Evolvability of a Viral Protease
Experimental Evolution of Catalysis, Robustness and Specificity

THOMAS SHAFFEE

Department of Biochemistry and Gonville and Caius College,
University of Cambridge

Under the supervision of F. Hollfelder and R. Minter

A thesis submitted for the degree of Doctor of Philosophy

September of 2013
Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This work is not being submitted for another degree or qualification at any academic institution and does not exceed 60,000 words.

The work presented in Chapter C has been submitted to *Nature Genetics* as ‘The evolvability of nucleophile exchange’ (Shafee, T., Gatti-Lafraconi, P., Minter, R., Hollfelder, F. 2013)

The work in Chapters D & E have been combined and prepared for submission to Molecular Biology and Evolution as ‘Neutral evolution – two combined mechanisms generate robust and evolvable enzymes’ (Shafee, T., Gatti-Lafraconi, P., Minter, R., Hollfelder, F. 2013)

Thomas Shafee, 27th of September 2013
I am very grateful for the help and guidance I’ve received throughout my PhD. I have been given great intellectual freedom by my supervisor, Dr Florian Hollfelder, who has always greeted my ideas and research with enthusiasm, encouragement and support. I would also like to thank my second supervisor, Dr Ralph Minter, for his insightful suggestions and help through my projects.

In particular, I would like to thank Dr Pietro Gatti-Lafranconi who has helped me with everything from pipetting to manuscript preparation and his advice and input has been invaluable in all of my projects. Additionally, Nigel Miller’s patient help with flow cytometry has made many of my experiments possible.

I also thank the excellent PostDoc community in the group for lively debate and useful comments. Specifically, I have appreciated the attention to detail of Dr Sean Devenish, the encyclopaedic literature knowledge of Dr Stephane Emond, the level-headed advice of Dr Maren Butz, the tireless enthusiasm of Dr Albert Kwok, and the chemical know-how of Dr Mark Mohammed. I’ve also been fortunate to have the support and companionship of a superb peer group of PhD students, in particular Charlie Miton who has always gone out of her way to be helpful and Chris Bayer with whom I’ve had many a fascinating discussion.

I have been in Cambridge for seven years of undergraduate and graduate life and am hugely grateful to the people who have contributed to this personally and intellectually stimulating time in my life. I would like to thank my parents, brother, uncles and aunts (biological and appointed) and Grandmouse for their fantastic support and for being interested in my work even if, on occasion, they weren’t sure what I was talking about. Lastly, I would like to thank my wonderful partner Katrina, without whom my world would be a less loving, less fascinating and disappointingly sensible place.
Abstract

The aim of this thesis is to investigate aspects of molecular evolution and enzyme engineering using the experimental evolution of Tobacco Etch Virus cysteine protease (TEV) as a model. I map key features of the local fitness landscape and characterise how they affect details of enzyme evolution.

In order to investigate the evolution of core active site machinery, I mutated the nucleophile of TEV to serine. The differing chemical properties of oxygen and sulphur force the enzyme into a fitness valley with a $>10^4$-fold activity reduction. Nevertheless, directed evolution was able to recover function, resulting in an enzyme able to utilise either nucleophile. High-throughput screening and sequencing revealed how the array of possible beneficial mutations changes as the enzyme evolves. Potential adaptive mutations are abundant at each step along the evolutionary trajectory, enriched around the active site periphery.

It is currently unclear how seemingly neutral mutations affect further adaptive evolution. I used high-throughput directed evolution to accumulate neutral variation in large, evolving enzyme populations and deep sequencing to reconstruct the complex evolutionary dynamics within the lineages. Specifically I was able to observe the emergence of robust enzymes with improved mutation tolerance whose descendants overtake later populations.

Lastly, I investigate how evolvability towards new substrate specificities changed along these neutral lineages, dissecting the different determinants of immediate and long-term evolvability. Results demonstrate the utility of evolutionary understanding to protease engineering.

Together, these experiments forward our understanding of the molecular details of both fundamental evolution and enzyme engineering.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>In Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Angstrom (10^{-10} meters)</td>
</tr>
<tr>
<td>Aa</td>
<td>Amino acid</td>
</tr>
<tr>
<td>AraC</td>
<td>Arabinose inhibitor</td>
</tr>
<tr>
<td>Bla</td>
<td>β-lactamase</td>
</tr>
<tr>
<td>BLIP</td>
<td>β-lactamase inhibitory protein</td>
</tr>
<tr>
<td>Cat</td>
<td>Chloramphenicol acetyltransferase</td>
</tr>
<tr>
<td>CFP</td>
<td>Cyan fluorescent protein</td>
</tr>
<tr>
<td>C-Y(x)</td>
<td>Cyan and yellow fluorescent protein, linked by sequence ‘x’</td>
</tr>
<tr>
<td>DE</td>
<td>Directed evolution</td>
</tr>
<tr>
<td>DFE</td>
<td>Distribution of fitness effects</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxy nucleotide tri-phosphate</td>
</tr>
<tr>
<td>E. coli</td>
<td>Escherichia coli bacteria</td>
</tr>
<tr>
<td>epPCR</td>
<td>Error-prone polymerase chain reaction</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence activated cell sorting</td>
</tr>
<tr>
<td>FRET</td>
<td>Fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>HT-SAS</td>
<td>High throughput screening and sequencing</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>LacI</td>
<td>Lac inhibitor</td>
</tr>
<tr>
<td>Lt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein database (a repository of solved protein structures)</td>
</tr>
<tr>
<td>pH</td>
<td>Hydrogen ion potential (measure of bulk solution acidity)</td>
</tr>
<tr>
<td>pKa</td>
<td>Acid dissociation constant (measure of specific residue acidity)</td>
</tr>
<tr>
<td>Population</td>
<td>A population of enzymes having undergone ‘x’ rounds of neutral evolution and ‘y’ rounds of adaptive evolution</td>
</tr>
<tr>
<td>(x,y)</td>
<td></td>
</tr>
<tr>
<td>TEM1</td>
<td>A class of β-lactamases commonly found in <em>E. coli</em></td>
</tr>
<tr>
<td>TEV</td>
<td>Tobacco Etch Virus nuclear inclusion 1 cysteine protease</td>
</tr>
<tr>
<td>wtTEV</td>
<td>In the context of discussing population evolution</td>
</tr>
<tr>
<td>TEV^{Cys}</td>
<td>In the context of discussing different nucleophiles</td>
</tr>
<tr>
<td>TEV^{Ala}</td>
<td>C151A mutant of TEV</td>
</tr>
<tr>
<td>TEV^{Ser}</td>
<td>C151S mutant of TEV</td>
</tr>
<tr>
<td>T_M</td>
<td>Temperature at which 50% of a protein has melted</td>
</tr>
<tr>
<td>YFP</td>
<td>Yellow fluorescent protein</td>
</tr>
</tbody>
</table>
# Table of Contents

## Chapter A - Introduction

1 Summary

## 2 Enzymes and Evolution

- 2.1 Catalysis of chemical reactions
- 2.2 Nucleophilic catalysis and catalytic triads
- 2.3 Promiscuity
- 2.4 Divergent molecular evolution
- 2.5 Protein sequence space and fitness landscapes
- 2.6 The distribution of fitness effects

## 3 Experimental Evolution of Proteins

- 3.1 Generating variation
- 3.2 Detecting fitness differences
- 3.3 Ensuring heredity
- 3.4 High throughput sequencing

## 4 Robustness

- 4.1 Genomic and protein robustness
- 4.2 Neutral networks

## 5 Evolvability

- 5.1 Definition and quantification
- 5.2 Determinants of evolvability
- 5.3 Evolution of evolvability
- 5.4 Theoretical and applied significance

## 6 TEV Protease – a Model Enzyme

- 6.1 Origin
- 6.2 Structure and function
- 6.3 Specificity

## 7 Overview of results

## Chapter B - Selection and screening methods for protease directed evolution

1 Summary

2 Introduction (Inhibition relief selection)

- 2.1 β-lactamase Inhibitor Protein and inhibition-relief selection

3 Results and Discussion (Inhibition relief selection)
Chapter C - Handicap-recover evolution to accommodate a new nucleophile

1 Summary

2 Introduction

3 Results and Discussion

4 Conclusions

Chapter D - Neutral evolution and emergent robustness

1 Summary

2 Introduction

3 Results and Discussion

4 Conclusions
Table of Figures

Figure A-1 | Energetics of enzyme catalysis ___________________________ | 2
Figure A-2 | Common elements of nucleophilic catalysis using a catalytic triad | 4
Figure A-3 | Crosswise catalytic promiscuity in the alkaline phosphatase superfamily | 7
Figure A-4 | Neo- and sub-functionalization after gene duplication | 8
Figure A-5 | Protein co-evolution sectors | 9
Figure A-6 | Example fitness landscape | 11
Figure A-7 | Mutation epistasis and landscape ruggedness | 12
Figure A-8 | Fitness peak separation by valleys | 13
Figure A-9 | Possible distributions of fitness effects | 15
Figure A-10 | Directed evolution cycle | 17
Figure A-11 | Directed evolution of a cytochrome P450 | 17
Figure A-12 | Mutagenesis methods | 19
Figure A-13 | Genotype-phenotype linkage methods | 21
Figure A-14 | Experimentally measured fitness landscape of DNA binders | 22
Figure A-15 | Sequence-function relationship mapped on protein structure | 23
Figure A-16 | Neutral network | 25
Figure A-17 | Experimentally measured distributions of fitness effects | 27
Figure A-18 | Abundance of beneficial mutations | 27
Figure A-19 | Measured determinants of evolvability | 29
Figure A-20 | Filamentous porovirus | 32
Figure A-21 | TEV protease β-barrels | 33
Figure A-22 | Substrate bound in the active site tunnel | 35
Figure B-1 | BLIP in complex with β-lactamase | 39
Figure B-2 | Scheme of how an active protease is coupled to cell survival | 39
Figure B-4 | Structure of BLIP showing insert sites | 42
Figure B-5 | Comparison of wtBLIP and iBLIP | 43
Figure B-6 | Comparison of STEV and STEVAla | 45
Figure B-7 | Purification and cleavage of iBLIP | 45
Figure B-8 | Proteolytic cleavage of a Dual-GFP fusion FRET-pair | 48
Figure B-9 | pC-Y plasmid map | 49
Figure B-10 | pMAA TEV plasmid map (±MBP) | 51
Figure B-11 | Effect of TEV protease on fluorescence in CFP-YFP FRET system | 52
Figure B-12 | Time-course of lysate FRET ratios | 52
Figure B-13 | Sample lysate screen of TEV mutant library | 53
Figure B-14 | Concentration dependence of FRET ratio | 54
Figure B-15 | Equilibration of C-Y FRET ratio | 54
Figure B-16 | Concentration dependent cleavage of C-Y(tev) by TEV and kinetic fitting | 55
Figure B-17 | Flow cytometric measurement of FRET in E. coli | 56
Figure B-18 | Sample flow cytometric of TEV mutant library | 57
Figure C-1 | Differences in cysteine and serine proteolysis mechanisms | 63
Figure C-2 | PA clan phylogeny | 65
Figure C-3 | Monophasic and biphasic curve fitting of TEV_Cys and TEV_Ser | 67
Figure C-4 | Promiscuous phosphatase activity of TEV_Cys and TEV_Ser | 68
Figure C-5 | Evolutionary recovery of TEV_Ser→TEV_SerX lineage | 71
Figure C-6 | Nucleophile activity tradeoff | 72
Figure C-7 | Stability losses during recovery of TEV_Ser lineage | 74
Figure C-8 | Mutations positions | 75
Figure C-9 | Schematic of sequence space flanking trajectory ............................................ 78
Figure C-10 | Mutation enrichments for TEV<sup>Ser</sup> .......................................................... 79
Figure C-11 | Mutation enrichments for the TEV<sup>Ser</sup>→TEV<sup>Ser</sup>X lineage libraries .......... 79
Figure C-12 | Enrichments of notable mutations ................................................................. 80
Figure C-13 | Locations of beneficial or deleterious mutations ............................................. 81
Figure C-14 | Adaptive potential along the TEV<sup>Ser</sup>→TEV<sup>Ser</sup>X lineage ...................... 81
Figure C-15 | The effect of mutations in the protein core ....................................................... 82
Figure C-16 | Beneficial mutations more common in triad second shell .................................. 83
Figure D-1 | Protein stability effects on fitness ................................................................. 88
Figure D-2 | Robustness of a miRNA fold ............................................................................ 90
Figure D-3 | Viral robustness after evolution at different co-infection rates ......................... 91
Figure D-4 | Mutation tolerance over neutral evolution of P450 oxygenase ...................... 92
Figure D-5 | Enriched mutations after neutral evolution of TEM1 β-lactamase ...................... 92
Figure D-6 | Bimodal distribution of fitness effects and purified mutant characterisation ......... 95
Figure D-7 | Mutation accumulation rate ....................................................................... 97
Figure D-8 | Population mutation tolerance ................................................................. 98
Figure D-9 | Mutation accumulation by structural region (N<sub>e</sub>=10<sup>4</sup> lineages) ............. 100
Figure D-10 | Population average activity (µ=1.5%, N<sub>e</sub>=10<sup>4</sup> lineage) ......................... 101
Figure D-11 | Population average soluble expression (µ=1.5%, N<sub>e</sub>=10<sup>4</sup> lineage) ............ 102
Figure D-12 | N<sub>e</sub>=10<sup>4</sup> Lineage phylogenies ................................................................. 104
Figure D-13 | Reconstructed node mutation locations (N<sub>e</sub>=10<sup>4</sup> lineages) ...................... 105
Figure D-14 | Robustness of reconstructed node sequences ............................................... 106
Figure D-15 | Mutation tolerance (N<sub>e</sub>=10<sup>2</sup> lineages) ...................................................... 108
Figure D-16 | Mutation accumulation by structural region (N<sub>e</sub>=10<sup>2</sup> lineages) ............... 109
Figure D-17 | N<sub>e</sub>=10<sup>2</sup> lineage phylogenies and node mutation locations ........... 112
Figure E-1 | Two overlapping neutral networks ............................................................ 118
Figure E-2 | Protease therapeutic concept .................................................................... 122
Figure E-3 | Identification of flexible target loop regions .............................................. 125
Figure E-4 | Target structures and loop sequences ...................................................... 126
Figure E-5 | Summary and nomenclature of C-Y<sup>(veg)</sup> evolution lineages ...................... 128
Figure E-6 | Promiscuity on randomised substrate libraries .............................................. 129
Figure E-7 | Immediate evolvability along neutral evolution ............................................ 130
Figure E-8 | Long-term evolvability on target sequences of neutrally evolved populations ...... 132
Figure E-9 | Activity on C-Y<sup>(veg)</sup> and C-Y<sup>(pdl)</sup> for wtTEV and evolved variants ............ 134
Figure E-10 | Activity of variants after 3 rounds of evolution on target sequence ............... 134
Figure E-11 | Stepping stone substrate cleavage in lysate by wtTEV .................................... 135
Figure E-12 | Mutation locations and viral substrate tunnel loop ........................................ 137
Figure E-13 | Substrate P6 binding by dual asparagines ..................................................... 138
Figure E-14 | Phylogeny of C-Y<sup>(veg)</sup> active mutants ...................................................... 139
Figure F-1 | Bimodal distribution of fitness effects as a landscape ...................................... 147
Figure F-2 | Explicit transect of a fitness landscape TEV<sup>Ser</sup>→TEV<sup>Ser</sup>X ......................... 148
Figure F-3 | Schematic of fitness landscape flanking TEV<sup>Ser</sup>→TEV<sup>Ser</sup>X trajectory .............. 149

Table C-2 | Variant property summary ................................................................. 70
Table C-3 | Regression of curve fitting of evolved variants ........................................... 71
Table D-1 | Reconstructed node summary (N<sub>e</sub>=10<sup>4</sup> lineages) ..................................... 105
Table D-2 | Reconstructed node summary (N<sub>e</sub>=10<sup>2</sup> lineages) ..................................... 112
Table E-1 | Previously measured immediate evolvabilities .......................................... 119
Table E-2 | Previously achieved TEV protease specificity changes .................................. 121
Table E-3 | Target sequences ................................................................. 126
Table E-5 | CFP-YFP target linker variants .............................................................. 127
Table E-6 | Population average promiscuous activity on C-Y<sup>(veg)</sup> .............................. 130
Table E-7 | Stepping stone substrate CFP-YFP linker variants ....................................... 135
Table G-2 | Substrate plasmids ................................................................. 155
Table G-3 | Primers mentioned in report ................................................................. 156
Table G-4 | Multiplex identifier sequences .............................................................. 157
Table G-5 | Sequences added to DALI alignment by BLASTp ........................................... 164
Chapter A - Introduction

1 SUMMARY

This section introduces the key concepts of enzymes and their evolution that underpin the later chapters. Additionally it describes experimental evolution methods and explains some of what these have revealed about evolution and evolvability. It concludes with a brief overview of the results chapters.
2 ENZYMES AND EVOLUTION

2.1 CATALYSIS OF CHEMICAL REACTIONS

Enzymes are the biological catalysts that perform almost all chemical reactions that occur in an organism. They are produced as a linear polymer of amino acids but fold up into a 3D structure (defined by their sequence) to bring together a few key amino acids into a precise arrangement (the active site) to perform chemistry\(^1\). These residues are responsible for binding a substrate and lowering the activation energy required for its conversion to product\(^2\) (Figure A-1). This lower activation energy determines the rate of the catalysed reaction \((k_{\text{cat}}/K_M)\)\(^3\). Enzymes do this so efficiently that some reactions that naturally take millions of years can be catalysed in seconds\(^4\).

![Figure A-1 | Energetics of enzyme catalysis](image)

A free energy profile \((\Delta G)\) of an abstract reaction of substrate to product \((S\rightarrow P)\) catalysed by enzyme \((E)\). Uncatalysed the reaction passes through a high energy transition state \((S^\oplus)\). When catalysed by an enzyme, however, it binds to form a complex \((E\cdot S)\) which passes through a lower energy transition state \((E\cdot S^\oplus)\). Finally the product complex \((E\cdot P)\) dissociates to release free product. The lower transition state energy means that the catalysed rate \(k_{\text{cat}}/K_M\) is much faster than the uncatalysed rate \(k_{\text{uncat}}\). (Adapted from reference\(^3\))

Activation energy is lowered by a combination of factors, some or all of which are used in different enzymes: Electrostatic stabilisation of
charge build-up on high energy transition states, the provision of general acids and bases, exclusion of solvent, orientation of substrates in reactions with more than one, and sometimes formation of a transient covalent bond. These effects are achieved using the backbone of the protein, the side chain functional groups and organic and inorganic cofactors such as metals and cytochromes.

The active site, is made up of a large, interconnected network of cooperative residues\textsuperscript{5–7}. Some directly contact the substrate and are involved in lowering transition state energy by the mechanisms listed above. Others play a supporting role in orienting, supporting and tuning the electronegativity of these. Although active site residues must be precisely orientated, they are commonly polar or charged\textsuperscript{1} and hence repel each other. Internal active site repulsion is compensated for by the enzyme having a stable structure around them provided by a well packed, hydrophobic core scaffold\textsuperscript{8,9} which accounts for much of the rest of the protein.

Despite this stable fold, it’s important to note that enzymes are not static entities. They’re not even particularly rigid entities\textsuperscript{8,10}. Enzymes are highly flexible and hence exist as an ensemble of conformations that are constantly interconverting. These motions range from very fast and small (such as side chain rotation\textsuperscript{11,12}) to large and slow (such as loop or whole domain movements\textsuperscript{13–15}). The role of these dynamics in activity is controversial\textsuperscript{16–19} but what is clear is that many enzymes have several distinct conformations which optimise the various requirements of substrate binding, catalysis and product release. Some enzymes will change between these states during their catalytic cycle\textsuperscript{2,13–15,19}.
One method of lowering a reaction’s activation energy is to split it into two easier reactions by nucleophilic, covalent catalysis. In this specific case, a single residue has great importance to the mechanism since it is tuned by its neighbours to lower its pKa and increase the availability of the deprotonated form. This nucleophile firstly attacks the substrate to form a covalent bond which is subsequently hydrolysed by activated water. Each of these two steps requires lower energy than the uncatalysed reaction, increasing its rate.

A common method for generating a nucleophilic residue for catalysis is by using an Acid-Base-Nucleophile catalytic triad. The triad forms a charge-relay network to polarise and activate the nucleophile, which attacks the substrate, forming a covalent intermediate which is then hydrolysed to regenerate free enzyme (Figure A-2).

**Figure A-2 | Common elements of nucleophilic catalysis using a catalytic triad**

(a) The enzyme (black) performs a nucleophilic attack on substrate (red) via a tetrahedral intermediate to eject the first product, leaving an acyl-enzyme intermediate. (b) The intermediate is hydrolysed by an activated water molecule via a second tetrahedral intermediate to release the second product and regenerate free enzyme.
This motif is so effective that the same geometry has arisen at least 25 times independently\textsuperscript{21,23} as a means of catalysing group transfer\textsuperscript{24,25} and hydrolysis\textsuperscript{22} reactions (including proteolysis). The two most commonly used nucleophiles are the oxygen (alcohol) of serine and the sulphur (thiol) of cysteine\textsuperscript{21,22}. I introduce the evolution and significance of different nucleophiles in greater detail in Chapter C.

In addition to covalent catalysis, the triad and surrounding residues coordinate additional mechanistic effects. General acid catalysis protonates leaving groups. General base catalysis activates the water that hydrolyses the intermediate. Backbone electrostatics in the oxyanion hole stabilise charge build-up on transition states and intermediates along the reaction pathway. Additional residues assist in binding substrate and, in some enzymes, coordinating large structural movements involved in the catalytic cycle such as lid opening and closing\textsuperscript{13,14}.

2.3 PROMISCUITY

Having described how enzymes achieve catalysis of their target substrate, it is useful to introduce catalysis of alternative substrates. Classically, enzymes have been viewed as catalysing only one reaction. Over recent decades, however, it has become increasingly clear that besides efficiently catalysing their target reaction, enzymes often also have a set of side-activities. The ability to catalyse more than one reaction with the same active site is known as promiscuity\textsuperscript{26}. Promiscuity is subcategorised into 3 classes describing in what way these reactions differ. Catalytic promiscuity refers to an enzyme that can cleave two different types of chemical bond, for example the hydrolysis of phosphonate and sulphate bonds by \textit{Pseudomonas aeruginosa} arylsulfatase\textsuperscript{27}. Substrate promiscuity is when an enzyme can bind several related substrates to perform the same catalysis on
them, for example some P450 enzymes can oxidise hundreds of substrates\textsuperscript{28}. Lastly, product promiscuity occurs when an enzyme converts a substrate into a high energy intermediate (e.g. the conversion of E,E farnesyl diphosphate to a reactive cation by γ-humulene synthase), which can then be resolved to a variety of different products\textsuperscript{29}.

The origins of promiscuity are still unclear. In many enzymes there is no evidence that promiscuity has been specifically selected for. In these cases it seems likely that promiscuous activities are simply the accidental by-products of extreme catalytic activity for the native reaction and the difficulty of discriminating between similar substrates. Lack of strong counter-selection against promiscuous activities allows their evolutionary persistence\textsuperscript{30}. For example clotting proteases such as thrombin merely need to be specific enough not to cause accidental cleavage.

Conversely some promiscuous enzymes are likely the result of direct selection for multiple activities such as enzymes that scavenge nutrients or remove environmental toxins (e.g. P450 oxidases\textsuperscript{28}, serum paraoxonases\textsuperscript{31} and phosphotriesters\textsuperscript{32}). Occasionally promiscuity can be exquisitely refined such as the ribosome. To function it must catalyse the polymerisation of any of 20 natural amino acids (substrate promiscuity), but specifically organised, based on the sequence of an mRNA template rather than mere random polymerisation. In this way controlled promiscuity is necessary for its function.

\section*{2.4 DIVERGENT MOLECULAR EVOLUTION}

Divergent evolution of proteins has created a wide array of enzymes descended from a common ancestor, but now encompassing a wide range of activities\textsuperscript{33,34}. The widest evolutionary category that we can currently place around a set of enzymes is a superfamily which
describes a group of enzymes with common ancestry inferred by fold and mechanistic similarity. The divergence of gene sequences occurs through two main processes: Adaptive evolution (natural selection for improved or altered function\textsuperscript{35,36}) and neutral evolution (the accumulation of mutations that have little or no effect on function leading to genetic drift\textsuperscript{37,38}).

Evidence for the importance of promiscuous activities as a starting point for adaptive evolution comes, in part, from comparative studies of existing proteins\textsuperscript{39}. The promiscuous activity of one member of a protein superfamily is often the native activity of another member of that family such as in the aminotransferase\textsuperscript{40} or alkaline phosphatase superfamilies\textsuperscript{41–43} (Figure A.3). Cross-wise promiscuity implies that divergent evolution has (partially) specialised each member towards one of a set of similar activities. For example, the pesticide-degrading phosphotriesterase still shows promiscuous activity towards the substrate of the enzyme from which it is descended\textsuperscript{44}. Similarly, type III antifreeze is descended from sialic acid synthase, which itself shows promiscuous ice-binding activity\textsuperscript{45}.

The promiscuity of enzymes is a key feature in modern models of molecular evolution. All models of gene diversification by duplication...
rely on some level of promiscuous activity either before or after gene duplication. Classical neofunctionalisation models postulate that after a gene duplicates, one copy is released from selective pressure for the parental function. This paralog accumulates mutations that either deactivate it (pseudogenisation, Figure A-4a) or create new activities (Figure A-4b). Conversely, subfunctionalisation models assume some level of promiscuous activity before duplication such that when a gene duplicates each copy can specialise towards one of the parent activities (Figure A-4c).

Both neo- and sub- functionalization theories fail to explain some aspects of molecular evolution. In neofunctionalisation, the relative abundance of deleterious mutations over beneficial mutations makes the likelihood that beneficial mutations will accumulate in a duplicated gene before it succumbs to pseudogenisation very low. Conversely, for subfunctionalisation to explain the development of the large range of contemporary functions, requires the assumption that a large number of genes are promiscuous. Indeed it has sometimes been misinterpreted to imply that ancestral genes were all promiscuous and have linearly
become more specialised over time, however several papers have attempted to clear up this confusion\textsuperscript{30,51}.

Modern models tend to lie somewhere in between these extremes whereby some degree of promiscuity likely exists before duplication. Which model best explains divergent evolution is likely to be protein specific and depend on the relative strength of selection, the prevalence of beneficial promiscuity, and the trade-offs between activities\textsuperscript{50}. Duplication may be neutral or be selected for gene copy amplification\textsuperscript{52} or relief of constraints if optimisation of the new activity trades-off with the native\textsuperscript{45,53,54}.

Through these divergent evolution processes, all new genes are generated (with the rare exception of de-novo gene birth\textsuperscript{55}). The evolution of enzyme active sites is constrained by chemical requirements on function. Other residues are less constrained and so their evolutionary divergence is more dominated by genetic drift (i.e. accumulation of neutral mutations). Therefore, active site residues tend to evolve more slowly except during periods where there is selection for change in function\textsuperscript{56,57}. Additionally, whilst unconstrained residues evolve independently, some residues are highly dependent on the identity of their neighbours. Interactions between residues cause them to co-evolve as ‘protein sectors’\textsuperscript{58,59} (Figure A-5).

![Figure A-5 | Protein co-evolution sectors](image-url)

Structure of rat trypsin (PDB 3tgi) with co-evolutionary sectors shown as surfaces and substrate as sticks. Blue sector is the β-barrel scaffold. Green sector is the triad and barrel interface. Red sector is the main substrate binding pocket. (Adapted from reference\textsuperscript{58})
In order to understand evolution of the enormous diversity of protein sequence and function, it is helpful to talk of sequence space – an imagined array of all protein sequences. Formally, each residue in a protein is a dimension with 20 possible positions along that axis corresponding to the possible amino acids. Hence there are 400 possible di-peptides arranged in a 20x20 space but that expands to $10^{130}$ for even a small protein of 100 amino acids arranged in a space with 100 dimensions. Despite the diversity of protein superfamilies, however, sequence space is extremely sparsely populated by functional proteins\(^60\). Most random protein sequences have no fold or function. Enzyme superfamilies, therefore, exist as tiny clusters of active proteins in a vast empty space of non-functional sequence.

A fitness landscape consists of adding an extra axis of fitness. For organisms, fitness formally describes the number of viable offspring that a particular genome sequence has. When discussing enzymes, however, it is common to define fitness as a particular function of interest\(^61\). Explicit fitness landscapes are overwhelmingly multidimensional but features of the topology are often represented in a 3D simplification (e.g. by principal component analysis). This generates a landscape with two horizontal axes of sequence space and a vertical axis of fitness\(^62\) (Figure A-6).
Since each point is a different enzyme sequence, mutations link adjacent points on the fitness landscape. Adaptive evolution of a single gene can therefore be thought of as a gene moving uphill from point to point in the fitness landscape. Similarly a population of gene variants inhabit a cloud of sequences which evolve such that higher points are more likely to proliferate and lower points are purged, leading the population to move uphill each generation. In this way, the structure of the fitness landscape guides evolutionary trajectories and evolutionary trajectories report on fitness landscape structure.

In these landscapes there are a number of key features. The maximum peak height (the global maximum) represents the most active possible enzyme sequence. Some enzymes likely exist near these as they operate at the diffusion limit (i.e. diffusion of substrate in and product out limits rate) and so fundamentally cannot be any faster. If all mutations are additive, they can be acquired in any order (Figure A-7a). The landscape, then, is perfectly smooth, with only one peak and all sequences can evolve uphill to it (Figure A-7d).

Conversely, if mutations interact with one another (epistasis Figure A-7b,c), the fitness landscape becomes rugged as the effect of a mutation depends on the genetic background of other mutations.
Such interactions occur whenever one mutation alters the local environment of another residue (either by directly contacting it, or by inducing changes in the protein structure). For example in a disulphide bridge, a single cysteine has no effect on protein stability until a second is present at the correct location. At its most extreme interactions are so complex that the fitness is ‘uncorrelated’ with gene sequence and the topology of the landscape is random (Figure A-7f).

Most landscapes are likely in between these two extremes of ruggedness (Figure A-7e). Such landscapes contain many local maxima which, though not as high as the global maximum, are nevertheless the highest accessible nearby peak. These are sometimes referred to as multiple evolutionary solutions to a problem\(^68,69\). The number and separation of these peaks defines how accessible they are from each other (Figure A-8).
Additionally, clustering of peaks in regions of the fitness landscape (i.e. landscape structure) indicates that the solutions contained are similar. Such structure comes about when chemical or physical constraints cause some regions of a fitness landscape to be particularly sparse in functional sequences, whilst others regions are rich.

Several important caveats exist. It is important to note that, since the human mind struggles to think in greater than three dimensions, 3D topologies can mislead. In particular it is not clear whether peaks are ever truly separated by fitness valleys in such multidimensional landscapes, or whether they are connected by vastly long neutral ridges. Additionally, the fitness landscape is not static in time but dependent on the changing environment and evolution of other genes. It is hence more of a seascape, further affecting how separated
adaptive peaks can actually be. Finally, since it is common to use function as a proxy for fitness when discussing enzymes, any promiscuous activities exist as overlapping landscapes that together will determine the ultimate fitness of the organism.

With these limitations in mind fitness landscapes can still be an instructive way of thinking about evolution. It is fundamentally possible to measure (even if not visualise) some of the multidimensional parameters of landscape ruggedness and of peak number, height, separation, and clustering. Simplified 3D landscapes can then be used relative to each other to visually represent the relevant features (see section 3.4).

2.6 THE DISTRIBUTION OF FITNESS EFFECTS

The result of mutations to a particular sequence can be described by the Distribution of Fitness Effects (DFE), the spectrum of fitnesses of mutants. Mutations can be deleterious, neutral or beneficial (each to different degrees) and so the DFE describes the local fitness landscape within a 'one mutation radius' of the mutagenised sequence. The relative frequencies of deleterious, neutral and beneficial mutation have large consequences for evolution. Selectionist theories proposed that almost all mutations were either deleterious or beneficial and hence variation in nature was exclusively adaptive. The neutral theory postulated that, in fact, many mutations were neutral and hence variation was the product of non-directional drift. This was later refined in the nearly-neutral theory that posited that mutations that are only mildly deleterious are also frequently fixed in populations.
Studies of fitness landscapes have been largely theoretical, relying either on mathematical and computation models or on bioinformatic measurements of natural diversity. Nevertheless, there are increasing attempts to study evolution in an experimental setting. Subsequent sections of this chapter will introduce some of the experimental work that is beginning to measure aspects of fitness landscapes and their effects on enzyme evolution.
The study of evolution is largely based on extant organisms and their genes. However, research is often limited by the lack of fossils (and particularly the lack of ancient DNA sequences) and incomplete knowledge of ancient environmental conditions. Experimental evolution has been used in various formats to understand underlying evolutionary processes in a controlled system. Experimental evolution has been performed on multicellular and unicellular eukaryotes, prokaryotes, viruses as well as individual enzyme, ribozyme, and replicator genes. Since experiments address topics as varied as the evolution of multicellularity, maximum flight speed or heat adaptation, I will concentrate here on the experimental evolution of enzyme genes.

Evolution requires three things to occur: variation between replicators, that variation causes fitness differences upon which selection acts, and that this variation is heritable. Directed evolution (DE) is a mimic of the natural evolution cycle in which a single gene is evolved by iterative rounds of mutagenesis, selection or screening, and amplification. Directed evolution (DE) is a mimic of the natural evolution cycle in which a single gene is evolved by iterative rounds of mutagenesis, selection or screening, and amplification (Figure A-10). It is frequently used for protein engineering as an alternative to rational design, but can also be used to investigate fundamental questions of enzyme evolution.

As a protein engineering tool, DE has been most successful in three areas. Firstly, improving enzyme stability for biotechnological use at high temperatures or in harsh solvents. Secondly, for improving binding affinity of antibody therapeutics and the activity of de novo designed enzymes. Thirdly in altering substrate specificity of existing enzymes, again, often for use in industry.
Of course, selecting for improvement in the assayed function simply generates improvements in the assayed function. To understand how these are achieved, the properties of the evolving enzyme have to be measured. Improvement of the assayed activity can be due to improvements in enzyme catalytic activity (Figure A-11) or enzyme concentration.

Figure A-10 | Directed evolution cycle
An example of a directed evolution experiment with comparison to natural evolution. Inner circle indicates the 3 stages of the cycle with the natural process being mimicked in brackets. The outer circle demonstrates a typical experiment.

Figure A-11 | Directed evolution of a cytochrome P450
Directed evolution of P450-BM3 (hydroxylase of fatty acids) for hydroxylation of propane to propanol (round 9 selection for increased stability). First panel shows how the assayed activity of turnover number increased. Second and third panels show the improvements of kinetic parameters $k_{cat}$ (catalysis) and $K_M$ (binding). Final panel shows the concomitant loss of $T_{SD}$ (thermal denaturation temperature). (Adapted from reference62)
Additionally, there is no guarantee that improvement on one substrate will improve activity on another. This is particularly important when the desired activity cannot be screened or selected for and so a ‘proxy’ substrate is used which can lead to evolutionary specialisation to the proxy without improving the desired activity. Consequently, choosing appropriate screening or selection conditions is vital for successful directed evolution. In sections 3.1-3.3 I will outline the variety of techniques that can be used for each of the three steps in the directed evolution cycle.

3.1 Generating Variation

The first step in performing a cycle of DE is the generation of a library of variant genes. Random sequence is vast (10^{130} possible sequences for a 100 amino acid protein) and extremely sparsely populated by functional proteins. Neither experimental\textsuperscript{106}, nor natural\textsuperscript{107} evolution can ever get close to sampling so many sequences. Of course, natural evolution samples variant sequences close to functional protein sequences and this is imitated in DE by mutagenising an already functional gene (Figure A-12).

The starting gene can be mutagenised by random point mutations (by chemical mutagens\textsuperscript{108} or error prone PCR\textsuperscript{109}) and insertions and deletions (by transposons\textsuperscript{110}, and S. Emond unpublished data). Gene recombination can be mimicked by DNA shuffling\textsuperscript{111} of several sequences (usually of more than 70\% homology) to jump into regions of sequence space between the shuffled parent genes. Finally, specific regions of a gene can be systematically randomised\textsuperscript{112} for a more focused approach based on structure and function knowledge. Depending on the method, the library generated will vary in the proportion of functional variants it contains. Even if an organism is used to express the gene of interest, by mutagenising only that gene,
the rest of the organism’s genome remains the same and can be ignored for the evolution experiment (to the extent of providing a constant genetic environment).

![Mutagenesis methods](image)

3.2 DETECTING FITNESS DIFFERENCES

Two main categories of method exist for isolating functional variants. Selection systems directly couple protein function to survival of the gene, whereas screening systems individually assay each variant and allow a quantitative threshold to be set for sorting a variant or population of variants of a desired activity.

Selection for binding activity is conceptually simple. The target molecule is immobilised on a solid support, a library of variant proteins is flowed over it, poor binders are washed away, and the remaining bound variants recovered to isolate their genes. Binding of an enzyme to immobilised covalent inhibitor has been also used as an attempt to isolate active catalysts. This approach, however, only selects for single catalytic turnover and is not a good model of substrate binding or true substrate reactivity. If an enzyme activity can be made necessary for cell survival, either by synthesizing a vital metabolite, or
destroying a toxin, then cell survival is a function of enzyme activity\textsuperscript{114,115}. Such systems are generally only limited in throughput by the transformation efficiency of cells. They are also less expensive and labour intensive than screening, however they are typically difficult to engineer, prone to artefacts and give no information on the range of activities present in the library.

An alternative to selection is a screening system. Each variant gene is individually expressed and assayed to quantitatively measure the activity (most often by a colourgenic or fluorogenic product). The variants are then ranked and the experimenter decides which variants to use as temples for the next round of DE. Even the most high throughput assays usually have lower coverage than selection methods but give the advantage of producing detailed information on each one of the screened variants. This disaggregated data can also be used to characterise the distribution of activities in libraries which is not possible in simple selection systems. Screening systems, therefore, have advantages when it comes to experimentally characterising adaptive evolution and fitness landscapes.

3.3 Ensuring Heredity

When functional proteins have been isolated, it is necessary that their genes are too, therefore a genotype-phenotype link is required\textsuperscript{116} (Figure A-13). This can be covalent, such as mRNA display where the mRNA gene is linked to the protein at the end of translation by puromycin\textsuperscript{106}. Alternatively the protein and its gene can be co-localised by compartmentalisation in living cells\textsuperscript{117} or emulsion droplets\textsuperscript{118}. The gene sequences isolated are then amplified by PCR or by transformed host bacteria. Either the single best sequence, or a pool of sequences can be used as the template for the next round of mutagenesis. The repeated cycles of Diversification-Selection-
Amplification generate protein variants adapted to the applied selection pressures.

![Diagram of genotype-phenotype linkage methods](image)

**Figure A-13 | Genotype-phenotype linkage methods**
As the protein is expressed, it can either be covalently linked to its gene (as in mRNA, left) or compartmentalized with it (cells or artificial compartments, right). Either way ensures that the gene can be isolated based on the activity of the encoded protein.

3.4 **HIGH THROUGHPUT SEQUENCING**

The maximum library size that can be screened for active variants (also called library coverage) is limited by the throughput of the assay. Additionally, the number of active variants chosen to be taken through to the next round of DE is limited by data analysis of their sequences. If a single variant is selected from one round to the next, each member of the lineage can be characterised in molecular detail but stochastic events by genetic drift have a larger effect. Conversely, at each round a pool of improved variants can be selected however it has, until recently, not been possible to know the sequences present in this gene pool. This trade-off between depth and breadth in DE studies has been partly alleviated by the rise of ‘next generation’ sequencing technologies. It has become feasible to sequence $10^5$ or more gene variants and so it is possible to keep track of the composition of large, mixed pools of variants.

Additionally, by sequencing a library of variants before and after an experimental selection, it is possible to measure the enrichment of functional sequences and the purging of non-functional sequences. Indeed, the degree of enrichment for each variant (its enrichment...
factor) indicates the relative fitnesses of variants and therefore which residue mutations are adaptive to that selection.

This technique has been used in two relevant ways, with a focus either on DNA sequence or on protein structure. If screening is performed in combination with next generation sequencing, each sequence can have a fitness ascribed to it and be laid out as an explicit fitness landscape\textsuperscript{119}. For example, when 30nt DNA molecules were evolved to bind a protein target\textsuperscript{120}, the local fitness landscape could be mapped to show two distinct broad optima, one higher than the other (Figure A-14). Sequences were observed to evolve uphill towards these peaks over 8 rounds. These works are revealing the complexity of real fitness landscapes and in particular their tendency to have many peaks of similar fitness\textsuperscript{79,120–122}.

When adaptive and deleterious mutations are mapped on the protein’s structure they can indicate which regions are enriched in functional variation and which are constrained to the wild type amino acid\textsuperscript{122–126}.
(Figure A-15). This can show structural and functional constraints that prevent adaptive mutation. Additionally, it can uncover regions where beneficial mutations are more common, acting as evolutionary hotspots.

Together, these technologies and methods are beginning to allow the understanding of complex evolutionary processes that have traditionally not been possible to investigate in detail. Examples of such processes are the evolution of robustness and evolvability, the structure of real fitness landscapes and the experimental evolution of large populations. The rest of this chapter will introduce these topics and the rest of this thesis will consist of their investigation.
4 ROBUSTNESS

4.1 GENOMIC AND PROTEIN ROBUSTNESS

Genomes mutate by environmental damage and imperfect replication, yet they display remarkable tolerance. For example >95% of point mutations in *C. elegans* have no detectable effect\textsuperscript{127} and even 90% of single gene knockouts in *E. coli* are non-lethal\textsuperscript{128}. This comes from robustness both at the genome level and protein level. Since Muller’s ratchet (the fixation of slightly deleterious mutations) tends to reduce robustness, selection pressures must exist that act to increase it (see Chapter D, 2.1).

There are many mechanisms that provide genome robustness. For example, genetic redundancy\textsuperscript{129} reduces the effect of mutations in any one copy of a multi-copy gene. Additionally the flux through a metabolic pathway is typically limited by only a few of the steps, meaning that changes in function of many of the enzymes have little effect on fitness\textsuperscript{130,131}. Similarly metabolic networks haven multiple alternative pathways to produce many key metabolites\textsuperscript{132}. Protein mutation tolerance is the product of two main features: the structure of the genetic code and protein structural robustness\textsuperscript{133,134}.

Proteins are resistant to mutations because many sequences can fold into highly similar structural folds\textsuperscript{135}. A protein adopts a limited ensemble of native conformations because those conformers have lower energy than unfolded and misfolded states ($\Delta\Delta G$ of folding)\textsuperscript{136,137}. This is achieved by a distributed, internal network of cooperative interactions (hydrophobic, polar and covalent)\textsuperscript{138}. Protein structural robustness results from few single mutations being sufficiently disruptive to compromise function. Proteins have also evolved to avoid aggregation\textsuperscript{139} as partially folded proteins can
combine to form large, repeating, insoluble fibrils and masses\textsuperscript{140}. There is evidence that proteins show negative design features to reduce the exposure of aggregation-prone $\beta$-sheet motifs in their structures\textsuperscript{141}.

Additionally, there is some evidence that the genetic code itself may be optimised such that most point mutations lead to similar amino acids (conservative)\textsuperscript{142,143}. These factors create a DFE of mutations that contains a high proportion of neutral and nearly-neutral mutations. Robustness is sometimes measured as the proportion of mutations that are neutral (roughly two thirds in proteins so far measured\textsuperscript{50,144,145}).

4.2 NEUTRAL NETWORKS

This resistance to deleterious mutation allows access to a large neutral network of phenotypically equivalent genes\textsuperscript{146–148}. Since many sequences have equivalent function, they can be imagined as existing on a broad, flat plateau on a fitness landscape. The more robust a protein, the more neutral neighbours it has\textsuperscript{149}. Since some sequences have more neutral neighbours than others, the population is predicted to evolve towards these robust sequences\textsuperscript{149} (Figure A-16). This is sometimes called circum-neutrality and represents the movement of populations away from cliffs in the fitness landscape\textsuperscript{150}. Experiments are beginning to confirm these predicted processes (see Chapter D).
5 EVOLVABILITY

5.1 DEFINITION AND QUANTIFICATION

Just as evolution is the change of gene sequences over time, evolvability is the propensity for change over time. Early definitions were mostly concerned with heritability of variation and so whether that variation could lead to evolution. Later definitions are broader – the latent potential for further evolution. Evolvability is sometimes quantified as the percentage of mutations that lead to a particular trait. At its minimum, zero evolvability indicates that the evolution of a trait is impossible (e.g. enzyme rates faster than the diffusion limit or insects thicker than 16 cm\(^{151,152}\)). Evolvability of 10\(^{-3}\) would mean that in a 200 amino acid enzyme, only a single point mutation is adaptive. Higher evolvability indicates multiple possible solutions. In this sense, evolvability is dependent on the DFE of mutations. Neutral evolvability will depend on the proportion of mutations that are not deleterious, i.e. robustness. Adaptive evolvability will depend on the much smaller number of mutations that are beneficial to function and fitness.

Experimentally measured DFEs are, in general, roughly bimodal (Figure A-17) whether for individual proteins\(^ {79}\) or whole organisms\(^ {153,154}\). Some studies have also identified a discrete nearly-neutral set and measured the amount by which these mutations reduce fitness\(^ {155}\). Additionally, the distribution of the rare set of beneficial mutations seems to be exponential with mutations of small effect being more common than mutations of large effect\(^ {156–158}\).
Figure A-17 | Experimentally measured distributions of fitness effects
In a DFE, the selection coefficient describes the relative enrichment by selection where 0 indicates neutrality, -1 indicates lethality, and values >0 indicate benefit. (a) The proportions of different mutation effects expected by models and the observed DFE from mutagenesis of 9 residues of Heat Shock Protein 90 in yeast to all possible amino acids. The red line shows the DFE of non-synonymous mutations, the grey line shows the DFE of synonymous mutations. Observed DFE in (b) 66 Tobacco Etch Virus mutants and (c) 45 \( \phi X174 \) virus mutants. (Adapted from references 79, 154)

Estimates of the abundance of beneficial mutations span 5 orders of magnitude in different systems\(^{50,97,123,126,157–162}\) (Figure A-18). This range is unsurprising since one can imagine different selection pressures for which it would be more or less difficult to find adaptive mutations.

Figure A-18 | Abundance of beneficial mutations
The proportion of spontaneous mutations that are adaptive to different selection pressures in different organisms. E3 ligase and \( \beta \)-lactamase evolution was performed by mutagenesis of that gene only, other evolutions by whole genome mutagenesis. The ‘lab conditions’ of the Drosophila evolution are detailed in reference\(^{161}\) and references therein. (Data from references\(^{50,97,123,157–162}\)
5.2 Determinants of Evolvability

Evolvability depends on two aspects of a fitness landscape. Firstly, the population’s movement speed through a fitness landscape (mutation and recombination rate). Secondly, the structure of that landscape (pleiotropy, chemical/physical constraints and epistasis) can limit the number of accessible fitness peaks.

Since functional sequences are not evenly distributed in sequence space, some biochemical properties are more readily accessible from some points in sequence space. Some key determinants the structure of the fitness landscape are detailed below.

Pleiotropy occurs when a single mutation affects multiple traits. It was originally used to describe genes or mutations that affect multiple traits in organisms such as butterfly wing spots. In biochemistry it is used for underlying protein properties. For example, it is commonly observed that mutations increasing activity decrease stability (negative pleiotropy). Conversely, proteins often struggle to discriminate between similar substrates and mutations improving an adaptive activity may increase an anti-adaptive one (positive pleiotropy). For example when β-isopropylmalate dehydrogenase is evolved to use NADP⁺ instead of the native NAD⁺ as a cofactor, it is unable to distinguish between NADP⁺ and the more abundant NADPH (Figure A-19a). The NADPH acts as an inhibitor and inability to uncouple their affinities leads to constraint on efficiency with the NADP⁺ cofactor. Similarly, RuBisCO cannot perfectly distinguish between carbon dioxide and oxygen, causing energy wastage.

Pleiotropic effects, therefore, reduce the power of selection to independently optimise enzyme properties.
Likewise, physical and chemical constraints prevent the evolution of some features. For example some proteases use threonine as their nucleophile. This has to be at the N-terminus of the protein since the methyl group sterically clashes with other residues in the enzyme in order to accommodate the substrate\(^2\) (Figure A-19b). Consequently, serine proteases can never have their nucleophile replaced with
threonine. In the same way, the speed of diffusion limits the maximum attainable rate of an enzyme or transport protein\textsuperscript{4,64,65} (Figure A-19c).

Finally, epistasis is usually considered a constraining factor on evolution as the lack of a smooth landscape makes it harder for evolution to access fitness peaks\textsuperscript{172}. In highly rugged landscapes, fitness valleys block access to some genes, and even if ridges exist that allow access, these may be rare or prohibitively long. For example, 5 mutations together enable TEM1 \(\beta\)-lactamase to cleave cefotaxime (a 3\textsuperscript{rd} generation antibiotic). However, of the 120 possible pathways to this 5-mutant variant, only 7\% were accessible to evolution as the remainder passed through fitness valleys\textsuperscript{170} (Figure A-19d). Conversely, a 5-mutant variant of epoxide hydrolase with altered specificity can be accessed by 46\% of pathways\textsuperscript{179}.

5.3 Evolution of Evolvability

Of particular contention is whether evolvability is itself evolvable\textsuperscript{163,180,181} or whether it is an intrinsic, unselected property of biological systems\textsuperscript{172}. Initially the idea of evolvability as being, itself, evolvable made many biologists uncomfortable as it seems teleological (the product of foresight)\textsuperscript{182–184}.

Recent works are beginning to suggest that there may be mechanisms whereby evolvability can evolve. In some instances there is evidence that evolvability evolved as the accidental, pleiotropic by-product of direct selection of another trait\textsuperscript{185}. Alternatively, an evolvability-enhancing mutation might be indirectly selected by hitchhiking on the fitter variants it facilitates\textsuperscript{186} (second-order selection).

These works have largely focussed on evolution of evolvability by the altered rate of mutation and recombination. There is now increasing interest in experimentally exploring whether populations can also
evolve towards regions of a fitness landscape\textsuperscript{187,188}, which promote evolvability by being less constrained (see \textbf{Chapter D}) and hence possibly richer in beneficial mutations (see \textbf{Chapter E}).

5.4 \textsc{Theoretical and Applied Significance}

The study of evolvability has fundamental importance for understanding very long term evolution of protein superfamilies\textsuperscript{71,72} and organism phyla and kingdoms\textsuperscript{180,189,190}. A thorough understanding of the details of long term evolution will likely form part of the Extended Evolutionary Synthesis\textsuperscript{180,191–193} (the update to the Modern Synthesis). In addition, these phenomena have two main practical applications. For protein engineering we wish to increase evolvability, and in medicine and agriculture we wish to decrease it.

Firstly, for protein engineering it is important to understand the factors that determine how much a protein function can be altered. In particular, both design and DE approaches aim to create changes rapidly through mutations with large effects\textsuperscript{194,195}. Such mutations, however, commonly destroy enzyme function or at least reduce tolerance to further mutations\textsuperscript{134,176}. Identifying evolvable proteins\textsuperscript{196} and manipulating their evolvability will be increasingly necessary in order to achieve ever larger functional modification of enzymes.

Secondly, many human pathologies are not static phenomena but capable of evolution. In medicine, bacteria, fungi, viruses and cancers evolve to escape host immune responses and to develop drug resistance\textsuperscript{197–199}. In agriculture, plants are becoming resistant to common herbicides\textsuperscript{200} and insects to insecticides\textsuperscript{201}. It is possible that we are facing the end of the effective life of most available antibiotics\textsuperscript{202}. Predicting evolvability of our pathogens\textsuperscript{203} and devising strategies to slow or circumvent\textsuperscript{204} it will require deeper knowledge of the complex forces driving evolution at the molecular level.
6. **TEV Protease – a Model Enzyme**

For the work presented in this thesis, I use the nuclear inclusion cysteine protease of Tobacco Etch Virus (TEV protease) as a model enzyme which I shall introduce here.

6.1 **Origin**

In order to minimise genome size, many viruses express their entire genome as one massive polyprotein which is subsequently cleaved into functional units by proteases\textsuperscript{205}. Tobacco etch virus is a filamentous plant potyvirus (Figure A-20) for which proteolysis is largely performed by the nuclear inclusion cysteine protease\textsuperscript{206} to generate mature viral proteins. TEV protease can be expressed as an active but non-toxic, 27kDa protein in \textit{E. coli}\textsuperscript{207} and is commonly used as a biochemical tool for the sequence-specific cleavage of fusion proteins\textsuperscript{207,208}. It is consequently well characterised and shows several properties that make it a useful model system for experimental evolution.

![Figure A-20 | Filamentous potyvirus](image)

The structure of the closely related tobacco mosaic virus based on (a) the X-ray crystal structure of the capsid protein (2tmv) and (b) scanning electron microscopy. (Adapted from references\textsuperscript{209,210})

6.2 **Structure and Function**

The structure of TEV protease has been solved by X-ray crystallography\textsuperscript{211}. It is comprised of two β-barrels and a flexible C-
terminal tail (Figure A-21) and displays homology to the chymotrypsin superfamily of proteases (see Chapter C). Covalent catalysis is performed with an Asp-His-Cys triad, split between the two barrels (Asp on β1 and His and Cys on β2). The substrate is held as a β-sheet, forming an antiparallel interaction with the cleft between the barrels and a parallel interaction with the C-terminal tail. The enzyme therefore forms a binding tunnel around the substrate and side chain interactions control specificity (detailed description in section 6.3). Catalysis proceeds as described in Figure A-2.

As a biochemical tool, TEV protease has several limitations. It is prone to deactivation by self-cleavage (autolysis), though this can be abolished through a single S219V mutation in the internal cleavage site. The protease expressed alone is also poorly soluble however several attempts have been made to improve its solubility through DE and computational design. It has also been shown that expression can be improved by fusion to Maltose Binding Protein (MBP) which acts a solubility enhancing partner. In the work
presented in later chapters I use the S219V variant of TEV fused to MBP for evolution.

6.3 SPECIFICITY

The reason for the use of TEV protease as a biochemical tool is its high sequence specificity. This specificity allows the controlled cleavage of proteins when the preferred sequence is inserted into flexible loops. It also makes it relatively non-toxic in vivo\(^{208}\) as the recognized sequence scarcely occurs in proteins.

The preferred cleavage sequence was first identified by examining the cut sites in the native polyprotein substrate for recurring sequence\(^{220}\). The consensus for these native cut sites is ENLYFQ\(\backslash S\) where ‘\(\backslash\)’ denotes the cleaved peptide bond. Residues of the substrate are labelled P6 to P1 before the cut site and P1’ after the cut site. Early works also measured cleavage of an array of similar substrates to characterise how specific the protease was for the native sequence\(^{205,221}\). Studies have subsequently used sequencing of cleaved substrates from a pool of randomised sequences to determine preference patterns\(^{222,223}\).

Though ENLYFQ\(\backslash S\) is the optimal sequence, the protease is active to a greater or lesser extent on a range of substrates (i.e. shows some substrate promiscuity). The highest cleavage is of sequences closest to the consensus EXLY\(\Phi\)Q\(\varphi\) where X is any residue, \(\Phi\) is any large or medium hydrophobe and \(\varphi\) is any small hydrophobe\(^{205,221–223}\).

Specificity is endowed by the large contact area between enzyme and substrate. Proteases such as trypsin have specificity for one residue before and after the cleaved bond due to a shallow binding cleft with only one or two pockets that bind the substrate side chains. Conversely, viral proteases such as TEV protease have a long C-terminal tail which completely covers the substrate to create a binding
tunnel (Figure A-22a). This tunnel contains a set of tight binding pockets such that each side chain of the substrate peptide (P6 to P1’) is bound in a complementary site (S6 to S1’) (Figure A-22b). In particular, peptide side chain P6-Glu contacts a network of three hydrogen bonds; P5-Asn points into the solvent, making no specific interactions (hence the absence of substrate consensus at this position); P4-Leu is buried in a hydrophobic pocket; P3-Tyr is held in a hydrophobic pocket with a short hydrogen bond at the end; P2-Phe is also surrounded by hydrophobes including the face of the triad histidine; P1-Gln forms four hydrogen bonds; and P1’-Ser is only partly enclosed in a shallow hydrophobic groove.

Although rational design has had limited success in changing protease specificity, DE has been used to change the preferred residue either before or after the cleavage site. I discuss these works in Chapter E, 2.3, and approach specificity from an evolutionary point of view, attempting to change the specificity of multiple substrate residues simultaneously.
OVERVIEW OF RESULTS

In Chapter B, I describe the development of both a selection system and a screening system for protease activity detection and quantification. The screening system was then used for the subsequent evolution experiments of TEV protease.

Chapter C addresses the evolution of key active site residues by mutating the nucleophile from cysteine to serine. DE using small libraries recovered protease activity and combining this with HT-SAS, I characterised the evolutionary trajectory and its flanking fitness landscape.

I use high throughput DE and HT-SAS in Chapter D to investigate the evolution of protein robustness and tolerance to mutations. By selection only for retention of wild-type-like activity, I neutrally evolve a population of TEV protease genes. I use deep sequencing to reconstruct the complex phylogeny and show that robust sequences appear and that their tolerance to mutations allows their descendants to dominate later populations.

To test whether these populations also show altered evolvability, in Chapter E I look at adaptive evolution towards new (possibly clinically relevant) substrate sequences. By quantifying changes in evolvability of the population I separate out two distinct mechanisms whereby neutral mutations can increase different aspects of evolvability.

Finally, in Chapter F I draw these experiments together and comment on the characterisation of the fitness landscape surrounding TEV protease and its implications for evolvability and enzyme engineering.
SUMMARY

High throughput protease assays are necessary to detect fitness differences between variants and isolating those required for directed evolution. Both selection and screening systems were attempted. A cell survival selection system was constructed, in which the cleavage recognition sequence of Tobacco Etch Virus protease was successfully cloned into the toxic β-lactamase inhibitory protein. Though inhibitory activity was retained, co-expression of the protease did not reduce this inhibition to restore cell survival. In parallel, a screening system was adapted, based on the proteolytic destruction of fluorescence resonance energy transfer in a fusion of donor and acceptor fluorescent proteins. The assay was first optimised for 96-well plate lysate screening (in conjunction with P. Gatti-Lafranconi) and then for fluorescence-activated cell sorting. The 96 well lysate screen proved the more sensitive method, but the cell sorting screen was high throughput enough to be mainly limited by cell transformation efficiency.
2 INTRODUCTION (SELECTION)

In a directed evolution cycle of diversification, selection/screening and amplification, only the second step is protein specific. As discussed in Chapter A, 3.2, this requires either for the particular activity to be coupled to gene enrichment (selection) or for a high throughput assay (screening). As proteases are active on peptides, the substrate can be synthesised directly by ribosomes in vivo. This affords the opportunity of a cell-survival selection system based on the destruction of a toxic protein. Because the preference sequence is so long, it occurs by random once every $20^7=10^9$ times and so is highly unlikely to occur in a protein. Additionally, TEV protease cleaves in unstructured regions, so the sequence needs to be in a flexible loop to be accessible.

2.1 β-LACTAMASE INHIBITOR PROTEIN AND INHIBITION-RELIEF SELECTION

In general, selection systems are high throughput and allow a fast and crude sweep of a large library\(^\text{225}\). By engineering a cleavage site into a lethal protein, cleavage by the protease should restore vitality and allow isolation of active genes. An inhibitor protein of β-lactamase was identified as a good candidate by P. Gatti-Lafranconi as the selection pressure could be tuned by varying the ampicillin concentration.

_E. coli_ cells expressing β-lactamase are able to survive on ampicillin as β-lactamase hydrolyses the antibiotic. Co-expression of β-lactamase Inhibitor Protein (BLIP) in these cells prevents cell survival on ampicillin\(^\text{226}\). The structure of the interaction complex has been solved\(^\text{227}\) (Figure B-1) and the residues critical to this interaction mapped by mutagenic studies\(^\text{228}\). These demonstrated that BLIP competitively inhibits β-lactamase by binding over its active site and
inserting residues into the active site pocket. As β-lactamase is localised to the periplasm, export of BLIP is critical to function.

The proposed system consists of coupling the activity of the protease to cell survival by engineering a BLIP protein to act as a substrate of the protease (Figure B-2). The sequence must be inserted at a point that will not affect the binding of BLIP to β-lactamase whilst still being accessible to the protease. Cleavage of BLIP by the protease must destroy the interaction with β-lactamase. Thus, an active protease that is expressed in the periplasm would cleave BLIP, which could no longer inhibit β-lactamase, allowing cell survival on ampicillin.

The ampicillin concentration can be used to control the stringency of selection so that selection pressure can be adapted to the protease library being used. Selection generates a population enriched in functional variants. By mutating the inserted cleavage site, the system would be adaptable to any other protease that is sequence-specific enough to be non-toxic in *E. coli* but active enough to cleave BLIP.
3 RESULTS AND DISCUSSION (SELECTION)

3.1 BLIP PLASMID ASSEMBLY

For a functional selection system, it was important to ensure inhibition of β-lactamase by BLIP that is efficient enough to abolish ampicillin resistance but sensitive enough to have that resistance restored by BLIP’s proteolysis. To achieve this it was necessary to construct a custom plasmid that ensures 1:1 stoichiometry of BLIP to β-lactamase. Additionally, the lactose inhibitor (LacI) was required for controlling inducible expression and the chloramphenicol acetyltransferase (CAT) as an alternative antibiotic resistance marker when BLIP is being expressed.

I therefore designed several cloning strategies, eventually constructing the plasmid pSB-BLIP (Figure B-3). β-lactamase was amplified from pUC18 by PCR (using primers blaF+blaR) which was ligated into pJET1.2/BLUNT to make pJET-bla. The pJET-Bla plasmid was subsequently digested and β-lactamase inserted into the multiple cloning site of pSU18. To insert wtBLIP as well as each loop variant, the BLIP gene and LacI were amplified by PCR from pGR32-BLIP (using primers blipF+blipR) and also inserted into the plasmid. The final plasmids were designated pSB-BLIP.

Both β-lactamase and BLIP are preceded by the TEM1 signal sequence for export to the periplasm. Chloramphenicol acetyltransferase was included so that the plasmid can be maintained when ampicillin selection is not being applied.
Chapter B - Selection and screening methods for protease directed evolution

3.2 INSERTION OF CLEAVAGE SEQUENCE INTO BLIP PRESERVES FUNCTION

In order for the ampicillin inhibition-relief selection system (Figure B-2) to function, a cleavable BLIP protein needed to be engineered. As previously stated, the position of the insert has to satisfy the following criteria:

- The uncleaved loop in BLIP must not affect β-lactamase interaction
- The loop must be exposed and accessible for the protease to cleave
- Upon cleavage, the BLIP fragments must not inhibit β-lactamase

I identified two possible sites based on analysis of the structure of the BLIP/β-lactamase complex (Figure B-4). Loop regions were chosen in the hope that they would better tolerate the presence of insertions and minimally disrupt the rest of the structure. Both insert sites are situated away from the interaction surface with β-lactamase. Finally both sites roughly bisect the protein so that residues essential to the interaction^228,229 with β-lactamase are split between the two halves.
Loops were inserted using the USER cloning method\textsuperscript{12} with primers ins1F+ins1R and ins2F+ins2R to generate iBLIP and i2BLIP (Table B-1) in the original pGR32 plasmid.

The inserted loops are required to not impair inhibition of $\beta$-lactamase. I therefore developed an assay (with P. Gatti-Lafranconsi) of the efficiency of BLIP inhibition of $\beta$-lactamase based on the serial dilution of a bacterial culture, spotted onto solid media. At high cellular concentration, low $\beta$-lactamase activity allows cell survival because ampicillin is cleared by the cumulative activity of many cells. As cell concentration decreases, selection is more stringent. Addition of IPTG
induces expression of the BLIP construct. Cells expressing wtBLIP, had to be roughly 100-fold more concentrated in order to achieve the same survival on 100 µg/µl ampicillin hence BLIP reduces β-lactamase activity by approximately 100-fold in this assay (Figure B-5, upper half).

Having BLIP and β-lactamase expressed from a single plasmid (pSB-BLIP) gave repeatable results by reducing the effects of relative gene copy number variation. Cells expressing i2BLIP showed no decreased survival on ampicillin indicating that the insert has disrupted BLIP inhibition of β-lactamase. Conversely, iBLIP was only 2- to 4-fold less effective than wtBLIP at inhibiting growth on ampicillin (Figure B-5, lower half). This indicates that iBLIP still functions as an effective inhibitor and hence the insert loop minimally interferes with the BLIP/β-lactamase interaction. Given this success, iBLIP was further characterised for cleavability by TEV.

![Figure B-5 | Comparison of wtBLIP and iBLIP](image)

Top10 cells are transformed with pSB-wtBLIP or pSB-iBLIP. As the cell culture is diluted there is no change in survival when IPTG is absent and BLIP is not induced. However, when IPTG is added cells are unable to survive at low cell density. Reduction in survival is comparable between wtBLIP and iBLIP. 100 µg/µl ampicillin, 400 µM IPTG (as indicated) at 37°C.

**3.3 TEV PROTEOLYSIS OF BLIP DOES NOT RESTORE CELL SURVIVAL**

The final goal of this system was to select for active TEV variants so the sensitivity of iBLIP to proteolysis and restoration of ampicillin
resistance was crucial. Several plasmids were constructed that encode TEV protease but vary in its fusion state, localisation and activity.

MBP-TEV – TEV protease was provided on plasmid pMAT10 as an MBP fusion (P. Gatti-Lafranconi). As the resistance gene of pMAT10 is β-lactamase, (β-lactamase is present on pSB-iBLIP so cannot be used as the resistance gene on the TEV protease-containing plasmid) I cloned the MBP-TEV fusion into pET28a which encodes kanamycin resistance to produce the plasmid pET28-MBPTEV.

Signal sequence-TEV – As both β-lactamase and BLIP are exported to the periplasm, TEV needs to be co-localised in order to function. I therefore also cloned TEV protease alone into pET22b which contains the PelB signal sequence. The signal sequence and TEV protease sequence were then cloned together into pET28a (pET28-STEV).

Inactive TEV – An appropriate negative control was needed for defining the system’s dynamic range. I therefore made the active site C151A mutation (TEVAla, previously shown to be inactive\textsuperscript{212}) by targeted mutagenesis of pET28-STEV (primers tevaF+tevaR). Clones were confirmed by sequencing (T7F primer). The TEVAla gene was subsequently cloned into the pET28-MBPTEV plasmid.

For the selection system to function, active protease must relieve iBLIP inhibition of β-lactamase. When co-expressed, TEV protease expressed in the cytosol (pET28-MBPTEV) has no effect on cell survival. Surprisingly, however, co-expression of TEV in the periplasm (pET28-STEV) also did not increase cell survival (Figure B-6) and was as ineffective as the inactive C151A mutant (pET28-STEVAla). This indicates that the expected inhibition relief of iBLIP cleavage by TEV does not occur. Likely explanations for this negative result include TEV mis-translocation, failure of TEV to cleave iBLIP, continued inhibitory function of the cleaved iBLIP.
In order to understand which part of the pathway was not operating as planned, I expressed and purified both TEV and iBLIP. Purifications of iBLIP from whole lysate and periplasm only were performed. Expression of iBLIP was low (Figure B-7a) and multiple bands appeared around its expected size (21kDa) though it is unclear which, if any, was iBLIP. To characterise whether iBLIP is cleavable by TEV in vitro, the proteins were incubated together in TEV activity buffer.

Co-incubation of the iBLIP preparation with TEV in vitro gave no evidence of cleavage (Figure B-7b) since concentrations were so low. Together these results indicate that although the insertion of a cleavage loop into BLIP to create iBLIP was successful, this construct was not inactivated by co-expression of TEV. Troubleshooting of these
negative results was hindered by poor expression of the iBLIP construct.

Given the difficulty in addressing the failure point in the system, a more manageable system might be the cleavage of a cytosolic, modular repressor with an accessible, flexible, cleavable linker between the DNA binding and transcription inhibition domains. A fully unstructured region would ensure TEV cleavage and separate domains would ensure separation and inactivation after cleavage. In fact, during the course of this work, just such a system was published\textsuperscript{114} with either histidine biosynthesis, antibiotic resistance or fluorescent protein genes downstream of the repressor.
4 INTRODUCTION (SCREENING)

In parallel with work on the BLIP-based selection system, I worked with P. Gatti Lafranconi on a screening system. This was envisaged as both an alternative to cell survival selection as well as a possible lower throughput but higher detail addition for DE.

4.1 DUAL-GFP FRET PAIR AND FLUORESCENCE SCREENING

There are many assays for measuring the activity of purified protease. The throughput of most is not high enough for use in directed evolution\textsuperscript{208,221} (>\textsuperscript{10^2} variants screened per round). High throughput in vivo assays of protease activity tend to be constructed for sensitive detection of inactive proteases for the screening of inhibitors. A few methods, however, are good candidates for conversion to a screening system for directed evolution. A dual GFP-variant Fluorescence Resonance Energy Transfer (FRET) system\textsuperscript{117,230,231} was chosen for optimisation for protease library screening.

FRET occurs when the emission wavelength of one fluorophore coincides with the excitation wavelength of another nearby fluorophore such that the energy is directly transferred across. Thus excitation of the first fluorophore leads to emission by the second. FRET decreases as the sixth power of the distance between the fluorophores. In the case of covalently bonded fluorophores, cleavage can be considered to completely abolish FRET.

A dual-GFP FRET protease sensor\textsuperscript{117,230,231} has been adapted (with P. Gatti-Lafranconi) to be suitable for the directed evolution of Tobacco Etch Virus protease (TEV). A cyan GFP-variant (CyPet or CFP) is fused to a yellow GFP-variant (YPet or YFP) by a linker including the TEV substrate recognition sequence. The emission wavelength of CFP
and the excitation wavelength of YFP are very similar. When CFP is excited by a photon at 414 nm, the energy is transferred to YFP and emitted at 527 nm. Cleavage by TEV protease abolishes FRET (Figure B-8 Error! Reference source not found.) such that when excited at 414 nm, light is emitted from CFP at 475 nm.

Cells co-expressing both the FRET sensor and TEV are induced and incubated so that cleavage occurs in vivo (in the cytosol). When CFP is excited, the ratio of CFP to YFP is used as an indicator of FRET efficiency and hence protease activity. This fluorescence ratio can either be measured in the cell lysate of a clonal population representing each variant (96 well plate screening) or in each individual cell by flow cytometry and Fluorescence Activated Cell Sorting (FACS).

A particular benefit is that modifications of the same core assay can be used for in vitro kinetic characterisation of purified enzymes as well as the high-throughput in vivo screen ensuring consistency of substrate.
5 RESULTS AND DISCUSSION (SCREENING)

5.1 C-Y(TEV) FRET-PAIR PLASMID ASSEMBLY

I constructed a screening system based on the change in fluorescence of a substrate upon proteolysis. The CFP-YFP gene has previously been optimised for intramolecular FRET\(^{117}\) and caspase-3 cleavage. The initial stages of adapting the system for directed evolution was performed by P. Gatti-Lafranconi (introducing the TEV cleavage linker and hanging induction to L-arabinose to allow co-induction of substrate and protease. The substrate gene is on a plasmid also containing chloramphenicol acetyltransferase (Figure B-9) and the protease gene on a plasmid containing β-lactamase. Cells have to maintain both these plasmids when grown on ampicillin and chloramphenicol.

![Figure B-9 | pC-Y plasmid map](Image)

The plasmid used to express C-Y. The wedges indicate which starting plasmids were used to construct the final pC-Y plasmid (the gap indicates synthetic oligonucleotides). Gene abbreviations: CFP = Cyan fluorescent protein; YFP = Yellow fluorescent protein; MBP = Maltose binding protein; AraC = Arabinose inhibitor; Cat = Chloramphenicol acetyltransferase.
In order to facilitate the easy manipulation of linker sequence between CFP and YFP, I introduced restriction sites (XhoI, SpeI). Whole gene synthesis was necessary as mutagenic PCR was impossible due to the high sequence identity of CFP and YFP causing non-specific primer annealing. These sites allow the DNA encoding the protease cleavage sequence to be inserted with synthetic oligonucleotides. The linker (discussed in more detail in Chapter E) can be either a single sequence or a set of sequences with randomised positions (NNS codon). To begin with, I inserted the cleavage preference sequence of TEV protease to create pC-Y(tev):

```
...GGSGGSENLYFQSGGSGGS...
```

Spacer – **Cleavage seq** - Spacer

5.2 Fluorescence of C-Y(TEV) is changed by TEV proteolysis in cell lysate

A cell lysate screen was optimised for the parallel assay of TEV mutant libraries of $10^2$. Due to low solubility, TEV was expressed as a MBP fusion protein to provide a constant solubility buffer\textsuperscript{219} (Figure B-10a). Both the enzyme and substrate were cloned into plasmids that give L-arabinose inducible expression so that co-expression is induced simultaneously. For expressing TEV without the MBP fusion for biophysical characterisation of the unfused enzyme, it was also cloned into pMAA NoMBP (Figure B-10b).
I ran control screens comparing the FRET ratio of cells co-expressing the C-Y(tev) substrate with either wild type TEV, the C151A mutant (TEV\textsuperscript{Ala}), or MBP. In the absence of TEV, there was an emission peak at 525 nm (the emission wavelength of YFP indicative of FRET) (Figure B-10).
In the presence of TEV, emission was mostly at 475 nm (the emission wavelength of CFP). This gives a dynamic range of a 7-fold drop in FRET efficiency (525/475 nm) in the presence of TEV. This demonstrates that the FRET pair functions with the TEV cleavage sequence as a linker and that it can be successfully proteolyed.

The FRET ratio was time dependent (Figure B-12). Over time, expression of the substrate increased FRET ratio, similarly longer co-expression times with TEV lead to a lower FRET ratio. Readings were more repeatable at later time points as the signal to noise ratio drops.
The assay can be performed in parallel in a 96 well plate on a library of protease variants generated by error prone PCR (epPCR) (Figure B-13a). Next generation sequencing (see Chapter D, 3.1) confirmed that variants each contained 1.5 nucleotide mutations on average (spread even along the gene) with 34% of these being silent mutations and <1% being frameshifts. The lysate assay allowed for screening of mutant library sizes of $10^2$. The FRET ratios of mutants mostly cluster around the FRET ratios of cells co-expressing wild type TEV (cleaved substrate) or cells co-expressing MBP (uncleaved substrate) (Figure B-13b).

5.3 TEV CLEAVES PURIFIED C-Y(TEV) IN VITRO

All proteases and C-Y(TEV) variants contain a His$_6$ tag for nickel affinity purification. C-Y(TEV) variants could be purified to high concentrations
(up to 90 µM) but TEV had poor solubility (maximum 4 µM). At low concentrations the fluorescence of C-Y(tev) variants was insufficient to measure FRET ratio however at concentrations above 0.5 µM FRET ratio was constant (Figure B-14).

A concentration of 1 µM substrate was therefore used in all further kinetics in line with the concentrations used in the original report of the FRET pair\(^\text{117}\). When thawed, purified C-Y(tev) variants did not exhibit their maximum FRET ratio (Figure B-15).

A >4 hour equilibration period at the assay temperature was therefore required in order to give a constant negative control. FRET ratio remained stable for several days after thawing. It is only the FRET ratio that equilibrates in this fashion, not the total fluorescence. I would
hypothesise that this is indicative of slow structural rearrangements of the relative orientations of the fluorophores.

Cleavage of C-Y(tev) by purified TEV is specific (neither thrombin nor TEV\textsuperscript{Ala} showed substrate turnover) and concentration dependent (Figure B-16a). This indicates specific hydrolysis of the C-Y(tev) linker by TEV using the catalytic cysteine nucleophile. Kinetic traces could be globally fitted with the Michaelis Menten function (Methods eq. G-2) in order to determine the composite kinetic activity parameter, $k_{cat}/K_M$ (Figure B-16b).

Figure B-16 | Concentration dependent cleavage of C-Y(tev) by TEV and kinetic fitting
Activity of different 0.5, 1, 2, 4 $\mu$M TEV (red), 5 $\mu$M TEV\textsuperscript{Ala} (green) and 30 $\mu$M Thrombin (blue) on 1 $\mu$M C-Y(tev). FRET ratio used to determine product concentration by fitting to eq. C-1 (Methods) (b) Data for 1 $\mu$M TEV activity fitted with eq. C-2 (Methods)

5.4 TEV PROTEOLYSIS OF C-Y(TEV) IS DETECTABLE IN INDIVIDUAL CELLS

The lysate co-expression assay was modified to increase throughput from $10^2$ to $10^6$. To test the dynamic range of the assay in FACS, TOP10 E. coli containing pC-Y(tev) were transformed with either pMAA(TEV) or pMAA. After overnight co-expression the clonal populations were assayed by FACS. The populations were
distinguishable by the difference in FRET ratio (After excitation of CFP, the emission of YFP divided by emission of CFP - FL5/FL4) however there was significant overlap (Figure B-17a). YFP was used as a reporter for overall expression since the emission of YFP after its excitation (FL1) is independent of cleavage of the CFP-YFP fusion. When separated by FL1 it can be seen that cells with higher expression also have less signal variance due to higher signal:noise (Figure B-17b,c,d).

Since the measurement of cleavage is ratiometric it is concentration independent, however high fluorescence gives higher signal:noise in the FRET ratio. Due to the small size of E. coli, achieving high enough substrate C-Y concentration required expression optimisation. It has previously been reported that low temperature expression improves fluorescence by reduced aggregation\textsuperscript{117}. I found that overnight expression at 21 °C gave the highest signal. Additionally slow flow rate (<1000 cells s\textsuperscript{-1}) to give longer laser beam passage time and high laser
power (>100mW) were necessary for high signal:noise ratio. Though these increase the signal:noise ratio and give a reduced variance in the FRET ratio, the median ratio remains the same. The ratio was also insensitive to the concentration of inducer (0.02%-1.00%) which is important for consistent library screening.

A library of TEV variants generated by epPCR was co-expressed with C-Y(tev) and assayed by flow cytometry. The library showed two populations that correspond to the FRET ratios of populations co-expressing TEV or expressing MBP (Figure B-18). The distribution of FRET ratios of cells in the top 10% of FL1 was roughly bimodal. Bimodality indicates that, in this assay, mutations are either effectively neutral or fully deleterious to TEV function (discussed in Chapter D).

![Flow cytometry plots showing FL1 (Emission from YFP after excitation of YFP) against FL5/FL4 (FRET ratio after excitation of CFP) for (a) cells co-expressing C-Y(tev) and either MBP (blue) or TEV (red) (b) a density plot of a library of TEV mutants.](image)

The majority of noise in the system is biological rather than technical. If a specific subpopulation of cells are sorted and immediately reassayed, they show very little spread. If the sorted cells are grown and re-expressed, however, they produce a pattern identical to the original, unsorted population indicating that the re-expression has re-introduced biological variation.
6 CONCLUSIONS

6.1 SELECTION BY BLIP CLEAVAGE IS UNSUITABLE FOR DIRECTED EVOLUTION

The plasmids and genes required to test the system were successfully constructed. Of the two loops targeted, only one allowed retention of inhibitory activity after insertion of the TEV cleavage sequence as demonstrated by the reduction in survival when iBLIP is expressed (Error! Reference source not found.). This demonstrates that loop 1 tolerates the insertion of seven amino acids without disrupting function. Co-expression of TEV, however, fails to restore survival (Figure B-6) indicating that TEV fails to successfully inactivate BLIP and restore vitality. There could be a number of reasons for this failure:

- TEV is not soluble or active in the periplasm
- TEV is not exported to the periplasm by PelB export sequence
- TEV does not cleave iBLIP because the loop is not exposed
- Cleavage of iBLIP does not destroy the interaction with β-lactamase

The main problem with the system is that it is difficult to isolate the point of failure in the multi-step pathway and the low expression of BLIP has made purification for in vitro analysis impractical. Since the main benefit of a cell survival selection system is its high throughput, the success of the high throughput FRET screening system in FACS system rendered this selection system obsolete.

6.2 FRET SCREENING PROVIDES MULTIPLE PLATFORMS FOR DIRECTED EVOLUTION

The CFP-YFP FRET-pair lends itself well to a variety of assay formats, each specialised for a different function. Purified proteins can be used to determine kinetic parameters by whole curve fitting (Figure B-16), cell lysate screening is a sensitive screen for medium throughput directed evolution ($10^2$) (Figure B-13) and flow cytometry allows for high throughput screening ($10^6$) (Figure B-18). Using the same
substrate for both screening and detailed characterisation avoids some artefacts common in directed evolution studies. In particular, directed evolution is rarely possible on the desired substrate and so one or more proxy colourmetric or fluorometric substrates are used. Improvements in turnover of the assayed substrate may not correlate to activity on the desired substrate as evolution optimises binding the colourmetric or fluorometric group or catalysis relies on their artificially high pKa values. The FRET system is suitable for high and low throughput screening as well as subsequent kinetic characterisation.

The substrate cleavage sequence can also be manipulated to create individual variant substrates or a whole class of related substrates by randomising one or more codons, allowing the study of specificity by measuring the set of cleaved substrates.

The mutant library screened by lysate and FACS gives a good comparison between the assay formats. The overall agreement is good, both showing a bimodal distribution of mutant activities (Figure B-13 & Figure B-18). The lysate screening system has higher dynamic range and greater sensitivity at discriminating between similar variants. Additionally the signal:noise ratio is high because the lysate screens the average FRET ratio of a large clonal population of E. coli. It is therefore well suited to the evolution from a low starting activity by isolating the single best variant from each round. The FACS system, conversely, was able to screen 4 orders of magnitude more variants and gives the ability to sort large populations. It is therefore better suited to experiments where an evolving population is generated.

These FRET-based screening systems enabled the directed evolution of TEV and are used in the further experiments laid out in this dissertation.
Chapter C - Handicap-recover evolution to accommodate a new nucleophile

1 SUMMARY

To understand the evolution of active sites, handicap-recover directed evolution was performed to convert Tobacco Etch Virus cysteine protease to a functional serine protease. Catalysis using serine was improved $>10^3$-fold but rather than re-specialise, the resulting enzyme was able to use either nucleophile. Description of the fitness landscape flanking the evolved trajectory revealed features that constrain and guide the evolution of active site architecture. In particular, high epistasis creates a rugged landscape but there are consistently several available uphill pathways at each step along the trajectory. Though mutations in the protein core are penalised, the active site second periphery is enriched in adaptive potential.

Parts of this chapter have been submitted for publication:

Shafee, T., Gatti-Lafranconi, P., Minter, R., Hollfelder, F.
The evolvability of nucleophile exchange Nature Genetics. (2013)
INTRODUCTION

2.1 THE EFFECT OF NUCLEOPHILE IDENTITY ON ENZYME MECHANISM

As introduced in Chapter A, nucleophilic enzymes use an interconnected set of active site residues to achieve catalysis. The sophistication of the active site network causes residues involved in catalysis, and residues in contact with these, to be the most evolutionarily conserved within their families. In catalytic triads, the most common nucleophiles are serine (an alcohol) or cysteine (a thiol). Compared to oxygen, sulphur’s extra d orbital makes it larger (by 0.4 Å), softer, form longer bonds ($d_{C-X}$ and $d_{X-H}$ by 1.3-fold) and have lower $pK_a$ (by 5 units). Here I concentrate on chemical differences between cysteine and serine proteases on catalytic chemistry, however similar issues affect hydrolases and transferases in general.

The $pK_a$ of cysteine is low enough that some cysteine proteases (e.g. papain) have been shown to exist as an $S^-$ thiolate ion in the ground state enzyme and many even lack the acidic triad member. Serine is also more dependent on other residues to reduce its $pK_a$ for concerted deprotonation with catalysis by optimal orientation of the acid-base triad members. The low $pK_a$ of cysteine works to its disadvantage in the resolution of the first tetrahedral intermediate as unproductive reversal of the original nucleophilic attack is the more favourable breakdown product. The triad base is therefore preferentially oriented to protonate the leaving group amide to ensure that it is ejected to leave the enzyme sulphur covalently bound to the substrate N-terminus. Finally, resolution of the acyl-enzyme (to release the substrate C-terminus) requires serine to be re-protonated whereas cysteine can leave as $S^-$. 
Figure C-1 | Differences in cysteine and serine proteolysis mechanisms
As in Figure A-2, the protease (black) performs a nucleophilic attack on peptide substrate (red) to form a tetrahedral intermediate. This breaks down by ejection of the first product, the substrate C-terminus, to form the acyl-enzyme intermediate. Water replaces the first product and hydrolysis occurs via a second tetrahedral intermediate to regenerate free enzyme. Indicated differences are (a) the deprotonated cysteine, (b) aspartate (grey) not present in all cysteine proteases, (c) concerted deprotonation of serine, (d) aspartate hydrogen bonding, (e) amide protonation of the first leaving group, (f) alcohol protonation of the serine leaving group.
Sterically, the sulphur of cysteine also has longer bonds and a bulkier Van der Waals radius to fit in the active site\textsuperscript{237} and a mutated nucleophile can be trapped in unproductive orientations. For example the crystal structure of thio-trypsin indicates that cysteine points away from the substrate, instead forming interactions with the oxyanion hole\textsuperscript{234}.

The evolutionary specialisation of enzymes around the needs of their nucleophile makes it unsurprising that nucleophiles cannot be interconverted in extant proteases\textsuperscript{212,238–243} (nor in most other enzymes\textsuperscript{244–249}) and the large activity reductions (>10\textsuperscript{4}) observed can be explained as a result of compromised reactivity or structural misalignment.

\subsection*{2.2 Nucleophile exchange in natural evolution}

Although extant enzymes cannot switch nucleophiles, there are several superfamilies that show divergent evolution of serine and cysteine proteases from a common ancestor. I use the PA clan\textsuperscript{250} of chymotrypsin-like proteases as a model superfamily to study nucleophile switches in evolution. The PA clan contains a diverse array of proteases from eukaryotes, prokaryotes and viruses. It also encompasses varied functions including blood clotting (e.g. thrombin), digestion (e.g. trypsin), snake venoms (e.g. pit viper haemotoxin), bacterial toxins (e.g. exfoliative toxin) and viral polyprotein processing (e.g. polio, norovirus, and TEV proteases).

Despite retaining as little as 10\% sequence identity, PA clan members isolated from viruses, prokaryotes and eukaryotes show structural homology. Members were therefore aligned by structural similarity (with DALI\textsuperscript{251}). The alignment was expanded by adding sequences without known structure by sequence homology (with BLASTp, see
methods Table G-5) and this final alignment used to generate a phylogeny (Figure C-2). My phylogeny suggests three independent nucleophile switching events (crossed circles), though nodes 1 and 2 are so deeply rooted that they may be a single event. Cellular proteases use serine as the nucleophile, but both cysteine and serine protease families are found in the viruses (indicated by red and black shading respectively in inner circle). Despite the fact that replacing the nucleophile has detrimental effects on the activity of extant proteases\textsuperscript{212,238–243}, the PA clan shows that the distinct active sites of cysteine and serine proteases can arise from divergent as well as convergent evolution. There is currently, however, no model that explains evolution of such core components of the catalytic machinery.

![Figure C-2 | PA clan phylogeny](image-url)

The phylogeny of PA clan of proteases indicates that nucleophile changes have occurred in evolution (crossed circles). The colour of the branches and inner ring indicates the identity of the nucleophile (cysteine in red, serine in black), outer ring indicates organism (viral in white, prokaryote in light blue, eukaryote in blue). Proteases with known structure (marked by black dots) were aligned to TEV\textsuperscript{Cys} on the basis of structure using DALI, sequences of unknown structure were added by sequence alignment with BLASTp; the maximum likelihood phylogeny was generated using MEGAS.
3 RESULTS AND DISCUSSION

For this chapter, Tobacco Etch Virus nuclear inclusion cysteine protease will be referred to as TEV<sup>Cys</sup>. The nucleophile identity of all variants will also be indicated by the superscript.

To investigate nucleophile switches in evolution I mutated the nucleophile of Tobacco Etch Virus nuclear inclusion cysteine protease (TEV<sup>Cys</sup>) to serine (TEV<sup>Ser</sup>). I performed 10 rounds of Directed Evolution (DE) and characterised the recovery of activity by two complementary methodologies. Firstly, I used biochemistry to investigate changes in catalytic activity and protein stability. Secondly I used High Throughput Screening and Sequencing (HT-SAS) to characterise the local fitness landscapes flanking the DE trajectory.

3.1 REPLACEMENT OF NUCLEOPHILE DECREASES TEV ACTIVITY 10<sup>4</sup>-FOLD

Targeted mutagenesis was used to introduce the C151S mutation into TEV<sup>Cys</sup> to create TEV<sup>Ser</sup> (primers tevsF+tevsR). The C151S mutation causes a large reduction in kinetic activity of the purified enzyme (TEV<sup>Ser</sup>) in line with other proteases<sup>212,238–243</sup>. Additionally, the replacement of the nucleophile with serine induces biphasic kinetics (Figure C-3), indicating a fast first step followed by a slower, rate-limiting step. A standard steady-state interpolation (Equation C-1) can only accurately fit $k_{cat}/K_M$ for TEV<sup>Cys</sup> because the first step is rate-limiting. For TEV<sup>Ser</sup> and variants, the modified Equation C-2 was used, as it allows measurement of the rate of all steps before and after the rate-limiting step ($k_{obs1}$ and $k_{obs2}$) (Table C-1). Compared to the $k_{cat}/K_M$ of TEV<sup>Cys</sup>, the $k_{obs1}$ and $k_{obs2}$ of TEV<sup>Ser</sup> were found to be 80 and 20,000-fold lower, respectively, indicating that a strong handicap is associated with nucleophile replacement in TEV<sup>Cys</sup>. Given the mechanism shown in section 2.1, it appears likely that the slow second step refers to de-
acylation of the covalent intermediate. It could also be a slowed product-release step after catalysis, however this seems less likely given the nature of the mutation.

\[
[P] = [P]_{\text{max}} - e^{-[E]k_{\text{cat}}/K_M t}
\]

**Equation C-1 | Standard Michaelis Menten interpolation**

- \([P]\) = concentration of product
- \([P]_{\text{max}}\) = final concentration of product (should be = starting substrate)
- \([E]\) = concentration of enzyme
- \(k_{\text{cat}}/K_M\) = rate of overall reaction
- \(t\) = time

\[
[P] = [P]_{\text{max}} - (A_1 e^{-[E]k_{\text{obs1}} t} + A_2 e^{-[E]k_{\text{obs2}} t})
\]

**Equation C-2 | Modification of Eq 1 for biphasic reaction**

- \([P]\) = concentration of product
- \([P]_{\text{max}}\) = final concentration of product (should be = starting substrate)
- \(A_1\) = Jump amplitude (ie. how much turnover is performed in first kinetic step)
- \(A_2\) = Second jump amplitude (should = \([P]_{\text{max}} - A_1\))
- \([E]\) = concentration of enzyme
- \(k_{\text{obs1}}\) = rate of first kinetic step
- \(k_{\text{obs2}}\) = rate of second kinetic step
- \(t\) = time

**Figure C-3 | Monophasic and biphasic curve fitting of TEV\text{Cys} and TEV\text{Ser}**

Kinetic data for cleavage of 1 µM substrate by (a) 4 µM TEV\text{Cys} and (b) 20 µM TEV\text{Ser} fit with single exponential (Eq. C-1) or double exponential (Eq. C-2).

**Table C-1 | Regression of curve fitting of TEV\text{Cys} and TEV\text{Ser}**

Summary of kinetic parameters and \(R^2\) values for fits to single or double exponential model.
3.2 NUCLEOPHILE MUTATION HAS MINIMAL EFFECT ON CATALYTIC PROMISCURITY

Trypsin shows a range of catalytic promiscuity. I therefore checked for promiscuous hydrolysis of a range of substrates: phosphate, phosphonate, phospho-diester, sulphate, sulphonate, acetate, phosphoryl-choline and xylopyranoside. Of these, only activity on phosphonates was detectable (Figure C-4) and found to be 2.7 min⁻¹M⁻¹, 10⁴ worse than the native protease activity. Surprisingly, the activity of TEV<sup>Ser</sup> was 0.8 min⁻¹M⁻¹ indicating that this promiscuous activity is less sensitive to the nucleophile mutation than the native activity.

![Promiscuous phosphonatase activity of TEV<sup>Cys</sup> and TEV<sup>Ser</sup>](image)

Figure C-4 | Promiscuous phosphonatase activity of TEV<sup>Cys</sup> and TEV<sup>Ser</sup>
Hydrolysis of different concentration of pNp-phenolphosphonate by 10 µM enzyme in TEV activity buffer at 25 °C.

3.3 DIRECTED EVOLUTION RESTORES PROTEOLYTIC FUNCTION USING SERINE

In order to investigate how the enzyme can compensate for the handicap of using a non-native nucleophile, 10 rounds of directed evolution (DE) were performed on TEV<sup>Ser</sup> by selecting for activity recovery in the presence of the mutated nucleophile.

Libraries of variant genes were generated by error prone PCR (epPCR), such that on average each resulting gene contained 1.3±0.4 amino acid mutations. Libraries were screened by the lysate assay detailed in
Chapter B, 5.2. Briefly, libraries were transformed into *E. coli* Top10 cells bearing the CFP-YFP substrate plasmid and the lysates from 350 individual colonies co-expressing protease and substrate were screened in each round. The single most active variant was used as the template for the next epPCR library and any S151C revertant rejected to force the recovery to follow a ‘forward’ evolutionary pathway. Revertants were predicted to occur once per thousand variants and results agreed with a revertant found in every second round (on average).

In this ratiometric, co-expression assay, the effect of the handicapped nucleophile caused a more moderate 6-fold cleavage reduction. Due to the sensitivity of the lysate assay, this was enough to allow directed evolution. A co-expression time of 4 hours gave good signal strength (high enough substrate concentration) whilst also giving good dynamic range (no saturation).

Over 10 rounds of evolution (TEV$^{Ser}$→TEV$^{Ser/X}$), the *in vivo* activity was recovered from 17 to 75% of TEV$^{Cys}$ in the cell lysate assay (Figure C-5a & Table C-2). To determine which enzyme properties were responsible for the improvement, I purified each enzyme variant for kinetic characterisation. *In vitro* kinetics indicated that the *in vivo* activity increase is due to a larger burst amplitude and an improvement in both $k_{obs1}$ (2x10$^3$-fold) and $k_{obs2}$ (3x10$^3$-fold) (Figure C-5b), with diminishing increments in later rounds, typical of DE trajectories$^{53}$. Despite the recovery, kinetics of all isolated variants were still best fit to the biphasic model (Table C-3). Conversely, S151C revertants could be fit to monophasic kinetics.
Table C-2 | Variant Property Summary

<table>
<thead>
<tr>
<th>Round</th>
<th>Details</th>
<th>Serine Mutations</th>
<th>Cysteine Mutations</th>
</tr>
</thead>
</table>
| I     | K6N, H28L, R50G, E223G, A231V | 6.4% 20% 1% | 39.4% 25% 1%
| II    | K6N, H28L, R50G, T113S, K215R, E223G, A231V | 3.3% 17% 7% | 30.4% 14% 7%
| III   | K6N, H28L, R50G, T113S, K215R, E223G, A231V | 2.2% 10% 3% | 43.6% 20% 3%
| IV    | K6N, H28L, R50G, T113S, L210M, K215R, E223G, A231V | 1.1% 6% 2% | 38.5% 14% 2%
| V     | K6N, H28L, R50G, T113S, L210M, K215R, E223G, E194D | 1.5% 5% 2% | 54.4% 25% 2%
| VI    | K6N, H28L, R50G, T113S, A206T, L210M, K215R, E194D, E223G | 0.5% 3% 1% | 39.5% 14% 1%
| VII   | K6N, H28L, R50G, T113S, A206T, L210M, K215R, E194D, E223G, (Frameshift) | 0.4% 2% 1% | 30.4% 12% 1%
| VIII  | K6N, H28L, R50G, T113S, A206T, L210M, K215R, E194D, E223G, (Frameshift) | 0.4% 2% 1% | 30.4% 12% 1%

For serine variants: \( k_{\text{obs1}} \) (min \(-1\) M \(-1\)), burst amplitude (Proportion of [E]), \( k_{\text{obs2}} \) (min \(-1\) M \(-1\)), soluble expression (% of wt), and thermal stability (°C). For serine variants: burst amplitude (Proportion of [E]), soluble expression (% of wt), and thermal stability (°C). For the cysteine back-mutants: \( k_{\text{cat}}/K_{\text{M}} \) (min \(-1\) M \(-1\)), soluble expression (% of wt), and thermal stability (°C).

*Mutations K229R, E230K and Δ231-236 are all caused by a 1nt frameshift.
The process of transcribing and translating a gene to make an enzyme is not perfect with an error made every $10^4$ residues. It is possible that the apparent activity of TEV$^{\text{Ser}}$ (and variants) is due contamination by enzymes with the original cysteine nucleophile. To rule this out, the identity of the nucleophile was confirmed in the proteins TEV$^{\text{Ser}}$ and TEV$^{\text{Ser}}$ IV by mass spectrometry and sensitivity to PMSF (a serine protease inhibitor) but not Io-Ac (a cysteine protease inhibitor).

Table C-3 | Regression of curve fitting of evolved variants
Summary of $R^2$ values for fits to single exponential (Eq. 1) or double exponential (Eq. 2) model for all variants. Kinetic data for cleavage of 1 µM substrate 20 µM TEV$^{\text{Ser}}$ and 4 µM of all other variants.
3.4 The evolved protease can use either nucleophile

Natural enzymes are sensitive to nucleophile mutations because of the evolutionary specialisation of their active site\textsuperscript{244–249}. Therefore, DE mutations were expected to re-specialise the enzyme for the serine nucleophile. However, nucleophile S151C back-mutants of all TEV\text{Ser}→TEV\text{Ser}X variants (TEV\text{Cys}→TEV\text{Cys}X) retained \textit{in vivo} activity levels similar to TEV\text{Cys} in the lysate assay (Figure C-5a). The kinetics of purified enzyme variants show that the modifications to the enzyme active site architecture prove nearly-neutral to activity with the original cysteine nucleophile. Consequently while the $k_{obs1}$ and $k_{obs2}$ of TEV\text{Ser}X are 10\textsuperscript{3}-fold improved over TEV\text{Ser}, the $k_{cat}/K_M$ of TEV\text{Cys}→TEV\text{Cys}X variants fluctuated by less than 4-fold (Figure C-5b), giving a trajectory that lacks a strong trade-off in nucleophile use (Figure C-6).

This neutrality may be because the mutations that accumulate to improve serine’s nucleophilicity have little effect on cysteine catalysis due to its higher intrinsic nucleophilicity\textsuperscript{22}. Mutations did not occur that fully specialise the active site towards serine (by failing to create a thiolate zwitterion\textsuperscript{239}, sterically clashing\textsuperscript{237}, or causing hyper-reactive oxidisation to sulphinic acid\textsuperscript{252}).

Figure C-6 | Nucleophile activity tradeoff
Plot of $k_{cat}/K_M$ activity using the original cysteine nucleophile versus (a) $k_{obs1}$ and (b) $k_{obs2}$ with a serine nucleophile for lineage.
Chapter C - Handicap-recover evolution to accommodate a new nucleophile

3.5 Catalytic activity is restored at the expense of structural integrity

Previous works have observed trade-offs between improved activity and structure destabilisation\textsuperscript{61,176,253}. Function-altering mutations close to the active site also tend to be in the core, which does not tolerate mutations well. Since the total amount of active protein present in a cell is a complex function of expression, folding, unfolding and aggregation, I used a set of techniques to characterise the stability of variants.

Differential scanning fluorimetry measures protein melting temperature ($T_M$). For this, enzyme variants had to be expressed and purified without the MBP fusion partner since this would mask the signal from thermal unfolding of TEV. Secondly, the soluble expression of variants still fused to MBP was measured by gel band densitometry and indicates the total amount of active protein present in the cell. Finally an in silico folding energy estimate could be evaluated by FoldX which calculates the total energy difference between the folded and unfolded states ($\Delta G$ of folding) as well as the change in this difference upon individually mutating of each residue to alanine ($\Delta \Delta G$ of alanine mutation).

The nucleophile mutation did not affect melting temperature ($T_M$) of purified protein (Figure C-7a) and quantification of soluble expression in \textit{E. coli} showed only a 10% decrease in TEV\textsuperscript{Ser} (Figure C-7b). FoldX also only predicted a minor destabilisation of the folded state (Figure C-7c). In contrast, during the evolution, there was a negative trade-off between activity and soluble expression (Figure C-7b), a frequently reported compromise\textsuperscript{61,253}. FoldX calculations agreed with this by predicting increasing destabilisation of the folded state along the DE trajectory (Figure C-7c). By the tenth round (TEV\textsuperscript{SerX}), the >1000-fold increase of $k_{obs1}$ and $k_{obs2}$ was accompanied by a 6-fold drop in soluble expression. In rounds where solubility decreased, mutations were in
residues with low solvent accessibility (0.1) whereas the average solvent accessibility of mutations in rounds where solubility increased was higher (0.6).

This negative pleiotropy was temporarily broken in rounds V and VI by two surface mutations (W130C and E194D) and a C-terminal truncation (due to a frameshift), which allowed the increase of both activity and solubility.

![Figure C-7](image)

**Figure C-7 | Stability losses during recovery of TEV\textsuperscript{Ser} lineage**

(a) Soluble expression (gel band quantification) and (b) thermal denaturation temperature of selected variants (Differential scanning fluorimetry without MBP fusion partner). (c) Energy difference between the folded and unfolded states ($\Delta G$) for all evolved variants (grey) and S151C revertants (white). Values are normalised to TEV\textsuperscript{Cys}, higher values indicate lower stability. (d) Distribution of the destabilisation effects of alanine mutations to all residues in TEV\textsuperscript{Cys}. Higher values indicate greater destabilisation by mutation to alanine. All variants show similar profiles. Error bars represent standard deviation of: 2 biological repeats of soluble expression, 3 technical repeats of DSF, 2 technical repeats of FoldX.

3.6 **Mutations that recover activity cluster around the active site**

To examine the distribution of mutations within the protein structure I used the published structure\textsuperscript{211} (PDB code 1lvm) to separate the enzyme into concentric shells emanating outwards from the catalytic triad. The first shell includes all residues within 4 Å of His46, Asp81 and Ser151. The second shell is all residues 4 Å from the first shell, and
so on. Mutations accumulated in the DE lineage (Table C-2) were not evenly distributed through the structure (p=0.005). The triad first shell did not accumulate any mutations. The triad second shell, however, contained 54% of mutations despite comprising only 23% of the protein (Figure C-8). Shells further than the second, conversely, were depleted in mutations. These observations suggest a strong selective pressure for the improvement of amino acid interactions around the active site, consistent with the idea that the non-native nucleophile had to be re-tuned.

Given that more than one mutation accumulated in each round, I performed a back-shuffle of the round II, V and VIII variants with TEV\textsuperscript{Ser} to see which mutations could be neutrally reverted to wild type at different rounds. This involved digesting the variant genes into 50 nt fragments, mixing them together with the similarly digested TEV\textsuperscript{Ser} gene and using PCR to reassemble the shuffled products. For reassembly, each evolved variant was mixed with TEV\textsuperscript{Ser} in a ratio such that 2-3 of the evolved mutations would be randomly reverted in each member of the shuffled library. Sequencing of 10 active variants from each shuffled library was expected to reveal some mutations that are
universally required for improved function (and so active variants never show reversion) whereas others would be neutral (and so would be reverted in variants retaining high activity). Mutations close together, however, rarely recombined and so the effects of some mutations could not be separated (e.g. A206T & L210M).

What was actually found was evidence of epistasis between the mutations accumulated in the lineage (Table C-4). This is consistent with other previous work on evolution of active site residues which looked in detail at the epistatic interactions between three residues. Reversions of H28L, L210M and K215R were tolerated from TEV Ser but not from TEV Ser VIII. Similarly, reversion of E224G in TEV Ser II did not reduce in vivo activity but its removal from TEV Ser VIII does. There are therefore no mutations that remain completely neutral. The finding that some mutations are nearly-neutral in early rounds yet adaptive in later rounds suggests epistatic dependence of later adaptive mutations on these early nearly-neutral substitutions. This makes sense as the mutations are tightly clustered, with many contacting each other in the structure. Unfortunately, multiple attempts to crystallise variants along the DE trajectory were unsuccessful. Even the wild type protease TEV Cys is reportedly hard to crystallise due to poor reproducibility.

<table>
<thead>
<tr>
<th>Table C-4</th>
<th>Back-shuffle mutation retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of 10 variants that retain activity after back-shuffle with TEV Ser. Shown are the proportion of active variants that retain each mutation. Back-shuffles were performed by shuffling TEV Ser with each of TEV Ser II, TEV Ser V and TEV Ser VIII such that 2-3 reversions occurred on average per variant. Mutations K229R, E230K and Δ231-236 are all caused by a 1nt frameshift and so are included in the table as 'STOP'.</td>
<td></td>
</tr>
</tbody>
</table>

The $>10^3$ increase in activity using serine was achieved by relatively few mutations (TEV Ser X has 92.4% sequence identity to TEV Cys) even
though the most closely related natural serine proteases have only 15% identity.

The closeness in protein sequence space of this nucleophile-permissiveness to an extant enzyme explains how divergent evolution of core catalytic machinery can occur within evolutionary superfamilies (such as the PA clan, Figure C-2). The ancestral gene basal to the divergent serine and cysteine branches of the PA clan could have been nucleophile-permissive despite modern members being unable to tolerate the switch.

3.7 Measurement of Local Fitness Landscapes Along the Trajectory

Directed evolution follows a single pathway through a fitness landscape and is a useful tool to understand how a protein can evolve, and what properties the evolving enzyme gains and loses\(^6\). The focus on detailed characterisation of a single trajectory, however, misses the broader context of the surrounding fitness landscape. In particular, it would be useful to know which properties of the evolved trajectory are general, and which occurred by chance in that specific lineage. I therefore used the Fluorescence Activated Cell Sorting (FACS) assay (Chapter B, 5.4) for High Throughput Screening and Sequencing (HTSAS) to characterise the local fitness landscapes flanking the trajectory. I aimed to address to what extent the following observations were stochastic events, or reflective of features of the fitness landscape:

- Rapid evolutionary recovery of activity using the new nucleophile
- Clustering of adaptive mutations around the active site periphery
- Trade-offs between improved activity and protein stability
- Epistatic interactions between the accumulated mutations

For each of the mutagenised libraries generated in rounds I to X of the evolutionary trajectory, \(\approx 10^6\) members were re-screened by flow cytometry. The library members exhibiting the top 1% activities were
isolated and a representative sample of $\approx 10^3$ mutants subjected to 454 sequencing. Frequency of mutations before and after selection was measured and used to calculate the enrichment factor (for each position of the sequence). Mutations that are enriched are beneficial under the selection pressures applied, those that are purged are deleterious. Linking sequence and function provides insight into the series of local fitness landscapes flanking the DE trajectory (Figure C-9). Furthermore, mapping functional effects on the protein structure indicates which regions of the structure are constrained and which are enriched in adaptive potential.

![Schematic of sequence space flanking trajectory](image)

Figure C-9 | Schematic of sequence space flanking trajectory
A 2D simplification of the multi-dimensional space of all possible protein sequences. Similar sequences are closer together in this space. (a) The TEV$^{\text{Ser}}$→TEV$^{\text{Ser}}$X trajectory generated by directed evolution. Only the trajectory members (large circles) are characterised. (b) Many related sequences (small circles) surround members of the directed evolution trajectory. High throughput selection and sequencing allows some of their properties to be characterised.

After HT-SAS, 29% of TEV$^{\text{Ser}}$ residues showed greater than 2-fold enrichment of mutations (Error! Reference source not found.). The protein surface and regions away from the active site were particularly enriched (52% and 36% respectively). The same analysis was performed for all variants along the DE lineage up to TEV$^{\text{Ser}}$IX (Figure C-11) to build an overlapping series of local fitness landscapes.
3.8 Mutations have strong epistatic dependence on their predecessors

Due to differences between the lysate and FACS screening techniques, the selection pressures in HT-SAS are not identical to those in DE (Chapter B 6.2).

Despite the lower sensitivity of the HT-SAS analysis at the single residue level, some striking correlations were evident (Figure C-12): the H28L mutation, introduced in TEV<sup>Ser</sup>IV, showed enrichment of alternative residues in the two subsequent rounds but is increasingly selected for retention in later rounds. The back mutation of W130C in TEV<sup>Ser</sup>V also coincided with its purge from the library. Finally the N174K of TEV<sup>Ser</sup>VIII showed enrichment (both N174K and N174Y) before it is selected in the lineage, after which further mutations were no longer enriched. The previous observation, that nearly-neutral
mutations can become adaptive for later evolution, appears to represent a more general property of the local fitness landscapes.

The real power of HT-SAS is not in analysing individual residues, but rather analysing trends in large sets of residues\textsuperscript{90}. Results in section 3.6 suggested that there are epistatic interactions between the mutations accumulated in the DE lineage in line with some previous work\textsuperscript{255}. To see if this was a general effect, I looked at whether the mutations accumulated in the DE lineage affected which residues subsequently showed adaptive mutations. The enrichment profiles (Figure C-11c) indicated that mutations at very few positions are universally enriched or universally purged throughout the DE lineage. The S151C nucleophile reversion was always enriched (average 9.8-fold, p=0.0004) and any other S151X mutation was selected against (average 43-fold, p<0.0001) (Figure C-11c & Figure C-12). Only 9 other residues (4\%) showed consistent mutation enrichment, on the surface (L98, T191), in the binding pocket (N177, S208, V209, V216) and in the unstructured C-terminus (V228, E230, L235). Similarly, only 29 (12\%) positions are so constrained that they were consistently purged of any variation, mostly buried, hydrophobes (Figure C-13). The mutations accumulated in the DE lineage, therefore, had large effects on the adaptiveness of other possible mutations in the protein. At each round, the set of mutations that are adaptive appeared to change. The shifting
spectrum of adaptive mutations is consistent with a rugged fitness landscape in which the mutations accepted along the DE trajectory drastically influence the possible uphill routes by epistasis.

![Figure C-13](image)

Figure C-13 | Locations of beneficial or deleterious mutations
(a) Structure of TEV_{Cys} (PDB 1lvm) showing enrichment of non-wt amino acids for each residue averaged over whole TEV_{Ser}→TEV_{Ser\x} lineage. Thick, blue regions indicate positive enrichment of non-wt amino acids, thin red region indicate purging of non-wt amino acids. (b) Structure of TEV (PDB code 1lvm) overlaid with 29 consistently purged mutations (red), 6 consistently enriched mutations (blue). Also shown are catalytic triad (red sticks) and N-terminal product (black sticks).

3.9 Adaptive Mutations are Common at All Steps of the Trajectory

Epistasis is usually considered a constraining factor on evolution as the lack of a smooth landscape makes it harder for evolution to access fitness peaks\(^{172}\). Despite this, at all steps along the adaptive trajectory there were residues (average 22% of 236) that display beneficial mutations (Figure C-14), indicating multiple uphill pathways in the flanking fitness landscape.

![Figure C-14](image)

Figure C-14 | Adaptive potential along the TEV_{Ser}→TEV_{Ser\x} lineage
Proportion of residues that show >2-fold enrichment for mutation in each library along evolutionary lineage.
That many alternatives to improve activity exist is consistent with the rapid evolutionary compensation of the mutated nucleophile handicap. It indicates that the abundance of adaptive mutations was sufficient to overcome the small number of variants screened at each round of DE.

3.10 SOME PROTEIN REGIONS ARE HOTSPOTS FOR ADAPTIVE MUTATIONS

The mutations accumulated in the DE lineage were overrepresented in the triad second shell. However, the available adaptive mutations vary from round to round. To resolve this apparent conflict, I looked at whether different regions of the protein were more or less likely to contain adaptive mutations.

As stated previously, the 12% of positions that were consistently constrained to the wild type residue are mostly located in the protein core. In fact, mutations in buried residues tended to be more deleterious on average ($p=0.01$ Figure C-15), supporting the previous observations of a trade-off between activity and solubility (Figure C-7).

Additionally, adaptive mutations are disproportionately located in the active site periphery (triad second shell). Firstly beneficial mutations were more common in this region (average 28% of 53, $p<0.0001$, Figure C-16a). Secondly, the average enrichment of triad second shell
Chapter C - Handicap-recover evolution to accommodate a new nucleophile

Mutations was higher than for the protein as a whole \( (p=0.04, \text{Figure C-16b} \& \text{Figure C-13b}) \). The second shell encompasses some core residues which, if removed from the analysis, enhance this trend \( (p=0.01) \). Conversely, residues contacting the nucleophile (nucleophile first shell) were strongly constrained showing few beneficial mutations \( \text{(average 16\% of 11, } p<0.0001 \text{)} \) and certainly more detrimental on average \( (p<0.0001) \). The triad’s first shell showed the largest fluctuations, indicative of the strong effects the interactions have on the core catalytic machinery. Finally, mutations in shells further away \( \text{(triad } \geq \text{third shell)} \) showed no more beneficial mutations than the protein on average \( (p=0.18) \).

Together, these results indicate that even though the set of adaptive residues available changes from round to round, there are some underlying trends in their locations.
Underlying trends in the adaptive potential of different regions of the enzyme explain the biased accumulation of mutations in this region during the adaptive DE trajectory (Figure C-8). Although few individual residues show consistent effect upon mutation, several regions of the protein do. Therefore, despite high epistasis, the active site second shell is repeatedly enriched in adaptive mutations. This is consistent with meta-analysis of directed evolution experiments which show that the largest catalytic changes are often caused by mutations 4-8Å from the catalytic centre\textsuperscript{254}.

Conversely, residues that showed improvement in activity using serine do not correlate to those found in natural serine proteases in the PA clan. In the PA clan, therefore, the adaptive differences between serine and cysteine proteases are buried by drift, consistent with their ancient divergence.
4 CONCLUSIONS

The handicap-recover directed evolution demonstrated that it is possible to convert TEV\textsubscript{Cys} into a functional serine protease TEV\textsubscript{SerX}. In addition to biochemically characterising this specific uphill trajectory (TEV\textsuperscript{Ser} \rightarrow TEV\textsuperscript{SerX}), I generate complementary data on the flanking fitness landscape by HT-SAS. These combined methods reveal how local fitness landscape topology guides the evolutionary trajectory.

4.1 EVOLUTIONARY CONSTRAINTS ON THE EVOLVED TRAJECTORY

The trajectory highlights several evolutionary constraints. The activity improvement negatively trades-off with structural integrity. The adaptive mutations accumulated along the evolutionary trajectory to TEV\textsuperscript{SerX} destabilise the structure and promote aggregation. Characterising local fitness landscapes throughout the trajectory reveals that buried residues rarely contain adaptive mutations. Likewise, the 12% of residues consistently purged of mutations are mostly in the protein core. These biophysical constraints bias evolution away from these regions of sequence space.

Previous work on active site evolution found two supporting mutations that enabled exchange of an active site residue in phosphoglycerate kinase\textsuperscript{255} but that their effects were contingent on their order of introduction. Of eight possible routes to the three mutation variant, only one did not pass through a fitness valley, suggesting that rare mutation combinations dominate and constrain the evolutionary trajectory. Similarly my results are consistent with high epistasis between mutations within the trajectory and between trajectory mutations and the spectrum of possible mutations. This creates a rugged fitness landscape where the acceptance of each mutation
drastically alters which subsequent mutations are adaptive. This ruggedness reduces repeatability of evolution such that replicate trajectories would widely deviate.

4.2 HIGH ADAPTIVE POTENTIAL GENERATES A VERSATILE ACTIVE SITE

Despite these constraints, the $10^4$-fold drop in activity was compensated for remarkably easily. In only 10 rounds, screening 350 variants per round, activity was recovered by $>10^3$ to improve the activity of TEV$^{\text{SerX}}$ to within an order of magnitude of the original TEV$^{\text{Cys}}$. The results indicate that the proportion of adaptive mutations is of the order of $10^{-1}$-$10^{-2}$, at the high end of previous evolvability measurements (Chapter A, 5.1) and therefore that extra rounds of evolution might further increase activity. Additionally, this increase does not trade-off with the ability to use the original nucleophile, cysteine, generating an enzyme that efficiently uses either. This was achieved through very few mutations compared to the sequence difference between natural serine and cysteine proteases. These mutations cluster around the active site periphery and is representative of an underlying trend for triad second shell mutations to be rich in adaptive potential.

Indeed analysis of the flanking fitness landscape shows that the ease of this evolution reflects the large number of possible adaptive pathways at each step. By HT-SAS, we show that there are many possible beneficial mutations available at each round. Hence, whilst routes to any given solution may be rare$^{255}$, the abundance of solutions results in high overall evolvability of nucleophile exchange. Thus the easily evolved enzyme TEV$^{\text{SerX}}$ is likely one of a great many similar solutions.
Chapter D - Neutral evolution and emergent robustness

1 SUMMARY

Populations of TEV protease genes were neutrally evolved in order to investigate how they spread through the local neutral network. Next generation sequencing revealed that evolution generated a number of highly robust variants whose descendants come to dominate the later populations due to their increased tolerance to mutations. These variants show improved structural robustness that correlates with a preferential distribution of mutations in exposed, disordered structural regions away from the active site that have minor impacts on protein stability. This shows how long-term neutral evolution contains implicit selection for robustness as genes diffuse through the local neutral network to extinction-resistant regions with more neutral neighbours.

Parts of this chapter and Chapter E have been combined and prepared for publication:

Shafee, T., Gatti-Lafraconi, P., Minter, R., Hollfelder, F.
Neutral evolution – two combined mechanisms generate robust and evolvable enzymes. MBE. (2013)
INTRODUCTION

2.1 PROTEIN ROBUSTNESS AND TOLERANCE TO MUTATIONS

As introduced in more detail in Chapter A, 4.1, proteins achieve stability through an internal network of interactions. Measurements of protein tolerance to random amino acid substitution ($\tau_{\text{sub}}$) have found an average $\tau_{\text{sub}}=0.65\pm0.02^{50,79,144,256}$. This indicates that only a third of random mutations destroy function in proteins so far studied.

In this robustness there is an apparent contradiction: proteins are apparently stable enough to tolerate most mutations, yet they are only marginally stable before mutation, with $\Delta G_{\text{folding}}$ of around $-10$ kcal/mol$^{257-259}$. This marginal stability has been explained by both adaptive and neutral mechanisms (Figure D-1). Firstly by selection to avoid accumulation of aggregates and be protease resistant$^{139,259}$. Secondly there is evidence that excess stability can reduce the flexibility necessary for protein activity$^{14,17,18,61}$. Finally excess stability above what can be selected for can be eroded by genetic drift$^{8,260}$ since mutations are more likely to reduce than increase stability.

![Figure D-1 | Protein stability effects on fitness](image)
A deeper understanding of the evolution of robustness may be able to resolve the apparent paradox that proteins tolerate mutations surprisingly well despite being perpetually on the edge of unfolding even before those mutations.

2.2 Theoretical predictions of robustness evolution

Mathematical modelling in silico has been used to investigate neutral network structure and its evolutionary consequences and predicts that neutral networks of equivalent proteins are not homogeneous\textsuperscript{133,149,150,261–265}. Some sequences in the network are more highly connected (have more neutral neighbours) and form central nodes in the network, whilst others form sparsely connected edge regions. Both abstract mathematics\textsuperscript{149,265} and explicit models\textsuperscript{133,263,264} predict that large, neutrally-evolving populations can evolve towards regions of the network with more neutral neighbours, i.e. with increased robustness. This can be understood in terms of the disproportionate extinction of variants that, though perfectly fit, have fewer neutral neighbours and so are more likely to be deactivated by mutation. As a counterpoint, models of lethal mutagenesis therapy in viruses indicate that, although robustness can increase as an adaptation to high mutation rates\textsuperscript{266}, it may not be high enough to prevent extinction\textsuperscript{267}.

Classical population genetics also makes a number of predictions on the conditions under which robustness could evolve. For example, population size (N=total number of variants) and effective population size (N\textsubscript{e}=total number of breeding variants) have a strong effect on the balance of selection versus genetic drift in an evolving population. Small populations (low N) have a smaller pool of variation, important when beneficial mutations are rare and the rate limiting step of evolution is the generation of adaptive variation, not its fixation by
selection. In small effective populations (low $N_e$), the stochastic effect of random chance has a larger impact on the differential survival of mutants and can overwhelm selection. Consequently, the small selection coefficients hypothesised to be important in the evolution of robustness likely require large populations\textsuperscript{149}.

2.3 Empirical Evidence for the Evolution of Robustness

Although the theoretical foundations of the evolution of robustness have been in place for some time, experimental and empirical evidence is currently limited.

Bioinformatic analysis indicates that eukaryote miRNA hairpins are more robust to mutations (Figure D-2a) than random sequences which fold into the same secondary structure\textsuperscript{268} (Figure D-2b). This indicates that natural RNA structures have evolved to be tolerant to mutations.

Additionally, experimental evolution of an RNA virus (phage $\phi 6$) has looked at the erosion of robustness under reduced selection pressure. Populations were either evolved under co-infection conditions (when complementation is expected to buffer deleterious mutations), or under single-infection conditions. Populations evolved with high co-infection were sensitive to mutagenesis which caused greater
reductions in fitness than mutation of the control populations (Figure D-3). This was interpreted to suggest that buffering by co-infection weakened selection for robustness.

Two experimental protein evolution works have addressed how the accumulation of neutral mutations can increase robustness during neutral evolution.\textsuperscript{187,188} This process is sometimes called ‘neutral drift’ although I will avoid this term for now, since it implies that the process is completely non-directional which is not necessarily the case.

Firstly, an experiment on P450 oxidase confirmed that robustness can be generated as an emergent property of a neutrally evolving population.\textsuperscript{187} A lysate activity endpoint screening system was used to assay either 1 or 435 variants and sort all variants with >75% activity. It was found that when only 1 variant was selected from round to round the mutation tolerance fell from 48% to 39% (Figure D-4).

The larger evolving population became highly polymorphomic with few mutations shared in multiple variants. In this case, the tolerance to mutations increased very slightly from 48% to 50% and the enzymes were found to be more thermostable and resistant to urea degradation.
Figure D-4 | Mutation tolerance over neutral evolution of P450 oxygenase

The fraction of variants that retain function after mutagenesis of $\mu=0.3$ non-synonymous changes per codon i.e. roughly one mutation per 460 a.a gene. The $N=1$ lineage (white) had a single mutant screened at each round, if it retained >75% activity, it was selected for the next. The $N=10^2$ lineage (grey) had 435 variants screened at each round and all variants retaining >75% activity were selected for the next. (Data from reference\textsuperscript{187})

Secondly, an experiment on $\beta$-lactamase concentrated on altering the strength of selection pressure\textsuperscript{188}. A cell survival selection system was used to evolve the gene on either 250 or 12.5 µg/ml of ampicillin. At each round $>10^6$ variants were selected, however tolerance to mutations was not quantified and merely reported to be in general $>50\%$. Several mutations were enriched (Figure D-5a) which were predicted to be stabilising. Some of these mutations were characterised and indeed increased thermostability (Figure D-5b) and some were able to compensate for an introduced destabilising mutation.

Figure D-5 | Enriched mutations after neutral evolution of TEM1 $\beta$-lactamase

(a) After neutral evolution for 18 rounds, the most enriched mutations ($\chi^2 p<0.05$) are shown here. Selection on 0 µg/ml (black bars), 12.5 µg/ml (blue bars), 250 µg/ml (red bars), * indicates variants chosen for biophysical characterisation. (b) Thermal denaturation temperatures of the chosen variants are higher than the wild type TEM starting enzyme. ((a) Adapted from reference\textsuperscript{188} and (b) data from reference\textsuperscript{188})
Together, these two evolution experiments support the core theoretical prediction that mutations which increase robustness can be accumulated in neutral evolution.

In my work I have added to this body of results by providing evidence of how robust variants appear, some of their biochemical and biophysical properties, and how their descendants take over the population. Using a high throughput screening system I was able to generate disaggregated data on the distribution of activities after mutation (as with the P450 evolution), in addition to assaying large populations (as with the TEM1 evolution). This has enabled observation of the accumulation of mutation diversity in greater breadth and detail than previously possible. Use of phylogenetic methods and ancestor reconstruction allowed confirmation of the hypothesis that the ancestors of late round populations genuinely were more robust and mutation tolerant than the wild type starting point.
3 RESULTS AND DISCUSSION

For this chapter, Tobacco Etch Virus nuclear inclusion cysteine protease will be referred to as wtTEV and derived populations will be denoted by the number of rounds of evolution.

In Chapter C, I performed DE via small libraries using the lysate assay developed in Chapter B, 5.2. Local fitness landscapes flanking the trajectory were assessed by HT-SAS using a higher throughput, flow cytometric assay, described in Chapter B, 5.4. In this chapter, I have used the flow cytometric assay to perform DE by screening large libraries and selecting a pool of variants at each round that maintain wild type function.

3.1 THE DISTRIBUTION OF FITNESS EFFECTS OF TEV MUTANTS

The wtTEV gene was randomly mutagenised by epPCR, the library of mutants was cloned into pMAA and co-expressed with the C-Y(tev) substrate in E. coli. When assayed by FACS, the distribution of phenotypes was bimodal. This corresponds to a DFE where mutant enzymes either have activity similar to wild-type TEV, or are inactive (Figure D-6a). This agrees with previous findings that mutations in whole organisms\textsuperscript{153,154} and individual enzymes\textsuperscript{79} often show a bimodal distribution. The rate at which mutations were introduced did not alter the bimodality of the distribution, but affected the relative ratio between the two sub-populations.
Mutation tolerance of an enzyme or enzyme population can be calculated by the relative abundance of wt-like enzymes after error prone PCR (epPCR). Since epPCR generates some variants with 1 mutation, some 2, some 3 etc. the mutation range follows a Gaussian distribution. This spread can be used to calculate what proportion of random mutations are neutral ($\tau_{sub}$) using \textbf{Equation D-1} (adapted from reference\textsuperscript{144}). The equation makes a number of assumptions: Mutations are either neutral or lethal, and their effects are additive (i.e. two neutral mutations are always neutral together).

$$S = \sum_{n=0}^{\infty} f_n (\tau_{sub})^n = f_0(\tau_{sub})^0 + f_1(\tau_{sub})^1 + f_2(\tau_{sub})^2 + \cdots + f_n(\tau_{sub})^n$$

\textbf{Equation D-1} | The probability of enzyme activity retention after random mutation

- $S$ = the fraction of enzyme variants displaying wt-like activity
- $n$ = number of mutations
- $f_n$ = the fraction of enzyme variants with $n$ mutations
- $\tau_{sub}$ = the probability that a random mutation will be neutral

Nb. Frameshift mutations will cause underestimation $\tau_{sub}$, however, these were rare enough that they made no difference to these estimates.
When mutations were introduced at either $\mu=0.5\%$ or $\mu=1.5\%$ non-synonymous changes per codon ($\approx 1$ or $3$ amino acid mutations per enzyme) in wtTEV, the probability of activity retention after a random amino acid substitution ($\tau_{\text{sub}}$) was 0.78 and 0.74, respectively. This is higher than reported previously for other proteins (average $\tau_{\text{sub}}=0.65\pm0.02$)\textsuperscript{50,79,144,256} meaning that in wtTEV only a quarter of random amino acid changes destroy activity. This may be because viral proteins are hypothesised to have higher structural robustness\textsuperscript{269} to compensate for the lack of genomic robustness, high mutation rates and strong selection pressures inherent to the viral life cycle\textsuperscript{266,270,271}.

I attempted to investigate the underlying protein properties by purifying 92 members of the $\mu=1.5\%$ mutant library. Poor expression of the protease lead to a high degree of noise in the data. However, some broad observations could be made. The underlying distribution of purified enzyme specific activities appeared less defined (Figure D-6b) and the range of expression showed no bimodality (Figure D-6c). The lack of obvious bimodality in the activity of purified enzyme variants may indicate that the bimodal distribution of phenotypes was partly the effect of the ‘end-point’ style phenotype assay. This is consistent with the observation that changes in enzyme $k_{\text{cat}}/K_M$ have non-linear effects on pathway flux\textsuperscript{130} and fitness\textsuperscript{272}. There also appeared to be a negative trade-off between activity and soluble expression in the library (Figure D-6d) in line with results in Chapter C.

3.2 Generating neutral evolution lineages

To investigate how mutation tolerance can change during evolution, I performed several neutral evolution experiments. For each generation, a population of $10^6$ mutant enzymes was expressed, assayed, and a homogeneous sub-population of functionally neutral mutants isolated by FACS (with $>99\%$ of selected clones retaining activity on the native
cleavage sequence). In this way, mutations that were neutral were accepted for the next round of evolution. Better-than-wt cleavage was not preferentially selected, and so mutations that increase activity were not preferentially selected.

Population size was controlled at $N=10^6$ by screening $10^6$ variants at each round. Similarly I controlled effective population size at $2 \times 10^4$ by limiting the number of neutral variants selected for the next round of mutagenesis. Parallel neutral evolution lineages compared the effect of mutation rate ($\mu=0.5\%$ or $\mu=1.5\%$ non-synonymous changes per codon per round). As mentioned in Chapter B, 5.2, mutations were evenly distributed through the gene with very few frameshifts. Keeping population and effective population sizes, mutation rate and selection pressure constant ensured that changes seen were due to the duration of neutral evolution, not changing conditions.

Since not all mutations are neutral, they were accumulated at a slightly lower rate than they were introduced (i.e. some mutations are deleterious and purged from the evolving populations). The $N_e=10^4$ lineages accumulated mutations at roughly linear rates of 0.4\% and 0.7\% amino acid variation per round (at $\mu=0.5\%$ and 1.5\% respectively Figure D-7a) although the rate lowered after round 7. These rates were slightly higher than in the equivalent $N_e=10^2$ lineages (0.3\% and 0.7\% Figure D-7b).

![Figure D-7](image-url)
Initially, as the populations accumulate mutations, their tolerance to further mutations (defined as $\tau_{0.5\%}$ and $\tau_{1.5\%}$ for $\mu=0.5\%$ and $\mu=1.5\%$, respectively) decreased (Figure D-8). By the time populations had accumulated $\approx2\%$ sequence diversity, mutation tolerance fell from $\tau_{0.5\%}=0.60$ to 0.08 and from $\tau_{1.5\%}=0.28$ to 0.06. This indicates that these populations contain variants that, though neutral in function, are less robust. This would be expected by a model in which proteins have stability in excess of some threshold$^{134,273}$. Mutations, on average, erode this excess stability such that a greater proportion of the population sits on the stability threshold and is affected by mutations (less robust)$^{91,134}$.

From this low point in mutation tolerance, the decline halted. In the $\mu=0.5\%$ lineage, tolerance initially fluctuated and then recovered to near wild type levels by later rounds. The trend in the $\mu=1.5\%$ lineage was more pronounced and showed a general increase in the tolerance to mutations that begins at round 3 and reached $\tau_{1.5\%}>0.50$ at rounds 9-11.

Consistent with the threshold hypothesis discussed above, once the population reaches a critical stability level, selective pressure to improve robustness sets in. In the case of wtTEV, this leads to the
recovery and stabilisation of mutation tolerance with efficiency that is relative to the rate at which mutations are introduced. The stronger increase in mutation tolerance in the \( \mu=1.5\% \) lineage is consistent with a greater selection pressure for robustness under higher mutation rate. It may also explain why this tolerance increase is greater than that seen in previous work on P450 oxidase\(^{187} \) (Figure D-4).

### 3.3 Enriched Mutations Confer Population Robustness

In order to investigate which mutations were accumulating in the evolving lineages, the populations of TEV variant genes were deep sequenced (1200±500 sequences per population). In early rounds, most residues accumulated diversity at rates at or below the mutation rate: increased diversity at individual positions in the sequence indicates the accumulation of neutral mutations is favoured.

Conversely, positions where non-wt mutations were purged indicates that those mutations are deleterious and under negative selection. In both lineages, after diversity accumulated to 2\% (5 rounds at \( \mu=0.5\% \) or 3 rounds at \( \mu=1.5\% \) (Figure D-9a,g), previously neutral mutations started to be strongly enriched, indicative of positive selection. Residues in shells further from the active site triad accumulated more mutations (Figure D-9b,h). This is in sharp contrast to results in Chapter C, 3.10 where mutations in the triad second shell were enriched in variants more active with an alternative nucleophile.

Additionally, residues predicted by FoldX to be minimally destabilised by alanine mutation (indicating their importance in stable folding) accumulate fewer mutations (Figure D-9c,i). Later rounds also showed a bias towards mutation accumulation outside of the enzyme core and in disordered regions (Figure D-9d,j).
Figure D-9 | Mutation accumulation by structural region (N=10^4 lineages)
(a) Changes in frequency of non-wt residues in the µ=0.5% lineage plotted by sequence.
(b) Mutation accumulation separated by triad interaction shell with residues <4Å from the catalytic triad (light red), 4-8Å from the catalytic triad (red) and >8Å from the catalytic triad (dark red).
(c) Mutation accumulation separated by alanine mutation sensitivity (FoldX prediction) with residues predicted to give >3kJmol⁻¹ destabilisation (dark blue) and <1kJmol⁻¹ destabilisation (light blue).
(d) Mutation accumulation separated by disorder prediction (IUPred) with residues with disorder prediction >0.6 (light purple) and <0.2 (dark purple).
(e) Mutation accumulation separated by residues with sequence conservation within the viral C4 proteases of >0.9 (green) and all residues (black).
(f) Round 11 non-wt residue frequency mapped onto structure (residues purged of variation in red, enriched in blue).
(g-l) Same data for µ=1.5%.
However, there was no correlation between the accumulated mutations and sequence conservation in an alignment of naturally occurring viral proteases (Figure D-9e,k). This disagrees with previous experiments that suggested that back-to-ancestor mutations were key to the emergent robustness in an experimentally evolving population\textsuperscript{188,274}. These trends could also be seen when mutation accumulation was plotted on the structure as mutations were purged in the core, whereas the surface contained neutrally accepted and actively enriched mutations (Figure D-9f,l).

Although neutral to the selected phenotype, mutations did affect the underlying properties that give rise to that phenotype. To analyse this in more detail, I focused on the $\mu=1.5\%$, $N_e=10^4$ lineage. The average $k_{cat}/K_M$ activity of the population fell to 16\% of wtTEV by round 3 after which it was never above 30\% (Figure D-10). Average soluble expression of the populations was always lower than wtTEV (Figure D-11a) and showed increasing dependence on the MBP fusion partner (Figure D-11c).

![Figure D-10](image_url)

**Figure D-10 | Population average activity ($\mu=1.5\%$, $N_e=10^4$ lineage)**

Averaged Activity ($k_{cat}/K_M$) of populations in the $\mu=1.5\%$ lineage. The whole population of enzyme variants was expressed and purified. Activity assays of the purified mixed enzyme reflects average activity.

This indicates that (at least at the whole-population level) selection for increased activity or concentration of active enzyme has not occurred and in fact both properties have fallen slightly, likely reflective of permissive selection for neutral function. A similar phenomenon has
been seen in other neutral evolution experiments where the activity of populations falls slightly to match the selection threshold\textsuperscript{187,188}.

While the soluble expression of the evolving populations was always less than the wtTEV starting level (Figure D-11a), their soluble expression after mutagenesis is higher than wtTEV in later rounds (Figure D-11b). This indicates that later populations retain soluble expression after mutagenesis better than wtTEV. Therefore, despite the fall in average soluble expression in later populations, these suffered a smaller loss in solubility upon mutation. This suggested that the enriched mutations (Figure D-9a) were likely selected for their effect on improving enzyme robustness that increased mutation tolerance and extinction-resistance (Figure D-8).

Figure D-11 | Population average soluble expression ($\mu=1.5\%$, $N_e=10^4$ lineage)  
Population average soluble expression for (a) evolved populations, (normalised to wtTEV) (b) evolved populations after mutagenesis (normalised to wtTEV) (c) evolved populations after mutagenesis without MBP fusion (normalised to wtTEV after mutagenesis without sMBP fusion).
3.4 LATER POPULATIONS ARE DESCENDED FROM ROBUST NODE SEQUENCES

High-throughput sequencing allowed phylogenies of the evolving lineages to be generated. A standard phylogeny only shows the sequences of contemporary, extant sequences (in this case, the final round of the experimental evolution). Since lineages in this work were sequenced throughout evolution, it is possible to use these ‘fossil’ DNA sequences to view evolutionary dynamics on a phylogeny.

When viewed as a phylogeny it can be seen that early populations have wtTEV as last common ancestor (Figure D-12) in agreement with mutations being accumulated across the gene. Later populations, however, are descended from non-wt nodes, consistent with the enrichment of mutation cohorts in Figure D-9. The co-enrichment of mutation cohorts is common in situations of high mutation rate and strong selection\textsuperscript{275}.

The phylogeny shows that these mutation cohorts were enriched, not as the result of a single sequence becoming more common, but due to the descendants of a particular sequence being more common.

In the case of the $\mu=1.5\%$ lineage (Figure D-12b), node sequence TEV-\(\beta_1\) appears in round 3 and its descendants proliferate until they are outcompeted by descendants of TEV-\(\beta_2\) in rounds 9-11. A similar trend is seen for the $\mu=0.5\%$ lineage (Figure D-12a) but with TEV-\(\alpha_1\) and TEV-\(\alpha_2\) cofounding later populations, i.e. neither are extinct by round 11. These phylogenetic topologies suggests that the node sequences ($\alpha_1$, $\alpha_2$, $\beta_1$ & $\beta_2$) are more extinction-resistant than wtTEV and other members of their population.
The observation that these node sequences appeared to be more evolvable could be confirmed by reconstructing and characterising the node enzymes. Ancestral gene reconstruction is increasingly used to check predictions about the properties of ancient enzymes. Phylogenetic methods implicitly generate the most likely sequence of each node anyway and whole gene synthesis is becoming economical. The same technique is even more applicable to the phylogenies.

Figure D-12 | \( N_e=10^4 \) Lineage phylogenies
Phylogenetic trees of (a) the \( \mu=0.5\% \) lineage and (b) the \( \mu=1.5\% \) lineage. Small circles indicate 100 sequences from each of rounds 1,3,5,7,9,11 that were aligned and used to generate each phylogeny. Large circles indicate wtTEV (red) and reconstructed nodes (grey).
generated here. The deep sequencing of very similar genes gives a high
degree of certainty when inferring node sequences.

I therefore did this for the genes of four ‘ancestral’ node genes, two
from each of the $\mu=0.5\%$ and $\mu=1.5\%$ lineages (Table D-1 & Figure
D-13). The ancestral node sequences were inferred by the phylogenetic
trees of late round populations (7, 9 & 11). I was able to observe all
predicted node sequences in the sequencing results of earlier rounds,
confirming that they genuinely existed. Since the forward and reverse
halves of the gene were sequenced separately, the two half-reads were
stitched together based on tree topology. Finally, the correct pairing of
the two gene halves was confirmed by full length Sanger sequencing of
individual descendants from later rounds.

<table>
<thead>
<tr>
<th>NAME</th>
<th>FIRST SEEN IN ROUND</th>
<th>AMINO ACID MUTATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>7</td>
<td>P8Q, P13T, Q58R, N68T, F139L, S153N, R203G,</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>9</td>
<td>K6M, P8S, H61Q, F132S, D136E, F162L, T191R</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>3</td>
<td>Q58R, S123N, W202R</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>5</td>
<td>S15A, T29S, K67R, D148V, S200R, E223STOP</td>
</tr>
</tbody>
</table>

The reconstructed nodes each contained 9-10 nucleotide mutations
compared to wtTEV, only 3 to 7 of which were non-synonymous (Table
D-1). The only recurring mutation was Q58R independently evolved in α1 and β1 (using different codons). P8 was also mutated to polar residues in α1 and α2. The C-terminal truncation at residue 223 in β2 is interesting as a similar truncation was also observed in the TEV evolution experiment for increased activity using a mutant nucleophile reported in Chapter C. This convergence suggests that the C-terminal truncation in TEV^serVI may have been an adaptation to increase robustness in that system as well.

As hypothesised, the enzymes that these genes encode displayed higher tolerance to mutations (Figure D-14a). At low mutation rates ($\mu=0.5\%$) all reconstructed nodes showed $>90\%$ of variants that still retained wt-like activity upon mutation. Even at high mutation rates ($\mu=2.0\%$) which leave only $10\%$ of wtTEV variants active, reconstructed nodes displayed $>50\%$ active mutants. Using the $\mu=2.0\%$ mutation rate, I calculate $\tau_{\text{sub}}=0.89$ for β1, $\tau_{\text{sub}}=0.95$ for α1 and $\tau_{\text{sub}}=0.96$ for α2 and β2 (for wtTEV $\tau_{\text{sub}}=0.74$). This validates the hypothesis that these node sequences encode particularly mutation tolerant enzymes.

Indeed, the smaller improvement in mutation tolerance of β1 over wtTEV may explain why it was outcompeted by β2. Conversely, α1 and
α2 have almost the same mutation tolerance in agreement with the persistence of descendants of both variants in later rounds.

In order to assess the underlying biophysical and biochemical determinants of the observed robustness, I purified the enzymes for characterisation. All reconstructed nodes had reduced $k_{\text{cat}}/K_M$ relative to wtTEV (Figure D-14b) in line with the generally reduced population average activity (Figure D-10). However, they showed increased soluble expression levels (Figure D-14c). The variants with the highest mutation tolerance (α2 and β2) showed both the highest activity and soluble expression among reconstructed nodes. This agrees with the high retention of solubility after mutation (Figure D-11b) in the populations descended from these nodes.

Mutations frequently reduce the stability of enzymes\textsuperscript{169,278} and it is known that wtTEV is poorly soluble and prone to aggregation\textsuperscript{216–218}. Together, these results imply that structure robustness, particularly resistance to aggregation, is a key aspect of the extreme mutation tolerance observed in the reconstructed enzymes.

3.5 Emergence of robustness requires large effective population size

By way of comparison, analogous experiments were performed in which a smaller effective population of only $2 \times 10^2$ variants ($N_e=10^2$) were selected at each round. As mutation rates ($\mu=0.5\%$ or $1.5\%$) and screened library size ($N=10^6$) were kept constant, these parallel lineages report on the effects of bottlenecks in the effective population size only. In this way it is possible to disentangle the two effects of a smaller variation pool ($N$ kept constant) and the effect of drift ($N_e$ varied by two orders of magnitude) and address the theoretical predictions introduced in section 2.2.
Under these conditions, neither of the lineages recovered tolerance (Figure D-15). At \(\mu=0.5\%\), tolerance remained constantly below \(\tau_{1.5}<0.15\). When \(\mu=1.5\%\), tolerance only transiently increased, peaking at \(\tau_{1.5}=0.75\) in round 8 (\(\approx 4\%\) amino acid variation) but quickly fell to \(\tau_{1.5}=0.00\) causing lineage extinction. This contrasts with the results of the \(N_e=10^4\) lineages (Figure D-8) and indicates that restriction in the number of variants used leads to, in the best scenario, stochastic variations in robustness and does not allow the fixation extinction-resistant populations.

Figure D-15 | Mutation tolerance \((N_e=10^2\) lineages\)

Change in mutational tolerance over evolution at mutation rates \(\mu=0.5\%\) (white circles) or \(\mu=1.5\%\) (grey circles) non synonymous mutations per codon per round in lineages with \(N_e=10^2\). Tolerance measured as the proportion of variants retaining wt-like activity after mutagenesis (i.e. \(\tau_{0.5}\%\) for \(\mu=0.5\%\) of codons per gene and \(\tau_{1.5}\%\) for \(\mu=1.5\%\)).
The restricted effective population size \((N_e=10^2)\) had effects on the efficacy of selection for both populations. The lack of highly robust variants generated in the \(\mu=0.5\%\) lineage may be due to the low accepted mutation rate (total amino acid diversity is never >2\%) and therefore insufficient sequences were sampled to find effective, robust variants. Alternatively it may be due to the stochastic purging of otherwise robust variants by genetic drift in the small effective populations.

Figure D-16 | Mutation accumulation by structural region \((N_e=10^2\) lineages)
Mutations were tolerated in regions similar to those seen in in the $N_c=10^4$ lineages. These regions were away from the catalytic triad (Figure D-16b), in residues predicted to be only marginally destabilised by alanine mutation (FoldX) (Figure D-16c), and in disordered regions (Figure D-16d). However, for the $\mu=1.5\%$ lineage, the trend was weak and mutations were accumulated more evenly through the structure (Figure D-16h-l). In particular mutations close to the triad were not purged as they were in other lineages. It is tempting to speculate that the lack of confinement of mutations to certain structural regions was the cause of the extinction of this lineage.

I also constructed phylogenies for the $N_c=10^2$ lineages. The phylogeny indicates that the $\mu=0.5\%$ lineage is dominated by the descendants of a variant (TEV-α3) that contains a single mutation V199D (Figure D-17a & Table D-2). This did not particularly increase the population’s mutation tolerance (Figure D-15 Error! Reference source not found.) and therefore is likely a chance fixation by genetic drift.

Conversely, the phylogeny of the $\mu=1.5\%$ lineage shows dominance of round 7 by descendants of TEV-β3, a variant with 3 amino acid substitutions (Table D-2b). Later rounds are dominated by the descendants of TEV-β4, a direct descendant of β3 with 9 extra substitutions and a truncation. The truncation is in fact a convergence with the truncation observed in β2. Radiation from β4 is limited in the final rounds due to the extinction of the whole lineage when no variants remained active after epPCR.
Therefore the phylogeny topologies for the $N_e=10^2$ lineages reflects their failure to evolve increased robustness, in line with predictions that large populations are required. Since the number of mutants screened at each round was the same ($N=10^6$), this is likely not due to a smaller pool of variation, but due to the greater stochastic effects of drift in agreement with the reduced purging of mutations in the active site and regions vital to structure (particularly when $\mu=1.5\%$). It is worth noting, however, that previous work performed on a P450 oxidase compared $N_e=1$ versus $N_e=10^2$ found that $10^2$ was sufficient for the evolution of robustness$^{187}$. Theoretical work suggest that $10^3$ is roughly the threshold where emergent robustness is expected in populations of these mutation rates$^{149}$. 
Table D-2 | Reconstructed node summary (N_e=10^2 lineages)
List of the mutations in each reconstructed node sequences relative to wtTEV

<table>
<thead>
<tr>
<th>NAME</th>
<th>FIRST SEEN IN ROUND</th>
<th>AMINO ACID MUTATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>α3</td>
<td>9</td>
<td>V199D</td>
</tr>
<tr>
<td>β3</td>
<td>3</td>
<td>D78E, S120N, D127E</td>
</tr>
</tbody>
</table>

Figure D-17 | N_e=10^2 lineage phylogenies and node mutation locations
Phylogenetic trees of (a) the μ=0.5% lineage and (b) the μ=1.5% lineage. Small circles indicate 100 sequences from each of rounds 1, 3, 5, 7, 9, 11 that were aligned and used to generate each phylogeny. Large circles indicate wtTEV (red) and reconstructed nodes (grey). Amino acid mutation locations in α3, β3 and β4 (c-e) mapped onto the structure of wtTEV (PDB 1lvm). Substrate in black, catalytic triad in red, mutations in grey.
4 CONCLUSIONS

By coupling high-throughput experimental evolution and deep sequencing I was able to reconstruct a complex sequence of evolutionary events. The results presented here indicate that neutrally evolving populations of enzymes can find robust enzyme variants in the local neutral network around wtTEV. These robust variants have higher mutation tolerance and hence their descendants are overrepresented in later populations. In this way robustness can evolve by simply selecting for preservation of wild-type function in the face of mutations. The robustness of natural enzyme folds may, in part, be an emergent property of such forces. Additionally, these results also indicate that mimicking neutral processes can be used as an enzyme engineering method to accumulate mutations to increase robustness and so manipulate evolvability.

wtTEV already has higher mutation tolerance ($\tau_{\text{sub}}=0.76$) than previously measured proteins$^{50,79,144,256}$ (average $\tau_{\text{sub}}=0.65\pm0.02$), possibly an adaptation to the high mutation rate of viral genomes$^{267,269,279}$. Viruses are known to have lower tolerance to random nucleotide mutations than cellular organisms$^{154}$. In the absence of robustness mechanisms such as genetic redundancy and alternative metabolic pathways, higher protein structure robustness may be an adaptation to achieve viable mutation tolerance.

Deep-sequencing allowed me to follow the populations’ evolution in molecular detail. This showed how mutations were purged from enzyme regions essential for its function$^{254}$ (close to the catalytic triad) and its stability (those predicted to have high order propensity and to be key to fold stability). It also revealed how some mutations were enriched above the background rate of mutagenesis, indicating positive selection.
In both high and low mutation rate lineages, selection for neutrality lead to the erosion of mutation tolerance until amino acid diversity of \( \approx 2\% \) accumulated. Once this point was reached most variants, although functional, were sensitive to mutation (average \( \tau_{0.5\%}=0.08 \) and \( \tau_{1.5\%}=0.06 \)) and those variants eventually went extinct. However, when the effective population size is large \( (N_e=10^4) \), populations contain a small number of variants with increased robustness.

Deep-sequencing also allowed the phylogenies of the evolving lineages to be built. Using the lineage phylogenies, I reconstructed four node enzymes. Only a single mutation convergently evolved in two, highlighting the diverse set of solutions to the problem of mutation tolerance. Despite the lack of common mutations, there are common properties. All have reduced \( k_{cat}/K_M \) activities but increased soluble expression compared to wtTEV. Additionally their descendants retain solubility after mutation better than wtTEV. The wtTEV enzyme is known to be highly insoluble and this resistance to aggregation both before and after mutation indicates greater protein structure robustness.

This structural robustness drastically increases mutation tolerance of the reconstructed enzymes. Three of these \((\alpha_1, \alpha_2, \beta_2)\) have tolerance to random amino acid mutation of \( \tau_{sub}>0.95 \) (compared to \( \tau_{sub}=0.76 \) for wtTEV and \( \tau_{sub} \approx 0.66 \) for an average protein \(^{144}\)). This mutation tolerance indicates that they have a high proportion of neutral neighbours and are therefore in a more interconnected region of the neutral network. Though these robust variants exist in the population only transiently, the high retention of fitness by their mutant descendants causes disproportionate neutral sequence radiation. This sequence radiation demonstrates how second order selection for robustness can result from direct selection for maintenance of wild type activity.
Chapter E - Adaptive evolvability and changing cleavage specificity

1 SUMMARY

Building on the results of Chapter D, I demonstrate how the accumulation of mutations neutral to activity on the native substrate can increase evolvability for sequence specificity changes. During neutral evolution, the frequency of fortuitous adaptive mutations fluctuated stochastically, with pronounced effects on the ‘immediate evolvability’ in response to selection. Additionally, later, robust populations had increased ‘long-term evolvability’ enabling higher activity after several rounds of adaptive evolution. In addition to shedding light on the evolution of evolvability, these results suggest ways of manipulating the sequence specificity of proteases for therapeutic and biotechnological applications.

Parts of this chapter and Chapter D have been combined and prepared for publication:

Shafee, T., Gatti-Lafranconi, P., Minter, R., Hollfelder, F.
Neutral evolution – two combined mechanisms generate robust and evolvable enzymes. MBE. (2013)
2 INTRODUCTION

2.1 EVOLVABILITY OF ENZYMES

Enzyme evolvability is the propensity of an enzyme (or population of enzyme variants) to acquire new functions by mutation and selection. The generation and improvement of, promiscuous functions provides a starting source of adaptive variation for evolution. In experimental measurements of adaptive evolvability of enzymes, several outcomes affect the response to a novel selection pressure. Even if the whole population does not display a promiscuous activity, some individual variants may be adapted, or have high potential for adaptive evolution. I will therefore separate ‘immediate’ and ‘long-term’ evolvability.

**Immediate** evolvability is how well a population can adapt by selection from standing variation alone. It is dependent on the population size (number of variants), the number of those variants that are unique (heterogeneity), and whether any variants have new promiscuous activities that do not compromise the native activity. **Immediate** evolvability, therefore, dictates the maximum promiscuous activity that can be selected from that population without further mutations.

**Long-term** evolvability determines the maximum activity after several, iterative rounds of mutagenesis and selection. It partly depends on adaptive standing (as with immediate evolvability) variation but also on the frequency and magnitude of beneficial, *de novo* mutations. For example, robustness could increase evolvability if a stable scaffold can compensate for destabilising adaptive mutations in the active site.\textsuperscript{273,280–282} Experiments suggest that this is the case if the active site and scaffold are separated in the structure,\textsuperscript{72,273,283} allowing ‘global suppressor’ mutations.
2.2 **Neutral Mutations and Cryptic Variation**

Interest in the interplay between drift and selection started in the 1930s when the shifting-balance theory proposed that, in some situations, genetic drift could facilitate later adaptive evolution. Although this theory was largely discredited it drew attention to possibility that drift can generate cryptic variation that, though neutral to current function, may affect selection for new functions.

In terms of sequence space, current theories predict that if the neutral networks for two activities overlap, a neutrally evolving population may diffuse to regions of the neutral network of the first activity that allow it to access a second (Figure E-1). This would only be the case when the distance between activities is smaller than the distance that a neutrally evolving population can cover. The degree of inter-penetration of the two networks will determine how common cryptic variation for the promiscuous activity is in sequence space.

**Cryptic variation**

Genetic variation that introduces promiscuous activities to some members of a population

---

**Figure E-1 | Two overlapping neutral networks**

A large set of 'lattice model' protein genes generated in silico (top) contain two neutral networks of Structure A (blue) and Structure B (red). Each black dash represents a sequence, the lines connecting them represent mutations. When viewed as a fitness landscape (bottom), they form two overlapping peaks. It is possible for a neutrally evolving sequence to mutate to the edge of network A (a1→a2→a/b) and arrive at a promiscuous sequence that exists in both neutral networks. From there it is possible to evolve to new structure B specialist (a/b→b1). (Adapted from reference)
Evidence is accumulating that mutations neutral to native function can be rich in cryptic variation that though unselected, may affect immediate evolvability \(^{188,290,291}\). For example, accumulation of neutral mutations in P450s, caused up to 4-fold changes in six measured promiscuous activities\(^{290}\) (Table E-1). Similarly, in β-lactamase, neutrally evolved populations contained slightly more variants that conferred resistance to cefotaxime than the mutagenised wild type\(^{188}\). Conversely, experiments on immediate evolvability of β-glucuronidase indicate that pools of neutral mutants can be less rich in adaptive mutations than random mutants. In this case, the most adaptive mutations and mutation combinations trade-off with the original activity and so are not present in neutrally evolved populations\(^{26,53,292,293}\).

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>Rounds of neutral evolution</th>
<th>Target substrate</th>
<th>Most improved variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>P450 oxidase</td>
<td>15</td>
<td>propranolol</td>
<td>1.6 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11-phenoxyundecanoic acid</td>
<td>2.2 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-amino-5-chlorobenzoxazole</td>
<td>3.2 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,2-methylenedioxybenzene</td>
<td>2.6 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2- phenoxyethanol</td>
<td>4.1 - fold</td>
</tr>
<tr>
<td>β-glucuronidase</td>
<td>4</td>
<td>β-galactose</td>
<td>3.9 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>paraxoxon</td>
<td>2.2 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-naphtylacetate</td>
<td>5.0 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dihydrocoumarin</td>
<td>1.8 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0-acetoxy-7-hydroxycomarin</td>
<td>2.1 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-thiobutyl butyrolactone</td>
<td>1.9 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7-O-diethylphosphoryl-3-cyano-7-hydroxycomarin</td>
<td>2.4 - fold</td>
</tr>
<tr>
<td>Paraoxonase 1</td>
<td>1-3 mixed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>************</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FREQUENCY INCREASE</td>
</tr>
<tr>
<td>β-lactamase</td>
<td>18</td>
<td>0.5 µM cefotaxime</td>
<td>1.7 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8 µM cefotaxime</td>
<td>1.8 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0 µM cefotaxime</td>
<td>1.1 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.0 µM cefotaxime</td>
<td>1.1 - fold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.0 µM cefotaxime</td>
<td>*</td>
</tr>
</tbody>
</table>

Table E-1 | Previously measured immediate evolvabilities
Improvement in a promiscuous activity in the best variant isolated from neutrally evolved populations. Fold improvements calculated from reported \(k_{cat}/K_M\) activities for P450 oxidase and β-glucuronidase, calculated from ‘specific activity’ for paraoxonase 1. Since activities of individual variants were not reported for β-lactamase, the frequency of variants surviving a particular antibiotic concentration (normalised to an error prone PCR library) was used. * indicates that the wt enzyme library had no surviving colonies, the neutrally evolved population had 1.5% survival. (Data from references\(^{188,290,291,293}\))
Observations that neutral evolution can increase enzyme robustness\textsuperscript{187} (see Chapter D) indicate that long-term evolvability might also be affected as more mutation-tolerant enzymes may better tolerate the introduction of adaptive, but destabilising mutations\textsuperscript{180,273,281}. Long-term evolvability has been described for organisms including bacteria\textsuperscript{186} and algae\textsuperscript{294} but comparative studies of protein long-term evolvability are largely theoretical\textsuperscript{72,295,296}. Substrate specificity is therefore a good model to study evolvability, although available studies have not, so far, addressed immediate and long term evolvability independently.

2.3 Protease Specificity Engineering

Directed evolution has proven to be a successful method for altering protease specificity\textsuperscript{297}. In the last two years, three publications on TEV protease have shown changes in specificity for the residue before\textsuperscript{224} and after\textsuperscript{114,214} the cleaved peptide bond at the P1 and P1’ positions (Table E-2).

A cell survival selection system was created based on cleavage of a modular transcription repressor allowing expression of a kanamycin resistance gene. This system isolated a TEV variant active on a substrate with aspartate at the P1’ position\textsuperscript{114} after 3 rounds of selection. The final variant had 14 mutations and was 3-fold more active on the target substrate and 92-fold less active on the native.

A GFP-RFP fluorescence assay has been used to select for activity on a substrate with arginine at the P1’ position\textsuperscript{214}. Though never kinetically quantified, the isolated variant after 4 rounds had almost 2-fold higher specific activity on the target sequence and showed general broadening of sequence preference at the P1’ position. This was achieved through
a mutation that may increase flexibility of the loop forming the substrate binding tunnel.

Finally, altered specificity at the P1 position has been attempted\textsuperscript{224}. This was a much greater challenge than the previous two works listed as the specificity at this position is strict and evolutionarily conserved. The Yeast Endoplasmic reticulum Screening System (YESS) couples protease activity to display of affinity tags on the surface of the yeast cell are then fluorescently labelled. A library of substrates randomised at the P1 position was screened against a library of proteases randomised around the P1 pocket whilst also counter-selecting against cleavage of the native sequence. The system isolated variants active on either P1 glutamate or histidine. The efficacy of the screening system led to one variant which had 500-fold improvement on the glutamate substrate with a 10-fold reduction on the native, whilst another was 20-fold improved on the histidine substrate and 50-fold reduced on the native.

<table>
<thead>
<tr>
<th>TARGET SEQUENCE</th>
<th>DIFFERENCES TO NATIVE</th>
<th>SELECTION/Screening SYSTEM</th>
<th>MOST IMPROVED VARIANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENLYFQD</td>
<td>1</td>
<td>Split repressor</td>
<td>3.4 - fold</td>
</tr>
<tr>
<td>ENLYFQR</td>
<td>1</td>
<td>GFP-RFP FRET</td>
<td>1.8 - fold</td>
</tr>
<tr>
<td>ENLYFES</td>
<td>1</td>
<td>YESS</td>
<td>500 - fold</td>
</tr>
<tr>
<td>ENLYFHS</td>
<td>1</td>
<td>YESS</td>
<td>20 - fold</td>
</tr>
</tbody>
</table>

Table E-2 | Previously achieved TEV protease specificity changes
Improvment in a promiscuous cleavage activity in a variant isolated through directed evolution towards that sequence. Fold improvements calculated from reported $k_{cat}/K_M$ activities for split repressor and YESS systems, calculated from ‘specific activity’ for GFP-RFP FRET screening system. (Data from references\textsuperscript{114,214,224})

A next step in controlling protease specificity is activity on substrates with multiple residue differences to the native preference sequence. Such substrates represent a far greater evolutionary challenge and so improved understanding of evolutionary potential and how to influence it is required. I therefore use protease specificity changes as a suitable measure of adaptive evolution when investigating evolvability.
2.4 CLINICAL APPLICATIONS

Protease therapeutics and biotechnology tools provide a suitable benchmark for the practical application of evolvability research. The foremost application for the control of protease specificity is as an alternative to antibody therapeutics. Antibodies are protein binders that evolved in vertebrates as part of the immune system to identify and neutralise foreign objects in the body. Antibody therapeutics work by binding to a pathogenic target cell or protein to prevent its action. Use of antibodies as drugs started in 1989 with Muromonab-CD3, a mouse anti-CD3 antibody that acts as an immunosuppressant. Since then, antibodies have proved highly successful with Adalimumab the 7th highest grossing drug (£41 billion in 2011) and over a hundred antibody drugs currently in clinical trials.

In a similar way, if a protease could be made to be specific to a single, chosen target (to avoid off-target effects), it should be able to perform much the same function by cleaving that target (Figure E-2). This would have the benefit over antibody therapeutics of being capable of multiple catalytic turnover, and so require a lower dose.

Figure E-2 | Protease therapeutic concept
A schematic of possible protease targets. A protease therapeutic could cleave a pathogenic soluble protein such as an overexpressed signal ligand. Alternatively, it could cleave the extracellular portion of a receptor in order to down-regulate it. (Adapted from reference)
Generating bespoke antibodies is now routine as it is possible to manipulate the biological systems that already exist in nature for generating specific protein binders. Additionally, established selection protocols allow for affinity maturation in vitro to further enhance binding. Conversely, there are currently no such protocols for manipulating protease specificity (as mentioned in Section 2.3). The development of protease cleavage engineering will advance it as an alternative technology to binding-based biological therapeutics.

In this work I apply experimental evolutionary techniques in an attempt to understand the evolution of evolvability and, in so doing, develop methods for protease specificity engineering. I use the TEV protease lineages that were neutrally evolved on the native sequence (generated in Chapter D) to investigate the changes in promiscuity, immediate evolvability and long-term evolvability as a product of neutral evolution.
3 RESULTS AND DISCUSSION

3.1 TARGET SEQUENCE CHOICE

As suitable targets for adaptive evolution, I searched for sequences in human proteins that had the following properties:

- The sequence shows some level of promiscuous cleavage by wtTEV
- Appears in a region that is flexible enough to be protease accessible
- Cleavage of the protein is likely to have a biological effect

I used the list of therapeutic antibody Stage 3 and Stage 4 targets (provided by Ralph Minter, Table E-3). If stoichiometric binding by an antibody gives a biological effect, then multiple turnover by a protease also has a high chance of effect. Therapeutic antibodies were excluded if the effect has been shown to not be due to simple binding.

<table>
<thead>
<tr>
<th>Target Protein</th>
<th>Target</th>
<th>Disease</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEGF-A</td>
<td>Cancer</td>
<td>Bevacizumab</td>
<td>Avastin</td>
</tr>
<tr>
<td>IgE</td>
<td>Asthma</td>
<td>Omalizumab</td>
<td>Xolair</td>
</tr>
<tr>
<td>TNFa</td>
<td>RA</td>
<td>Adalimumab</td>
<td>Humira</td>
</tr>
<tr>
<td>IL6</td>
<td>RA</td>
<td>Tocilizumab</td>
<td>Actemra</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target Protein</th>
<th>Target</th>
<th>Disease</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>a4b1</td>
<td>MS</td>
<td>Natalizumab</td>
<td>Tyasbi</td>
</tr>
<tr>
<td>EGFR</td>
<td>Cancer</td>
<td>Cetuximab</td>
<td>Erbitux</td>
</tr>
<tr>
<td>HER2</td>
<td>Cancer</td>
<td>Trastuzumab</td>
<td>Herceptin</td>
</tr>
</tbody>
</table>

Table E-3 | Target protein list

Targets of antibody therapeutics (as of 2012). Excluded are antibody/complement-directed cellular cytotoxicity targets and agonist mAb targets. Highlighted are the targets found to have sequences in loop regions which contain the sequence pattern [EDVQRSGT]-x-[LIVAHTYP]-[YLIHWFMVRKAT]-x-Q-(P) by ExPASy ScanProsite (Table E-4).
Flexible loop regions were identified by examining the published structures for regions with high B-factor (Figure E-3a). These loops were cross-correlated with computational predictions of disorder (IUPred\textsuperscript{306}, Figure E-3b) and compiled into a list (Figure E-3c).

In order to maximize the probability of finding some level of promiscuous activity by wtTEV I searched this list for sequence patterns of residues similar to the canonical cleavage sequence of wtTEV\textsuperscript{211,222,223}:

\[[\text{EDVQRSQT}]-[\text{x}]-[\text{LIVAHTYP}]-[\text{YLIHWFMVRKAT}]-[\text{x}]-[\text{Q}]-[\text{P}]\]
Residues are separated by dashes. Allowed residues are contained in [] brackets and disallowed residues in {} brackets. ‘x’ indicates any amino acid. Four target proteins contained loops satisfying these criteria (Figure E-4), each with $\geq 4$ differences to the native ENLYFQ\$S sequence (Table E-4).

Figure E-4 | Target structures and loop sequences
The 4 target proteins identified as having putative cleavable loops (shown in red). (a) Programmed death inhibitory ligand, (b) Vascular endothelial growth factor receptor 2, (c) B lymphocyte stimulator, (d) Interleukin-23. (e) Cutaway of the binding tunnel of wtTEV (PDB code 1lvb). Substrate shown with residues in alternating light and dark blue and labelled. Triad shown in red.

Table E-4 | Target sequences
The 4 target sequences identified as having putative cleavable loops by wtTEV. Sequence similarity is denoted by ‘*’ = identical. ‘.’ = conservative. ‘.’ = semi-conservative.

For eventual practical use of evolved TEV variants, the enzyme would need to cleave the sequence in the context of the true target structure. These experiments, however, concentrate on the evolution and manipulation of enzyme sequence specificity and so the sequences are only assayed in an idealized context of the flexible CFP-YFP linker.
3.2 PROMISCUOUS CLEAVAGE OF TARGET SEQUENCES

The 4 loop sequences were cloned into the linker sequence between CFP and YFP in the pC-Y plasmid by synthesising the oligonucleotide sequences encoding the linker and ligating that directly into the digested plasmid. The substrate names and linker sequences are listed in Table E-5. Activity of wtTEV on C-Y(pdl) and C-Y(veg) was 97- and 13,000-fold lower respectively compared to activity on the native sequence, C-Y(tev). Conversely, there was no detectable cleavage of the substrates with 5 residue differences, C-Y(il23) and C-Y(blys).

<table>
<thead>
<tr>
<th>GENE NAME</th>
<th>LINKER SEQUENCE</th>
<th>DIFFERENCE TO NATIVE</th>
<th>(K_{cat}/K_M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Y(tev)</td>
<td>ENLYFQS</td>
<td>-</td>
<td>1.9x10^4</td>
</tr>
<tr>
<td>C-Y(pdl)</td>
<td>EDLKVQH</td>
<td>4</td>
<td>2.0x10^2</td>
</tr>
<tr>
<td>C-Y(veg)</td>
<td>RDLKTSQ</td>
<td>4</td>
<td>1.5x10^9</td>
</tr>
<tr>
<td>C-Y(blys)</td>
<td>ETPTIQK</td>
<td>5</td>
<td>&lt; 1x10^-3</td>
</tr>
<tr>
<td>C-Y(il23)</td>
<td>DILKDOQ</td>
<td>5</td>
<td>&lt; 1x10^-3</td>
</tr>
</tbody>
</table>

Table E-5 | CFP-YFP target linker variants
Amino acid sequences of C-Y linkers. Residues differing from the native preference sequence are underlined. \(k_{cat}/K_M\) in min^{-1} M^{-1}.

Since C-Y(veg) had the lowest detectable activity it was chosen as the largest challenge for the most in depth analyses, however directed evolution results will be presented on all 4 substrate sequences.

3.3 NEUTRAL EVOLUTION ERODES AVERAGE PROMISCUITY

TEV was previously neutrally evolved on the native cleavage sequence C-Y(tev) in Chapter D to generate lineages of neutral variants. For this specificity evolution study, I have used the \(\mu=1.5\%\), \(N_e=10^4\) lineage (mutations introduced at 1.5% per codon per round, \(10^4\) variants retaining wt-like activity selected each round). This population showed the greatest final tolerance to mutations (Chapter D & Figure D-8) and so was chosen to investigate the effects of neutral mutations on adaptive evolution for new substrate specificities. In this chapter I will refer to populations as ‘x,y’ where ‘x’ is the number of rounds of
neutral evolution on C-Y(tev), and ‘y’ is the number of rounds of adaptive evolution on C-Y(veg) (Figure E-5). Hence, population 7,3 indicates that seven rounds of evolution on C-Y(tev) were followed by three rounds of evolution on C-Y(veg), and so on.

To test for detectability of promiscuous activities of wtTEV in this system, I first looked at cleavage of three ‘substrate libraries’. In each substrate library, two of the substrate residues were randomised (NNS codon) to generate 400 variant substrates. The library of substrates could be co-expressed with the wtTEV protease in the same way that previously I have expressed libraries of proteases with a single substrate and measures by FACS.

Substrates showed a roughly bimodal distribution of cleavage, as seen in libraries of proteases in Chapter D. The percentage of substrates cleaved by wtTEV is an indicator of how many of these sequences are promiscuous substrates (Figure E-6a). Equally, the same experiment with libraries of proteases shows the average promiscuity of the enzyme population on the substrate library. Although cleavage of these substrate libraries does not directly demonstrate which sequences were cleaved, sorting and sequencing cleaved substrates would allow measurement of sequence specificity.
Since C-Y(P34xx) showed an intermediate number of substrates cleaved I used that substrate library to test how promiscuity of the neutrally evolving lineage changed over multiple rounds of neutral evolution (Figure E-6b). The neutrally evolving lineage showed slightly reduced and fluctuating promiscuity towards the C-Y(P34xx) library.

The mixed enzyme populations of the neutral lineage (at rounds 0, 1, 3, 5, 7, 9 & 11) were also purified in order to measure average $k_{cat}/K_M$ activity on C-Y(veg) in vitro. The populations did not show any increase in activity on the promiscuous substrate, in fact the average promiscuous $k_{cat}/K_M$ activity on C-Y(veg) falls below the detection threshold (Table E-6). Together, the reduction in promiscuity on the substrate C-Y(veg) and the substrate library C-Y(P34xx) indicate that the absence of selection leads to loss of promiscuity, not specificity.
<table>
<thead>
<tr>
<th>POPULATION</th>
<th>$K_{cat}/K_M$ C-Y(VEG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wtTEV</td>
<td>$2.0 \times 10^2$</td>
</tr>
<tr>
<td>1.0</td>
<td>$&lt; 1 \times 10^1$</td>
</tr>
<tr>
<td>3.0</td>
<td>$&lt; 1 \times 10^1$</td>
</tr>
<tr>
<td>5.0</td>
<td>$&lt; 1 \times 10^1$</td>
</tr>
<tr>
<td>7.0</td>
<td>$&lt; 1 \times 10^1$</td>
</tr>
<tr>
<td>9.0</td>
<td>$&lt; 1 \times 10^1$</td>
</tr>
<tr>
<td>11.0</td>
<td>$&lt; 1 \times 10^1$</td>
</tr>
</tbody>
</table>

Table E-6 | Population average promiscuous activity on C-Y(veg) $K_{cat}/K_M$ values (M$^{-1}$ min$^{-1}$) in vitro for mixed purified enzyme population cleavage of C-Y(veg).

3.4 **Neutral evolution generates fluctuating immediate evolvability**

Although the populations lost activity on C-Y(veg) on average, there might be minority members that do show promiscuous activity. To test whether promiscuity towards the target sequence C-Y(veg) changed significantly during the neutral evolution lineage, I performed a selection for C-Y(veg) cleavage on the enzyme populations without further mutagenesis (Figure E-7). Selection pressure was applied by 2 rounds of FACS enrichment of the top 1% most active variants on C-Y(veg). The response to selection for a particular substrate by a population is a measure of its immediate evolvability as it depends on whether there are fortuitous adaptive variants in the standing variation.

![Figure E-7](image_url) **Figure E-7 | Immediate evolvability along neutral evolution**

Median FRET ratio of populations being selected for activity on C-Y(veg) (two rounds of enrichment) after 0-11 rounds of neutral evolution on C-Y(tev). FRET ratio normalised to the median ratio of cells expressing wtTEV and C-Y(tev). The population from which variant Va was isolated is indicated.
Populations 5,0 and 11,0 responded to this selection pressure with an increase in the population average activity on C-Y(veg), demonstrating their high immediate evolvability. From population 5,0 (with the highest immediate evolvability) I isolated the most active single variant (TEV-Va) which showed a 35-fold improvement in $k_{cat}/K_M$ over wtTEV on C-Y(veg). This is somewhat larger than the typical 2 to 5-fold promiscuous activity improvements found in variants isolated from standing variation in a neutrally evolved lineages.\textsuperscript{188,290,291,293} (Table E-1). It is also higher than most previous attempts at protease specificity engineering\textsuperscript{114,214,224}. The best of these experiments reported a 500-fold improvement\textsuperscript{224}, however this was from first randomising the substrate and selecting to ensure an easy target and only altering a single residue of specificity.

The non-linear increase in immediate evolvability as the population neutrally evolves indicates that mutations adaptive for C-Y(veg) activity are not directly selected for during neutral evolution. Lack of a consistent trend along the neutral lineage confirms that C-Y(veg) activity is truly an unselected promiscuous activity, i.e. is lost due to lack of selection. Therefore, even though the populations on average have no activity on C-Y(veg), some populations do contain members that do show improved activity.

The immediate evolvability towards C-Y(veg) also does not correlate with the general promiscuity towards C-Y(P34xx), although both measures fluctuate along the neutral lineage (Figure E-6b & Figure E-7). Therefore, the level of promiscuity in the population does not appear to be a good predictor of whether promiscuity for a particular target substrate is more likely to be present.
3.5 **Neutral Evolution Increases Long-Term Evolvability**

The *immediate* evolvability so far presented describes the prevalence of fortuitous mutations in a population and is the facet of evolvability most commonly addressed in previous studies. To expand upon these works, I also measured the potential for further beneficial mutations by performing 3 rounds of adaptive evolution on C-Y(veg) (Figure E-8a). At each round, diversity was introduced (µ=0.5% of codons per gene per round) after which the top 1% most active variants were sorted, re-expressed and re-sorted to ensure stringent selection (N=10^6 screened, N_e=10^4 selected). As starting points I used wtTEV and populations 5,0 and 7,0 from the neutral evolution lineage. These populations compared the *long-term* evolvability of populations which previously showed very different levels of *immediate* evolvability (Figure E-6c).

**Figure E-8 | Long-term evolvability on target sequences of neutrally evolved populations.**

Median FRET ratio on C-Y(veg) of populations being selected for (a) activity on C-Y(veg), (b) C-Y(pdl), (c) C-Y(blys), (d) C-Y(il23). Black circles indicate the lineage starting from wtTEV (0,0→0,3). Grey circles indicate the lineage starting from population neutrally evolved on C-Y(tev) for 5 rounds (5,0→5,3). White circles indicate the lineage starting from population neutrally evolved on C-Y(tev) for 7 rounds (7,0→7,3). The populations from which variants Vb, Vc, Vd and Pa were isolated are indicated.
For C-Y(veg), each round shows small increase in population median activity, however lineage 7,0→7,3 shows a fastest rate increase. The most active variant isolated from population 0,3 (TEV-Vb) has a 40-fold faster $k_{cat}/K_M$ than wtTEV. This is barely higher than that of Va indicating that, by the measure of most active isolatable variant, long-term evolvability of wtTEV is barely higher than the immediate evolvability present in population 5,0 from variation generated by neutral evolution alone.

Average cleavage increased faster for populations that had been neutrally evolved. In particular the lineage 5,0→5,3 improves rapidly but plateaus whereas lineage 7,0→7,3 reaches a higher average activity. The most active variants from populations 5,3 (TEV-Vc) and 7,3 (TEV-Vd) were 37 and 203-fold improved respectively (Figure E-9 & Figure E-10a). Of note is that Vc (from population 5,3) is directly descended from Va (from population 5,0) but shows minimal activity improvement, merely slightly increased solubility. Therefore populations that have been neutrally evolved show faster adaptive evolution and yield more active variants (i.e. long-term evolvability).

Long-term evolvability was also compared between wtTEV and population 7,0 for 3 rounds of adaptive evolution on C-Y(pdl), C-Y(il23) and C-Y(blys). After 3 rounds of evolution on C-Y(pdl), population 7,3(pdl) contained a variant 25-fold improved on C-Y(pdl) (TEV-Pa), whereas 3 rounds of evolution from wtTEV yielded no improved variants (Figure E-8b). Neither wtTEV nor population 7,0 showed any improvement on C-Y(il23) or C-Y(blys) after 3 rounds of evolution on those substrates (Figure E-8c,d). The higher long-term evolvability of population 7,0 appears to be an intrinsic property, rather than specific to C-Y(veg) since population 7,0 also showed higher long-term evolvability for C-Y(pdl) activity than wtTEV. Both C-Y(il23) and C-Y(blys) proved too great an evolutionary challenge for this system.
The purified variants, though showing improved $k_{\text{cat}}/K_M$ on their target sequences, have comparatively little change in the $k_{\text{cat}}/K_M$ on C-Y(tev) (Figure E-10b). This indicates that interactions with the new substrate are specific rather than a general increase in reactivity.

Figure E-9 | Activity on C-Y(veg) and C-Y(pdl) for wtTEV and evolved variants
Activity ($k_{\text{cat}}/K_M$) (a) wtTEV and (b) evolved variants on substrates C-Y(pdl) and C-Y(veg). Substrate indicated by bar colour (C-Y(tev) in white, C-Y(pdl) in purple, C-Y(veg) in green).

Figure E-10 | Activity of variants after 3 rounds of evolution on target sequence
(a) $k_{\text{cat}}/K_M$ improvement of evolved variants on substrates C-Y(pdl) and C-Y(veg). (b) $k_{\text{cat}}/K_M$ improvement on the original C-Y(tev) substrate of evolved variants. Top row indicates rounds of neutral evolution on C-Y(tev). Bottom row indicates rounds of adaptive evolution on C-Y(target).
3.6 **STEPPING-STONES DID NOT AID EVOLUTION ON UN-CLEAVABLE SEQUENCES**

An attempt was also made at circumventing the constraints that prevent the evolution of activity on C-Y(il23) and C-Y(blys). Since both C-Y(pdl) and C-Y(veg) are cleaved to some extend by wtTEV it was possible to narrow down two residues in each substrate that were likely the main factors preventing activity. Each of these residues was individually introduced into C-Y(tev) as well as being individually removed from C-Y(il23) or C-Y(blys) (*Table E-7*). Residues that were preventing cleavage could be spotted by their strong reduction in activity when added to C-Y(tev), or their restoration of activity when removed from C-Y(target) (*Figure E-11*).

Some residues such as isoleucine in position P6 gave no change in cleavage either when added to C-Y(tev) or when removed from C-Y(il23), indicating that it is not the cause of cleavage inhibition. Addition of proline at P4 almost completely prevented cleavage of C-Y(tev), however its removal from C-Y(blys) failed to increase cleavage. Therefore, although the P4 proline is inhibitory (likely as a β-sheet
breaker), the blys sequence must contain several other residues that restrict cleavage.

Of the 4 residues tested, the aspartate in P2 of C-Y(il23) seemed the most promising as it caused a 50% reduction in activity when added to the native sequence and restored 15% activity when removed from C-Y(il23) (Figure E-11). Therefore, I performed 3 rounds of adaptive evolution on each of C-Y(t+D2) and C-Y(i-D2). I did this starting both from wtTEV and from population 7,0. After 3 rounds, however, none of these 4 lineages showed any detectable increase in cleavage.

Together these results indicate that some target sequences are simply too challenging for the approaches used here. For activities too far away (e.g. C-Y(blys) and C-Y(il23)), neither robust variants nor the wild-type showed any evolvability, although it is possible that populations other than 7,0 might have greater evolvability for this target. Additionally, attempting several stepping stone sequences between C-Y(tev) and C-Y(target) could make finding stepwise solutions more likely.

3.7 Mutations that underlie immediate and long-term evolvability

The isolated improved variants (TEV-Va, Vb, Vc & Vd) contain between 5 and 13 amino acid mutations (Figure E-12a-c). Populations previously neutrally evolved on C-Y(tev) yielded variants with the most mutations due to the accumulated neutral mutations before the adaptive evolution rounds.

The isolated variants all showed mutations in their C-termini which forms the flexible lid region of the substrate binding tunnel in viral proteases (Figure E-12d-f). This has previously been shown to be important in substrate recognition and mutations have previously been observed in directed evolution or specificity switched in TEV\textsuperscript{214,224}. 
Additionally, all isolated variants contained N171D or N176I (Figure E-12a-c). These two asparagines are responsible for the recognition of the aspartate of the native cleavage sequence (ENLYFQS) (Figure E-13a). In C-Y(veg) the substrate sequence contains arginine in the place of aspartate rendering these residues counter-productive and removing cationic residues from the P5 binding pocket likely reduces repulsion of the C-Y(veg) substrate sequence.

When either the N171D or N176I mutations were introduced into wtTEV, it increased activity on the C-Y(veg) substrate by more than an order of magnitude (Figure E-13b). N176I was neutral to activity on the native sequence but N171D changed specificity by reducing activity on C-Y(tev) by an order of magnitude (likely due to negative charge repulsion). Both mutations also reduced soluble expression of the protease (Figure E-13c).

In order to understand the sequence changes giving rise to fluctuating immediate evolvability, I reanalysed data from Chapter D, 3.3. Looking at the population sequence variation in the neutrally evolving lineage,
it can be seen that position 171 is mostly purged of mutations (Figure E-13d) indicating strong constraint at that position. No single mutation accumulated above 1.5% and only in conjunction with known stabilising mutations (Figure E-13e). This agrees with the reduction in activity on the native substrate sequence by the N171D mutation.

Conversely, position 176 accumulated amino acid diversity at about the average rate for the whole gene during the neutral evolution (Figure E-13d). Because there was no direct selection for or against N176I, its concentration in the population fluctuated over the course of neutral evolution. The N176I mutation was most prevalent in population 5,0 (Figure E-13f), the population with the highest immediate evolvability towards C-Y(veg) (Figure E-6c). The same N176I mutation was
therefore convergently evolved during adaptive evolution starting from wtTEV. Thus as neutral or slightly deleterious mutations stochastically fluctuate in concentration within the population, they have profound effects on immediate evolvability.

When viewed as a phylogeny, it can be seen that Vd and Pa (the most active variants found for their target substrates) are both descended from $\beta_2$ (Figure E-14), the robust variant identified in Chapter D. Variant Vc, however, is directly descended from Va and so offers an explanation for why the 5,0→5,3 lineage plateaued in average cleavage, and yielded only a 37-fold improved variant. By being highly adapted, Va outcompeted less the well adapted but more robust variant, $\beta_2$. Therefore the lineage had high immediate evolvability, enriching Va, but low long-term evolvability as robust variants were absent.

The exceptionally high mutation tolerance of $\beta_2$ may aid adaptive evolution both by having fewer deleterious mutations, and a greater proportion of beneficial mutations by buffering beneficial but
destabilising mutations. This would explain both the fast rate of adaptation and highly active variant evolved from population 7,0.

In this way, the causal factors for immediate and long-term evolvability can be antagonistic. Fortuitous standing variation leads to high immediate evolvability but its enrichment may come at the cost of robust variants that, though they are not active towards the new substrate, would have given higher long-term evolvability.
4 CONCLUSIONS

4.1 EVOLUTION OF PROMISCUITY, AND IMMEDIATE AND LONG-TERM EVOLVABILITY

In this chapter I show how neutral evolution affects immediate evolvability though genetic drift of sequences and long-term evolvability through the emergence of robustness.

TEV protease is known for its high sequence specificity so unsurprisingly has very low promiscuous activity on alternative substrate sequences. Along the neutral evolution lineage, average promiscuity towards a set of 400 test substrates fell slightly indicating the general loss of promiscuous activities as did promiscuity towards a specific test substrate sequence, C-Y(veg). This is perhaps explained by the slight drop in native activity (Chapter D) which may cause some minor promiscuous activities to drop below the detection threshold.

However, as the population evolved through the local neutral network, adaptive standing variation (e.g. N176I) appeared and disappeared introducing chance increases in promiscuous cleavage activities into some members of the population (cryptic variation). These members (e.g. TEV-Va from population 5,0) could be isolated by selection to indicate the large fluctuations in immediate evolvability.

Simultaneously, later populations contained variants with mutations that improved robustness (Chapter D, 3.4) and influenced long-term evolvability. The mutation tolerance lead to a faster increase in population average activity and the isolation of the most active variants, Vd and Pa, after 3 rounds of adaptive evolution on C-Y(veg) and C-Y(pdl) from population 7,0. Therefore, in this case, longer durations of neutral evolution did increase long-term adaptive evolvability due to the emergence of robustness.
Immediate and long-term evolvability do not correlate as they represent two orthogonal library properties (likelihood of bearing a variant with altered specificity vs. improvement in mutational tolerance). Evidence comes from the long-term evolvability of population 7,0, which is higher than wtTEV despite population 7,0 starting with no greater immediate evolvability. Indeed there is evidence of antagonism between the two properties as population 5,0 shows the highest immediate evolvability, yielding variant Va, which outcompetes robust variants (such as β2) and lowers long-term evolvability.

4.2 Relevance to Protein Engineering

These experiments compare favorably with previously reported evolution attempts. Although I was able to isolate the enzyme Va (35-fold improved on C-Y(veg)) from standing variation alone (also somewhat higher than is typical for adaptive variation in neutral evolution Table E-1), the inconsistency of immediate evolvability explains the limited adaptive potential of some previous neutral evolution experiments. However, in later rounds, the enrichment of the descendants of robust variants does improve long-term evolvability. Starting from wtTEV, 3 rounds of adaptive evolution produced no improved variants on C-Y(pdl) and only a 40-fold improved variant on C-Y(veg) (Vb). The same adaptive evolution starting from population 7,0 found a 25-fold improved Pa (on C-Y(pdl)) and the >200-fold more active Vd (on C-Y(veg)), both descended from the robust variant β2.

Therefore, characterising the exploration of the local neutral network by an evolving population of genes allowed insight into the relationship between neutral evolution and adaptive evolvability. Additionally, it highlights the generation of robust enzymes as a key goal in improving long-term evolvability in practical applications of experimental evolution.
Chapter F - Discussion and outlook

1 SUMMARY

In this chapter I will draw together the results presented in this thesis into two themes. Firstly I present how my results advance the understanding of protein evolution by describing relevant features of the fitness landscape surrounding TEV protease. Secondly I discuss the relevance of these findings to future protein engineering efforts.
Development of the multi-format FRET screening system in Chapter B allowed the generation of an integrated set of data using the same fluorescent assay. It was therefore possible to use the same substrate for high throughput FACS screening all the way to detailed kinetics studies. Although the results presented here are specific to TEV protease, in conjunction with other published works they shed light on several general aspects of enzymology.

2.1 Neutral Fitness Plateau

In the case of TEV protease the distribution of mutant phenotypes is bimodal with mutations either reducing activity to background levels or affecting it so little that a wt-like phenotype remains (in several different assay formats – Chapter D). This indicates a local fitness landscape that is mainly comprised of flat, high plateaus (a neutral network) traversed by deep fitness canyons (sudden, complete loss of function) (Figure F-1a,b).

This is similar to previous observations of a bimodal distribution of fitness effects of mutations on proteins and organisms\textsuperscript{79,153,278}. In such landscape it has been predicted that populations will evolve towards regions of higher mutational tolerance\textsuperscript{149}. This robustness is equivalent to the population moving in the landscape to regions with fewer nearby fitness cliffs and more so neutral neighbours\textsuperscript{263} and as such, the evolution of robustness can be an emergent property of landscape topology.

One way of moving away from fitness cliffs if to move towards the centre of a neutral network. A relatively smooth, additive landscape
forms a neutral network with a coherent peak (Figure F-1c). In this case it would be expected that intermediate sequences between the starting enzyme (wtTEV) and the robust variant (e.g. TEV-β2) would be progressively more robust. In the lineage that produced β2, the phylogeny gives no evidence of particularly mutation tolerant intermediates. This result is more in line with a rugged, epistatic landscape which creates a complex, irregular neutral network in which the population is funnelled towards a different area with fewer cliffs (Figure F-1d) and would not necessarily show more robust intermediates.

Figure F-1 | Bimodal distribution of fitness effects as a landscape
A representation of a broad fitness peak (a) and a detailed view of the top plateau of a fitness peak (b) shows that even at high fitness, many single mutations can completely inactivate the enzyme causing cliffs that take fitness from ≈1 to ≈0. (c) A neutral network generated by a smooth, additive landscape that forms a ‘round’ peak. Each circle represents a functional gene variant and lines represents point mutations between them. Light grid-regions have low fitness, dark regions have high fitness. White circles have few neutral neighbours, black circles have many. Light grid-regions contain no circles because those sequences have low fitness (as in Chapter A, Figure A-16). (d) An alternative scenario, in which a large neutral network is criss-crossed with ‘fitness cliffs’ due to epistasis (a neutral network representation of (b)). In this case it is also possible for a population to evolve to regions with fewer local cliffs.
2.2 RUGGED FITNESS VALLEY

By forcing TEV into a local fitness valley (TEV\textsuperscript{Ser}) and experimentally evolving for activity recovery in Chapter C, I have been able to map out an uphill trajectory (TEV\textsuperscript{Ser}→TEV\textsuperscript{Ser}X) which lies parallel to a nearly flat neutral trajectory (TEV\textsuperscript{Cys}↔TEV\textsuperscript{Cys}X). In this case I can plot an explicit transect of a fitness landscape (Figure F-2) showing that the mutations accumulated in TEV\textsuperscript{Ser}X are nearly-neutral to activity using the original cysteine nucleophile. In this case TEV\textsuperscript{Ser}X could be described as being robust to nucleophile exchange, an alternative and far more specific facet of robustness than the more general structural robustness to random mutagenesis described in Chapter D.

These observations contrast with the weak negative trade-offs between old and new function that are typical for evolution for altered substrate specificity or catalytic activity\textsuperscript{26,54}. When selective pressure is applied for new function, typically the lack of selection the old activity leads to its loss by genetic drift as mutations re-organise the active site\textsuperscript{53}. Given the specialisation of natural serine protease towards their nucleophile, we might expect that further evolution would start to show a trade-off.
Mutations that are adaptive, or even neutral towards use of the new serine nucleophile may gradually erode the potential to use the original cysteine. However the main constraints in this case were observed in the enzyme core and mutations in contact with the catalytic triad. The lack of a trade-off (so far) between old and new function indicates that active site optimisation around the serine nucleophile doesn’t disrupt the use of cysteine. It also suggests that neutral evolution could bridge widely differing active sites in natural evolution.

High throughput screening and sequencing in conjunction with traditional directed evolution approaches allowed characterisation of the local fitness landscapes flanking this trajectory (Figure F-3).

The analysis characterises the fitness valley around TEVSer as having a rugged topology yet an abundance of uphill pathways directed by constrains in core structural residues and adaptive potential in the active site perimeter. This underlying landscape guided the trajectory towards one possibility in a set of many but similar solutions. Therefore, although only one evolutionary pathway was characterised
in detail, it is likely representative of a large class of equivalent trajectories.

2.3 ALTERNATIVE FITNESS PEAKS

Finally, in Chapter E, I show evolution towards several nearby fitness peaks with alternative sequence specificities. By measuring both immediate and long-term evolvability I show that some variants with altered specificity could be found within the neutrally evolving lineage (e.g. TEV-Va), but that movement of population to regions of high robustness gave eventual access to better variants (e.g. TEV-Vc).

Fluctuating immediate evolvability also aids our understanding of divergent evolution after gene duplication. It suggests that in the evolutionary history of a population of enzyme variants, promiscuous activities are constantly being introduced and purged at random. As genes duplicate at a roughly constant rate, redundant copies are mostly inactivated by pseudogenisation. The chance periods of high promiscuity, however, may lead to greater likelihood of sub- and neo-functionalisation.

Immediate and long-term evolvability have different requirements. Immediate evolvability is achieved by a diverse, polymorphic population in order to fortuitously find rare beneficial mutations. Long-term evolvability, however, is improved by a robust starting enzyme (Chapter D) that, at the least has fewer deleterious mutations, and at most may have more beneficial ones compared to wtTEV. Consequently, it is possible that these even trade-off against one another if some active variants (higher immediate evolvability) displace less active but robust variants (higher long term evolvability) early in an adaptive evolutionary period.
3 Enzymes and Engineering

3.1 Neutral Evolution and Controlling Protease Specificity

The relevance of ‘neutral drift’ for protein engineering has been a contentious topic with conflicting results as to whether the accumulation of neutral mutations increases evolvability, especially in a protein engineering context. By measuring the sequence evolution of populations (Chapters D & E) and reconstructing their phylogenies I was able to partly resolve this issue. Specifically, neutral evolution displayed two distinct processes with independent evolvability effects.

The first process is a classical genetic drift as the population accumulated neutral and nearly-neutral mutations. As an engineering tool this method has several drawbacks. Firstly, mutations have to be neutral to the original function as well as beneficial to the new. Though this is true of many mutations, it is certainly not true of all. Additionally, neutral mutations appear and disappear stochastically in the population as they are under no specific selection for retention (Chapter E). This leads to large fluctuations in the immediate evolvability from standing variation and arbitrary lengths of neutral evolution are not guaranteed to explore productive regions of sequence space. Hence, screening for new activities after an arbitrary number of generations may miss adaptive potential that was present in earlier rounds. This stochastic element of neutral evolution can accurately be called ‘neutral drift’ since it describes non-directional diffusion in the neutral network. As an engineering technique then, the ‘drift’ part of neutral evolution is inefficient, as observed for previous experimental evolution of β-glucuronidase to β-galactosidase.
The second process is a directional movement of populations towards enzymes with higher robustness (Chapter D). This is a process of second-order selection for extinction resistance\textsuperscript{186} and so the label of ‘drift’\textsuperscript{188,274,290,291,293} is misleading. The neutral evolution of robustness is more promising as an engineering technique since it reduces the proportion of mutations that are deleterious to function and enables faster adaptation to a new activity as well as reaching a more active final variant, improving long-term evolvability.

Finally, none of the variants evolved in Chapter E showed re-specialisation to the new sequence, activity was improved on the new sequence, sometimes by several orders of magnitude, but the native activity never varied more than 2-fold. This may be in issue in applications where tight specificity is required but could be solved with appropriate counter-selection\textsuperscript{224}. However, the >200-fold improvement of TEV-Va on its target sequence (4 residue differences to native sequence) compares favourably with previous works that change the substrate specificity at only one substrate residue. This gives some optimism that the large specificity changes desired in protease engineering are possible.

3.2 Manipulating active site chemistry

Chapter C shows the sensitivity of the enzyme to nucleophile mutation, exchange of a single atom amongst the tens of thousands in the enzyme. Results are consistent with previous nucleophile mutagenesis studies\textsuperscript{212,238–249} and with the difficulties of de novo enzyme design in which a designed active site\textsuperscript{307} is docked into existing protein structures\textsuperscript{308,309}. Together, these results highlight that engineering of optimised catalysis is more than precise orientation, but also correct activation and tuning of chemistry\textsuperscript{101,310,311}. 

152
Nevertheless, directed evolution of the enzyme was able to adapt the active site to accommodate a new nucleophile despite constraints such as epistasis and negative pleiotropy between activity and solubility. The proportion of beneficial mutations is at the high end of previous measurements and reflects that the local sequence space is much richer in solutions to this chemical problem than originally expected. The adaptive mutations for nucleophile re-optimisation were disproportionately located in the active site periphery (in strong contrast to the neutral evolution lineages in Chapter D where such mutations were purged).

Additionally, the lack of a strong trade-off between use of the original and new nucleophile supports manipulation of core catalytic machinery as an avenue of protein engineering\textsuperscript{105}. In particular it may allow access to alternative chemistries, not normally associated with a particular protein scaffold or specificity.

4 SYNOPSIS

To close, I have demonstrated the utility of integrating traditional and high-throughput techniques both for directed evolution (Chapters D & E) and for measuring fitness landscapes to aid the understanding of the process and outcomes of directed evolution (Chapter C). I have used these techniques to help create an integrated picture of neutral evolution, the emergence of robustness, evolvability, active site evolution, and substrate specificity. Together these aid our understanding both of enzyme evolution and engineering.
Chapter G - Materials and methods

1 PLASMIDS

<table>
<thead>
<tr>
<th>NAME</th>
<th>RELEVANT FEATURES</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>pET28-MBPTEV</td>
<td>Kanmx, T7 lac promoter, MBP-TEV fusion</td>
<td>This project</td>
</tr>
<tr>
<td>pET28-STEV</td>
<td>Kanmx, T7 lac promoter, PelB export sequence, TEV</td>
<td>This project</td>
</tr>
<tr>
<td>pET28-STEVa</td>
<td>Kanmx, T7 lac promoter, PelB export sequence, C151A TEV</td>
<td>This project</td>
</tr>
<tr>
<td>pMAA</td>
<td>Bla, Ara promoter, MBP, AraC</td>
<td>Novagen</td>
</tr>
<tr>
<td>pMAT(MBPTEV)</td>
<td>Bla, Ara promoter, MBP-TEV fusion, AraC</td>
<td>Gatti-Lafranconi</td>
</tr>
<tr>
<td>pMAA(MBPTEV)</td>
<td>Bla, Ara promoter, MBP-TEV fusion, AraC</td>
<td>Gatti-Lafranconi</td>
</tr>
<tr>
<td>pMAA(MBPTEVs)</td>
<td>Bla, Ara promoter, MBP-TEVSer fusion, AraC</td>
<td>This project</td>
</tr>
<tr>
<td>pMAA(MBPTEVa)</td>
<td>Bla, Ara promoter, MBP-TEVAla fusion, AraC</td>
<td>This project</td>
</tr>
<tr>
<td>pMAA(TEV)</td>
<td>Bla, Ara promoter, TEV, AraC</td>
<td>Gatti-Lafranconi</td>
</tr>
<tr>
<td>pMAA(TEVs)</td>
<td>Bla, Ara promoter, TEVSer, AraC</td>
<td>This project</td>
</tr>
</tbody>
</table>

Table G-1 | Protease plasmids
pMAA(MBPTEVs) was used to express all variants in Chapter C except for protein expression for thermal denaturation, when protease genes were subcloned into pMAA(TEVs). Similarly for Chapter C pMAA(MBPTEV) and pMAA(TEV) were used.

<table>
<thead>
<tr>
<th>NAME</th>
<th>RELEVANT FEATURES</th>
<th>SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>pC-Y(cleavage)</td>
<td>CAT, Dual-GFP FRET fusion, AraC</td>
<td>Gatti-Lafranconi</td>
</tr>
<tr>
<td>pGR32</td>
<td>CAT, lac promoter, BLIP, Lac</td>
<td>Palzkill</td>
</tr>
<tr>
<td>pJET1.2/Blunt</td>
<td>Blunt ligation plasmid</td>
<td>Fermentas</td>
</tr>
<tr>
<td>pSU18</td>
<td>CAT,</td>
<td>Novagen</td>
</tr>
<tr>
<td>pUC18</td>
<td>Bla,</td>
<td>Novagen</td>
</tr>
<tr>
<td>pJET-bla</td>
<td>Bla (+ required restriction sites)</td>
<td>This project</td>
</tr>
<tr>
<td>pSB-wtBLIP</td>
<td>Bla, CAT, LacI, lac promoter, BLIP</td>
<td>This project</td>
</tr>
<tr>
<td>pSB-iBLIP</td>
<td>Bla, CAT, LacI, lac promoter, iBLIP</td>
<td>This project</td>
</tr>
</tbody>
</table>

Table G-2 | Substrate plasmids
pC-Y(cleavage) represents a set of plasmids all encoding the CFP-YFp substrate but with different cleavage sequences as the linker. The full table of linker sequences is detailed in Chapter C.
## Primers

<table>
<thead>
<tr>
<th>Name</th>
<th>$T_M/°C$</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ins1F</td>
<td>71</td>
<td>ACTUCCAGAAATGCCCGACGCTCACCTCAGC</td>
</tr>
<tr>
<td>ins1R</td>
<td>71</td>
<td>AAGUACAGATTCCTCTCGGCGCCAGCAGCTTCGCC</td>
</tr>
<tr>
<td>ins2F</td>
<td>69</td>
<td>ACTUCCAGAAATGCCCGACGCGAAGGTGGACTCG</td>
</tr>
<tr>
<td>ins2R</td>
<td>71</td>
<td>AAGUACAGATTCCTCTCGGCGCCAGCAGCTTCGCC</td>
</tr>
<tr>
<td>blaF</td>
<td>60</td>
<td>GATCGATCAAGTTGCTGCTGCTAGTGAGCGGAGCGTACG</td>
</tr>
<tr>
<td>blaR</td>
<td>58</td>
<td>GCTGCTGATGTCGAGGCTGCACTGAGGTGGAAGC</td>
</tr>
<tr>
<td>blipF</td>
<td>68</td>
<td>GCCATTGGAGATTCCTCTCCTCGGAGCGCATATGCTTCAG</td>
</tr>
<tr>
<td>blipR</td>
<td>68</td>
<td>GCCATTGGAGATTCCTCTCCTCGGAGCGCATATGCTTCAG</td>
</tr>
<tr>
<td>tevF</td>
<td>60</td>
<td>GTTCGAGGGGATCCAGG</td>
</tr>
<tr>
<td>tevR</td>
<td>59</td>
<td>CCCCGATCCCTCGGAGGC</td>
</tr>
<tr>
<td>restF</td>
<td>66</td>
<td>TGAATAGAGATCTTATGACTCGAGCGGCTTATCCAGGCTACG</td>
</tr>
<tr>
<td>restR</td>
<td>67</td>
<td>AGGCTACAGATCTCTGATATCGATCTCGAATCAAGGCTCCACCTGAGGC</td>
</tr>
<tr>
<td>tevaF</td>
<td>63</td>
<td>GGGGACGGCGCCATGGG</td>
</tr>
<tr>
<td>tevaR</td>
<td>64</td>
<td>CTAATGAGGGATCCAGCCTCGGCTTGCTGCTGACG</td>
</tr>
<tr>
<td>tevsF</td>
<td>63</td>
<td>GGGGACGGCGCCATGGG</td>
</tr>
<tr>
<td>tevsR</td>
<td>64</td>
<td>CTAATGAGGGATCCAGCCTCGGCTTGCTGCTGACG</td>
</tr>
<tr>
<td>shufF</td>
<td>60</td>
<td>GAAACATCCCGAGATGCTGCC</td>
</tr>
<tr>
<td>shufR</td>
<td>59</td>
<td>GTTATGCTAGTATGCTCAGG</td>
</tr>
<tr>
<td>seqXF</td>
<td>60</td>
<td>CGTATCGCCTCCCTCGGCCATCAG (MID-X) GTTCGAGGGGATCCAGG</td>
</tr>
<tr>
<td>seqXR</td>
<td>59</td>
<td>CGTATCGCCTCCCTCGGCCATCAG (MID-X) GTTCGAGGGGATCCAGG</td>
</tr>
</tbody>
</table>

**Table G-3 | Primers mentioned in report**

Primers were designed with NetPrimer (primer $T_M$ was calculated under standard conditions, 1.5nM Mg$^{2+}$) and synthesised by Invitrogen Life Technologies. Sequencing results were analyzed with FinchTV and CLC sequence viewer 6. Primers seqXF and seqXR contain multiplex identifier sequences (MID-X) detailed in Table G-4.
### Table G-4 | Multiplex identifier sequences

Standard multiplex sequences that are each differ by at least two nucleotides. Used to sequence multiple different populations simultaneously by 454 sequencing. The mixed sequences can then be separated afterwards (de-multiplexed) by grouping sequences by their MID. Each group then corresponds to a population before the samples were mixed for sequencing.

<table>
<thead>
<tr>
<th>MID-X</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACGAGTGCCT</td>
</tr>
<tr>
<td>2</td>
<td>ACGCTCGACA</td>
</tr>
<tr>
<td>3</td>
<td>AGACGACTCTC</td>
</tr>
<tr>
<td>4</td>
<td>AGCACTGTAGG</td>
</tr>
<tr>
<td>5</td>
<td>ATCAGACACG</td>
</tr>
<tr>
<td>6</td>
<td>ATATCCGCGAG</td>
</tr>
<tr>
<td>7</td>
<td>CGTGCTCTCTA</td>
</tr>
<tr>
<td>8</td>
<td>CTGCGGTGTCA</td>
</tr>
<tr>
<td>9</td>
<td>TCTCTATGCG</td>
</tr>
<tr>
<td>10</td>
<td>TGATAGCTCT</td>
</tr>
<tr>
<td>11</td>
<td>CATAGTAGTG</td>
</tr>
<tr>
<td>12</td>
<td>CGAGAGATAC</td>
</tr>
<tr>
<td>13</td>
<td>ATACGACGTA</td>
</tr>
<tr>
<td>14</td>
<td>TCACGTACTA</td>
</tr>
<tr>
<td>15</td>
<td>CGTCTAGTAC</td>
</tr>
<tr>
<td>16</td>
<td>TCTACGAGC</td>
</tr>
<tr>
<td>17</td>
<td>TGACTACTC</td>
</tr>
<tr>
<td>18</td>
<td>ACGTCTACAG</td>
</tr>
<tr>
<td>19</td>
<td>CGTAGACTAG</td>
</tr>
<tr>
<td>20</td>
<td>TACGAGTATG</td>
</tr>
</tbody>
</table>
3 Buffers

Lysis
50 mM Tris
300 mM NaCl
5 mM Imidazole
pH 8

Wash
50 mM Tris
300 mM NaCl
50 mM Imidazole
pH 8

Elution
50 mM Tris
300 mM NaCl
300 mM Imidazole
pH 8

96W-Lysis
50 mM NaH₂PO₄
300 mM NaCl
pH 8

96W-Wash
83 mM NaH₂PO₄
500 mM NaCl
10 mM Imidazole
pH 7

96W-Elution
50 mM NaH₂PO₄
300 mM NaCl
150 mM Imidazole
pH 7

96W-MES
20 mM MES
100 mM NaCl
150 mM Imidazole
pH 5

TEV Storage
50 mM Tris
1 mM EDTA
10% sucrose
pH 8

TEV Activity
50 mM Tris
1 mM EDTA
10% sucrose
pH 8 with HCl
1 mM DTT (freshly added)

RF1
100 mM RbCl
30 mM Potassium acetate
10 mM CaCl₂
50 mM MnCl₂
10% glycerol
pH 5.8 with acetic acid
Filter sterilised

RF2
10 mM MOPS
10 mM RbCl
75 mM CaCl₂
10% glycerol
pH 6.8 with acetic acid
Filter sterilised

PBS
137 mM NaCl
2.7 mM KCl
10 mM NaH₂PO₄
1.8 mM KH₂PO₄
pH 8
4 MOLECULAR BIOLOGY

4.1 GENE AMPLIFICATION

PCR using primers containing the desired restriction sites was performed in an Eppendorf Mastercyler (25 rounds of: 30 s at 94 °C, 30 s at (T_m-5) °C, 1 min per kb of amplicon at 72 °C) using 0.4u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs, 100 ng total plasmid template in 20 µl. Refer to Table G-3 for primer T_m values.

4.2 USER MUTAGENESIS

Performed as published312. Briefly, the plasmid was amplified by PCR in an Eppendorf Mastercyler (30 s at 94 °C, 30 s at 56 °C, 60 s at 72 °C) using Pfu Cx polymerase according to the manufacturer's instructions and uracil-containing primers. DpnI (1 µl) and USER (1 µl) were then added to the crude product (20 µl) and incubated 1hr at 37 °C. DNA was purified using the QIAquick PCR purification kit and eluted in H_2O (30 µl). DNA ligase buffer and T4 DNA ligase (1 µl) was added and incubated 12 h at 16 °C.

4.3 TARGETED MUTAGENESIS

For targeted mutagenesis, a two-step, overlap-extension method was used. This requires each half of a gene to be amplified using outside primers and inside mutagenic primers that introduce the desired mutation. The two gene fragments are then used as primers for each other to generate the full length mutant gene. PCR was performed in an Eppendorf Mastercyler using flanking primers and primers containing the desired mutations tevF+(mutagenic primer) and tevR+(mutagenic primer) was performed (10 rounds of: 30 s at 94 °C, 30 s at (T_m-5) °C, 60 s at 72 °C) using 0.4u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs, 100 ng amplicon template in 20 µl. These amplificates were mixed and overlap extension PCR with the flanking primers tevF+tevR completed the gene (10 rounds of: 30 s at 94 °C, 30 s at (T_m-5) °C, 60 s at 72 °C) using 0.4u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs, 100 ng total plasmid template in 20 µl. The product was ligated into pMAA as described below.

4.4 RANDOM MUTAGENESIS

Error prone PCR in an Eppendorf Mastercyler using flanking primers tevF+tevR was performed (30 rounds of: 30 s at 94 °C, 30 s at 50 °C, 90 s at 72 °C, 30 rounds) using 0.4u Mutazyme II polymerase (Agilent), 1µM each primer, 0.8 mM dNTPs, 400ng plasmid template in 20 µl. The product was ligated into pMAA as described below.

4.5 SHUFFLING

The genes of interest were each amplified by PCR using external primers shufF+shufR (30 rounds of: 30 s at 94 °C, 30 s at (T_m-5) °C, 90 s at 72 °C) using 4u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs, 1 µg amplicon template in 200 µl. The PCR products for the two genes to be shuffled were mixed such that product genes would have 2-3 amino 'back-mutations' from the earlier round shuffle parent on average, (total 4 µg) and digested in 50 µl with DNAase (0.005 units) for 10 minutes. The digested DNA was separated by gel electrophoresis and fragments between 50 and 70 bp were excised and purified. Assembly of the purified fragments was achieved by a 35 round PCR as above, except using a modified annealing step (65, 62, 59, 56, 53, 50, 47, 44, 41 °C each for 30s) using 1u Pfu Turbo (Agilent), 4 mM dNTPs, in 50 µl. 10% of the product was used as the template for a PCR reaction using the nested primers tevF+tevR to amplify the gene for cloning (30 rounds of: 30 s at 94 °C, 30 s at (T_m-5) °C, 60 s at 72 °C) using 1u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs, in 50 µl. The product was ligated into pMAA as described below.
4.6 LIGATION AND CLONING

Unless otherwise stated in results sections, the following protocol was used as a standard protocol for ligating genes into plasmids (library creation/evolution/shuffling etc). PCR product was purified by spin column (Zymo research), digested with XhoI, NcoI and DpnI (1u each, Fermentas) at 37°C for 1 hr, purified again, and ligated into the XhoI, NcoI digested pMAA vector at 16 °C for 16 h, 1u T4 ligase (NEB), 10 ng of each vector and insert in 10 µl. This was used to transform electro-competent _E. coli_ (lucigen) as per the manufacturer’s instructions. Restriction digests were performed according to the manufacturer’s instructions. DNA was analysed by agarose (0.5%) gel electrophoresis and stained with SybrSafe (1%) against a Hyperladder I (Bioline) standard. DNA was purified after each enzymatic reaction (QIAquick PCR purification kit) according to manufacturer’s instructions. DNA concentrations were determined by measuring the absorption at 260 nm with a Nanodrop spectrophotometer (Thermo Scientific).

 Ligations into JET1.2/Blunt vector were performed according to manufacturer’s instructions. All other ligations were performed using equimolar vector and insert DNA (10 ng) with T4 DNA ligase (1 µl) in ligase buffer (total 10 µl) and incubation for 16 h at 16 °C. When synthetic oligonucleotides were used as insert, 20 ng (in 10 µl H₂O) was prepared by heating the mix of forward and reverse oligonucleotides to 95 °C, cooling to 37 °C over 1 h and then phosphorylating by addition of PNK (3u) and T4 ligase buffer followed by incubation for 1 h at 37 °C.

4.7 PREPARATION OF CHEMO-COMPETENT CELLS

LB (500 ml, containing the appropriate antibiotics) was inoculated with cells (5 ml from overnight inoculum) and incubated at 37 °C (200 rpm) until OD₆₀₀=0.4. The culture was divided into 50 ml aliquots and incubated on ice for 15 min then pelleted (25 min centrifugation at 3500 rpm). Pellets were resuspended in RF1 solution (8 ml per aliquot), incubated on ice again for 15 min and re-pelleted as before. Pellets were resuspended in RF2 solution (3 ml per aliquot), incubated on ice again for 15 min and 50 µl aliquots in PCR tubes were snap-frozen using liquid nitrogen.

4.8 PREPARATION OF ELECTRO-COMPETENT CELLS

LB (1 L, containing the appropriate antibiotics) was inoculated with cells (10 ml from overnight inoculum) and incubated at 37 °C (200 rpm) until OD₆₀₀=0.4. The culture was divided into 50 ml aliquots and incubated on ice for 30 min then pelleted (25 min centrifugation at 3500 rpm). Pellets were resuspended in distilled H₂O (25 ml per aliquot), incubated on ice again for 15 min and re-pelleted. The previous step is repeated, with resuspension in 12.5 ml, 2.5 ml then 100 µl of distilled H₂O +10% glycerol. 50 µl aliquots in PCR tubes were snap-frozen using liquid nitrogen.

4.9 TRANSFORMATION

Chemo-competent cells (50 µl) were incubated with DNA (1-100 ng) on ice for 15 min, heat-shocked for 30s at 42 °C then 60s on ice and incubated in LB (1 ml) for 1 h at 37°C before plating on LB-agar containing the appropriate antibiotic.

Electro-competent cells (50 µl) were incubated with DNA (1-100 ng) on ice for 1 min in an electroporation cuvette. Electroporation of 1800v was followed immediately by the addition of Invitrogen ‘Recovery Medium’ (300 µl) and incubation for 1 h at 37 °C before plating on LB-agar containing the appropriate antibiotic.
BIOCHEMISTRY AND ASSAYS

5.1 PROTEIN EXPRESSION AND PURIFICATION

LB (1 l, containing the appropriate antibiotics) was inoculated with cells (from an overnight inoculum) and incubated at 37 °C and 200 rpm for 4 hours. The cells were induced by addition of L-arabinose (0.2%) and incubated at 30 °C for a further 4 hours. The cells were pelleted and frozen for storage. The cell pellet was thawed and resuspended in 10 ml lysis. Cells were lysed on ice by sonication (2 s on, 10 s off, 5 min) and pelleted. The supernatant was loaded onto a HisTrap (GE Life Sciences, 5ml) column, washed in wash buffer (6 column volumes) and eluted with elution buffer (500 µl fractions). Protein content was monitored by A_280. Fractions containing the protein of interest were pooled and exchanged into TEV Storage buffer using a HiTrap desalting column (GE Life Sciences, 5 ml). 100 µl aliquots in PCR tubes were snap-frozen using liquid nitrogen.

5.2 96- WELL EXPRESSION AND PURIFICATION

In a 96-deepwell plate, LB (1 ml per well, containing the appropriate antibiotics) was inoculated with the cells to be tested and incubated for 12 hours at 30 °C. The cells were induced by addition of 50 µl L-arabinose (0.2% final concentration) and incubated at 25 °C for a further 16 hours. Cells were pelleted, then resuspended in 300 µl PBS + Bugbuster (Novagen) and incubated for 1 hour at 25 °C. Finally, cells were pelleted and the supernatant separated and 700 µl 96w-lysis buffer added to each.

A 96-well TALON resin plate (BD Biosciences) was used for parallel purification of enzymes. Liquid was drawn through using a vacuum manifold (BD Biosciences). The plate was prepared by washing with H_2O (3 times 300 µl per well) then washing with 96w-lysis buffer (3 times 300 µl per well). The cell supernatants (1 ml each) were loaded into the wells and incubated on ice (10 min). The supernatant was then drawn through using the vacuum manifold and washed with 96w-lysis buffer (3 times 300 µl per well) then washed with 96w-wash buffer (6 times 300 µl per well). 300 µl 96w-elution buffer was added to each well and incubated on ice (10 min). The elutant was then drawn through using the vacuum manifold into individual tubes. Protein concentration of each sample was assayed by adding 30 µl eluant to 200 µl Bradford reagent (Sigma) and absorbance was measured by spectrophotometer (Spectromax M5 at 595 nm). The TALON plate was washed with 96w-MES (3 times 300 µl per well) then washed with H_2O (3 times 300 µl per well).

5.3 FLUORESCENCE 96- WELL PLATE SCREENING

In a 96-deepwell plate, LB (1 ml per well, containing the appropriate antibiotics) was inoculated with the cells to be tested and incubated for 12 hours at 30 °C. The cells were induced by addition of 50 µl L-arabinose (0.2% final concentration) and incubated at 25 °C for a further 6 hours. Cells were pelleted, then resuspended in 300 µl PBS + Bugbuster (Novagen) and incubated for 1 hour at 25 °C. Finally, cells were pelleted and the supernatant was assayed in a Spectramax M5 (Molecular Devices, excitation 414 nm, observation 475 nm and 525 nm).

5.4 FLUORESCENCE IN VITRO KINETICS

CFP-YFP substrate was thawed overnight at 4 °C. Unless otherwise stated, 1 µM substrate was added to 1, 2, 4 and 8 µM purified enzyme in 200 µl TEV Storage buffer + 1 mM DTT in a 96-well plate. Solutions were assayed by spectrophotometer (Spectromax M5 - excitation 414 nm, observation 475 nm and 525 nm) reading every minute for 12 hours at 25 °C for 12 hours. Concentration of product was calculated using Equation G-1. Kinetic traces were fit first to simple Michaelis-Menten kinetics, Equation G-2.
\[ [P] = [S] \frac{R_0 - R_t}{R_0 - R_\infty} \]

**Equation G-1 | Product concentration**

\([P]\) = Concentration of product

\([S]\) = Concentration of substrate

\(R_0\) = FRET ratio of substrate (at \(t=0\))

\(R_\infty\) = FRET ratio of product (at \(t=\infty\))

\(R_t\) = FRET ratio at time \(t\)

\[ [P] = [P]_{\text{max}} - e^{-[E]k_{\text{cat}}/K_M}t \]

**Equation G-2 | Michaelis Menten interpolation (also C-1)**

\([P]\) = Concentration of product

\([P]_{\text{max}}\) = Final concentration of product (should be = starting substrate)

\(A_1\) = Jump amplitude (ie. how much turnover is performed in first kinetic step)

\(A_2\) = Second jump amplitude (should = \([P]_{\text{max}}-A_1\))

\([E]\) = Concentration of enzyme

\(k_{\text{cat}}/K_M\) = Rate of overall reaction

\(k_{\text{obs1}}\) = Rate of first kinetic step

\(k_{\text{obs2}}\) = Rate of second kinetic step

\(t\) = Time

### 5.5 Fluorescence Activated Cell Sorting (FACS)

Transformed cells were scraped off the agar plate in LB (2 ml), 4 µl of which was used to inoculate fresh culture in a 96-deepwell plate, LB (1 ml per well, containing the appropriate antibiotics) After 4 hours at 37 °C, 4 µl of this culture was used to inoculate 1 ml fresh LB (+antibiotic) and incubated for 4 hours at 30 °C. The cells were induced by addition of L-arabinose (0.2%) and incubated at 20 °C overnight (16 hours). For FACS a MoFlo MLS (BeckmanCoulter) was used. Samples were diluted 50:50 with PBS and flowed at 1000-3000 cells per second (70 µm nozzle, 60 psi (4 bar), 514 nm ssc trigger). Excitation at 406 nm (100 mW Krypton laser) was monitored using filters at 470±10 nm (FL4) and 546±5 nm (FL5). Excitation at 514 nm (200 mw Argon laser) was monitored by filters at 546±5 nm (FL1). Cells expressing pMAA + pCtY and pMAA-TEV + pCtY were used as negative and positive controls, respectively. Sorting was performed by gating firstly above FL1 to ensure expression and secondly on the FL5/FL4 to select the 1% of cells with the lowest FRET ratio. Cells were sorted into LB (containing 100 µg/ml ampicillin and 37 µg/ml chloramphenicol) and incubated overnight at 37 °C before being harvested.

### 5.6 Gel Band Densitometry

In a 96-deepwell plate, LB (1 ml per well, containing the appropriate antibiotics) was inoculated with the cells to be tested and incubated for 12 hours at 30 °C. The cells were induced by addition of L-arabinose (0.2%) and incubated at 25 °C for a further 4 hours. Cells pelleted by centrifugation and then resuspended in PBS (250 µl) + Bugbuster (Novagen) and incubated for 1 hour at 25 °C. Cells were pelleted and the supernatant (10 µl) was run on a polyacrylamide gel (1.5%). Gels were stained with InstantBlue (Expedion) according the manufacturer’s instructions. Stained gels were photographed on a lightbox and the image analysed with ImageJ.
5.7 DIFFERENTIAL SCANNING FLUORIMETRY

Sypro Orange (Sigma) was added to purified protein (20 µl of 100 mg/ml). Samples were analysed in a real time PCR machine (Corbett Research 2plex HRM) by increasing temperature from 25 °C to 95 °C in 0.5 °C steps over 70 minutes (i.e. increasing 1 °C/min). The fluorescence curves were fitted to Equation G-4 to calculate the temperature at which half of the protein is melted (T_m).

\[
Y = Y_{\text{min}} + \frac{Y_{\text{max}} - Y_{\text{min}}}{1 + \left(\frac{T}{T_m}\right)^{-h}}
\]

Equation G-4 | Thermal denaturation sigmoidal curve
Y = Fluorescence
Ymin = Minimum fluorescence
Ymax = Maximum fluorescence
T = Temperature
T_m = Melting temperature (infection point)
h = Hill coefficient (unfolding cooperativity)
6 BIOINFORMATICS AND DATA ANALYSIS

6.1 PA CLAN ALIGNMENT AND PHYLOGENY

An alignment based on structural homology to TEV protease (PDB 1lvm) was generated using DALIprotein\textsuperscript{254}. From this primary homology assessment, sequences lacking known structures were added based on amino acid sequence similarity to several proteases from the phylogeny (Table G-5) using BLASTp search of non-redundant sequences. Columns with >90% gaps were removed (Gapstreeze\textsuperscript{313} online). The amino acid sequence alignment was used to generate a maximum likelihood phylogeny in MEGA5\textsuperscript{314}.

<table>
<thead>
<tr>
<th>BLAST QUERY</th>
<th>PDB</th>
<th>NO. SEQUENCES ADDED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tobacco etch</td>
<td>1lvm</td>
<td>75</td>
</tr>
<tr>
<td>Astrovirus</td>
<td>2w5e</td>
<td>14</td>
</tr>
<tr>
<td>Rhinovirus 2A*</td>
<td>2hrv8</td>
<td>14</td>
</tr>
<tr>
<td>Sesbania mosaic</td>
<td>1zyo</td>
<td>12</td>
</tr>
<tr>
<td>Rhinovirus 3C*</td>
<td>1cqqA</td>
<td>10</td>
</tr>
<tr>
<td>Alkaline Protease</td>
<td>3cp7</td>
<td>8</td>
</tr>
<tr>
<td>Norovirus</td>
<td>2fyq</td>
<td>6</td>
</tr>
<tr>
<td>Achromobacter</td>
<td>1arc</td>
<td>5</td>
</tr>
<tr>
<td>Hepatitis C</td>
<td>2obq</td>
<td>4</td>
</tr>
<tr>
<td>Human BB fragment</td>
<td>1rtk</td>
<td>3</td>
</tr>
<tr>
<td>Equine arteritis</td>
<td>1mbm</td>
<td>3</td>
</tr>
<tr>
<td>Hepatitis A</td>
<td>1hav</td>
<td>2</td>
</tr>
</tbody>
</table>

Table G-5 | Sequences added to DALI alignment by BLASTp

The Protein sequences of each query were used to search the non-redundant sequence data base by BLASTp. * Rhinovirus has two paralogous copies of its protease. Protease 2A is smaller than 3C and has a different substrate specificity.

6.2 HIGH THROUGHPUT SEQUENCING AND DATA ANALYSIS

The plasmids to be sequenced by 454 were amplified by PCR using tevF+tevR (10 rounds of: 30 s at 94 °C, 30 s at (T\textsubscript{M}-5) °C, 60 s at 72 °C) using 0.4u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs, 100 ng template in 20 µl. This product was purified to remove primer and used as the template for a second PCR to add the 454 recognition sequences seqXF+seqXR, each population using the appropriate multiplex identifier sequences (see Table G-4) (10 rounds of: 30 s at 94 °C, 30 s at (T\textsubscript{M}-5) °C, 60 s at 72 °C) using 0.4u Pfu Turbo (Agilent), 1 µM each primer, 4 mM dNTPs in 20 µl. The products were mixed in equimolar ratios and sequenced using a GS Junior (Roche) by the Cambridge University Biochemistry Department Sequencing Facility. Sequencing generated forward reads for the first 400nt and reverse reads for the last 400nt of the 800nt total amplicon. Reads were aligned to the TEV\textsuperscript{56} DNA sequence using AVA aligner (Roche). Columns with >90% gaps were removed (Gapstreeze online) and remaining gaps replaced with 'n's in order to prevent frameshifts upon translation to protein. The alignments were used to generate a quantitative consensus (Jalview) which was exported to Excel (Microsoft) for analysis. Enrichment scores were calculated for each codon in the gene as the proportion of non-wt residue at each position after selection divided by the proportion before selection. To reflect the sensitivity of sequencing and to avoid dividing by zero, the minimum frequency of non-wt residues is set as 0.001. Protein regions were identified using PyMol\textsuperscript{313} for nucleophile and triad shells, IUPred\textsuperscript{306} for disordered regions, GetArea\textsuperscript{316} for solvent accessibility, Jalview for conserved regions, and FoldX\textsuperscript{317} for regions sensitive to alanine mutation in silico. Statistical significance was calculated by heteroscedastic t-test (since
equal variance could not be guaranteed) when comparing the average effect of a subset of mutations in a protein region to the whole protein. Similarly, $\chi^2$ was used when comparing the number of beneficial mutations in a protein region to the whole protein. A 3-category $\chi^2$ test was used for analysing the distribution of the 15 mutations accumulated in the DE lineage.

6.3 FOLDX

PDB file 1lvm was edited to remove the second chain and subjected to iterative rounds of the FoldX "repair" routine until energy levels reached a stable minimum. This PDB file was then used for the determination of total energies of all mutants in the serine lineage (and corresponding revertants) by FoldX. Two structures were generated for each mutant, and average values used for analysis. As the sequence of TEV Cys used in this work differs from 1LVM, the $\Delta$G of TEV Cys was calculated and used as reference (set to zero). A single PDB file for each mutant was then used in the alanine scanning routine to calculate the contribution of each position to global stability.

6.4 FACS MUTATION TOLERANCE ANALYSIS

The FL5/FL4 ratio of cells above the FL1 threshold was recorded (Summit 4.3) and exported to Excel. The histogram was fitted to Equation G-5 to calculate the relative populations of active and inactive variants.

$$y = \frac{A_1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu_1}{\sigma_1}\right)^2} + \frac{A_2}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

Equation G-5 | Bimodal normal distribution

$y = $ frequency
$x = $ FRET ratio
$[A1] = $ Size of active population
$[A2] = $ Size of inactive population
$\sigma1 = $ standard deviation of active population
$\sigma2 = $ standard deviation of inactive population
$\mu1 = $ mean FRET ratio of active population
$\mu2 = $ mean FRET ratio of inactive population

When the sequence before mutagenesis is a single gene and the library afterwards has been sequenced, the proportion of variants that retained wt-like activity was fit to Equation D-1 to determine the tolerance to random amino acid substitution. When a library was generated from a mixed population it is not possible to know which variants acquired which number of mutations and so $\tau_{sub}$ cannot be calculated. In these cases tolerance is simply given at the mutation rate ($\tau_{0.5}$ for $\mu$=0.5% and $\tau_{1.5}$ for $\mu$=1.5%).

$$S = \sum_{n=0} f_n (\tau_{sub})^n = f_0(\tau_{sub})^0 + f_1(\tau_{sub})^1 + f_2(\tau_{sub})^2 \ldots + f_n(\tau_{sub})^n$$

Equation G-6 | The probability of enzyme activity retention after random mutation
(also D-1)
$S = $ the fraction of enzyme variants displaying wt-like activity
$n = $ number of mutations
$f_n = $ the fraction of enzyme variants with n mutations
$\tau_{sub} = $ the probability that a random mutation will be neutral

Nb. Frameshift mutations will cause underestimation $\tau_{sub}$, however, these were rare enough that they made no difference to these estimates.

6.5 PROCRASTINATION

Avoiding work was mostly achieved through the news sites: BBC, Ars Technica and Nature news. On slow news days Facebook was occasionally required. For a full list see browser history.
Chapter H - References


176


265. Kaneko, K. Evolution of Robustness. 3–8


Chapter H - References


313. Gapstreeze. at <http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html>


315. PyMol at <http://www.pymol.org/>
