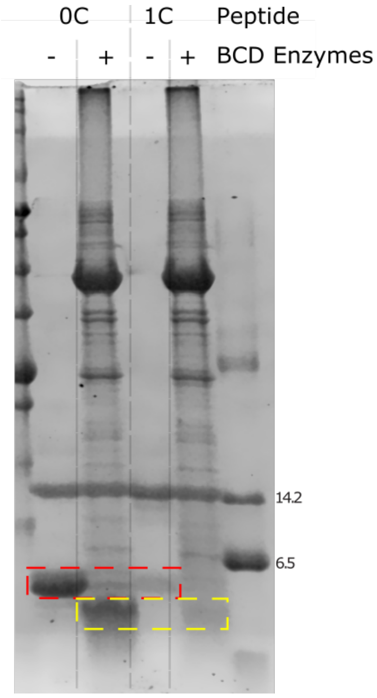


Figure 5.8: TOMM Synthase activity assay electrophoretically separated on a polyacrylamide gel. When incubated for 72 hours in the presence of Factor Xa-cleaved BamB, BamC and BamD, a mobility shift was observed for both 0C and 1C bands. Both peptides had higher mobility after incubation with the BCD complex. "-" indicates the absence of BCD enzymes in a reaction and "+" indicates the presence of BCD enzymes in a reaction. The peptide bands prior to incubation with BCD enzymes are enclosed in a red dashed box. The peptide bands observed after incubation with BCD enzymes are enclosed in a yellow dashed box. Expected mass for the 0C and 1C peptides is 6.7 Da

The unclear result from the initial assay developments tests led to a verification of all construct designs to ensure the correct proteins were being used at all points. However, a truncation was detected in the leader sequence inserted in the synthetic substrate peptides, when compared to leader sequences used in previous studies (Scholz et al., 2011). This error occurred due to a misidentification of the start codon in the BamA substrate peptide gene from the *B. amyloliquefaciens* genomic sequences and required a redesign of the substrate constructs (Fig. 5.9). In addition to the 0C,1C,2C and 3C peptides previously mentioned, a peptide with the plantazolicin core region — RCTCTTIISSSSTF containing 10 heterocyclisation sites and named BPSIC — was also constructed as a positive control for *B. amyloliquefaciens* complex activity. The peptides BPSIC, B0SIC and B2SIC were the first to be successfully cloned with the new leader sequence, so these were used for assay development while the remaining variants were being constructed.

Truncated	MTQIKVPTALIASVHGEGQHLFEPMAA
Scholz *et al.* 2011	MEEVTIMTQIKVPTALIASVHGEGQHLFEPMAA

FIGURE 5.9: Sequences of the truncated and corrected plantazolicin leader peptide, from Scholz et al., 2011.

All three substrates were expressed in high soluble yields and purified by chitin affinity chromatography followed by in-column cleavage of the intein-CBD tag (Fig. 5.10a, data shown for BPS only). Free BPS was isolated by a two-step ultrafiltration protocol, removing most of uncleaved BPSIC by filtering the eluted fraction through a 30 kDa cutoff filter. This was followed by removal of remaining higher molecular weight proteins by filtration through a 3 kDa cutoff filter that did not retain the free peptide of expected mass equal to 9.4 kDa, which was interpreted as normal variation in the mass cutoff of the filter device. The migration of the BPS peptide in Tricine SDS-PAGE was faster than expected for its mass, but this was interpreted as anomalous migration known to occur in Tricine SDS-PAGE, especially with peptides containing high frequencies of hydrophobic amino acids

(Schagger, 2006) — of which there are 21 out of 33 residues in the plantazolicin leader peptide. To verify whether the two independent indications of lower molecular weight than expected for the BPS peptide were correct, the purified peptides were submitted to MALDI-TOF mass spectrometry, which identified a clear peak at 5543.3 Da and supporting the anomalous migration and unexpected behaviour in filtration to indicate a likely truncation in the purified peptide (Fig. 5.10 c).

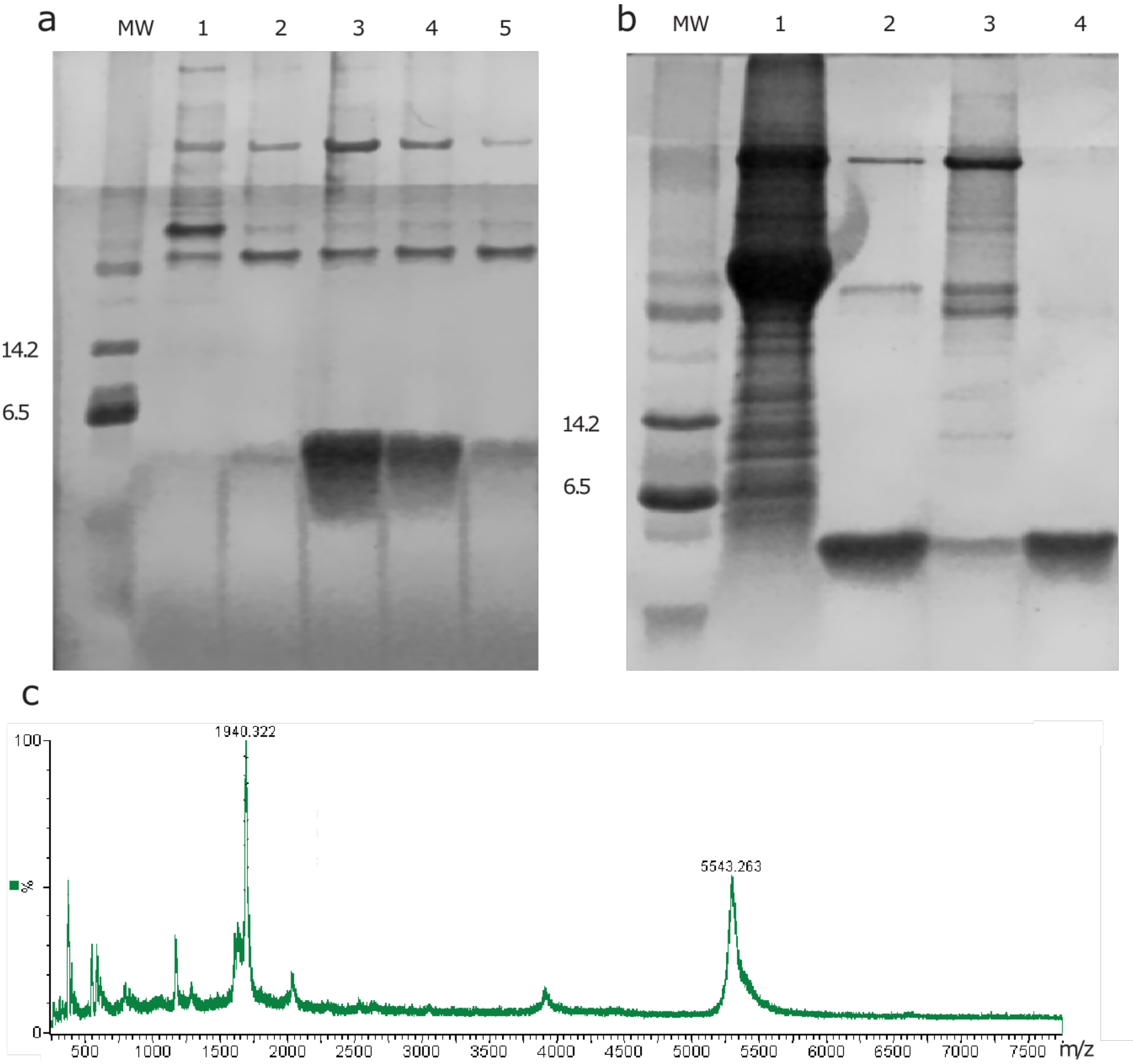


Figure 5.10: BPSIC substrate peptide purification and MALDI-TOF detection. A - Eluted fractions from the chitin resin column. MW - Molecular weight marker. Lanes 1-5 Eluted fractions, with the highest concentration of cleaved peptide in lanes 3 and 4. Expected MW for BPSIC = 8.6kDa. B - Removal of uncleaved fusion constructs by ultrafiltration. Lane 1 - Retentate in 30kDa cutoff filter. Lane 2 - Filtrate collected from 30kDa cutoff filter. Lane 3 - Retentate in 3kDa cutoff filter. Lane 4 - Filtrate collected from 3kDa cutoff filter and subsequently used as pure BPSIC peptide. C - MALDI-TOF spectrum of purified BPSIC. Expected mass for the peptide is 9428 Da

To verify whether this truncation was caused by the changes in the leader peptide portion of the construct, B0S and B2S were isolated by the same procedure as BPS and analysed by MALDI-TOF, along with TOMM synthase reactions using each substrate and the full BCD complex (Fig. 5.11). As observed for the BPS substrate, the expected peak for each free substrate peptide could not be observed and other unexpected peaks were also seen in the lower regions of the spectra. Incubation with BCD enzymes for heterocyclisation produced new peaks, but none corresponding to expected shifts of -20 Da or -40 Da for the two-cysteine B2S peptide and no shift for the negative control B0S peptide (Fig. 5.11)

TOMM synthase activity assays were also attempted with the purified BPSIC peptide and a similar electrophoretic mobility shift to the one observed in Fig. 5.8 could be observed after BPSIC was incubated with Factor Xa-cleaved BamB, BamC and BamD (Fig. 5.12a). The reaction mixtures were then analysed by MS and a shifted peak was detected (Fig. 5.12c), but this peak did not match the peak observed after BPSIC purification (Fig. 5.10c). Also, the observed shift was of approximately 332 m/z units, while the expected shift from heterocyclisation of all 10 residues in the core peptide would be only 200Da. Even if the two threonine residues in the Strep tag were also heterocyclised to methyloxazoles, the expected shift would still only be 240Da, which indicates that the observed shifts in MALDI-TOF spectra were not due to TOMM synthase activity.

The electrophoretic mobility shift after incubating substrate peptides with the TOMM synthase complex was reproducible with all of the substrate peptides constructs used until this point, initially interpreted as evidence of an active enzyme complex. However, since the MS results were inconclusive and the direct detection methods (chemical labelling of free cysteines or antibody detection of thiazoles) were not attempted for reaction products of this complex, there still remained a possibility that the apparent mass change was a technical artefact and not related to TOMM synthase activity.

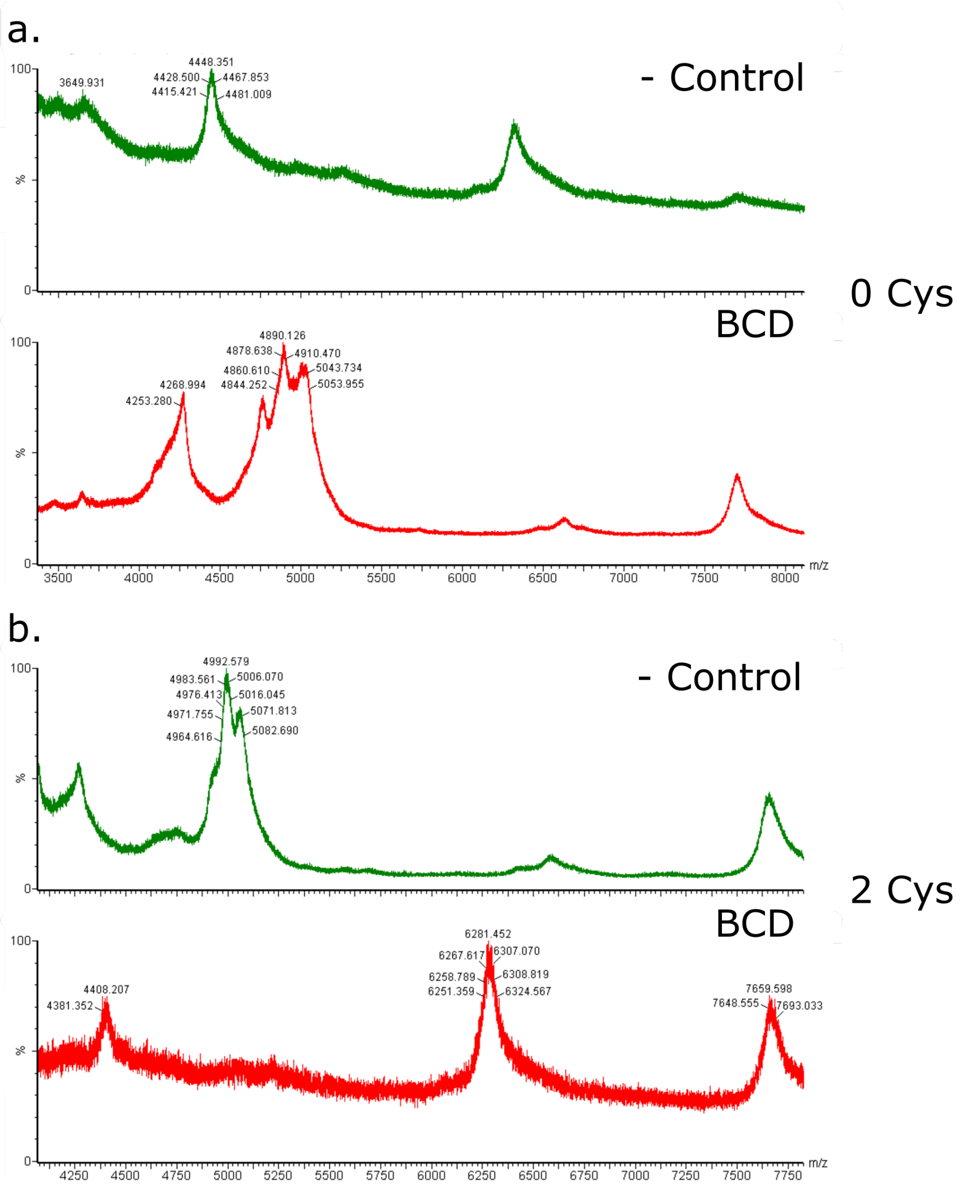


Figure 5.11: MALDI-TOF spectra of a TOMM synthase activity assay with corrected leader peptide sequences. Spectra in green represent negative controls incubated in the absence of TOMM synthase complexes and spectra in red were incubated in the presence of purified and Factor Xa-cleaved BamB, BamC and BamD. (a) Spectra obtained for the B0S non-cyclisable peptide. (b) Spectra obtained for the B2S substrate peptide containing two cyclisable cysteines. Plots were zoomed into the region of the spectra that contained peaks, no peaks were observed outside this region. B0S expected MW = 8454.2 Da. B2S expected MW = 8497.1 Da. Each cyclisation event was expected to produce a loss of 20 Da, to a total of 40 Da if both Cys residues in B2S were cyclised.

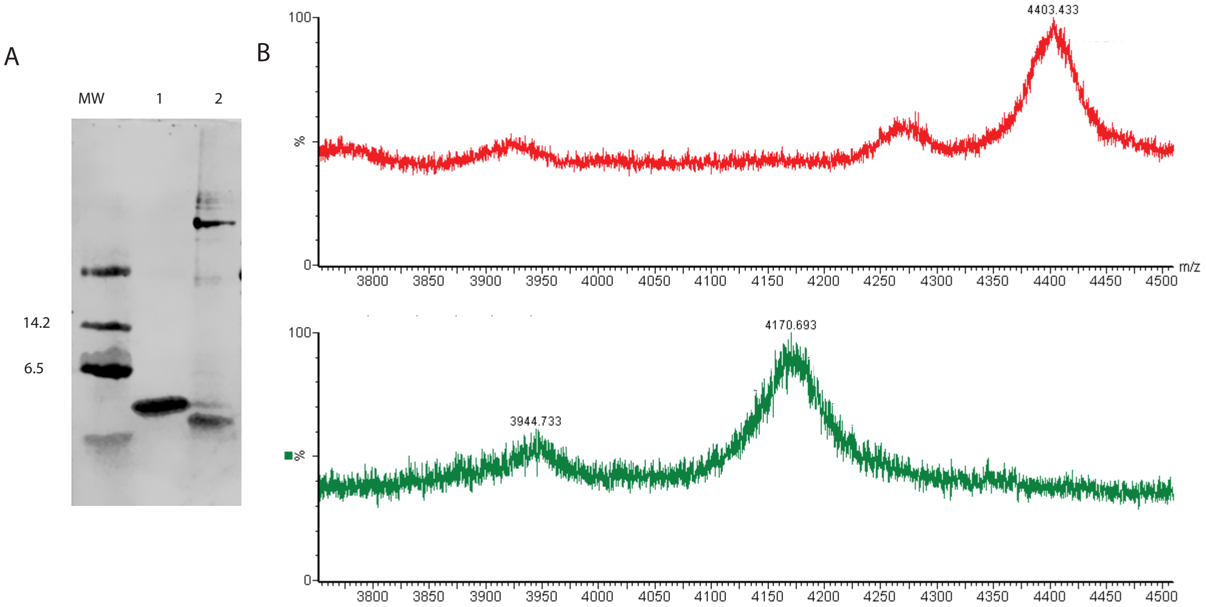


FIGURE 5.12: Synthase assay using BPSIC substrate. A - Gel shift assay of the reaction products. MW - molecular weight marker. 1 - BPSIC peptide incubated in the absence of enzymes. 2 - BPSIC peptide incubated in the presence of Factor Xa-cleaved BCD complex. B - MALDI analysis of the reaction products from A. The top spectrum represents the unreacted peptide and the bottom spectrum is the peptide fraction after incubation with the active enzymes. Plots were zoomed into the region of the spectra that contained peaks, no peaks were observed outside this region. Expected MW for BPSIC = 9428 Da. Each cyclisation event was expected to produce a loss of 20 Da, to a total of 200 Da if all the cyclisation sites in BPS were cyclised.

One possible confounding factor would be a contaminant protease in the synthase enzyme preparations used in the assays. To eliminate this possible source of error, a synthase reaction was set up with uncleaved MBP-fused BamB, BamC and BamD, which are likely to be inactive (as described by Milne et al. (1999); Melby et al. (2012)). After a 72-hour incubation, no shift was observed in the MBP fusion lane, while the Factor Xa-cleaved lane had the same shift that was observed in previous experiments (Fig. 5.13), further indicating that the observed shift is due to TOMM synthase activity. Cleavage of the substrate peptides by Factor Xa itself is unlikely as none of the substrate constructs have Factor Xa sites and the assays are conducted in 5mM DTT, which would reduce the disulfide bonds needed to maintain the Factor Xa subunits in their active form (Waugh, 2011).

Despite the indication of activity from the shift in electrophoretic mobility, no clear evidence of heterocyclisation activity could be obtained for the *B. amyloliquefaciens* plantazolicin synthase system. Commercially-synthesised peptides could be used to eliminate the possibility of the anomalous masses being caused by intracellular degradation during peptide expression in *E. coli.* Also, synthase reactions using the BamBCD complex could be analysed with tandem MS/MS methods, fragmenting the peptides to confirm their sequence and also accurately detect locations of any heterocyclisation event that occurred, as done by Melby et al. (2012) in the initial description of this enzyme complex. These alternative approaches to confirm this evidence of activity were not pursued due to other sets of TOMM synthase complexes and substrates becoming available for use, as described in the following section.

5.1.2 *In vitro* reconstitution of the *Bacillus sp.* Al-Hakam TOMM synthase

While efforts were being made to obtain a robust and consistent assay system for the *B. amyloliquefaciens* TOMM synthase complex, contact was established with the group

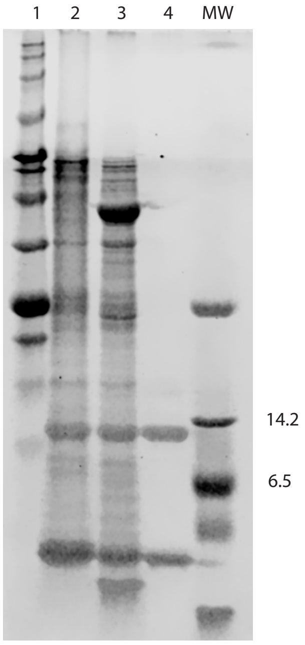


Figure 5.13: Control experiment for protease contamination in BamB, BamC and BamD preparations. Lane 1 - NEB Broad Range marker. Lane 2 - 0C peptide incubated with uncleaved TOMM synthase enzymes. Lane 3 - 0C peptide incubated with Factor Xa-cleaved enzymes. Lane 4 - 0C peptide incubated with no enzymes. MW - Low Molecular weight marker

of Prof. Douglas A. Mitchell (University of Illinois at Urbana-Champaign) for assistance in method development. Prof. Mitchell advised the use of the *Bacillus sp.* Al-Hakam (Balh) TOMM synthase system, due to its greater tolerance to mutations in the core peptide (Melby et al., 2012; Deane et al., 2016) and higher *in vitro* activity (personal communication). Expression constructs for all the proteins in the Balh system were kindly donated, along with constructs for the *E. coli* microcin B17 system as a second backup strategy.

5.1.2.1 Expression and purification of the *Bacillus sp.* Al-Hakam TOMM synthase complex and substrates

All of the constructs for expression of the Balh TOMM synthase system contain TEV protease-cleavable N-terminal MBP fusions under transcriptional control of the T7 promoter, in pET-based vectors (Fig. 5.14). Among these were two substrate constructs:

BalhA1-WT coding for the wild-type TOMM precursor and BalhA1-NC coding for a non-cyclisable negative control variant, in which all Cys, Ser, and Thr residues after the leader sequence were replaced with Ala (Fig. 5.14a-b). The cyclodehydratase BalhD and the docking protein BalhC are encoded by the native sequence from *Bacillus sp.* Al-Hakam (Fig. 5.14c-d). However, the dehydrogenase used in Melby et al. (2012) and subsequent work in the Mitchell Group (Melby et al., 2014; Dunbar et al., 2014) is the cognate BcerB from *Bacillus cereus*, since the FMN cofactor required for activity is not loaded when the native BalhB is expressed in *E. coli*. All of the vectors were transformed into *E. coli* NEB 10-β for amplification, plasmids were extracted and confirmed by sequencing, followed by transformation into T7 Express LysY/Iq for expression.

The first expression tests in 10 ml culture volumes were carried out using the conditions previously established for the Bam system (Section 5.1.1.1) and also the conditions used by the Mitchell group: induction of the enzyme constructs with 0.4 mM IPTG at 22°C for 16 h and induction of the substrate constructs with 1 mM IPTG at 22°C for 1 h (Fig. 5.15). All of the proteins had higher soluble yields under the conditions established in this work for the Bam system, so these conditions were scaled up to 200 mL cultures in shake flasks for purification.

All cultures were lysed by sonication and purified by amylose-affinity chromatography (Fig. 5.16) according to the methods in Section 2.4.5. Good yields were obtained for BcerB and BalhC, but BalhD recovery was approximately 30-fold lower (Fig. 5.16a), reflecting the lower soluble expression yield (Fig. 5.15). Multiple cultures of BalhD were carried out and eluted fractions were pooled before concentration by ultrafiltration, to compensate for this lower yield.

All three proteins of the TOMM synthase complex were concentrated in MBP storage buffer to at least 1 mg/ml and stored at 4°C for up to two weeks. Unexpectedly, concentrated solutions containing BcerB dehydrogenase did not have the strong yellow colour

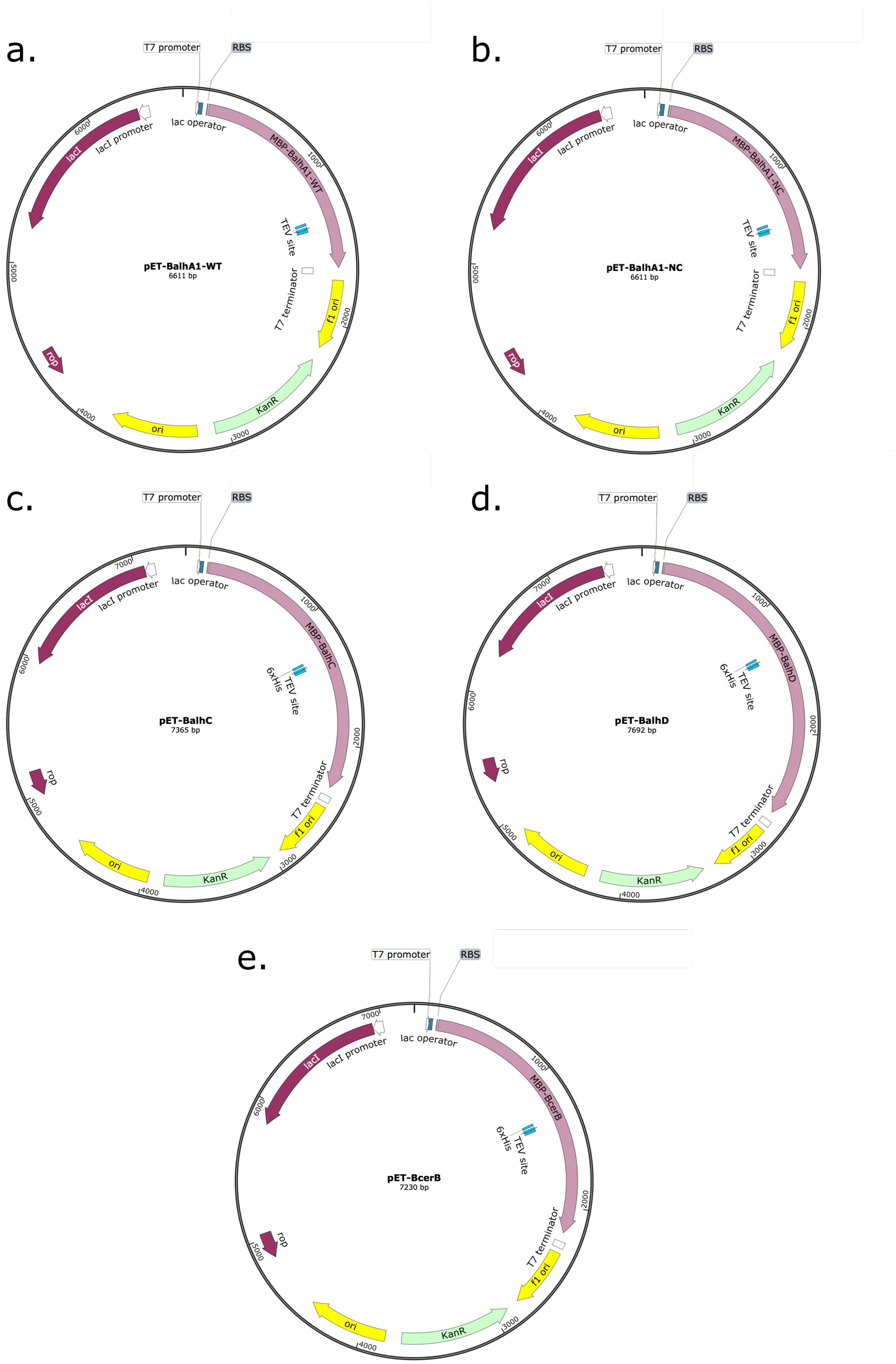


FIGURE 5.14: Constructs used for the expression of the *Bacillus sp.* Al-Hakam TOMM synthase complex and substrates - All constructs were fused to MBP at their N-termini and a TEV Protease recognition site was included to allow the production of fusion-free proteins. (a)BalhA1-WT wild-type substrate peptide. (b) BalhA1-NC negative control peptide with no cyclisable residues. (c) BalhC docking protein. (d) BalhD cyclodehydratase. (e) BcerB dehydrogenase from *B. cereus*. Constructs kindly donated by Prof. Douglas A Mitchell.

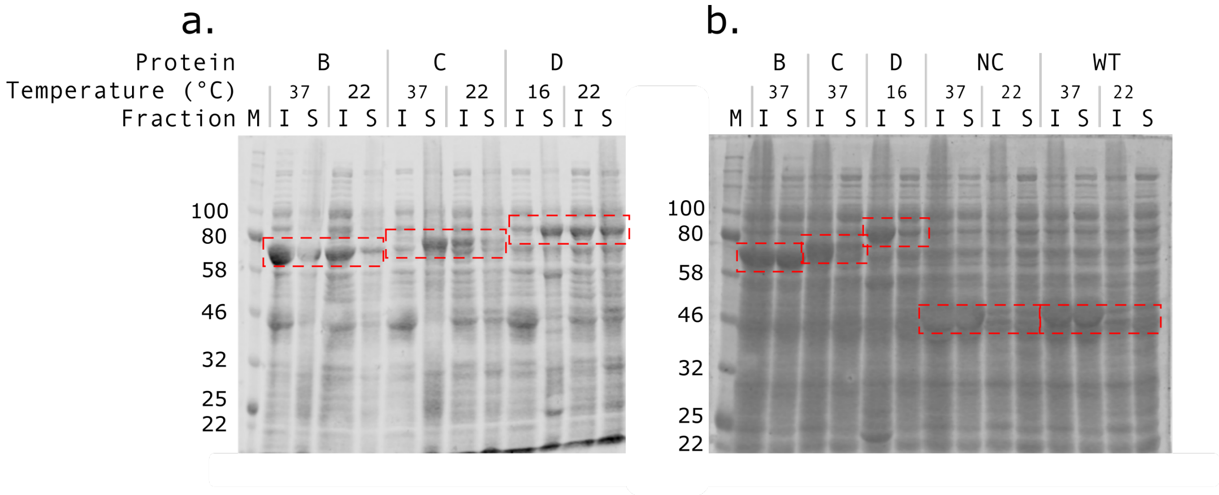


Figure 5.15: Induction tests for proteins from the *Bacillus sp.* Al-Hakam TOMM synthase complex. All constructs were induced under the conditions optimised for the equivalent proteins in the *B. amyloliquefaciens* complex and the conditions described in (Melby et al., 2012). (a) Induction tests for TOMM synthase complex enzymes. (b) Confirmation of selected conditions for TOMM synthase complex enzymes and induction tests for substrate peptides. Bands that correspond to the proteins of interest are enclosed in red dashed boxes. B - BcerB dehydrogenase (Expected MW = 75.8kDa). C - BalhC docking protein (Expected MW = 80.4 kDa). D - BalhD cyclodehydratase (Expected MW = 92.5 kDa). NC - BalhA1-NC negative control peptide (Expected MW = 53 kDa). WT - BalhA1-WT wild-type substrate peptide (Expected MW = 53.3 kDa). M - NEB broad range protein standard. I - Insoluble fraction. S - Soluble fraction. The induction times after addition of IPTG were 3 hours for 37 °C, 1 hour for 22 °C for NC/WT, 16 hours for 22 °C for B/C/D, and 16 hours for 16°C.

characteristic of FMN-bound TOMM dehydrogenases (Melby et al., 2012; Gonzalez et al., 2010b), despite being used by Melby et al. (2012) to correct this issue with the native BalhB enzyme and being expressed in media supplemented with 50 µg/ml riboflavin. Repeated inductions and purifications under the same conditions and exactly following the expression conditions of Melby et al. (2012) did not yield different results (data not shown), so assay development continued to determine whether there was any detectable residual activity in these BcerB preparations, while expression of the *E. coli* microcin B17 system was started as an alternative strategy (Section 5.1.3).

The substrate peptides were also recovered in high yield, but with strong free MBP bands (of approximately the same intensity as the MBP-BalhA1 fusion bands) indicating possible degradation of the core peptide region by proteases *in vivo* or after lysis (Fig. 5.16b). Similar bands are also present in the enzyme purifications, but in lower relative proportion.

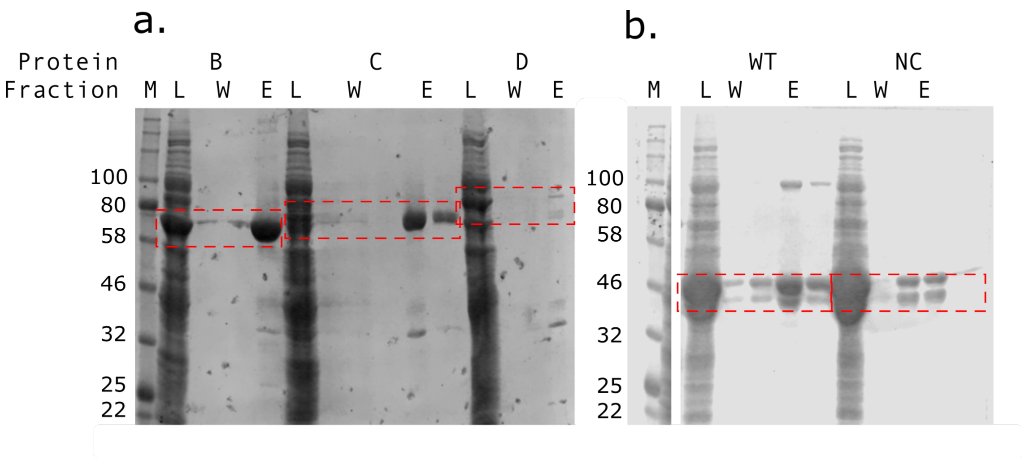


FIGURE 5.16: Purification of proteins for the *Bacillus sp.* Al-Hakam TOMM synthase complex - All constructs were purified by amylose affinity chromatography. (a) Purification of TOMM synthase complex enzymes with a yield for the D protein at least 30-fold lower than for B or C. (b) Purification of substrate peptides, with an apparent doublet band visible for the WT substrate. Bands that correspond to the proteins of interest are enclosed in red dashed boxes. B - BcerB dehydrogenase (Expected MW = 75.8kDa). C - BalhC docking protein (Expected MW = 80.4 kDa). D - BalhD cyclodehydratase (Expected MW = 92.5 kDa). NC - BalhA1-NC negative control peptide (Expected MW = 53 kDa). WT - BalhA1-WT wild-type substrate peptide (Expected MW = 53.3 kDa). M - NEB broad range protein standard. L - induced culture lysate. W - column wash fractions. E - Elution fractions. The image in (b) has been cropped to remove intervening lanes unrelated to the current experiment.

Subsequent induction experiments included a protease inhibitor cocktail in the lysis buffer (cOmplete, EDTA-free Protease Inhibitor Cocktail, Roche) and produced lower levels of degradation products (Section 5.1.3.1).

An apparent dimer band was also visible in the BalhA1-WT purifications (Fig. 5.16), which would appear to be dimerisation induced by disulfide bridges between the Cys residues in the WT peptide, absent in the NC variant. However, all samples were pre-incubated in SDS-PAGE loading buffer containing 10 mM TCEP at 95°C, which should be sufficient to reduce all disulfides in proteins (Burns et al., 1991). Since this apparent dimer is a relatively minor component (less than 20% of the intensity of the monomeric MBP fusion) of the purified fractions, no effort was made to isolate the monomeric band before starting assay development.

At this point, all components needed for *in vitro* activity assays of the *Bacillus sp.* Al-Hakam synthase complex were successfully expressed and purified, enabling assay development to begin.

5.1.2.2 Development of *in vitro* assays for the *Bacillus sp.* Al-Hakam TOMM synthase complex

Initially, fluorescent labeling of free thiols in the peptide substrates and MALDI-TOF detection were carried out on TEV-protease-cleaved BalhA1 substrate variants to confirm these peptides had the expected mass and free thiols. The iodoacetamide labeling reagent used in previous experiments was substituted for Fluorescein-5-Maleimide, due to the higher specifity for thiol groups over the undesired side reaction with amino groups (L.C. Cheah, unpublished data). Free BalhA1-WT peptide has an expected MW of 6034.8 and six cysteines which should be labeled by the maleimide-fluorescein reaction. On the other hand, free BalhA1-NC has an expected MW of 5807.0 Da and no free thiols to react with the maleimide reagent. However, both peptides showed similar levels of labelling upon incubation with maleimide-fluorescein (Fig. 5.17), indicating similar reactivity of the free peptides to the maleimide group in these conditions, corresponding to side reactions. Since TCEP or other reducing agents were excluded from labeling mixtures to prevent scavenging of the dye by these compounds (Shafer et al., 2000), it is possible that the thiols in the WT peptide were not accessible due to intra- or intermolecular disulfide bonds and the observed labelling was due to off-target reactions. This is corroborated by the continued presence of the apparent dimer band of approximately 100 kDa only in the WT peptide lanes.

MALDI-TOF spectra of these two peptides contained peaks in the expected m/z region for singly-charged species (Fig. 5.17), suggesting that the free peptides observed in tricine SDS-PAGE corresponded to the expected sequences of the expressed constructs. Synthase assays were carried out by incubating WT and NC peptides in the presence of TEV-cleaved BcerB, BalhC and BalhD, followed by MALDI-TOF detection, which should show a shift of -20 Da in the BalhA1-WT peak for every heterocyclisation event (-100 Da for conversion into the final product containing 5 azole heterocycles). However, only the non-cyclised peak was ever observed in these experiments, meaning that no azoles were produced. This could result from a complete lack of heterocyclisation activity on the BalhA1-WT purified substrate or from acid hydrolysis of azolines (Melby et al., 2012), that were not oxidised by the BcerB enzyme due to the lack of bound FMN cofactor. Azolines can be detected by labeling free cysteines after incubation with the BCD enzymes, followed by incubation in acidic conditions to hydrolyse azolines and MS/MS fragmentation. Any residues that were not heterocyclised would be labeled, the oxidised thiazoles would remain intact and thiazolines would be detected as free Cys residues after hydrolysis.

Due to the inability to express the BcerB enzyme with the FMN cofactor required for its activity and the lack of specific cysteine labelling in the BalhA1-WT peptide, efforts were shifted towards establishing the *E. coli* microcin B17 synthase complex for *in vitro* assay development. Activity assays with the FMN-bound BamB enzyme from *B. amyloliquefaciens* and the BalhCD proteins were not attempted since they were not described in any reports from the Mitchell group and the microcin B17 system was deemed a safer option to achieve successful activity detection.

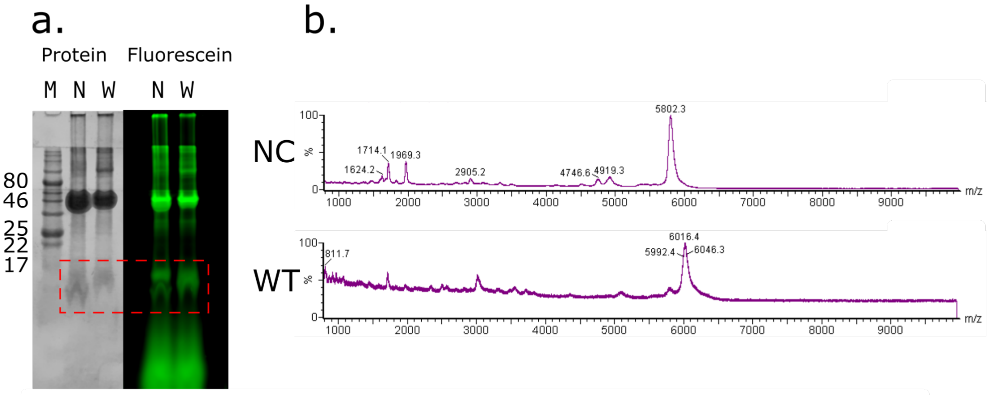


Figure 5.17: Development of *In vitro* activity detection assays for the Balh TOMM synthase complex - (a) Maleimide-fluorescein labeling of cysteines. Total protein was detected by coomassie staining (left) and Maleimide-fluorescein-tagged proteins were detected by fluorescence (right) No increase in fluorescent labeling was detected for the WT peptide over the NC peptide (red box). M - Molecular weight marker. N - Non-Cyclisable peptide. W - Wild-type peptide (b) MALDI-TOF spectra of NC (expected MW = 5807.0 Da) and WT (expected MW = 6034.8 Da) peptides. Each cyclisation event was expected to produce a loss of 20 Da, to a total of 100 Da all the modifiable sites in WT were cyclised.

5.1.3 *In vitro* reconstitution of *Escherichia coli* microcin B17 biosynthesis

This section will describe experiments carried out with the aim of overexpressing the *E. coli* TOMM synthase complex and measuring its activity *in vitro*.

5.1.3.1 Expression and purification of the *E. coli* microcin B17 TOMM synthase complex and substrates

The constructs kindly donated by the Mitchell group for expression of the *E. coli* microcin B17 (Mcb17) system components have the same design used for the *Bacillus sp.* Al-Hakam proteins. N-terminal MBP fusions are present in all the constructs, with a TEV protease site to produce free proteins and under the control of the T7 promoter. (Fig. 5.18). Only one substrate construct was used, coding for the full-length McbA precursor peptide. All of the enzymes are the native versions from the microcin B17 biosynthesis operon: McbB docking protein, McbC dehydrogenase and McbD cyclodehydratase. Vectors were transformed into NEB 10-βcells for amplification, confirmed by sequencing and transformed into the expression strain T7 Express LysY/Iq.

Initial induction tests were carried out using the conditions established for the Balh proteins, producing satisfactory levels of soluble expression for all proteins but the McbD cyclodehydratase (data not shown). To improve the soluble yield of McbD, a range of culture conditions and media were screened (Fig. 5.19). The highest soluble yields were obtained using both TB and 2xTY media, with 16 h inductions at 25°C. All cultures were induced in shake flasks at 200 ml scale for purification.

All cultures were lysed in the presence of protease inhibitor cocktail and purified by amylose-affinity chromatography (Fig. 5.20a), following the methods in Section 2.4.5. During the purification runs, expression of McbC dehydrogenase with its required FMN cofactor was evident from the yellow colour of eluted fractions, becoming more intense after

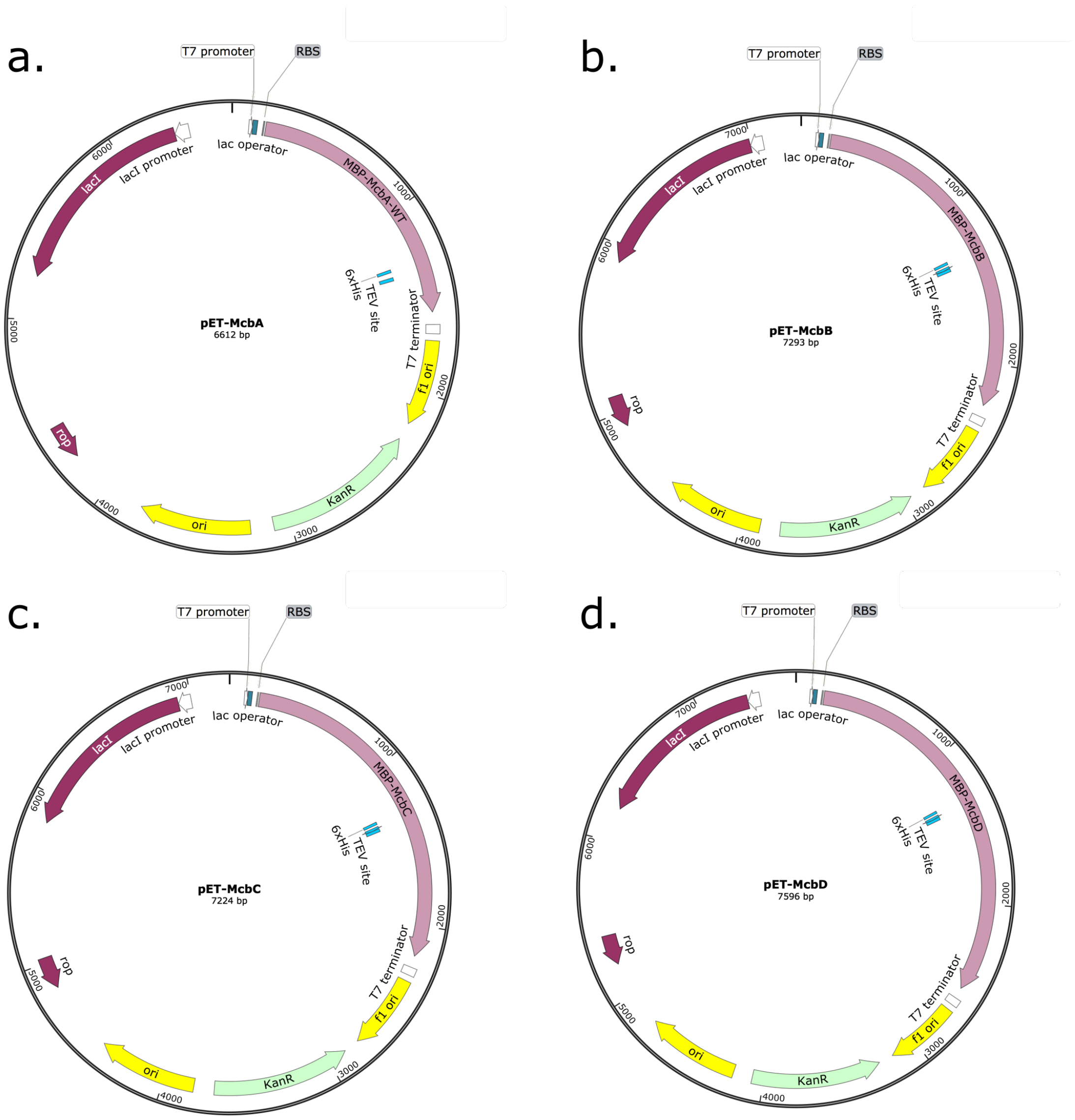


FIGURE 5.18: Constructs used for the expression of the *E. coli* microcin B17 synthase complex and substrates - All constructs were fused to MBP at their N-termini and a TEV Protease recognition site was included to allow the production of fusion-free proteins. (a)McbA substrate peptide. (b) McbB docking protein (c) BalhC dehydrogease. (d) BalhD cyclodehydratase. Constructs kindly donated by Prof. Douglas Mitchell.

concentration by ultracentrifugation. The relative proportion of free MBP from apparent degradation of fusion products was less than 10% of the total protein in the McbA lane, indicating that proteases did not cause significant degradation of fusion proteins. On the other hand, an apparent dimer of the MBP-fused substrate was also observed in this system — representing approximately 15% of the total protein in eluted lanes — but assay development was continued in the presence of this additional band.

McbD expressed in optimised conditions was still recovered in approximately 50-fold

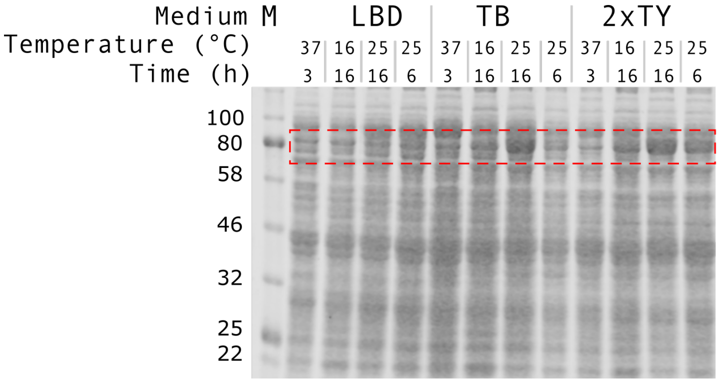


FIGURE 5.19: Induction test for the *E. coli* McbD cyclodehydratase. Induction at 25°C for 16h in 2xTY medium was used for subsequent cultures of this construct. Bands that correspond to the protein of interest are enclosed in red dashed boxes. LBD - LB with 5% glucose. TB - Terrific Broth

lower concentration than the other proteins in the system (Fig. 5.20). A possible cause for these lower purification yields is reduced amylose-binding affinity of the MBP fused to McbD compared to other fusion partners, caused by MBP-fusion partner interactions that make the interaction between MBP and affinity resin less favourable (Park et al., 1998). Previous experiments by other group members with the *Pyrococcus furiosus* TOMM synthase complex suggested a capacity for heterodimerisation of the cyclodehydratase and the docking protein by copurifying both proteins by affinity chromatography when only one of the proteins was tagged with MBP (J. Ewan Coates, unpublished data). To test this possibility, the McbB docking protein and McbD cyclodehydratase were co-purified by mixing their induced lysates in 1:2 volumetric ratio before injecting into the chromatography column (to account for the higher soluble yield of McbB) (Fig. 5.20b), yielding approximately the a 2:1 McbB to McbD mass ratio in the eluted fraction, which was deemed acceptable for assay development.

5.1.3.2 Development of *in vitro* assays for the microcin B17 TOMM synthase complex

The purified proteins of the *E. coli* microcin B17 synthase system were used to start development of MS and cysteine-labelling assays. Initially, the substrate alone was digested

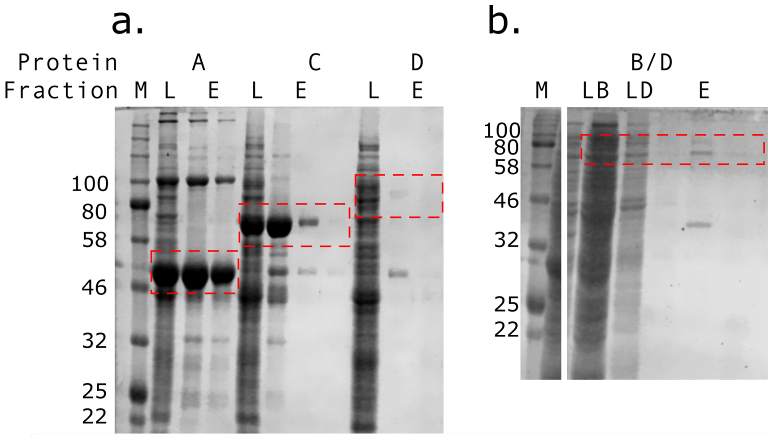


Figure 5.20: Purification of proteins for *E. coli* microcin B17 synthesis - All constructs were purified by amylose affinity chromatography. (a) purification of McbA substrate peptide, McbC dehydrogenase and McbD cyclodehydratase. The yield for McbD was at least 50-fold lower than for the other proteins. (b) Copurification of McbB and McbD, yielding both proteins in an approximate 2:1 mass ratio (B:D). Bands that correspond to the proteins of interest are enclosed in red dashed boxes. A - McbA substrate peptide (Expected MW = 49.6 kDa). C - McbC dehydrogenase (Expected MW = 74.4 kDa). D - McbD cyclodehydratase (Expected MW = 86.5 kDa). B - McbB docking protein (Expected MW = 77.6 kDa). M - NEB broad range protein standard. L - induced culture lysate. W - column wash fractions. E - Elution fractions. LB - induced McbB culture lysate. LD - induced McbD culture lysate. The image in (b) has been cropped to remove intervening lanes unrelated to the current experiment.

with TEV protease to establish conditions for production of the free peptide. Overnight digestion of 20 µg MBP-McbA with 8 U TEV Protease led to 30% cleavage of the fusion protein detectable by SDS-PAGE (Fig. 5.21a) and the remaining material was split into two equal fractions and precipitated by TCA-acetone for detection of free McbA peptide by Tricine SDS-PAGE. Due to the known poor solubility of the McbA substrate (Yorgey et al., 1993; Sinha Roy et al., 1998), pellets were resuspended in acetonitrile and MBP storage buffer (Fig. 5.21 b). No clear band corresponding to the free McbA peptide could be observed, but the lane containing acetonitrile-resuspended material had an amount of total protein that was at least 5-fold higher than was obtained in the aqueous resuspension. Precipitation followed by resuspension in acetonitrile coupled to a three-fold increase in the TEV protease concentration (24 U TEV for 20 µg protein) were used to produce material for establishment of conditions for MS detection of the McbA peptide and its heterocyclised products.

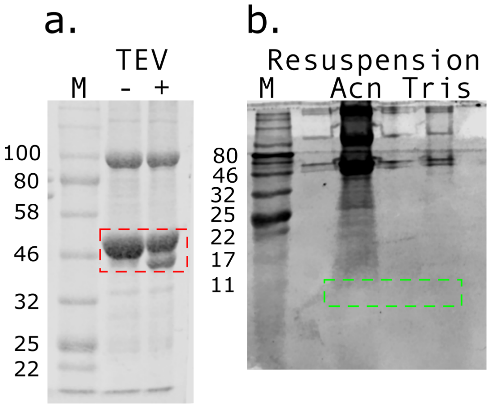


Figure 5.21: Production of free McbA substrate peptide by cleavage with TEV protease. (a) MBP-McbA fusion cleaved with TEV protease. Only partial cleavage was observed, as a lower band corresponding to free MBP (both bands enclosed in a red dashed box). The free McbA peptite (expected MW = 6.2 kDa) is expected to migrate at the same rate as the dye front in these conditions and, therefore, would not be detected. (b) Tricine SDS-PAGE of cleaved McbA peptide after precipitation with TCA/Acetone and resuspension in acetontrile (Acn) or MBP storage buffer (Tris). The expected position for the free McbA peptide is enclosed in the green dashed box. M - NEB broad range protein standard.

Attempts at detecting the cleaved McbA peptide by MALDI-TOF, directly from TEV cleavage reactions or after precipitation and resuspension in acetonitrile, failed to produce any signal close to the expected mass value of 6239.8 Da (data not shown). This could be caused by the low recovery of peptides by acetonitrile resuspensions, compounded by losses of material during sample preparation, resulting in a final concentration of peptides in the MALDI-TOF plates that is too low for detection.

To remedy this issue, detection by Liquid Chromatrography coupled to Mass Spectrometry (LC/MS) was used to exploit the analyte concentration that occurs during chromatography for improved detection. Two separate reactions were carried out, one repeating the conditions used for MALDI-TOF detection of TEV-cleaved McbA and a separate reaction containing all the components of the Mcb17 synthesis machinery and a threefold increase in the MBP-McbA substrate concentration (60 µg MBP-McbA in 100 µl reaction volume) to attempt detection of synthase complex activity. Both reactions were TCA-acetone precipitated and resuspended in 10 µl acetonitrile for LC/MS injection. A single peak was detected for the TEV cleavage reaction at 6241 Da and multiple peaks consistent with multiple heterocyclised intermediates were visible in the TOMM synthase reaction, with a single larger peak at 6242 Da, very close to the expected MW of 6239.8 Da (Fig. 5.22). These spectra are the strongest indication of *in vitro* TOMM synthase activity that were obtained in this work and they suggest the presence of species containing a maximum of four heterocyclised residues (peak at approximately 6161 Da, mass 80 Da smaller than than the unmodified peak) out of a total of eight sites that must be heterocyclised in mature Mcb17. However, they are not sufficient to conclude unequivocally that heterocyclisation ocurred in these experiments and must be corroborated by other forms of evidence.

The commercial polyclonal antibody specific for the drug sulfathiazole (LSBio) was tested for its ability detect thiazole moieties installed in the McbA peptide. TOMM synthase reactions were carried out with the purified Mcb components, following the conditions that led to detection of possible heterocyclic species by LC/MS and using two separate batches of purified McbA peptide. To prevent loss of poorly-soluble material during sample handling for Western blotting, dot blots were carried out with the entire volume of reaction mixtures loaded onto the blotting membrane. However, no colorimetric signal specific to the reactions expected to contain thiazole modifications could be detected (Fig. 5.23). The difference in signal intensity between reactions containing BCD enzymes and the negative control was lower than 10%. The differences observed between spots could be explained by the higher substrate peptide concentration in purification batch 2 (approximately 2 µg/ml) compared with batch 1 (approximately 1 µg/ml), suggesting that the signal observed was a result of nonspecific binding to protein rather than specific recognition of thiazoles by the primary antibody.

Due to the solubility issues encountered while carrying out experiments with the full-length McbA substrate peptide, constructs were made by inverse PCR to express the truncated variant McbA$_{1\text{-}46}$ (Sinha Roy et al., 1998; Belshaw et al., 1998) with inserted residues at the C-terminus to increase solubility. Two different variants were made, one with a

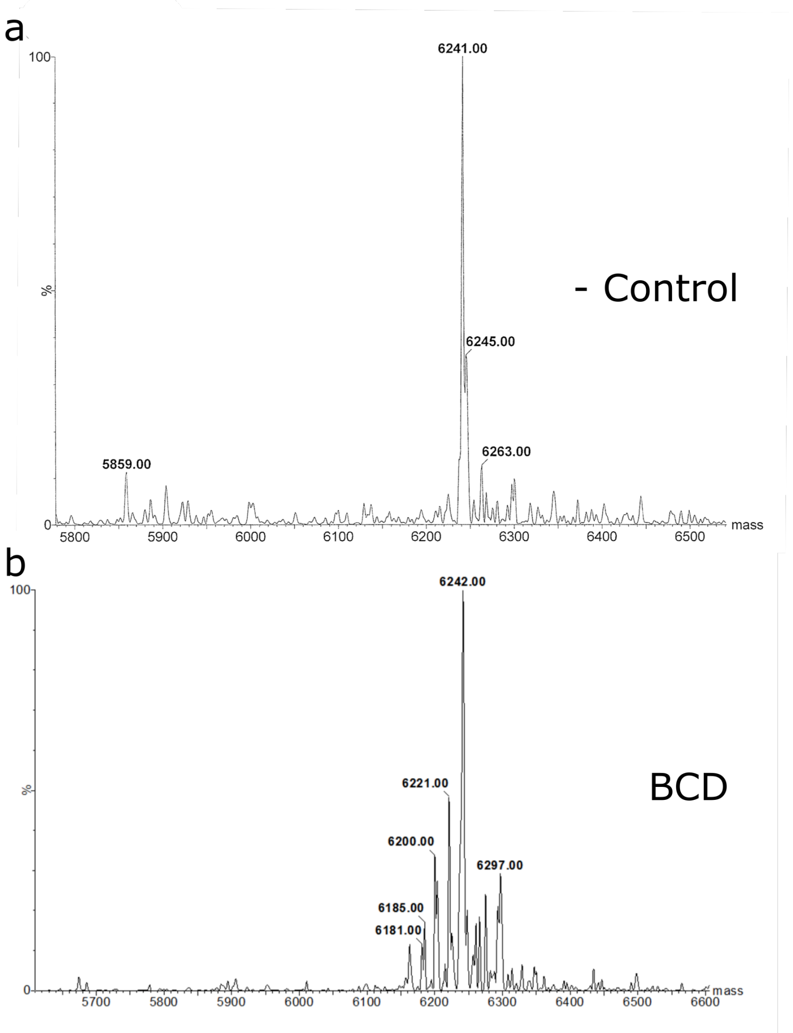


FIGURE 5.22: LC-MS detection of McbA heterocyclisation. (a) Negative control spectrum with free McbA peptide produced by TEV protease cleavage. (b) Heterocyclisation reaction using free McbA and TOMM synthase complex enzymes. Peaks with approximately 20 Da spacing between them indicate the formation of multiple heterocyclised intermediates. Expected MW of uncyclised McbA = 6370.9 Da. Plots were shifted to align the unmodified peak. Each cyclisation event was expected to produce a loss of 20 Da, to a total of 160 Da all the modifiable sites in McbA were cyclised.

KKD insertion (McbA$_{1\text{-}46}$ KKD) and a second variant with a QMQ insertion (McbA$_{1\text{-}46}$ QMQ). The length of the insertions was chosen arbitrarily and the sequences were selected to produce one variant with a highly-charged insertion — KKD — that is expected to have a strong impact on solubility and one variant with a mostly polar but uncharged insertion — QMQ — that would be expected to have a smaller contribution to solubility (Trevino et al., 2007). The McbA$_{1\text{-}46}$ truncation reduces the number of Gly stretches — which are likely to contribute strongly to the aggregation tendency of the peptide — and also contains

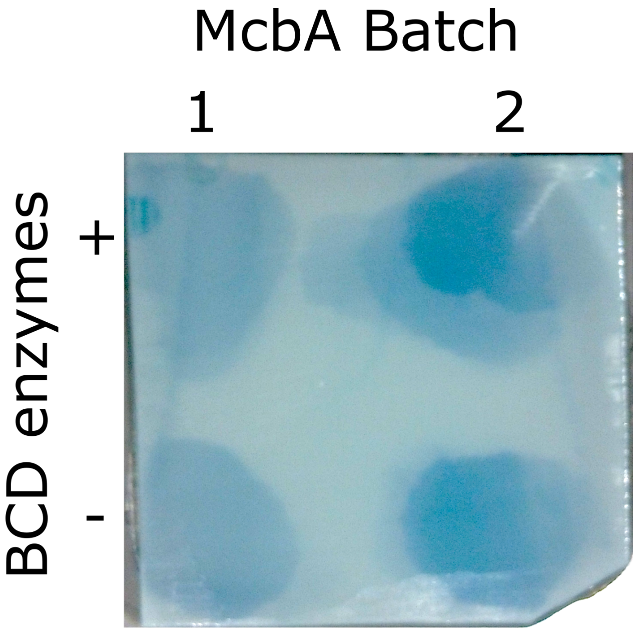


Figure 5.23: Dot blot of Microcin B17 synthase reactions using a polyclonal sulfathiazole primary antibody - Colorimetric signal was similar for reactions containing the synthase complex and the negative controls

only a single SC bisheterocyclisation site, simplifying the interpretation of results during assay development by MS detection of mass losses from heterocyclisation events or cysteine labelling. Negative control variants were also made for each construct by mutating both cyclisable residues to Gly.

After expression and purification, MBP-McbA$_{1\text{-}46}$ variants were cleaved with TEV protease, labelled with maleimide-fluorescein and resolved by Tricine SDS-PAGE. Distinct labeled bands corresponding to free McbA$_{1\text{-}46}$ peptide were visible for both insertion variants (Fig. 5.24). Importantly, no fluorescein signal was detected for the negative variants, indicating that maleimide labeling was specific for cysteines in the substrate peptides in this experiment, unlike the results observed for the BalhA1 substrates (Fig. 5.17a). Faint free peptide bands were also visible after staining for proteins, indicating an increase in aqueous solubility of the new variants. These results suggest that mutations to increase the hydrophilic character of substrate peptides are a viable strategy to produce material in high quantities for higher-throughput assays. However, the activity of the heterocyclisation machinery on these insertion variants was not measured and could be impaired especially by the charged insertions, since the WT sequence contains mostly hydrophobic residues surrounding the heterocyclisation sites.

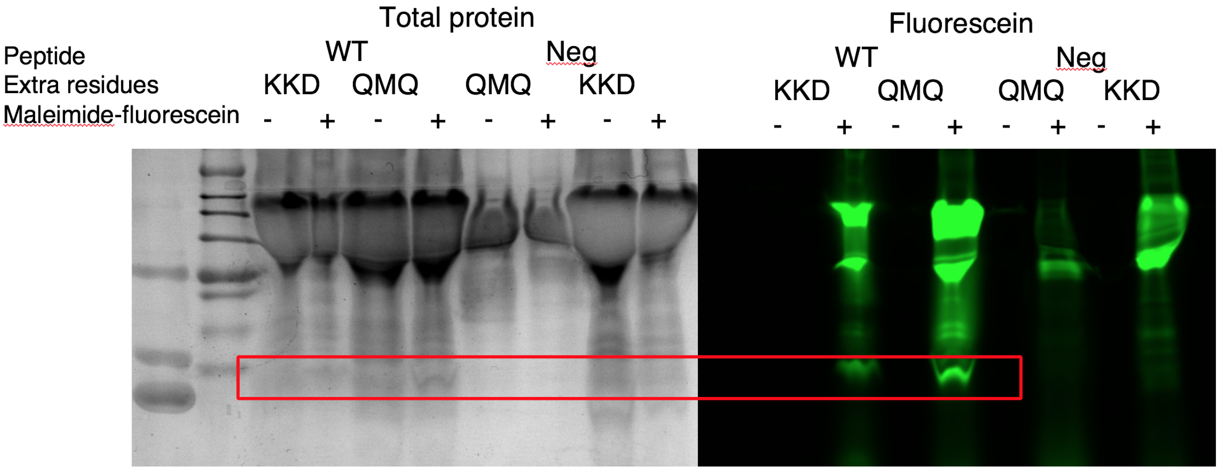


Figure 5.24: Maleimide-fluorescein labeling of McbA substrate variants for improved solubility. Tricine SDS-PAGE of McbA substrate variants KKD and QMQ. Differential labeling was observed between variants containing cysteines and negative control variants. Bands that correspond to the proteins of interest are enclosed in red dashed boxes. WT - McbA 1-46 substrate. Neg - McbA 1-46 substrate with S40G and C41G substitutions.

5.2 Conclusions

Despite the efforts by multiple groups to characterise the function of distinct TOMM synthase complexes and the range of products they are able to synthesise, only a few species have had their biosynthetic machinery for this class of compounds reconstructed *in vitro* (Li et al., 1996; Lee et al., 2008; Mitchell et al., 2009; Scholz et al., 2011; Melby et al., 2012). The relatively low interest in the study and engineering of the TOMM biosynthetic mechanism for the production of novel compounds can be at least partly attributed to the difficulties encountered here during attempts at establishing a robust *in vitro* system for detection of TOMM synthase activity and also reported by other groups.

The enzymes of the *B. amyloliquefaciens* complex responsible for the production of plantazolicin were solubly expressed and purified by affinity chromatography, but their activity could not be detected due to apparent cleavage of the substrate peptides. Difficulties in employing *E. coli* for the heterologous production of short peptides — especially those with antimicrobial activity — are widely acknowledged, including a tendency towards degradation by intracellular proteases (Piers et al., 1993) and toxicity against the host if genes related to immunity are not coexpressed (Li, 2009).

While the BalhA1 substrate peptides of the *Bacillus sp.* Al-Hakam complex were successfully expressed and purified in this work with their expected molecular weights, the dehydrogenase BcerB could not be produced with the bound FMN cofactor required for its activity. Curiously, this orthologous enzyme from *B. cereus* was selected by Melby et al. (2012) for its soluble expression with the bound cofactor to replace the BalhB enzyme, which was never successfully expressed with FMN. The failure to produce active BcerB protein illustrates the common issue of inconsistent expression of proteins across groups and even at different times, often requiring meticulous experimentation to determine the sets of conditions required for successful expression of a protein (Faiq et al., 2014).

A lack of bound FMN was also encountered by another member of the Pinheiro group, when attempting to produce the TOMM synthase complex from *Pyrococcus furiosus* (J. Ewan Coates, unpublished data). Milne et al. (1999) also attempted the coexpression of a chaperone shown to copurify with the complex, with the aim of increasing the concentration of active enzymes. However, a decrease in TOMM synthase activity was observed (Milne et al., 1999), suggesting that as-yet unidentified factors from native organisms could be needed for efficient folding and activity of the complex. Another possible solution to this issue is to extend the complementation approach used by Melby et al. (2012) and attempt to substitute non-functional dehydrogenases with the proteins that were successfully expressed with bound FMN such as McbC from *E. coli* and BamB from *B. amyloliquefaciens.*

The greatest degree of success in *in vitro* reconstruction was achieved here with the Mcb17 synthase complex from *E. coli.* All the proteins involved in Mcb17 heterocyclisation were successfully expressed and purified, including McbC dehydrogenase with bound FMN, and indications of azole formation were obtained by LC/MS detection of *in vitro* reactions. However, the poor solubility of the free substrate peptide McbA made handling of these activity assays difficult and also produced inconsistent results. The poor aqueous solubility of free McbA was previously acknowledged (Yorgey et al., 1993; Sinha Roy et al., 1998) and attempts in this work to produce substrate variants with additional polar residues for increased solubility suggested a degree of success in producing variants that would be more amenable to characterisation. Tolerance of the McbBCD complex toward these mutations would still need to be tested, along with any cytotoxic activity against sensitive *E. coli* retained by these variants.

An *in vivo* biosynthesis system could also be employed to characterise the plantazolicin biosynthetic system from *B. amyloliquefaciens*. This compound has been identified as a narrow-spectrum antimicrobial with a high specificity for *B. anthracis* and low activity on the non-pathogenic closely-related species *B. cereus* (Molohon et al., 2016), making its study relevant in the understanding of this pathogen and in drug development. While virulent strains of *B. anthracis* are classified as biosafety level 3 and most groups involved in protein engineering do not possess the facilities and training required to handle these organisms, plantazolicin is still active on avirulent strains classified as biosafety level 2 (Molohon et al., 2016). Therefore, the functional prediction results obtained here could be transferred to the plantazolicin biosynthetic complex for product and complex engineering with the aim of creating tools for the study of *B. anthracis* pathogenicity and, eventually, novel compounds for drug development.

Characterisation and engineering of TOMM synthase complexes and their products is still an open challenge in the field. The work described here represents incremental steps toward the aim of establishing a platform for the introduction of heterocycles into arbitrary positions in synthetic peptides, which would enable the quick and efficient exploration of the possible structures that exist in this class of compounds.

Chapter 6

InDel assembly as a framework for employing length and compositional variation in directed evolution

6.1 Introduction

This chapter covers the development of a strategy for the directed evolution of synthetic TOMM products, which is comprised of (1) a novel library assembly method to produce synthetic TOMM precursor peptides with targeted diversity in length and composition and (2) an analysis framework to detect enrichment of motifs from selection experiments employing these libraries. In the absence of a functional *in vitro* TOMM biosynthesis platform, the strategy was employed to engineer a flexible substrate-recognition loop in TEM-1 β-lactamase, generating seven novel variants of diverse lengths, with activity against a non-cognate substate.

6.1.1 Development of a method to produce targeted length and compositional diversity

Given the limitations of current methods for the production of protein coding sequences containing length and sequence diversity (Section 1.5.1.3), a novel DNA assembly strategy was conceived as a tool to produce heterometric libraries of variants for screening novel bioactive TOMMs. Significant structure-function analysis has not been reported for any already characterised TOMM. Therefore, such a method would ideally be able to produce tunable distributions of length and composition. Using this flexibility, screening or selection for novel compounds can start using broad diversity to identify candidate motifs, followed by design of targeted libraries to explore the sequence space around hits and further rounds of enrichment to obtain improved variants.

Here, an enzymatic method — called InDel Assembly — was established to produce tunable heterometric libraries based on enzymatic cycles of restriction and ligation in which codon-length assembly blocks are added sequentially to template oligos (Fig. 6.1). Biotinylated dsDNA oligonucleotides — named template oligos — are immobilised on paramagnetic streptavidin beads, allowing for quick exchange of reaction components and washes between steps. Assembly starts with digestion of bead-bound oligos (Fig. 6.1 Step i) by a Type IIS restriction endonuclease, leaving a $5'$ overhang (Fig. 6.1 Step ii). Assembly block oligos with three degenerate base (NNN) overhangs are then annealed to the exposed template overhangs and ligated, increasing the length of the coding sequence by a single codon (Fig. 6.1 Step iii). The assembly blocks added for ligation in the example extend the sequence by a single codon (Fig. 6.1 Step iii), which ensures the lack of frameshifts in protein coding sequences. Along with the single codon, each assembly block also contains a recognition site for the same Type IIS endonuclease, allowing a new assembly cycle to begin.

Mixtures of assembly blocks containing different codons can be added in Fig. 6.1 Step ii, theoretically leading to incorporation of each codon according to its relative concentration in the assembly mixture. At the end of the assembly process, a final cycle is carried out with a capping block containing a primer binding site to enable amplification of the library for cloning into a vector of choice.

The assembly pipeline proposed here is reminiscent of the steps used by Sloning (Van den Brulle et al., 2008) and ProxiMAX (Ashraf et al., 2013) technologies, but these make use of

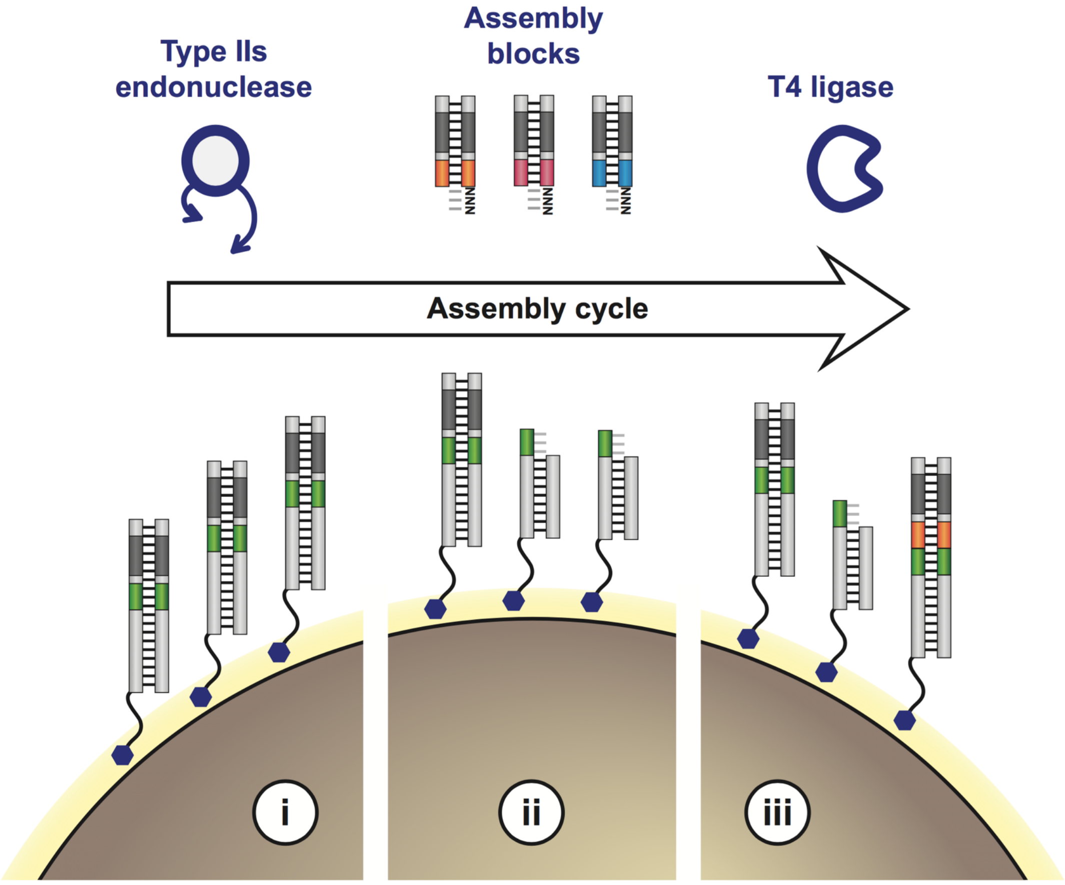


FIGURE 6.1: Mechanism for proposed InDel Assembly strategy - An assembly cycle starts with bead-bound template (i) oligos which are digested with a type IIS endonuclease (ii) to produce overhangs for annealing and ligation of building block oligos (iii). Both enzymatic steps have < 100% efficiencies, leading to a range of product sizes at the end of the assembly. According to Tizei et al. (2017).

elution from beads or amplification to enrich for sequences that were successfully extended in every assembly cycle, reducing the likelihood of any length variations.

The key difference between the two methods cited above and the proposed workflow for InDel Assembly is that digestion and ligation of bead-supported oligos in every cycle of the assembly process are not carried out to completion, with only a fraction of all template oligos being extended after each full assembly cycle. This is analogous to the strategy used in the COBARDE technology (Osuna et al., 2004) for producing libraries containing length diversity, but using partial enzymatic cycles in place of partial deprotection during solid-phase synthesis of oligos. Once the assembly cycles are completed, the lengths of sequences produced should follow a binomial distribution determined by the number of assembly cycles and the combined efficiency of the restriction and ligation enzymatic reactions.

6.1.2 Enrichment analysis of libraries containing heterometric diversity

In addition to the difficulties involved in producing libraries containing targeted heterometric diversity, the analysis of results obtained during directed evolution employing such libraries also poses a challenge. The analysis strategies discussed in Section 1.5.3 for detecting motifs enriched by selection using NGS data compare variants using sequence alignments, which are known to generate unreliable results when significant length variation exists (Nuin et al., 2006). Even when length variation is acknowledged as an important factor in determining function — such as the specificity-determining loops in antibodies — sequences of equal length are grouped in analysis to ensure better quality for alignments, discarding any information that could be gained by comparing motifs of similar sequence that differ by indels (Ravn et al., 2013). The lack of available methods capable of efficiently detecting enriched sequences containing length variation is also acknowledged by important groups in the field of protein engineering (Toth-Petroczy and Tawfik, 2014b) and this gap motivated the development of an analysis strategy along with the library assembly method, to create an integrated framework to accelerate directed evolution using heterometric libraries.

Sequence comparisons on the basis of subdivisions of larger sequences into "words" or k-mers have become widely adopted with the advent of large-scale sequencing datasets that became accessible with NGS. Any sequence of characters can be represented as a list of all its component k-mers, which are shorter strings of k characters. K-mer-based comparison strategies been developed for phylogenetic reconstruction using large datasets (Gardner and Hall, 2013) and *de novo* assembly of genomes from NGS data (Zerbino and Birney, 2008; Bankevich et al., 2012), with the latter becoming the *de facto* standard tools of the field.

Genome assembly from k-mers is a problem of reconstructing a sequence from its component k-mers, obtained from subdivisions of the reads produced by NGS methods. Algorithms such as de Bruijn graphs are used to perform this reconstruction, representing k-mers as edges in a graph and generating a reconstructed sequence by finding paths connecting overlapping k-mers (Compeau et al., 2011).

Since the component k-mers of any given sequence enriched by selection during directed evolution also become enriched, a strategy capable of detecting sets of overlapping k-mers that became enriched during selection could be used to reconstruct functional motifs in a process that is analogous to the ones used in genome assembly. Such a method would also recover shared motifs contained in enriched sequences of different lengths, since the component k-mers of these motifs would also be shared — thereby avoiding the known limitations of the commonly-used alignment-based enrichment analysis methods.

6.1.3 Selection of a model directed evolution target as a proof-of-concept for InDel Assembly

Since a TOMM synthase *in vitro* modification system could not be established (see Chapter 5), libraries produced with the system would need to be validated against a proof-of-concept biological system. Ideally, the selected system would produce a phenotype that can be easily screened in a high throughput assay or selected, that is already well-characterised, and with known variants caused by indels. Among the phenotypes considered in this search were the change in the activity of phosphotriesterase (PTE) when one of its loops is deleted (Afriat-Jurnou et al., 2012), the shifts in kinetic parameters caused by truncation of an active site loop in a cellulase (von Ossowski et al., 2003), the improved folding of a single residue deletion within GFP (Arpino et al., 2014) or improved binding after loop extensions in affinity maturation of certain antibodies (Krause et al., 2011). However, establishing a system to detect and enrich for a desired phenotype in these proteins were not compatible with the timeframe available for the PhD project.

The β-lactamase TEM-1 is a widely-used model for directed evolution (Jones, 2005; Kipnis et al., 2012; Palzkill et al., 1994a; Petrosino and Palzkill, 1996), due to the antibiotic resistant phenotype that it confers upon *E. coli* cells when periplasmically expressed. The wild-type enzyme confers resistance to penicillin-derived antibiotics such as ampicillin by hydrolysing the β-lactam ring in their structure, rendering these molecules inactive. In addition to the ease of selection, two other main factors contribute to making TEM-1 a versatile tool in directed evolution studies. First, there is a wide range of commercially available cognate and non-cognate substrates, many of which are commonly-used antibiotics in human or livestock therapeutic applications. Second, there is a repertoire of previously-described variants from experimental and clinical work with varying substrate specificities.

TEM-1 belongs to class A β-lactamases, α-β-α sandwich enzymes with a structure that is very similar to the peptidoglycan biosynthesis enzymes targeted by β-lactam antibiotics (Kelly et al., 1986). TEM-1 hydrolytically inactivates its substrates employing a catalytic mechanism that resembles the one found in serine proteases (Minasov et al., 2002). In TEM-1, Glu166 and Lys73 act in concert to activate Ser70 to attack the amide bond in the substrate, forming an acyl-enzyme intermediate (Meroueh et al., 2005). The same Glu residue then acts in the deacylation reaction by activating a water molecule that attacks the ester to regenerate the free enzyme (Adachi et al., 1991).

The level of resistance conferred by a given enzyme variant against an antibiotic is commonly described by the Minimum Inhibitory Concentration (MIC) of a strain expressing that variant, which is defined as the smallest concentration of the antibiotic that prevents growth of the strain. This value — and consequently the ability of a strain to survive selection under a given antibiotic concentration — is determined by two factors: the specific activity of the enzyme on the antibiotic and the amount of active protein exported to the periplasm per cell. Therefore, it is important for selection systems to produce homogeneous expression levels across a population, to minimise the influence of expression variance on the resistance of cells and ensure a greater likelihood that enzyme variants with higher specific activities are recovered.

Among the regions of TEM-1 that have been targeted in directed evolution studies, the active-site Ω-loop is especially relevant because many known variants of this sequence produce enzymes with activity on the non-cognate substrate ceftazidime (See Fig. 6.2 for a comparison to the cognate substrate ampicillin), including mutations at the catalytic Glu166 residue. Known ceftazidime-active variants of this loop differ from the wild-type (WT) sequence in length as well as composition, including variants with long insertions of up to 19 residues (Palzkill et al., 1994b; Petrosino and Palzkill, 1996; Hayes et al., 1997). Circular permutation of the TEM-1 sequence has also been exploited to produce novel functional enzymes and one variant which has its N- and C- termini at the Ω-loop was found to be active on cefotaxime, an antibiotic that belongs to the same β-lactam class as ceftazidime (Guntas et al., 2012). The highest activity mutant described for this loop is a triple-mutant of residues 165-167, from the wild type WEP to YYG (Petrosino and Palzkill, 1996).

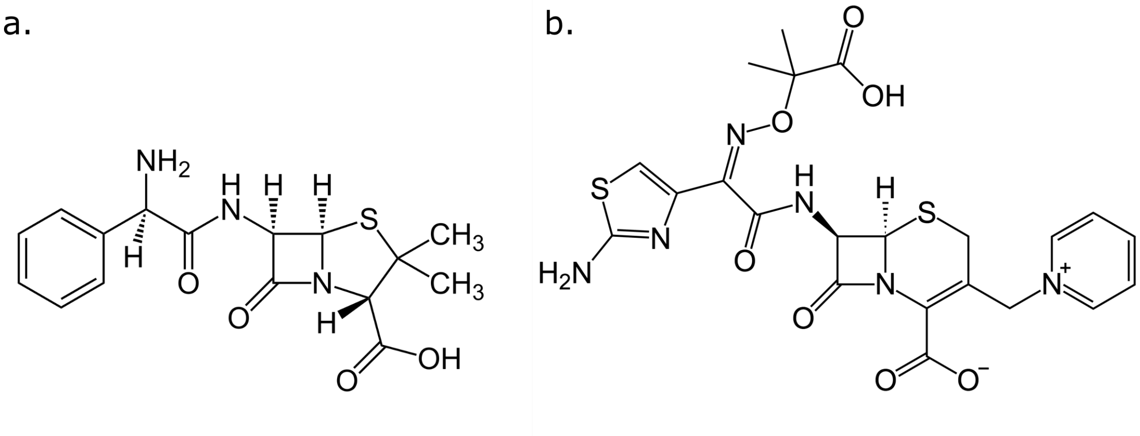


FIGURE 6.2: Structures of the β-lactam antibiotics (a) ampicillin and (b) ceftazidime. Public domain images obtained at www.wikipedia.org.

The structure of the $_{165}$YYG$_{167}$ mutant (hereafter referred to as $_{164}$RYYGE$_{168}$ to include the adjacent residues) was determined and the substantial shift (> 8 Å) in the conformation of the loop in the triple mutant suggests that the Ω-loop tolerates the large changes to its main chain that are needed to accomodate the bulkier cephalosporin substrate, while still producing a functional protein (Fig. 6.3, structures from Stojanoski et al. (2015)). Interestingly, the mutated Glu 166 residue is important for both acylation and deacylation steps in the WT enzyme and the structure of the triple-mutant suggests that Tyr 166 is able, at least, to partially substitute for those functions in ceftazidime hydrolysis (Stojanoski et al., 2015).

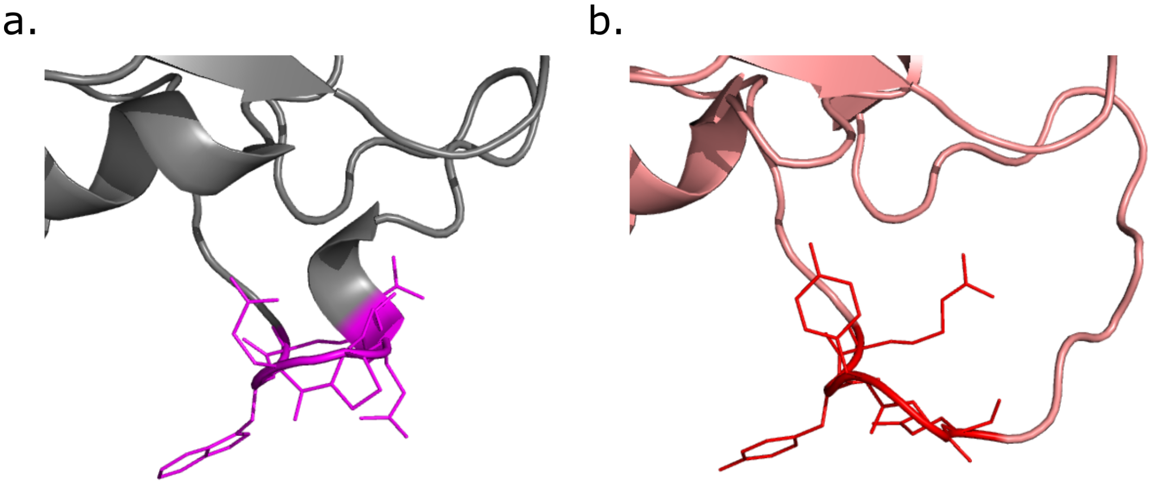


Figure 6.3: (a) Structure of the wild-type (highlighted residues in magenta and the remainder of the structure in grey cartoon form; PDB structure 1XPB) and $_{164}$RYYGE$_{168}$ (highlighted residues in red and the remainder of the structure in light pink cartoon form; PDB structure 4RVA) TEM-1 Ω-loop. The three residues mutated in the $_{164}$RYYGE$_{168}$ variant are highlighted along with the adjacent fixed residues and their side chains displayed. The region downstream of the mutations has a distinct shift in conformation when compared to the wild-type structure.

Given the role of residues 165-167 of Ω-loop loop in determining TEM-1 substrate specificity and the known length and composition variants with selectable activity on ceftazidime, this region was selected as a proof-of-concept for libraries produced by InDel assembly. The initial goal was to produce a heterometric library with diversity centered around the known $_{164}$RYYGE$_{168}$ variant — to validate the system by recovering this known positive variant and at the same time explore the sequence space around it for other variants active on ceftazidime that could not be easily produced by other diversification strategies. Two parallel strategies were then pursued to refine the search of the Ω-loop sequence space: purifiying selection with increasing stringency to isolate any superior variants and deep sequencing of diversified libraries before and after selection steps — to identify motifs which contributed to activity against ceftazidime and inform further rounds of diversification and selection.

The main requirement for establishing a selection system for TEM-1 function is periplasmic expression of the β-lactamase active protein, conferring resistance to β-lactam antibiotics in the culture medium by hydrolysing these molecules before they can act on cells and employing whole cells as the genotype-phenotype linkage. Previous selection experiments were carried out using high copy number vectors with the TEM-1 gene in a plasmid under control of the native constitutive promoter (Palzkill and Botstein, 1992).

The initial design attempted here for the InDel proof-of-concept had the TEM-1 gene under the control of the tunable arabinose-inducible pBAD promoter in the low-copy number pBAD-30 plasmid (Guzman et al., 1995). This strategy was chosen to allow precise control of β-lactamase expression levels for fine tuning of selection strengths (Guzman et al., 1995), without producing the very high expression levels that can be obtained by other inducible promoters such as the T7 promoter (Rosano and Ceccarelli, 2014). Low levels of induction using L-arabinose would then be used to induce production of TEM-1 protein at levels similar to the ones obtained from the constitutive promoter used in previous studies (Palzkill and Botstein, 1992) and there should be no ampicillin resistance when cells are grown in the absence of inducer. Diversified Ω-loop regions would be amplified from InDel-assembly derived libraries with primers containing overhangs for the Type II restriction enzyme BsaI. These fragments could then be inserted into the constant region of the TEM-1 coding sequence by inverse-PCR amplification of entire plasmids with the appropriate restriction sites and overhangs for the library fragments to be ligated and transformed into *E. coli* for selection.

6.1.4 Residue numbering for length-variant sequences

To ensure consistent numbering of residues when comparing sequence variants of divergent lengths, a numbering scheme inspired on the standards used for antibodies (Abhinandan and Martin, 2008) was adopted here. In this scheme, residues outside the diversified region maintain the same position numbers as the wild-type sequence and any changes in length within the diversified region lead to the introduction of gaps or additional symbols in the numbering. Since there is no information to reconstruct the full sequence of events that generated any given sequence in an InDel library and assign a position to insertions or deletions, all changes in length are defined to occur from the C-terminal end of the variable region. This means that variants shorter than the wild-type length are represented as deleted positions between the last residue in the variable region and the first residue of the downstream invariant region. Inversely, in sequences longer than the wild-type length, the inserted residues are given the same number as the last position of the variable region in the wild-type sequence followed by a lower-case letter to continue the count in alphabetical order. Examples of this numbering scheme applied to TEM-1 Ω-loop variants can be found in Table 6.1.

Length	**Variable region**	**Downstream invariant residue**
= WT	$_{164}$RYYGE$_{168}$	169L
< WT	$_{164}$RYGA$_{167}$	169L
> WT	$_{164}$RYYGEMG$_{168b}$	169L

Table 6.1: Examples to illustrate the residue numbering scheme adopted for length variants produced by InDel Assembly

6.2 Results and Discussion

6.2.1 Simulation of libraries produced by InDel Assembly

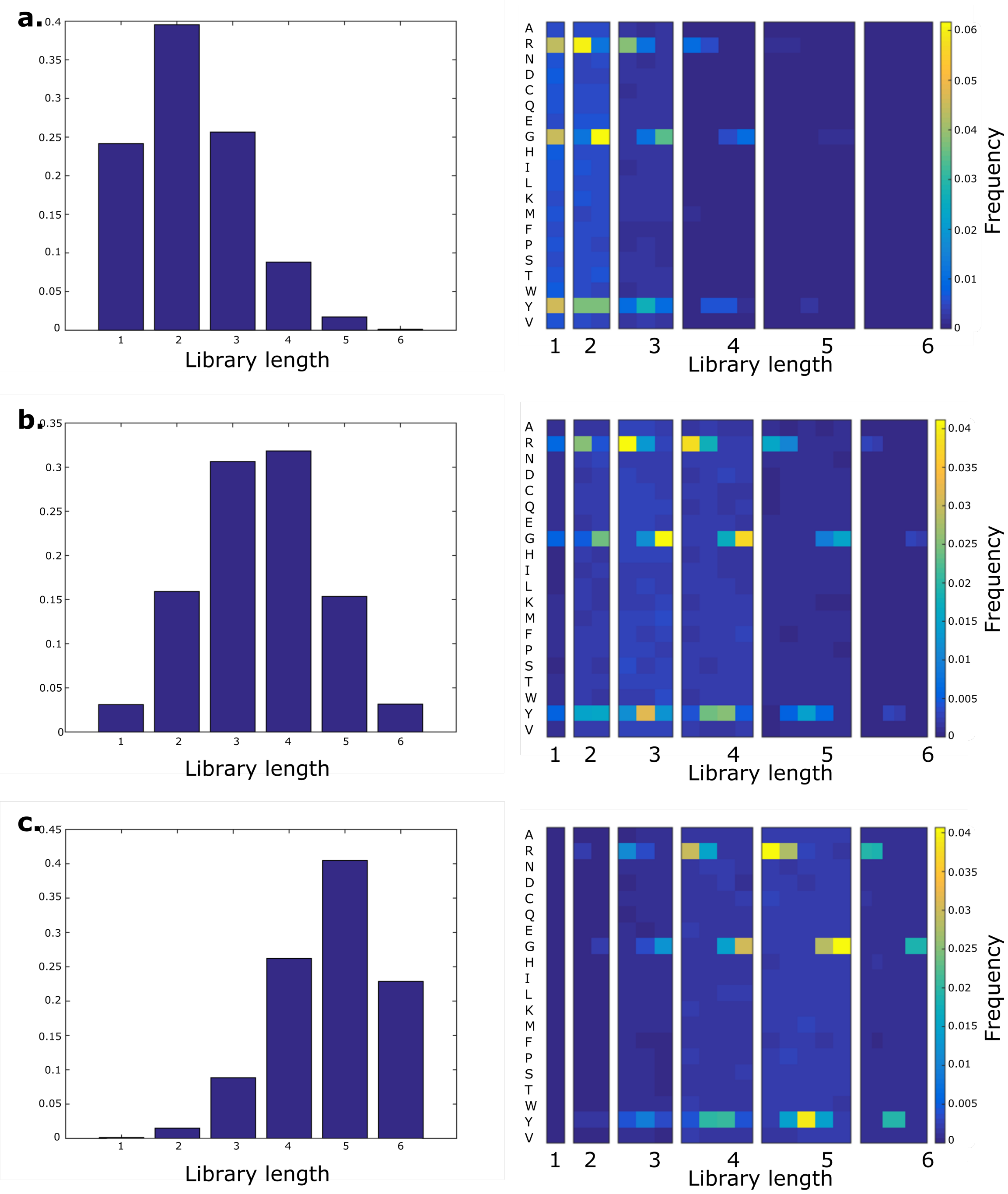
A simulation tool written by Dr. Vitor Pinheiro in MATLAB was used to explore the capabilities of the proposed method by generating hypothetical InDel Assembly libraries, based on cycles of restriction and ligation with efficiencies lower than 100%. This model uses a single per-cycle efficiency parameter to encompass the restriction and digestion steps, being represented as the probability that a codon from the chosen assembly pool would be incorporated into each sequence in the library. The identity of the inserted codons in each step is determined randomly, with each codon having a likelihood of incorporation proportional to its relative concentration within the pool used for that step. Overall, the input parameters for the simulations are per-cycle efficiency, number of cycles in the assembly, number of sequences in the simulated library, and relative concentrations of each of the 20 codons in each assembly cycle.

Rounds of simulations were carried out to explore the parameter space for the model and analyse the impact of varying parameters on the length and composition distributions in the simulated libraries. For each run, the length distribution of the products is represented in a histogram and the composition distribution is summarised by heatmaps showing how composition is distributed along the length of sequences of each possible length. The simulations shown here aimed to produce diversity in length and composition around an arbitrarily-chosen target sequence: RYG.

A set of representative simulated library assemblies exploring the sequence space around RYG was made, with the intention of reproducing this sequence and producing diversity around it (Fig. 6.4). The per-cycle efficiency was varied between 25-75% in these runs, the number of sequences in each library was set to 10,000, and six assembly cycles were simulated for each run. The assembly block mixture used in each step was composed of 50% of the codon for the corresponding residue in the RYG target and 50% of the remaining 19 codons, with each position corresponding to two steps in the assembly (Fig. 6.4d). This library design strategy produces length diversity with a mean of three residues and frequency peaks of three and four residues when the per-cycle efficiency is 50% (Fig. 6.4b). The higher and lower efficiencies produce length distributions with peaks around two (Fig. 6.4a) and five residues (Fig. 6.4c), with corresponding shifts in the composition distributions.

The shifts in the length distributions for efficiencies higher or lower than 50% can be compensated by altering the library design. For efficiencies lower than 50%, a higher number of assembly steps per position in the target are needed to produce a library centered around the desired length (Fig. 6.5a). Similarly, decreasing the number of assembly steps and simultaneously creating overlapped biases in the assembly block mixtures towards adjacent residues in the target RYG sequence can produce similar results for efficiencies higher than 50% (Fig. 6.5b). However, both of these strategies produce asymmetries in the length distributions, with broadly-dispersed longer insertions (Fig. 6.5a) or a truncated distribution containing only shorter insertions (Fig. 6.5b), along with lower frequencies for any length that is not three or four in both runs.

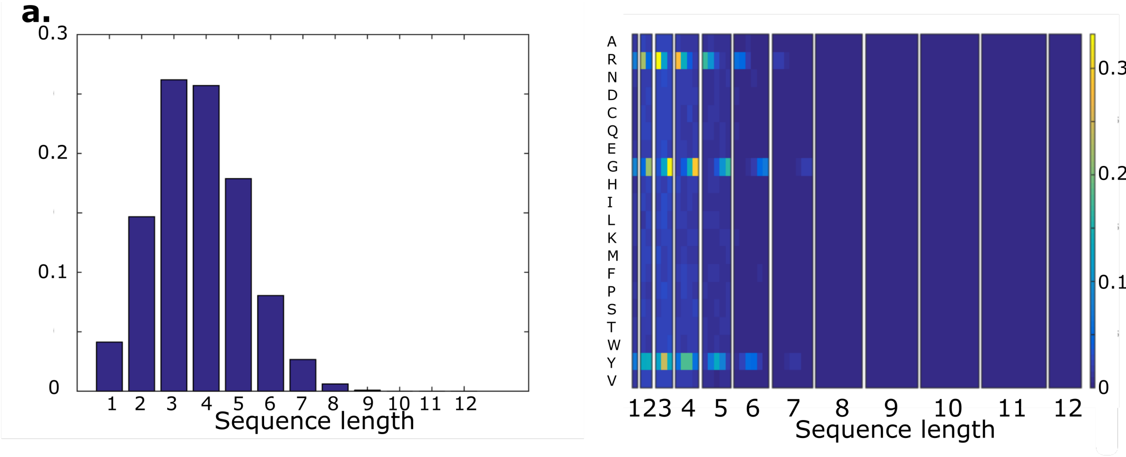
The simulations using 50% insertion efficiency per cycle demonstrate a simple mapping between number and composition of assembly cycles and the length and composition in the final library: every two assembly cycles will contribute mostly to one position in the final library, with indels and substitutions distributed symmetrically around the target sequence length and composition. If prior knowledge about the functional sequence space is available for a given target system, different efficiencies could also prove useful — especially if functional sequences are known to be more frequent at shorter or longer ends of the distribution. However, the complex mapping between assembly cycles and composition distribution observed when the efficiency was higher than 50% makes library design less

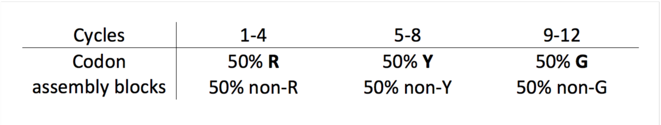


d.

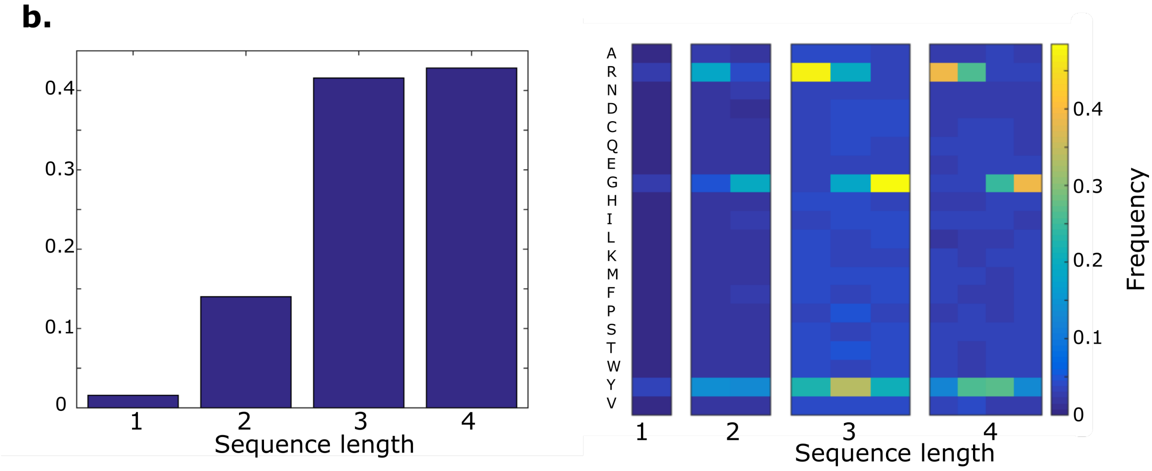
Cycles	1-2	3-4	5-6
Codon assembly blocks	50% **R** 50% non-R	50% **Y** 50% non-Y	50% **G** 50% non-G

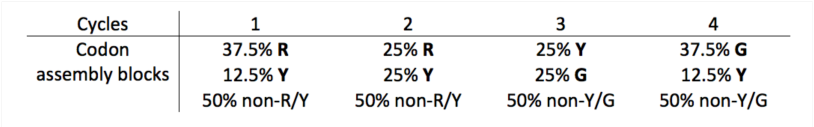
FIGURE 6.4: Simulation of InDel Assembly libraries with varying per-cycle efficiencies - Length (histograms) and composition distributions (heatmaps) for simulated InDel Assembly products with 25% (a), 50% (b), or 75%(c) insertion efficiency per cycle. The composition of the assembly block mixtures for each reaction cycle was kept constant for all simulation runs (d).





Cycles	1-4	5-8	9-12
Codon assembly blocks	50% **R** 50% non-R	50% **Y** 50% non-Y	50% **G** 50% non-G





Cycles	1	2	3	4
Codon assembly blocks	37.5% **R** 12.5% **Y** 50% non-R/Y	25% **R** 25% **Y** 50% non-R/Y	25% **Y** 25% **G** 50% non-Y/G	37.5% **G** 12.5% **Y** 50% non-Y/G

FIGURE 6.5: Simulation of InDel Assembly libraries with varying assembly block mixtures - Four cycles biased towards each residue in the target sequence were used in the 25% per-cycle insertion efficiency library (a). For the 75% efficiency simulation (b), only four cycles in total were used and the codon biases for each cycle encompassed adjacent codons in the RYG target sequence. Part of the composition heatmaps for 12-residue length in the 25% efficiency simulation were truncated for clarity.

tractable, requiring computational calculation of optimal assembly mixtures to efficiently target libraries.

6.2.2 Development of a strategy for the detection of length-variable enriched motifs

The idea of employing k-mer based analysis in the reconstruction of sequence motifs enriched by rounds of selection was proposed by the author of this thesis and the analysis framework was developed collaboratively in discussion with Dr. Vitor Pinheiro. MATLAB scripts used in simulated selections and the k-mer-based analysis were written by Dr. Vitor Pinheiro, with adaptations by the author to enable the analysis of data derived from NGS datasets.

As stated in Section 6.1.2, an analysis strategy was developed to detect selected length-variable motifs by analysis and reconstruction of enriched k-mers after rounds of selection in directed evolution experiments. Due to the short length of the targeted region in the Ω-loop of TEM-1 — only 5 residues in the wild-type and $_{164}$RYYGE$_{168}$ variant — a size of three residues was selected for subdivision of sequences in all the analyses described in this thesis. This value was chosen as a compromise in discriminative power for sequence reconstruction. For instance, repeats of length k or longer cannot be correctly reconstructed by k-mer based analysis alone and the known ceftazidime-active variant $_{164}$RYYGE$_{168}$ already has a Tyr repeat of length 2, preventing the recovery of this sequence and similar variants by an analysis based on 2-mers. On the other hand, longer k-mers can resolve longer repeats and make reconstruction easier by requiring fewer overlapping k-mer comparisons, but also reduce the number of k-mers that each sequence is subdivided into — a sequence of length 5 is only composed of two distinct 4-mers — and can make the analysis prone to failing to account for indels like alignment-based strategies.

A masking strategy was also adopted in the k-mer decomposition (Vinga and Almeida, 2003), in which one position of each k-mer was represented as a wildcard character "_" to reduce computational burden — there are 8000 possible 3-mers but only 800 possible masked 3-mers — and also to detect positions within functional motifs that allow variation — the enrichment of two 3-mers that only differ at one position would then be considered as enrichment of a single 3-mer masked at that position. In the masked 3-mer representation adopted here, the 3-mer $X_1X_2X_3$ would be represented as X_1X_2_ and X_1_X_3. Finally, the first non-diversified residues surrounding the targeted region in each library were represented as “Z” characters to allow identification of enriched motifs that spanned the entire region or at least reached one of its ends — also increasing the number of possible masked 3-mers to 882. The subdivision of an example sequence into its component masked 3-mers is shown in Fig. 6.6 a.

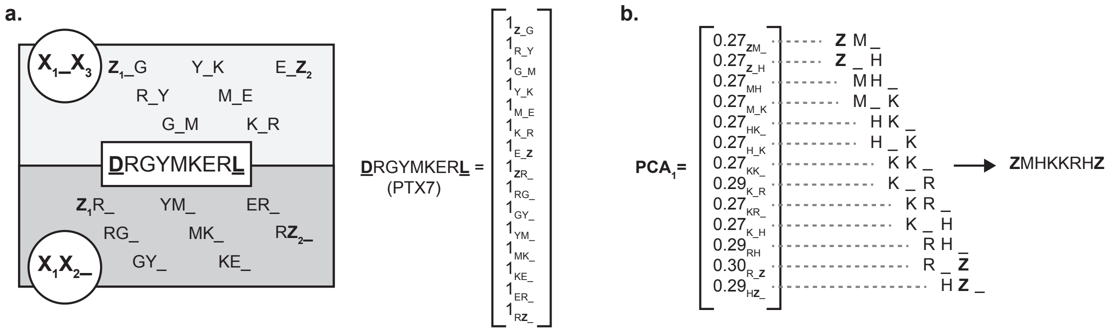


Figure 6.6: k-mer-based motif detection. (a) Sequences are subdivided into all possible masked 3-mers, as shown for the variable sequence RGYMKER (underlined residues represent fixed ends). The counts for each residue form an 882-dimension vector (zero values not shown for clarity). Vectors are normalised and multiplied by the enrichment score, then PCA identifies enriched kmers, allowing reconstruction of sequences. (b) Demonstration of sequence reconstruction from a set of enriched kmers. The sequence can be reconstructed by the overlapping kmers from the initial fixed “Z” character until the terminal residue. According to (Tizei et al., 2017)

Each sequence within a sequencing dataset (pre- or post-selection for each round of selection) was then represented as a column vector of 882 dimensions — one for each of the masked 3-mers, including the terminal “Z” characters — with masked 3-mer count as the value in each dimension (Fig. 6.6 a). Vectors were normalised by transformation into unit vectors and then multiplied by an enrichment score, calculated by subtracting the pre-selection count of the sequence by its post-selection count. The vectors for all the sequences in a round of selection represent it in an 882-dimensional space, with the magnitude of each vector as a proxy for its enrichment by selection and, consequently, its function.

Since an 882-dimensional dataset is too complex for direct analysis, principal component analysis (PCA) was used to identify which of these dimensions representing the masked 3-mers had the strongest contributions to the enrichment observed in a round of selection. Deconvolution by PCA maps the observed variation in the dataset onto a new set of coordinates in descending order of the fraction of total variance explained by each component (or dimension).

The masked 3-mers with the strongest contributions to each component compose the most highly enriched sequences that contributed to that component and, therefore, can be used to reconstruct shared motifs in those sequences. Importantly, enriched sequences that differ in length but contain a shared motif would cluster together in the 882-dimensional space and contribute to the same components in the PCA. An arbitrary contribution threshold of absolute value 0.1 (i.e. values higher than 0.1 or lower than -0.1 were considered) was selected for inclusion of masked 3-mers in further analysis. Motif reconstruction for each component was carried out by manual inspection of the masked 3-mers that met the selected threshold value to identify 3-mers that are part of larger motifs by their overlaps.

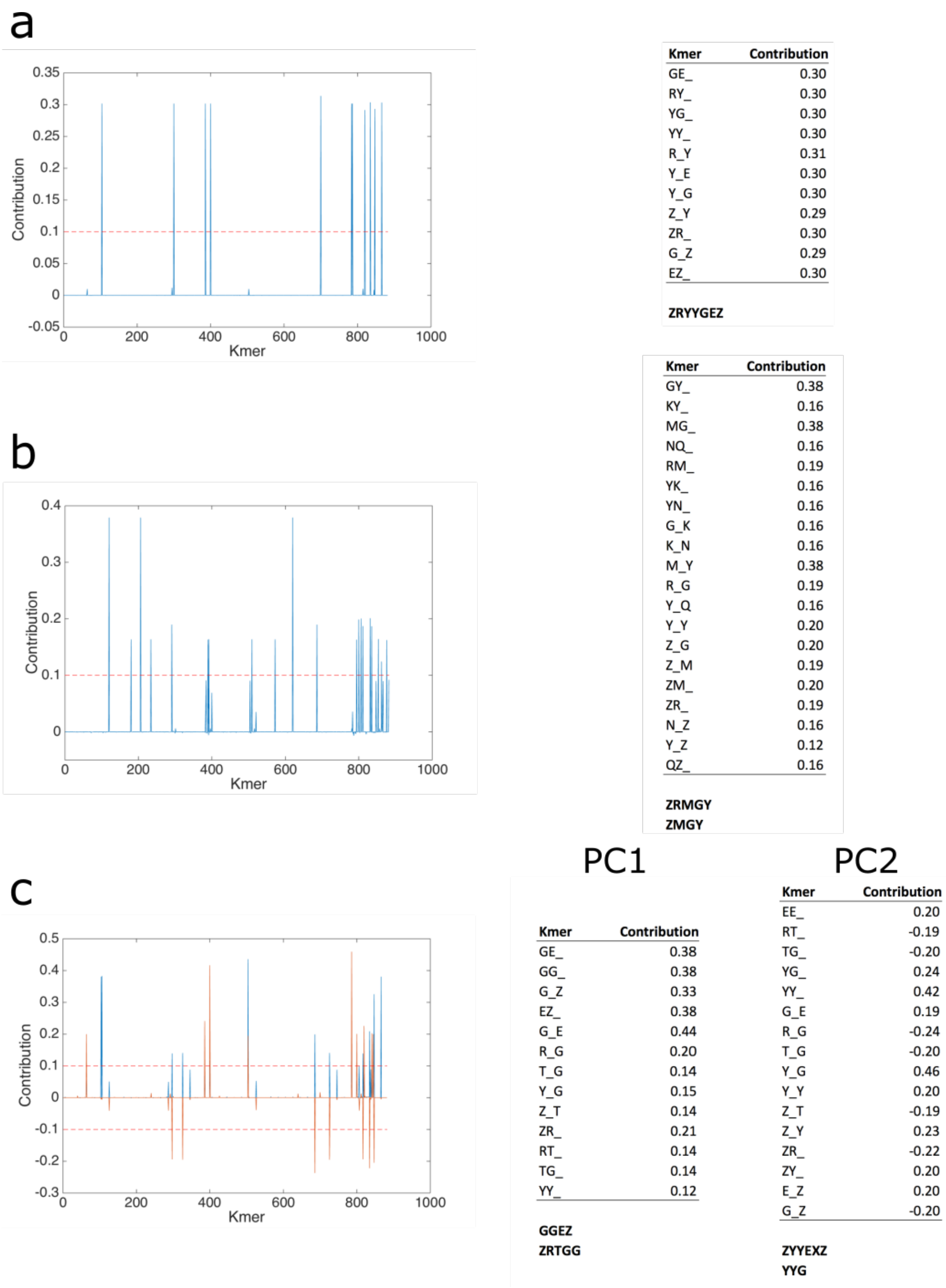
The motif detection workflow was tested with simulated selections against ten-cycle library assemblies biased towards the sequence "RYYGE" (as in Fig. 6.4 b). Selection was represented by multiplication of the frequency of any variant that exactly matches a chosen motif in any part of its sequence by a user-specified enrichment factor — with positive factors for enrichment of functional variants and negative factors for simulated depletion of deleterious mutations. Enrichment factors and the k-mer PCA were then calculated to determine whether selected motifs could be reconstructed.

Initially, selection of the RYYGE motif targeted by the library synthesis was attempted, by selecting this sequence with a 200-fold enrichment factor. As expected, the RYYGE sequence itself was the most highly enriched variant, followed by variations with one or more N- or C- terminal insertions, that all had lower pre-selection frequencies than RYYGE (data not shown). The first component of the PCA accounted for $> 96\%$ of all variation in the selected dataset and the sequence RYYGE could be clearly reconstructed from the 3-mer contributions in that component (Fig. 6.7 a). This demonstrated the ability of the motif detection strategy to detect enrichment of a sequence that was present in relatively high frequency ($> 1\%$) in the initial library.

Next, motif detection was attempted for a sequence that was not part of the RYYGE assembly target, to determine whether lower frequency variants enriched out of a background of higher frequency variants could still be recovered. The motif MGY was enriched 200-fold in this test and it was represented by the masked 3-mers with the highest contributions of the first principal component, but many other 3-mers not matching the selected motif were also present (Fig. 6.7 b). Since this component only accounted for 44% of the variation contained in this dataset, the selected 3-mers in the second and third components were also inspected (82% of the total variation in the first three components). Similar patterns were visible in these components, with the MGY motif (or parts of it) detectable and other unrelated 3-mers.

The spurious 3-mers detected by the PCA in this selection could be a result of poor sampling of the sequence space, due to all the enriched sequences having lower frequency values in the initial dataset than in the previous example — the highest initial frequencies among the enriched variants were 0.003% for variants MGYGE and MGY, which represent only 3 counts for each. Repetition of the simulated assembly, selection and analysis corroborated this obervation, producing a clear signal for enrichment of the 3-mers that make up the MGY motif and disperse contributions from unrelated 3-mers.

A third simulated selection was carried out with two separate enrichment patterns: 200-fold enrichment of GYY-containing sequences and 200-fold depletion of the motif GGE



Kmer	Contribution
GE_	0.30
RY_	0.30
YG_	0.30
YY_	0.30
R_Y	0.31
Y_E	0.30
Y_G	0.30
Z_Y	0.29
ZR_	0.30
G_Z	0.29
EZ_	0.30

ZRYYGEZ

Kmer	Contribution
GY_	0.38
KY_	0.16
MG_	0.38
NQ_	0.16
RM_	0.19
YK_	0.16
YN_	0.16
G_K	0.16
K_N	0.16
M_Y	0.38
R_G	0.19
Y_Q	0.16
Y_Y	0.20
Z_G	0.20
Z_M	0.19
ZM_	0.20
ZR_	0.19
N_Z	0.16
Y_Z	0.12
QZ_	0.16

ZRMGY
ZMGY

PC1

Kmer	Contribution
GE_	0.38
GG_	0.38
G_Z	0.33
EZ_	0.38
G_E	0.44
R_G	0.20
T_G	0.14
Y_G	0.15
Z_T	0.14
ZR_	0.21
RT_	0.14
TG_	0.14
YY_	0.12

GGEZ
ZRTGG

PC2

Kmer	Contribution
EE_	0.20
RT_	-0.19
TG_	-0.20
YG_	0.24
YY_	0.42
G_E	0.19
R_G	-0.24
T_G	-0.20
Y_G	0.46
Y_Y	0.20
Z_T	-0.19
Z_Y	0.23
ZR_	-0.22
ZY_	0.20
E_Z	0.20
G_Z	-0.20

ZYYEXZ
YYG

FIGURE 6.7: Motif detection in simulated InDel libraries. The plots show the contributions of all k-mers to each principal component (PC1 in blue and PC2 in orange). The 0.1 cutoff for analysis of scores is indicated by the red dashed line. Contributions of the selected kmers are listed in the tables, with reconstructed motifs for each principal component written in bold under each table. (a) First principal component for the 200-fold enrichment of RYYGE motif, accounting for 96% of the variation in the simulation. (b) First principal component for the 200-fold enrichment of the MGY motif, accounting for 44% of the variation in the simulation. (c) First (blue) and second (orange) principal components for the 200-fold enrichment of GYY motif and 200-fold depletion of GGE motif, accounting for 57% and 18% of the variation in the simulation, respectively. The Z character denotes the fixed regions flanking the sequence diversified in the simulated library.

(Fig. 6.7 c). The depleted motif was detected in the first component (accounting for 57% of variation) and the presence of the C-terminal “Z” fixed region character associated with it indicates it was mainly in the C-terminus of the enriched variants. Given the RYYGE target sequence for the assembly, a GGE sequence would be expected to be more frequent at the C-terminus of the variable region due to the biased GE positions, requiring only a single additional insertion of a Gly codon to produce the depleted motif. On the other hand, the enriched GYY motif could not be reconstructed in either of the next two components of the PCA that accounted for $> 85\%$ of the variation (only first and second components shown in Fig 6.7 c for clarity). The initial frequencies for the enriched sequences were approximately 20% of the frequencies for the depleted variants, which could be the cause for no clear enrichment signal being detected.

These simulations demonstrated that the masked 3-mer PCA can recover selected motifs from enrichment patterns in libraries, but the results obtained are strongly affected by the initial composition of the library. Another confounding factor in these simulations is the artificial nature of the enrichment criteria — variant counts were multiplied by a fixed factor to represent enrichment or depletion and the context of the motif within the remaining variant sequence was completely ignored. Therefore, further investigation into the capabilities of the detection strategy were done once data from selections of TEM-1 variants for activity against ceftazidime became available (See Sections 6.2.6.2 and 6.2.6.3).

6.2.3 InDel Assembly protocol development and optimization

Initial versions of the InDel Assembly protocol will be described here as changes were made to the method to improve its performance. The final, optimised version of the method was described in Section 2.2.11.

Template oligos were synthesised with a biotin and $(PEG)_4$ spacer at their $5'$ end and also harboured a fluorescein label in a portion of the sequence that is not recognised by the restriction endonuclease — to ensure the labeled base did not interfere with restriction enzyme recognition (Fig 6.8). The assembly blocks were single-stranded oligonucleotides designed to fold into a hairpin with a degenerate NNN overhang at their 5′ end, which were heat-cool annealed to ensure formation of the correct secondary structure before 5′ phosphorylation with T4 PNK. This design was preferred since an assembly block composed of two separate single-stranded oligos would require additional handling steps to ensure that only the the correct 5′ end of the duplex is phosphorylated or costly custom synthesis options, such as a 5′ phosphorylated base in one strand or a blocked 5′ in the complementary strand.

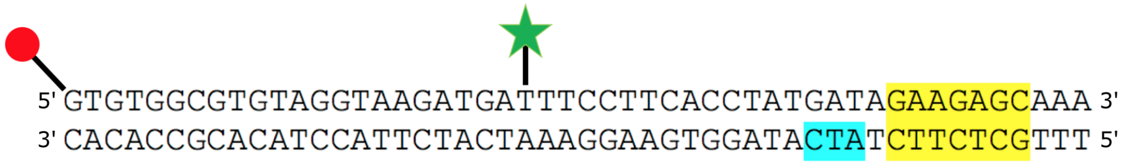

5' GTGTGGCGTGTAGGTAAGATGATTTCCTTCACCTATGATAGAAGAGCAAA 3'
3' CACACCGCACATCCATTCTACTAAAGGAAGTGGATACTATCTTCTCGTTT 5'

FIGURE 6.8: Design of template block used for optimisation of InDel Assembly - 5′ biotin TEG group is represented as a red circle, the fluorescein label is represented as a green star, the recognition site for SapI is highlighted in yellow and the region left as a 5′ overhang in the complementary strand after restriction is highlighted in cyan. The complementary strand (bottom) contains no modifications.

The initial conditions tested in development of the InDel Assembly protocol were based on prior experience of the Pinheiro group with analogous restriction and ligation reactions in solution, as well as other protocols employing biotinylated oligos captured on paramagnetic streptavidin-coated beads (Pinheiro et al., 2012b). A saturating amount of template oligos (50 pmol) was added to 5 µl prewashed MyOneC1 beads (maximal binding capacity is 5 pmol oligos / µl resuspended bead slurry according to the manufacturer). Digestion was carried out with 10U SapI for 10 minutes at 37°C and ligation was carried out with 400U T4 DNA Ligase for 5 minutes at 25°C. The reaction volume was 100 µl for both steps and the buffers supplied by the manufacturer were used. Assembly was carried out for three full cycles of digestion, annealing, and ligation, removing a 1 µl aliquot from the bead suspension after every step. Reaction efficiency was measured by denaturing electrophoresis in polyacrylamide gels (denaturing PAGE) coupled to fluorescence scanning to detect extension of the fluorescein-labeled template strand.

The efficiency of both enzymatic steps in the assembly was very low under the initial conditions (Fig. 6.9 a). A high gain setting in the fluorescence detector was needed for visualisation of any ligated bands, which led to saturation of the unligated bands and made accurate quantitation of the bands unreliable. However, it is clear that less than 50% of the initial template was digested in each cycle and only a very small portion of the digested templates were ligated in each cycle (Fig. 6.9 a). Since these conditions produced insertion yields very far below the desired 50% value, conditions for the assembly reactions were optimised, starting with the restriction reaction.

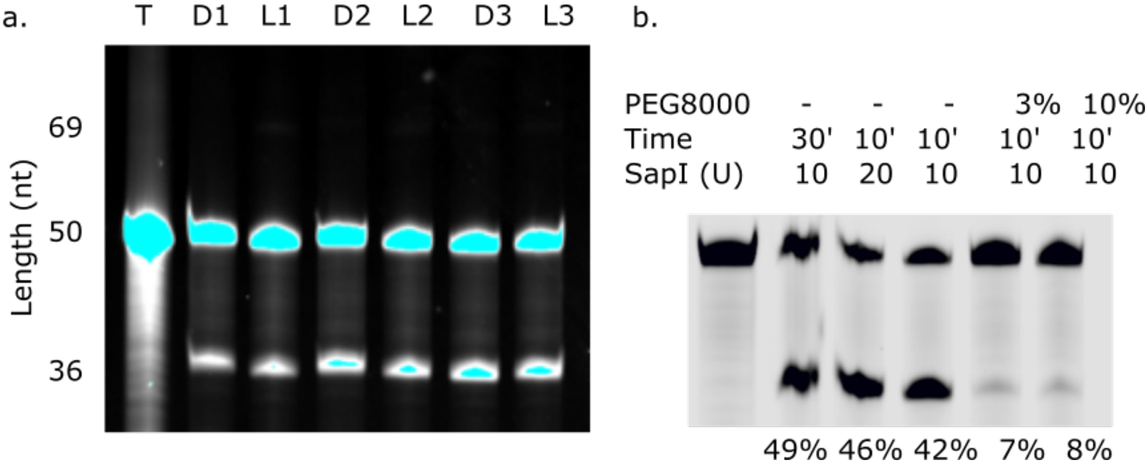


FIGURE 6.9: Optimisation of the digestion step in InDel Assembly. (a) Three full cycles of InDel Assembly carried out using the initially proposed reaction conditions. Lane T contains only the template oligo. DX and LX represent the output of the digestion and ligation steps, respectively, of each assembly cycle. Saturated bands in the image prevented accurate quantitation of reaction products (b) SapI digestion of the template oligo under varying reaction conditions. Digestion efficiencies are under each lane and were calculated as the ratio between the intensity of the digested band and the sum of the digested and undigested bands. The template oligo was loaded in the first lane on the left. The template oligo is 50 nt, the first digestion product is expected to be 36 nt and the first ligation product is expected to be 69 nt.

The hairpin design for the assembly block oligos was abandoned before this experiment and replaced by two annealed oligos with an overlap of twenty base pairs after the SapI recognition site (represented in Fig. 6.1), to ensure efficient binding of the enzyme to extended oligos. Changes in reaction time, enzyme concentration and concentration of the crowding agent PEG-8000 were evaluated for their effect on the SapI reaction step. Aliquots from each reaction were loaded on a denaturing PAGE and efficiency was measured by band densitometry using ImageJ. Both the increase in reaction time and enzyme concentration had an effect on the efficiency of digestion (Fig. 6.9 b), so these conditions were combined and subsequent development of the assembly used 2-hour digestions with 20U SapI / µl in a reaction volume of 100 µl. The efficiency obtained in these conditions was deemed sufficient to start the optimisation of the ligation step.

Digestion was carried out as described above to produce the overhang needed for the ligation step, with the exception of a digestion time of 2 hours aiming to increase the digestion efficiency further, since a longer reaction time increased efficiency in the digestion test. Unlike the observed inhibition of SapI digestion by PEG-8000, this crowding agent is known to increase the activity of T4 DNA ligase (Sambrook and Russell, 2001), so it was added to a final concentration of 3% in the ligation mixture. A time course reaction was carried out by periodically removing aliquots from the same reaction tube incubated at 25°C. As expected the intensity of the ligated band increased over time, with a peak of approximately 40% ligation efficiency after 20.5 hours of reaction time (Fig. 6.10). Since more than 20% of the template had been ligated after only one hour of reaction time, this time point was selected for further development as a compromise between ligation efficiency and the number of InDel cycles that could be carried out within a single day. An unexpected doublet band was observed for the ligated product in this reaction (Fig. 6.10 a), with both bands in the doublet increasing simultaneously in intensity. This was interpreted as blunt ligation of a second assembly block oligo at the $3'$ end of the growing chain, which would be removed by the restriction enzyme digestion.

As shown in Fig. 6.4, a ligation efficiency of only 20% (close to 25% shown in the simulation) would require four three-hour long assembly cycles to be carried out for every

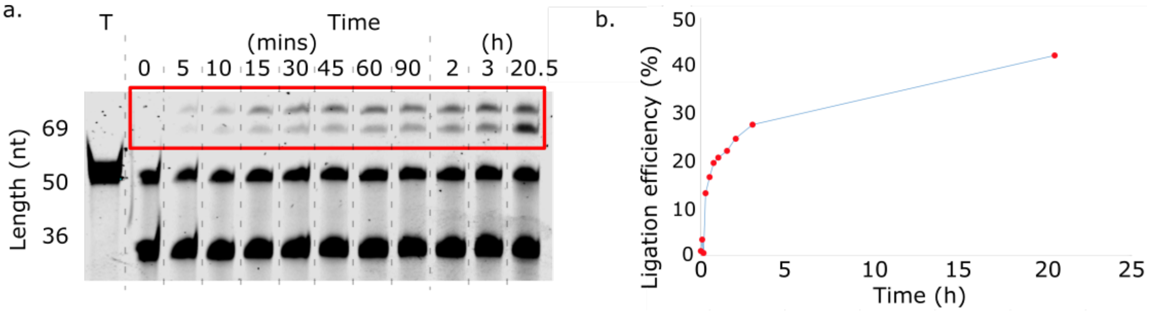


FIGURE 6.10: Optimisation of the ligation step in InDel Assembly. (a) Denaturing PAGE of a ligation timecourse. The intensity of both bands within the red box were added together for ligation efficiency measurements for each time point. (b) Plot of ligation efficiencies at each time point.

unit increase in mean sequence length and would also create asymetric length distributions. Therefore, the published and patent literature were searched for additional conditions that could enhance ligation activity and obtain efficiencies closer to the 50% target. Patents from New England Biolabs report increased efficiency for T4 DNA ligase in the presence of the small molecule 1,2-propanediol (Kucera and Evans, 2014) and the accessory enzyme 5′ deadenylase, which removes adenylated side products of T4 DNA ligase activity that inhibit ligation (Lohman and Nichols, 2015). A full cycle of InDel Assembly was carried out using the previous conditions, 12% 1,2-propanediol and 0.5 U/µl 5′ deadenylase. These conditions led to approximately 50% insertion yield after one full assembly cycle (Fig. 6.11). Therefore, these conditions were used for the production of the first sequence libraries for directed evolution of the TEM-1 Ω-loop.

6.2.4 Establishing the selection system for TEM-1 variants

Since pBAD-30 already has a copy of TEM-1 as its transformation marker, this was replaced by the CAT (chloramphenicol acetyltransferase) chloramphenicol resistance cassette and a new copy of TEM-1 was inserted downstream of the pBAD promoter by Type IIS restriction enzyme-based cloning, producing vector pBT1-WT 6.12. The 165-YYG-167 variant from the Palzkill group (Petrosino and Palzkill, 1996) was made by inverse PCR of pBT1 with oligos containing the mutations from the wild-type sequence WEP to YYG, which was

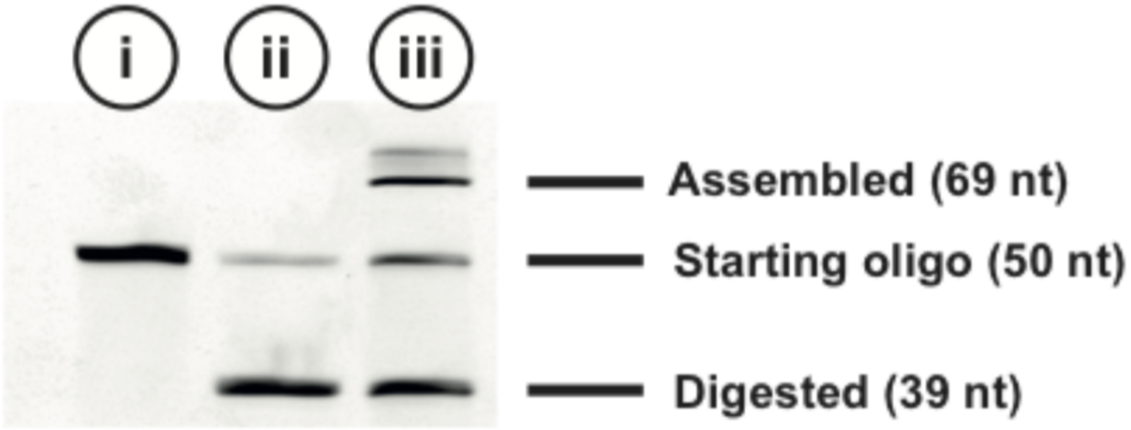


FIGURE 6.11: InDel assembly reaction in optimised conditions. (i) Starting template oligo. (ii)Products of digestion step. (iii) Products of Ligation step. According to (Tizei et al., 2017).

blunt-ligated to produce pBT1-RYYGE. In addition, the M182T stabilising mutation was added by the same procedure to all mutant versions of the TEM-1 protein, since it is established in the literature that mutations that increase activity in non-cognate substrates tend to be destabilising and M182T can compensate for some of this effect (Sideraki et al., 2001). This stabilisation would lead to higher levels of folded and active enzyme per cell and, therefore, the $_{164}$RYYGE$_{168}$ variant expressed by this construct would be expected to have a MIC towards ceftazidime higher than 64 µg/ml that was observed by Petrosino and Palzkill (1996).

These vectors should confer cells with constitutive resistance to chloramphenicol and arabinose-inducible resistance to AMP (pBT1-WT) or ceftazidime (pBT1-RYYGE). Since a standardised method for measurement of MICs had not yet been established in this work, comparative tests were carried out between WT and $_{164}$RYYGE$_{168}$ variants to establish culture conditions that would differentiate variants active on CAZ from those with activity against this drug similar to WT. Images were not recorded for display of these results, so only growth comparisons will be described here.

Plating both constructs on media containing the corresponding antibiotic (100 µg/ml AMP for WT and 10 µg/ml CAZ for $_{164}$RYYGE$_{168}$) in presence or absence of 0.1% (w/v)

L-arabinose inducer produced robust growth when the construct was induced but growth was also visible in the absence of the inducer, albeit with reduced numbers of colonies. No growth was observed when the WT construct was plated on CAZ plates or when the ${}_{164}$RYYGE${}_{168}$ construct was plated on AMP plates, repeating the respective sensitivity towards ampicillin and ceftazidime of the ${}_{164}$RYYGE${}_{168}$ variant and WT, observed by Petrosino and Palzkill (1996). However, variability in colony sizes was visible when cells were plated under inducing conditions (data not shown), suggesting the existence of cell-to-cell expression heterogeneity of the pBAD promoter in these constructs, which has been previously characterised (Siegele and Hu, 1997).

Therefore, a new construct was made replacing the resistance marker of pUC19 with CAT and inserting the TEM-1 coding sequence under control of its own promoter by Type IIS restriction enzyme-based cloning (Fig. 6.12). As expected, strains carrying these vectors with the WT (pTEM1-WT) and ${}_{164}$RYYGE${}_{168}$ (pTEM1-RYYGE) variants were constitutively resistant to their corresponding antibiotics, but sensitive to ceftazidime (for the WT construct) or ampicillin (for the ${}_{164}$RYYGE${}_{168}$ construct). Since the construct and strain used were distinct from the system constructed by Petrosino and Palzkill (1996), the ${}_{164}$RYYGE${}_{168}$ construct was plated on media containing increasing concentrations of ceftazidime up to 500 µg/ml ceftazidime and growth was observed at 100 µg/ml ceftazidime but not at 200 µg/ml ceftazidime, indicating that the MIC of this variant for ceftazidime was between these two values and much higher than the value reported by Petrosino and Palzkill (1996). These results demonstrated that the pTEM1 constructs are robust for selection of variants and, therefore, these vectors were used as templates for library assembly, and as negative (WT) and positive (${}_{164}$RYYGE${}_{168}$) controls for ceftazidime resistance.

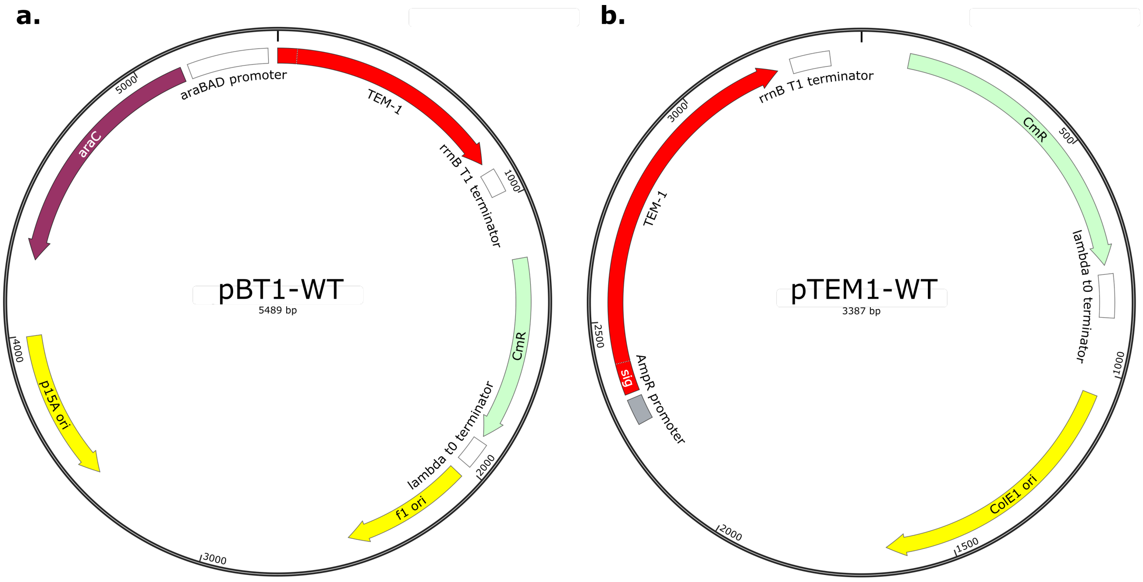


FIGURE 6.12: Vectors used for expression of TEM-1 during development of InDel assembly and rounds of selection. (a) First design of the TEM-1 expression vector with the TEM-1 gene under control of the pBAD arabinose-inducible promoter. (b) TEM-1 expression vector used for all selection experiments, with TEM-1 under the control of its own promoter. Both vectors also carry the chloramphenicol resistance marker that was used for cloning.

6.2.5 Directed evolution of TEM-1 using libraries generated by InDel Assembly

6.2.5.1 First round of diversification and selection

The goal for the first round TEM-1 directed evolution was to demonstrate that libraries produced by InDel could effectively explore the sequence space around a target sequence, with simultaneous variation in length and composition. The library for this round was designed to contain the $_{164}$RYYGE$_{168}$ variant at a frequency of approximately 0.1% and ensure its recovery as a known positive variant, while at the same time exploring the sequence space — in both length and composition — around this peak. This strategy also aimed to determine whether $_{164}$RYYGE$_{168}$ is an isolated fitness peak in its sequence neighbourhood or whether other similar sequences could produce similar or even higher levels of activity on ceftazidime. The targeted region was expanded to contain residues Arg164 and Glu168, since deletions and mutations at these residues had also been previously shown to produce strains with higher MIC than WT on ceftazidime (Palzkill et al., 1994a;

Sowek et al., 1991). To achieve this, the library was assembled in ten InDel cycles with the optimised reaction conditions (from Section 6.2.3) to produce a distribution of sequence lengths centred around the wild-type length of five residues. In each cycle, a mixture of codon building blocks composed of 50% of the codon present in the equivalent position in the $_{164}$RYYGE$_{168}$ variant and 50% of the remaining 19 codons was added in the annealing step.

Once the library was assembled and transformed into *E. coli* NEB 10-β (see Section 2.2.6), cells were plated into agar plates containing 34 µg/ml chloramphenicol as a transformation control and 10-200 µg/ml CAZ. Colony density was visually estimated, with near-confluent growth at 10 µg/ml CAZ, fewer than 20 colonies at 50 µg/ml CAZ and none on the higher concentrations — suggesting that the assembled library had a high proportion of low-activity mutants. The variants from three arbitrarily selected colonies from the 50 µg/ml plate were sequenced and all three showed no insertions in the targeted loop, indicating that only the final cap block had been ligated onto these variants after ten cycles of assembly. These results suggested that the measured per-cycle assembly efficiency from the method optimisation (Fig. 6.4 b) was overestimated since — according to simulations (see Section 6.2.1) — the majority of the assembled sequences were expected to contain five or more residues and the $_{164}$RYYGE$_{168}$ variant should have been recovered in the plates containing 50 µg/ml ceftazidime, as it was expected to represent approximately 0.1% of all assembled sequences. A possible explanation for this large discrepancy is that test assemblies during optimisation were only carried out for a single cycle and did not account for factors that could reduce reaction efficiencies over multiple assembly rounds, such as carryover of enzymes and digested oligo fragments into subsequent reaction steps or loss of enzyme activity after being reused for multiple cycles within a day. Another possibility is that saturating concentrations of oligos bound to the bead surface can hinder the accessibility to enzymes and assembly block fragments, reducing the maximum efficiency that can be obtained per step.

The three identical variants are effectively full-deletions of the five targeted residues in the Ω-loop and were more resistant to ceftazidime than the WT. Since this variant was not isolated in previous publications of TEM-1 variants, one of the clones was named Δ5 and stored for later characterisation along with other variants isolated in later selection steps.

A new assembly was made following the strategy described above, while reducing the template oligo bound to the streptavidin beads to 10 pmol. This parameter was changed as an attempt to reduce the proportion of oligos that were not ligated to any codon building blocks during all the assembly cycles (i.e. increase the overall per-cycle efficiency) and, ultimately, obtain a library of higher quality with a length distribution closer to what was predicted in simulations (Fig. 6.4). Once this new library was assembled, it was transformed into NEB 10-β, cells were plated into 24.5 cm x 24.5 cm LB agar plates containing 50 µg/ml ceftazidime for selection and incubated overnight at 37°C.

This selection yielded more than one thousand colonies, which were harvested and grown in liquid media containing the same concentration of ceftazidime until visibly turbid (OD600 of approximately 1), to amplify the selected transformants and further enrich for the most highly active variants. Plasmids were extracted from this culture for later deep sequencing and an aliquot was frozen in glycerol at -80°C. Approximately 10^4 CFU of this culture were plated onto LB agar plates containing increasing concentrations of ceftazidime up to 500 µg/ml. The ceftazidime concentrations above the estimated MIC for $_{164}$RYYGE$_{168}$ (between 100 and 200 µg/ml) represent a selection step of sufficient stringency to remove this variant from the selected pool, allowing only variants that confer higher MICs to be recovered.

A single transformant was isolated in the plate containing 300 µg/ml ceftazidime and sequenced, which showed that its Ω-loop had the sequence $_{164}$RGYMKER$_{168\text{b}}$ (named

PTX7) replacing the target $_{164}$RYYGE$_{168}$ sequence. In a single InDel diversification and high-stringency selection cycle, a variant with apparently higher fitness on ceftazidime was obtained that is two residues longer than the length of the wild-type sequence and the $_{164}$RYYGE$_{168}$ variant.

6.2.5.2 Control library

In the same assembly round that produced the succesful $_{164}$RYYGE$_{168}$ diversification library (see Section 6.2.5.1), a small control library was generated in parallel to detect possible assembly biases — the assembly simulations in section 6.2.1 all assumed unbiased selection of codons for incorporation — and also obtain an independent measure of assembly efficiency. To achieve this, five assembly cycles were carried out with no overlap between the building blocks used in each cycle. The building blocks in each cycle were always mixed in equimolar amounts and the mixtures had the composition shown in Table 6.2. At the end of the process, the assembly products were amplified and stored for later deep sequencing along with the selection libraries, to characterise the population of assembled products in detail.

Cycle	Residue	Codon
1	Ala	GCA
2	Cys	TGC
	Asp	GAC
3	Glu	GAG
	Phe	TTC
4	Gly	GGT
	His	CAC
	Leu	CTG
	Pro	CCA
5	Gln	CAG

Table 6.2: Composition of assembly block mixtures in the control library

An aliquot of the amplified products were cloned in a small-scale transformation and isolated in the absence of selection for TEM-1 function — by plating on media containing chloramphenicol. Three colonies were arbitrarily selected from this transformation and sequenced to determine the number of successful building block insertions in each assembly product. Two of the products had sequences corresponding to two insertions and the other sequence contained no building blocks inserted between the template and cap blocks. While a χ^2 goodness-of-fit test for this small sample did not reject the null hypothesis that the observed values came from a population with the expected mean of 2.5 insertions for a five-cycle assembly, further improvement to the assembly efficiency was sought in the assembly step for the second round of directed evolution targeting the Ω-loop.

6.2.5.3 Second round of diversification and selection

Since the PTX7 variant is divergent from $_{164}$RYYGE$_{168}$ in sequence and length, another round of directed evolution was carried out to explore the sequence neighbourhood around the PTX7 variant isolated in the first round and determine whether there are more undiscovered peaks of activity against ceftazidime in the Ω-loop sequence space. A more aggressive diversification strategy was used to maximise the breadth of the sequence space covered by this new library around the apparent peak obtained after selecting the first-round library. Since the sequence of PTX7 ($_{164}$RGYMKER$_{168b}$) had an N-terminal Arg residue and the variants with the highest activity against ceftazidime from Palzkill et al. (1994b) and Petrosino and Palzkill (1996) all had an N-terminal Arg, this residue was kept constant in the second round libraries.

The remaining six residues were targeted for variation in the form of insertions, deletions and substitutions at each position in the PTX7 sequence. The volume of streptavidin bead slurry used to capture template oligos in the beginning of the assembly was increased from 5 µl to 60µl, in an attempt to further increase the efficiency of restriction and ligation reactions by reducing crowding effects that could have occurred at bead surfaces in the conditions used for the previous assembly rounds. The assembly strategy was composed of thirteen cycles with the following sequence of building block mixture additions

XG*XY*XM*XK*XE*XR*X, where X represents codons for all 20 aminoacids in equal frequencies (100%X) and letters followed by asterisks represent the corresponding codon at 50% frequency and the remaining 19 codons in equal frequencies (X*). Assembly simulations (assuming that per-cycle efficiency reaches 50% with the change in reaction conditions) carried out for this strategy in comparison to the one used previously show that the fully-degenerate cycles increase the coverage for residues not present in PTX7, which was the objective for this round of diversification (Fig. 6.13). The PTX7 variant should represent approximately 0.1% of all assembly products when the diversification strategy of the first round is used in simulation (Fig. 6.13 a) and only around 0.01% of all sequences produced under the more aggressive assembly strategy (Fig. 6.13 b).

This library was transformed into NEB 10-β and selected on 200 µg/ml ceftazidime, yielding approximately 80 colonies, which were grown in LB containing 200 µg/ml ceftazidime for plasmid extraction and selection at higher stringencies to isolate any highly-active clones. At 500 µg/ml ceftazidime, a single colony was obtained and its Ω-loop was sequenced. This variant had the sequence $_{164}\underline{\text{R}}\text{GYKEERD}_{168c}$ (the underlined residue is the N-terminal Arg that was removed from the diversified region in this round) and was named PTX8.

This novel variant was longer than both $_{164}\text{RYYGE}_{168}$ and the best clone obtained from the previous library (PTX7), but it also diverged from both sequences in composition. Increasing selection stringency revealed the existence of another fitness peak — apparently with even higher activity on ceftazidime than $_{164}\text{RYYGE}_{168}$ and PTX7, since neither of these variants have been observed to form colonies at 500 µg/ml ceftazidime. The only sequences longer than PTX8 that have been explored in TEM-1 directed evolution were produced in a pentapeptide insertion experiment, yielding variants with higher activity than WT on ceftazidime, but much lower than the activity shown by $_{164}\text{RYYGE}_{168}$ (Hayes et al., 1997).

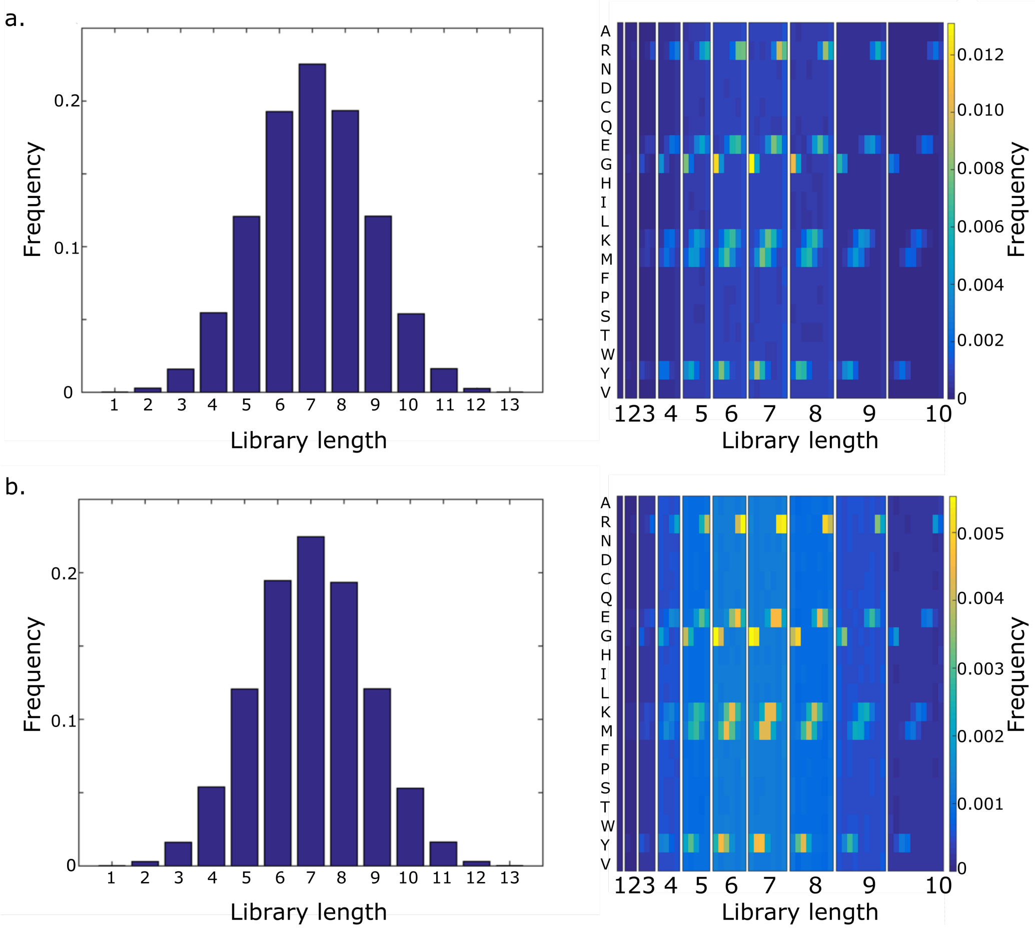


FIGURE 6.13: Simulation of two strategies for assembly libraries for the second of selection - The sequence distributions of library simulations using two different diversification strategies for thirteen assembly cycles are shown. (a) Pairs of consecutive assembly cycles with 50% of their building blocks represented by each position in PTX7, followed by a final cycle 100% X. (b) Alternating cycles of 100%X and X*. The composition heatmaps for lengths of eleven and twelve residues were removed for clarity.

To obtain PTX8 ($_{164}\underline{\text{R}}$GYKEERD^{168c}) from PTX7 ($_{164}$RGYMKER$_{168b}$) without a library made by InDel Assembly — using single steps of indels or substitutions — would require: deletion of the Met residue from PTX7, followed by insertion of a second Glu residue following the one already present and finally insertion of the C-terminal Asp residue for a total of 3 mutational events separating these two variants with high activity on ceftazidime.

6.2.6 Deep sequencing of libraries produced by InDel Assembly

Samples from the control library and both rounds of TEM-1 directed evolution (before and after selection) were sequenced on an Illumina MiSeq instrument at the UCL Institute of

Child Health genomics facility and the resulting datasets were processed according to the steps described in Section 2.6.5. The analysis for each dataset will be described in the following sections.

6.2.6.1 Control library

Quantification of the assembled sequences from the control library clearly show that the per-cycle efficiency estimated by band densitometry analysis was overestimated (Fig. 6.14). Over the five assembly cycles of the control library, an average efficiency of only 3.7% was obtained, producing the length distribution in Fig. 6.14. This result corroborates the non-significant trend towards insertion frequencies lower than the expected 50% value that was observed by Sanger sequencing of three clones from this library, prior to deep sequencing (Section 6.2.5.2)

Some insertion bias was detectable in this dataset (Table 6.3), but it cannot be solely attributed to differences in GC content producing variable binding affinities in the 3-base overhangs using the available data. In the only cycle which had a mixture of codons of varying GC content (third cycle), there was a near threefold excess of the codon with two GC bases (GAG for glutamate), but similar variations were also observed in the fourth cycle which only had codons with two GC bases. The codons from the second cycle in the control library had markedly lower counts than all others, possibly due to improper resuspension of beads or other manipulation error during the assembly. Since this dataset is derived from a single assembly run, its results are sufficient to show that different codons do get incorporated with varying efficiencies, but it is not possible to quantitate these biases reliably. More accurate quantitation of incorporation bias could be obtained by sequencing of multiple independent assembly reactions containing all twenty codons in equal proportions, to determine their relative incorporation efficiencies and whether there are any

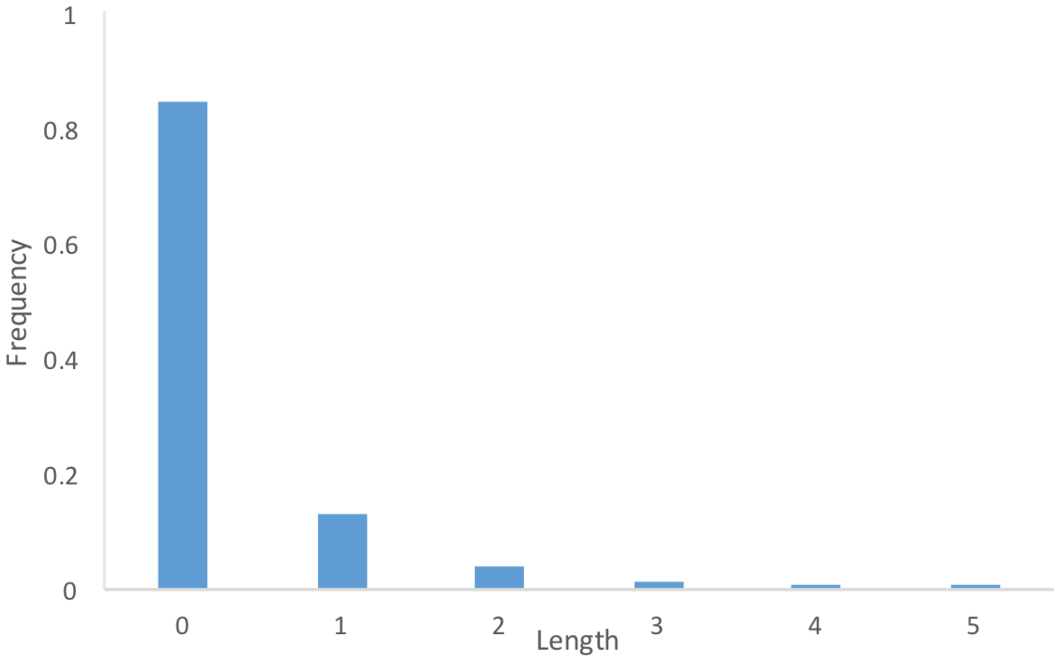


Figure 6.14: Length distribution of the control library assembled products

Cycle	Codon (residue)	GC bases	Count	Fraction	Bias[1]
1	GCA (A)	2	4758	1	-
2	TGC (C)	2	193	0.29	-
	GAC (D)	2	472	0.71	2.4
3	GAG (E)	2	3158	0.74	2.9
	TTC (F)	1	1083	0.26	-
4	CGT (G)	2	899	0.11	-
	CAC (H)	2	1596	0.20	1.8
	CTG (L)	2	2656	0.34	3.0
	CCA (P)	2	2687	0.34	3.0
5	CAG (Q)	2	4547	1	-

[1] Bias was defined as the ratio of counts of a given codon to the codon with the lowest frequency in the same assembly step

Table 6.3: Estimates of insertion bias from the control library

effects between adjacent codons caused by the annealing of degenerate NNN overhangs to the previous codon in each step.

The pre-selection datasets from the first and second directed evolution rounds could also be used to estimate insertion biases for different codon assembly blocks. The expected frequencies of each residue in both TEM-1 libraries were calculated from the frequencies of codon blocks in each assembly step and compared with the frequencies observed in the NGS data. It is clear that there was divergence between these values, with some residues being over-represented in one library while others seemed to be incorporated at a lower efficiency (Fig. 6.15 a, c). However, when these divergences are compared between libraries, it is evident that several of the trends observed in one round are inverted in the other round, which would indicate random variation around around a mean value. This is more evident when the data is analysed in the form of the observed bias for each residue, defined as $(observed - predicted)/predicted$ (Fig. 6.15 b, e). These values show that all of the strongest positive or negative biases present in the first-round library are inverted to the opposite signal or greatly reduced (i.e. Serine maintains a positive bias of approximately 0.5, while Glutamate changes from a negative bias to a strong positive bias). While more replicate experiments would be needed to estimate the bias toward any specific residue accurately, the data shown here indicate systematic insertion bias does not occur.

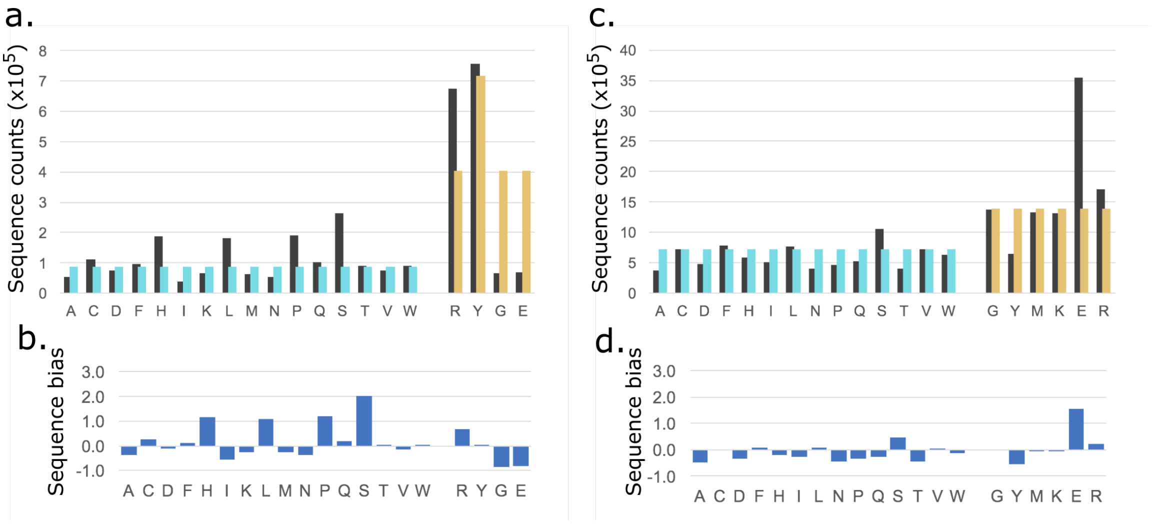


FIGURE 6.15: Estimates of insertion bias in pre-selection (a,b) first- and (c,d) second-round libraries. (a,c) Predicted and observed counts for each residue. Residues added only in equimolar proportions as part of degeneracies at each site are labeled in cyan, while residues targeted in each assembly are shown in orange. Observed counts for each residue are shown in black. (b,d) Incorporation bias for each amino acid.

6.2.6.2 First round of selection

DNA amplified directly from the paramagnetic beads used in the first round assembly and the plasmids extracted from the selected clones of the same library were submitted for high-throughput sequencing, to characterise the initial library and to obtain information from the enrichment of sequences to help direct further rounds of directed evolution. Reads were trimmed, selected for high quality, translated and counted. This yielded 73,084 unique sequences out of 3.3x10^6 reads for the pre-selection library and 9,840 unique sequences from 3.3x10^5 reads in the post-selection library. This number of unique sequences in the post-selection dataset is much higher than the estimated 1,000 colonies recovered from the selection plate, which could be caused by sequencing errors that persist after the filtering steps in the analysis.

The average efficiency per cycle of assembly was approximately 8%, slightly higher than observed for the control library (produced using the same reaction conditions), and this yielded a length distribution with a strong bias towards very short sequences (Fig. 6.16). The sequence length of the wild-type and $_{164}$RYYGE$_{168}$ variants was only represented by 0.2% of this library and the only 0.005% of the reads recovered in this dataset had a length equal to PTX7.

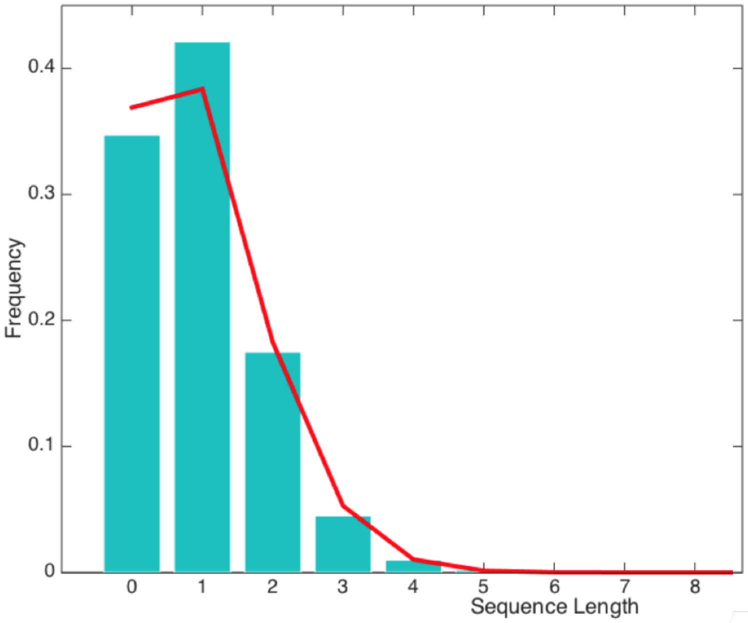


FIGURE 6.16: Length distribution of assembled sequences in the first round library for TEM-1. Frequencies for each class observed in the NGS dataset (bars) compared with the expected frequencies from a binomial distribution using the per-cycle efficiency derived from the dataset.

The exploration of sequence space around the target $_{164}$RYYGE$_{168}$ sequence was characterised by examining the compositional diversity and also the coverage of the possible sequences for each length category, before and after selection (Fig. 6.17). The compositional diversity was evenly distributed throughout the libraries, with only a small bias towards arginine codons in the first position, which was intentional in the library design. The sequence landscapes shorter than $_{164}$RYYGE$_{168}$ were relatively well-covered, with almost all possible 3-mers and around 10% of all 4-mers present in the pre-selection library. However, 5-mers had under 1% coverage because of the lower-than-expected efficiency from the assembly and the higher number of possible variants of this length. The high number of enriched sequences with Hamming distance greather than or equal to one shows that the sequence space around the $_{164}$RYYGE$_{168}$ target is populated with functional sequences, that were recovered by selection. This is also corroborated by the logo plots, which show a preference for Arg at the first position, Tyr at the third position and Gly at the fourth position. This data also confirmed that the sequence space around the seven-residue PTX7 variant was sparsely sampled and similar activities with high activities could still be found by a second round of diversification and selection.

The k-mer based motif detection analysis was used to probe the enriched sequences from this round of selection. The linear enrichment score was replaced by a Z-score derived from comparison between pre- and post-selection libraries as Poisson distributions, which was expected to provide accurate representation of enrichment than was observed in some of the simulated selections (Section 6.2.2). The score was defined as follows:

$$Z = \frac{(cX - Y) - (c\theta_x - \theta_y)}{\sqrt{C^2\theta_X + \theta_Y}} \tag{6.1}$$

here c is the ratio in size between post- and pre-selection libraries as a correction for sampling, X is the number of counts for a sequence post-selection, Y is the number of

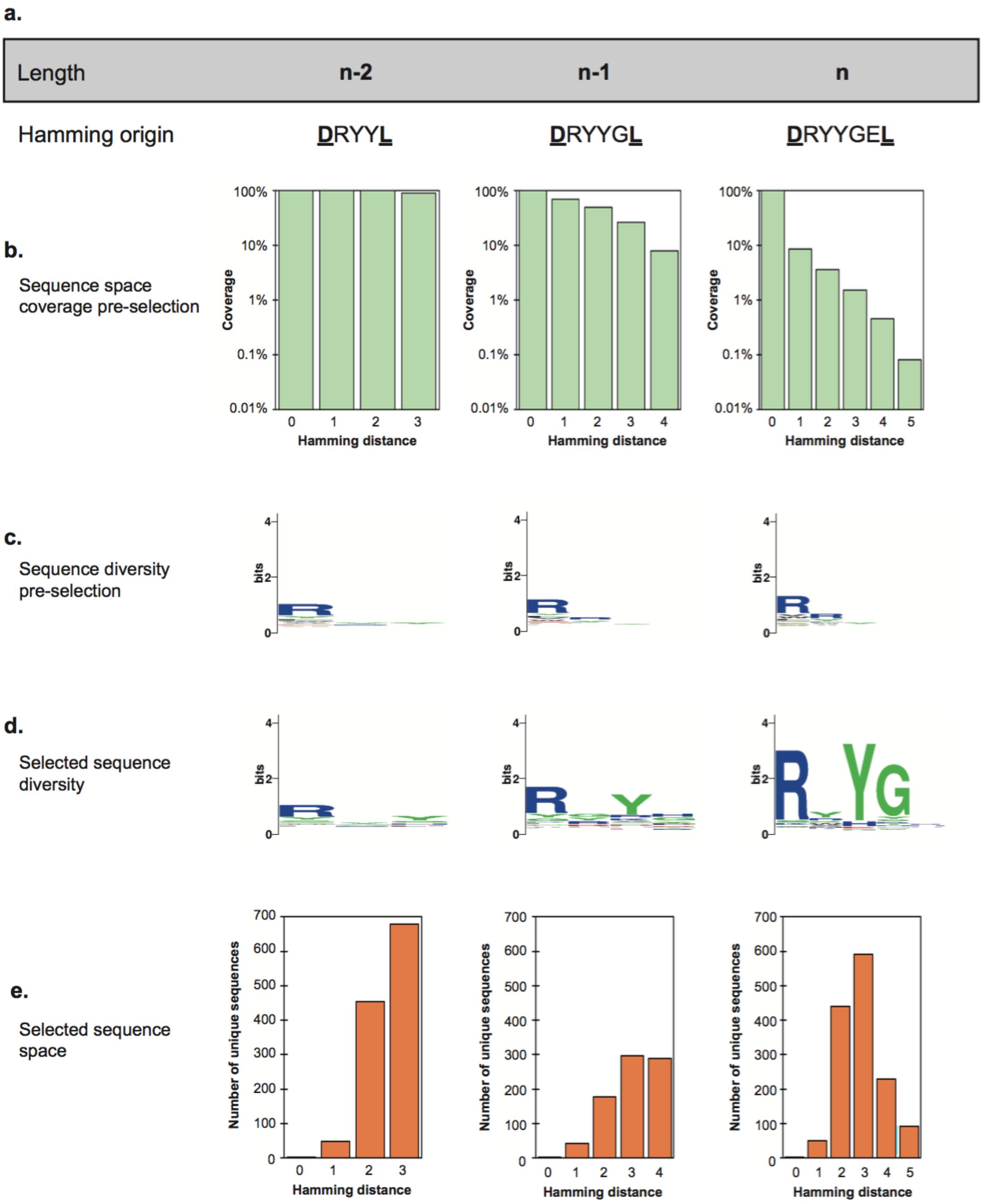


FIGURE 6.17: Exploration of the sequence space around $_{164}$RYYGE$_{168}$ before and after selection. (a) The possible sequence space is divided into fixed lengths and the most frequent variant derived from $_{164}$RYYGE$_{168}$ at that length was used as origin for Hamming distance calculation. (b) The sequence around $_{164}$RYYGE$_{168}$ was efficiently covered at the shorter sequence lengths. (c) Sequence logos show minor biases in the pre-selection library for arginine residues at the N-terminal end of the variable region. (d) Selection produced strong enrichment of sequences at the higher lengths. (e) Sequence diversity recovered post-selection at each length.

counts for that sequence pre-selection, θ_X and θ_Y are estimated Poisson parameters defined as X and Y counts as a fraction of the total number of reads for the respective libraries. The Z-score indicates divergence between the post- and pre-selection distributions, with very high positive values indicating the most highly enriched sequences.

Comparisons between the PCA-derived motifs and the most frequent sequences directly from the post-selection NGS dataset are in Table 6.4. While the first dimension of the PCA was covered by the three most frequent sequences, the second dimension recovered the $_{164}$RYYGE$_{168}$ sequence along with shorter variants RYYG and RYY. Other slightly more divergent motifs were also picked out by the PCA, such as GGW (PCA$_6$), RGYH (PCA$_8$), and SYHZ (PCA$_9$).

Component	PCA-derived motif	Most frequent match	Ranking
1	**Z**(D/T/P)**Z**	**Z**D**Z**	1
		ZT**Z**	2
		ZP**Z**	3
2	**Z**R(Y/G)YG**Z**	**Z**RYYG**Z**	14
	ZR(Y/G)YGX**Z**	**Z**RYYGE**Z**	16
	ZR(Y/G)Y**Z**	**Z**RYY**Z**	324
3	**Z**D**Z**	**Z**D**Z**	1
4	**Z**(D/S)**Z**	**Z**D**Z**	1
		ZS**Z**	4
	ZYS**Z**	**Z**YS**Z**	5
5	**Z**P**Z**	**Z**P**Z**	3
6	**Z**S**Z**	**Z**S**Z**	4
	ZG(G/S)**Z**	**Z**GG**Z**	12
	ZGGX**Z**	**Z**GGW**Z**	45
7		**Z**S**Z**	4
	Z(S/Q)**Z**	**Z**Q**Z**	7
	Z(R/S)YYG	**Z**RYYGE**Z**	16
	ZYYG	**Z**SYYG**Z**	225
		ZYYGH**Z**	4568
8	**Z**S**Z**	**Z**S**Z**	4
	ZRGYX**Z**	**Z**RGYH**Z**	10
	ZXH**Z**	**Z**PH**Z**	18
9	**Z**(S/G/N)(Y/H/G)**Z**	**Z**GG**Z**	12
	ZXYY(G/H)**Z**	**Z**RYYG**Z**	14
	Z(S/G/N)YH**Z**	**Z**SYH**Z**	31
10	**Z**(Q/M)**Z**	**Z**Q**Z**	6
	ZRGYX**Z**	**Z**RGYH**Z**	10

TABLE 6.4: Enriched motifs detected by k-mer based PCA of the first selection round

6.2.6.3 Second round of selection

The pre-selection library and the plasmids extracted from the selected clones of the second round library were also submitted for high-throughput sequencing, to characterise the initial library and to identify motifs that were enriched in this round of selection. Reads were trimmed, selected for high quality, translated and counted. This yielded 731,711 unique sequences out of 7.9x10^6 reads for the pre-selection library and 765 unique sequences from 6.5x10^4 reads in the post-selection library. Similarly to what was observed in the first round dataset, 765 unique sequences were recovered in the post-selection NGS data, out of approximately 80 colonies recovered from the selection. This near tenfold excess of sequences over the number of colonies could be due to sequencing errors, mutations accumulated during outgrowth or degradation of ceftazidime during outgrowth in liquid media leading to a lower selection stringency and recovery of inferior clones.

The average efficiency per cycle of InDel assembly was approximately 20% for this library, higher than was observed in the control and first-round libraries (Fig. 6.18). The wild-type and $_{164}$RYYGE$_{168}$ sequence length of 5 residues constituted 7.5% of the total sequences in the library and the length of PTX8 — 8 residues in total, but the N-terminal Arg was kept constant in this assembly and is not counted in the plot — was approximately 0.8% of the total sequences.

It is interesting to note that the proposed single-mutation path from PTX7 to PTX8 (Section 6.2.5.3) was fully represented in the pre-selection library. The initial Met deletion ($_{164}$RGYKER$_{168a}$) had 26 reads, the Glu insertion ($_{164}$RGYKEER$_{168b}$) had 6 reads and the final selected variant ($_{164}$RGYKEERD$_{168c}$) had a single read. However, none of the intermediates were detected in the post-selection dataset and initial PTX7 sequence was also not represented in the selected variants. This is only one of the many possible single-mutation paths that can bridge these two sequences (including different orders of these

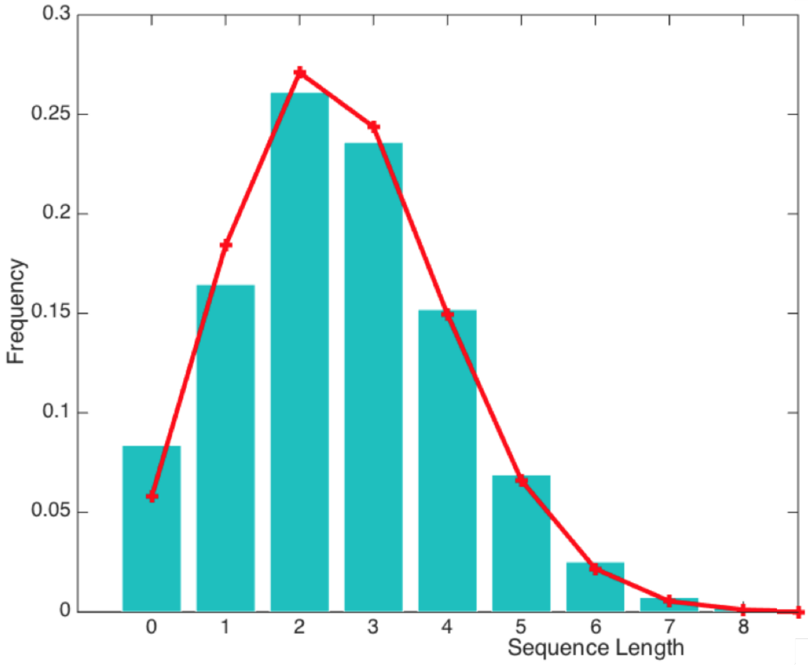


FIGURE 6.18: Length distribution of assembled sequences in the second round library for TEM-1. Frequencies for each class observed in the NGS dataset (bars) compared with the expected frequencies from a binomial distribution using the per-cycle efficiency derived from the dataset.

steps and also other mutations followed by reversals), but it is a demonstration of the potential inherent in using InDel Assembly-derived libraries for protein engineering. Since none of the proposed intermediates was enriched in the second round, it is likely that they did not confer sufficient resistance to survive the higher selection stringency and, therefore, could not be obtained by consecutive single mutations while being selected at 200 µg/ml ceftazidime.

As was done for the first-round library, the exploration of sequence space was characterised by the compositional diversity and fraction of possible sequences in each length category (Fig. 6.19). In this round, there was some bias towards glutamate and arginine residues in the last two positions, as well as a weaker bias for glycine in the first position. The higher assembly efficiency and higher number of cycles in this round led to a higher degree of coverage of sequence space. Nearly 50% of all possible four-residue sequences were detected before selection, compared with less than 10% in the first library. The longer landscapes have greatly increased numbers of possible sequences and, therefore, were more sparsely covered. However, as in the previous round, the clone isolated at the highest selection stringencies belongs to a length category that had less than 0.1% of its possible sequences sampled. The higher per-cycle insertion efficiency enabled this library to cover the wild-type and $_{164}$RYYGE$_{168}$ sequence length (four diversified residues in this assembly) much more efficiently than the first round, which did not reach 1% coverage for this length category.

Since this round of selection was carried out at a higher stringency, the post-selection diversity was greatly reduced and the sequence logos show one or two motifs being clearly dominant in each landscape. The number of unique sequences post selection was also smaller, with less than 100 sequences recovered post-selection in each landscape. The difference between the number of colonies obtained in the library selection plate and the number of unique sequences recovered by sequencing is likely a result of sequencing artifacts generated in library preparation or the sequencing process itself, due to the existence of multiple low-frequency sequences diverging from the highly-enriched variants by only a single residue.

The clear trends in selection visible in the broad view of library exploration were compared with the results produced by the k-mer based PCA for motif detection (Table 6.5). The first four PCA dimensions match the most-frequent sequences exactly, but there are changes in ranking starting from the fifth dimension and some sequences with lower read counts appear higher in the PCA motifs due to similarity to other more strongly-enriched sequences. Sequences $_{164}$RMHKKRH$_{168b}$, $_{164}$REYGEQ$_{168a}$, $_{164}$RRYGT$_{168}$ and $_{164}$RGERQ$_{168}$ were selected for further characterization, along with other variants isolated during selection rounds and controls.

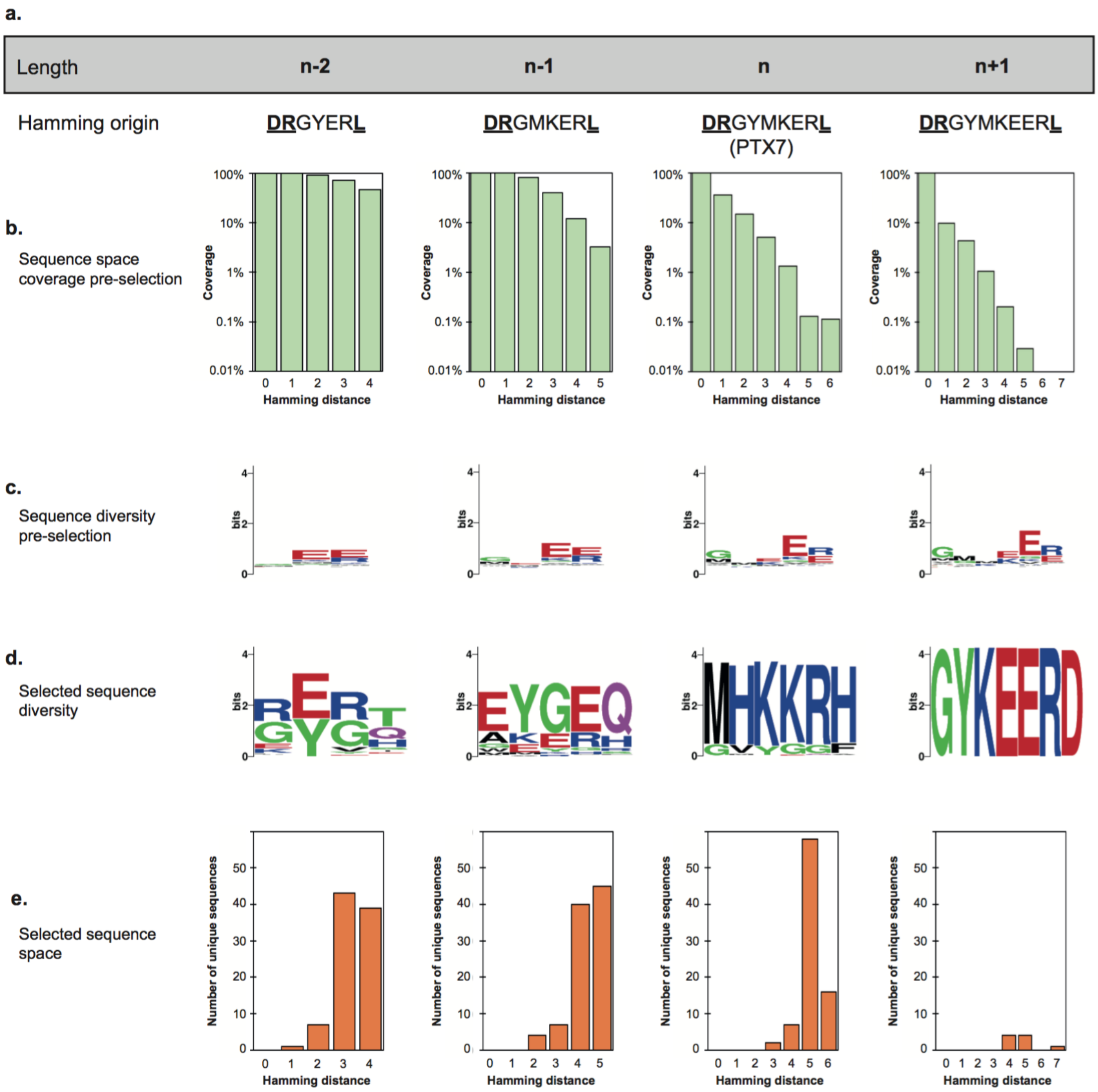


Figure 6.19: Exploration of the sequence space around PTX7 before and after selection. (a) The possible sequence space is divided into fixed lengths and the most frequent variant derived from PTX7 at that length was used as origin for Hamming distance calculation. (b) The sequence space around PTX7 was efficiently covered at the shorter sequence lengths. (c) Sequence logos show minor biases in the pre-selection library for glutamate residues towards the C-terminal end of the variable region. (d) Selection produced strong enrichment of sequences at each length. (e) Sequence diversity recovered post-selection at each length.

Component	PCA-derived motif	Most frequent match	Ranking
1	**Z**MHKKRH**Z**	**Z**MHKKRH**Z**	1
2	**Z**EYGEQ**Z**	**Z**EYGEQ**Z**	2
3	**Z**RYGT **Z**	**Z**RYGE**Z**	3
4	**Z**GERQ**Z**	**Z**GERQ**Z**	4
	ZGE**Z**		
5	**Z**GVYGGF**Z**	**Z**GVYGGV**Z**	7
	ZGVY**Z**	**Z**GVY**Z**	8
6	**Z**AKERH**Z**	**Z**AKERH**Z**	5
	GVY	**Z**GVY**Z**	8
	Z E(V/K/R)X**Z**	**Z**EKERH**Z**	108
7	**Z**AKEXH	**Z**AKERH**Z**	5
	Z(A/G)YV**Z**	**Z**GYV**Z**	6
8	**Z**EEVH**Z**	**Z**EEHV**Z**	9
9	**Z**(A/W)(E/Y)EHR**Z**	**Z**WEGRQ**Z**	10
	ZWEGR(Q)**Z**	**Z**AYEHR**Z**	12
	VEGRQ		
10	VX**Z**		
	ZGVY**Z**	**Z**GVY**Z**	8
	Z(A/G)Y(Y/E)HR**Z**	**Z**AYEHR**Z**	12
	ZGAYEHR**Z**		

TABLE 6.5: Enriched motifs detected by k-mer based PCA of the second selection round

6.2.7 Phenotypic measurements and comparisons between selected variants

A small number of candidates from the PCA analysis was selected for characterisation by measurement of the levels of resistance conferred to different classes of antibiotics. Among these were WT and $_{164}$RYYGE$_{168}$ as controls, PTX7, PTX8, Δ5 and the set of four variants extracted from the top four dimensions of the second round library PCA to validate the functional motif detection based on sequence enrichment: $_{164}$RMHKKRH$_{168b}$, $_{164}$REYGEQ$_{168a}$, $_{164}$RRYGT$_{168}$, and $_{164}$RGERQ$_{168}$. Only sequences from the second round were selected for this characterisation with the goal of identifying variants with the highest levels of resistance to ceftazidime, which these were expected to possess due to being selected at higher stringency.

Two simple measurements from clinical microbiology assays (Balouiri et al., 2016) were chosen to compare the phenotypes of variants isolated from selection and observed in the second-round sequencing dataset to WT and $_{164}$RYYGE$_{168}$, requiring development of protocols in the group to produce these measurements. Minimal Inhibitory Concentrations (MICs) were estimated by the broth microdilution assay, in which cultures are grown in the presence of dilutions of antibiotics in 96-well plates and growth is measured by OD600, defining MIC as the lowest concentration of a specific antibiotic at which no growth was observed for a given strain. Disc Diffusion Assays were also carried out, by measuring the diameter of the zones of inhibition created around filter paper discs containing different antibiotics.

Initial attempts at measuring MICs by growth in 96-well plates resulted in noisy data due to evaporation from wells on the edges of the plate and cell clumps in sub-MIC concentrations reducing the accuracy of absorbance measurements (Fig. 6.20). The evaporation issue was solved by not carrying out cultures in the outer rows and columns of the plates and filling those wells with 200 µl sterile H_2O. The effect of cell clumping was reduced by mixing all cultures using a multichannel pipette prior to reading on the spectrophotometer.

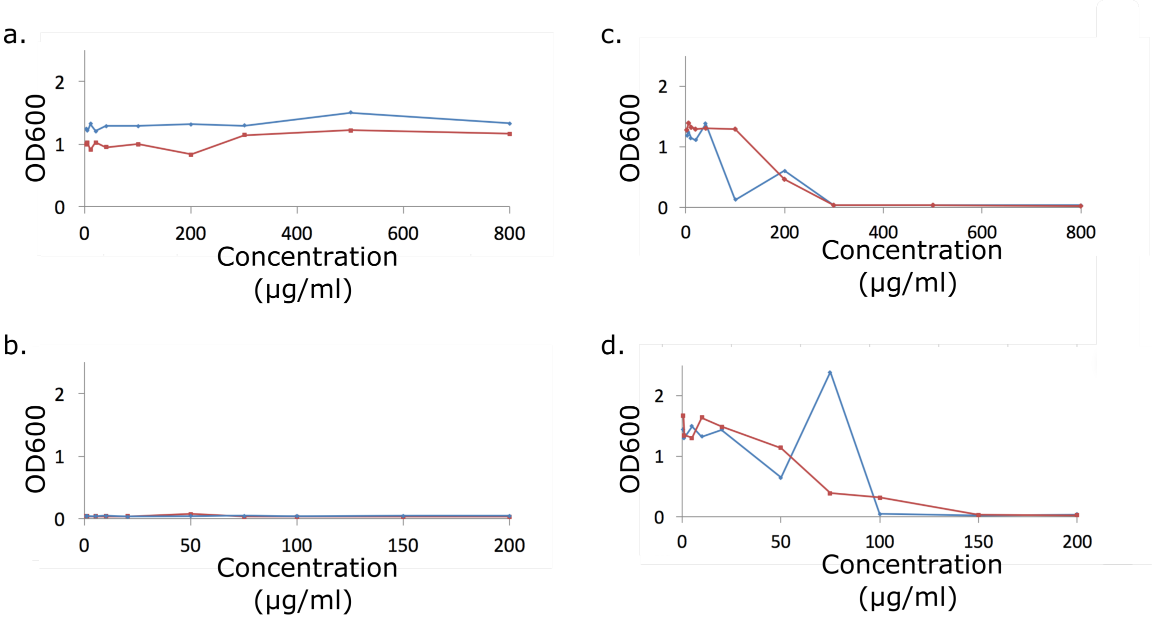


FIGURE 6.20: Initial conditions for MIC measurements in 96-well plates - Growth of TEM-1 WT (a and b) and $_{164}$RYYGE$_{168}$ (c and d) variants grown in varying concentrations of ampicillin (a and c) or ceftazidime (b and d) were measured by absorbance measurements at 600 nm. Each condition was grown as two independent biological replicates.

The variants chosen for these phenotypic comparisons were all recloned to ensure that the measured phenotypes were solely a consequence of changes in the Ω-loop rather than any undetected mutations which could have occurred elsewhere in the TEM-1 coding sequence, in the vector backbone or in the genome of isolated clones. These new clones were produced by commercial synthesis of the variable regions for each construct as separate single strands containing BsaI sites. Pairs of single stranded oligos coding for each variant were annealed, digested and inserted into pTEM1 vector amplified by iPCR with primers containing BsaI sites with compatible overhangs.

Characterisation of the substrate spectrum of isolated TEM-1 variants was carried out by MRes student Emma Harris and the results are in Fig. 6.21. In the disc-based assay, PTX7 and $_{164}$RYYGE$_{168}$ did not have their growth inhibited by the ceftazidime discs, while the WT was strongly inhibited and all other variants (including PTX8) had intermediate levels of resistance. A similar pattern was seen for cefotaxime, which is also a non-cognate cephalosporin. The cognate substrates ampicillin and carbenicillin did not affect the growth of the WT, with the remaining clones showing varying degrees of resistance to these antibiotics. Interestingly, the $_{164}$RGERQ$_{168}$ variant was more resistant to these antibiotics than all other selected motifs, while having similar or lower activity on the cephalosporins. As expected, the stabilised $_{164}$RYYGE$_{168}$ variant characterised here had a MIC of 150 µg/ml, apparently higher than 64 µg/ml that was described by Petrosino and Palzkill (1996), but direct *in vitro* activity comparisons of the two proteins would be needed to eliminate the confounding effects of strain differences and changes in expression constructs on the relative amount of active enzyme produced by each strain.

Imipenem, belonging to the carbapenem class, was also included in the panel but, as expected, none of the variants conferred resistance to this drug which is clinically used to treat resistant infections, due to the very inefficient hydrolysis of this compound by many TEM-1 variants. The results obtained in the microplate-based assay mirrored the disc-based assay closely, with the main difference being the slightly higher MIC of the $_{164}$RYYGE$_{168}$ variant in ceftazidime (150 µg/ml) compared to PTX7 (100 µg/ml). Since both variants had null values for growth inhibition in the single-dose disc assay, the range of concentrations used in the microplate assay allowed these two variants to be distinguished and demonstrated that the original $_{164}$RYYGE$_{168}$ variant confers higher levels of resistance onto *E. coli* than the novel variants selected at higher ceftazidime concentrations.

Significance tests were carried out for the disc-based assays, employing *t*-tests with Bonferroni corrections for multiple testing. All the variants were significantly more resistant to ceftazidime and cefotaxime than WT, except for $_{164}$RGERQ$_{168}$ against cefotaxime, and all were significantly less resistant to ampicillin and carbenicillin than WT. None of the variants had statistically-significant differences in resistance to imipenem when compared to WT. The $_{164}$RGERQ$_{168}$ variant was also significantly more resistant to carbenicillin than PTX7, the variant with the second lowest inhibition radius for this antibiotic.

As expected, the results for drugs of the same class — penems (ampicillin and carbenicillin) and cephalosporins (ceftazidime and cefotaxime) — were very similar, reflecting the structural differences between classes. The results obtained for this panel of variants corroborates the previous observation of the E166Y mutation in the active site leading to higher activity on ceftazidime (Sowek et al., 1991). It is interesting to note that of the variants tested in this panel, only $_{164}$RGERQ$_{168}$ has a Glu residue in the same position as WT. This is the selected variant with the highest resistance to penems and lowest resistance to cefotaxime, suggesting a role for the glutamate residue in favouring activity in penems. All the other variants have a Tyr residue in the same position as $_{164}$RYYGE$_{168}$, except for $_{164}$RMHKKRH$_{168b}$ which has His166 or Lys167/168 as possible candidates for the general base in the catalytic mechanism of this mutant.

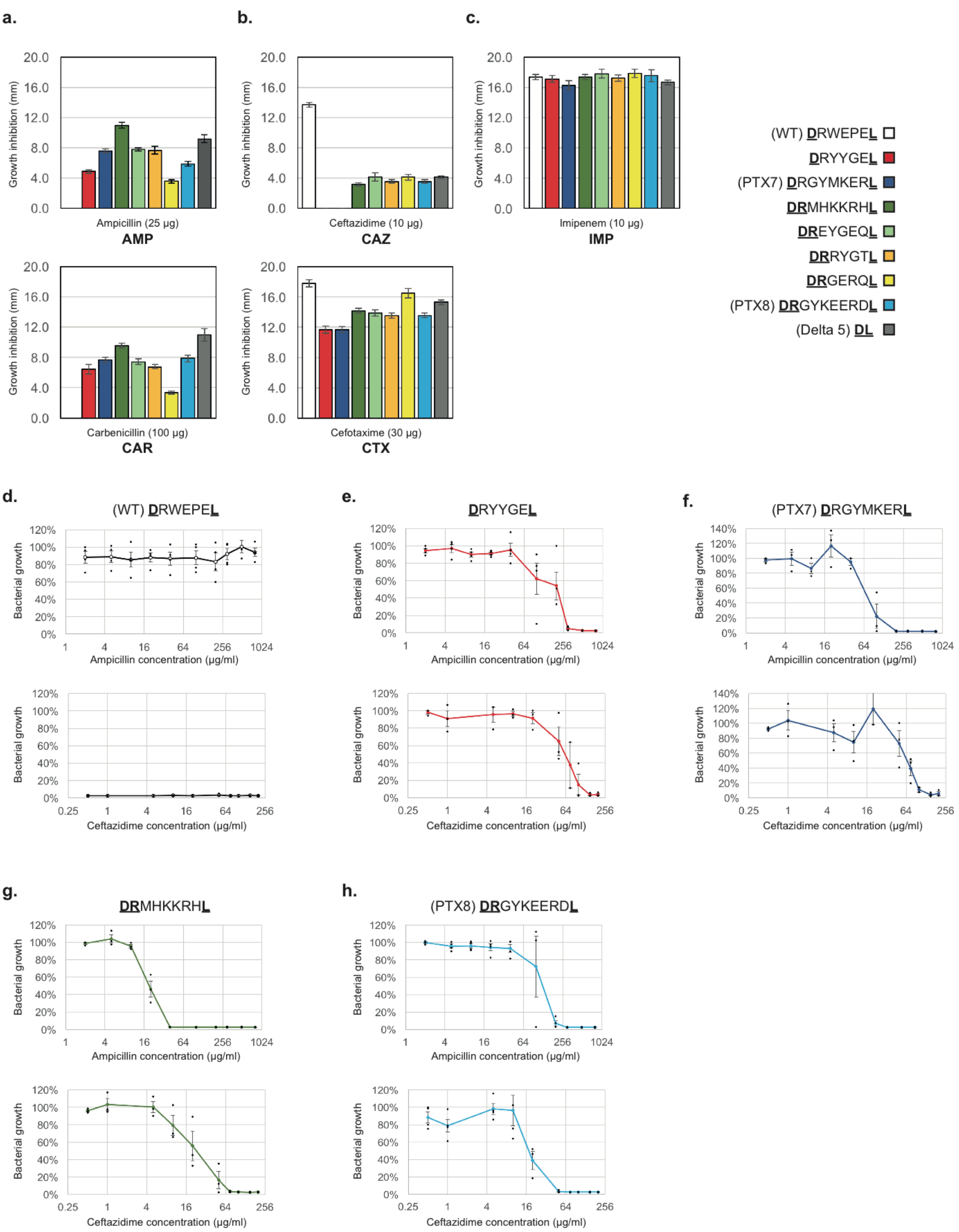


FIGURE 6.21: Resistance spectrum of selected TEM-1 variants produced by selection of InDel-derived libraries. Resistance of all clones to (a) penem, (b) cephalosporin, and (c) carbapenem antibiotics was measured by growth inhibition around antibiotic-soaked paper discs. Lower values in this assay indicate higher resistance. Resistance of selected clones to ampicillin and ceftazidime: (d) WT, (e) ${}_{164}\text{RYYGE}_{168}$, (f) PTX7 , (g) ${}_{164}\text{RMHKKRH}_{168b}$ and (h) PTX8. Results produced by MRes student Emma Harris, using the assay methodology developed in this work.

6.2.8 Selection parasites

Since the recloned PTX7 and PTX8 variants were unable to grow at the ceftazidime concentrations used for high stringency selection of the libraries (300 µg/ml ceftazidime for PTX7 and 500 µg/ml ceftazidime for PTX8) and that allowed the original clones — hereafter named PTX7S and PTX8S — to be isolated, they possibly represent selection parasites (Tizei et al., 2016). These are variants that are enriched by selection, but possess a distinct phenotype from what selection was intended to enrich, due to secondary effects that increase their chance of survival in the selection system. In the specific case of selecting Ω-loop libraries for increased ceftazidime hydrolysis by TEM-1, a parasite can be a mutation in an untargeted region of the TEM-1 sequence, a mutation in the plasmid that increases the concentration of active β-lactamase or a genomic mutation that synergistically increases resistance to ceftazidime through another mechanism.

The plasmid backbones of clones PTX7S and PTX8S were fully sequenced to identify any secondary mutations which could explain the differences in phenotype. Each clone had a single substitution in the entire plasmid sequence, in addition to the targeted TEM-1 Ω-loop region, and both mutations were outside the TEM-1 coding sequence. This suggests that the higher tolerance of PTX7S and PTX8S when compared to the recloned variants could be a consequence of more efficient expression of the TEM-1 gene, rather than an increase in specific activity of the enzyme.

PTX7S had a G to T substitution 49 base pairs upstream of the N-terminal Met codon of TEM-1, between the AmpR promoter and TAAGGAG Ribosome Binding Site (RBS). The Salis lab RBS calculator (Salis et al., 2009) did not predict any new Ribosome Binding Sites in the mutated region, which could produce higher levels of protein. Therefore, the isolation of PTX7S at concentrations higher than the MIC of PTX7 could be explained by another undetected regulatory effect or buffering of ceftazidime concentrations by increased production of Penicillin Binding Proteins (Sauvage and Terrak, 2016).

PTX8S had a C to T substitution in position 89 of the ColE1 origin of replication. This site is transcribed from both strands to produce the RNAI and RNAII regulators of ColE1 replication initiation. Stable base-pairing interactions between these loops in these RNA molecules prevents the initiation of replication because the $3'$ end of RNAII is the primer that starts replication, so a mutation in this region could have an impact on plasmid copy number. This substitution has been previously reported as responsible for runaway plasmid replication, by disturbing the interaction between the second loop of RNAI and RNAII (Lacatena and Cesareni, 1981; Camps, 2010). This was corroborated by the observation that plasmid extractions from PTX8S cultures yielded approximately threefold more DNA than was routinely obtained from all other clones in the pTEM1 vector. An increased plasmid copy number could lead to an increase in production of PTX8 mRNA and protein, increasing the total ceftazidime-hydrolysing activity of carrier cells.

These secondary mutations could also be part of the reason why the $_{164}$RYYGE$_{168}$ variant — which confers levels of resistance against ceftazidime similar to PTX7 and PTX8 — was not isolated in either round of high-stringency selection, while at least being present in both pre-selection libraries. Since the stringent selection steps in both rounds were carried out at concentrations far higher than the MIC of $_{164}$RYYGE$_{168}$ (300µg/ml and 500 µg/ml CAZ in the selections, while its MIC was 150 µg/ml), mutations of similar effect as the ones found in PTX7S and PTX8S would need to occur in the strain carrying a $_{164}$RYYGE$_{168}$ copy to increase the likelihood of this variant being isolated under high-stringency conditions.

Alternatively, the high-stringency plates could have contained lower effective concentrations of the antibiotic due to degradation by TEM-1 variants secreted during growth of selection pools in liquid culture, allowing growth of strains with MICs lower than the intended ceftazidime concentration. In this situation, any changes in relative number of cells of each variant in the pool placed into liquid cultures could be amplified during growth and the sample of only 10^4 cells in the high-stringency plates could not be sufficient to recover all the functional variants from the pool after growth. This effect could be reduced by washing the colonies harvested from the initial selection plates in fresh medium to ensure no residual TEM-1 is carried over and eliminating the liquid growth step before plating into high-stringency conditions, to ensure a more even representation of the harvested variants.

6.3 Conclusions

The results obtained from the two rounds of selection for TEM-1 using libraries made with InDel assembly demonstrate the power of this novel approach to quickly obtain functional variants from unexplored regions of sequence space. Despite the per-cycle insertion efficiencies being lower than desired — even after additional rounds of protocol improvement — libraries produced by this strategy allowed for efficient exploration of the sequence neighbourhood around known positive variants of TEM-1 while at the same time producing novel functional variants that are highly divergent in length and composition. Producing similar diversity by other strategies would require the construction of multiple libraries to cover all possible insertions and deletions from target sequences (16 separate libraries to cover just 1 residue longer or shorter than PTX7), which can become costly and experimentally intractable. Due to the ease of producing customisable libraries that cover length-variant regions of sequence space, InDel Assembly has potential for use in discovery of binders such as antibodies or ones based on different protein scaffolds (e.g, bicyclic peptides (Heinis et al., 2009)).

The fact that both clones isolated from the high-stringency selection steps were selection parasites illustrates one of the main pitfalls in employing selection in directed evolution. As selective pressures are increased to isolate a small number of variants — or even a single variant — from a large initial population, the likelihood of obtaining such a false positive result also tends to increase, especially when selection is carried out *in vivo* by differential growth of cells that are adaptable and can spontaneously generate variants that circumvent the artificial selection system constructed for directed evolution. The appearance of these parasites is a possibility that must always be evaluated when designing selection strategies for directed evolution and, if possible, researchers should aim to predict which possible phenotypes could represent parasites within the selection, to implement strategies to mitigate their effect (Tizei et al., 2016).

In this context, the enrichment analysis framework presented here can also be a powerful strategy to avoid such parasites, since neither PTX7 nor PTX8 were identified in the motifs from the top ten dimensions of the PCAs (see Sections 6.2.6.2 and 6.2.6.3). The fact that more than one variant in the post-selection library can contribute to the enrichment detected in the PCA reduces the likelihood of a parasite receiving a high score, since a secondary beneficial mutation would have to occur simultaneously in all the variants that contributed to the predicted motif.

Libraries produced with InDel Assembly coupled to the k-mer based enrichment analysis can help accelerate the engineering of biological systems, reducing the selection or screening effort require to obtain the desire phenotype by using information from previous rounds of selection to inform the design of subsequent libraries. Previous strategies to analyse the output of antibody selections stratified the enriched variants into discrete length categories, discarding valuable information about functional motifs that could be exploited to direct further improvements more robustly to the obtained binders (Ravn et al., 2013). In addition to the strategy described here for loop engineering of TEM-1, considerable effort has been directed towards obtaining methods to intelligently restrict search space in directed evolution, producing libraries enriched for the desired functions and — consequently — reducing screening or selection effort (Currin et al., 2015; Woldring et al., 2017).

So far, InDel assembly has only been tested for codon-by-codon assembly, but the method design can accommodate differently-sized building blocks, which also opens the possibility of producing libraries to engineer phenotypes on different scales. Short structural elements can be combinatorially assembled to produce protein domains, domains can be shuffled in large modular proteins, and even metabolic pathway engineering could be done with longer elements encompassing coding sequences and regulatory regions.

The framework presented in this chapter, though validated by engineering an enzyme as a proof-of-principle, can be directly employed for the engineering of TOMMs and other classes of compounds derived from ribosomal peptides. As long as a robust selection or screen is available to represent the test stage of the synthetic biology cycle (Baldwin et al., 2015), InDel Assembly can be used to produce a library — targeted if starting candidates are known, or with no bias — and the k-mer based analysis strategy can detect functional motifs to inform further rounds of diversification and selection.

Due to the cycle time of over three hours and the fact that multiple cycles need to be carried for every unit increase in average length of the library, the current codon-by-codon assembly strategy would only be adequate for relatively short stretches of residues in a hypothetical TOMM precursor library. However, longer fragments can be built up using longer assembly blocks as mentioned above and by designing library amplification primers to allow multiple copies of the library to be cloned in tandem, even up to the 69-residue length of the McbA precursor of Microcin B17.

Even in the absence of a robust *in vitro* platform for the heterocyclisation of modified TOMM precursors, it would be possible to employ InDel libraries to further probe the requirements for function in the Microcin B17 system. By cloning a low diversity library built using fragments containing heterocyclisable residues flanked by glycines (most of the cyclised positions in Mcb17 are in G(S/C)(S/C)G and G(S/C)G motifs) and linker fragments into the biosynthetic operon, it would be possible to carry out medium throughput screening on 96-well plates to obtain more information on the functional effects of the order and number of the heterocyclic moieties within the compound.

The inconsistency between results obtained for DNA gyrase inhibition and antibacterial activity for truncated Microcin B17 variants (Shkundina et al., 2014; Thompson et al., 2014) suggests an as-yet-uncharacterised target for this molecule. In this regard, if a screen of InDel-derived variant precursors heterocyclised *in vivo* produces novel Microcin B17 analogs, these could be employed as tools to elucidate further the structure-function relationship of this microcin. One possible strategy to determine whether there is a second target of Microcin B17 that is unknown would be to screen microcin variants against a panel of *E. coli* gene deletion mutants and any mutant that presents a response different to the WT to microcin variants would indicate a candidate gene for a gyrase-independent effect. The proteins encoded by these candidate mutants could then be identified and the effect of microcin variants on their function measured by appropriate assays.

Chapter 7

Conclusion and Perspectives

7.1 Conclusions

The work presented here contributed in distinct aspects to the mechanistic understanding of TOMM synthases and their products, representing foundational work towards the establishment of this biosynthetic pathway as a tool for the discovery of novel azole-containing bioactive compounds and also with broader applicability in many other biological systems.

A sequence-based functional prediction was developed and tested for a well-characterised protein family, the bacterial solute binding proteins involved in the transport of amino acids and other small molecules from the external medium. The three metrics employed in the prediction detected residues known to be relevant to function of the glutamine binding protein GlnBP, including ones that have been targeted by previous work to alter the substrate specificity of this protein (Looger et al., 2003). The NoSE metric proposed here is of special interest, being designed to detect determinant residues, conserved only within a subset of sequences within a mixed-function alignment that are known to share a single function. Previous methods to detect residues with this evolutionary signature relied on the construction of alignments with clearly delimited functional subgroups (Kalinina et al., 2004), which is often not possible if insufficient functional annotation is available for the family of interest.

The prediction strategy was then applied to the TOMM synthases and predictions were validated for the dehydrogenase McbC in the biosynthetic complex for Microcin B17 of *E. coli*. A simple bioassay was established for validation, showing an impact on the *in vivo* production of this TOMM for eight out of the sixteen McbC mutants constructed. Among the putative functions for the mutants with a negative phenotype inferred from the literature and homology modeling are candidate FMN cofactor-interacting residues, sites with apparent structural roles, and others along a putative homodimerisation surface. The functional results obtained for McbC and the predictions made for the remaining proteins in the complex can assist in producing novel insight into the functioning of the TOMM synthase complex, by targeting experimental characterisation effort at sites already known to posses a functional role, or at least predicted to have a role.

Three distinct TOMM synthase complexes — from *B. amyloliquefaciens*, *Bacillus sp.* Al-Hakam, and *E. coli* — were heterologously expressed and purified for *in vitro* reconstitution and deeper characterisation of their activity. Activity detection was attempted by mass spectrometry, fluorescent labeling of free cysteines, and employing a commercial polyclonal antibody. Unequivocal and consistent detection of *in vitro* TOMM synthase activity could not be obtained for any of the three complexes due to difficulties with the expression and purification of proteins in the complex and the substrate peptides. However, evidence of activity was found for the complex from *E. coli* by MS detection and substrate peptide variants with increased solubility were obtained that avoid the peptide solubility issues encountered in attempts to detect the activity of this complex. Further investigation into these substrate variants can yield a more robust assay system for TOMM synthases, producing consistently reproducible results.

Finally a novel integrated framework was developed for the directed evolution of flexible regions of proteins, employing sequence libraries containing diversity in both composition and length. The assembly method produces customisable high-quality libraries containing heterometric diversity and the analysis strategy detects motifs enriched among sequences of variable lengths after rounds of selection. Validation of this system by two rounds of directed evolution of the β-lactamase TEM-1 led to the isolation of seven novel variants active on the non-cognate substrate ceftazidime spanning a range of lengths — the diversified region of these variants had sequences of zero to eight residues in length — that could not be easily obtained by previous methods. This framework also has immediate applicability in the engineering of proteins such as antibodies (Ravn et al., 2013) and enzymes with active sites composed of loops (Afriat-Jurnou et al., 2012).

7.2 Perspectives

Each of the avenues of research explored in this thesis raised relevant new questions that could be pursued by future members of the Pinheiro group and could also be taken up by other research groups working in related fields. The tools developed in this work with the aim of furthering the understanding of TOMM synthases and establishing their engineering can also be applied to other biological systems.

The functional prediction strategy proposed here — including the NoSE metric for detection of determinant residues — was able to recover sites known to be important in the solute binding protein GlnBP along with additional candidate residues that have not yet been investigated and could be relevant to its native function, as well as for engineering of novel phenotypes. Other families of proteins that maintain enough sequence conservation for construction of alignments but diverge in phenotype could be targeted by this strategy for prediction of functionally relevant sites. Among these possible targets are the highly diverse families of proteins involved in eukaryotic signaling pathways such as protein kinases with their wide range of phosphorylation substrates (Bradley et al., 2017) and families of G-Protein Coupled Receptors that include members acting as receptors for small molecules and light receptors (Wolf and Grunewald, 2015).

The initial characterisation obtained here for the dehydrogenase McbC from the *E. coli* TOMM synthase complex can be expanded to the remaining proteins in the complex for identification of more candidate residues for complex engineering. In parallel, the mutants confirmed to affect dehydrogenase function in the *in vivo* assays can be further characterised to determine the nature of their effect on McbC function. Overexpression and purification of the mutants can quickly reveal whether FMN binding was affected, by the visible lack of yellow colour in solutions of proteins that not contain this cofactor. Structural integrity can be assayed by denaturation-based assays such as thermofluor (Ericsson et al., 2006) or circular dichroism — the latter also yielding information on any changes in secondary structure content caused by mutations.

In addition, the microcin B17 bioassay system can be employed to investigate interactions between components of the TOMM synthase complex by varying relative expression levels of genes within the operon to identify rate-limiting steps in overall TOMM biosynthesis and how these are affected by the mutations obtained here. The approach employed by Melby et al. (2012) to reconstruct a functional *Bacillus sp.* Al-Hakam TOMM synthase complex with the addition of a functional dehydrogenase from a different species can be mirrored in the *in vivo* system, enabling the identification of putative enzymes from related organisms that can rescue TOMM biosynthesis.

In vivo engineering of TOMM production can be further expanded by the development of a biosensor for heterocyclised moieties in peptides, analogously to so,e of the recently-developed small-molecule biosensors (Feng et al., 2015b; Skjoedt et al., 2016). Such biosensors could be used in logic circuits linking intracellular presence of heterocyclised peptides to growth or expression of a selectable marker, enabling the exploration of the sequence space for substrates and synthase complexes with the aim of obtaining promiscuous systems capable of delivering diverse libraries of synthetic TOMMs for activity screening.

Finally, the efficient exploration — at a low cost — of length-variable sequence space that was enabled by InDel Assembly coupled to the motif detection framework can be employed to shift protein engineering efforts away from their almost-exclusive focus on constant sequence lengths (Toth-Petroczy and Tawfik, 2014a), more faithfully reflecting the sampling of indels by biological processes (Tenaillon et al., 2016; Toprak et al., 2011). The assembly method itself can also be made even more accessible by optimising the amount of time and reagents required for library construction, as well as translation into an automated format with computer-guided library design and automated liquid handling for the assembly steps. Changes in the size of the sequence inserted with each assembly cycle can also be explored to enable the assembly of combinatory libraries of secondary structure elements, domains, and even non-coding elements such as regulatory regions.

Chapter 8

Appendix A - List of Reagents

Table 8.1: Reagents used in this work.

Reagent (Abbreviations or stock solution used in this work)	Manufacturer
α-Cyano-4-hydroxycinnamic acid (CHCA matrix, 10 mg/mL solution in Methanol / 0.1% Formic acid)	Sigma-Aldrich, USA
1kb DNA ladder N3232L	New England Biolabs, USA
100 bp DNA ladder N3231L	New England Biolabs, USA
20% Sodium Dodecyl Sulfate solution (SDS)	Fisher Scientific, USA
2-Butanol	Sigma-Aldrich, USA
37% Hydrochloric Acid (HCl)	VWR International, USA
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	Sigma-Aldrich, USA
5-(Iodoacetamido)fluorescein (IAA-FITC, 10 mM stock solution in DMF)	Santa Cruz Biotechnology, USA

Reagent (Abbreviations or stock solution used in this work)	**Manufacturer**
Maleimide-5-Fluorescein (Mal-FITC)	Thermo Scientific, USA
50% Sodium Hydroxide solution (NaOH)	Sigma-Aldrich, USA
Absolute Ethanol (Ethanol)	VWR, USA
Acetone	VWR International, USA
Acrylamide/Bisacrylamide 37.5:1, 40% solution	Alfa Aesar, USA
Adenosine triphosphate (ATP)	Roche Diagnostics GmbH, Germany
Agarose	Fisher Scientific, USA
Amicon Ultra-0.5 Centrifugal Filter Unit with Ultracel membrane	Merck Millipore, Germany
Ammonium acetate (4M solution in dH_2O)	AppliChem GmbH, Germany
Ammonium persulfate (10% (w/v) solution in dH_2O)	Acros Organics, USA
Ampicillin, sodium salt (100 mg/mL solution in dH_2O)	AppliChem GmbH, Germany
Anti-sulfathiazole polyclonal serum	LSBio, USA
Bacto Tryptone (Tryptone)	Becton Dickinson, USA
Bacto Yeast Extract (Yeast Extract)	Becton Dickinson, USA
Betaine	Sigma-Aldrich, USA
Boric acid	Fisher Scientific, USA
Bovine Serum Albumin (BSA)	Sigma-Aldrich, USA
Bromophenol Blue	Alfa Aesar, USA

Reagent (Abbreviations or stock solution used in this work)	**Manufacturer**
Chitin Resin	New England Biolabs, USA
Color Protein Standard, Broad Range P7712L	New England Biolabs, USA
cOmplete Mini EDTA-Free Protease Inhibitor Cocktail	Roche Diagnostics GmbH, Germany
D-Glucose	Sigma-Aldrich, USA
Diethyl Ether	Sigma-Aldrich, USA
Difco Agar (Agar)	Becton Dickinson, USA
Dithiothreitol (DTT)	Thermo Scientific, USA
Dimethylformamide (DMF)	Sigma-Aldrich, USA
Dimethylsulfoxide (DMSO)	Thermo Scientific, USA
Ethylenediaminetetraacetic acid (EDTA), trisodium salt	Sigma-Aldrich, USA
Factor Xa protease	New England Biolabs, USA
Formic acid	Sigma-Aldrich, USA
GeneJET Plasmid Miniprep Kit	Thermo Scientific, USA
GeneJET PCR Purification Kit	Thermo Scientific, USA
Glycerol	Fisher Scientific, USA
Glycine	Sigma-Aldrich, USA
Glycogen Azure	Sigma-Aldrich, USA

Reagent (Abbreviations or stock solution used in this work)	**Manufacturer**
Immobilon-P transfer membrane	Merck Millipore, Germany
InstantBlue	Expedeon Ltd, UK
Isopropyl β-D-1-thiogalacto-pyranoside (IPTG)	Glycon Bioch. GmbH, Germany
L-Cysteine	Acros Organics, USA
MBPTrap HP 1mL chromatography columns	GE Healthcare, Sweden
MBPTrap HP 5mL chromatography columns	GE Healthcare, Sweden
Methanol	Merck KGaA, Germany
Monarch DNA Gel Extraction Kit	New England Biolabs, USA
Orange G	TCS Biosciences, UK
Phenol-Chloroform-Isoamyl Alcohol (25:24:1)	Acros Organics, USA
Phenylmethanesulfonylfluoride (PMSF)	AppliChem GmbH, Germany
Pierce 660 nm Protein Assay Reagent	Thermo Scientific, USA
Pierce TBM HRP substrate	Thermo Scientific, USA
Potassium Chloride (KCl)	Fisher Scientific, USA
ProTEV Plus protease	Promega, USA
Rabbit anti-goat IgG, HRP	Thermo Scientific, USA
Sequazyme Peptide Mass Standards Kit	LIfe Technologies, USA
Sodium Chloride (NaCl)	Fisher Scientific, USA
SYBR Safe DNA Gel Stain	Life Technologies, USA

Reagent (Abbreviations or stock solution used in this work)	**Manufacturer**
Tetramethylethylenediamine (TEMED)	Sigma-Aldrich, USA
Triethylammonium bicarbonate buffer (TEAB, 1M solution in H_2O, pH 8.5)	Sigma-Aldrich, USA
Trichloroacetic acid (TCA)	Fisher Scientific, USA
Tricine	Acros Organics, USA
Tris (2-carboxyethyl) phosphine (TCEP)	Sigma-Aldrich, USA
Triton X-100	Sigma-Aldrich, USA
Trizma Base (tris(hydroxymethyl) aminomethane) (Tris)	Sigma-Aldrich, USA
Ultra Low Range Molecular Weight Marker M3546	Sigma-Aldrich, USA

Chapter 9

Appendix B - List of Oligonucleotides used in this work

The list of oligos referred to in the text is contained in the AppendixB.xls file of the enclosed disc.

Chapter 10

Appendix C - Scripts written for data analysis

The scripts written for data analysis in this work, including dataset construction, functional score prediction, NGS data treatment and k-mer based motif detection, are in the "scripts" folder of the attached disc.

Bibliography

Abhinandan, K. R. and Martin, A. C. (2008). Analysis and improvements to kabat and structurally correct numbering of antibody variable domains. *Mol Immunol*, 45(14):3832–9.

Adachi, H., Ohta, T., and Matsuzawa, H. (1991). Site-directed mutants, at position 166, of rtem-1 beta-lactamase that form a stable acyl-enzyme intermediate with penicillin. *J Biol Chem*, 266(5):3186–91.

Afriat, L., Roodveldt, C., Manco, G., and Tawfik, D. S. (2006). The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry*, 45(46):13677–86.

Afriat-Jurnou, L., Jackson, C. J., and Tawfik, D. S. (2012). Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry*, 51(31):6047–55.

Agarwal, V., Pierce, E., McIntosh, J., Schmidt, E. W., and Nair, S. K. (2012). Structures of cyanobactin maturation enzymes define a family of transamidating proteases. *Chem. Biol.*, 19(11):1411–22.

Aharoni, A., Gaidukov, L., Khersonsky, O., Mc, Q. G. S., Roodveldt, C., and Tawfik, D. S. (2005). The 'evolvability' of promiscuous protein functions. *Nat Genet*, 37(1):73–6.

Alberts, I. L., Nadassy, K., and Wodak, S. J. (1998). Analysis of zinc binding sites in protein crystal structures. *Protein Sci*, 7(8):1700–16.

Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. a., Bugni, T. S., Bulaj, G., Camarero, J. a., Campopiano, D. J., Challis, G. L., Clardy, J., Cotter, P. D., Craik, D. J., Dawson, M., Dittmann, E., Donadio, S., Dorrestein, P. C., Entian, K.-D., Fischbach, M. a., Garavelli, J. S., Göransson, U., Gruber, C. W., Haft, D. H., Hemscheidt, T. K., Hertweck, C., Hill, C., Horswill, A. R., Jaspars, M., Kelly, W. L., Klinman, J. P., Kuipers, O. P., Link, a. J., Liu, W., Marahiel, M. a., Mitchell, D. a., Moll, G. N., Moore, B. S., Müller, R., Nair, S. K., Nes, I. F., Norris, G. E., Olivera, B. M., Onaka, H., Patchett, M. L., Piel, J., Reaney, M. J. T., Rebuffat, S., Ross, R. P., Sahl, H.-G., Schmidt, E. W., Selsted, M. E., Severinov, K., Shen, B., Sivonen, K., Smith, L., Stein, T., Süssmuth, R. D., Tagg, J. R., Tang, G.-L., Truman, A. W., Vederas, J. C., Walsh, C. T., Walton, J. D., Wenzel, S. C., Willey, J. M., and van der Donk, W. a. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.*, 30(1):108–60.

Arnold, F. H. (1998). Design by directed evolution. *Accounts of Chemical Research*, 31(3):125–131.

Arora, A. and Scholar, E. M. (2005). Role of tyrosine kinase inhibitors in cancer therapy. *J Pharmacol Exp Ther*, 315(3):971–9.

Arpino, J. A., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J., and Jones, D. D. (2014). Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on gfp folding and structure. *Structure*, 22(6):889–98.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., and Ben-Tal, N. (2016). Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*, 44(W1):W344–50.

Ashraf, M., Frigotto, L., Smith, M. E., Patel, S., Hughes, M. D., Poole, A. J., Hebaishi, H. R., Ullman, C. G., and Hine, A. V. (2013). Proximax randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons. *Biochem Soc Trans*, 41(5):1189–94.

Baker, K., Bleczinski, C., Lin, H., Salazar-Jimenez, G., Sengupta, D., Krane, S., and Cornish, V. W. (2002). Chemical complementation: a reaction-independent genetic assay for enzyme catalysis. *Proc Natl Acad Sci U S A*, 99(26):16537–42.

Baldwin, G., Bayer, T., Dickinson, R., Ellis, T., Polizzi, K., Stan, G. B., Freemont, P., and Kitney, R. (2015). *Synthetic Biology - a Primer.* World Scientific, London, UK.

Balouiri, M., Sadiki, M., and Ibnsouda, S. (2016). Methods for in vitro evaluating antimicrobial activity: A review. *Journal of Pharmaceutical Analysis*, 6(2):71–79.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19(5):455–77.

Barendt, P. A., Ng, D. T., McQuade, C. N., and Sarkar, C. A. (2013). Streamlined protocol for mrna display. *ACS Comb Sci*, 15(2):77–81.

Belshaw, P. J., Roy, R. S., Kelleher, N. L., and Walsh, C. T. (1998). Kinetics and regioselectivity of peptide-to-heterocycle conversions by microcin B17 synthetase. *Chem. Biol.*, 5(7):373–384.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). Genbank. *Nucleic Acids Res*, 41(Database issue):D36–42.

Bent, A. F., Koehnke, J., Houssen, W. E., Smith, M. C. M., Jaspars, M., and Naismith, J. H. (2013). Structure of PatF from Prochloron didemni. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*, 69(Pt 6):618–23.

Bharatham, K., Zhang, Z. H., and Mihalek, I. (2011). Determinants, discriminants, conserved residues–a heuristic approach to detection of functional divergence in protein families. *PLoS One*, 6(9):e24382.

Bibikova, M., Beumer, K., Trautman, J. K., and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. *Science*, 300(5620):764.

Blind, M. and Blank, M. (2015). Aptamer selection technology and recent advances. *Molecular Therapy-Nucleic Acids*, 4.

Bommarius, A. S., Blum, J. K., and Abrahamson, M. J. (2011). Status of protein engineering for biocatalysts: how to design an industrially useful biocatalyst. *Curr. Opin. Chem. Biol.*, 15(2):194–200.

Bordner, A. J. (2009). Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics*, 10:312.

Borneman, A. R., Desany, B. A., Riches, D., Affourtit, J. P., Forgan, A. H., Pretorius, I. S., Egholm, M., and Chambers, P. J. (2011). Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of saccharomyces cerevisiae. *PLoS Genet*, 7(2):e1001287.

Bradley, D., Vieitez, C., Rajeeve, V., Cutillas, P., and Beltrao, P. (2017). Global analysis of specificity determinants in eukaryotic protein kinases. *bioRxiv*.

Breaker, R. R. and Joyce, G. F. (1994). A dna enzyme that cleaves rna. *Chem Biol*, 1(4):223–9.

Burns, J. A., Butler, J. C., Moran, J., and Whitesides, G. M. (1991). Selective reduction of disulfides by tris(2-carboxyethyl)phosphine. *Journal of Organic Chemistry*, 56(8):2648–2650.

Caffrey, D. R., Lunney, E. A., and Moshinsky, D. J. (2008). Prediction of specificity-determining residues for small-molecule kinase inhibitors. *BMC Bioinformatics*, 9:491.

Cakar, Z. P., Seker, U. O., Tamerler, C., Sonderegger, M., and Sauer, U. (2005). Evolutionary engineering of multiple-stress resistant saccharomyces cerevisiae. *FEMS Yeast Res*, 5(6-7):569–78.

Camps, M. (2010). Modulation of cole1-like plasmid replication for recombinant gene expression. *Recent Pat DNA Gene Seq*, 4(1):58–73.

Capra, J. a. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–82.

Carrigan, P. E., Ballar, P., and Tuzmen, S. (2011). Site-directed mutagenesis. *Methods Mol Biol*, 700:107–24.

Chen, C.-Y., Georgiev, I., Anderson, A. C., and Donald, B. R. (2009a). Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U. S. A.*, 106(10):3764–9.

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. (2009b). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol*, 26(7):1523–31.

Chen, R. (2001). Enzyme engineering: rational redesign versus directed evolution. *Trends Biotechnol*, 19(1):13–4.

Chou, H. H. and Keasling, J. D. (2013). Programming adaptive control to evolve increased metabolite production. *Nat Commun*, 4:2595.

Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A. J., and Voytas, D. F. (2010). Targeting dna double-strand breaks with tal effector nucleases. *Genetics*, 186(2):757–61.

Christiansen, A., Kringelum, J. V., Hansen, C. S., Bogh, K. L., Sullivan, E., Patel, J., Rigby, N. M., Eiwegger, T., Szepfalusi, Z., de Masi, F., Nielsen, M., Lund, O., and Dufva, M. (2015). High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci Rep*, 5:12913.

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kuhn, R. (2015). Increasing the efficiency of homology-directed repair for crispr-cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol*, 33(5):543–8.

Cirino, P. C., Mayer, K. M., and Umeno, D. (2003). Generating mutant libraries using error-prone pcr. *Methods Mol Biol*, 231:3–9.

Collin, F., Thompson, R. E., Jolliffe, K. A., Payne, R. J., and Maxwell, A. (2013). Fragments of the bacterial toxin microcin b17 as gyrase poisons. *PLoS One*, 8(4):e61459.

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nat Biotechnol*, 29(11):987–91.

Crameri, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. (1998). Dna shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, 391(6664):288–91.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–90.

Currin, A., Swainston, N., Day, P. J., and Kell, D. B. (2015). Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev*, 44(5):1172–239.

Czarnetzky, E. J., Morgan, I. M., and Mudd, S. (1938). A STABLE HEMOLYSIN-LEUCOCIDIN AND ITS CRYSTAL-LINE DERIVATIVE ISOLATED FROM BETA HEMOLYTIC STREPTOCOCCI. *J. Exp. Med.*, 67(4):643–57.

Dattelbaum, J. D. and Lakowicz, J. R. (2001). Optical determination of glutamine using a genetically engineered protein. *Anal Biochem*, 291(1):89–95.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. (2017). Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res*, 45(D1):D289–D295.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat Rev Genet*, 14(4):249–61.

de Lorimier, R. M., Smith, J. J., Dwyer, M. A., Looger, L. L., Sali, K. M., Paavola, C. D., Rizk, S. S., Sadigov, S., Conrad, D. W., Loew, L., and Hellinga, H. W. (2002). Construction of a fluorescent biosensor family. *Protein Sci*, 11(11):2655–75.

Deane, C. D., Burkhart, B. J., Blair, P. M., Tietz, J. I., Lin, A., and Mitchell, D. A. (2016). In vitro biosynthesis and substrate tolerance of the plantazolicin family of natural products. *ACS Chem Biol*, 11(8):2232–43.

Deane, C. D., Melby, J. O., Molohon, K. J., Susarrey, A. R., and Mitchell, D. A. (2013). Engineering unnatural variants of plantazolicin through codon reprogramming. *ACS Chem. Biol.*, 8(9):1998–2008.

Dellus-Gur, E., Toth-Petroczy, A., Elias, M., and Tawfik, D. S. (2013). What makes a protein fold amenable to functional innovation? fold polarity and stability trade-offs. *J Mol Biol*, 425(14):2609–21.

Dickinson, B. C., Packer, M. S., Badran, A. H., and Liu, D. R. (2014). A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nat Commun*, 5:5352.

Dill, K. A. and MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110):1042–6.

Doi, T., Yoshida, M., Shin-ya, K., and Takahashi, T. (2006). Total synthesis of (R)-telomestatin. *Org. Lett.*, 8(18):4165–7.

Dunbar, K. L., Chekan, J. R., Cox, C. L., Burkhart, B. J., Nair, S. K., and Mitchell, D. a. (2014). Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis. *Nat. Chem. Biol.*, 10(10):823–9.

Dunbar, K. L., Melby, J. O., and Mitchell, D. a. (2012). YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nat. Chem. Biol.*, 8(6):569–75.

Dunbar, K. L. and Mitchell, D. a. (2013). Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *J. Am. Chem. Soc.*, 135(23):8692–701.

Echave, J., Spielman, S. J., and Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*, 17(2):109–21.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Ericsson, U. B., Hallberg, B. M., Detitta, G. T., Dekker, N., and Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem*, 357(2):289–98.

Eswar, N., Eramian, D., Webb, B., Shen, M. Y., and Sali, A. (2008). Protein structure modeling with modeller. *Methods Mol Biol*, 426:145–59.

Faiq, M. A., Ali, M., Dada, T., Dada, R., and Saluja, D. (2014). A novel methodology for enhanced and consistent heterologous expression of unmodified human cytochrome p450 1b1 (cyp1b1). *PLoS One*, 9(10):e110473.

Feng, J., Jester, B. W., Tinberg, C. E., Mandell, D. J., Antunes, M. S., Chari, R., Morey, K. J., Rios, X., Medford, J. I., Church, G. M., Fields, S., and Baker, D. (2015a). A general strategy to construct small molecule biosensors in eukaryotes. *Elife*, 4.

Feng, J., Jester, B. W., Tinberg, C. E., Mandell, D. J., Antunes, M. S., Chari, R., Morey, K. J., Rios, X., Medford, J. I., Church, G. M., Fields, S., and Baker, D. (2015b). A general strategy to construct small molecule biosensors in eukaryotes. *Elife*, 4.

Fernandez-Gacio, A., Uguen, M., and Fastrez, J. (2003). Phage display as a tool for the directed evolution of enzymes. *Trends Biotechnol*, 21(9):408–14.

Fiedurek, J. and Gromada, A. (1997). Selection of biochemical mutants of aspergillus niger with enhanced catalase production. *Appl Microbiol Biotechnol*, 47(3):313–6.

Firth, A. E. and Patrick, W. M. (2005). Statistics of protein library construction. *Bioinformatics*, 21(15):3314–5.

Fowler, D. M. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat Methods*, 11(8):801–7.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Brief Bioinform*, 7(3):225–42.

Fukunishi, Y. and Nakamura, H. (2011). Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. *Protein Sci*, 20(1):95–106.

Fukusaki, E., Hasunuma, T., Kajiyama, S., Okazawa, A., Itoh, T. J., and Kobayashi, A. (2001). Selex for tubulin affords specific t-rich dna aptamers. systematic evolution of ligands by exponeential enrichment. *Bioorg Med Chem Lett*, 11(22):2927–30.

Gardner, S. N. and Hall, B. G. (2013). When whole-genome alignments just won't work: ksnp v2 software for alignment-free snp discovery and phylogenetics of hundreds of microbial genomes. *PLoS One*, 8(12):e81760.

Garrido, M. C., Herrero, M., Kolter, R., and Moreno, F. (1988). The export of the DNA replication inhibitor Microcin B17 provides immunity for the host cell. *EMBO J.*, 7(6):1853–62.

Ghadessy, F. J. and Holliger, P. (2007). Compartmentalized self-replication: a novel method for the directed evolution of polymerases and other enzymes. *Methods Mol Biol*, 352:237–48.

Glavinas, H., Mehn, D., Jani, M., Oosterhuis, B., Heredi-Szabo, K., and Krajcsi, P. (2008). Utilization of membrane vesicle preparations to study drug-abc transporter interactions. *Expert Opin Drug Metab Toxicol*, 4(6):721–32.

Gonzalez, D. J., Lee, S. W., Hensler, M. E., Markley, A. L., Dahesh, S., Mitchell, D. a., Bandeira, N., Nizet, V., Dixon, J. E., and Dorrestein, P. C. (2010a). Clostridiolysin S, a post-translationally modified biotoxin from Clostridium botulinum. *J. Biol. Chem.*, 285(36):28220–8.

Gonzalez, D. J., Lee, S. W., Hensler, M. E., Markley, A. L., Dahesh, S., Mitchell, D. A., Bandeira, N., Nizet, V., Dixon, J. E., and Dorrestein, P. C. (2010b). Clostridiolysin s, a post-translationally modified biotoxin from clostridium botulinum. *J Biol Chem*, 285(36):28220–8.

Goto, Y., Ito, Y., Kato, Y., Tsunoda, S., and Suga, H. (2014). One-pot synthesis of azoline-containing peptides in a cell-free translation system integrated with a posttranslational cyclodehydratase. *Chem. Biol.*, 21(6):766–74.

Gruenwald, K., Holland, J. T., Stromberg, V., Ahmad, A., Watcharakichkorn, D., and Okumoto, S. (2012). Visualization of glutamine transporter activities in living cells using genetically encoded glutamine sensors. *PLoS One*, 7(6):e38591.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320(2):369–87.

Guntas, G., Kanwar, M., and Ostermeier, M. (2012). Circular permutation in the omega-loop of tem-1 beta-lactamase results in improved activity and altered substrate specificity. *PLoS One*, 7(4):e35998.

Guzman, L. M., Belin, D., Carson, M. J., and Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose pbad promoter. *J Bacteriol*, 177(14):4121–30.

Hanes, J. and Pluckthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A*, 94(10):4937–42.

Hayashi, Y., Morimoto, J., and Suga, H. (2012). In vitro selection of anti-Akt2 thioether-macrocyclic peptides leading to isoform-selective inhibitors. *ACS Chem. Biol.*, 7(3):607–13.

Hayes, F., Hallet, B., and Cao, Y. (1997). Insertion mutagenesis as a tool in the modification of protein function. extended substrate specificity conferred by pentapeptide insertions in the omega-loop of tem-1 beta-lactamase. *J Biol Chem*, 272(46):28833–6.

Heck, A. J. (2008). Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Methods*, 5(11):927–33.

Heinis, C., Rutherford, T., Freund, S., and Winter, G. (2009). Phage-encoded combinatorial chemical libraries based on bicyclic peptides. *Nat Chem Biol*, 5(7):502–7.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9.

Heo, L., Shin, W. H., Lee, M. S., and Seok, C. (2014). Galaxysite: ligand-binding-site prediction by using molecular docking. *Nucleic Acids Res*, 42(Web Server issue):W210–4.

Herrero, M. and Moreno, F. (1986). Microcin b17 blocks dna replication and induces the sos system in escherichia coli. *J Gen Microbiol*, 132(2):393–402.

Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., van den Boom, D., Cantor, C. R., and Palsson, B. O. (2006). Comparative genome sequencing of escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nat Genet*, 38(12):1406–12.

Hess, G. T., Fresard, L., Han, K., Lee, C. H., Li, A., Cimprich, K. A., Montgomery, S. B., and Bassik, M. C. (2016). Directed evolution using dcas9-targeted somatic hypermutation in mammalian cells. *Nat Methods*, 13(12):1036–1042.

Hoesl, M. G., Oehm, S., Durkin, P., Darmon, E., Peil, L., Aerni, H. R., Rappsilber, J., Rinehart, J., Leach, D., Soll, D., and Budisa, N. (2015). Chemical evolution of a bacterial proteome. *Angew Chem Int Ed Engl*, 54(34):10030–4.

Hoinka, J., Berezhnoy, A., Dao, P., Sauna, Z. E., Gilboa, E., and Przytycka, T. M. (2015a). Large scale analysis of the mutational landscape in ht-selex improves aptamer discovery. *Nucleic Acids Res*, 43(12):5699–707.

Hoinka, J., Berezhnoy, A., Dao, P., Sauna, Z. E., Gilboa, E., and Przytycka, T. M. (2015b). Large scale analysis of the mutational landscape in ht-selex improves aptamer discovery. *Nucleic Acids Res*, 43(12):5699–707.

Hong, P., Koza, S., and Bouvier, E. S. (2012). Size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. *J Liq Chromatogr Relat Technol*, 35(20):2923–2950.

Hospital, A., Goni, J. R., Orozco, M., and Gelpi, J. L. (2015). Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem*, 8:37–47.

Hurst, J. M., McMillan, L. E., Porter, C. T., Allen, J., Fakorede, A., and Martin, A. C. (2009). The saapdb web resource: a large-scale structural analysis of mutant proteins. *Hum Mutat*, 30(4):616–24.

Huse, W. D., Sastry, L., Iverson, S. A., Kang, A. S., Alting-Mees, M., Burton, D. R., Benkovic, S. J., and Lerner, R. A. (1989). Generation of a large combinatorial library of the immunoglobulin repertoire in phage lambda. *Science*, 246(4935):1275–81.

Jimenez, J. I., Xulvi-Brunet, R., Campbell, G. W., Turk-MacLeod, R., and Chen, I. A. (2013). Comprehensive experimental fitness landscape and evolutionary network for small rna. *Proc Natl Acad Sci U S A*, 110(37):14984–9.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–21.

Jones, D. D. (2005). Triplet nucleotide removal at random positions in a target gene: the tolerance of tem-1 beta-lactamase to an amino acid deletion. *Nucleic Acids Res*, 33(9):e80.

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006.

Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C., and Thornton, J. M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci*, 7(2):233–42.

Kalinina, O. V., Mironov, A. a., Gelfand, M. S., and Rakhmaninova, A. B. (2004). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, 13:443–456.

Kalyon, B., Helaly, S. E., Scholz, R., Nachtigall, J., Vater, J., Borriss, R., and Süssmuth, R. D. (2011). Plantazolicin A and B: structure elucidation of ribosomally synthesized thiazole/oxazole peptides from Bacillus amyloliquefaciens FZB42. *Org. Lett.*, 13(12):2996–9.

Kan, S. B., Lewis, R. D., Chen, K., and Arnold, F. H. (2016a). Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science*, 354(6315):1048–1051.

Kan, S. B., Lewis, R. D., Chen, K., and Arnold, F. H. (2016b). Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science*, 354(6315):1048–1051.

Kan, S. B., Lewis, R. D., Chen, K., and Arnold, F. H. (2016c). Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science*, 354(6315):1048–1051.

Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., Haussler, D., and Sali, A. (2005). Ls-snp: large-scale annotation of coding non-synonymous snps based on multiple information sources. *Bioinformatics*, 21(12):2814–20.

Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically useful: what the rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–98.

Kebschull, J. M. and Zador, A. M. (2015). Sources of pcr-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res*, 43(21):e143.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10(6):845–58.

Kelley, L. a. and Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, 4(3):363–371.

Kelly, J. A., Dideberg, O., Charlier, P., Wery, J. P., Libert, M., Moews, P. C., Knox, J. R., Duez, C., Fraipont, C., Joris, B., and et al. (1986). On the origin of bacterial resistance to penicillin: comparison of a beta-lactamase and a penicillin target. *Science*, 231(4744):1429–31.

Kelly, W. L., Pan, L., and Li, C. (2009). Thiostrepton biosynthesis: prototype for a new family of bacteriocins. *J. Am. Chem. Soc.*, 131(12):4327–34.

Khan, T., Douglas, G. M., Patel, P., Nguyen Ba, A. N., and Moses, A. M. (2015). Polymorphism analysis reveals reduced negative selection and elevated rate of insertions and deletions in intrinsically disordered protein regions. *Genome Biol Evol*, 7(6):1815–26.

Kipnis, Y., Dellus-Gur, E., and Tawfik, D. S. (2012). Trins: a method for gene modification by randomized tandem repeat insertions. *Protein Eng Des Sel*, 25(9):437–44.

Kiss, G., Pande, V. S., and Houk, K. N. (2013). Molecular dynamics simulations for the ranking, evaluation, and refinement of computationally designed proteins. *Methods Enzymol*, 523:145–70.

Koehnke, J., Bent, A., Houssen, W. E., Zollman, D., Morawitz, F., Shirran, S., Vendome, J., Nneoyiegbe, A. F., Trembleau, L., Botting, C. H., Smith, M. C. M., Jaspars, M., and Naismith, J. H. (2012). The mechanism of patellamide macrocyclization revealed by the characterization of the PatG macrocyclase domain. *Nat. Struct. Mol. Biol.*, 19(8):767–72.

Koehnke, J., Bent, A. F., Zollman, D., Smith, K., Houssen, W. E., Zhu, X., Mann, G., Lebl, T., Scharff, R., Shirran, S., Botting, C. H., Jaspars, M., Schwarz-Linek, U., and Naismith, J. H. (2013). The cyanobactin heterocyclase enzyme: a processive adenylase that operates with a defined order of reaction. *Angew. Chem. Int. Ed. Engl.*, 52(52):13991–6.

Koplovitz, G., McClintock, J. B., Amsler, C. D., and Baker, B. J. (2011). A comprehensive evaluation of the potential chemical defenses of antarctic ascidians against sympatric fouling microorganisms. *Mar. Biol.*, 158(12):2661–2671.

Krause, J. C., Ekiert, D. C., Tumpey, T. M., Smith, P. B., Wilson, I. A., and Crowe, J. E., J. (2011). An insertion mutation that distorts antibody binding site architecture enhances function of a human antibody. *MBio*, 2(1):e00345–10.

Kucera, R. and Evans, T. (2014). Ligation enhancement. *USPTO*, (US20120283144 A1).

Lacatena, R. M. and Cesareni, G. (1981). Base pairing of rna i with its complementary sequence in the primer precursor inhibits cole1 replication. *Nature*, 294(5842):623–6.

Lahiry, P., Torkamani, A., Schork, N. J., and Hegele, R. A. (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat Rev Genet*, 11(1):60–74.

Lane, M. D. and Seelig, B. (2014). Advances in the directed evolution of proteins. *Curr Opin Chem Biol*, 22:129–36.

Lee, H., Popodi, E., Tang, H., and Foster, P. L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A*, 109(41):E2774–83.

Lee, J., Saddler, J. N., Um, Y., and Woo, H. M. (2016). Adaptive evolution and metabolic engineering of a cellobiose- and xylose- negative corynebacterium glutamicum that co-utilizes cellobiose and xylose. *Microb Cell Fact*, 15:20.

Lee, S. W., Mitchell, D. a., Markley, A. L., Hensler, M. E., Gonzalez, D., Wohlrab, A., Dorrestein, P. C., Nizet, V., and Dixon, J. E. (2008). Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc. Natl. Acad. Sci. U. S. A.*, 105(15):5879–84.

Leemhuis, H., Stein, V., Griffiths, A. D., and Hollfelder, F. (2005). New genotype-phenotype linkages for directed evolution of functional proteins. *Curr Opin Struct Biol*, 15(4):472–8.

Li, G. Y., Zhang, H., Sun, Z. T., Liu, X. Q., and Reetz, M. T. (2016). Multiparameter optimization in directed evolution: Engineering thermostability, enantioselectivity, and activity of an epoxide hydrolase. *Acs Catalysis*, 6(6):3679–3687.

Li, Y. (2009). Carrier proteins for fusion expression of antimicrobial peptides in escherichia coli. *Biotechnol Appl Biochem*, 54(1):1–9.

Li, Y., Milne, J. C., Madison, L. L., Kolter, R., and Walsh, C. T. (1996). From Peptide Precursors to Oxazole and Thiazole-Containing Peptide Antibiotics: Microcin B17 Synthase. *Science*, 274(5290):1188–1193.

Lilien, R. H., Stevens, B. W., Anderson, A. C., and Donald, B. R. (2005). A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *J Comput Biol*, 12(6):740–61.

Lin, Y. R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., and Baker, D. (2015). Control over overall shape and size in de novo designed proteins. *Proc Natl Acad Sci U S A*, 112(40):E5478–85.

Linder, J., Garner, T. P., Williams, H. E. L., Searle, M. S., and Moody, C. J. (2011). Telomestatin: formal total synthesis and cation-mediated interaction of its seco-derivatives with G-quadruplexes. *J. Am. Chem. Soc.*, 133(4):1044–51.

Lipovsek, D., Antipov, E., Armstrong, K. A., Olsen, M. J., Klibanov, A. M., Tidor, B., and Wittrup, K. D. (2007). Selection of horseradish peroxidase variants with enhanced enantioselectivity by yeast surface display. *Chem Biol*, 14(10):1176–85.

Liu, C. E., Liu, P. Q., and Ames, G. F. (1997). Characterization of the adenosine triphosphatase activity of the periplasmic histidine permease, a traffic atpase (abc transporter). *J Biol Chem*, 272(35):21883–91.

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

Livingstone, C. D. and Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 9(6):745–56.

Lockless, S. W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–9.

Lohman, G. and Nichols, N. (2015).

Looger, L. L., Dwyer, M. A., Smith, J. J., and Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, 423(6936):185–90.

Lutz, S. (2010). Beyond directed evolution–semi-rational protein engineering and design. *Curr Opin Biotechnol*, 21(6):734–43.

Lv, D. S., Gong, W. K., Zhang, Y., Liu, Y., and Li, C. H. (2017). A coarse-grained method to predict the open-to-closed behavior of glutamine binding protein. *Chemical Physics*, 493:166–174.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., and Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*, 105(27):9272–7.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y., and Bryant, S. H. (2017). Cdd/sparcle: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*, 45(D1):D200–D203.

Marino Buslje, C., Teppa, E., Di Doménico, T., Delfino, J. M., and Nielsen, M. (2010). Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol.*, 6(11):e1000978.

Marliere, P., Patrouix, J., Doring, V., Herdewijn, P., Tricot, S., Cruveiller, S., Bouzon, M., and Mutzel, R. (2011). Chemical evolution of a bacterium's genome. *Angew Chem Int Ed Engl*, 50(31):7109–14.

Marson, C. M. and Saadi, M. (2006). Synthesis of the penta-oxazole core of telomestatin in a convergent approach to poly-oxazole macrocycles. *Org. Biomol. Chem.*, 4(21):3892–3.

Martin, L. C., Gloor, G. B., Dunn, S. D., and Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124.

Melby, J. O., Dunbar, K. L., Trinh, N. Q., and Mitchell, D. A. (2012). Selectivity, directionality, and promiscuity in peptide processing from a Bacillus sp. Al Hakam cyclodehydratase. *J. Am. Chem. Soc.*, 134(11):5309–16.

Melby, J. O., Li, X., and Mitchell, D. A. (2014). Orchestration of enzymatic processing by thiazole/oxazole-modified microcin dehydrogenases. *Biochemistry*, 53(2):413–22.

Meroueh, S. O., Fisher, J. F., Schlegel, H. B., and Mobashery, S. (2005). Ab initio qm/mm study of class a beta-lactamase acylation: dual participation of glu166 and lys73 in a concerted base promotion of ser70. *J Am Chem Soc*, 127(44):15397–407.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46.

Miller, D. A., Luo, L., Hillson, N., Keating, T. A., and Walsh, C. T. (2002). Yersiniabactin synthetase: a four-protein assembly line producing the nonribosomal peptide/polyketide hybrid siderophore of Yersinia pestis. *Chem. Biol.*, 9(3):333–44.

Miller, M. C., Parkin, S., Fetherston, J. D., Perry, R. D., and Demoll, E. (2006). Crystal structure of ferric-yersiniabactin, a virulence factor of Yersinia pestis. *J. Inorg. Biochem.*, 100(9):1495–500.

Milne, J., Roy, R., Eliot, A., and Kelleher, N. (1999). Cofactor requirements and reconstitution of microcin B17 synthetase: a multienzyme complex that catalyzes the formation of oxazoles and thiazoles in the antibiotic. *Biochemistry*, 38(15):4768–81.

Milne, J. C., Eliot, A. C., Kelleher, N. L., and Walsh, C. T. (1998). ATP/GTP hydrolysis is required for oxazole and thiazole biosynthesis in the peptide antibiotic microcin B17. *Biochemistry*, 37(38):13250–61.

Minasov, G., Wang, X., and Shoichet, B. K. (2002). An ultrahigh resolution structure of tem-1 beta-lactamase suggests a role for glu166 as the general base in acylation. *J Am Chem Soc*, 124(19):5333–40.

Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., ten Hoopen, P., Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., Sterk, P., and Finn, R. D. (2016). Ebi metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res*, 44(D1):D595–603.

Mitchell, D. a., Lee, S. W., Pence, M. a., Markley, A. L., Limm, J. D., Nizet, V., and Dixon, J. E. (2009). Structural and functional dissection of the heterocyclic peptide cytotoxin streptolysin S. *J. Biol. Chem.*, 284(19):13004–12.

Moelbert, S., Emberly, E., and Tang, C. (2004). Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci*, 13(3):752–62.

Molohon, K. J., Blair, P. M., Park, S., Doroghazi, J. R., Maxson, T., Hershfield, J. R., Flatt, K. M., Schroeder, N. E., Ha, T., and Mitchell, D. A. (2016). Plantazolicin is an ultra-narrow spectrum antibiotic that targets the bacillus anthracis membrane. *ACS Infect Dis*, 2(3):207–220.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, 108(49):E1293–301.

Moretti, S., Armougom, F., Wallace, I. M., Higgins, D. G., Jongeneel, C. V., and Notredame, C. (2007). The M-Coffee web server: A meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Res.*, 35(2):645–648.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–91.

Mulder, N. J. and Apweiler, R. (2002). Tools and resources for identifying protein families, domains and motifs. *Genome Biol*, 3(1):REVIEWS2001.

Muteeb, G. and Sen, R. (2010). Random mutagenesis using a mutator strain. *Methods Mol Biol*, 634:411–9.

Nguyen, U. T., Bittova, L., Muller, M. M., Fierz, B., David, Y., Houck-Loomis, B., Feng, V., Dann, G. P., and Muir, T. W. (2014). Accelerated chromatin biochemistry using dna-barcoded nucleosome libraries. *Nat Methods*, 11(8):834–40.

Nixon, A. E. and Firestine, S. M. (2000). Rational and "irrational" design of proteins and their use in biotechnology. *IUBMB Life*, 49(3):181–7.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217.

Nuin, P. A., Wang, Z., and Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, 7:471.

Oh, B. H., Ames, G. F., and Kim, S. H. (1994). Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. *J Biol Chem*, 269(42):26323–30.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23:263–286.

Osuna, J., Yanez, J., Soberon, X., and Gaytan, P. (2004). Protein evolution by codon-based random deletions. *Nucleic Acids Res*, 32(17):e136.

Pace, N. J. and Weerapana, E. (2014). Zinc-binding cysteines: diverse functions and structural motifs. *Biomolecules*, 4(2):419–34.

Packer, M. S. and Liu, D. R. (2015). Methods for the directed evolution of proteins. *Nat Rev Genet*, 16(7):379–94.

Palzkill, T. and Botstein, D. (1992). Probing beta-lactamase structure and function using random replacement mutagenesis. *Proteins*, 14(1):29–44.

Palzkill, T., Le, Q. Q., Venkatachalam, K. V., LaRocco, M., and Ocera, H. (1994a). Evolution of antibiotic resistance: several different amino acid substitutions in an active site loop alter the substrate profile of beta-lactamase. *Mol Microbiol*, 12(2):217–29.

Palzkill, T., Le, Q. Q., Venkatachalam, K. V., LaRocco, M., and Ocera, H. (1994b). Evolution of antibiotic resistance: several different amino acid substitutions in an active site loop alter the substrate profile of beta-lactamase. *Mol Microbiol*, 12(2):217–29.

Park, J. H., Choi, E. A., Cho, E. W., Hahm, K. S., and Kim, K. L. (1998). Maltose binding protein (mbp) fusion proteins with low or no affinity to amylose resins can be single-step purified using a novel anti-mbp monoclonal antibody. *Mol Cells*, 8(6):709–16.

Patani, H., Bunney, T. D., Thiyagarajan, N., Norman, R. A., Ogg, D., Breed, J., Ashford, P., Potterton, A., Edwards, M., Williams, S. V., Thomson, G. S., Pang, C. S., Knowles, M. A., Breeze, A. L., Orengo, C., Phillips, C., and Katan, M. (2016). Landscape of activating cancer mutations in fgfr kinases and their differential responses to inhibitors in clinical use. *Oncotarget*, 7(17):24252–68.

Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B., and Keasling, J. D. (2012). Microbial engineering for the production of advanced biofuels. *Nature*, 488(7411):320–8.

Petrosino, J. F. and Palzkill, T. (1996). Systematic mutagenesis of the active site omega loop of tem-1 beta-lactamase. *J Bacteriol*, 178(7):1821–8.

Piers, K. L., Brown, M. H., and Hancock, R. E. (1993). Recombinant dna procedures for producing small antimicrobial cationic peptides in bacteria. *Gene*, 134(1):7–13.

Pinheiro, V. B., Taylor, A. I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J. C., Wengel, J., Peak-Chew, S. Y., McLaughlin, S. H., Herdewijn, P., and Holliger, P. (2012a). Synthetic genetic polymers capable of heredity and evolution. *Science*, 336(6079):341–4.

Pinheiro, V. B., Taylor, A. I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J. C., Wengel, J., Peak-Chew, S.-Y., McLaughlin, S. H., Herdewijn, P., and Holliger, P. (2012b). Synthetic genetic polymers capable of heredity and evolution. *Science*, 336(6079):341–4.

Pires, D. E., Chen, J., Blundell, T. L., and Ascher, D. B. (2016). In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep*, 6:19848.

Pistolesi, S. and Tjandra, N. (2012). Temperature dependence of molecular interactions involved in defining stability of glutamine binding protein and its complex with l-glutamine. *Biochemistry*, 51(2):643–52.

Polz, M. F. and Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate pcr. *Appl Environ Microbiol*, 64(10):3724–30.

Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32(Database issue):D129–33.

Prassler, J., Thiel, S., Pracht, C., Polzer, A., Peters, S., Bauer, M., Norenberg, S., Stark, Y., Kolln, J., Popp, A., Urlinger, S., and Enzelberger, M. (2011). Hucal platinum, a synthetic fab library optimized for sequence diversity and superior performance in mammalian expression systems. *J Mol Biol*, 413(1):261–78.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1:S71–7.

Ralser, M., Querfurth, R., Warnatz, H. J., Lehrach, H., Yaspo, M. L., and Krobitsch, S. (2006). An efficient and economic enhancer mix for pcr. *Biochem Biophys Res Commun*, 347(3):747–51.

Ravn, U., Didelot, G., Venet, S., Ng, K. T., Gueneau, F., Rousseau, F., Calloud, S., Kosco-Vilbois, M., and Fischer, N. (2013). Deep sequencing of phage display libraries to support antibody discovery. *Methods*, 60(1):99–110.

Ravn, U., Gueneau, F., Baerlocher, L., Osteras, M., Desmurs, M., Malinge, P., Magistrelli, G., Farinelli, L., Kosco-Vilbois, M. H., and Fischer, N. (2010). By-passing in vitro screening–next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res*, 38(21):e193.

Reetz, M. T. and Carballeira, J. D. (2007). Iterative saturation mutagenesis (ism) for rapid directed evolution of functional enzymes. *Nat Protoc*, 2(4):891–903.

Reis, A. C. and Salis, H. (2017). An automated model test system for systematic development and improvement of gene expression models. *bioRxiv*.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6):276–7.

Robertson, D. L. and Joyce, G. F. (1990). Selection in vitro of an rna enzyme that specifically cleaves single-stranded dna. *Nature*, 344(6265):467–8.

Rosano, G. L. and Ceccarelli, E. A. (2014). Recombinant protein expression in escherichia coli: advances and challenges. *Front Microbiol*, 5:172.

Ruffner, D. E., Schmidt, E. W., and Heemstra, J. R. (2014). Assessing the combinatorial potential of the RiPP cyanobactin tru pathway. *ACS Synth. Biol.*

Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*, 27(10):946–50.

Sambrook, J. and Russell, D. (2001). *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press.

Sandana Mala, J. G., Kamini, N. R., and Puvanakrishnan, R. (2001). Strain improvement of aspergillus niger for enhanced lipase production. *J Gen Appl Microbiol*, 47(4):181–186.

Sardar, D., Pierce, E., McIntosh, J. A., and Schmidt, E. W. (2014). Recognition Sequences and Substrate Evolution in Cyanobactin Biosynthesis. *ACS Synth. Biol.*, 24(17):1639–41.

Sauvage, E. and Terrak, M. (2016). Glycosyltransferases and transpeptidases/penicillin-binding proteins: Valuable targets for new antibacterials. *Antibiotics (Basel)*, 5(1).

Schagger, H. (2006). Tricine-sds-page. *Nat Protoc*, 1(1):16–22.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nat Methods*, 9(7):676–82.

Schmidt, E. W., Nelson, J. T., Rasko, D. a., Sudek, S., Eisen, J. a., Haygood, M. G., and Ravel, J. (2005). Patellamide A and C biosynthesis by a microcin-like pathway in Prochloron didemni, the cyanobacterial symbiont of Lissoclinum patella. *Proc. Natl. Acad. Sci. U. S. A.*, 102(20):7315–20.

Scholz, R., Molohon, K. J., Nachtigall, J., Vater, J., Markley, A. L., Süssmuth, R. D., Mitchell, D. a., and Borriss, R. (2011). Plantazolicin, a novel microcin B17/streptolysin S-like natural product from Bacillus amyloliquefaciens FZB42. *J. Bacteriol.*, 193(1):215–24.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The foldx web server: an online force field. *Nucleic Acids Res*, 33(Web Server issue):W382–8.

Seyfang, A. and Jin, J. H. (2004). Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round. *Anal Biochem*, 324(2):285–91.

Shafer, D. E., Inman, J. K., and Lees, A. (2000). Reaction of tris(2-carboxyethyl)phosphine (tcep) with maleimide and alpha-haloacyl groups: anomalous elution of tcep by gel filtration. *Anal Biochem*, 282(1):161–4.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623–656.

Shchemelinin, I., Sefc, L., and Necas, E. (2006). Protein kinases, their function and implication in cancer and other diseases. *Folia Biol (Praha)*, 52(3):81–100.

Shin-ya, K., Wierzba, K., Matsuo, K.-i., Ohtani, T., Yamada, Y., Furihata, K., Hayakawa, Y., and Seto, H. (2001). Telomestatin, a novel telomerase inhibitor from Streptomyces anulatus. *J. Am. Chem. Soc.*, 123(6):1262–3.

Shkundina, I., Serebryakova, M., and Severinov, K. (2014). The c-terminal part of microcin b is crucial for dna gyrase inhibition and antibiotic uptake by sensitive cells. *J Bacteriol*, 196(9):1759–67.

Sideraki, V., Huang, W., Palzkill, T., and Gilbert, H. F. (2001). A secondary drug resistance mutation of tem-1 beta-lactamase that suppresses misfolding and aggregation. *Proc Natl Acad Sci U S A*, 98(1):283–8.

Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, 329(5989):309–13.

Siegele, D. A. and Hu, J. C. (1997). Gene expression from plasmids containing the arabad promoter at subsaturating inducer concentrations represents mixed populations. *Proc Natl Acad Sci U S A*, 94(15):8168–72.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7(539).

Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M., and Marino Buslje, C. (2013). MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res.*, 41(Web Server issue):W8–14.

Singh, J., Kumar, M., Mansuri, R., Sahoo, G. C., and Deep, A. (2016). Inhibitor designing, virtual screening, and docking studies for methyltransferase: A potential target against dengue virus. *J Pharm Bioallied Sci*, 8(3):188–94.

Singh, T., Biswas, D., and Jayaram, B. (2011). Aads–an automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. *J Chem Inf Model*, 51(10):2515–27.

Sinha Roy, R., Belshaw, P. J., and Walsh, C. T. (1998). Mutational analysis of posttranslational heterocycle biosynthesis in the gyrase inhibitor microcin B17: distance dependence from propeptide and tolerance for substitution in a GSCG cyclizable sequence. *Biochemistry*, 37(12):4125–36.

Sinha Roy, R., Kelleher, N. L., Milne, J. C., and Walsh, C. T. (1999). In vivo processing and antibiotic activity of microcin B17 analogs with varying ring content and altered bisheterocyclic sites. *Chem. Biol.*, 6(5):305–18.

Sinha Roy, R., Kelleher, N. L., Milne, J. C., and Walsh, C. T. (1999). In vivo processing and antibiotic activity of microcin b17 analogs with varying ring content and altered bisheterocyclic sites. *Chem Biol*, 6(5):305–18.

Skjoedt, M. L., Snoek, T., Kildegaard, K. R., Arsovska, D., Eichenberger, M., Goedecke, T. J., Rajkumar, A. S., Zhang, J., Kristensen, M., Lehka, B. J., Siedler, S., Borodina, I., Jensen, M. K., and Keasling, J. D. (2016). Engineering prokaryotic transcriptional activators as metabolite biosensors in yeast. *Nat Chem Biol*, 12(11):951–958.

Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–7.

Sowek, J. A., Singer, S. B., Ohringer, S., Malley, M. F., Dougherty, T. J., Gougoutas, J. Z., and Bush, K. (1991). Substitution of lysine at position 104 or 240 of tem-1ptz18r beta-lactamase enhances the effect of serine-164 substitution on hydrolysis or affinity for cephalosporins and the monobactam aztreonam. *Biochemistry*, 30(13):3179–88.

Stanley, D., Chambers, P. J., Stanley, G. A., Borneman, A., and Fraser, S. (2010). Transcriptional changes associated with ethanol tolerance in saccharomyces cerevisiae. *Appl Microbiol Biotechnol*, 88(1):231–9.

Stemmer, W. P. (1994). Dna shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A*, 91(22):10747–51.

Stojanoski, V., Chow, D. C., Hu, L., Sankaran, B., Gilbert, H. F., Prasad, B. V., and Palzkill, T. (2015). A triple mutant in the omega-loop of tem-1 beta-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis. *J Biol Chem*, 290(16):10382–94.

Stoltenburg, R., Reinemann, C., and Strehlitz, B. (2007). Selex–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng*, 24(4):381–403.

Su, T., Liu, F., Gu, P., Jin, H., Chang, Y., Wang, Q., Liang, Q., and Qi, Q. (2016). A crispr-cas9 assisted non-homologous end-joining strategy for one-step engineering of bacterial genome. *Sci Rep*, 6:37895.

Süel, G. M., Lockless, S. W., Wall, M. a., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10(1):59–69.

Sun, Y. J., Rose, J., Wang, B. C., and Hsiao, C. D. (1998). The structure of glutamine-binding protein complexed with glutamine at 1.94 a resolution: comparisons with other amino acid binding proteins. *J Mol Biol*, 278(1):219–29.

Sunyaev, S., Vasily, R., Koch, I., Lathe III, W., Kondrashov, A., and Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, 10(6):591–597.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, 43(Database issue):D447–52.

Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M., and Jiang, R. (2012). Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *Biotechniques*, 52(3):149–58.

Tauchi, T., Shin-Ya, K., Sashida, G., Sumi, M., Nakajima, A., Shimamoto, T., Ohyashiki, J. H., and Ohyashiki, K. (2003). Activity of a novel G-quadruplex-interactive telomerase inhibitor, telomestatin (SOT-095), against human leukemia cells: involvement of ATM-dependent DNA damage response pathways. *Oncogene*, 22(34):5338–47.

Taylor, W. R. (1986). The classification of amino acid conservation. *J Theor Biol*, 119(2):205–18.

Tee, K. L. and Wong, T. S. (2013). Polishing the craft of genetic diversity creation in directed evolution. *Biotechnol Adv*, 31(8):1707–21.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Medigue, C., Schneider, D., and Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615):165–70.

Teunissen, A., Dumortier, F., Gorwa, M. F., Bauer, J., Tanghe, A., Loiez, A., Smet, P., Van Dijck, P., and Thevelein, J. M. (2002). Isolation and characterization of a freeze-tolerant diploid derivative of an industrial baker's yeast strain and its use in frozen doughs. *Appl Environ Microbiol*, 68(10):4780–7.

Thompson, R. E., Collin, F., Maxwell, A., Jolliffe, K. A., and Payne, R. J. (2014). Synthesis of full length and truncated microcin b17 analogues as dna gyrase poisons. *Org Biomol Chem*, 12(10):1570–8.

Tian, J., Liu, Q., Dong, S., Qiao, X., and Ni, J. (2010). A new method for multi-site-directed mutagenesis. *Anal Biochem*, 406(1):83–5.

Tizei, P. A. G., Csibra, E., Torres, L., and Pinheiro, V. B. (2016). Selection platforms for directed evolution in synthetic biology. *Biochem Soc Trans*, 44(4):1165–75.

Tizei, P. A. G., Harris, E., Renders, M., and Pinheiro, V. B. (2017). Efficiently exploring functional space in loop engineering with variations in length and composition. *bioRxiv*.

Toprak, E., Veres, A., Michel, J. B., Chait, R., Hartl, D. L., and Kishony, R. (2011). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet*, 44(1):101–5.

Toth-Petroczy, A. and Tawfik, D. S. (2014a). Hopeful (protein indel) monsters? *Structure*, 22(6):803–4.

Toth-Petroczy, A. and Tawfik, D. S. (2014b). The robustness and innovability of protein folds. *Curr Opin Struct Biol*, 26:131–8.

Tran, J. H., Jacoby, G. a., and Hooper, D. C. (2005). Interaction of the plasmid-encoded quinolone resistance protein Qnr with Escherichia coli DNA gyrase. *Antimicrob. Agents Chemother.*, 49(1):118–25.

Trevino, S. R., Scholtz, J. M., and Pace, C. N. (2007). Amino acid contribution to protein solubility: Asp, glu, and ser contribute more favorably than the other hydrophilic amino acids in rnase sa. *J Mol Biol*, 366(2):449–60.

Trindade, S., Perfeito, L., and Gordo, I. (2010). Rate and effects of spontaneous mutations that affect fitness in mutator escherichia coli. *Philos Trans R Soc Lond B Biol Sci*, 365(1544):1177–86.

Valdar, W. S. (2002). Scoring residue conservation. *Proteins*, 48(2):227–41.

Valdar, W. S. and Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–24.

van Bloois, E., Winter, R. T., Kolmar, H., and Fraaije, M. W. (2011). Decorating microbes: surface display of proteins on escherichia coli. *Trends Biotechnol*, 29(2):79–86.

Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S., Fuhrmann, M., Schatz, O., O'Connell, T., O'Connell, D., Auckenthaler, A., and Schwer, H. (2008). A novel solid phase technology for high-throughput gene synthesis. *Biotechniques*, 45(3):340–3.

Vant-Hull, B., Gold, L., and Zichi, D. A. (2000). Theoretical principles of in vitro selection using combinatorial nucleic acid libraries. *Curr Protoc Nucleic Acid Chem*, Chapter 9:Unit 9 1.

Velmurugan, D., Mythily, U., and Rao, K. (2014). Design and docking studies of peptide inhibitors as potential antiviral drugs for dengue virus ns2b/ns3 protease. *Protein Pept Lett*, 21(8):815–27.

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics*, 19(4):513–23.

von Ossowski, I., Stahlberg, J., Koivula, A., Piens, K., Becker, D., Boer, H., Harle, R., Harris, M., Divne, C., Mahdi, S., Zhao, Y., Driguez, H., Claeyssens, M., Sinnott, M. L., and Teeri, T. T. (2003). Engineering the exo-loop of trichoderma reesei cellobiohydrolase, cel7a. a comparison with phanerochaete chrysosporium cel7d. *J Mol Biol*, 333(4):817–29.

Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., and Church, G. M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, 460(7257):894–8.

Wang, K. and Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, 7:385.

Waterhouse, A. M., Procter, J. B., Martin, D. M. a., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191.

Waugh, D. S. (2011). An overview of enzymatic reagents for the removal of affinity tags. *Protein Expr Purif*, 80(2):283–93.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A*, 106(1):67–72.

Weld, J. T. (1934). THE TOXIC PROPERTIES OF SERUM EXTRACTS OF HEMOLYTIC STREPTOCOCCI. *J. Exp. Med.*, 59(1):83–95.

Whitehead, T. A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., Myers, C. A., Kamisetty, H., Blair, P., Wilson, I. A., and Baker, D. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol*, 30(6):543–8.

Wilson, D. S., Keefe, a. D., and Szostak, J. W. (2001). The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl. Acad. Sci. U. S. A.*, 98:3750–3755.

Wipf, P. and Uto, Y. (1999). Total synthesis of the putative structure of the marine metabolite trunkamide A. *Tetrahedron Lett.*, 40(28):5165–5169.

Wisselink, H. W., Toirkens, M. J., del Rosario Franco Berriel, M., Winkler, A. A., van Dijken, J. P., Pronk, J. T., and van Maris, A. J. (2007). Engineering of saccharomyces cerevisiae for efficient anaerobic alcoholic fermentation of l-arabinose. *Appl Environ Microbiol*, 73(15):4881–91.

Woldring, D. R., Holec, P. V., Stern, L. A., Du, Y., and Hackel, B. J. (2017). A gradient of sitewise diversity promotes evolutionary fitness for binder discovery in a three-helix bundle protein scaffold. *Biochemistry*, 56(11):1656–1671.

Woldring, D. R., Holec, P. V., Zhou, H., and Hackel, B. J. (2015). High-throughput ligand discovery reveals a sitewise gradient of diversity in broadly evolved hydrophilic fibronectin domains. *PLoS One*, 10(9):e0138956.

Wolf, S. and Grunewald, S. (2015). Sequence, structure and ligand binding evolution of rhodopsin-like g protein-coupled receptors: a crystal structure-based phylogenetic analysis. *PLoS One*, 10(4):e0123533.

Wong, A., Rodrigue, N., and Kassen, R. (2012). Genomics of adaptation during experimental evolution of the opportunistic pathogen pseudomonas aeruginosa. *PLoS Genet*, 8(9):e1002928.

Yamaguchi, J., Naimuddin, M., Biyani, M., Sasaki, T., Machida, M., Kubo, T., Funatsu, T., Husimi, Y., and Nemoto, N. (2009). cdna display: a novel screening method for functional disulfide-rich peptides by solid-phase synthesis and stabilization of mrna-protein fusions. *Nucleic Acids Res*, 37(16):e108.

Yorgey, P., Davagnino, J., and Kolter, R. (1993). The maturation pathway of microcin B17, a peptide inhibitor of DNA gyrase. *Mol. Microbiol.*, 9(4):897–905.

You, C. and Percival Zhang, Y. H. (2012). Easy preparation of a large-size random gene mutagenesis library in escherichia coli. *Anal Biochem*, 428(1):7–12.

Young, T. S., Dorrestein, P. C., and Walsh, C. T. (2012). Codon randomization for rapid exploration of chemical space in thiopeptide antibiotic variants. *Chem. Biol.*, 19(12):1600–10.

Young, T. S. and Walsh, C. T. (2011). Identification of the thiazolyl peptide GE37468 gene cluster from Streptomyces ATCC 55365 and heterologous expression in Streptomyces lividans. *Proc. Natl. Acad. Sci. U. S. A.*, 108(32):13053–8.

Yu, C. Y., Li, X. X., Yang, H., Li, Y. H., Xue, W. W., Chen, Y. Z., Tao, L., and Zhu, F. (2018). Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. *Int J Mol Sci*, 19(1).

Zamble, D. B., Miller, D. A., Heddle, J. G., Maxwell, A., Walsh, C. T., and Hollfelder, F. (2001). In vitro characterization of dna gyrase inhibition by microcin b17 analogs with altered bisheterocyclic sites. *Proc Natl Acad Sci U S A*, 98(14):7712–7.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–9.

Zhao, H. and Arnold, F. H. (1997). Combinatorial protein design: strategies for screening protein libraries. *Curr Opin Struct Biol*, 7(4):480–5.

Zinchenko, A., Devenish, S. R., Kintses, B., Colin, P. Y., Fischlechner, M., and Hollfelder, F. (2014). One in a million: flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Anal Chem*, 86(5):2526–33.