

Ancestral sequence reconstruction as an accessible tool for the engineering of biocatalyst stability

Submitted by Adam Christopher Thomas to the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Biological Sciences

In December 2018

This thesis is available for library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature:

Abstract

Synthetic biology is the engineering of life to imbue non-natural functionality. As such, synthetic biology has considerable commercial potential, where synthetic metabolic pathways are utilised to convert low value substrates into high value products. High temperature biocatalysis offers several system-level benefits to synthetic biology, including increased dilution of substrate, increased reaction rates and decreased contamination risk. However, the current gamut of tools available for the engineering of thermostable proteins are either expensive, unreliable, or poorly understood, meaning their adoption into synthetic biology workflows is treacherous. This thesis focuses on the development of an accessible tool for the engineering of protein thermostability, based on the evolutionary biology tool ancestral sequence reconstruction (ASR). ASR allows researchers to walk back in time along the branches of a phylogeny and predict the most likely representation of a protein family's ancestral state. It also has simple input requirements, and its output proteins are often observed to be thermostable, making ASR tractable to protein engineering.

Chapter 2 explores the applicability of multiple ASR methods to the engineering of a carboxylic acid reductase (CAR) biocatalyst. Despite the family emerging only 500 million years ago, ancestors presented considerable improvements in thermostability over their modern counterparts. We proceed to thoroughly characterise the ancestral enzymes for their inclusion into the CAR biocatalytic toolbox.

Chapter 3 explores why ASR derived proteins may be thermostable despite a mesophilic history. An *in silico* toolbox for tracking models of protein stability over simulated evolutionary time at the sequence, protein and population level is built. We provide considerable evidence that the sequence alignments of simulated protein families that evolved at marginal stability are saturated with stabilising residues. ASR therefore derives sequences from a dataset biased toward stabilisation.

Importantly, while ASR is accessible, it still requires a steep learning curve based on its requirements of phylogenetic expertise. In chapter 4, we utilise the evolutionary model produced in chapter 3 to develop a highly simplified and accessible ASR protocol. This protocol was then applied to engineer CAR enzymes that displayed dramatic increases in thermostability compared to both modern CARs and the thermostable AncCARs presented in chapter 2.

Table of Contents

<i>Abstract</i>	2
<i>Acknowledgements</i>	14
<i>Preface</i>	16
Chapter 1	17
Introduction	17
<i>1.1 Enzymes in Synthetic biology</i>	18
1.1.1 Defining synthetic biology	18
1.1.2 Enzymes	23
1.1.3 Enzyme engineering	30
1.1.4 Moving synthetic biology beyond proof of concept	37
<i>1.2 Thermostable enzymes for the advancement of industrial synthetic biology</i>	39
1.2.1 Defining enzyme stability	39
1.2.2 Most proteins are marginally stable	43
1.2.3 Nature's strategies for protein stabilisation	47
1.2.4 Stable enzymes in synthetic biology	49
1.2.5 Obtaining stable enzymes	56
1.2.6 Optimising the engineering of stable enzymes for synthetic biology	62
<i>1.3 ASR as an efficient tool to engineer protein stability</i>	70
1.3.1 Methodologies for ancestral sequence reconstruction	70
1.3.2 Probing patterns in protein evolution with ASR - thermostability	83
1.3.3 ASR as an engineering tool	86
<i>1.4 Aims and Objectives</i>	91
1.4.1 Chapter 2	91
1.4.2 Chapter 3	92
1.4.3 Chapter 4	92
<i>1.5 – Addendum</i>	93

Chapter 2	96
Thermostable carboxylic acid reductases generated by ancestral sequence reconstruction	96
2.1 <i>Authors</i>	97
2.2 <i>Preface</i>	97
2.3 <i>Abstract</i>	98
2.4 <i>Introduction</i>	98
2.5 <i>Results</i>	102
2.5.1 Ancestral reconstruction of CARs produces functional enzymes	102
2.5.2 AncCARs have a broad substrate range	105
2.5.3 Ancestral CARs show dramatic increases in stability	111
2.5.4 AncCARs vary in their loop-based properties	114
2.6 <i>Discussion</i>	116
2.7 <i>Conclusion</i>	120
2.8 <i>Methods</i>	120
2.9 <i>Supplementary Figures</i>	126
3.10 <i>Supporting information</i>	148
Chapter 3	149
Survivor bias drives overestimation of stability in ancestral proteins	149
3.1 <i>Authors</i>	150
3.2 <i>Preface</i>	150
3.3 <i>Abstract</i>	151
3.4 <i>Introduction</i>	151
3.5 <i>Hypotheses – tenets of survivor bias</i>	155
3.5.1 Tenet 1: A mutation’s contribution to protein stability is derived from a normal distribution with a negative mean	155
3.5.2 Tenet 2: The majority of proteins are marginally stable	155
3.5.3 Tenet 3: Contemporary proteins contain fewer significantly destabilizing amino acids	156

3.5.4 Tenet 4: Ancestors sample from a stabilizing mutation space, despite a destabilizing global mutational landscape	156
<i>3.6 Methods</i>	157
<i>3.7 Results and discussion</i>	161
3.7.1 Stability can be tracked as the protein evolves within PESST.	161
3.7.2 A stability threshold biases the distribution of Δr , α values in the evolving dataset	165
3.7.3 Marginality causes overestimation of stability in ancestral sequences	168
3.7.4 Marginality causes overestimation of stability in consensus sequences	174
<i>3.8 Conclusion and perspective</i>	176
<i>3.9 Supplementary methods</i>	178
<i>3.10 Supplementary figures</i>	181
<i>3.11 Supporting information</i>	204
Chapter 4	205
Simplified ancestral sequence reconstruction – an accessible tool for engineering protein stability.	205
<i>4.1 Authors</i>	206
<i>4.2 Preface</i>	206
<i>4.3 Abstract</i>	206
<i>4.4 Introduction</i>	208
<i>4.5 Results</i>	212
4.5.1 Ancescon produces ancestors with high stability in simulations	212
4.5.2 High weighted node balance is a strong indicator of stability in simulations	214
4.5.3 sASR produces functional CAR enzymes	217
4.5.4 AspCARs are thermostable CAR variants	221
<i>4.6 Discussion</i>	223
<i>4.7 Conclusion</i>	226
<i>4.8 Methods</i>	226
<i>4.9 Supplementary figures</i>	231

4.10 Supporting information	242
Chapter 5	243
General discussion	243
5.1 ASR's place as an engineering tool	244
5.2 Protein stabilization with ASR	245
5.3 Accessible protein engineering with ASR	248
5.4 Developments in the modelling of protein evolution	250
5.5 Conclusions	250
Chapter 6	252
Bibliography	252
Chapter 7	299
Appendices	299
7.1 Permissions for figure 5	300
7.2 Permissions for figure 6	300
7.3 Characterization of carboxylic acid reductases in the toolbox of synthetic chemistry	300

Table of Figures

FIGURE 1 – GENETIC CIRCUITS ARE ANALOGOUS TO SIMPLE ELECTRONIC CIRCUITS	20
FIGURE 2 - THE CURRENT STATE OF DNA READING AND DNA WRITING	22
FIGURE 3 - ENZYMES SIGNIFICANTLY LOWER THE ΔG^\ddagger OF A REACTION	26
FIGURE 4 - A TWO DIMENSIONAL PROTEIN FITNESS LANDSCAPE	33
FIGURE 5 - DIRECTED EVOLUTION IS THE ITERATIVE UPHILL WALK TO OPTIMA IN FUNCTIONAL SPACE	35
FIGURE 6 - MUTATIONS IN PROTEINS ARE ON AVERAGE DESTABILISING	46
FIGURE 7 - PDB 1J2V – CUTA1 FROM <i>PYROCOCCUS HORIKOSHII</i>	49

FIGURE 8 - A CONSENSUS SEQUENCE GENERATED FROM AN ALIGNMENT OF INSULIN HOMOLOGUES	67
FIGURE 9 - A HYPOTHETICAL PHYLOGENETIC TREE	72
FIGURE 10 - A SINGLE MCMC SIMULATION APPROACHING CONVERGENCE IN PROBABILITY SPACE	78
FIGURE 11 - MARGINAL AND JOINT LIKELIHOOD ANCESTRAL RECONSTRUCTIONS ARE NOT EQUIVOCAL	80
FIGURE 12 - THE STABILITY OF EXTANT CAR ENZYMES ANALYSED IN FINNIGAN <i>ET AL.</i> (2017)	95
FIGURE 13 - BAYESIAN INFERENCE OF ACTINOMYCETE CAR PHYLOGENY AND ASR	105
FIGURE 14 - ANCCARS SHOW ACTIVITY ON CANONICAL CAR SUBSTRATES	107
FIGURE 15 – HOMOLOGY MODELS SUGGEST SLIGHT DISORDERING OF ANCCAR ACTIVE SITES	110
FIGURE 16 - ANCCARS ARE THERMOSTABLE ENZYMES	113
FIGURE 17 - ANCCAR TOLERANCE TO SOLVENTS LOOP DEPENDENT ENVIRONMENTAL FACTORS	115
FIGURE 18 - PESST EVOLUTIONARY ALGORITHM PSEUDOCODE	158
FIGURE 19 - MEAN STABILITY OF THE POPULATION SIMULATED IN PESST TENDS TOWARD ϵ DURING EVOLUTION	162
FIGURE 20 - MEAN STABILITY CHANGE OF THE EVOLVING DATASET TENDS TOWARD THE MEAN OF THE INITIAL MATRIX (Δ) FROM WHICH THE PROTEIN IS DEFINED	164
FIGURE 21 - IMPOSITION OF A STABILITY THRESHOLD LEADS TO EQUILIBRATION OF MEAN PROTEIN STABILITY, AND POSITIVE-BIAS IN THE POPULATION'S $\Delta_{R,A}$ DISTRIBUTION	167
FIGURE 22 - SURVIVOR BIAS DRIVES INCREASING STABILITY IN ANCESTRAL PROTEINS	173
FIGURE 23 - SURVIVOR BIAS DRIVES THE STABILIZATION OF CONSENSUS SEQUENCES	175
FIGURE 24 - ANCESCON INTRODUCES STABILITY BIAS IN PESST SIMULATIONS OF PROTEIN EVOLUTION	214
FIGURE 25 - NODE WEIGHTED BALANCE IS AN ACCURATE PREDICTOR OF HIGH ANCESTOR STABILITY	216
FIGURE 26 - TREE OUTPUT FROM ANCESCON BUILT WITH THE WEIGHBOUR METHOD	218
FIGURE 27 - ASPCARS ARE CAR ENZYMES WITH ACTIVITY ON CANONICAL SUBSTRATES	220
FIGURE 28 - ASPCARS ARE HIGHLY THERMOSTABLE ENZYMES	222
FIGURE 29 - CALCULATING WEIGHTED BALANCE	227

Table of Tables

TABLE 1 – EC CLASSIFICATION OF THE SIX MAJOR ENZYME SUBFAMILIES	24
TABLE 2 - CLASSIFICATION OF AMINO ACIDS BY HYDROPHOBICITY	40
TABLE 3 - INNOVATIONS FOR THE STABILISATION OF PROTEINS IN THERMOPHILES	48
TABLE 4 - A NON-EXHAUSTIVE LIST OF FREELY AVAILABLE, COMMONLY USED MSA TOOLS FOR PROTEIN SEQUENCES	66
TABLE 5 - COMMONLY USED SUBSTITUTION MATRICES DERIVED FROM LARGE DATABASES OF ALIGNED PROTEINS	73
TABLE 6 - COMMONLY UTILISED ALGORITHMS FOR INFERENCE OF MAXIMUM LIKELIHOOD PHYLOGENIES	76
TABLE 7 - EMPIRICAL ANCESTRAL RECONSTRUCTION ALGORITHMS	83
TABLE 8 – ANCCAR CO-FACTOR KINETICS	108
TABLE 9 – ANCCAR SUBSTRATE KINETICS	109
TABLE 10 - DEFAULT PARAMETERS IN PESST	159
TABLE 11 - STABILITY INCREASES IN COMMERCIALY IMPORTANT PROTEINS ENGINEERED BY ASR, AND THEIR POTENTIAL UTILIZATION IN CURRENT BIOINDUSTRIAL WORKFLOWS	210
TABLE 12 - ASPCAR KINETICS	220

Table of Supplementary Figures

SUPPLEMENTARY FIGURE 1 - CURRENT PROPOSED CAR REACTION MECHANISM	126
SUPPLEMENTARY FIGURE 2 - ALIGNMENT OF ANCCAR PROTEIN SEQUENCES	128
SUPPLEMENTARY FIGURE 3 – DIFFERENCE BETWEEN ASR ALGORITHM OUTPUTS IS NOT ALTERNATIVE SAMPLING OF POSTERIOR PROBABILITY TABLES.	129
SUPPLEMENTARY FIGURE 4 - REGIONS OF SIGNIFICANT SEQUENCE VARIATION BETWEEN ANCCARS AND 5MST/5MSP	130
SUPPLEMENTARY FIGURE 5 - ALL CAR ENZYMES ARE SOLUBLE	131
SUPPLEMENTARY FIGURE 6 - ANCCARS ARE LESS PROTEASE SENSITIVE THAN EXCARS	132
SUPPLEMENTARY FIGURE 7 - TRIS INHIBITS ANCCAR ENZYMES	133
SUPPLEMENTARY FIGURE 8 - ANCCARS HAVE EQUIVALENT SUBSTRATE RANGES	135

SUPPLEMENTARY FIGURE 9 - ANCCAR KINETICS IN ATP AND NADPH	136
SUPPLEMENTARY FIGURE 10 - ANCCAR CARBOXYLIC ACID SUBSTRATE KINETICS	140
SUPPLEMENTARY FIGURE 11 - ANCCAR ACTIVE SITE COMPARISONS TO 5MST	142
SUPPLEMENTARY FIGURE 12 - ANCCAR STABILITY IN NA ₂ CO ₃	143
SUPPLEMENTARY FIGURE 13 - NADPH STANDARD CURVE	144
SUPPLEMENTARY FIGURE 14 - REPRESENTATIVE Δ MATRIX GENERATED IN PESST SIMULATIONS	182
SUPPLEMENTARY FIGURE 15 - IMPLEMENTATION SCHEME OF BIFURCATIONS IN PESST	182
SUPPLEMENTARY FIGURE 16 - TRANSITION RATES AT EACH SITE DERIVED FROM A MODIFIED LG MODEL IMPLEMENTED INTO PESST	185
SUPPLEMENTARY FIGURE 17 - FOUR INDEPENDENT RATE CATEGORIES DEFINED BY THE MEDIAN VALUES OF FOUR QUANTILES FROM 10,000 SAMPLES OF THE GAMMA DISTRIBUTION	187
SUPPLEMENTARY 18 - SUMMARY DATA FOR RUNS PRESENTED IN FIGURE 1B AND C	188
SUPPLEMENTARY FIGURE 19 - DURING SIMULATED EVOLUTION, THE DISTRIBUTION OF $\Delta_{R,A}$ VALUES IN EVOLVING DATA APPROACHES THE DISTRIBUTION OF $\Delta_{R,A}$ VALUES IN THE STABILITY MATRIX	189
SUPPLEMENTARY FIGURE 20 - SIMULATIONS WITH IMPOSED Ω REACH EQUILIBRIUM SLIGHTLY ABOVE Ω	190
SUPPLEMENTARY FIGURE 21 - DURING SIMULATED EVOLUTION WHERE $\Omega > \epsilon$, THE DISTRIBUTION OF $\Delta_{R,A}$ VALUES IN EVOLVING DATA IS POSITIVELY BIASED COMPARED TO THE DISTRIBUTION OF $\Delta_{R,A}$ VALUES IN THE GLOBAL STABILITY MATRIX Δ	191
SUPPLEMENTARY FIGURE 22 - COMPARING HEAT MAPS OF $\Delta_{R,A}$ VALUES IN SIMULATIONS THAT HAVE NO Ω IMPOSED TO SIMULATIONS WITH IMPOSED Ω CLEARLY SHOW THE POSITIVE SHIFTED BIAS TOWARD STABILITY UNDER Ω	192
SUPPLEMENTARY FIGURE 23 - DISTRIBUTION OF $\Delta_{R,A}$ VALUES IN EVOLVING DATA DOES NOT CONVERGE WITH $\Delta_{R,A}$ VALUES IN GLOBAL STABILITY MATRIX Δ WHEN EVOLVED AT MARGINALITY	193
SUPPLEMENTARY FIGURE 24 - SIMULATIONS WHERE GLOBAL STABILITY MATRIX $\Delta_{r, \alpha} = 0$ AND $\Delta_{r, \alpha} = 1$ ARE ABLE TO EXPLORE A WIDE STABILITY SPACE DUE TO RELEASE OF SELECTIVE PRESSURE	194
SUPPLEMENTARY FIGURE 25 - REPRESENTATIVE PHYLOGENIES GENERATED FROM PESST SIMULATIONS USED FOR FIGURES 22 AND 23	196

SUPPLEMENTARY FIGURE 26 - REPRESENTATIVE SIMULATIONS SHOWING THAT RECONSTRUCTED ANCESTORS EXPLORE A HIGHER STABILITY SPACE THAN THAT OF THEIR EVOLUTIONARY HISTORY WHEN A BIDIRECTIONAL SELECTIVE PRESSURE IS PRESENT	198
SUPPLEMENTARY FIGURE 27 - PHYLOGENY FOR SIMULATION THAT EVOLVES FROM HIGH STABILITY TO A THRESHOLD	199
SUPPLEMENTARY FIGURE 28 - CONSENSUS SEQUENCES FROM SIMULATIONS INITIATED AT <i>T_{0high}</i> ARE SIGNIFICANTLY MORE STABLE THAN THOSE FROM <i>T₀Ω + 5; Ω + 25</i>	200
SUPPLEMENTARY FIGURE 29 - PESST SIMULATIONS USED IN ANALYSIS OF ANCESCON RECONSTRUCTIONS	232
SUPPLEMENTARY FIGURE 30 - NODES PREDICTED BY SASR ARE CONSIDERABLY BIASED TOWARD STABILITY. HIGH STABILITY NODES CAN BE PREDICTED WITH NODE BALANCE	236
SUPPLEMENTARY FIGURE 31 - RAW TREE OUTPUT FROM ANCESCON	237
SUPPLEMENTARY FIGURE 32 - ASPCAR SEQUENCES RECONSTRUCTED WITH ANCESCON	238
SUPPLEMENTARY FIGURE 33 - SDS-PAGE GELS OF ASPCAR-A43 AND ASPCAR-A50	239
SUPPLEMENTARY FIGURE 34 - SATURATION CURVES FOR ASPCAR KINETICS	241

Table of Supplementary Tables

SUPPLEMENTARY TABLE 1 - ROOT MEAN SQUARED VALUES OF ALPHA CARBON ATOM DISPLACEMENT IN ANCCAR PROTEIN MODELS - SHOWING GOOD FIT OF DATA	145
SUPPLEMENTARY TABLE 2 - ANCCAR KINETICS ON CARBOXYLIC ACID SUBSTRATES	146
SUPPLEMENTARY TABLE 3 - ANCCAR TOLERANCE TO VARIOUS SOLVENTS	147
SUPPLEMENTARY TABLE 4 - ADDITIONAL PARAMETERS HANDLED BY PESST	201
SUPPLEMENTARY TABLE 5 - SEEDS FOR THE SIMULATIONS USED FOR FIGURES 19 AND 20	202
SUPPLEMENTARY TABLE 6 - SEEDS FOR THE SIMULATIONS USED FOR FIGURE 21	203
SUPPLEMENTARY TABLE 7 - SEEDS FOR FIGURES 22 AND 23	203
SUPPLEMENTARY TABLE 8 - SIMULATION P-VALUES IN FIGURE 22	204
SUPPLEMENTARY TABLE 9 – SEEDS FOR PESST SIMULATIONS	242

List of Abbreviations

DNA – Deoxyribonucleic Acid
ASR – Ancestral Sequence Reconstruction
CAR – Carboxylic acid reductase
PESST – Protein Evolution Simulation with Stability Tracking
MCMC – Markov Chain Monte Carlo
MSA – Multiple Sequence Alignment
ATP – Adenosine triphosphate
NADPH - Nicotinamide adenine dinucleotide phosphate reduced
LG – Le and Gascuel
SDS-PAGE – Sodium dodecyl sulfate – polyacrylamide gel electrophoresis
BLAST – Basic Local Alignment Software Tool
EC – Enzyme Classification
NAPD - Nicotinamide adenine dinucleotide phosphate
PCR – Polymerase Chain Reaction
GFP – Green Fluorescent Protein
PDB – Protein Database
ACS – Acetyl-CoA Synthetase
GCS - Glycolyl-CoA Synthetase
PROSS – Protein Repair One Stop Shop
FRESCO – Framework for Rapid Enzyme Stabilisation by Computational libraries
DMSO – Dimethyl Sulfoxide
CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats
MAFFT – Multiple Alignment using Fast Fourier Transform
MUSCLE – MULTiple Sequence Comparison by Log-Expectation
T-Coffee – Tree-based Consistency Objective Function for Alignment Evaluation
JTT – Jones Taylor Thornton
WAG – Whelan and Goldman
SH – Shimohaira and Hasegawa
aLRT – approximate Likelihood-Ratio Test
PHYMLIP – PHYLogeny Inference Package
PAUP – Phylogenetic Analysis Using Parsimony
PhyML – Phylogenies by Maximum Likelihood
RAxML – Randomized Axelerated Maximum Likelihood
MCMCMC – Metropolis Coupled Markov Chain Monte Carlo
ML – Maximum Likelihood
PAML – Phylogenetic Analysis by Maximum Likelihood
FastML – Fast algorithm for Maximum Likelihood
Ancescon – ANCEStral sequence reconstruction
LUCA – Last Universal Common Ancestor
EF-Tu – Elongation Factor-Themounstable

NDK – Nucleoside Diphosphate Kinase
REAP – Reconstructed Evolutionary Adaptive Paths
FRET – Förster Resonance Energy Transfer
HLD – Haloalkane Dehalogenase
CYP3 – Cytochrome P450 family 3
AMP – Adenine monophosphate
NADP⁺ - Nicotinamide adenine dinucleotide phosphate
ANL – Acyl-CoA ligase, Non-ribosomal peptide synthetase, Luciferase
PPT – Phosphopantetheine
WAG+I+G – Whelan and Goldman + Invariant + Gamma
AncCAR – Ancestral Carboxylic acid reductase
rmsd – Root Mean Squared Deviation
HEPES – 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
Tris - Tris(hydroxymethyl)aminomethane
ExCAR – Extant Carboxylic Acid Reductase
YASARA – Yet Another Scientific Artificial Reality Application
DSF – Differential Scanning Fluorimetry
LB – Luria-Bertani broth
FF – FastFlow
IPTG - Isopropyl β -D-1-thiogalactopyranoside
MOPS - 3-(N-morpholino)propanesulfonic acid
MES - 2-(N-morpholino)ethanesulfonic acid
KOH – Potassium hydroxide
Å – Angstrom
WT – Wild Type
PON – Paraoxonase
LG+I+G – Le and Gascuel + Invariant + Gamma
sASR – simplified Ancestral Sequence Reconstruction
ivSc – in vitro *Saccharomyces cerevisiae*
KARI – Ketol-Acid Reducto-Isomerase

Acknowledgements

I would like to first thank my supervisors: Nicholas Harmer and Mark van der Giezen for their extensive support throughout this entire process. I appreciate that their input has allowed me to grow considerably as a scientist over the last 4 years. I appreciate that ideas were always allowed to flow freely, and many excellent scientific discussions were had in the synthesis of work presented in this thesis. Finally I appreciate the encouragement and freedom I have received in to progress my future career during my PhD tenure.

Particularly I would like to thank Nic for his exceptional support throughout this process, his patience, his mentorship and his continued enthusiasm for each new project that evolved out of an initial experiment that worked, even though “in hindsight we should have done it *very differently*”.

I would also like to thank the BBSRC for their funding for this project, and the SWBio DTP, a rapidly growing community of excellent early career researchers. Particularly, I would like to thank Samantha Southern for providing her continued support to this community, and for organising the invaluable thesis bootcamp.

Thank you for Emily Leproust of Twist Bioscience for recommending me for such an exciting and valuable internship at the company, and to Maria Ramirez and Paddy Finn for your mentorship while undertaking the placement.

I would like to thank the Harmer group members Sumita, Rhys and Alice, and past members Will and Erin for always making labs fun, and for always being able to help out whenever help was needed. Also, thank you to Will for teaching me many essential lab techniques in my first year, and for allowing me to co-author on my first published research paper.

I would like to thank Laurence Higgins, Emily Shepperd, Jamie Gilman and Laura Reffo for their constant friendship along this crazy ride – for great food at the Higgins’, for countless impromptu beers, for rocket launch parties, for 7 am squash matches, and for all just being nothing but supportive!

I would also like to thank Kyran and Hayley at the Pursuit of Hoppiness, Exeter, for providing wonderful beer, and being even more wonderful friends!

Thank you to my parents, Kim, Chris, and Sally for their love, support, and constant encouragement.

And finally, thank you does not cut it, but thank you Annie. You have been my rock, my partner, and my closest friend throughout my PhD. Thank you for your constant patience, even with all of the late nights. Thank you for driving me to the lab at 2 am to check experiments. Thank you for constantly keeping me on track, even from 6000 miles away during my internship, and from 200 miles away in my last year. Thank you for never questioning my passions, and always pushing me to be better. Thank you for holding me up, and for grounding me. And thank you for being with me as we carve this path. I can't wait to start our next chapter.

Preface

Chapters 3, 4 and 5 of this thesis are reformatted manuscripts that have either been submitted to scientific journals, or are prepared for submission to scientific journals. Each of these chapters will contain a preface explaining author contributions and the publication status of the works. Bibliographies for each manuscript are collated in the bibliography chapter.

Chapter 1 consists an extended introduction, introducing and detailing each manuscript's context in their broader fields. Some repetition between the introduction and the introduction and the subsequent chapters may occur as each manuscript is individually introduced.

An additional publication by the thesis author is presented in the appendices (Finnigan *et al.*, 2017). The foundational work in this publication preludes chapters 3 and 5, and has been included as a complete manuscript in PDF format. A brief description of this work is presented as an addendum to chapter 1.

Chapter 1

Introduction

1.1 Enzymes in Synthetic biology

1.1.1 Defining synthetic biology

1.1.1.1 A clear definition

Synthetic biology, colloquially SynBio, is a famously difficult to define field (Serrano, 2007). In a European Union report from a high-level expert group of scientists, the agreed upon definition of synthetic biology reads as “the engineering of biology: the synthesis of complex, biologically based (or inspired) systems which display functions that do not exist in nature” (Directorate-General for Research European Commission, 2005). In the United Kingdom’s 2012 roadmap for synthetic biology, the field was defined as “the design and engineering of biologically based parts, novel devices and systems as well as the redesign of existing, natural biological systems” (Clarke *et al.*, 2012).

To expand both accepted definitions, synthetic biology distils life to a complex but tangible system that takes on inputs, and in response produces outputs. Life as a unit system contains a nested set of interacting hierarchical systems that can all be engineered (Cardinale and Arkin, 2012). At the base level is an organism’s DNA. DNA encodes proteins, which perform cellular functions including metabolic pathways. Metabolic pathways are multi-step chemical reactions that provide energy for cell survival and reproduction. In higher taxa, cells form multi-cellular structures. Whole organisms can also converge into consortia (Endy, 2005, Lee *et al.*, 2012; Cardinale and Arkin, 2012). At its core, synthetic biology is the use of engineering frameworks to enact modifications in the basal system (DNA) in order to induce non-natural effects at higher systematic levels (Endy, 2003). Synthetic biologists aim to identify, standardise and wholly characterise individual genetic elements, converting them into modular parts with predictable outputs (Smolke *et al.*, 2018). Such predictability allows for computational modelling of part interactions. Modelling allows for the informed design of genetic circuits, providing mathematical directive for engineering decisions (Chandran *et al.*, 2008). Synthetic biology also includes the focusing and simplification of biological systems. For example single metabolic pathways can be isolated *in vitro* to generate routes for complex biocatalytic conversions without extraneous factors (Shi *et al.*, 2017; Lu, 2017). Furthermore, the definition of synthetic biology can be

expanded to include the use of single or multiple biological elements to produce entirely novel functionality. For example DNA base-pairing can be utilised to generate complex, novel and functional three dimensional structures by DNA origami (Benn *et al.*, 2018; Hong *et al.*, 2017).

1.1.1.2 “First wave” synthetic biology

Synthetic biology sits at the intersection of three broader fields: engineering, biology and chemistry. Taking an engineering perspective, synthetic biology is a field for generating standardised tools and protocols that enable the bottom-up engineering of life at every level to produce defined and predictable outputs (Martin *et al.*, 2009). From a biology perspective, tools to probe the plasticity of life provide avenues to understand and test natural biological systems at every level (Keller, 2009). From a chemistry perspective, synthetic biology is the optimisation over every system level to efficiently generate chemical products at high efficiency (Hall *et al.*, 2012; Luo *et al.*, 2013).

Synthetic biology’s origins can be traced back to 1961 (Cameron *et al.*, 2014). Early experiments on the *lac* operon in *E. coli* showed that environmental inputs lead to defined cellular outputs based on the control of gene expression (Jacob and Monod, 1961). This work led to the hypothesis that genetics that underpin the functionality of life are commensurate to electrical circuitry (Cameron *et al.*, 2014). In its simplest sense, a genetic circuit involves a number of genetic elements that interplay to produce a response or output that is predictable, consistent and controlled. Consider a simple electrical circuit containing a switch, a resistor and a bulb. Direct parallels can be drawn to a genetic circuit that contains a promoter, a ribosome binding site and a gene for a fluorescent protein (figure 1). Importantly, electrical circuitry is modular and compatible, and each part’s function and specification is well documented. Therefore the combination of well understood parts allows for the generation of complex networks that have predictable outputs.

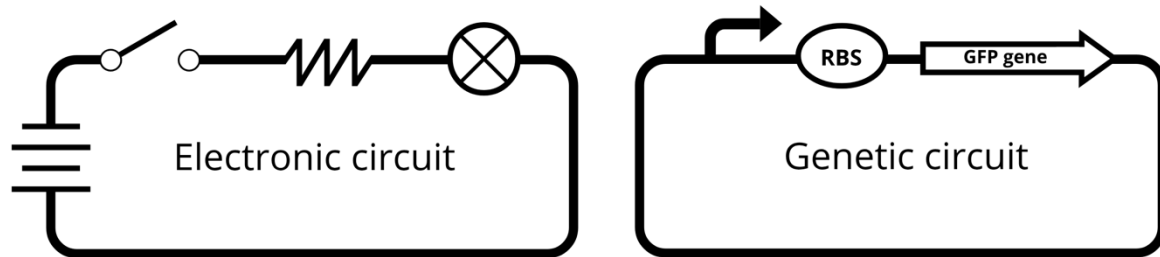


Figure 1 – Genetic circuits are analogous to simple electronic circuits

Simple electronic circuits used for teaching electronics consist of standard, modular parts with well-defined and predictable function. Basic synthetic biology aims to construct genetic circuits in the same manner, where each part is fully characterised and the output from the pathway can be predicted. A promoter and ribosome binding site together act as a switch and a resistor. In the context of a resistor, properties of the promoter and ribosome binding site set the rate of transcription and translation of the gene (the brightness of the bulb in the circuit).

Purnick and Weiss (2009) describe two waves of contemporary synthetic biology. In the first wave, the concept of modularity is a central tenet. For simple and accessible development, a unified DNA assembly process is utilised to combine functional genetic elements into circuits. All new parts are subsequently developed to work with the assembly process, providing ease of prototyping. Multiple parts that work together in composite to produce a predictable output can together be considered a single part, that is equally combinatorial with other parts and modules (Canton *et al.*, 2008; Serrano, 2007). Additionally, early guiding principles outlined the importance of clearly defined chassis organisms into which engineered biological systems can be integrated. These organisms must be malleable to engineering workflows, and present a number of tools that allow for engineering at multiple levels. The chassis is the platform into which novel circuitry is integrated, tested and run (Adams, 2016). Organism chassis act as both the physical housing for the circuitry, but also provide a pool of metabolites and pre-existing circuitry onto which new parts can be plugged into. *Escherichia coli* is an obvious heavy-use example in the synthetic biology lab. The organism benefits from a 20 minute doubling time, ease of genetic modification through the utilisation of plasmids, and the huge body of research underlying its application (Cameron *et al.*, 2014).

As genetic parts are encoded in DNA, the first wave synthetic biology grew out of significant advancements in the accessibility and affordability of synthetic DNA developed around 20 years ago. In a keynote speech at the 2018 international Genetically Engineered Machine Giant Jamboree, early synthetic biology adopter Professor George Church of Harvard University showed a receipt from 1980 for two 10 base pair strands of DNA, priced at \$13,000 (Church, 2018). The cost of DNA synthesis has since dropped rapidly based on advances in synthesis technology (Katz *et al.*, 2018). Considering the cheapest synthesis companies on the market at the time of writing, that same DNA costs \$1.40 today (Twist Bioscience, 2018). Over the same timeframe, similar trends in cost and throughput have seen the accessibility of DNA sequencing data dramatically improve (figure 2). Large databases of DNA sequences are freely available online. At the time of writing, Genbank is in its 228th release (Benson *et al.*, 2013). It contains approximately 280 trillion base pairs of DNA sequence data from 200 million sequences, all of which is searchable both manually, and by sequence similarity with the Basic Local Alignment Software Tool (BLAST; Altschul *et al.*, 1990). Therefore, ready access to reading and writing of DNA allows for the accessible design and formation of modular genetic parts to specification (Katz *et al.*, 2018).

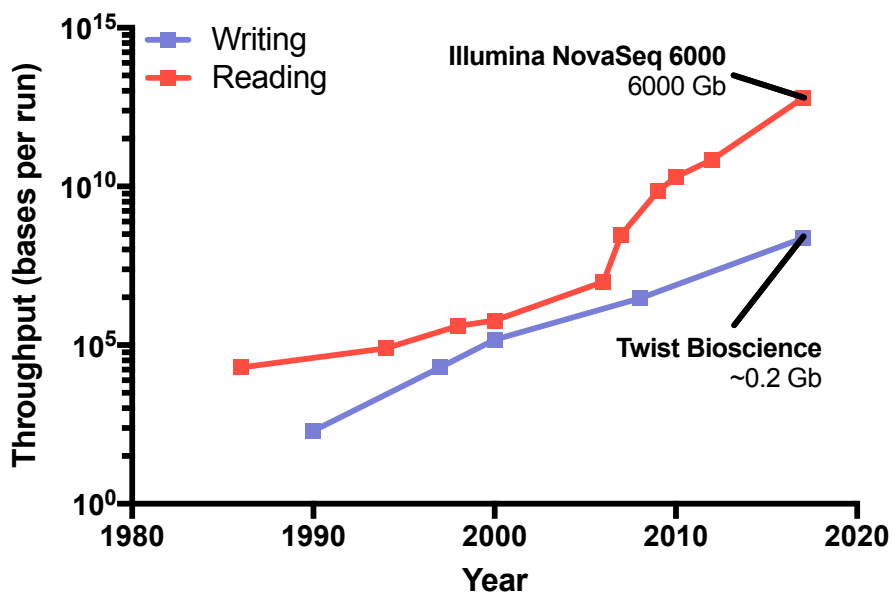


Figure 2 - The current state of DNA reading and DNA writing

Over the last 20 years, access to reading and writing DNA has significantly improved. Contemporary data represents the 2018 market leaders for throughput (Illuminina, 2018; Twist Bioscience 2018; Cox and Chen, 2018). Past data is adapted from Smolke *et al.* (2018).

1.1.1.3 “Second wave” synthetic biology and commercialisation

Within the first wave, research was largely foundational. Whereas within the second wave of synthetic biology research is instead translational – i.e. how can we use synthetic biology as a solution to real problems and generate tangible value from the technologies developed in the first wave (Amos, 2014; Chubukov *et al.*, 2016; Erb *et al.*, 2017). Purnick and Weiss (2009) were the first to identify the inflection point between first and second wave synthetic biology. While the shuffling of parts can produce desirable outputs and responses, considerable bottlenecks occur when complex systems are moved beyond a proof of principal (Kitney and Bradley, 2012; Hodgman and Jewett, 2012; Boehm and Bock, 2019; Liu *et al.*, 2018). The second wave of synthetic biology is abstracted from the generation and discovery of individual parts, and instead is focused on the optimisation of parts and total systems to meet project requirements and specification (Schmidt-Dannert and Lopez-Gallego, 2016). Therefore, the second wave of synthetic biology also concerns the commercialisation of genetic circuitry (Purnick and Weiss, 2009; Amos, 2014). Due to the complexity of life, it is currently poorly understood how cells respond and interact with

synthetic circuits. For this reason extensive work still ongoing for the design, construction and testing of genetic parts (Smolke *et al.*, 2018; Casini *et al.*, 2015). Therefore contemporary synthetic biology runs in two parallel strands, one for the development of parts, and one for the optimisation of systems.

Development of commercialised synthetic biology has important economic implication (Schmidt-Dannert and Lopez-Gallego, 2016; Chubukov *et al.*, 2016). In the United Kingdom's 2016 Synthetic Biology Strategic Plan, a roadmap was laid out to grow the field into a £10 billion market by 2030 (Synthetic Biology Leadership Council, 2016). Refactoring life allows for the production of useful, marketable products. As such, synthetic biology can be considered a platform onto which new industrial processes can be developed, impacting multiple market sectors (Clarke *et al.*, 2012). Non-exhaustively, these include farming for the generation of resilient or enhanced crops, enhancing the scope of recycling programs by utilising waste products as feedstock, the development of pharmaceuticals including antibiotics, drugs and biologics like antibodies, and the development of fine chemicals from low cost feed stocks (Synthetic Biology Leadership Council, 2016). At the heart of each of these applications, these chemical bioconversions are performed by re-routing an organism's metabolism, or isolating the enzymes driving metabolism of the *feedstock in vitro*. Either way, these practices remove the need for harsh catalysts and multi-step purification paradigms (Chubukov *et al.*, 2016).

1.1.2 Enzymes

1.1.2.1 Basic enzyme principals

Enzymes are proteins innovated for the catalysis of biologically important chemical reactions, and are central to many synthetic biology applications. The first evidence of the existence of enzymes came in the early 18th century, when sugar factory scientists Payen and Persoz (1833) discovered that germinating barley could turn starch to sugar. Payen ground and filtered germinating barley, and extracted a white flocculant material by alcohol precipitation. When solubilised, this material was shown to break down the glycosidic bonds

in starch, producing solubilised sugar. Diastase was later renamed amylase, the same enzyme secreted by salivary glands for the digestion of dietary starch (Armstrong, 1933).

Since diastase, it has been discovered that almost all chemical reactions required for life are facilitated by the activity of enzymes. Some have suggested that based on their ubiquity and essential role in protein synthesis, aminoacyl-tRNA synthetases were among the very first enzymes to have evolved (Woese *et al.*, 2000; Chaliotis *et al.*, 2017). Since, enzymes have evolved into six major classes, distinguished by enzyme classification (EC) identifiers (table 1; McDonald and Tipton, 2014).

Identifier	Name	Catalysis	Examples
EC1	Oxidoreductase	Donor-acceptor reaction where a donor molecule is oxidised, donating hydrogen to an acceptor molecule that is reduced. Both donors and acceptors can be cofactors (e.g. NADP+ or NADPH).	Dehydrogenase, reductase, oxygenase, peroxidase, dismutase, luciferase
EC2	Transferase	Donor-acceptor reaction where a chemical group (i.e. methyl) is transferred from a donor to an acceptor. Transfer typically occurs by exchange with another group (smallest unit: Hydrogen).	Methyltransferase, glycosyltransferase, transaminase, transketolase, acetyltransferase, phosphotransferase, riboflavin synthase
EC3	Hydrolase	Hydrolytic cleavage of chemical bonds, most commonly C-O, C-N, C-C, S-S.	Amylase, lipase, esterase, protease, nucleosidase, glycosidase, peptidase, helicase, GTPase, ATP synthase
EC4	Lyase	Cleavage of chemical bonds without hydrolysis or oxidation. Can form ring structures.	Aldolase, dehydratase, decarboxylase, adenylyl cyclase, tryptophan synthase, ferrochelatase
EC5	Isomerase	Conversion of a compound from one isomer to another by intermolecular rearrangement, for example stereochemistry inversion.	Racemase, epimerase, cis-trans isomerase, tautomerase, cyclase, decyclase, cycloisomerase, mutase
EC6	Ligase	Joining of two larger molecules with hydrolysis of a diphosphate in a nucleotide triphosphate.	DNA ligase, chelatase, aminoacyl-tRNA synthetase, Acetyl-CoA synthetase, thiokinase, ubiquitinase

Table 1 – EC classification of the six major enzyme subfamilies

Enzymes are able to catalyse both unimolecular and multimolecular reactions. For simplicity unimolecular reactions will be discussed below (Fersht *et al.*, 2017). Typically, enzymes contain a conserved active site that is responsible for the catalysis of substrate to product in a highly selective manner. Selectivity is defined by a combination of the residues constituting the active site surface, the overall fold and flexibility of the active site, and the shape and size of the substrate tunnel if the active site is buried (Kingsley and Lill, 2015; Weng *et al.*, 2011). Compatible substrates interact with the active site in a flexible-lock and key fashion, where the enzyme's dynamic structure is stabilised by complementary interactions between the active site and the substrate (Koshland, 1958). This same interaction also holds the substrate in a conformation that minimises the Gibbs free energy of activation (ΔG^\ddagger) for the reaction (Fersht *et al.*, 2017). ΔG^\ddagger is defined as the difference between the Gibbs free energies of the substrate and the transition state. Gibbs free energy of a given system is defined as:

$$\Delta G = \Delta H - T\Delta S$$

Where ΔH denotes enthalpy, T denotes temperature, and ΔS denotes entropy (Kuriyan *et al.*, 2012). Enzymes are exceptional at minimising the Gibbs free energy of activation, even for highly stable molecular structures. For example, aqueous orotic acid spontaneously decarboxylates with a half-life of approximately 78 million years. Orotidine decarboxylase increases the rate of this same reaction 10^{17} -fold (Fersht, 2017). Orotic acid has such a long half-life as release of the covalently bound carboxyl group requires energy. For this to occur, a molecule in a favourable ground state becomes an intermediate in an unfavourable transition state (figure 3). Enzymes circumvent the energy requirement of such a process by creating a favourable electrostatic environment that stabilises the transition state (Warshel *et al.*, 2006), by directly reacting with the substrate providing alternative routes to the transition state (Nelson and Cox, 2013), or by distorting the substrate structure to destabilise the ground state (Benkovic and Schiffer, 2003).

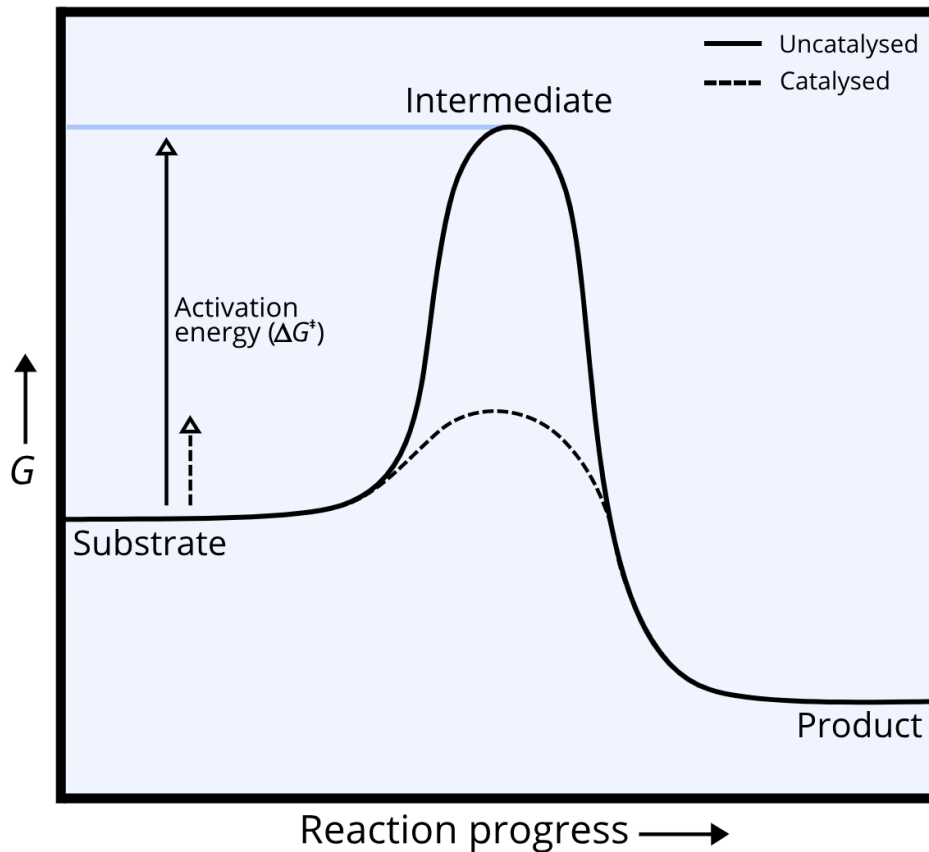


Figure 3 - Enzymes significantly lower the ΔG^\ddagger of a reaction

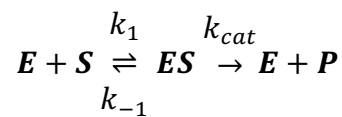
An energy diagram showing the difference between ΔG^\ddagger in uncatalysed (solid line) and catalysed (dashed line) reactions. Reaction shown is the simple conversion of substrate to product. Without a catalyst, the energy required for a reaction to access a high energy intermediate represents a barrier for the reaction to progress to the product. In this demonstration, an enzyme has significantly lowered the energy required to access the transition state. Figure redrawn from Cooper (2000).

1.1.2.2 Enzymes as a tool

Enzymes have been utilised by humans since before recorded history. Approximately 7,000 years ago, Neolithic farming communities developed the art of cheesemaking, using rennet (chymosin) from the stomach of ruminants to curdle milk (Salque *et al.*, 2013). Recently, 13,000 year old stone mortars in Raqefet Cave, Israel, were found to show evidence of purposeful alcohol fermentation by hunter-gatherer settlers (Liu *et al.*, 2018). However, it took until the early 20th century for the first pure enzyme preparations to be achieved, with the crystallisation of urease (Sumner, 1926). Commercial preparations of pure enzymes came to market in the 1960s, where proteases were added to biological washing powders

for improved stain removal (Gurung *et al.*, 2013). In 2014, the enzyme market was valued at \$4.2 billion, with a compound annual growth rate of 7%. A market value of \$6.2 billion is projected for 2020 (Singh *et al.*, 2016). Today, purified enzymes are utilised in the pharmaceutical, food, agricultural, paper, leather, textile, cosmetic, detergent, chemical, waste, biofuel and green polymer industries (Singh *et al.*, 2016; Raveendran *et al.*, 2018; Klein-Marcuschamer *et al.*, 2011). Enzymes provide considerable commercial value as they provide a safe, low cost, enantioselective route to chemistry that typically generates little to no toxic by-product. Additionally, modern access to synthetic DNA and recombinant technologies means the synthesis of enzymes can be trivial given the enzyme is easily soluble (Smolke *et al.*, 2018).

An important metric when considering the economic viability of an enzyme is the rate at which its reaction progresses, and how that rate changes with respect to substrate concentration. Michaelis-Menten kinetics describe this relationship (Fersht, 2017), which is schematically described as:



Where E denotes the enzyme, S denotes the substrate, P denotes the product and k denotes a rate constant. Here the enzyme binds reversibly to the substrate to form the enzyme-substrate complex at rates k_1 and k_{-1} for the forward and reverse processes respectively. Subsequently, the enzyme and the product are released at the catalytic rate k_{cat} . From these rate constants, the Michaelis-Menten equation describing the rate of a reaction (v) can be derived:

$$v = V_{max} \frac{[S]}{K_M + [S]}$$

where:

$$V_{max} = k_{cat}[E]_0$$

and:

$$K_M = \frac{k_{-1} + k_{cat}}{k_1}$$

Where K_M denotes the Michaelis constant, and $[E]_0$ denotes the concentration of enzyme. Importantly, these formulae are only valid under the assumptions that ES rapidly reaches steady state on reaction initiation, that S is far in excess of E , and P is absent on initiation. From these formula, it is possible to calculate optimal substrate concentrations for maximum reaction rate, and the concentration of enzyme required to achieve economically viable processes (Cooper, 2000; Fersht, 2017).

An organism's metabolism is effectively a cascade of enzymatic reactions that convert an input from the environment into energy (primary metabolism) or an ancillary beneficial product (secondary metabolism). A key pursuit in synthetic biology is the engineering of an organism's endogenous metabolism to create synthetic metabolic cascades from heterologous enzymes that generate useful, market valuable products (Erb *et al.*, 2017; Na *et al.*, 2010; Chubukov *et al.*, 2016). Increasing demands for difficult to produce fine chemicals, the recent advent of cheaper DNA synthesis and improved molecular biology tools have meant complex enzyme cascades are becoming increasingly tractable as a solution for optically pure and novel chemical synthesis (Carbonell *et al.*, 2016; Schmidt-Dannert and Lopez-Gallego, 2016; Keasling, 2012). A notable recent example introduced four genes from the *Artemisia annua* plant into yeast to re-route its native mevalonate pathway for the synthesis of artemisinic acid, a precursor to the antimalarial drug artemisinin (Paddon and Keasling, 2014). The remarkable stereoselectivity of enzymes allows for the generation of homochiral molecules, where heterochirality considerably lowers the efficiency and output of traditional chemical synthesis (Martin *et al.*, 2009; Chubukov, 2016). Additionally, enzymes function in innocuous conditions, ablating the requirement of harsh solvents or chemicals, moving toward a "green chemistry" paradigm (Anastas and Eghbali, 2010).

In theory, low cost starting products are converted to metabolic intermediates, which then pass through the synthetic pathway without by-product build-up, mitigating the need for

step-wise purifications that often cause bottlenecks in traditional fine chemical synthesis (Martin *et al.*, 2009). Such pathways are optimised over a series of design-build-test cycles that maximise output based on pushing each enzyme's kinetic properties, and refining the chassis organism to carry out high volume synthesis (Schmidt-Dannert and Lopez-Gallego, 2016; Erb *et al.*, 2016). Synthesis of the antimalarial drug precursor artemisinic acid in yeast is one of the most successful industrial implementations of a multi-enzyme pathway (Keasling, 2012). Amorphadiene synthase converts farnesyl pyrophosphate from the mevalonate pathway into amorpha-4,11-diene. A cytochrome P450 monooxygenase from *Artemisia annua* then converts amorpha-4,11-diene into artemisinic acid in a three-step process (Ro *et al.*, 2006; Keasling, 2012). Chemicals company Amyris hold the current patent for this pathway (EP2565197A1; Seeberger *et al.*, 2013).

1.1.2.3 Enzyme toolboxes

The dramatically increasing capacity of high throughput DNA sequencing is enabling more rapid discovery of new biological parts. "Catalytic toolboxes" have been developed for the rapid screening for useful parts for accelerated prototyping (Martin *et al.*, 2009; Winkler, 2018). Toolboxes are typically curated panels of natural enzymes that perform similar reactions on varied but well-characterised substrates at varied rates (O'Reilly and Turner, 2015). Efforts in expansion of the scope of enzyme toolboxes are seen as an important pursuit in lowering the development costs for industrial scale biocatalysis (Keasling, 2012). Ideally, future chemical manufacturers will have access to bulk quantity enzymes that can be mixed and matched in high throughput. However, such a goal requires the expansion of the consortia of enzymes available to synthetic biology (Schmidt-Dannert and Lopez-Gallego, 2016). Enzymes derived from nature are rarely optimal for utilisation in biocatalytic workflows, as they have evolved to function optimally in a specific cellular environment, often compliant to a highly specific role. Therefore, a key approach to the expansion of the enzyme toolbox is enzyme engineering (Endy, 2005; Schmidt-Dannert and Lopez-Gallego, 2016). Enzyme engineering involves the modification of the primary protein sequence to imbue novel or improved functionality at the tertiary level. Mechanism-guided engineering processes can guide enzymes with no, or low activity on a given substrate into highly

optimised catalysts that meet industrial requirements (Schmidt-Dannert and Lopez-Gallego, 2016).

1.1.3 Enzyme engineering

1.1.3.1 Space and landscapes as a concept

A common notion encountered when discussing the modification and the nature of enzyme sequences is “space” (i.e. Dryden *et al.*, 2008; Povolotskaya and Kondrashov, 2010; Buchholz *et al.*, 2018). In the context of an enzyme’s sequence, “space” is a conceptual device describing the complete set of states an enzyme can possess. This can be distilled into hierarchical categories:

Global sequence space

Global sequence space is the sum total of all possible sequences given a particular length of amino acids. Global sequence space is highly dimensional, as every amino acid position can conform to one of 20 possible states (Buchholz *et al.*, 2018). Very quickly sequence space begins to deal in extraordinarily large numbers. For example, the average protein size in *E. coli* is approximately 277 amino acids (Skovgaard *et al.*, 2001). A 277 amino acid protein has a global sequence space of 2.16×10^{297} total possible sequences. To provide (rather inadequate) comparative scale, recent estimates suggest there are $\sim 5.3 \times 10^{79}$ atoms in our observable universe (Planck collaboration *et al.*, 2015).

Contained within this exceptionally vast sequence space in our arguably average protein example is every functional 277 amino acid protein that has ever existed, as well as every functional 277 amino acid protein that has never existed. At the time of writing, protein sequence database UniProt is on release 2018_10. The total number of proteins on the database, as an estimation of the total number of proteins known to science, is a miniscule fraction of possible sequence space at 1.3×10^8 sequences (Uniprot, 2018). Over the 4Gya life has been evolving, it has been estimated that between 4×10^{21} and 4×10^{43} total protein sequences have been explored by life (Dryden *et al.*, 2008) – still an infinitesimally small fraction of the possible explorable space in our above example. For these reasons, global

sequence space makes for a poor conceptual tool when it comes to the study of enzymes and their potential. Instead, it is typical to focus on more detailed descriptions.

Functional space

Functional space contains the set of all possible functional sequences within a given global sequence space (Povolotskaya and Kondrashov, 2010). This is a difficult number to estimate, as we are not able to know the total set of all possible amino acid sequences that fold to form a functional molecule. This concept is further confounded by our inability to know the complete set of possible functions that proteins are able to perform. However, it is understood that the majority of global sequence space is catalytically deficient (Povolotskaya and Kondrashov, 2010; Dryden *et al.*, 2008). It is estimated that a randomised sequence library of order 10^{24} variants would be required before one should expect to obtain functional biocatalysts (Taylor *et al.*, 2001). In one study, four functional ATP binding proteins (ANBPs) were identified from a library of 6×10^{12} random proteins (Keefe and Szostak, 2001; Lo Surdo *et al.*, 2004). However, these sequences had no notable catalytic activity suggesting that functional catalysts represent a minute fraction of sequence space.

Functional and structural protein sequence space are related

Consider a point in sequence space that contains a functional protein. Then consider the sequence space around that point. As a functional protein is typically able to tolerate sequence variation at points besides essential fixed residues, it can be considered that a given protein exists in a densely populated cluster of viable sequence space (Nardo *et al.*, 2018). Now consider a single, functional protein that folds into a three-dimensional structure. A protein's structural and functional sequence space can be defined as every amino acid sequence that can produce an equivalent, viable fold and/or function. In evolution, such densely populated regions of sequence space consist of all viable homologues of a given protein structure (Shakhnovich *et al.*, 2005). Such sequences derive from the same evolutionary origin, whereby a fitness gain to the ancestor fixed a specific protein structure in a population, and drift or divergent selective pressures led to the diversification of sequence space.

Property space

Protein property space is the set of protein properties that are possible given structural and functional space. Properties denote the function defining traits a protein possesses, for example stability, flexibility, ion binding capacity, multimerisation potential, substrate range and turnover. In nature, a protein's property space is intrinsically linked to a protein's fitness contribution to a given organism (Boucher *et al.*, 2014). It is possible to envisage property space as a fitness landscape, whereby the set of properties defined by the protein sequence confers a phenotype (Kondrashov and Kondrashov, 2015). When a protein is evolving, its set of properties is optimised to function adequately in its given setting. Fitness landscapes are multi-dimensional, as they are defined by the sum of all protein properties. However, as a convenient conceptual tool, we shall hone in on a single property, and distil a fitness landscape down to a two-dimensional space, where the x-axis denotes the protein sequence, and the y-axis denotes the sequence's fitness contribution to the parent organism (figure 4; de Visser and Krug, 2014; Boucher *et al.*, 2015). A fitness landscape therefore contains peaks of high fitness and troughs of low fitness. Evolution has traversed this landscape to obtain a sequence that confers a fitness adequate for survival given a selective pressure (Pál *et al.*, 2006). Pervasive selective pressures then fix sequences and structures in the population as their loss would significantly decrease fitness (Tokuriki and Tawfik, 2009A). In a given landscape, it is possible for multiple solutions to confer adequate fitness. While it is tempting to consider that it is beneficial for a protein to optimise for the highest peaks in a given landscape, it must also be considered that evolution is a non-preparative force that will only naïvely optimise for a given selective pressure (Taverna and Goldstein, 2002; Williams *et al.*, 2007). Therefore once a protein reaches a position in a fitness landscape that is sufficient for survival, the selective pressure requires no further optimisation. When combining all fitness landscapes, a protein can be considered as the adequate optimisation over all dimensions in the context of a set of selective pressures. This leads to trade-offs within the landscape, where the increase of fitness in one dimension decreases fitness in another, which can subsequently be compensated for by additional mutations (Brown *et al.*, 2010; Hartl *et al.*, 2014).

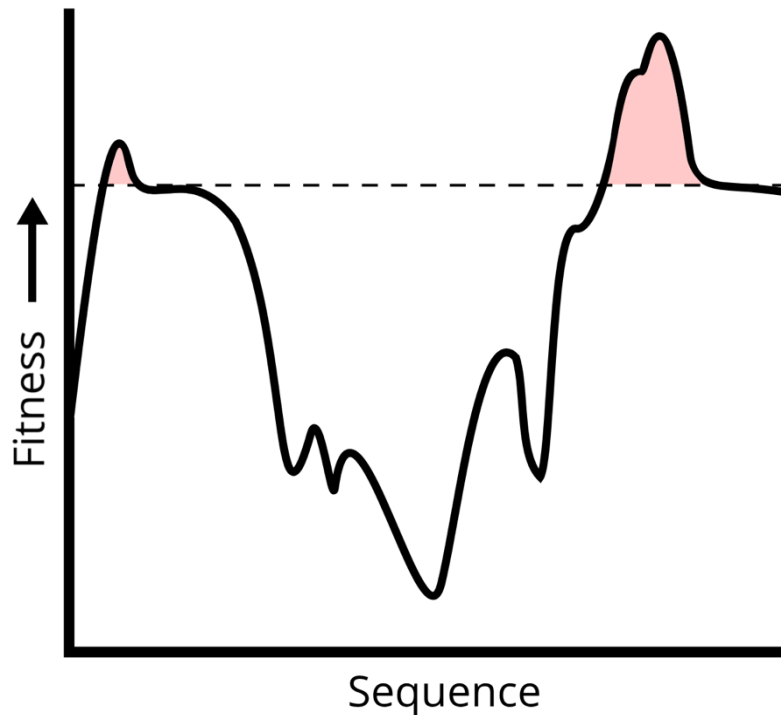


Figure 4 - A two dimensional protein fitness landscape

Protein fitness landscapes are multi-dimensional, as many protein properties contribute to host organism fitness. For conceptual simplicity, landscapes can be distilled to represent single properties (de Visser and Krug, 2014; Boucher *et al.*, 2014). In this instance the fitness landscape fluctuates based on the sequence, providing multiple peaks and troughs in a given sequence space. Dashed line represents the minimum viable fitness for a given parent organism. Therefore in this representation there are more than one route to viable fitness, with all viable sequence space represented by red shaded maximas.

1.1.3.2 Engineering is the meaningful traversal of property space

Throughout a protein's evolutionary history, its property space has been refined to provide specialized functionality that is adequate for survival of its parent organism. This is at odds with the requirements of an enzyme in synthetic biology, where maximal output in a non-natural setting is necessary for success. "Toolboxes" of natural enzymes, for example the set of Carboxylic Acid Reductases described by Winkler (2018), represent the library of sequence space that nature has developed. Such panels possess a limited activity space defined by the natural enzyme consortium (O'Reilly and Turner, 2015). Exploration of sequence space around select enzymes in these toolboxes is therefore a commonly adopted

solution for elucidating new enzymes with beneficial properties that lead to process optimization (Bommarius *et al.*, 2011).

Engineering a protein's amino acid sequence represents the traversal of sequence space toward more beneficial properties (Currin *et al.*, 2015). Conceptually, the fitness landscape becomes synthetic, where new dimensions are introduced based on required functionality. Additionally, the pressure to achieve optima is much higher, as value in synthetic biology is typically derived from the direct output of the synthetic constructs utilised (Nevozhay *et al.*, 2012). Traditional methods focus on iterative design and rational residue-wise point mutations calculated by computational modelling of a protein's active site (Kaufmann *et al.*, 2010). However, given the paucity in understanding around structure-function relationships in proteins, the method leads to unpredictable and expensive workflows with ill guaranteed success (Arnold, 2018). Contemporary technologies allow for the efficient traversal of sequence space by the bulk modification of the amino acid sequence (Currin *et al.*, 2015). For brevity the most pervasive technology, directed evolution, will be discussed further.

Directed evolution

Directed evolution aims to model evolutionary processes with a laboratory-scale complement of Darwinian survival (Arnold and Volkov, 1999; Bornscheur *et al.*, 2012; Arnold; 2018). The methodology hinges on a simple algorithm. An amino acid sequence of interest is subject to sequence randomisation through enzymatic or direct synthetic methods (Arnold, 2018; Li *et al.*, 2018A; Turner, 2009). Resultant sequence variant libraries are then subject to a selection criteria that is directly associated to the fitness landscape in the dimension of interest. By scattering sequences across the fitness landscape, and directly linking protein performance to survival, only the fittest sequences from the library are obtained (Renata *et al.*, 2015). Subsequent narrowing searches emulate an uphill iterative walk within the fitness landscape (figure 5; Shivange *et al.*, 2016). This iterative walk allows for the traversal of large functional hurdles in property space to be broken up into a series of smaller optimisations.

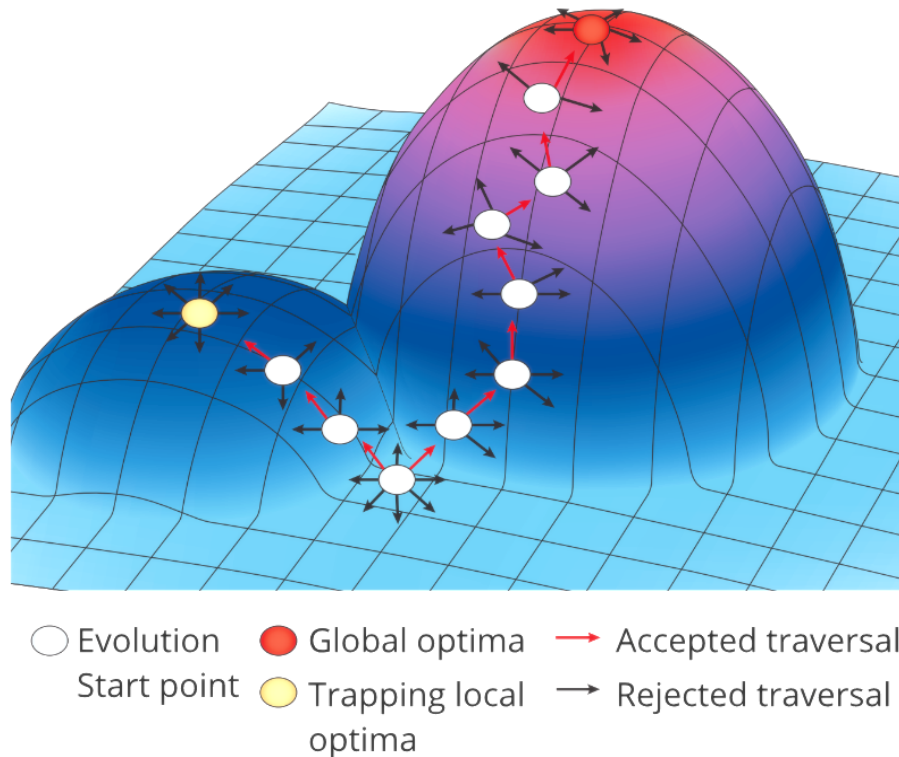


Figure 5 - Directed evolution is the iterative uphill walk to optima in functional space

Here the fitness landscape is represented in three dimensions. Libraries of sequences sample the surface of the landscape in proximity to each evolution start point. Beneficial mutations are selected allowing a protein to climb toward fitness peaks. Here, two peaks are shown separated by a fitness valley. As directed evolution is generally only upward climbing, it is possible for a directed evolution experiment to become trapped at local optima, without ever realising global optima. Image adapted from Packer and Liu, 2015. Image modified and reproduced with permission.

Traditional directed evolution experiments rely on sequence randomisation by error-prone PCR, chemical mutagenesis or use of a mutagenic strain (Cirino *et al.*, 2003; Neylon, 2004). However, chance dictates that random mutagenesis will rarely be representative of the global fitness landscape, and biases caused by PCR or mutagen activity can entirely ablate large portions of designed sequence diversity (Li *et al.*, 2018A). More recent methods include the semi-rational prediction of enzyme states based on a pre-existing crystal structure to produce libraries of highly targeted point mutations (i.e. CASTing or Rosetta; Steiner and Schwabb, 2012; Kaufmann *et al.*, 2010). Protein structure is linked to function by basing mutagenesis on assumptions about structure-function relationships derived from

crystal structures (Lutz *et al.*, 2010; Bornscheur *et al.*, 2012). As a result, semi-rational design must also make assumptions about the most important residues involved in the relationship between sequence, structure and phenotype (Winkler and Kao, 2014; Lutz, 2010), and is therefore at risk of missing sites not overtly linked to function that would still confer positive changes. This is compounded by the methodology utilised to generate the library. In a recent study (Li *et al.* 2018B), the massive sequencing of a library generated by error prone PCR for the engineering of enantioselectivity in limonene epoxide hydrolase showed that around 50% of the designed sequence space was not explored by the library. However, the direct synthesis of the library using solid-phase DNA synthesis technology led to the exploration of 97% of the same library, and the generation of twice the amount of beneficial mutants. Regardless of library generation method, optimised sequences are typically selected from the population by one of two common methods: selection or screening.

In selection pressure based experiments, some evolutionary driving force is applied to the protein population to link protein sequence to desired function. This is typically achieved by inducing a significant fitness cost to organisms that do not perform a provided function (e.g. a toxic product). Selection provides a quantitative measure of the ability of individuals in the library to perform a required function, allowing for the selection of the most optimal sequences within the population (Turner *et al.*, 2009). Survivors of selection rounds are deemed the most “fit” sequences in the landscape, the winner(s) of which can be put forward as the starting point for additional rounds of mutagenesis (Packer and Liu, 2015; Goldsmith *et al.*, 2012). This method has successfully been employed to evolve large biocatalysts toward various optimal properties (e.g. engineering 120 kDa cytochrome P450 monooxygenases for alkane hydroxylation; Fasan *et al.*, 2007).

On the other hand, high throughput screening aims to observe fitness directly based on activity, turnover or survival. Such methods are especially powerful if fitness and phenotype cannot be easily linked (Packer and Liu, 2015; Xiao *et al.*, 2015). A high throughput assay is required to quantitatively screen the entire sequence population. With the advent of microfluidics based screening systems, the volume of screening possible has increased. However, the success of a high throughput screen is a direct function of assay sensitivity and

high throughput feasibility (Ye *et al.*, 2017; Martínez and Schwaneberg, 2013; Wójcik *et al.*, 2015; Agresti *et al.*, 2009).

Regardless of library generation and screening method, the first tenet of directed evolution states “you get what you screen for” (Li *et al.*, 2018B). To efficiently screen for a property, the selective pressure or criteria applied must directly optimise the sequence landscape of interest. “Parasitic” sequences that are fit under a given selective criteria but do not harbour desired traits can be selected when the selection pressure rewards more than one possible phenotype (Tizei *et al.*, 2016). Indirect effects can lead to the optimisation of unintended properties, which may or may not undermine the overall engineering experiment. For example organisms grown under the presence of a toxin to select for more efficient toxin remediation machinery could cause indirect selection of variants more able to form biofilms or spores that protect against the toxin (Marlière *et al.*, 2011; Tizei *et al.*, 2016). In the object of efficiently traversing fitness landscapes, considerable care must be taken when choosing screening modes to ensure a protein sequence is pushed toward the desired outcome, without inducing risks of secondary unpredicted outcomes.

1.1.4 Moving synthetic biology beyond proof of concept

Compared to a traditional engineering challenge, optimising living systems is markedly more difficult and imprecise. Ideally all the features of the system should be synergised to maximise product yield in the face of input cost. Critical factors include the choice of chassis and its growth environment (Vinay-Lara *et al.*, 2016), the metabolism of the chosen chassis (including its stress responses; Lee *et al.*, 2008; Dahl *et al.*, 2013), the enzymes contained within the pathway (Bornscheur *et al.*, 2012), the expression and control of the cascade and its cofactors (Angelastro *et al.*, 2016), and the reaction process including the reaction vessel (Goundry *et al.*, 2017). A bioengineering project can be likened to building a computer by constructing each component from the ground up, with only sparse knowledge of how each component will interact with the whole system, leaving experimentation (and therefore many rounds of failure) as the only means of progression. Implementation of workflows like the archetypal design – build – test cycle, which itself can contain many nested, equivalent cycles, allow iterative forward engineering of such a process (Agapakis, 2014).

Predictably, the end-goal for development of such cascade-based pathways is integration into an industrial setting (Schmidt-Dannert and Lopez-Gallego, 2016). For a company, success not only requires a functioning enzyme cascade that synthesises a marketable product. Rather, the cascade must also be scalable to a market viable production volume (Langerack and Hultink, 2006). Historically, the scale-up process incurs high costs for development and implementation (Keasling, 2012). These are often caused by underestimation of scale-up timeframes, or unforeseen process bottlenecks or inefficiencies. Any young company looking to develop industrial enzymatic pathways is therefore of high risk to investors (Chubukov *et al.*, 2016; Sanford *et al.*, 2016). The same financial burden will also impede enzyme cascade development within larger companies attempting to integrate biocatalysis into their current chemical synthesis portfolios. The considerable development cost must be offset by an equally large payoff for integration (Lechner *et al.*, 2016; Langerack and Hultink, 2006). However, in terms of technology readiness levels (a market estimation of technology maturity), an analysis by Cambridge Consultants (Ho *et al.*, 2017) shows that several synthetic biology start-up companies are “market ready”. Successful examples of enzymatic pathway utilization are readily emerging in the bioindustrial market (Paddon and Keasling, 2014; Shetty, 2016). Notable examples include Bolt Threads, who sell apparel constructed from synthetic spider silks synthesised in yeast (DeFrancesco, 2017); Evolva who use yeast to ferment low value inputs into marketable compounds including Stevia™, nootkatone and resveratrol (Nyffenegger *et al.*, 2017); and Hyasynth who generate medicinal cannabinoids within recombinant microorganisms (Feeney and Punja, 2017). Notwithstanding, one contemporary challenge for synthetic biologists is the reduction of costs when developing enzyme technologies past their proof-of-concept, to increase biotechnology’s penetrance into the chemical industry.

1.2 Thermostable enzymes for the advancement of industrial synthetic biology

1.2.1 Defining enzyme stability

1.2.1.1 Protein folding and stability depend on entropy and enthalpy

A key determinant of an enzyme's applicability to a biotechnological process is its stability (Chapman *et al.*, 2018). Stability refers to how well the protein retains its fold and function over time, or under process specific conditions. An unstable protein will quickly lose function, impacting process efficiency. More specifically, stability refers to the energy input required to impede a folded protein's correct function through damage to its structure (unfolding; Pace, 1990). Therefore, the stability of an enzyme relates directly to the molecule's ability to maintain its tertiary or quaternary structure. A protein's tertiary or quaternary structure is the sum of van der Waals forces, hydrophobic interactions, water liberation, ionic interactions and disulphide bridges acting upon the protein's secondary structure (Dill, 1990; Pace *et al.*, 1996; Fersht, 2017). Additionally, the layer of water that surrounds the protein (its solvation shell) has a stabilising effect (Ebbinghaus *et al.*, 2008). Electronegative oxygen atoms in water undergo electrostatic interactions with the positively charged surface residues surrounding the protein, "holding" it in conformation (Pal *et al.*, 2002).

In free energy terms, a protein folds to minimise its Gibbs free energy (ΔG) of folding, given by: $\Delta G = \Delta H - T\Delta S$. The Gibbs free energy of folding can also be given by:

$$\Delta G = G_{folded} - G_{unfolded}$$

Where G_{folded} is the free energy of a folded state, and $G_{unfolded}$ is the free energy of the unfolded state (Fang, 2014). For protein folding to be spontaneous, the free energy of the folded protein has to be smaller than the free energy of the unfolded protein. A simple view of why proteins fold involves the consideration of the first formula. Conformational entropy is high in the unfolded state as it has high free range of movement. However, enthalpy also increases the closer the protein gets toward primary structure as there are exponentially increasing numbers of potential stabilising interactions possible between amino acid side

chains (Kuriyan *et al.*, 2012). Therefore, as enthalpy is low in the folded state, ΔG is negative and folding is maintained (Yang *et al.*, 2013). The order of amino acids directly controls whether folding collapses the molecule into one that is functional or not (Zwanzig, 1992).

Entropy also drives stabilisation considering the interaction of the protein with its surrounding water molecules (Kuriyan *et al.*, 2012). A protein is made up of amino acids with hydrophobic or hydrophilic side chains (table 2). Interactions with polar hydrophilic residues lead to the bulk ordering of water molecules around the protein forming the solvent shell, which has ice-like properties, and a lower entropy than bulk water (Ebbinghaus *et al.*, 2007). As the universe tends towards increased entropy, a protein can enter favourable states by packing non-polar hydrophobic residues in the centre of the protein, which allows for the release of water from around the residues into the surrounding system, dramatically increasing disorder (Kuriyan *et al.*, 2012). Mutations in the protein core are therefore on average slightly destabilising (Faure and Koonin, 2015). The mean change in protein melting temperature for a mutation has been calculated as $-5\text{ }^{\circ}\text{C}$ (Pucci and Rومان, 2016).

Property	Amino acids
Hydrophobic	Alanine, Isoleucine, Leucine, Methionine, Phenylalanine, Proline, Tryptophan, Valine
Hydrophilic	Arginine, Asparagine, Aspartate, Cysteine, Glutamate, Glutamine, Glycine, Histidine, Lysine, Serine, Threonine, Tyrosine

Table 2 - Classification of amino acids by hydrophobicity

It is also important to consider that a protein fluctuates between multiple favourable Gibbs free energy states due to the highly dynamic nature of the system. Dynamism stems from the sheer volume of possible interactions between residues within the folded protein, and the flexible nature of the bonds making up the protein's primary structure (Jaenicke, 1991). To increase a protein's stability, is therefore necessary to modify one or many of the aspects

that lead to its destabilisation. For example, by decreasing enthalpy in the system, the Gibbs free energy becomes more negative, and therefore the energy input required to achieve destabilisation is increased (Hilser *et al.*, 1996; Dagan *et al.*, 2013). The same net effect can be achieved by increasing the free energy of the denatured state (Sugita and Kitao, 1998). Mutations in the primary structure of a protein can have an impact on both the folded and denatured state (Shortle, 1996). Alternatively, changing how a protein interacts with its bulk surroundings, for example increasing its propensity for hydrophobic packing or minimising chain fluctuations, will typically have positive effects on overall stability (Jaenicke, 1991).

1.2.1.2 How to denature a protein

In the context of synthetic biology, protein stability is typically defined as its resistance to destabilising agents, especially temperature, solvents, pH, or salt. Temperature causes denaturation of proteins in two key steps, reversible and irreversible unfolding. In reversible unfolding, a reversibly inactive form of the enzyme forms an equilibrium with its active form (Daniel and Danson, 2013; Fersht, 2017). Temperature has a direct impact on the equilibrium between the active and inactive state reversible state, where higher temperatures push the enzyme toward the inactive reversible state (Peterson *et al.*, 2004; Daniel *et al.*, 2010). The conformational landscape sampleable by a protein during its folding is vast, and when a protein folds it continuously samples multiple viable states of low Gibbs free energy. It can therefore be assumed that the inactive reversible state is caused by conformational shifts in the enzyme that with increasing temperature lead to lowered Gibbs free energy (Fersht, 2017). As the flexibility of the primary structure is directly dependant on the temperature of the system, the entropy of the system will increase as temperature increases. The enthalpic change is small however, as the protein structure remains largely intact, but is more freely able to sample conformational space. This leads to new conformational states that do not form a functional active site at “optimally” low Gibbs free energy under high temperatures (Fersht, 2017; Daniel and Danson, 2013). However, this conformation must not be distant from functional conformational space, as the protein must be able to traverse back to the functional state both at the increased temperature (maintaining equilibrium), and as the temperature decreases (regaining function; Daniel *et al.*, 2010). As temperatures increase, it is expected that the number of possible non-

functional states increases (Daniel *et al.*, 2010; Daniel and Danson, 2013) until complete denaturation occurs.

Complete denaturation is the global, irreversible loss of structure. As temperature increases, there is enough energy in the system to break down the hydrogen bonds both in the network of water molecules holding the protein in place, and the hydrogen bonds in the core of the protein (Koizumi *et al.*, 2007). Looser conformations lead to decreased hydrophobic packing and increased water penetration in the centre of the protein (Groot and Bakker, 2016). As hydrogen bonds are not able to form at higher temperatures, the enthalpy in the system is decreased, and the unfolded state of the protein (closer to primary structure) has a considerably lower Gibbs free energy than the same unfolded state at lower temperatures.

At standard mesophilic temperatures, it is thought that the folding of the protein towards the lowest Gibbs free energy occurs rapidly due to a combination of nucleation into smaller macro-structures and hydrophobic collapse (Zwanzig *et al.*, 1992; Duan and Kollman 1998; Zhou *et al.*, 2004). At higher temperatures, there are increased numbers of sampleable folded states available to the protein. As the system cools, the number of possible states diminishes. However, cooling from high flexibility allows the protein to access thermodynamically favourable states that are not the native state, that require addition of energy to unfold. Therefore, as the protein refolds, it gets stuck on local minima of Gibbs free energy with no route to escape, leading to misfolding and loss of function (Strucksberg *et al.*, 2007). These same principals can be applied directly to the other synthetic biology-centric stability factors. Modifications to the pH, solvent and salt concentration in a solution modify the sum of zwitterionic states across the protein's surface. As the charge changes on the surface of the protein, the strength of the protein's interaction with the solvation shell also changes (Jungwirth and Cremer, 2014). Addition of solvents also displaces water in the solvation shell (Mattos, 2002; Schellman, 2003). A protein's solvation shell impacts the number of states the protein can sample based on how strong the interaction is, and increased protein flexibility is consistent with weaker solvation shell interactions (Mattos, 2002; Timasheff, 2002; Born *et al.*, 2008).

Relaxation of the solvation shell induces considerable vulnerability of the protein core to destabilisation (Dahanayake and Mitchell-Koch, 2018). Often, specific modules within the enzyme structure are separated by loop structures which extend to the surface of the protein. Loop structures typically have higher degrees of freedom of movement compared to the protein protein's core. (Papaleo *et al.*, 2016; Dominy *et al.*, 2002). As such, loop regions are key determinants on protein stability (Balasco *et al.*, 2013). Based on disruption of the solvation shell and the ionisation of the protein surface, destabilising agents kick off a destabilisation chain reaction in the protein structure (Fersht, 2017). Increased flexibility leads to penetrance of both water and destabilising agent, disrupting the core structure. Resultant core relaxation subsequently allows for the penetration of increasingly more water and destabilising agent, inducing further unfolding (Nestl and Hauer, 2014; Dominy *et al.*, 2002). Additionally, changes to zwitterionic states can lead to electrostatic repulsion within the protein, completely changing its folding landscape, increasing the likelihood that inactive states are sampled (Dominy *et al.*, 2002). A thermostable protein will often also be stable in the presence of destabilizing agents, as the principals underlying destabilization are largely equivalent in both instances (Hao and Berry, 2004; Razvi and Scholtz, 2006; Arabnejad *et al.*, 2017). Therefore, unless specified, "stability" will be discussed in terms of thermostability going forward, and increases or decreases in stability will generally be discussed in terms of a protein's melting temperature.

1.2.2 Most proteins are marginally stable

Logically, from a fitness perspective it should be beneficial for proteins to be highly stable. With stable proteins, the organism's resistance to fluctuating temperatures increases, and overall less resource would be required for protein synthesis as the probability for a given protein to be folded is high (Goldstein, 2011). However, in nature the opposite is consistently observed. For the majority of organisms, the majority of their proteins are marginally stable (Taverna and Goldstein, 2002; Williams *et al.*, 2007; Goldstein, 2011). That is, the Gibbs free energy of unfolding is only slightly negative, and small increases in temperature lead to the denaturation of proteins (Privalov and Khechinashvili, 1974). Considering a protein's functional space, the native stability of most proteins is far lower than its maximum possible stability (Taverna and Goldstein, 2002). The observation that

marginal stability is a near-ubiquitous trait amongst proteins therefore suggests that some evolutionary pressure favours less stable protein conformations (A phenomenon explored in Chapter 3).

Following early observations that proteins are marginally stable (e.g. Privalov and Khechinashvili, 1974), initial research considered marginality as an evolved trait (Jaenicke, 1991; Shoichet, 1995). While a protein has to evolve to be stable, it also has to evolve to be functional. Active sites can require the opening of the protein core to allow for penetration of water molecules, which could lead to destabilisation (Shoichet *et al.*, 1995). Furthermore, active proteins often require conformational changes to their active site, where the native structure of the protein and structure of a protein undergoing catalysis are often structurally different (Weng *et al.*, 2011; Todd *et al.*, 2002). Often, promiscuous enzymes can take on multiple substrate-dependent conformations (Pabis *et al.*, 2018). There is therefore a selective advantage conferred by active site flexibility, resulting in a trade-off between activity and stability (Tokuriki *et al.*, 2008). This trade-off manifests as the requirement for increased flexibility, leading to increased Gibbs free energy of the system due to the balance of entropy, enthalpy and dynamics. It has been observed on several independent studies that the introduction of activity decreasing mutations in the active site of a protein induce increases in a protein's stability (Tsou *et al.*, 1998; Závodsky *et al.*, 1998; Shoichet *et al.*, 1995; Schreiber, 1994; Daudé *et al.*, 2013; Sharma *et al.*, 2014; Shahid *et al.*, 2015; Martin *et al.*, 2018). It can be concluded that proteins sacrifice stability to a tolerable level (i.e. to ensure that they fold) to allow for increased flexibility, and concordantly increased activity or promiscuity.

However, the trade-off phenomenon is not ubiquitous. One of the clearest examples is Superfolder GFP, where the protein is both considerably more stable and brighter than wild type GFP (Pédelacq *et al.*, 2006; Steiner *et al.*, 2008). Taverna and Goldstein (2002) considered the impact of neutral theory on marginality. According to the neutral theory of evolution (Kimura, 1968; King and Jukes, 1969; Kimura 1983), a trait is only considered adaptive once all other mechanisms can be rejected. Additionally, the theory states that the majority of mutations have an effectively neutral impact on fitness, and sequences can

therefore drift toward spaces that allow for the generation of properties that do not fulfil an obviously adaptive purpose.

Consider then that three constraints define the evolution of a protein: does it sample conformational space to fold into a structure with minimal Gibbs free energy quickly; does it fold to produce a function; and does it fold to be stable in its environment (Taverna and Goldstein, 2002)? The latter is the only selective pressure pushing a protein's thermostability. As a protein evolves, once it's stability allows it to maintain its tertiary structure and function at environmental temperatures, there are no immediate fitness benefits imparted to the protein upon further stabilisation (Bershtein *et al.*, 2006; Bloom *et al.*, 2005). It therefore cannot be assumed that a selective pressure would exist to drive further stabilization. The minimum functional stability of the protein can be considered a stability threshold (Bloom *et al.*, 2005; Khersonsky *et al.*, 2018). A protein of a stability that fulfils the threshold criteria can therefore obtain increases in stability by mutational drift. However, proteins that drift toward decreased stability are selected against in the population, as they cause a fitness cost to the parent organism (loss of function, aggregation; Tokuriki and Tawfik, 2009A). It would therefore be reasonable to expect that the distribution of protein stabilities would cover a broad temperature range, yet as discussed this is not observed (Goldstein, 2011). This implies that there must also be some broad destabilising force exerted on evolving proteins.

A neural network trained on the stabilising effects of 1,600 mutation across 90 proteins further predicted that for all mutations across 25,000 proteins, the average mutation in a protein causes a -5 °C change in protein stability (Pucci and Rooman, 2016). This observation is corroborated in work by Tokuriki *et al.* (2007), where the Gibbs free energy change for every possible mutation in 21 globular proteins was calculated using foldX (Schymkowitz *et al.*, 2005) to model free energy changes caused by mutations. Across all proteins, the distribution of a stabilisation effects was observed to fit to almost identical Gaussian distributions that were negatively stabilising on average ($\Delta\Delta G$ is positive for the majority of mutations; figure 6). For core mutations, the distribution was broader, and considerably more destabilising on average in comparison to surface residues (which fit to a narrow distribution that was on average slightly destabilising). A broader scope study using the

same foldX algorithm (Faure and Koonin, 2015) modelled the effects of every possible amino acid substitution in all proteins in seven organisms including both hyperthermophiles and mesophiles. Again the average stability contribution was slightly destabilising, with a heavy tail of strongly destabilising mutations. Core-surface destabilisation relationships were similar to those previously observed. Additionally, for hyperthermophilic organisms, the average destabilisation contributions observed were significantly shifted toward destabilisation, suggesting thermophilic proteins approach maximally stabilising states.

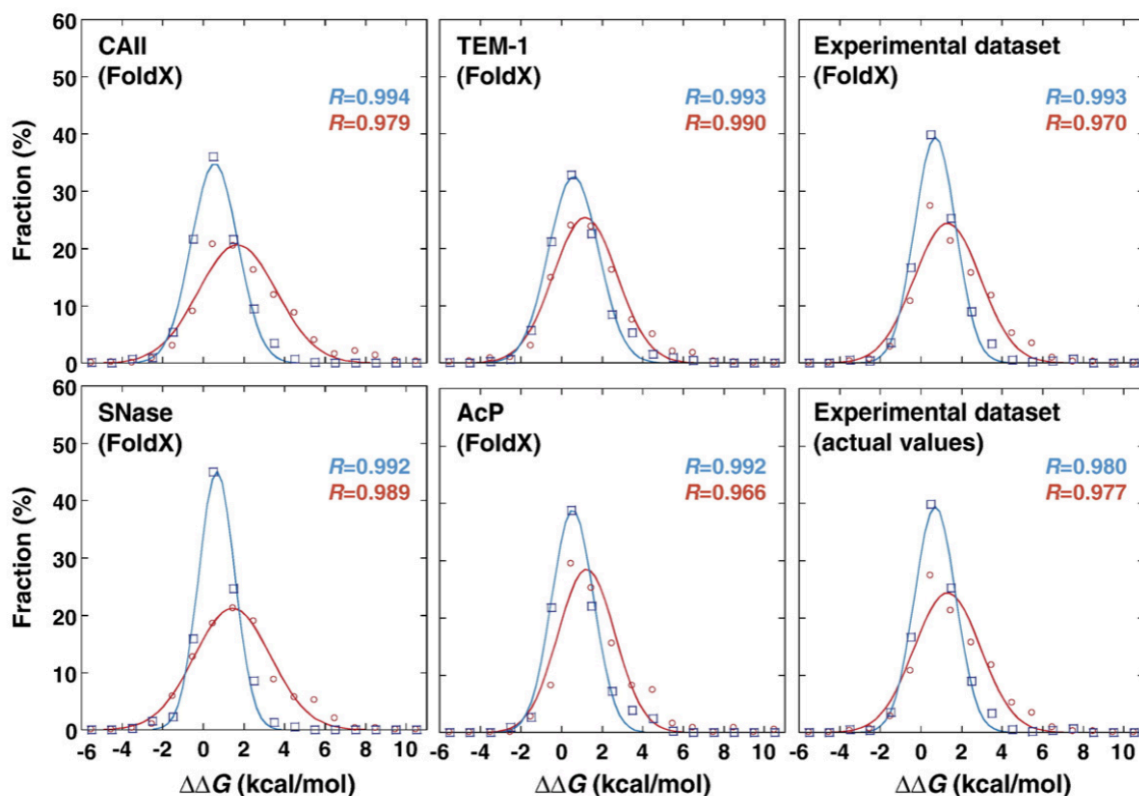


Figure 6 - Mutations in proteins are on average destabilising

Tokuriki *et al.* (2007) modelled the effects on protein stability induced by all possible mutations in four protein families and a database of mutations known to effect stabilization (ProTherm database; Gromiha *et al.*, 2002). Stabilization effects were observed to exhibit highly similar Gaussian distributions, confirmed by comparison with data within the ProtTherm database. It is observed that mutations on a protein's surface (blue curve) are on average neutral with regards to their conferred change in Gibbs free energy (a proxy for stability contribution). On the other hand, mutations in a protein's core (red curve) are destabilizing on average. Therefore, mutations in a protein are destabilizing on average. Image reproduced with permission.

Considering these data, a protein under drift is considerably more likely to accrue destabilising mutations over stabilising mutations. Hypothetically, if a protein from a hyperthermophile was horizontally transferred and fixed in a mesophilic population, it would be expected that the protein would randomly and rapidly drift toward marginality at the mesophilic stability threshold. Therefore, drift is an acceptable hypothesis to describe marginality in protein populations, and marginality arises in the absence of a selective pressure. Providing further evidence that marginality can evolve in the absence of selective pressure, Goldstein (2011) modelled the evolution of a single 300 amino acid protein whose free energy properties were modelled to the first 300 amino acids of 55 structurally diverse proteins. The only selective pressure placed on the protein was the requirement to fold into a given structure at a pre-specified temperature. The evolving protein quickly approached marginality, giving further weight to the reasoning that marginality is not adaptational.

1.2.3 Nature's strategies for protein stabilisation

For a hyperthermophilic organism to survive, their proteins need to be able to survive high temperatures. However, hyperstable proteins are constrained in their evolutionary plasticity, as the majority of mutations in hyperstable proteins are significantly destabilising (Faure and Koonin, 2015). In order for hyperstable proteins to achieve high stability, innovations must overcome free energy landscapes that are largely incompatible with high temperatures. Hyperstable proteins therefore adopt a number of strategies to ensure protein stabilisation. However, there are no universal strategies adopted to generate stabilising effects. An in depth coverage of all strategies is presented by Pucci and Rooman (2017). Table 3 provides a brief description of strategies observed stabilising thermostable proteins. Additionally, the truncation of loops, the formation of hydrogen bonds, hydrophobic packing and water release discussed above are all commonly observed in thermostable proteins.

Innovation	Description	Mechanism
Higher number of salt bridges (Bosshard <i>et al.</i> , 2004)	Electrostatic interaction between oppositely charged residues	Increase free energy of unfolding by increasing unfolded enthalpy
Higher number of cation-π interactions (Prajapati <i>et al.</i> , 2006)	Positive charge of Lys/Arg/His interacting with electron dense rings in Phe/Tyr/Trp	High bond energy interaction that increases at high temperatures. Typically occur on surface. Increase unfolded enthalpy.
Higher number of π-π interactions (Makwana and Mahalakshmi, 2015)	Interaction of two electron dense rings between Phe/Tyr/Trp	High bond energy sharing of electrons between two stacked aromatic rings. Tight packing. Increase unfolded enthalpy. Decrease flexibility
Disulphide bonds within cytoplasm (Betz, 1993)	Covalent Cysteine-Cysteine bonds formed by enzymatic intervention	Only broken by chemical/enzymatic means in biological conditions. Forced folding. Tight packing. Nucleate hydrophobic core.
Dense interaction network hubs (Sammond <i>et al.</i> , 2016)	Focused groups of above interactions	Cooperative foci of above mechanisms with large contribution from molecular packing

Table 3 - Innovations for the stabilisation of proteins in thermophiles

Some of the most thermostable proteins in nature belong to the ubiquitous CutA1 family of proteins, which sequester divalent cations conferring resistance to ionic metals (Matsuura *et al.*, 2015). Many members of the family show stability of over 100°C, with the most stable example to date from the hyperthermophilic archaeon *Pyrococcus horikoshii* denaturing at 150 °C (Tanaka *et al.*, 2006; figure 7). Stabilities far above their environmental temperature is characteristic for this family – Human CutA1 denatures at around 96 °C. *P. horikoshii* grows optimally at 98 °C. It has been observed that robustness is essential for the protein’s function, and is therefore under positive selection for hyperthermostability in all environmental conditions (Hirata *et al.*, 2012). Comparisons between crystal structures have unveiled a number of possible strategies that confer hyperstability into the family. Firstly, the protein forms a tightly compacted trimeric cylindrical structure with dense interaction hubs, minimal flexibility, and truncated surface loops leading to structures with both low enthalpy and extremely low entropy. (Sato *et al.*, 2010; Buchko *et al.*, 2015; Hirata *et al.*, 2012). Additionally, the hyperstability of the protein is highly susceptible to changes in flexibility caused by single amino acids, suggesting a highly optimized structure. Stability differences of over 70 °C between CutA1 variants are attributed to the insertion of a glutamine and proline into a beta sheet, causing a structural “kink” in one strand of the sheet structure (Hirata *et al.*, 2012; figure 7 red circle).

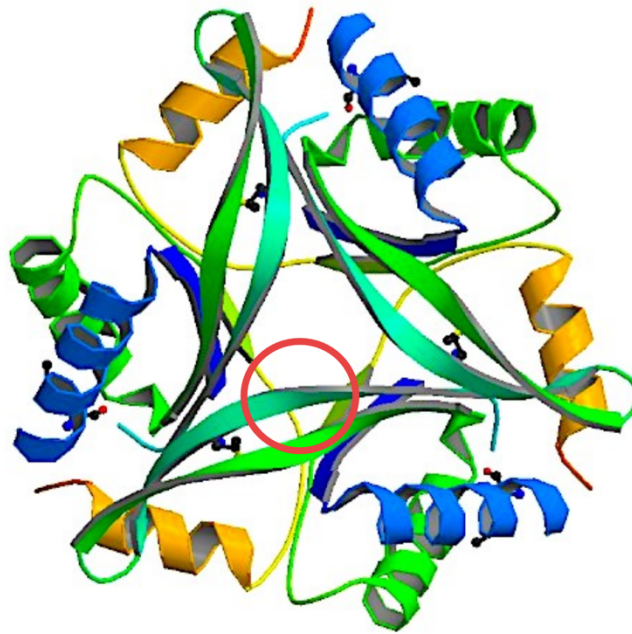


Figure 7 - PDB 1J2V – CutA1 from *Pyrococcus horikoshii*

Crystal structure of hyperthermostable CutA1 from *P. horikoshii* (Tanaka *et al.*, 2004). This variant of CutA1 denatures at 150 °C. Its structural rigidity is conferred by its tight cylindrical packing through a dense network of interaction hubs in the centre of the trimeric protein. Point mutations in the red circle decrease the denaturation temperature by up to 70 °C, showing the central beta sheets are key determinants of stability (Hirata *et al.*, 2012).

1.2.4 Stable enzymes in synthetic biology

1.2.4.1 Scale-up is treacherous

A key application for the development of synthetic metabolic pathways is penetrance into, and subsequent disruption of chemical markets (Le Feuvre and Scrutton, 2018). In theory, enzymes offer a highly efficient route to chemical synthesis through enantioselective, high purity bioconversions. Novel enzymes therefore allow for the synthesis of second and third generation biofuels, materials, and bulk chemicals for food, perfume, textile, and fine chemical industries (Bornscheuer *et al.*, 2012; Narancic and O’connor, 2017; Guerriero *et al.*, 2015; Le Feuvre and Scrutton, 2018). Additionally, enzymatic pathways can be engineered

to accept economically and socioeconomically important inputs, including plastics (Narancic and O'Connor, 2017; Wierckx *et al.*, 2015), non-food lignocellulosic biomass (Turner *et al.*, 2007; Hollinshead *et al.*, 2014; You *et al.*, 2013; Guerriero *et al.*, 2015), and the often touted “free” resource sunlight (Lips *et al.*, 2018; Baltes and Voytas, 2018; Boehm and Bock, 2019). While promising, the second wave of synthetic biology requires proof-of-principal pathways to be scaled-up into profitable processes. However, repeatedly, synthetic biology encounters problems with the scale-up step (Boehm and Bock, 2019; Liu *et al.*, 2018; Kwok, 2010).

As an example, reactors in the 100 m³ size range (i.e. fed batch bubble reactor) encounter a number of inefficiencies that are not accounted for in initial pathway prototyping undertaken at small scales (Moser *et al.*, 2012). At the macro-scale, systems are homogenised by constant mixing. However, micro-states form throughout the system from fluctuations in temperature, pH, waste accumulation, carbon accessibility and aeration. These lead to sub-optimal growth conditions for microorganisms, causing metabolic burdens, slowed growth rates, and increased ATP consumption (Hollinshead *et al.*, 2014; Hewitt and Nienow, 2007; Wang *et al.*, 2015). Stress responses to such conditions lead to inefficiencies caused by early onset starvation responses, reduction in amino acid synthesis, lowered exogenous pathway expression, shifted metabolite flux, slowed replication and precocious entry into the death phase (Hollinshead *et al.*, 2014). Such responses can also be amplified by quorum effects within the population (Ye *et al.*, 2016). How a synthetic pathway reacts and performs under such conditions is unpredictable until thoroughly tested (Moser *et al.*, 2012). Often, a given process will need to undergo significant refinement through many iterations at both the pathway and organism levels to enable market viable productivity (Park *et al.*, 2008; Julleson *et al.*, 2015; Sanford *et al.*, 2016). As such, it is common for scale-up timescales to be underestimated, leading to complex and high-risk investment roadmaps (Sanford *et al.*, 2016).

The synthetic biofuels market is exemplary of scale-up issues. In 2014, biofuel pioneers LS9 sold for a net loss of \$41 million after failing to obtain funding for scale-up of microbial biodiesel production. Their buyers, Renewable Energy Group, have inserted the LS9 developed technology into pathways for the development of fine chemicals (LaMonica,

2014). Around the same time, biofuels giant KiOR sold their \$215 million plant for \$3.7 million in their 2014 bankruptcy filing, after consistently failing to achieve promised yields and capacity (Fehrenbacher, 2015; Sanford *et al.*, 2016). Solazyme, who initially held partnerships with the US government to generate biofuels from algae, instead now gains commercial value from the sale of synthetic of bio-oils for nutritional products and the culinary industry as TerraVia, which recently sold to Corbion following bankruptcy in 2017 (Lane, 2016; Fehrenbacher, 2016; Corbion, 2017). The history of Amyris tells a similar tale: they initially developed systems for the biosynthesis of biofuel precursor farnesene. Following the company's stock price plummeting 95% (\$800 → \$40 per share) after taking \$34 million in losses in 2017, its fermentation plants were sold, shifting the company's focus to the generation of capital from fine chemical products like fragrances (Bomgardner, 2017; Amyris, 2018). While fuel industries are highly competitive, these issues are indicative of a disconnect between proof-of-concept and commercial application. In many young synthetic biology applications, value only exists in theory. Realisation of value requires effective scale-up to meet market needs and offset high and unpredictable set-up costs. Hence today marketable applied synthetic biology workflows typically target specialist market products of high value. For synthetic biology to ever penetrate, disrupt and maintain within broader markets, technologies that enable efficient biocatalysis at multiple scales are essential.

1.2.4.2 Thermostable workflows could enable scale-up – Lignocellulose case study

The use of stable enzymes and high temperature bioconversions offer a number of benefits over mesophilic systems that could enable more secure and successful scale-up processes (Jemli *et al.*, 2014; Eberhardt *et al.*, 2018; Chubukov *et al.*, 2016; Abdel-Banat *et al.*, 2010; Bommarius, 2015). High temperature catalysis initiated from lignocellulosic biomass is one of the best studied potential applications of thermostable enzymes for the scale-up of synthetic biology (Guerriero *et al.*, 2016).

For a synthetic biology application to be market viable, it has to provide a unique solution to an existing problem, and has to be market competitive enough to generate profit following development (Langerak and Hultink, 2006; Lechner *et al.*, 2016). Lignocellulose is the most abundant carbohydrate source in nature, and is the most readily available fully recyclable

raw material on earth (Guerriero *et al.*, 2016). Derived from plant cell walls, the cellulose component of lignocellulose makes up 40% total plant biomass, offering an low cost waste product derived carbon source to drive synthetic bioconversions (Turner *et al.*, 2007). In theory, lignocellulose bioconversions could be disruptive in many markets, as they drive down the cost of raw materials, allowing for the generation of innovations with consumer costs that provide market penetrative, and potentially market expanding pricing (Srinivasan *et al.*, 1997). Importantly, for this case, lignocellulose derived glucose (plus processing costs) must be cheaper than “off-the-shelf” glucose.

However, glucose is largely inaccessible when packaged into cellulose. Cellulose has a highly crystalline structure, packed by van der Waals forces and hydrogen bonds, making it recalcitrant to hydrolytic enzyme activity (Cheng *et al.*, 2011; Zhao *et al.*, 2012). Biomass utilisation immediately becomes costly due to the need for pre-treatment with high temperatures or chemicals (Hendriks and Zeeman, 2009). Even with pre-treatment, long depolymerisation and hydrolysis times with large volumes of cellulase and glycoside hydrolase enzymes are required for carbon release at useful scales, incurring high manufacturing costs (Fenila and Shastri, 2016). Additionally, pre-treatment processes can lead to the generation of fermentation inhibitors which need to be titrated out of solution (Xia *et al.*, 2013; Long *et al.*, 2018). Therefore, alternative processing strategies are required for biomass derived glucose to be commercially viable. Thermophilic enzymes are one candidate technology that may offer improved biomass utilisation.

Approximately half the projected costs of biomass conversion are associated with enzymatic processing. However, these projections are based on mesophilic systems (Yeoman *et al.*, 2010; Eberhardt *et al.*, 2018; Klein-Marcuschamer *et al.*, 2011). If high temperature pre-treatment regimens for depolymerisation are utilised, thermostable enzymatic desaccharification can potentially be performed immediately in the same pot without the requirement of extensive cooling (Turner *et al.*, 2007; Szijárto *et al.*, 2008; Yeoman *et al.*, 2010). Such high temperature systems have been reported that reach over 70% conversion of cellulosic biomass to respective sugars (Noordam *et al.*, 2018; Long *et al.*, 2018). Saccharified product would then be routed into subsequent fermentations. Optimal fermentation conditions rely on the thermotolerance of the fermentation strain, or stability

of biocatalysts (Yeoman *et al.*, 2010). Large volumes of water are therefore required to maintain consistent temperatures of both the fermenter and any additives to avoid heat stress when mesophilic fermentation chassis are utilised. High temperature bioconversions therefore offer cost-savings through the relaxation of cooling requirements, and the recycling of metabolic heat to maintain high process temperatures (Abdel-Banat *et al.*, 2010; Lin *et al.*, 2014). High temperatures would also decrease the viscosity of plant slurry, allowing for less energy expenditure in system homogenisation. Heat-dependant relaxation of the cellulose crystal structure combined lowered viscosity would increase enzymatic access to substrate, increasing overall process efficiency (Yeoman *et al.*, 2010; Kallionen *et al.*, 2014; Chatterjee *et al.*, 2015; Long *et al.*, 2018).

Additionally, the fermentation environment is hyper-rich in carbon-based energy sources, meaning microorganisms carried in slurries can lead to contamination of fermentations and considerable production bottlenecks. A case study by GE Life Sciences analysed the potential economic impact of a contamination event in a 2000L bioreactor that makes a “blockbuster” drug (Westman, 2017). It was assumed that it would take one month from discovery of contamination to resume bioreactor activity. Costs incurred by scrapping a fermentation batch approach \$1million from lost raw materials, sanitisation expenditure and labour hours. Constant quality assurance is required to ensure a spill over event would not contaminate the next batch once the fermenter is turned back on (Survana *et al.*, 2011). Depending on where in the drug synthesis and purification process the contamination has reached, an additional \$3 million expenditure could be incurred due to sanitising of purification apparatus, and discarding potentially damaged multi-use consumables. If costs were to be considered at a market-level, a ripple of impacts caused by loss of reputation, long lead times, loss of market share and litigation may lead to over \$1billion in lost revenue (Westman, 2017). At large-scales, in any biocatalytic workflow, it is imperative that contamination events are avoided. Considering biomass conversions, high temperature liquefactions and fermentations ensure that only the intended thermotolerant fermentation organisms can survive optimally in the carbon rich environment, especially when microorganisms may be carried with the biomass substrate (Yeoman *et al.*, 2010; Viikari *et al.*, 2008; Kallioinen *et al.*, 2014; Long *et al.*, 2018).

Finally, overall process efficiency is intrinsically linked with enzymatic half-life (specific activity). An enzyme's thermostability and half-life at temperature are typically positively correlated, where thermotolerant enzymes survive for longer in solution than their less stable homologues at the same temperature (Polizzi *et al.*, 2007; Yeoman *et al.*, 2010). An estimated 50% of process costs are related to the utilisation of enzymes. Therefore, decreasing the molar quantity of enzyme while attaining equivalent or improved processivity is essential for bioprocess commercial viability at larger scales (Long *et al.*, 2018; Wu and Arnold, 2013; Klein-Marcuschamer *et al.*, 2011).

1.2.4.3 Stable enzymes also enable scale-up by acting as favourable substrates for protein engineering

An important emerging application of thermostable proteins to synthetic biology scale-up is their use as a favourable engineering substrate (Bommarius, 2015). It has been highlighted that thermostable proteins harbour increased "mutational robustness" when they possess stabilities considerably above the stability threshold (Zhou *et al.*, 2008; Tokuriki and Tawfik, 2009B). This is important when considering the optimisation of proteins for maximum efficiency in a commercialised synthetic metabolic pathway. If proteins in the pathway are to be engineered, an attractive starting point for engineering would be a protein that operates at temperatures far above the native stability of the synthetic system (Khersonsky *et al.*, 2018; Pardo *et al.*, 2018). As discussed, the majority of mutations are destabilising (Faure and Koonin, 2015): a thermostable protein can therefore sample a greater sequence space over the mesophilic counterpart, as fewer mutations are significantly detrimental to protein folding.

In nature it is observed that the innovation of new function typically involves trade-offs with stability, where functional innovation requires proteins to accumulate simultaneous counter-stabilising mutations throughout the structure (Tokuriki *et al.*, 2008).

Destabilisation of engineered biocatalysts is also often encountered when new or improved functions are imparted (Martin *et al.*, 2018). Engineering new function into proteins through directed evolution therefore leads to stability induced optimisation plateaus after only a small number of mutagenesis rounds (Goldsmith *et al.*, 2017). From a fitness space

perspective, minima that exist because of protein destabilisation impede the introduction of mutations required for the traversal across property space. These minima are effectively flattened in stable enzymes until the protein accrues enough mutations for the engineered protein's stability to approach that of the intended system (Porebski and Buckle, 2016). Logically, larger traversals of property space are possible with thermostable proteins due to the shifts in property landscapes. The release of pressure caused by the requirement of counter-stabilising mutations should also make such traversals possible in fewer mutations.

Bloom *et al.* (2006) offer an applied example of this phenomena, in the engineering of novel substrate binding in the cytochrome P450_{BM3} monooxygenase. In this comparative study, P450_{BM3} homologues derived from mesophilic and thermophilic parents were subject to random mutagenesis, generating equivalent mutation distributions throughout the proteins. Activity was then screened on five novel, pharmaceutically relevant substrates. Three-fold more mutants with activity on any of these substrates were identified from the thermostable parent. Additionally, thermostable mutants with novel binding properties were destabilised up to 14 °C, whereas no mutants from the marginally stable protein incurred stability losses >3 °C. In the same study, the authors also exemplified benefits of high stability in the rational engineering of P450_{BM3}. By including a positively charged residue in the active site, binding and catalysis of the negatively charged substrate naproxen was achieved. Modification of the active site caused a 7 °C decrease in stability, which would not be permissible if the same engineering process was performed from the marginally stable homologue (Bloom *et al.*, 2006).

Goldsmith *et al.* (2017) also showed that the engineering optimisation plateau can be overcome by undertaking directed evolution on thermostable proteins. Phosphotriesterase was engineered to scavenge and neutralise V-type nerve agents in the body. Previous studies from a mesophilic counterpart achieved a plateau of ~500-fold improved activity on nerve agents VX and RVX with subsequent rounds of mutagenesis conferring rare and minimal improvements (Cherney *et al.*, 2013). Focused introduction of stabilising mutations into the evolving library allowed for the release from this plateau. 5000-fold improvements in activity on agent VX, and 17000-fold improvement on agent RVX compared to wild type

were subsequently attained – providing a viable candidate biopharmaceutical for low-dose post-exposure treatment of V-type nerve agents (Goldsmith *et al.*, 2017).

As such, very recent studies have begun to front-load existing enzyme engineering frameworks with stabilisation tools, with the aim of simplifying all downstream steps. Trudeau *et al.* (2018) needed to convert an acetyl-CoA synthetase (ACS) into a glycolyl-CoA synthetase (GCS) as a step in re-routing proto-respiration to bypass Rubisco for carbon conservative conversion of solar energy into biomass. As an initial step in the engineering process, the “Protein Repair One Stop Shop” (PROSS; Goldenzweig *et al.*, 2016) stabilisation tool was used to increase the stability of ACS by greater than 10 °C. ACS was then subject to active site engineering using library-based tools to engineer GCS with a 16-fold shift in activity in favour of glycolate. Both the benefits of high temperature systems, and the engineering opportunities afforded by thermostable proteins present necessity for advanced tools that enable access to stable enzymes for use in synthetic biology workflows (Rigoldi *et al.*, 2018).

1.2.5 Obtaining stable enzymes

1.2.5.1 Stable enzymes from nature

Thermophiles require thermostable enzymes in order to survive. Thermophiles therefore present an essential resource for obtaining enzymes with high stability for the use in synthetic biology workflows. Many examples of thermostable protein variants obtained from characterised thermophiles exist in the literature. Notable examples include L-aminoacylases (Toogood *et al.*, 2002; Tanimoto *et al.*, 2008) and γ -lactamases (Hickey *et al.*, 2009), both of which are utilised in the multi-ton biosynthesis of pharmaceutical precursors (Littlechild, 2015; Littlechild, 2017). Additional examples of note include the amylotransferases, which are essential for starch processing into plant-based products used in the food industry (Kaper *et al.*, 2004; Turner *et al.*, 2007), and aldo-keto reductases which are used in the synthesis of primary alcohols (Willies *et al.*, 2010).

In the advent of increasing sequencing capacities, metagenomes provide an important alternative resource for the identification of thermostable proteins (Helm *et al.*, 2018). Metagenomics is the study of mixed genetic pools obtained directly from environmental samples (Tringe *et al.*, 2005). In metagenomic analyses, everything in a given sample is sequenced simultaneously. Predictive algorithms then identify genes of interest in the metagenomic sample, either by analyses of homology with existing gene databases (Huson *et al.*, 2011), or by predictive methods that integrate computational deep learning with training datasets of well annotated genomic samples (Zhang *et al.*, 2017). Homology searches of metagenomic samples isolated from high temperature environments (e.g. hot-springs) have harboured numerous examples of thermostable enzymes that are used industrial workflows (Guazzaroni *et al.*, 2015). Examples include esterases from solfataric field mud holes in Indonesia (Rhee *et al.*, 2004), epoxide hydrolases from microbial mats growing in Russian hot-springs (Ferrandi *et al.*, 2015), xylanases from compost heaps (Verma *et al.*, 2013), and cellulases from elephant dung (Ilmberger *et al.*, 2012).

1.2.5.2 Engineering stable enzymes

While thermophiles offer a rich resource for thermostable enzymes, the discovery of enzymes with both thermostability and desirable activity is rare. As discussed, engineering new function into thermostable enzymes often leads to decreases in stability. Taking natural enzymes with desired function, and engineering them for stability is therefore a desirable workaround. However, engineering proteins for thermostability has been described as “one of the most challenging problems in protein science” (Suplatov *et al.*, 2015). A protein’s stability is the product of complex internal and external molecular interactions. Observations from random mutations scattered throughout protein families has suggested that stabilising mutations occur between one in every 300 to 1,000 random mutations screened (Bloom and Glassman, 2009). Such paucity mutations that increase stability has led to innovations that overcome low probabilities by screening considerably large numbers of mutations across a protein family to find stabilising sets of mutations (Arnold, 2018; Kaushik *et al.*, 2016; Ye *et al.*, 2017). Alternative tools identify refined libraries of select sites that are predicted to harbour stabilising mutations based upon

structural and sequence data (Goldenzweig *et al.*, 2018; Yang *et al.*, 2015). Finally, a newly emerging group of methodologies utilise alignments of protein sequences to inform the selection of amino acids from wider protein families to introduce stabilising bias throughout the protein sequence (Sternke *et al.*, 2018; Gumulya *et al.*, 2018).

Directed evolution for engineering stability

Directed evolution offers a powerful tool for engineering stability. Success in directed evolution hinges on the relationship between the selective pressure applied and the desired trait (Arnold, 2018). Fortuitously, engineering stability by directed evolution requires the application of an exceptionally simple selective pressure – heat (Eijsink *et al.*, 2005). By steadily increasing the temperature of the system, and selecting for functional sequences based on activity at a given temperature, rapid refinements to protein stability can be made (Cherry and Fidanstef, 2003; Denard *et al.*, 2015).

Directed evolution has successfully been employed to push a number of biocatalysts toward thermostability. Early studies quickly identified directed evolution as a method that can circumvent the common trade-off between stability and activity, by screening for both activity and stability in parallel. For example, Giver *et al.* (1998) focused on stabilisation of *p*-nitrobenzyl esterase, an important catalyst in the synthesis of cephalosporins. In their low-throughput methodology, single transformants from random mutagenesis were split into two screening populations. Parallel screening of mutants for both activity on substrate, and activity following heating allowed for the cross-referencing of libraries to identify double-benefit mutants. Experiments led to the identification of mutants with an approximately 15 °C increase in stability following incubation, and a 1.3x increase in specific activity at 30 °C. In another example, Wu and Arnold (2013) screened three generations of 2,800 colonies of a randomly mutated thermostable cellulase for high temperature lignocellulose degradation at 75 °C. While the wild type exhibited a half-life of only two minutes at this temperature, seven additional point mutations produced a variant with a half-life of 4.5 hours, and a 10-fold decreases in saccharification times at 75 °C.

It is generally accepted that thermostability increases of 15 °C and above are “outstanding”, and methodologies for engineering protein stability should strive for such increases (Wijma *et al.*, 2014; Bednar *et al.*, 2015). However, the majority of directed evolution attempts become stuck in the 2-14 °C optimisation range (Wijma *et al.*, 2014). Additionally, the directed evolution work discussed above exemplify how screening capacity is a limiting factor to the discovery of useful properties, where large volumes of sequences are slowly screened for enhancements (Packer and Liu, 2015). Therefore alternative “semi-rational” methods are being developed for the identification of optimal residues from refined libraries that address the numbers issue in traditional directed evolution (Reetz *et al.*, 2008; Kaushik *et al.*, 2016). Examples of recent technological advancements in semi-rational stability engineering are discussed below.

Hotspot identification by free energy modelling

One of the first tools available for the selection and identification of specific stability enhancing residues within a protein was Rosetta (Rohl *et al.*, 2004; Zangellini *et al.*, 2006; Kaufmann *et al.*, 2010). Rosetta is a software suite for the modelling of a protein’s structural behaviour under applied conditions (i.e. substrate binding), and for the prediction of structural changes following sequence modifications (Richter *et al.*, 2011). Given a protein crystal structure, Rosetta is able to sample the conformational space a folded protein can inhabit by sampling possible bond angles that do not disturb the proteins overall 3D conformation (Kaufmann *et al.*, 2010). While stabilisation is not the package’s core role, the Rosetta software suite is able to provide predictions of free energy changes when mutating target sites (Magliery, 2015). Energy calculations are based on structural data, models of solvation, electrostatic interactions, hydrogen bonding, and van der Waals forces (Kaufmann *et al.*, 2010). Mutants are then be scored based on their preference in a given condition against the wild type structure (Jia *et al.*, 2015). Additional stability centric applications have since emerged, for example FoldX (Schymkowitz *et al.*, 2005), where the sole purpose of the algorithm is free energy prediction based on a given crystal structure. FoldX provides atomic resolution interaction predictions to produce energetic calculations of both native state stability and the stability of subsequent mutants (Christensen and Kepp, 2012). It is common for FoldX and Rosetta to be utilised in tandem for false-positive

identification, where residues that both programs identify are considered true candidates for stabilisation improvement.

Stability prediction data from such modelling software are then utilised by third party applications to generate libraries of potential sequence mutations for enzyme stabilisation. One of the most mature adjunct applications is the identification of “hotspots” within the enzyme structure – point mutations that are likely to harbour the greatest stabilising effects (Yang and Wang, 2010). Hotspots are based on highly flexible residues that may lead to high rates of solvent access to the protein core, as well as residues that are highly mutable without damaging function (Bendl *et al.*, 2016). Hotspot identification tools like “Hotspot Wizard” and “FireProt” utilise Rosetta and FoldX in the validation of predicted mutants by modelling free energy changes, allowing for significant library refinement (Sumbalova *et al.*, 2018; Bednar *et al.*, 2015; Wijma *et al.*, 2014). Such highly refined libraries (10^2 - 10^4 variants) can still provide outstanding increases in protein stabilisation (>15 °C). For example, FireProt was validated on the haloalkane dehalogenase DhaA (Bednar *et al.*, 2015). In DhaA, 5,529 point mutations are viable. Corrected for functional determinants, destabilising positions, and false positives, 29 desirable single point mutations were identified based on energy calculations or consensus residues at given sites. This was refined to a final library of only 21 possible sites for the generation of variants based on scores from both FoldX and Rosetta. From this library, a number of stabilising mutants were obtained, including a point mutant that conferred a 16.3 °C improvement in thermostability. Bednar *et al.* (2015) then assessed combinations of predicted beneficial mutations for additivity, and isolated a DhaA variant with eleven mutations that presented a 24.6 °C improvement over wild type. The majority of these mutants introduced bulkier hydrophobic residues that improve packing, with a few residues predicted to improve rigidity.

FRESCO (the Framework for Rapid Enzyme Stabilisation by Computational libraries) is another FoldX and Rosetta driven method for enzyme stabilisation (Wijma *et al.*, 2018; Fürst *et al.*, 2018). FRESCO is similar in workflow to FireProt, but also includes an analysis of structure for potential sites that could permit disulphide bridges. Additionally the refinement of FoldX and Rosetta identified sites involves molecular dynamics to identify mutants that lead to an increased flexibility that may counteract potential stabilisation

predicted by free energy calculations (Rigoldi *et al.*, 2018). FRESCO also follows the same confirm-combine strategy as FireProt. It is predicted that over 10% of the mutants in a FRESCO library are stabilising, compared to the 1 in 1,000 conservative estimate for random mutagenesis (Wijma *et al.*, 2018; Bloom and Glassman, 2009). Wijma *et al.* (2014) provided the first validation of FRESCO, where a library of 64 limonene hydroxylase mutants yielded a final mutant with a 35 °C increase in stability over the wild type. Importantly, FRESCO proves robust for the engineering of other traits under the umbrella of stability. Arabnejad *et al.* (2017) utilised the FRESCO framework for the engineering of co-solvent compatibility in a halohydrin dehalogenase type C (HhdC). The resultant twelve residue mutant showed no decrease in relative activity after five hours incubation in 50% (v/v) methanol, acetonitrile, or dimethylformaldehyde. On the other hand, the wild-type enzyme showed negligible activity in the three solvents at the same concentration at time zero. Interestingly, the enzyme also showed a 28 °C improvement in thermostability, providing further evidence that engineering for thermostability is a powerful method to engineer for a swathe of stabilising effects.

B-fitting

B-fitting follows a similar philosophy to FireProt and FRESCO, whereby optimal targets for stabilisation are selected based on the protein's structure. B-fitting instead optimises B-factors instead of free energy changes (Yu and Huang, 2014). B-factors represent atomic displacement observed in crystal structures, and high B-factors signify regions of high flexibility in the protein (Parthasarathy and Murthy, 2000). In principal, B-fitting improves protein stability by identifying the most flexible regions of the protein and suggesting rigidifying mutations in place, solidifying interactions with the protein solvent shell, and restricting bulk solvent access to the protein core (Gao *et al.*, 2018; Yu and Huang, 2014; Reetz and Carballiera, 2007). In one exceptionally successful example from Reetz *et al.* (2006), B-fitting identified ten highly mobile residues throughout a mesophilic lipase (LipA). From these ten sites, eight saturation mutagenesis libraries were generated, which were subjected to stepwise selection based on the enzyme's ability to survive 15 minutes of incubation at increasing temperatures. The best mutant identified from this initial generation was only 4.3 °C more stable than wild type (50 °C). This hit was then taken

forward and subject to iterative saturation mutagenesis at the site that gave the second best increase in stability. Continued stepwise mutagenesis was then performed across the entire library, whereby stabilising mutations were continually accumulated. Two final eleven residue mutants showed no discernible loss in activity after 15 minutes of incubation at 100 °C, which remains one of the greatest engineered increases in protein stability published to date. Furthermore, half-life was improved approximately 490-fold, yet activity was not sacrificed. As with HhdC described above, Reetz *et al.*, 2010 showed that the same enzyme was also highly stable in the presence of harsh solvents compared to wild-type. Half-lives in 50% (v/v) acetonitrile, DMSO or dimethylformaldehyde were improved approximately 80-fold, 22-fold and 19-fold respectively.

1.2.6 Optimising the engineering of stable enzymes for synthetic biology

1.2.6.1 Stable protein engineering may not solve optimisation bottlenecks

In the first wave of synthetic biology, modularisation of genetic parts allowed for the accelerated prototyping of synthetic systems (Purnick and Weiss, 2009). This creates a bottleneck in the scale-up of synthetic biology processes, where treacherous systems optimisation finds many promising prototypes become non-viable economically based on the unpredictability at the protein and system level (Boehm and Bock, 2019; Liu *et al.*, 2018). Such unpredictability can be circumvented to some extent by the utilisation of stable enzymes, as they allow for efficient traversals of sequence space in protein optimisation, while providing process-benefiting properties (Jemli *et al.*, 2014). Therefore, broad access to stable enzymes may circumvent some of the bottlenecks preventing the progression of synthetic biology applications beyond a prototype (Bommarius, 2015).

It is apparent that none of the widely adopted methods discussed fit with the synthetic biology ethos. Access to stable proteins from nature rely on broadly accessible thermophile metagenomes (Guazzaroni *et al.*, 2015), which do not exist at the time of writing. Considerable expertise and investment in metagenomics are therefore required to extract candidate proteins. There is also no guarantee that desired enzymes will exist in a given metaproteome, nor is there guarantee that a candidate enzyme will possess stabilities or

activities that are “right for the job” (Lehmann *et al.*, 2000). While directed evolution and semi-rational based methodologies have been utilised to produce outstandingly stabilised protein variants (>15 °C; Giver *et al.*, 1998; Reetz *et al.*, 2006), the methodologies are also poorly accessible, and present a range of challenges (Bommarius, 2015). Directed evolution is slow, and requires considerable expertise to undertake (Ravikumar *et al.*, 2018). Semi-rational design, while computer aided, mandates access to protein crystal structures for success, which are not always available (Chen *et al.*, 2012). Even when they are, considerable expertise on the relationship between protein structure and function are required to achieve exceptional increases in stability (i.e. the requirement of checking for reasonable mutations by visual inspection (Wijma *et al.*, 2014). Furthermore, all methods require considerable time investment to undertake, with an unknowable number of revisions required to reach stabilisation (Wedge *et al.*, 2009). It is also not known whether exceptional stabilisation is possible with a select family, until optimisation plateaus are reached (Goldsmith *et al.*, 2017). The generation of stable enzymes, rather than the optimization of systems, produce another confounding factor in the unpredictable roadmap to scale-up.

Instead, it may be beneficial to once again apply focus to democratisation. Standard methodologies that enable the rapid and open source stabilisation of enzymes may allow for the widening of the optimisation bottleneck in synthetic biology. To draw an example from DNA technologies, CRISPR is considered to be one of the biggest innovations in the genetic engineering field based on its simple requirements, its high success rate and its low cost (Cong *et al.*, 2013). These factors have allowed for broad dissemination of the tool, and rapid innovation (Doudna and Charpentier, 2014). An equivalent tool for protein engineering would be easily accessible and expert-agnostic, allowing a broad user base to develop stable enzymes for implementation and testing in synthetic biology frameworks. Additionally, to be broadly accessible, the tool should be low cost, requiring both resource and time investment to be minimised (Endy, 2005). Finally, the tool should have short generation times, where potentially beneficial variants can be identified and tested rapidly (Sun *et al.*, 2014). Failure to obtain stable enzymes does not then generate huge process set-backs, and the optimisation roadmap can be more confidently predicted.

1.2.6.1 Alignment-only methods for stability engineering

Recent research has emerged on the utilisation of alignment-only methodologies to induce stabilising bias into sequences. Such methodologies require only multiple sequence alignments of proteins as input, and generate thermostable candidate proteins by selecting optimal states within variable sites in the alignment using freely available software (Durani and Magliery, 2013; Magliery *et al.*, 2015). Interestingly, all other stabilisation methods discussed take a bottom-up approach to enzyme engineering, implementing small numbers of highly beneficial mutations obtained over iterated screens into a final enzyme. In contrast, alignment-only methods are top-down. Large volumes of point mutations are scattered throughout the protein's highly variable sites. Despite the large volumes of mutation, the methods are theoretically likely to produce functional sequences as any inferred mutation is derived from functional residues already present in the multiple sequence alignment (Porebski and Buckle, 2016). However, the underlying mechanisms driving stabilisation, and factors defining successful stabilisation with alignment based technologies are poorly understood. Nevertheless, as the requirements for such methodologies are simply a multiple sequence alignment of homologues and freely available software, the barrier to access such methods is extremely low. Alignment based methods are therefore excellent candidates for democratised protein stabilisation tools (Sternke *et al.*, 2018; Porebski and Buckle, 2016). Alignment-based engineering can be split into two fields – consensus sequences and Ancestral Sequence Reconstruction (ASR).

Multiple sequence alignments

At the heart of alignment-based protein stabilisation are algorithms for generating accurate multiple sequence alignments. Such algorithms take lists of homologous sequences as input, and positionally align related amino acids between proteins to one another. Homologous proteins are typically identified utilising local alignment tools that search databases of sequences for hits with localised pairwise similarity to a query sequence (Altschul *et al.*, 1990). Basic Local Alignment Search Tool (BLAST) is the most commonly utilised software for homologue searches (Madden, 2013). Another commonly used tool is Pfam, which uses Hidden Markov models for the statistical inference of homology between proteins and their subdomains. Within the Pfam database, protein domains are grouped into families based on

their homology with domains of defined function (Finn *et al.*, 2014). Multiple sequence alignments of homologous sequences provide an understanding of historical relationships between sequences, and provide a visual tool for identifying conserved versus non-conserved sites across a protein structure. These in turn can help identify relationships between a protein sequence and its functional and fitness landscapes (Notredame, 2002; Phillips *et al.*, 2000).

Most modern alignment software utilises a progressive algorithm (Feng and Doolittle, 1987). Here, the most similar pair of sequences are first aligned. Progressively, the second most similar are aligned, and third most similar and so on until all sequences are aligned to one another. Typically, when computing pairwise alignments sequences are compared based on a comparison model (Notredame, 2002). This defines likelihood values for pairwise evolutionary relationships between residues, and defines criteria on which to infer a gaps in the alignment (denoting insertion-deletion [indel] events; Edgar, 2004; Phillips *et al.*, 2000). As searching for globally optimal alignments of sequences becomes exponentially more computationally expensive with each new sequences, alignment software utilises a heuristic approach that is tuned to give close-to-optimal answers. Typically, the algorithms will identify highly similar sets of sequences first, radiating outwards for the placement of more dissimilar regions. Pair-wise alignments are based on scoring systems which can be based on alignment motifs, amino acid properties, evolutionary history, models of amino acid replacement (Notredame, 2002). Logical likelihood scores are computed for each position, which in turn influence the alignment of future aligned residues (Notredame, 2002; Thompson *et al.*, 1994). Gaps are also penalised, with larger gaps incurring higher penalties, as indels become objectively rarer with length as a function of the protein's total length when assuming homology (Vingron and Waterman, 1994).

While these methods are typically computationally non-intensive, it is possible to propagate errors throughout the alignment if sequences are misaligned early on in the process (Dickson *et al.*, 2010). Therefore, multiple sequence alignments come with a caveat of being a "best guess", and should typically be accompanied by refinement of the alignment by eye (Notredame, 2002). Tools also exist for refinement, for example GBlocks (Talavera and Castresana, 2007), which is able to identify and remove ambiguously aligned regions from a

given alignment. Commonly used open source algorithms for the multiple sequence alignment of proteins are described in table 4. It should be noted that benchmarking finds the algorithms to have within 2% accuracy to one another, with some algorithms performing better or worse depending on the benchmarking dataset. Algorithm choice therefore often comes down to personal preference (Le *et al.*, 2017).

Algorithm	Method	Notes	Reference
Clustal	Scoring system with weighting, integration of substitution matrices, and gap penalties based on hydrophobicity.	Highly complex scoring algorithm allows for more biologically accurate alignments	Thompson <i>et al.</i> , 1994
MAFFT	Fast Fourier transform for rapid homology detection. Simplified version of Clustal scoring system.	CPU-light. Up to 100x faster than other methods when large numbers of sequences (>60) are to be aligned.	Katoh <i>et al.</i> , 2002
MUSCLE	K-mer counting for identifying homology. Tree-guided scoring. Refinement based on sub-trees.	Most accurate default algorithm in some benchmark tests. Decrease in accuracy with large alignments (Le <i>et al.</i> , 2017)	Edgar, 2004
T-Coffee	Makes all pairwise alignments first using Clustal. Alignments are then built based on the library of pairwise alignments by library extension.	Most accurate default algorithm in some benchmark tests (Le <i>et al.</i> , 2017). Most computationally intensive algorithm.	Notredame <i>et al.</i> , 2000

Table 4 - a non-exhaustive list of freely available, commonly used MSA tools for protein sequences

Consensus protein design

Consensus proteins are derived directly from a multiple sequence alignment. A novel protein sequence is constructed from the most common amino acid to exist at each alignment position (figure 8). Lehmann *et al.* (2000) were the first to identify that sequence-wide sets of consensus sequences can lead to protein stabilisation. The researchers posited that every amino acid contributes to stability, and therefore modifying large numbers of residues produces an increased chance of obtaining stabilising residues. A consensus phytase was derived from an alignment of thirteen sequences that possessed a 15 °C increase in stability over its constituents. Addition of six more sequences into the alignment generated a consensus sequence with a stability improvement of 22.4 °C. By conducting saturation mutagenesis at each site that varied between the two stable consensus proteins,

it was observed that not all substitutions are stabilising. However, the sum of all substitutions is net-stabilising, where each substitution has small effects (<3 °C) on overall protein stability. Such an effect highlights that consensus stabilisation is global, and not contingent on the discovery of small numbers of hyper-stabilising mutations like previously discussed semi-rational methodologies (Rigoldi *et al.*, 2018). Importantly, the majority of amino acid replacements in consensus phytases regions of highly disordered structure, where large ranges of sequence space are being rapidly sampled, and secondary structures were highly flexible. However, the authors state that it is was not possible to clearly characterise the underlying mechanisms of stabilisation (Lehmann *et al.*, 2000).

Salmon	MAFWLQAASLLVLLALSPG-ADAAAVQHLCGSHLVDALYLVCGEKGLFYNPKRDVDPLI-	57
Frog	MALWMQCLPLVIVLFFSTPN-TEALVNQHLCGSHLVEALYLVCGERGFFYYPKVKRDMEQ	59
Chicken	MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ	60
Mouse	MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED	60
Human	MALWMRLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED	60
Pig	MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN	60
	** : : * : * : : * . ***** ** :***** :*:** ** :	
Consensus	MALWMRLPLLALLALXGPXPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARRXXEX	60
Salmon	GFLSPKSAQDNEEFPFKQMEMMVKRGVSVEQCCHKPCNIFDLQNYCN	104
Frog	ALVSGPQDNELDGMQLQPQEYQKMKRGIVEQCCHSTCSLFQLESYCN	106
Chicken	PLVSSPLRGEAGVLPFQQEYEVKVRGIVEQCCHNTCSLYQLENYCN	107
Mouse	PQVEQELGGSPGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN	107
Human	LQ-----GSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN	98
Pig	PQAGAEELGGGLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN	107
	: * * * * : * * * . * : : : * : * * * *	
Consensus	PQVSEXLGGELGGLQFLALEXXKQKRGIVEQCCXSICSLYQLENYCN	107

Figure 8 - A consensus sequence generated from an alignment of insulin homologues

Sequences in the alignment were identified by BLASTp using the human insulin sequence as query (Altschul *et al.*, 1990). Sequences were aligned with Clustal Omega (Sievers *et al.*, 2011). Consensus sequences were generated within the Geneious v.10 sequence handling software suite based on the most common sequence at a given site (Kearse *et al.*, 2012). Black residues denoted X in the consensus marks sites where a majority cannot be reached.

Consensus protein design has since been utilised for the successful engineering of a number of protein families (i.e. Dai *et al.*, 2007; Sullivan *et al.*, 2011; Porebeski *et al.*, 2015; Paatero *et al.*, 2016; Okafor *et al.*, 2018 and Sternke *et al.*, 2018). However the true success rate of the methodology is difficult to discern as failures are unlikely to be published (Sternke *et al.*, 2018). In a recent study by Sternke *et al.* (2018), six consensus proteins were generated

from diverse protein families. Four of the six were identified as more stable than their constituent sequences, and all were functional, suggesting that consensus methods are broadly applicable and generally robust. Potential issues highlighted in the literature centre around the ratification of sites where there is ambiguous signal, and the correct “cut-off” required to generate optimal answers (Okafor *et al.*, 2018; Porebski and Buckle, 2016; Paatero *et al.*, 2016).

Figure 8 represents a consensus of insulin homologues. It can be seen that across the 107 amino acid alignment, ten sites do not form a consensus, and therefore no true consensus sequence can be identified. Typically, to circumvent this issue, residues are inserted as a random choice between the most likely sequences (Lehmann *et al.*, 2000), multiple consensus sequences are constructed (Sullivan *et al.*, 2011), or through alignments (>1000 sequences) are constructed so ambiguities are considerably rare (Sternke *et al.*, 2018; Porebski *et al.*, 2015). Defining the cut-off for consensus engineering involves the identification of specific sites to be input into a known protein sequence based on the consensus. By defining a “cut-off” of 50%, only residues where the consensus represents over 50% of amino acids at a position are output. It has been observed that restricting consensus residues at positions that are more conserved, ignoring those that are highly variable, can lead to improved stabilisation and success in producing functional consensus proteins (Bershtein *et al.*, 2008; Sullivan *et al.*, 2012; Durani and Magliery *et al.*, 2013). However, it is not possible to know the best cut-off for a given alignment until all possible cut-offs are performed (Okafor *et al.*, 2018).

Despite extensive successes, with stability improvements ranging from 5.5 °C (Dai *et al.*, 2007) to 32 °C (Paatero *et al.*, 2016), the consensus protein engineering field does not yet fully understand the underlying mechanisms behind stabilisation. In a review of consensus sequence engineering, Porebski and Buckle (2016) argue that a consensus residue typically has a greater stabilising effect at a given site than a random residue as consensus residues represent an ancestral state derived from some thermophilic ancestor. This is reasonable as it is thought that ancient life evolved in far warmer environments than those that exist today (Gaucher *et al.*, 2008). Unfortunately, falsifying or supporting this hypothesis is difficult as direct observation of ancient organisms and their environments is not possible.

Additionally, Bershtein *et al.* (2008) show that stabilising residues compensate for destabilising residues when proteins are subjected to multiple rounds of mutation, simulating drift. As a result marginally stable proteins tend to conserve stabilising residues between closely related families. Consensus sequences may therefore be the amalgamation of multiple conserved stabilising residues that were innovated at different points in the protein family's evolution (Porebski and Buckle, 2016). This is contentious as one could also expect compensated destabilising residues to be readily amalgamated in global-consensus sequences, especially as destabilising mutations are far more common than stabilising mutations (Bloom and Glassman, 2009). Therefore, as the accepted underlying stabilisation mechanism is not fully understood, and as there is no defined methodology to consistently produce functional stabilised consensus proteins, consensus sequences may not be the ideal broadly accessible solution for protein stabilisation (Okafor *et al.*, 2018).

Ancestral Sequence Reconstruction

Ancestral Sequence Reconstruction (ASR) is a tool for generating statistical predictions of a protein family's ancestry based on an alignment of sequences, a phylogenetic tree generated from this alignment, and a model of amino acid substitution. ASR will output sequences that represent a sequence with the maximum statistical likelihood to have evolved into all sequences in the alignment, or a related subpopulation of the alignment. ASR studies of diverse protein families have identified emergent properties of ancestral proteins, including increased thermal stability and altered substrate specificities (i.e. Okafor *et al.*, 2018; Nguyen *et al.*, 2017, Shih *et al.*, 2016, Wilson *et al.*, 2015; Risso *et al.*, 2015). Consequently, a series of studies have probed evolutionary history to isolate sites of interest to engineer enzymes with novel functionality (i.e. Alcolombri *et al.*, 2011; Conti *et al.*, 2014; Gonzalez *et al.*, 2014; Miyazaki *et al.*, 2001; Watanabe and Yamagishi, 2006). Additionally, traits that are useful to bio-industry such as thermostability, improved expression, broadened or tightened substrate range and structural simplicity are reported in the contemporary ASR literature. Therefore, there has been rising interest in the use of ASR primarily as an engineering tool (i.e. Babkova *et al.*, 2017; Gumulya *et al.*, 2017; Wilding *et al.*, 2017; Gumulya *et al.*, 2018; Risso *et al.*, 2018). As this thesis will focus on the utilisation and development of ASR as an engineering tool, an in depth analysis of ASR is presented.

1.3 ASR as an efficient tool to engineer protein stability

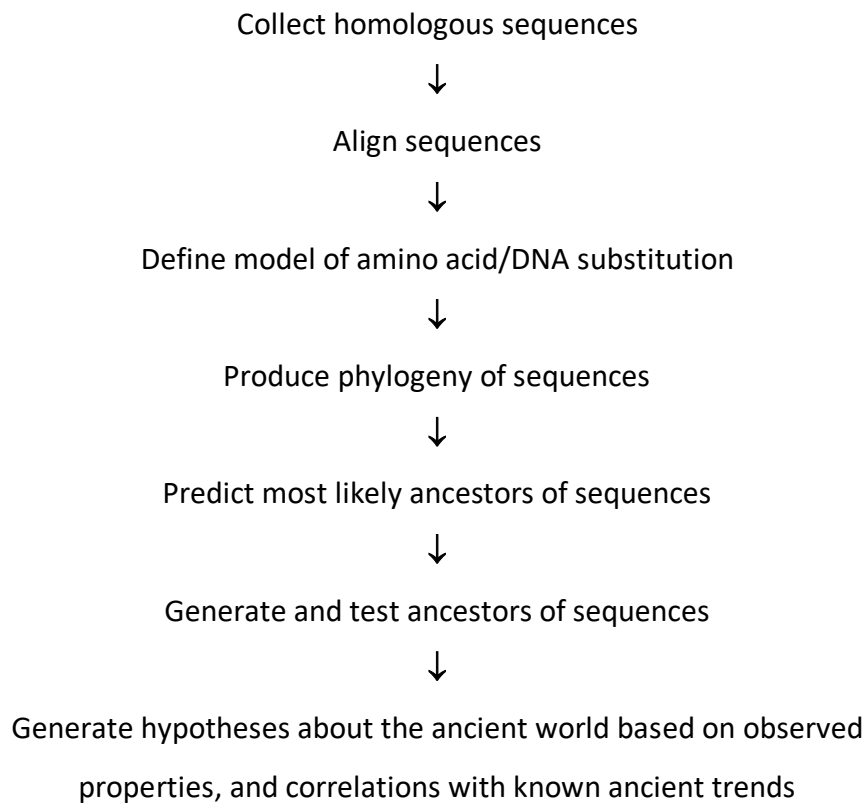
1.3.1 Methodologies for ancestral sequence reconstruction

1.3.1.1 Conception of ancestral sequence reconstruction

ASR was first described in conjecture by Pauling and Zuckerkandl (1963). In principal, homologous sequences contain information about their shared ancestry, and the degree of difference between the two sequences represents a measure of time at which two homologous sequences separated from one another (Joy *et al.*, 2016). Therefore predictions of ancient states can be made based on shared states between modern (extant) sequences. Probabilistic models of amino acid change are used to derive the most likely ancestral amino acid to have produced the modern state at any site that varies between two homologues (Pupko *et al.*, 2000). Pauling and Zuckerkandl were arguably visionary in their perspective, laying the groundwork for a field that would not take shape until considerable progress was made in both evolutionary biology, computational biology and DNA synthesis some 30 years later (Gaucher, 2007).

Pauling and Zuckerkandl's vision was not put to the test until Stackhouse *et al.* (1990) reconstructed ancestors of pancreatic RNases in ruminants. A parsimonious phylogeny (See section 1.3.1.2 for a discussion on phylogenetics) was generated describing the evolution of ruminant RNases. This prediction aimed to satisfy Occam's razor (predictions requiring less speculation are less spurious; Steel and Penny, 2000). Parsimony was used to predict ancestral residues at two positions, producing an RNase belonging to the ancestor of ox, swamp buffalo and river buffalo from the Pliocene era. Modular mutagenesis was used to generate the ancestral sequence, which was shown to be functional, and functionally equivalent in activity and stability to extant sequences (Stackhouse *et al.*, 1990). Jermann *et al.* (1995) extended this study to the ancestor of artiodactyl RNases, utilising the same method to predict ancestral amino acids across 24 sites. Ancestors possessed extant-equivalent activities. However, an increase in stability (≈ 2 °C) was observed in ancestors that occur after the innovation of foregut digestion in true ruminants. Increased stability was hypothesised to aid complex digestion of grasses where 20% of dietary nitrogen is obtained from digested nucleotides. The two seminal studies by Stackhouse *et al.* (1990),

and Jermann *et al.* (1995) form the groundwork for modern ancestral sequence reconstruction workflows, which can be summarised as:



1.3.1.2 Underlying principles for contemporary ASR

Phylogenetics – key principals

Central to any ancestral reconstruction is the phylogeny, where phylogenetic accuracy and reconstruction accuracy are directly related (Vialle *et al.*, 2018). A phylogeny is a graph describing the evolutionary relationship between individual DNA or amino acid sequences. Phylogenies describe the series of speciation events that were most likely to have produced the consortia of sequences provided in the sequence alignment (Wiley and Lieberman, 2011). These relationships are displayed as the distance between sequences where distance can be represented as either real time (calibrated to real data from the fossil record), or as the number of substitutions per site (Baum and Smith, 2013). Phylogenies are represented as trees, where the leaves of the tree are modern sequences, the branches of the tree represent the distance (amount of divergence) between sequences, and nodes represent

branching points between sequences (figure 9; Yang and Rannala, 2012). These relationships can only be inferred, as evolution is rarely observed due to million-year time scales. Exceptions include experimental phylogenies (Randall *et al.*, 2016), and phylogenies of extremely rapidly evolving sequences like viruses (Moratorio and Vignuzzi, 2018). Modern techniques to infer phylogenetic trees centre around two algorithms: Maximum likelihood and Bayesian inference (Lemmon *et al.*, 2009).

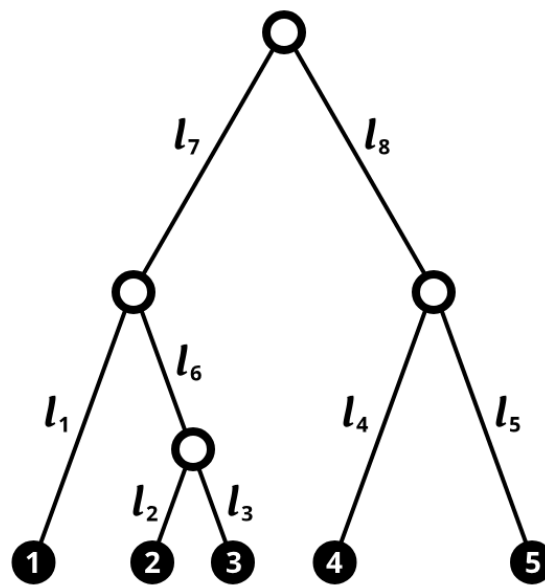


Figure 9 - A hypothetical phylogenetic tree

Modern sequences are represented by numbered circles. Branches are represented by a length (l) that is calculated as a distance from all other sequences in the tree. Nodes (block hollow circles) represent the branching point between homologues. The distance between sequence 2 and 3 is represented by $l_2 + l_3$. The distance between sequence 2 and sequence 4 is represented by $l_2 + l_6 + l_7 + l_8 + l_4$.

Both algorithms are classified as character-based methodologies, where all sequences in the alignment are considered simultaneously, and possible tree topologies are scored according to each individual site in the alignment (Huson and Bryant, 2005). In theory, the best scoring tree for all positions is the most accurate representation of the true phylogeny within the limitations of the algorithm and the alignment provided (Yang and Rannala, 2012). As the number of possible trees increases exponentially with each new branch, tree scores are

typically computed by heuristically searching through the global tree space. Heuristic searches assume that similar trees will have similar probabilities, therefore small rearrangements allow for step-wise optimisation of the tree score (Guindon and Gascuel, 2003). In order to define a basis for scoring, the algorithm must have some prior knowledge of evolutionary patterns. For protein trees (the focus of this thesis) relationships between sequences are described by a model of amino acid substitution.

Amino acid substitution models are matrices of likelihood values that define the probability of one amino acid changing into another. It has been observed that amino acids do not mutate into every other amino acid with equal likelihood, for two reasons (Yang *et al.*, 2000). Firstly, amino acid substitutions are driven by codon changes, therefore amino acid substitutions differing by a single base (e.g. valine to alanine) are more likely than substitutions requiring two, or occasionally three base changes (e.g. isoleucine to tryptophan). Secondly, substitutions are more likely between amino acids with functional and structural similarity, as such changes are less likely to disrupt the function or stability of their parent protein (Rodrigue *et al.*, 2010; Le and Gascuel, 2010; Betts and Russel, 2003). A number of substitution matrices have been generated in the literature that attempt to model the average substitution rates for all proteins (Le and Gascuel, 2008). Modern phylogenetics tends to lean on substitution matrices (table 5) derived from observed amino acid changes in large databases of protein families (Le and Gascuel, 2010; Yang and Rannala, 2012).

Model	Notes	Reference
JTT	First of its kind. Computed conditional probability of amino acid substitution based on mutations observed SWISS-PROT database in release 15.0 (approximately 17000 sequences). Computed from 3905 sequences.	Jones <i>et al.</i> , 1992
WAG	Incorporates phylogenetic information from alignments. Utilises approximate maximum likelihood based predictions Computed from approximately 50000 sequences in 3,912 alignments.	Whelan and Goldman, 2001
LG	Incorporated variable rates across sites when computing substitution rates.	Le and Gascuel, 2008

Table 5 - Commonly used substitution matrices derived from large databases of aligned proteins

As well as amino acid substitution rates not being uniform, the rate of mutation varies between positions across the protein, where some sites mutate far less, or far more than average (e.g. active site vs flexible loop; Yang, 1994; Gu *et al.*, 1995). It is typically observed that across-site substitution rates can be modelled to a gamma distribution (Yang, 1996). Gamma distributions are a two-parameter family of continuous probability distributions which are defined by a shape parameter (κ) and a scale parameter (θ). For the sake of computational simplicity, continuous gamma distributions are often distilled to a discrete distribution consisting of the median values of the distribution's quartiles or octiles (Golding, 1983; Yang 1994; Yang 1996). The shape of the gamma distribution is typically estimated from the number of changes present across the alignment (Yang, 1996). Furthermore, some residues essentially never change, and therefore invariability is also modelled by attributing some sites with a substitution rate of zero (Gu *et al.*, 1995). When defining models of amino acid substitution, the assistance of predictive tools is often required as it is not possible to know which model best fits a given alignment, and thus has the highest likelihood of generating accurate phylogenies. For example, ProtTest (Abascal *et al.*, 2005) utilises an Akaike information criterion to estimate the quality of an amino acid substitution model given an input alignment. Models can then be ranked allowing the user to make an informed decision on model utilisation in a phylogenetic experiment (Darriba *et al.*, 2011).

Maximum likelihood phylogenies

Maximum likelihood is a statistical methodology designed for the estimation of unknown parameters in a model. Therefore, given sets of data, maximum likelihood computes the likelihood of the data derived from the given constants. Conversely, given fixed data, maximum likelihood also computes the likelihood that a constant exists. In phylogenetic tree-building, the correct topology of the tree is treated as an unknown constant, and the known constants are the alignment and the model of amino acid substitution (Felsenstein, 1981). Maximum likelihood software will then heuristically search tree space to find the tree with the highest likelihood for a given alignment and substitution model (Guindon *et al.*, 2010). In a similar manner to which directed evolution can be described as a series of uphill walks that traverse toward optimal sequence space (figure 5), maximum likelihood inference of a phylogeny can be considered a hill-climbing optimization over tree topology

space. To minimise computational effort, it is typical to generate small subtrees from the global tree, and locally optimise their orientations first (Dhar and Minin, 2015). Then additional branches or optimised subtrees can be added, and the orientation optimised. By adding branches in different orders, with different starting subtrees, different highly optimal tree topologies are searched (Yang and Rannala, 2012). At the same time, given the model of amino acid substitution and the alignment, an optimisation of branch lengths is performed. Here, branch lengths are modified branch-wise or jointly, to optimise their likelihood. This is performed over many passes until significantly small optimisations are observed (Felsenstein, 1981; Dhar and Minin, 2015). The final maximum likelihood tree is generated as the largest product of site-wise probabilities in the alignment given the tree topology, computed when assuming unequal rates of evolution between sites (Felsenstein, 1981; Yang and Rannala, 2012).

During the heuristic search, the order in which data is considered can lead to alternative optimal topologies. It is therefore generally appropriate to provide some score of topological confidence at each node on the tree (often called branch supports; Stamatakis *et al.*, 2008). The most common method used to compute support values for maximum likelihood trees is non-parametric bootstrapping (Dhar and Minin, 2015). Bootstrapping re-samples optimal trees from the total dataset by generating a new alignment from random columns of the true alignment with replacement (Felsenstein, 1995). Trees are then optimised from this dataset using the same maximum likelihood approach. Comparisons of branching topology between the most optimal subtrees and the consensus tree, and the percentage of matching branching topologies are then reported for each node on the tree of the true alignment (Dhar and Minin, 2015; Yang and Rannala, 2012; Felsenstein, 1985). In theory, as columns in the alignment data are treated independently, an alignment that is robustly modelled by the tree should produce consistently similar tree orientations to the starting dataset orientation. Bootstrap values above 70% are typically accepted as accurate (Hills and Bull, 1993). Alternatively, statistical tests that assess the likelihood of sites given a set of bootstrap trees (SH test; Shimohaira and Hasegawa, 1999), or the confidence of branch positioning against a null hypothesis of a collapsed 0 length branch in the same position (aLRT; Anisimova *et al.*, 2011) can be utilised as a rapid alternative to bootstrapping (Guindon *et al.*, 2010). Studies show that statistical tests can be equally robust to

bootstrapping, and choice of topology confidence scoring method often depends on the algorithm being used and ultimately the researchers preference (Anisimova *et al.*, 2011). Commonly utilised algorithms for maximum likelihood tree-building are summarised in table 6 (Yang and Rannala, 2012).

Algorithm	Notes	Reference
PHYLIP	Maximum likelihood calculation of phylogenies by hill-climbing with both local and global rearrangements. Confidence can be computed by bootstrapping.	Felsenstein, 1995
PAUP	Hill-climbing that can search all trees by brute force, or search partial trees by adding branches or swapping branches. Confidence can be computed by bootstrapping.	Swofford, 2001
PhyML	Fast maximum likelihood implementation with simultaneous hill-climbing optimisation of branch lengths. Rapidly reaches optima. Confidence can be computed by bootstrapping, SH test, aLRT test.	Guindon and Gascuel, 2003
RAxML	Rapid and accurate algorithm that can also be parallelised for computation of large datasets. Optimises a parsimony tree that is pre-predicted by re-insertion of subtrees. Confidence can be computed with bootstrapping, rapid bootstrapping and SH test.	Stamatakis <i>et al.</i> , 2005

Table 6 - Commonly utilised algorithms for inference of maximum likelihood phylogenies

Bayesian phylogenies

While Maximum Likelihood phylogenies are a powerful and often used tool for phylogenetic inference, the calculation of their support values is contrived. Bootstrapping and statistical inference do not relate directly to the biological data (Yang and Rannala, 2012).

Considerable debate on the true meaning of bootstraps has been undertaken in the literature, and multiple definitions are acceptable (summarised in Berry and Gascuel, 1996). Instead, Bayesian frameworks for inferring phylogenies avoid this issue by providing posterior probability values of each node position, calculated directly from the tree and data (Huelsenbeck *et al.*, 2001; Mau *et al.*, 1999). The confidence scores provided in Bayesian analysis state “the likelihood of the tree is X” and are therefore considerably simpler to interpret than maximum likelihood trees (Yang and Rannala, 2012). Due to such benefits, and access to a highly parameterisable and accessible algorithm in MrBayes (Huelsenbeck and Ronquist, 2001), Bayesian inference is favoured by some researchers.

Bayesian inference of a phylogeny relies on Bayes's theorem (Huelsenbeck *et al.*, 2001):

$$f(A|X) = \frac{f(X|A)f(A)}{z}$$

Where A is the unknown parameter (the branch lengths and divergence times of the tree), X is some observed data (the multiple sequence alignment), and $f(A)$ is the prior distribution, or what we know about the unknown parameter before analysing the data (a tree topology and substitution model). $f(A|X)$ is the posterior probability (the likelihood of a parameter given the observed data) and, $f(X|A)$ is the information that is known about A due to the data observed. Finally z is the normalising constant that allows us to ensure that $f(A|X)$ is a proper statistical distribution (Nascimento *et al.*, 2017; Huelsenbeck *et al.*, 2001). An in depth descriptions of the underlying methods applying Bayes theorem to phylogenetics is beyond the scope of this thesis; a thorough description of the methodology and its application is presented between the works of Huelsenbeck *et al.* (2001) and Nascimento *et al.* (2017).

In brief terms, it is not possible to generate an exact calculation of the complete posterior distribution due to the existence of huge numbers of possible phylogenies for a given alignment. Therefore the posterior distribution is sampled using a Markov Chain Monte Carlo (MCMC) process that allows for the sampling to approach the most optimal tree topology (Nascimento *et al.*, 2017; Huelsenbeck and Ronquist, 2001). MCMC is iterative, where a random tree is first defined, and a likelihood is calculated for the given tree. Another tree is then defined in nearby tree space (a change in branch lengths or topology), and its likelihood is compared to that of the previous generation (Larget and Simon, 1999). If the newly calculated likelihood is very close to the previous likelihood, or an improvement over the previous likelihood, then the new tree is accepted (Yang and Rannala, 2012). Importantly, this process allows for downhill traversals if the penalty in likelihood is small enough, accounting for the total tree probability space being rough in nature (many small local optima; Larget and Simon, 1999). Additionally the algorithm is "forgetful" where it only remembers the last step it took (Nascimento *et al.*, 2017). It is therefore possible to imagine

that the most optimal tree topologies will be sampled rapidly, after a rapid traversal through probability space (figure 10).

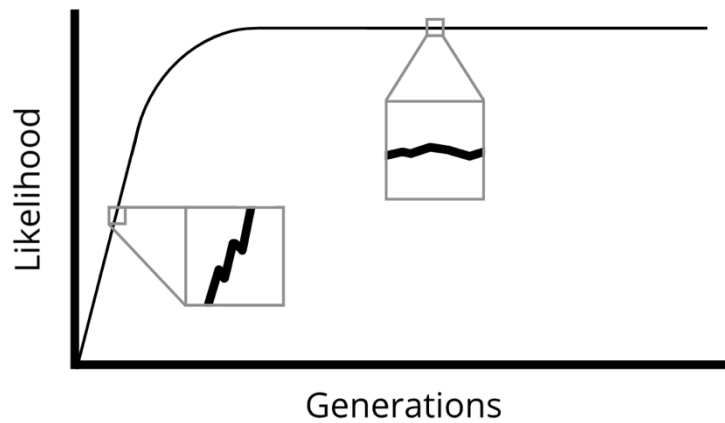


Figure 10 - A single MCMC simulation approaching convergence in probability space

A mock-up of traversals of probability space, where a low likelihood is rapidly traversed to a high likelihood. While detrimental steps are permitted in the MCMC analysis, they are comparatively small compared to possible positive traversals when approaching the improvement plateau (convergence). At convergence, likelihood values will fluctuate around the optima.

Once the optimum likelihood in tree space is reached (convergence), fluctuations around the optimum are maintained at a plateau (Nascimento *et al.*, 2017). If the chain is able to efficiently sample the posterior distribution at the plateau (i.e. the local space is not noisy) then multiple simultaneous simulations optimising the same tree space should show low variance between runs. Multiple simultaneous simulations also allow for more of the tree space environment to be sampled per MCMC run (Altekar *et al.*, 2004). Running and comparing multiple chains simultaneously is termed a Metropolis Coupled MCMC (MCMCMC; Metropolis *et al.*, 1953; Altekar *et al.*, 2004). “Heating” can also be applied to one or more of the chains (Marshall, 2010). A heated chain is allowed to step downhill further, effectively flattening the probability space. Therefore, peaks and troughs in stability space can be traversed more easily, avoiding local optima trapping (Nascimento *et al.*, 2017). As a rule of thumb, the lower the variance, the more accurate the simulation (Marshall, 2010). The simulation is typically run for hundreds of thousands to millions of

generations to give the best opportunity to reach convergence (Nascimento *et al.*, 2017). Once convergence is reached, an optimal tree topology can be generated from the summary of all tree topologies sampled. As it is desirable to only include the most optimal trees in the analysis it is typical to apply burn-in, where the initial steps in the chain are removed from prior analyses (Marshall, 2010). Posterior probabilities are also computed from the summary of probabilities sampled by the Markov Chain (Nascimento *et al.*, 2017; Yang and Rannala, 2012).

1.3.1.3 - ASR algorithm principals

Computational methods for ancestral reconstruction were first implemented on DNA sequences by Yang *et al.* (1995), where an empirical Bayesian approach was applied to maximise the likelihood of $p(\text{ancestors}|\text{contemporaries})$ (Pupko *et al.*, 2000). Here, the most likely sequence at a node on a phylogeny is computed given the set of sequences that radiate from that node. Confusingly, this approach can therefore be described as both ML and Bayesian. For clarity ML will be used in this thesis. Importantly, Yang *et al.* (1995) distinguished two flavours of ancestral reconstruction (marginal and joint) that can be computed in this framework, based on algorithmic approach to inferring ancestral sequences. Joint reconstruction attempts to find the set of character states that maximise the likelihood over the tree simultaneously. Marginal reconstruction attempts to maximise the state at the given node that maximises the likelihood given all other states at all other nodes (Pupko *et al.*, 2000).

Marginal reconstructions and joint reconstructions are not equivalent (Ashkenazy *et al.*, 2012). Consider the tree in figure 11, that displays the probability of character A (a base or amino acid) being derived from characters B or C, and so on to contemporary sequences E through I. The marginal likelihood algorithm to compute the state of A asks what the most likely immediate step along the branches of the tree is to have created A. In this example this would be $B \rightarrow A$, with a likelihood of 0.55 (blue arrow). Computing the state of A with joint likelihood probes the most likely set of steps to have created A across the entire tree. With a joint likelihood of 0.405, $I \rightarrow C \rightarrow A$ (red arrow) is the most likely global solution in figure 11, and therefore the maximum likelihood state of A is different depending on the

optimisation heuristics (Pupko *et al.*, 2000). The use of marginal versus joint reconstruction depends on the questions being asked in the experiment. If it is desirable to know the maximum likelihood sequence at a specific position based on its immediate surroundings, then marginal likelihood is most appropriate. However, if the aim is to describe the best set of all hypothetical taxa, then joint reconstruction is more appropriate (Joy *et al.*, 2016). For the majority of reconstructions, marginal reconstruction is utilised, as computing joint reconstructions is computationally intensive for large datasets. Joint reconstruction is therefore restricted to optimisation across the tree for the single maximum likelihood sequence at each node. On the other hand, marginal reconstruction deals with a sufficiently small sequence space to compute the likelihood of all possible states simultaneously, and algorithms will typically output a matrix of posterior possibilities for each state (Yang *et al.*, 1995; Pupko *et al.*, 2000; Merkl and Sterner, 2015). For this reason, marginal reconstruction is considered an appropriate estimate of joint reconstruction, and is a widely accepted method for reconstructing ancestral sequences (Merkl and Sterner, 2015; Joy *et al.*, 2016).

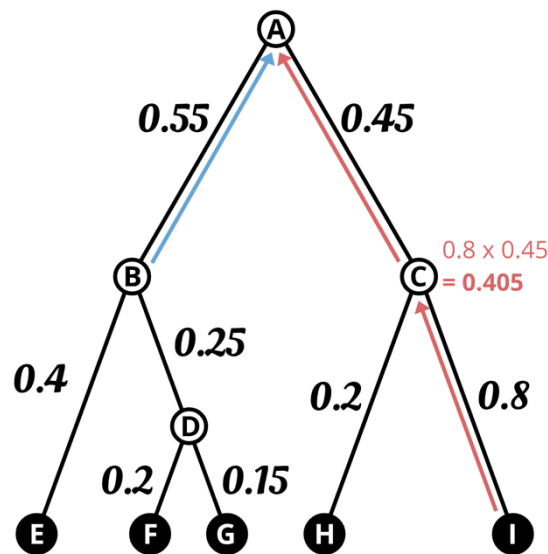


Figure 11 - Marginal and joint likelihood ancestral reconstructions are not equivocal

Hypothetical phylogeny showing the probabilities that state A is derived from states B through I.

Marginal likelihood will maximise the likelihood of the state given the immediate next states (blue arrows). Joint likelihood will maximise the likelihood of the state given the global set of states (red arrows). Therefore it is possible for joint and marginal likelihood to reconstruct different ancestors.

1.3.1.4 - Algorithm choice in ASR

For amino acid sequences, Bayes's theorem is implemented in a similar fashion to phylogenetic reconstruction when optimising $p(\text{ancestors}|\text{contemporaries})$ (Pupko et al., 2000). However, the number of states being optimised is significantly smaller in the ancestral reconstruction problem. Both node order and branch length are optimised simultaneously in MCMC simulations inferring tree topology. On the other hand, only one of a maximum of 20 possible states needs to be optimised at each position in the posterior probability of an ancestral reconstruction. Additionally, ancestral sequences are only optimised over the states available in the alignment, which rarely represent all 20 possible amino acids (Merckl and Sterner, 2015). A simple empirical Bayes algorithm is therefore sufficient to compute ancestral states until datasets get considerably large or complex, at which point a MCMC algorithm to infer the joint posterior distribution can be employed (i.e. >1,000 sequences; Joy et al., 2016; Gumulya et al., 2018). Importantly, to generate a matrix of posterior probabilities, Bayes's theorem requires a prior distribution what is known about the sites before analysis. As with phylogenetic reconstruction, priors are typically a multiple sequence alignment, a model of amino acid substitution rates, as well as a distribution describing rate heterogeneity amongst sites (Merckl and Sterner, 2015; Pupko et al., 2000).

Empirical Bayes assumes that the priors are absolutely accurate. In reality, ASR is an estimation over a set of estimations, and therefore truly accurate representations of ancestral history are unlikely due to carried error. Drawn conclusions from ASR experiments should therefore be cognisant of carried error that may impact the robustness of the experiment (Williams et al., 2006; Westesson et al., 2012). Eick et al. (2017) calculate that for a sequence with 90% of sites unambiguously predicted, there is a 1.2% chance that the rest of the ambiguous sites are accurately computed if all other sites have two possible states. Typically conclusions derived from reconstructed sequences therefore require comparison with real world calibration data to be considered confident (Gaucher et al., 2008; Eick et al., 2017).

This thesis will only discuss empirical algorithms. Three commonly used empirical algorithms have been developed that build upon the original work on maximum likelihood ratification of states by Pupko et al. (2000) and Koshi and Goldstein (1996), summarised in table 7. As

the core algorithms are largely similar, algorithm choice comes down to functionality. The largest difference between the algorithms is the method of gap placement. Gap placement has to be handled as an independent problem, as amino acid substitution matrices do not include likelihood values for indel events (Holmes, 2017). There is a paucity of methods to derive indel rates in contemporary phylogenetic tools, and none of the empirical ASR algorithms apply exact calculations of indel rates (Westesson *et al.*, 2012).

PAML does not provide any tools to handle gaps (Yang *et al.*, 2007), and gaps in the alignment are ignored by the algorithm. This is an incorrect assumption if not accounted for. Consider an alignment of sequences. One of the sequences has a single insertion that occurred very recently in its evolutionary history. This forces a gap to be placed in every other sequence in the alignment at this position. In PAML, all other gaps are ignored as they are considered as missing data, so only the insertion is considered by the algorithm. Any ancestor of this sequence will therefore contain a gap, leading to overestimation of ancestral sequence length (Yang, 2005; Yang *et al.*, 2007).

Ancescon handles gap placement by parsimony using supersedence-based placement, where gaps are considered before amino acids when encountered at a site. If a site has a gap, the branch containing that gap is pruned from the tree. If a node then has no or one child, then a gap is placed in the ancestor (Cai *et al.*, 2004). Parsimony based gap placement always assumes a gap to be more likely than an amino acid at a position, and also fails to account for multiple insertion events at a given position. It can therefore be expected to significantly over or underestimate gap placement depending on the alignment (Ashkenazy *et al.*, 2012).

FastML applies maximum likelihood to estimate gap placement, where gaps are considered as an additional parameter in the marginal likelihood reconstruction (Ashkenazy *et al.*, 2012). Each site is considered a binary state, where 0 denotes a site containing an amino acid, and 1 denotes a site containing a gap. Binary maximum likelihood is then applied to the dataset to determine the probability of any ancestral site containing a gap given the binary states in its immediate daughter nodes. Indel rates are derived based on a framework that estimates rates based on a variant of parsimony that integrates

stochasticity into its derivation criteria. However, this method is still contrived as it is not based upon true rate data from real datasets.

Algorithm	Reconstruction methods	Gap reconstruction	Rate heterogeneity	Type	Reference
Ancescon	Marginal and Joint	Supersedence with branch trimming	Alignment guided or maximum likelihood rate factors	Online server	Cai <i>et al.</i> , 2004
PAML (CodeML)	Marginal and Joint	No	Customisable or estimated gamma, uniform.	Command line	Yang <i>et al.</i> , 2007
FastML	Marginal and Joint	Binary maximum likelihood	Estimated gamma, uniform	Online server	Ashkenazy <i>et al.</i> , 2012

Table 7 - Empirical ancestral reconstruction algorithms

1.3.2 Probing patterns in protein evolution with ASR - thermostability

Ancestral sequence reconstruction allows researchers to walk back along the branches of a phylogeny, and generate proteins that represent the most likely sequence to have generated all descendent sequences in the provided dataset. Given access to synthetic DNA tools, it is possible to resurrect such proteins in the laboratory and infer conclusions about the evolution of traits in a protein family’s history based on the ancestral protein’s properties. As such, ASR has been utilised to probe a number of philosophical questions surrounding the early evolution of enzymes to better understand the diverse functionalities exhibited by these molecules today. By conducting ASR on promiscuous enzyme families, the adaptive paths defining substrate discrimination, specificity and plasticity have been analysed (Voordreckers *et al.*, 2012; Wheeler *et al.*, 2018; Pawlowski *et al.*, 2018; Babkova *et al.*, 2017). Additionally, contemporary proteins represent a single adaptive path through sequence space based on a given set of pressures. ASR has therefore been utilised to study the functional space represented by alternative adaptive paths, giving a small window into what protein evolution “could have been” (Starr *et al.*, 2017; Cole *et al.*, 2013).

Furthermore, at some or many points in the evolution of enzymes, dynamic interactions between multiple separately evolving protein surfaces began coevolving to generate functional homo and heteromultimeric proteins. ASR has therefore also been utilised to probe whether innovation of complexity is a stepwise or one-time event in the evolutionary history of a protein family (Lim and Marqusee, 2017, Prinston *et al.*, 2017; Hochberg and

Thornton, 2017; Finnigan *et al.*, 2011; Zinn *et al.*, 2015). Finally, there is considerable debate over the nature of very early life (>3.5 Gy ago), including the conditions required for life to evolve. Inferences about the temperature of ancient life have been made from a large body of research studying the stability of ancient proteins (Gaucher *et al.*, 2003; Gaucher *et al.*, 2008; Akamuna *et al.*, 2013; Butzin *et al.*, 2013; Hart *et al.*, 2014; Okafor *et al.*, 2018; Garcia *et al.*, 2017). The wider ancestral reconstruction field has been summarised in the combined works of Harms and Thornton (2010), Merkl and Sterner (2015), Joy *et al.* (2016) and Hochberg and Thornton (2017).

A fundamental question asked of protein families by ASR is whether their ancestral traits fundamentally differ from those of their contemporary counterparts (Gaucher *et al.*, 2008). An important target for study is protein stability. If ancestral protein stabilities are fundamentally different to those of today's proteins, it indicates that the ancestor evolved in an environment that is also fundamentally different to the modern environment as proteins typically evolve at marginality. The first study of stability in ancestral sequence was conducted in the seminal works by Jermann *et al.* (1995), where the ancestors of artiodactyl RNases were found to be more stable in the ancestors of true ruminants.

Miyazaki *et al.* (2001) were the first to consider the environment inhabited by the Last Universal Common Ancestor (LUCA) of life. At the time, there was debate around the LUCA's environment. One school of thought believed the ancient life evolved in temperatures similar to those that exist today, based on the instability of many biological compounds (Miller and Lazcano, 1995), including the DNA double helix (Galtier *et al.*, 1999). On the other hand, the deepest branches derived in ribosome-based species trees were represented by both thermophilic archaea and bacteria, leading others to suggest that ancient life evolved at high temperatures (Woese, 1987). This hypothesis is supported by evidence from meteor impacts 4.5-3.8 Gy ago that heated the ocean to >100 °C, allowing only the most thermophilic of organisms to survive around the conception of life (approximately 4 Gy ago; Nisbet and Sleep, 2001). Parsimony was used to derive the ancestors of the ubiquitous metabolic enzymes 3-isopropylmalate dehydrogenase and isocitrate dehydrogenase. Three highly conserved regions between the enzymes were reconstructed, and imparted into the natural 3-isopropylmalate dehydrogenase from

thermophile *Sulfolobus* “strain 7” by site directed mutagenesis. It was rationalised that stabilising residues are considerably rare in thermophiles, therefore any stabilisation by ancestral residues is significant. An increase in stability of approximately 4 °C was observed in the ancestor over the wild type, and half-life at 99 °C was doubled.

It was stated by Miyazaki *et al.* (2001) that their research supported the “hot-start” hypothesis of ancient life. However, stability is a trait defined by the whole protein structure. Therefore, despite the algorithm identifying a set of stabilised ancestors, the work does not conclusively illustrate the net-stabilisation across all ancestral residues. To alleviate this issue, Gaucher *et al.* (2003) applied the modern ASR algorithm (PAML) to the question of ancient stability by reconstructing the 3.5 Gyo LUCA of the ubiquitous elongation factor-thermostable (EF-Tu) protein family from modern mesophilic lineages. EF-Tu enzymes were synthesised in full. Their stabilities were approximately 12 °C higher than their modern mesophilic counterparts (43 → 55 °C), suggesting that the ancestor of modern mesophiles was a thermophile.

Further studies have consistently corroborated this finding, demonstrating equivalent trends in environmental cooling. Gaucher *et al.* (2008) again focused on the EF-Tu enzymes. This work reconstructed multiple nodes that existed between 3.5-0.5 Gyo to track stability when stepping back through epochs. Ancestors were based on highly detailed phylogenies derived from species tree data. Ancestral stabilities increased stepwise toward the LUCA, which denatured at 73.3 °C (a 30 °C increase over *E. coli*). Importantly, Gaucher *et al.* (2008) also showed significant correlation between the stabilities observed and the temperature predictions of the ancient earth based on ratios of $\delta^{18}\text{O}$ and $\delta^{30}\text{Si}$ isotopes in ocean cherts (Robert and Chaussidon, 2006). Peres-Jimenez *et al.* (2011), and Akanuma *et al.* (2013) independently showed equivalent trends in the thioredoxins and the nucleoside diphosphate kinases respectively. In both instances, enzyme stabilities stepped down toward mesophily from stabilities ≈ 30 °C higher than extant enzymes. From these data, it was inferred that environmental temperatures cooled at a rate of approximately 6 °C per billion years. Importantly, ancient thioredoxins were also considerably more stable at acidic pH, correlating with predictions for the acidity of ancient oceans (Perez-Jimenez *et al.*, 2011). Equivalent trends derived from nucleoside diphosphate kinase (NDK) ancestors were

further corroborated by Garcia *et al.* (2017), who reconstructed NDKs from eukaryotic and prokaryotic phototrophs to incorporate analyses over a range of possible ancestral environments beyond the ocean.

1.3.3 ASR as an engineering tool

ASR consistently produces proteins that display outstanding increases to stability, with various examples showing ≥ 30 °C improvements over extant counterparts (Gaucher *et al.*, 2008; Risso *et al.*, 2013; Perez-Jiminez *et al.*, 2011; Akanuma *et al.*, 2013). ASR could therefore offer a powerful yet simple tool for the engineering of protein stability. For example, reconstructions of Precambrian β -lactamases by Risso *et al.* (2013) observed an ancestor with a 35 °C increase in stability over its modern descendants (TEM-1). On the other hand, studies attempting to engineer TEM-1 by directed evolution have achieved an optimization plateau at 20 °C improvement (Hecky and Müller, 2005).

Chen *et al.* (2010) developed the first ASR driven protein engineering tool: Reconstructed Evolutionary Adaptive Paths (REAP). REAP is a semi-rational tool for the engineering of novel functional properties into proteins. The method first identifies sites in the protein undergoing functional divergence based on data in the phylogeny. ASR is then utilized to identify possible variants from the posterior probability of such divergent sites. Small (<100 variant) libraries are generated consisting multiple ancestral variant amino acids sampled from the posterior distribution (Lutz *et al.*, 2010). Such libraries offer a high chance of generating viable proteins as such poorly resolved residues often exist at highly variable sites in the alignment. Therefore, REAP libraries typically sample sites most resistant to deleterious mutation increasing the likelihood of obtaining functional proteins (Cole and Gaucher, 2011; Cole *et al.*, 2013). Chen *et al.*, 2010 utilized REAP to construct a library of 93 Taq polymerase variants containing up to 4 mutations, with the aim of engineering polymerase activity on the reversible terminator dNTP-ONH₂ for use in DNA sequencing (Chen *et al.*, 2013). Of this library, 30 variants possessed activity on the desired target, with two variants possessing “exceptional” activity that was suitable for incorporation into DNA sequencing protocols (Chen *et al.*, 2010). Many other studies have used evolutionary histories in this manner to isolate sites of interest to engineer enzymes with novel

functionality (Alcolombri *et al.*, 2011; Conti *et al.*, 2014; Gonzalez *et al.*, 2014; Miyazaki *et al.*, 2001; Watanabe and Yamagashi, 2006).

In a somewhat related fashion, ASR was used by Zakas *et al.* (2015) and Zakas *et al.* (2017) to compliment orthologue scanning of the protein drug coagulation factor VIII used in the treatment of haemophilia (Kempton and White, 2009). This study presented the ancestors as the direct engineering product, discovering coagulation factor ancestors that displayed improved stability, biosynthetic efficiency, activity and expression compared to other human factor VIII biologics that are currently available. Blanchet *et al.* (2017) used the same principals to engineer a small set of Mamba venom based protein biologics that displayed improved adrenoceptor selectivity, and identified three residues in the protein that modulate protein affinity from the libraries of variants generated from the posterior probabilities of the reconstruction experiment.

A review by Wijma *et al.* (2013) suggested that ASR could be utilized as a tool for the direct engineering of proteins for stability if the family possesses an ancestor that is considerably ancient (it evolved in a high temperature environment). Recent studies have explored this application of ASR for bioindustrially important proteins. Whitfield *et al.* (2015) reconstructed the ancestors of periplasmic amino acid binding proteins in the search for a robust L-arginine binding protein with high stability and substrate selectivity for use in Förster Resonance Energy Transfer (FRET) analyses. FRET experiments require robust sensor proteins that maintain their structure throughout the experiment course to avoid noise or false positive signals. Thermostable proteins are therefore sought after for such applications (Clifton *et al.*, 2017). Natural amino acid binding proteins, including those, identified from thermophiles, show poor specificity for L-arginine, with additional binding activity on structurally similar amino acids histidine, ornithine and lysine. Whitfield *et al.* (2015) reconstructed two ancient amino acid binding proteins, one representing the ancestor of glutamine binding proteins and one representing the ancestor of non-specific arginine binding proteins. The ancestor of the family of glutamine binding proteins was an extremely thermostable L-arginine specific biosensor, showing a 29 °C increase in stability over its descendants. This protein was subsequently utilized for *in situ* analysis of L-arginine concentrations in acute rat hippocampus slices.

Babkova *et al.*, 2017 were the first to engineer enzymes with synthetic biology relevance, targeting the haloalkane dehalogenases (HLDs). HLDs convert halogenated alkanes into a primary alcohol and a halide. A number of by-products from the chemical industry are halogenated, meaning HLDs have considerable importance in chemical waste treatment, and bioremediation (Dvorak *et al.*, 2014). Additionally, HLDs are important synthetic pathway components, able to produce optically pure primary alcohols (Koudelakova *et al.*, 2013; Prokop *et al.*, 2009). Babkova *et al.* (2017) reconstructed five HLD ancestors, representing ancestral sequences of some of the best studied HLDs. Ancestors displayed between an 8 and 24 °C increase in stability over their extant counterparts, and two of the ancestors represented the most stable HLDs reconstructed to date.

Gumulya *et al.*, 2018 made an important leap in the utilization of ASR as an engineering tool. Previous studies focused on the most ancient of sequences from early life to ensure stabilization (Gaucher *et al.*, 2008). Instead, Gumulya *et al.* (2018) tested how broadly applicable ASR engineering is to diverse protein families by reconstructing ancestors of the CYP3 family of cytochrome P450 monooxygenases. Importantly, the family was innovated by the first vertebrates. There is no evidence that proto-vertebrates existed in conditions much different to those observed today. Yet, CYP3 ancestors exhibited a 30 °C increase in thermostability over modern sequences. Such a result is particularly significant as directed evolution of the same protein family failed to produce exceptional increases in stability, with optimization plateauing at improvements of 4 °C (Li *et al.*, 2007). It is hypothesised that ancestral CYP3 stability derives from proteins existing in ancient organisms that inhabited warmer oceanic conditions, and stability was a trait carried over into the CYP3 ancestor (Gumulya *et al.*, 2018).

Importantly, work by Gumulya *et al.*, (2018) shows that ASR may be a broadly functional tool for the engineering of thermostable proteins. This work hints that relatively modern proteins that only encountered mild environmental temperatures for the majority of their history can still harbour exceptionally stabilized protein variants (Risso *et al.*, 2018). Additionally, like stabilization from consensus sequences, ASR is able to provide considerable increases in stability based on simple input requirements, in the absence of a

crystal structure (Wijma *et al.*, 2013). However, consensus engineering is contrived, with no guaranteed method for optimal engineering of stable proteins. Ancestral sequence reconstruction relies solely on the information held within the extant sequences, and uses Bayesian inference to define the maximum likelihood ancestor given an input dataset, providing a quantitative ratification of each amino acid site. ASR therefore provides a clearly defined methodology for the generation of stable sequences.

1.3.4 How does ASR generate stable sequences?

While ASR is seeing considerable early success as an engineering tool, it is not fully understood how ASR confers improvements to protein stability. Strong evidence exists for the stabilization of ancestral proteins that existed in organisms inhabiting oceans heated by meteor impacts (Gaucher *et al.*, 2008). However, it can be argued that the rationalization for stable ancestral CYP3 cytochrome P450 monooxygenases by Gumulya *et al.* (2018) is not parsimonious given contemporary understanding of marginal stability in protein evolution. For their hypothesis to be true, some ancestral thermostable state had to be maintained for billions of years before horizontally transferring into the proto-vertebrates. At the time of writing, evidence suggests that CYP3 never made such a jump, and instead evolved with other vertebrate CYP families from a 'genesis locus' in basal vertebrates (Nelson *et al.*, 2013). Therefore we find current hypotheses on the evolution of stability in CYP3 spurious.

Trudeau *et al.* (2016) provide an alternative hypothesis for the origin stability in ancestral proteins derived from mesophilic histories. Previously, reconstructed mammalian paraoxonases, used to generate ancestor libraries for tests of robustness in inferred ancestral space, were found to be highly thermostable (Bar-Rogovsky *et al.*, 2015). In line with other studies, a 30 °C increase in stability was observed, yet there is no evidence that the mammalian LUCA evolved in high temperature environments. This is an important finding as it provides further evidence of ASR being a broad spectrum engineering tool. Trudeau *et al.* (2016) hypothesised that stabilization was the result of a "consensus effect".

A consensus sequence is the average residue at any position, and consensus residues are net stabilizing when taken across the entire protein (Sternke *et al.*, 2018; Porebski and

Buckle, 2016). Mammalian ancestor paraoxonases showed 80% similarity to the consensus of the same alignment, whereas extant proteins were considerably more divergent (50-70%). The consensus effect is also shown to exist across ancestral proteins in the literature (a detailed review is provided by Trudeau *et al.*, 2016). By this merit, it is possible that the propensity of ancestral reconstruction to select for consensus residues biases an ancestral protein towards a more stable state. To better understand the consensus effect's impact on ancestral protein reconstruction, Okafor *et al.* (2018) compared ancestral EF-Tu with consensus EF-Tu. Despite differences in sequence, ancestor, extant and consensus proteins shared a conserved structure. Consensus sequences were considerably less stable than thermostable extant and ancestral sequences, suggesting that the consensus effect is not pervasive for all families. Furthermore, as discussed the leading hypothesis for the stabilization of consensus sequences relates to the introduction of residues derived from a stable ancient sequence (Porebski and Buckle, 2016). It can therefore be argued that the stable mammalian paraoxonase ancestor instead devalues the leading hypothesis for the stabilization of consensus proteins, and the introduction of consensus residues in ancestral proteins is simply an expectable trait of the methodology.

Despite ASR potentially being a powerful tool for the generation of empirically derived stable proteins, the generation of reliable inputs for the methodology still requires some level of expertise, as the probability of obtaining a functional protein is directly related to the quality of both the alignment and the phylogeny (Vialle *et al.*, 2018; Joy *et al.*, 2016). It is also not yet understood how algorithm choice effects the stabilization of ancestral proteins. Additionally, the lack of understanding of the underlying forces driving stabilization leads to each reconstruction being a "shot in the dark", requiring resource commitment without confidence that the experiment will produce desirable results. As with consensus sequence driven stabilization, the fidelity of engineering by reconstruction is difficult to gauge as failed experiments are unlikely to be published (Sternke *et al.*, 2018). Therefore, while ASR is a promising enabling tool for the scale-up in synthetic biology, more exploratory research is required before it can be considered widely accessible.

1.4 Aims and Objectives

As discussed in section 1.2, open and accessible tools for the engineering of thermostable proteins are highly desirable, as stable proteins can relax bottlenecks that exist in the scale-up of synthetic biology. As discussed in Section 1.3, ASR presents a poorly understood but promising engineering tool that consistently produces protein variants with outstanding stabilization from simple inputs. This thesis will investigate whether ASR can meet the requirements of a democratised protein engineering technology. By performing ASR experiments on both physical and *in silico* datasets, this thesis investigates the tool's simplicity of utilization, expertise requirements, input requirements, and broader potential applicability. Additionally, this thesis aims to gain a deeper understanding of the mechanisms underlying ASR's stabilizing properties. An argument that both ASR and consensus protein design stabilize proteins through equivalent, largely ubiquitous evolutionary forces will be generated from contemporary literature. This hypothesis will then be thoroughly investigated through the design and utilization of *In silico* protein evolution modelling tools. A further objective of this thesis is to use newfound evidence of ASR's stabilization mechanism to design an engineering centric protein reconstruction technology for highly accessible protein stability engineering. These aims and objectives will be explored over three chapters:

1.4.1 Chapter 2

Chapter 2 presents a prepared manuscript undergoing peer review in Nature Communications at the time of writing. In this study, we compliment work by Gumulya *et al.*, 2018 by further testing the suitability of ASR for the engineering of stability into protein families predicted to have evolved in mild environmental conditions throughout their evolutionary history. We target the bioindustrially important carboxylic acid reductase (CAR) family of enzymes as they have not been identified in thermophilic parent organisms, and present a significant engineering challenge (see section 1.5 – Addendum). As discussed in section 1.3, there are multiple empirical ASR algorithms available. However, it is not known whether each algorithm is a suitable engineering tool. Here, we reconstruct ancestral CARs with the three main empirical reconstruction algorithms discussed in section 1.3.1. We also aim to devise a simple method for the placement of gaps in PAML derived ancestors by transposition from other algorithms. We derive three highly thermostable CAR enzymes,

and report a number of industrially relevant enzyme traits, providing a full characterization dataset for future use as part of the CAR toolbox.

1.4.2 Chapter 3

Chapter 3 presents a manuscript prepared for submission to eLife. In chapter 2, ASR was shown to generate thermostable proteins from protein families that are unlikely to have ever inhabited high temperature environments in their history. As discussed in section 1.2.6.1, and section 1.3.4, the current hypotheses describing the underlying mechanisms of stabilization in alignment based stability engineering tools are incomplete. Here we aim to define, and provide evidence supporting an alternative hypothesis for the stabilization of alignment based engineering tools, termed the “survivor bias hypothesis”. As discussed in section 1.2.2, most proteins are marginally stable. We hypothesise that this property leads to stabilizing biases in the ancestral dataset due to the titration of significantly destabilizing residues from the alignment dataset. Using pure Python, we aimed to develop an open source tool for the modelling of stability changes in hypothetical protein populations that evolve according to the accepted evolutionary processes outlined in section 1.2.

Importantly, this tool allows for the tracking of stability changes across evolutionary time at the residue, protein and population level. We aimed to reconstruct ancestors of simulated protein populations that had evolved at marginality. From these populations we show evidence of significant stabilization bias in both ancestral sequences, and consensus sequences. More importantly, this chapter aimed to provide the first robust, and first unifying theory for the forces driving stabilization across all alignment based engineering tools.

1.4.3 Chapter 4

Chapter 4 presents a prepared manuscript that will be submitted to eLife as a sister study to the manuscript presented in chapter 3. Chapter 3 provided a toolbox with which to test alignment based engineering tools. As discussed in section 1.1, synthetic biology is enabled by broadly accessible tools. Additionally, as discussed in section 1.2.6, while stable proteins may allow for the relaxation in scale-up bottlenecks that stymie the progression of synthetic biology applications to market, there are no broadly accessible contemporary stability engineering tools. Section 1.3 discussed that ASR may be a broadly accessible engineering

tool. While chapter 2 showed that ASR could engineer highly thermostable proteins from challenging targets, the method still requires some expertise to perform, as phylogenetics is a steep learning curve. In this chapter we aim to design the first truly expertise-agnostic, broadly accessible stability engineering tool called “simplified ASR”, based on the ASR algorithm Ancescon. Ancescon can be run in a phylogeny-free mode, where the only input requirement is a multiple sequence alignment. We aimed to use the evolutionary model designed in chapter 3 to simulate sASR experiments to enable the design of a set of criteria that maximise success with the tool. We then aimed to ratify sASR by reconstructing the CARs. sASR was used to generate the most stable CAR biocatalyst reported to date, which were then characterised kinetically.

1.5 – Addendum

Carboxylic acid reductases

For the majority of this thesis, we focus on the carboxylic acid reductase family of enzymes (CARs; E.C. 1.2.1.30). CARs are important enzymes for bioindustry, catalysing the reaction:



via a phosphopantethiene intermediate that is covalently bound to a phosphopantethiene binding site within the core structure of the enzyme (Winkler, 2018). A full catalytic CAR mechanism is presented in chapter 2. CARs have important implications for industrial biocatalysis, replacing a complex chemical synthesis protocol with a simple enzymatic catalysis (Akhtar *et al.*, 2013). Traditional carbonyl reduction is performed with environmentally damaging metal hydride reducing agents. The reduction of carboxylic acids is energetically unfavourable, and strong reducing agents are required to reduce the carboxylic acid. However such reducing agents will often immediately reduce the aldehyde product into the corresponding primary alcohol. An additional oxidation step is therefore required to convert the alcohol back into an aldehyde (Gaylord, 1957). Replacing this reaction with an enzymatic process is therefore desirable.

CARs are promiscuous enzymes with known activity on over 100 aliphatic fatty acid and aromatic carboxylic acid substrates (Winkler, 2018). As a result, CARs have been utilized in a number of commercial biosynthesis pathways (e.g. Akhtar *et al.*, 2013; Kallio *et al.*, 2014; France *et al.*, 2016). However, the main CAR families evolved in actinomycetes around 500 mya, and to date no CAR has been identified from a thermophile (Lewin *et al.*, 2016; Winkler, 2018). Additionally, CARs are large (130 KDa), highly dynamic proteins with only partially resolved crystal structures (Gahloth *et al.*, 2017). CARs therefore present a complex engineering challenge, and serve as an excellent stress-test for ASR engineering protocols.

As a prelude to the research presented in this thesis, a thorough phylogenetic analysis of CARs was performed by the author and published in a study by Finnigan *et al.* (2017). The complete manuscript of this study is presented in the appendices (Chapter 7.3). In this work, CARs were identified as members of the ANL protein superfamily, allowing for key catalytic residues to be identified based on CAR alignments with sister families of acyl-CoA synthetases and non-ribosomal peptide synthases. An exhaustive (at the time of writing) dataset of 124 actinomycete CARs was also obtained from homology searches. All extant CAR sequences utilized in this thesis are derived from this database.

Finnigan *et al.* (2017) also lay the groundwork for a CAR activity assay used extensively in this thesis, that measures changes in NADPH absorbance at 340 nm. This work thoroughly characterized five carboxylic acid reductases in terms of substrate kinetics and stability. A CAR from *Mycobacterium phlei* (MpCAR) was reported as the most stable CAR identified to date, which begins losing activity after incubation for 30 minutes at 42 °C, and loses 50% activity after incubation at 48 °C. The CAR from *Nocardia iowensis* (NiCAR), the most widely studied CAR in the literature, began losing activity at 38 °C, and lost 50% activity following incubation at 42 °C (figure 12). Since, a CAR from *Mycobacterium avium* has been identified that begins losing activity following incubation at 48 °C, and retains 50% activity at 49 °C (Kramer *et al.*, 2018).

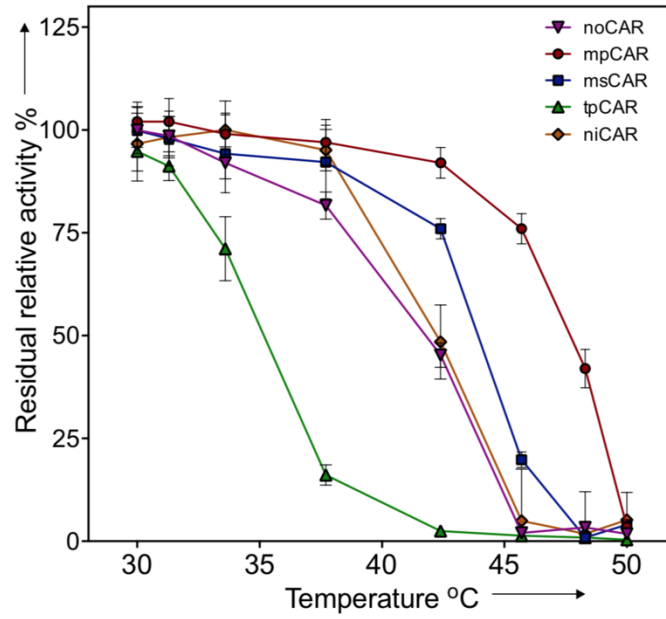


Figure 12 - The stability of extant CAR enzymes analysed in Finnigan *et al.* (2017)

The residual activity of CARs was assessed on 4-methylbenzoic acid following incubation at increasing temperatures. Points and error bars represent the average and standard deviation of three independent readings. Complete assay details are presented in chapter 2.

Chapter 2

Thermostable carboxylic acid reductases generated by ancestral sequence reconstruction

2.1 Authors

Adam Thomas^{1,2}, Rhys Cutlan^{1,2}, William Finnigan², Mark van der Giezen², Nicholas Harmer^{1,2}

1. Living Systems Institute, Stocker Road, Exeter EX4 4QD, U.K.
2. Department of Biosciences, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, U.K.

2.2 Preface

This chapter consists a reformatted manuscript for an article submitted for review to Nature Communications. Comments from peer review were received in early December. However we are unable to include the requested changes into the presented manuscript based on time constraints.

AncCARs developed in this study spawned the main question explored in Chapter 3 – why do ancestral proteins exhibit high stability even when the history of their family is unlikely to have encountered high temperature conditions?

AT, NH and MvdG conceived the study. AT wrote the article. All authors edited the article. AT conducted sequence handling, phylogenetics, ASR, protein purification and assays, and was involved in critical discussion throughout. RC conducted protein purification and assays, performed protein structure modelling, contributed to the writing of the article, and was involved in critical discussion throughout. WF provided substrates, helped develop the reduction assays, and developed purification protocols for CARs.

2.3 Abstract

Carboxylic Acid Reductases (CARs) are biocatalysts of industrial importance. However, the properties of this protein family, especially their poor stability, render them sub-optimal for use in a bioindustrial pipeline. Due to their size, dynamic structure and reaction complexity, traditional protein engineering methods for improving thermostability are not viable in the CAR enzymes. Here, we employed ancestral sequence reconstruction (ASR) – a burgeoning engineering tool that can identify stabilizing but enzymatically neutral mutations throughout a protein. We used a three-algorithm approach to reconstruct functional ancestors of the *Mycobacterium* and *Nocardia* CAR1 orthologues, representing the largest reconstructed proteins to date. Ancestral CARs (AncCARs) were confirmed to be CAR enzymes with a preference for aromatic carboxylic acids. Ancestors also showed varied tolerances to solvents, pH and *in vivo*-like salt concentrations. Compared to well-studied extant CARs, AncCARs had a T_m up to 34 °C higher, with half-lives up to nine times longer than the greatest previously observed. Using ancestral reconstruction we have expanded the existing CAR toolbox with three new thermostable CAR enzymes, providing access to the high temperature biosynthesis of aldehydes to drive new applications in biocatalysis.

2.4 Introduction

Many industries are placing increasing emphasis on achieving carbon neutral manufacturing. For the chemical industry, the sustainable catalysis of high-value chemicals through enzyme cascades (“green chemistry”) is a key opportunity (Nielsen and Moon, 2013; Kelley *et al.*, 2014). Enzymes generally provide high yields with few side products and do so at mild reaction conditions. Enzymes therefore mitigate the production of excessive chemical waste and the use of toxic catalysts, while also reducing energy and solvent usage (Sheldon, 2016). Nevertheless, enzymes are still poorly represented in the chemical synthesis market (Wallace and Balskus, 2014). Enzymes are generally highly evolved towards their biological role *in vivo*, and rarely have properties optimized for a green chemistry application. Limited enzyme stability, restricted substrate ranges, substrate flux sinks, and low turnover rates are common barriers to success (Finnigan *et al.*, 2012; Packer and Liu, 2015; Ye *et al.*, 2017; Ebert and Pelletier, 2017; Kaushik *et al.*, 2016).

Carboxylic Acid Reductases (CARs; E.C. 1.2.1.30) are a family of enzymes with increasing relevance to green chemistry. They catalyse the reduction of an aliphatic or aromatic acid to the respective aldehyde, using ATP and NADPH as cofactors (Finnigan *et al.*, 2017; Winkler, 2018). This reaction is otherwise challenging to achieve chemically or biochemically. Consequently, CARs are being used in biotechnology for the enantiopure biosynthesis of intermediates in enzyme cascades. Examples of these include biofuels (Akhtar *et al.*, 2013; Kallio *et al.*, 2014), replacement petroleum-based intermediates (Khusnutdinova *et al.*, 2017), pharmaceutical building blocks (France *et al.*, 2016), cosmetics (Gottardi *et al.*, 2017), and flavorings (e.g. vanillin; Hansen *et al.*, 2013).

There are currently four identified CAR subgroups: Subgroup I make up CARs of bacterial origin, whilst type II-IV make up CARs discovered in a broad spectrum of fungi (Khusnutdinova *et al.*, 2017; Stolterfoht *et al.*, 2017). CAR subgroup I can be further split into five families, of which family CAR1 (the focus of this study) is the best characterized. CARs consist of three distinct domains: an adenylation (A)/thiolation (T) domain, a phosphopantetheine (PPT) binding domain and a reductase (R) domain (supplementary figure 1; Gahloth *et al.*, 2017). The prevailing model for carboxylic acid reduction suggests the CAR reaction proceeds in four steps. The reaction is initiated in the A/T-domain by a nucleophilic attack of the acid on ATP, to form an AMP-acyl ester intermediate. Structural determination of CAR fragments indicate that an A/T-subdomain undertakes a 165° rotation characteristic to the superfamily to which this domain belongs (CL0378; Gottardi *et al.*, 2017; Gahloth *et al.*, 2017). Additionally, the PPT-binding domain undertakes a 75° rotation relative to the A/T-domain (Gahloth *et al.*, 2017). This dynamic re-orientation of the subunits relative to one another presents the AMP-acyl intermediate to the PPT, which displaces AMP to form a PPT-acyl thioester intermediate. This intermediate is then passed to the reductase domain. Here, the intermediate is reduced by NADPH to release an aldehyde product (Current model described in supplementary figure 1).

CAR1 family CARs have demonstrated diverse substrate ranges, with activity against over 100 carboxylic acids (Winkler, 2018; Stolterfoht *et al.*, 2017), including both aromatic acids (Finnigan *et al.*, 2017; Napora-Wijata *et al.*, 2014; Moura *et al.*, 2016), and aliphatic acids

(Finnigan *et al.*, 2017). This diverse substrate range and apparent substrate plasticity highlights CARs' broad potential in green chemistry. However, CARs lack some desirable properties. It has been highlighted that isolation of CARs with improved thermostability is an important goal to improve the carboxylic acid reductase toolbox (Winkler, 2018). Green chemistry pipelines benefit from operating at increased temperatures to improve substrate solubility and reaction rates while mitigating risks of contamination and costs from cooling (Asial *et al.*, 2013; Suplatov *et al.*, 2015; Wijma *et al.*, 2018). Additionally, stable enzymes can often be operated longer than their unstable counterparts, improving per-enzyme productivity per batch reaction, lowering the cost of the enzyme relative to the product (Sheldon, 2016; Asial *et al.*, 2013). Other desirable biocatalytic properties include solvent tolerance, broad substrate ranges, and ready evolvability. We previously reported that well characterised extant CARs (ExCARs) are barely suitable for reactions above 37 °C. The most stable extant CAR (from *Mycobacterium avium*) loses activity rapidly above 48 °C, and retains 50% of activity after incubation for 30 minutes at 49 °C (Kramer *et al.*, 2018). ExCARs also show short half-lives at 37 °C and will likely present a huge metabolic burden for biofactory strains (Finnigan *et al.*, 2017). Therefore, the current state of the CAR toolbox only services batch biocatalysis, which significantly reduces their scale-up potential, hampering their use in biotechnology.

Ancestral sequence reconstruction (ASR) is a popular tool to study the evolutionary histories of protein families. ASR studies of diverse protein families have identified emergent properties of ancestral proteins, including increased thermal stability and altered substrate specificities (Nguyen *et al.*, 2017; Shih *et al.*, 2016; Wilson *et al.*, 2015; Risso *et al.*, 2015). Consequently, a series of studies have used evolutionary histories to isolate sites of interest to engineer enzymes with novel functionality (Alcolombri *et al.*, 2011; Conti *et al.*, 2014; Gonzalez *et al.*, 2014; Miyazaki *et al.*, 2001; Watanabe and Yamagishi, 2006). When used as an engineering tool, ASR has produced enzymes with improved stability (Whitfield *et al.*, 2015), substrate ranges (Wilding *et al.*, 2017), or both (Babkova *et al.*, 2017). ASR differs from other engineering methods as it generates new sequences based upon probabilistic searches of non-conserved functional space, giving each output a high likelihood of being functional given an accurate sequence alignment input. Given enough variation in the input dataset, resulting ancestors can often vary considerably from extant sequences (<30%). This

allows for the discovery of beneficial mutations not accessible by other methods, including coordinated sets of mutations. These can modify traits determined by protein-wide sequence states, including stability under thermal or other stresses.

Notably, all studies to date focusing on ASR for engineering explore ancestral sequence space use a single reconstruction algorithm. Additionally, most available algorithms output “posterior probabilities” at each residue, providing a sequence space representing putative ancestors around a point in sequence space (Yang, 2007; Ashkenazy *et al.*, 2012). Variation within this space is a resource of both sequence and functional diversity (Gaucher *et al.*, 2008). When sampling through these posterior probabilities, there is no “ruleset” dictating the best probability cut-off to efficiently explore space – an issue that has presented in other alignment-based engineering methods (e.g. consensus alignment; Porebski and Buckle, 2016). To avoid this issue, we instead explored the algorithmic variation within the ASR toolbox as a source of sequence and functional diversity by deriving the most likely sequence from multiple maximum likelihood-based reconstruction algorithms. Each algorithm differs subtly, and therefore will output a different, absolute sampling of ancestral sequence space when given the same problem (Yang, 2007; Ashkenazy *et al.*, 2012; Randall *et al.*, 2016; Cai *et al.*, 2004).

Here, we demonstrate the use of ASR to identify three ancient actinomycete CAR biocatalysts that display a 16-34 °C shift in thermal stability compared to ExCARs. The three alternative putative ancestral proteins showed similar substrate ranges, and refined substrate preferences to extant CARs. Comparison of the output from different reconstruction algorithms showed dramatic variations in tolerance to environmental conditions effecting loop-associated traits between the putative ancestral proteins, including tolerance to *in vivo*-like salt concentrations, pH and protic and aprotic solvents. This study represents the largest enzyme to have been reconstructed successfully by ASR to date, and the first reconstruction of an enzyme with four mechanistic steps. This further demonstrates ASR’s potential application to biotechnology and green chemistry.

2.5 Results

2.5.1 Ancestral reconstruction of CARs produces functional enzymes

We previously reported a dataset of 124 CAR homologs identified from the CAR1 family (Finnigan *et al.*, 2017; Chapter 7.3). Of this dataset, 48 sequences representing distinguished clades containing a single genus were used to produce a phylogeny broadly covering CAR sequence space. An alignment of the 48 CAR sequences was created in MUSCLE. Removal of highly divergent regions in the alignment was conducted with the Gblocks algorithm (Talavera and Castresana *et al.*, 2007). ProtTest estimated the best fitting model of amino acid substitution for this alignment to be WAG with independent sites (+I) and a gamma distributed substitution rate (+G; Abascal *et al.*, 2005; Whelan and Goldman, 2001). As only CAR1 enzymes had been reported at the point of reconstruction, we aimed to reconstruct the ancestors of their best represented genera, from *Mycobacterium*, *Nocardia* and *Streptomyces*. To construct the phylogeny, we therefore treated well established sequences from *Tsukamurella* and *Segnilliparus* (the *Tsukamurella* clade) as paralogues, providing an outgroup to the *Mycobacterium*, *Nocardia* and *Streptomyces* clades. The resulting phylogeny was well supported throughout (figure 13A).

Within recent literature, marginal ancestral protein reconstruction has been shown to introduce novel functional properties into proteins (Babkova *et al.*, 2017; Akanuma, 2017; Wheeler *et al.*, 2016). As ancestral proteins typically trend towards increased stability when sampling from more ancient nodes (Gaucher *et al.*, 2008), we reconstructed the most recent common ancestor of the *Nocardia*, *Streptomyces* and *Mycobacterium* CARs. To explore differences in reconstruction algorithm choice had on the sequence and property space sampled in the ancestor, three marginal reconstruction algorithms with optimized likelihood scores were used: FastML (Ashkenazy *et al.*, 2012), PAML (Yang, 2007) and Ancescon (Cai *et al.*, 2004). This produced four putative ancestral proteins: AncCAR-A (Ancescon); AncCAR-F (FastML); and PAML variants with gaps reconstructed by cross-mapping from the other two algorithms producing AncCAR-PA and AncCAR-PF, respectively (supplementary figure 2). AncCARs possessed 95.1% pairwise identity, and 91% conservation across the four proteins, with much of the variation being held in the adenylation domain (figure 13B). Their identity to extant CARs (ExCARs) ranges between 55 and 76%. To explore whether algorithmic

variation was merely sampling variation from posterior probabilities, an ancestor was derived from the PAML output. The most probable residues were substituted with the second most probable residues in the posterior probability table, where the second most probable residues possessed a probability over 30% (AncCAR-P30). The resulting protein shared 93.6% pairwise identity to the algorithm-derived AncCARs. We observed that algorithm-derived variation differed considerably from the posterior probability derived variation (supplementary figure 3). Compared to the most likely AncCAR-P sequence, only 16% and 22% of total derived variation was shared between AncCAR-P30 and AncCAR-A, and AncCAR-P30 and AncCAR-F respectively (supplementary figure 3B). To give confidence that reconstructed ancestral CARs were accurate representations of CAR enzymes, we modelled AncCAR structures using homology modelling to crystal structures 5MST, 5MSD, 5MSP and 5MSO. Comparing all ancestor models to all extant structures, the average root mean squared deviation (rmsd) of alpha-carbon atom position is 0.86 ± 0.32 Å between ancestral models and extant adenylation domains, and 1.01 ± 0.15 Å between ancestral models and extant reductase domains, suggesting a good fit for each model (figure 13C, D; supplementary table 1). Comparison of the models to experimental crystal structures shows that most of the variation between ancestors occurs in surface loop regions (supplementary figure 4). Each AncCAR could be expressed in, and readily purified from *E. coli* to between 3 and 7 mg enzyme per litre (supplementary figure 5). The ancestral CAR proteins demonstrated some protease sensitivity. However, in comparison to extant CARs, they were more resistant to limited proteolysis by common proteases (supplementary figure 6).

Figure 13 - Bayesian inference of actinomycete CAR phylogeny and ASR

A) CAR phylogeny was constructed in MrBayes (Ronquist *et al.*, 2012) under the, WAG+I+G model of amino acid substitution (Whelan and Goldman, 2001), with the *Tsukamurella* clade constrained to the outgroup. The tree was configured in FigTree V1.4.3. The scale-bar represents amino acid changes per site. Node weights represent the posterior probability of a given node calculated from the MCMCMC analysis, with 1 being unequivocal. Red circle represents the target node for ancestral reconstruction. **B)** Identity barcode displaying the pairwise identity over 1,168 amino acid sites between the four ancestors. x-axis denotes residues 1-1168 sequentially, y-axis denotes pairwise identity at a site. Black bars denote pairwise identity (%) at each site. The final four ancestors are conserved at 91% of sites, with 95.1% pairwise identity. Between the ancestors, the greatest diversity is maintained in the A/T domain with a pairwise identity of 93%. Both the phosphopantetheine binding domain and the reductase domains have a higher conservation, at 97.5% and 97.3% identity respectively. AncCAR-F and PF are 1,161 aa in length, AncCAR-A and PA are 1,153 aa in length. Alignment data between ancestors was obtained in Geneious using MUSCLE and modified in Microsoft Excel. Domains are highlighted: **i** – Adenylation domain; **ii** – phosphopantetheine binding di-domain 1; **iii** - phosphopantetheine binding di-domain 2; **iv** – reductase domain. **C)** Model of AncCARs adenylation domain superimposed on extant CAR structure 5MST. Structures: Yellow – 5MST; Green - AncCAR-A; Blue - AncCAR-PA; Orange – AncCAR-F, Red: AncCAR-PF **D)** Model of AncCARs reductase domain superimposed on ExCAR structure 5MSO. Structures: Yellow – 5MSO; Green - AncCAR-A; Blue - AncCAR-PA/PF; Orange – AncCAR-F. Images were produced with PyMol.

2.5.2 AncCARs have a broad substrate range

Assays of AncCAR activity were performed in HEPES instead of the canonical CAR buffer system Tris, as HEPES is more suited to pH 7.5, and Tris was found to inhibit AncCAR activity above 50 mM (supplementary figure 7). AncCARs were screened for activity on 21 aromatic and aliphatic fatty carboxylic acids at 5 mM concentrations. No significant activity could be detected for AncCAR-F on any of these substrates. This protein was therefore eliminated from further kinetic analyses. The other three AncCARs show equivalent substrate ranges to one another across all substrates tested. Ten of the 21 substrates, including nine aromatic carboxylic acids and one aliphatic carboxylic acid showed a statistically significant NADPH

turnover ($P \leq 0.001$) compared to background rate for at least two of the three ancestors (supplementary figure 8). A subset of these are shown in figure 14.

Kinetic analysis of AncCAR activity was first conducted on NADPH and ATP in the presence of 5 mM (*E*)-3-phenylprop-2-enoic acid (supplementary figure 9). For NADPH, AncCAR K_M values were similar to those derived from ExCARs. On the other hand, observed K_M values for ATP were between 10 and 100 times lower than values derived for ExCARs (table 8; chapter 7.3; Finnigan *et al.*, 2017). This suggests the AncCARs bind ATP considerably tighter than modern CAR proteins.

AncCAR kinetics on substrates showing significant activity from background were then tested in saturating NADPH and ATP levels (supplementary figure 10A). The Michaelis constant of all AncCARs was typically determined to be approximately 10-fold higher than those previously reported for the ExCARs (table 9; supplementary figure 10B; chapter 7.3; Finnigan *et al.*, 2017). All AncCARs showed strong activity on canonical substrates: benzoic acid and its derivative 4-methylbenzoic acid. AncCARs have a clear preference for substrates with electron rich conjugated carboxyl groups, with turnovers being amongst the highest across all tested substrates for all ExCARs (table 9; Winkler, 2018). For example, AncCAR-PA turnover of 3-phenylpropionic acid is the highest turnover rate observed for any substrate across all four carboxylic acid reductase subgroups, 1.5-fold higher than that of any substrate reported for the CAR1s (468 min^{-1} ; table 9). Finally, whilst AncCARs are active on octanoic acid, AncCAR preference for fatty acid substrates is attenuated compared to ExCARs, with no activity seen for canonical 3-C and 5-C aliphatics. Octanoic acid was turned over by AncCARs at rates comparable to ExCARs (Finnigan *et al.*, 2017). However, each AncCAR enzyme showed an approximately 100-fold higher K_M (table 9). Octanoic acid also displayed substrate inhibition on AncCARs at high concentrations (supplementary figure 10).

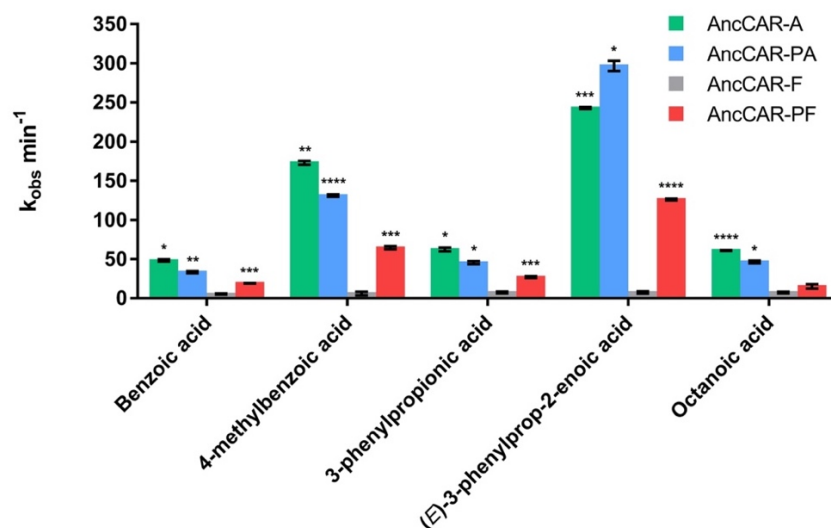


Figure 14 - AncCARs show activity on canonical CAR substrates

ATurnover of NADPH by AncCARs was measured with 24 unique carboxylic acids, of which five are shown. Bar chart shows activity on canonical acid substrates at 5 mM. Each substrate was tested in triplicate, and error bars represent the standard error. Asterisks represent degrees of significance from t-test of triplicate versus all controls (* = $0.0001 < P \leq 0.001$; ** $0.00001 < P \leq 0.0001$; *** = $0.000001 < P \leq 0.00001$; **** = $P \leq 0.000001$). Complete substrate screens are presented in supplementary figure 8.

		ATP	NADPH
AncCAR-A	k_{cat} (min ⁻¹)	340 ± 2.7	386.4 ± 11.1
	K_M (μM)	76.8 ± 4.3	54.8 ± 5.1
	k_{cat}/K_M (min ⁻¹ μM ⁻¹)	4.4 ± 0.3	7.1 ± 0.7
AncCAR-PA	k_{cat} (min ⁻¹)	392.3 ± 11.2	482.2 ± 15
	K_M (μM)	69.1 ± 6.7	58.5 ± 5.0
	k_{cat}/K_M (min ⁻¹ μM ⁻¹)	5.7 ± 0.6	8.2 ± 0.8
AncCAR-PF	k_{cat} (min ⁻¹)	219.2 ± 3.1	230.6 ± 2.5
	K_M (μM)	42.9 ± 2.3	29.0 ± 1.2
	k_{cat}/K_M (min ⁻¹ μM ⁻¹)	5.1 ± 0.3	8.0 ± 0.3
Extant CARs ¹¹	K_M (mM)	64–84	24–36

Table 8 – AncCAR co-factor kinetics

Rates of AncCAR activity on ATP and NADPH were determined using a 12 point, 1.7x dilution series of substrate, with concentrations starting at 800 mM. Each concentration was investigated in triplicate. Data were fitted to the Michaelis-Menten equation. Graphs in supplementary figure 8. Extant CAR data was derived from the literature.

		Benzoic acid	4-methylbenzoic acid	3-phenylpropionic acid	(E)-3-phenylprop-2-enoic acid	Octanoic acid
AncCAR-A	k_{cat} (min ⁻¹)	149.1 ± 7.1	398.4 ± 13.9	327.3 ± 16.9	203.5 ± 3.8	302.6 ± 19.2
	K_M (mM)	61.2 ± 5.6	6.5 ± 0.4	33.0 ± 3.4	0.9 ± 0.06	8.7 ± 1.1
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	2.4 ± 0.3	61.3 ± 4.3	9.9 ± 1.1	226.1 ± 15.7	34.9 ± 4.9
AncCAR-PA	k_{cat} (min ⁻¹)	176.4 ± 7.7	146.6 ± 6.8	468.1 ± 36.7	396.4 ± 5.6	344.1 ± 27.1
	K_M (mM)	81.4 ± 6.2	5.5 ± 0.5	68.2 ± 8.5	4.0 ± 0.1	11.0 ± 1.6
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	2.2 ± 0.2	26.7 ± 2.7	6.9 ± 1.0	99.1 ± 2.8	31.3 ± 5.1
AncCAR-PF	k_{cat} (min ⁻¹)	71.8 ± 2.7	61.9 ± 2.8	325.6 ± 26.3	193.8 ± 8.0	169.0 ± 13.7
	K_M (mM)	79.9 ± 5.4	7.0 ± 0.5	98.7 ± 11.5	3.8 ± 0.3	11.1 ± 1.7
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	0.9 ± 0.1	8.8 ± 0.7	3.3 ± 0.5	51.0 ± 4.5	15.2 ± 2.6
MpCAR (Finnigan et al., 2017)	k_{cat} (min ⁻¹)	140 ± 20	122 ± 3	21.5 ± 0.7	67 ± 2	58 ± 1
	K_M (mM)	20 ± 4	3.7 ± 2	3.0 ± 0.3	0.3 ± 0.02	2.0 ± 0.1
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	7 ± 1	33 ± 2	7.2 ± 0.7	240 ± 2	29 ± 2

Table 9 – AncCAR substrate kinetics

For ancestral CARs, kinetic rates on various aromatic and aliphatic compounds were determined using an 8 point, 1.7x dilution series of acid from near saturation in 125 mM HEPES. Each concentration was investigated in triplicate. Data were fitted to the Michaelis-Menten equation using GraphPad v.7.0. Graphs in supplementary figure 10. Corresponding kinetic values are presented in supplementary table 2 MpCAR data was derived from the literature.

In AncCAR homology models, we observed that the active site of the ancestors' adenylation domains appear to be slightly disordered compared to the extant structures (figure 15A; Stolterfoht *et al.*, 2017). This is most evident when comparing differences in a variable loop that stretches into the active site between positions 286-302. The catalytically essential His315 is positioned as a rotamer away from the substrate, suggesting this residue has a large sampling space within the active site of the AncCARs. In the model of AncCAR-PF (figure 15B), this loop region is significantly shortened and is unable to contact the substrate. Comparison of inactive AncCAR-F to ancestor models and ExCAR structures showed no obvious structural or functional residue changes that explain the loss of activity (supplementary figure 11).

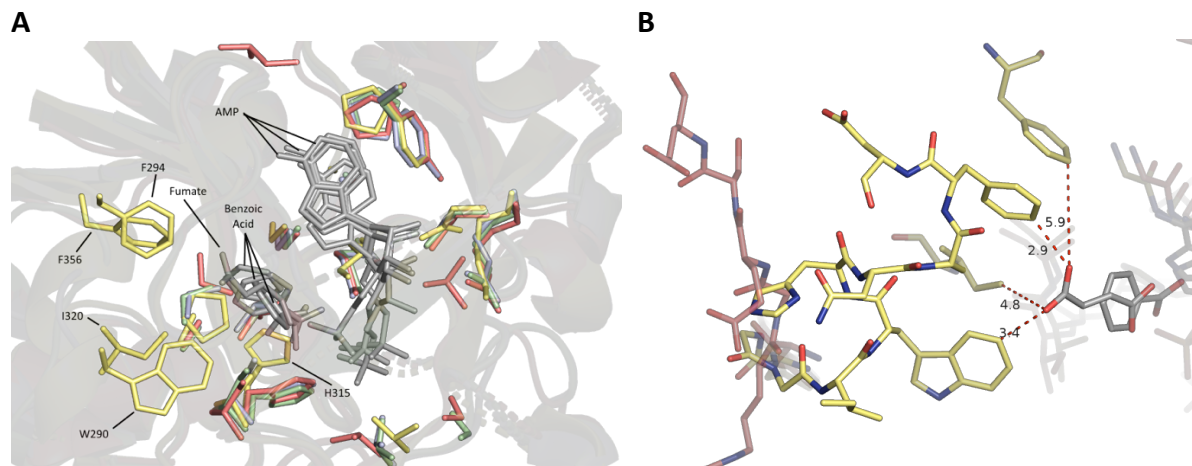


Figure 15 – Homology models suggest slight disordering of AncCAR active sites

A) The predicted active site structure of the adenylation domain. AncCARs A (green), PA (blue) and PF (red) are overlaid onto *S. rusogus* CAR (PDB ID: 5MST; yellow). By the change in substrate position (gray) and placement of the residues around the substrate, it can be seen that the shape of the active site varies between ancestors, compared to SrCAR. The residues lining the active site pocket of ExCARs (positions 246-250) are poorly resolved. **B)** In AncCAR-PF, the highly variable loop between positions 286-302 of the adenylation domain of AncCAR-PF (red) does not interact with the substrate, in contrast to SrCAR (5MST; yellow). Model structures of AncCAR-A, PA, and PF were produced in YASARA, and visualized in PyMol.

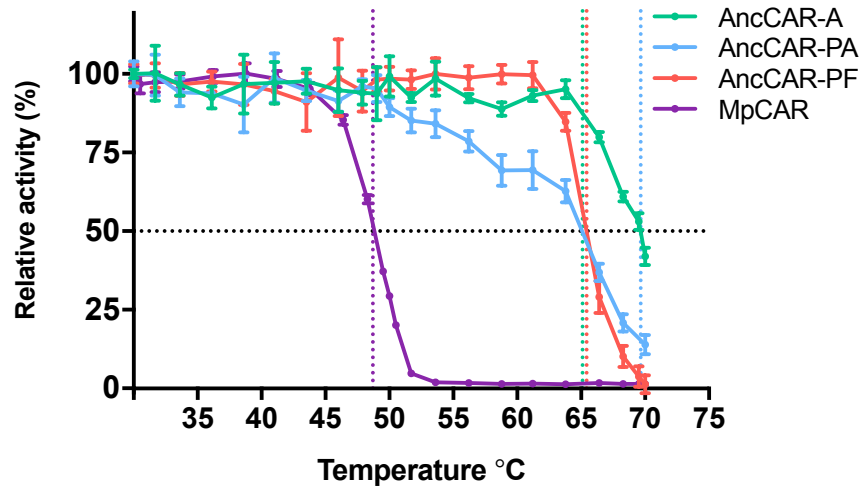
2.5.3 Ancestral CARs show dramatic increases in stability

Many ancestral proteins have displayed increased resistance to temperature (Whitfield *et al.*, 2015; Gaucher *et al.*, 2008; Akanuma, 2017; Hobbs *et al.*, 2012; Butzin *et al.*, 2013; Zakas *et al.*, 2015; Trudeau *et al.*, 2016; Okafor *et al.*, 2018). AncCAR-A is the most thermostable ancestor, and the most stable CAR protein reported to date, with an A_{50} of around 70 °C. 50% activity was retained for AncCAR-PA and AncCAR-PF at 65.1 °C and 65.4 °C respectively. Comparatively, MpCAR, one of the most stable ExCARs reported to date (Finnigan *et al.*, 2017; chapter 7.3), displayed an A_{50} of around 49 °C (figure 16A). AncCAR half-life at 37 °C in 50 mM HEPES was monitored by assessing their activity on 5 mM (*E*)-3-phenylprop-2-enoic acid at intervals over a period of 10 days. AncCAR-A showed a short half-life of less than 41 hours. This was of stark contrast to AncCAR-PA and AncCAR-PF, whose half-lives at 37 °C were between 168-216 hours. AncCAR-PA and AncCAR-PF display the longest half-lives reported to date in CARs, approximately 27-fold longer than the half-life observed for MpCAR (7 hours; figure 16B). Monitoring of AncCAR unfolding in real time with differential scanning fluorimetry (Senisterra *et al.*, 2011; Vivoli *et al.*, 2014) also corroborates that AncCARs are highly stable. All AncCARs showed the greatest rate of unfolding (T_m) between 67 and 68 °C (figure 16C).

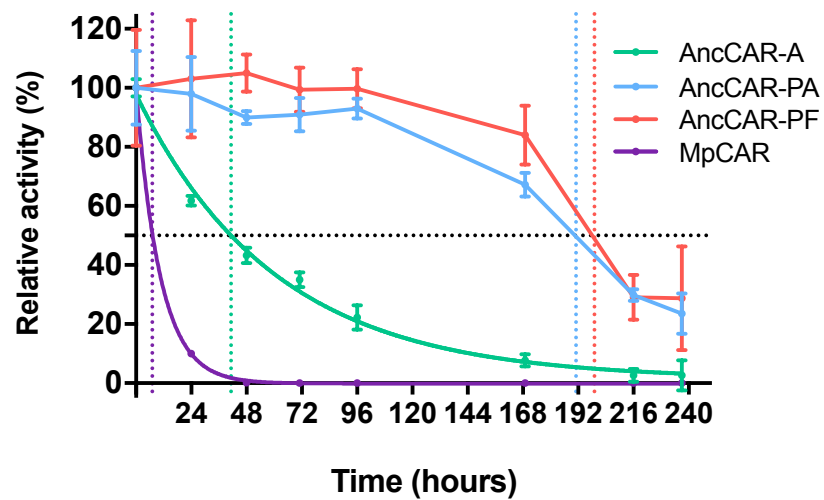
Importantly, biocatalysts are used for both *in vitro* and *in vivo* bioindustrial pipelines. Robust biocatalysts are therefore required to function in the highly ionic environments demanded by *in vivo* bioconversions. Ionic solutions can have either stabilizing or destabilizing effects on enzymes (Dominy *et al.*, 2002). To better characterize AncCARs for use in the CAR toolbox, their thermostability was assessed in a buffer simulating the ionic environment inside a *Saccharomyces cerevisiae* cell (van Eunen *et al.*, 2010). In these potentially challenging conditions, MpCAR slightly destabilized, with an A_{50} of around 47 °C. AncCAR-PA was the least thermostable ancestor, with an A_{50} of 45 °C – a 20 °C decrease over incubation in standard *in vitro* assay conditions. AncCAR-A showed a 16 °C decrease in stability over the salt free condition, presenting an A_{50} of approximately 54 °C, losing activity in a near linear fashion from around 40 °C. AncCAR-PF is the only ancestral protein observed to be tolerant to an ionic environment at temperature, with an equivalent A_{50} to the salt free condition at 65 °C (figure 16D). To confirm it was the presence of salt that was effecting stability in AncCARs and MpCAR, we repeated the experiment in the presence 500 mM NaCl.

Equivalent destabilizing effects to the *in vivo*-like conditions were observed (supplementary figure 12).

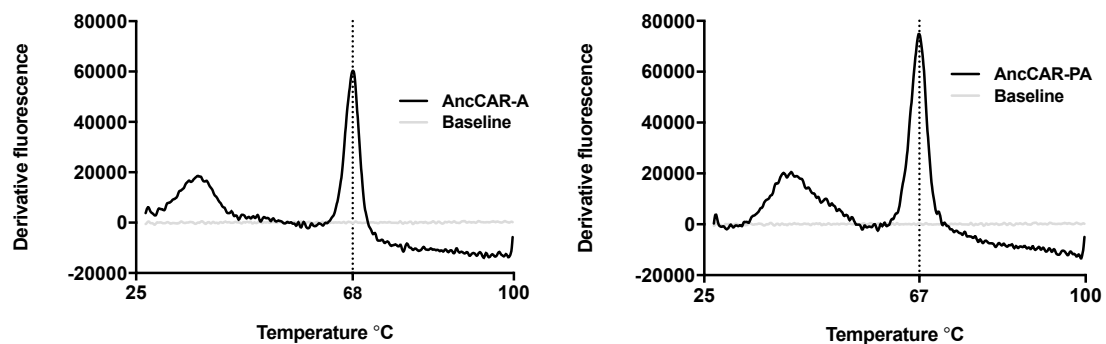
A



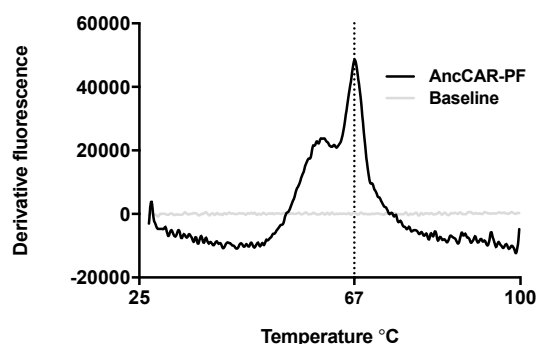
B



C



C (cont.)



D

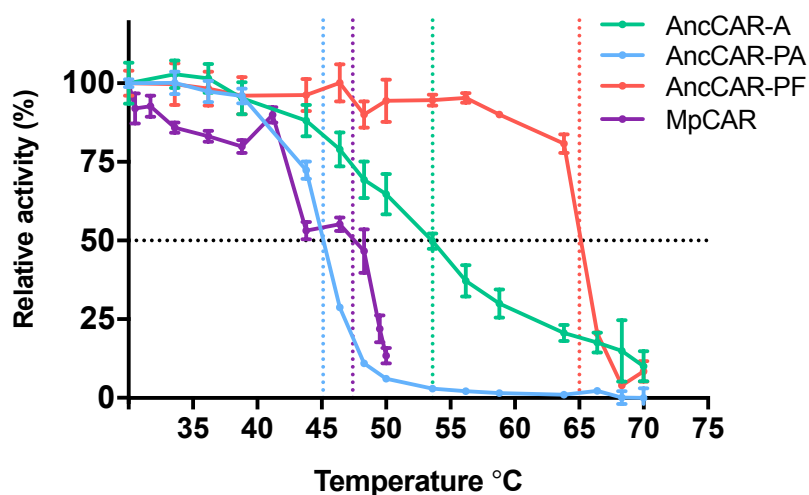


Figure 16 - AncCARs are thermostable enzymes

A) AncCARs and MpCAR were incubated in 50 mM HEPES at temperatures from 30 °C to 70 °C for 30 min. Each point represents the rate of NADPH oxidation in 5 mM (*E*)-3-phenylprop-2-enoic acid at temperature relative to the rate of NADPH oxidation in 5 mM (*E*)-3-phenylprop-2-ionic acid at 30 °C. Error bars represent the standard error (all data taken in triplicate). Black dotted horizontal line represents 50% activity (A_{50}). Coloured vertical dotted lines represent temperature at which A_{50} is reached. MpCAR retains 50% activity at 48.7 °C. AncCAR-A retains 50% activity at 69.7 °C. Both AncCAR-PA and AncCAR-PF show similar stability, with A_{50} (temperature where 50% of activity remains) of 65.1 °C and 65.4 °C respectively. **B)** To assess half-life at 37 °C, AncCARs and MpCAR were incubated at temperature over a period of 10 days. Relative activity versus a zero-time point was assessed by activity on 5 mM (*E*)-3-phenylprop-2-enoic acid. The black dotted horizontal line represents 50% activity. Coloured vertical dotted lines represent time taken to reach 50% enzyme activity compared to time zero. Error bars represent standard error, in all cases calculated from three experimental replicates. **C)** Differential scanning fluorimetry to assess AncCARs' critical unfolding temperature. Enzymes were incubated in HEPES and analysed between the temperature

of 25 °C and 100 °C. Thermal shift curves were drawn from raw DSF data in GraphPad. **D)** AncCARs have environment dependent temperature resistance. AncCARs and MpCAR were incubated in *in vivo*-like ionic concentrations that model the internal environment of a *S. cerevisiae* (van Eunen *et al.*, 2010) cell at temperatures from 30 °C to 70 °C. Data were determined and represented as in panel A.

2.5.4 AncCARs vary in their loop-based properties

Salt-tolerance has been proposed as a “loop-associated” trait, where net surface charge effects solvent penetrance (Dominy *et al.*, 2002). Following our observation that much of the variation between ancestors is loop based, we further investigated AncCAR’s resistance to other proposed loop associated conditions. Solvent tolerance is a common industrially relevant loop-associated property desired in biocatalysts. We assessed the AncCARs’ solvent tolerance in a range of protic and aprotic solvents at increasing solvent concentrations in comparison to MpCAR and NiCAR (supplementary table 3). There is no consistent trend observable between all ancestors on all solvents. AncCAR-A is the least solvent tolerant enzyme for all solvents besides DMSO and methanol. For all solvents besides acetone, AncCAR-PF is the most solvent tolerant, retaining 50% activity in the presence of over 25% methanol. Ancestors show the greatest variance to tolerance in methanol, with AncCAR-PF showing considerable increases in tolerable concentration of solvent compared to AncCAR-A (89%) and AncCAR-PA (119% increase). In protic solvents, AncCAR-PF performed similarly to the most solvent tolerant extant CAR. In contrast, in aprotic solvents, the AncCARs generally showed greater activity than extant CARs. This was particularly so in 10% DMSO (v/v), with all ancestral proteins retaining 86-92% activity, compared to 67-74% for the extant CARs (figure 17A). A wide pH tolerance for industrially relevant enzymes is another highly desirable loop-associated trait. All AncCARs displayed no loss of activity between 6.0 and 9.0 pH units (figure 17B). AncCAR-A lost activity in alkaline conditions above pH 9.0, whereas AncCAR-PF and AncCAR-PA maintained 100% activity up to pH 10.0. All ancestral CARs show a decrease in activity below pH 6.0 ($pK_1 \approx 5$ for all three enzymes). However, this feature is shared with extant CARs, with MpCAR showing even greater pH tolerance than AncCAR-PF, the most pH tolerant of the AncCARs (50% activity between pH 5.01 and pH 11.56 for AncCAR-PF, compared to pH 4.3 to 11.8 for MpCAR).

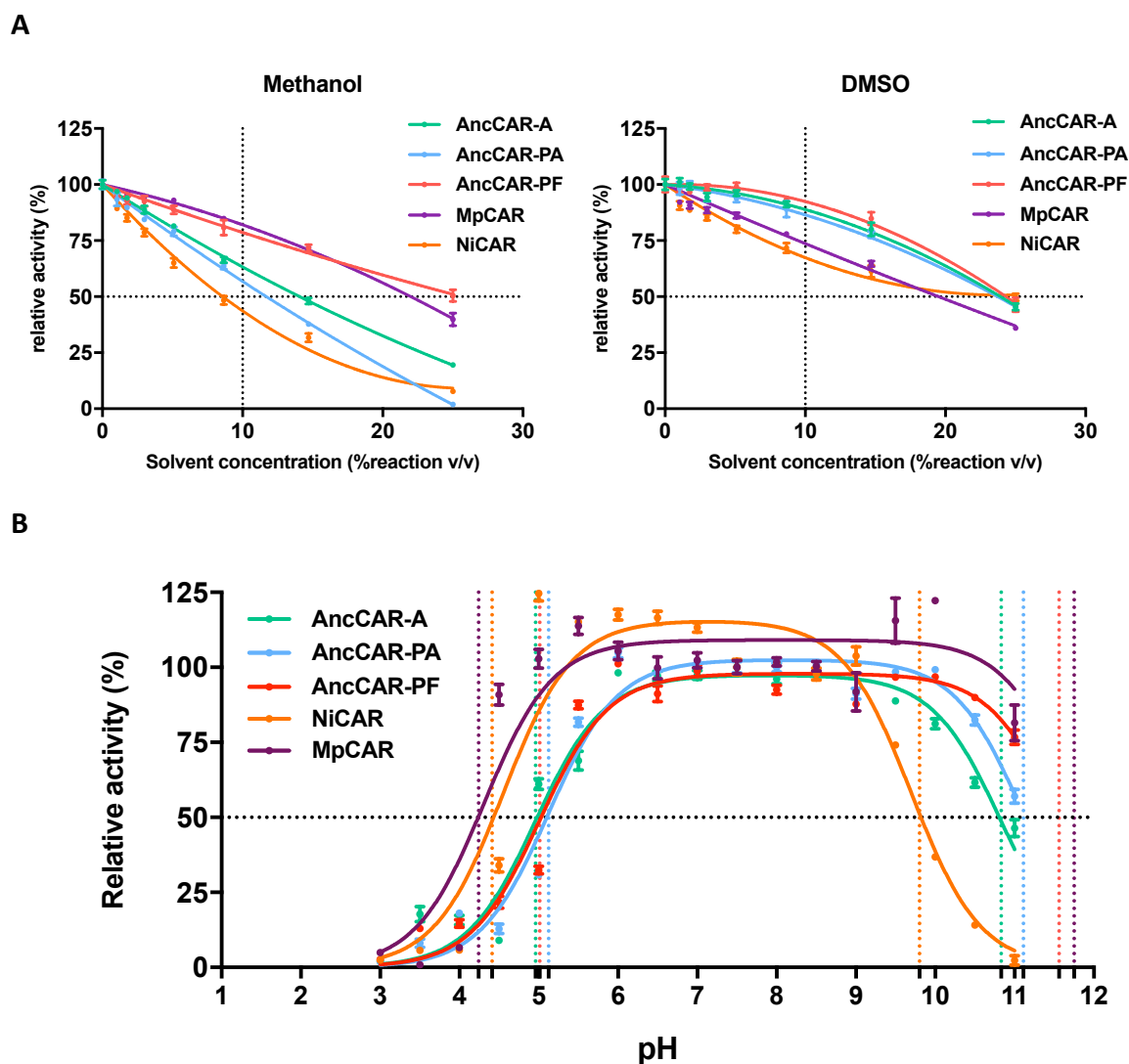


Figure 17 - AncCAR tolerance to solvents loop dependent environmental factors

A) AncCAR and ExCAR activity on 5 mM (*E*)-3-phenylprop-2-enoic acid was assessed in aprotic and protic solvents by solvent titration from 25% (v/v). Graphs represent relative activity of each AncCAR and ExCAR at increasing concentrations of solvent compared to 0% solvent. Error bars are standard error (three replicates). Data for all solvents can be found in supplementary table 3. **B)** To assess the resistance of AncCARs folding to pH, AncCARs were incubated for 30 minutes in 0.5 pH increments between pH 3 and 11, before being assayed for their turnover of NADPH in the presence of 5 mM (*E*)-3-phenylprop-2-enoic acid relative to turnover at pH 7.5 (100%). Data was analysed in GraphPad Prism 7.0. pK_1 and pK_2 values were calculated respectively as: AncCAR-A - 4.96 ± 0.06 and 10.83 ± 0.06 ; AncCAR-PA - 5.12 ± 0.05 and 11.11 ± 0.07 ; AncCAR-PF - 5.011 ± 0.06 and 11.56 ± 0.11 ; NiCAR - 4.55 ± 0.09 and 9.70 ± 0.09 ; MpCAR - 4.3 ± 0.1 and 11.8 ± 0.3 . Error bars represented standard error, calculated from three experimental replicates.

2.6 Discussion

Protein engineering for the optimization of application specific properties in enzymes is integral to the future green chemistry market. Limited understanding about the sequence-function relationship in biocatalysts presents a significant challenge for synthetic biology. This is exemplified by the CARs. Single amino acids that regulate CAR function and selectivity are starting to be uncovered, including active site point mutants that modulate substrate turnover (Stolterfoht *et al.*, 2017). Nevertheless, at present without significant innovation in the protein engineering field the semi-rational engineering of CARs with high-throughput approaches would remain prohibitively expensive. Furthermore, CARs with improved stability have been highlighted as an important potential addition to the CAR toolbox (Winkler, 2018). However, there are no defined rules available to guide the rational engineering of thermostability in any enzyme, let alone one as complex and poorly understood as CARs (Okafor *et al.*, 2018).

Here, we aimed to sample ancient sequence space using multiple ASR algorithms to engineer stability into CARs. CARs present as challenging targets for ASR: they are large (>1,100 amino acids), and undertake four catalytic steps including two large scale domain reorientations (Finnigan *et al.*, 2017; Gahloth *et al.*, 2017). In the first instance, it is therefore surprising that all four reconstructed enzymes could be readily expressed and purified in *E. coli* (supplementary figure 5). It is even more surprising that three of the four putative ancestors were functional CAR-like enzymes, showing unambiguous CAR activity against a range of standard CAR substrates (figure 14; supplementary figures 9 and 10). AncCARs identified were highly conserved (91% identity; figure 13B). Despite such high conservation, a broad functional space was identified. Our use of multiple reconstruction algorithms also allowed for the sampling of sequence space in an empirical manner. Homology modelling suggests that variation between the ancestors is concentrated at surface loops, mostly within the A domain (figure 13B; supplementary figure 4B). Loops are flexible regions within a protein that can exhibit large degrees of motion, are often tolerant to amino acid substitution and are a key determinants of protein stability (Papaleo *et al.*, 2016; Balasco *et al.*, 2013). We observe that AncCARs vary in their loop-dependent

properties, with variation in their tolerance to *in vivo* like salt concentrations (figure 16D), in their activity in protic and aprotic solvents (figure 17A; supplementary table 3), and in their tolerance to alkaline conditions (figure 17B). Such conditions modify the sum of zwitterionic states across the protein surface, causing repulsive forces within the protein's loop-regions. In turn, increased repulsion of loops expose the hydrophobic core of the protein to bulk solvent (Dominy *et al.*, 2002; Nestl and Hauer, 2014; Dill, 1990). As loop regions are resistant to the deleterious effects of mutation, they are more likely to vary in the extant dataset, allowing ASR-based searches of ancient sequence space to capture this variation at the functional level. These results highlight the potential of ASR as an engineering tool even for large, complex biomolecules that are otherwise less tractable for protein engineering.

The different ancestral reconstruction algorithms that we used apply subtly different gapping regimens. These are likely to partly explain the variation in both loop-based properties and reaction rates reported between ancestors. Altering loop lengths can modify structural flexibility, with a concomitant impact on stability as discussed above (Balasco *et al.*, 2013; Nestl and Hauer, 2014; Dill, 1990). This highlights the importance of gap reconstruction in ASR studies. We would encourage other ASR users to attempt reconstruction with multiple ASR algorithms when working with alignments that contain gaps, to confirm that gap placement is coordinated between methodologies. In cases where the sequence identity is sufficiently high to eliminate any ambiguity in gap locations, the use of multiple algorithms may be less important. We would also encourage future ASR engineering studies to include consideration of gap placement to expand understand of the impact this has on obtainable property space.

This study expands on previous work investigating ASR's use as a protein engineering tool, confirming its tractability to the engineering of large, mechanistically complex multi-domain proteins. Importantly, all functional CAR ancestors were found to be highly thermostable ($A_{50} > 65^{\circ}\text{C}$) in simple buffer conditions (figure 16A). AncCAR-A, with around 50% activity retained after incubation at 70°C , shows $21\text{-}34^{\circ}\text{C}$ greater thermostability than the best studied extant CARs (Finnigan *et al.*, 2017; Kramer *et al.*, 2018). However, ancestors showed considerable variation in their half-lives, with AncCAR-A losing 50% activity on just over 40 hours, whereas AncCAR-PA and PF maintained at least 50% activity for over a week (figure

16B). As we are not aware of a highly thermostable CAR variant that exists in the CAR toolbox, ancient CAR enzymes provide much needed functionality, providing a means to convert carboxylic acids into aldehydes within a high temperature biocatalysis. Overall, AncCAR-PF presents as an attractive, all-purpose CAR enzyme due to its extraordinarily hardy nature, and broad scale resistance to many challenging conditions. It is stable up to around 65 °C in both *in vivo* and *in vitro* conditions, it has a half-life of over a week, it has a pH range of 6.5 pH units and it exhibits the highest tolerance to solvent in all tested cases besides acetone. These collective properties are highly desired in CAR enzymes due to the poor solubility of their aldehyde products, ensuring efficient coupling to downstream bioconversions. On the other hand, AncCAR-A and AncCAR-PA appear to be excellent biocatalysts for the production of cinnamic aldehyde derivatives. AncCAR-PA's turnover of 3-phenylpropionic acid is the highest turnover rate observed to date for any CAR from any family on any substrate.

Importantly, such enzyme improvements are of broad industrial relevance as they were achieved with free software, without the prerequisite of an experimental structure, and without having to produce or screen a library of variants. ASR's delivery of large stability increases will therefore offer a cost and time saving opportunity in current protein engineering pipelines. We further anticipate that ASR will not replace existing engineering pipelines, but instead act as a front-end process. Enzymes with increased stability "smooth" the sequence-function landscape. This occurs as stable enzymes can permit the introduction of destabilizing mutations without cost to enzyme fitness, thus improving mutational robustness and introducing new avenues for property discovery (Suplatov *et al.*, 2015; Romero and Arnold, 2009). It therefore follows that ancestral enzymes could be more easily engineered for improved or refined activities (Wagner, 2008). Being able to rapidly "strip back" enzymes to a more plastic molecule may provide improved avenues for more complex protein engineering pipelines.

In terms of experimental design in ASR, our observation of a rich ancestral property space informs important considerations. It is commonplace in today's ancestral reconstruction literature, whether focused on engineering or on evolution, that ancestors are constructed from nodes in a single lineage or small number of lineages to understand their properties

(Zakas *et al.*, 2015; Blanchet *et al.*, 2017; Voordeckers *et al.*, 2012). To our knowledge, only benchmarking studies have assessed the difference between algorithms, and have done so on a very small number of enzyme targets (Randall *et al.*, 2016; Hanson-Smith *et al.*, 2010). However, our work shows that the properties of sequences derived from different algorithms differ based on ancestral reconstruction method, yet no algorithm can be argued to provide more confident representation of ancestral space. Therefore, in future ASR work, comparisons between ancestors made with different algorithms might provide better insight into ancestral property space. Additionally, we show that ancestors exhibit vastly different stability profiles, dependent on whether the proteins are being assayed within an *in vitro* and *in vivo*-like environment. To confidently conclude that thermostable proteins confer a high stability of ancient life, proteins must be seen to be thermostable in *in vivo* conditions, as the limits of protein stability within the cell environment define the environmental limits in which an organism can survive (Karshikoff *et al.*, 2015). To our knowledge, all ASR studies that address the temperature environment of early life only test their proteins using *in vitro* conditions (Nguyen *et al.*, 2017; Risso *et al.*, 2015; Gaucher *et al.*, 2008; Butzin *et al.*, 2013; Hobbs *et al.*, 2012; Trudeau *et al.*, 2016). We therefore encourage caution be taken when drawing conclusions about a protein's environment based on *in vitro* stability alone, as well as conclusions drawn from one representation of ancestral space at a given node.

2.7 Conclusion

ASR offered an attractive solution for engineering CARs, as their complexity makes them intractable to conventional protein engineering methods. ASR has a high likelihood of obtaining functional sequences, as every extant sequence referenced already contains permitted residues at each position. Here, using ASR, we have successfully engineered three functional carboxylic acid reductase enzymes with novel properties tractable to biotechnology. All three ancestors bring valuable properties to the CAR toolbox, providing novel enzymes with stable and robust properties. These properties unlock an entirely new array of biochemical capabilities for CAR reactions particularly in the applications of high temperature biosynthesis. Additionally, stable AncCARs may prove useful for future enzyme engineering studies with this enzyme. We show that ancestral reconstruction with multiple algorithms offers an important engineering technology for large and/or poorly understood protein families.

2.8 Methods

Sequence handling

Unless specified, all algorithms were performed under default settings. Multiple sequence alignments were performed in Geneious version 10.0.2 with MUSCLE (Kearse et al., 2012; Edgar, 2004). The resulting alignments were modified manually. These were then further modified by either: a) manually removing insertions represented by one, or very few leaves; and b) the GBlocks algorithm in the Phylogeny.fr program suite (Talavera and Castresana, 2007; Dereeper *et al.*, 2008), forming two distinct alignment datasets. Best fit models of amino acid replacement were identified using ProtTest version 3.4 (Abascal *et al.*, 2005). The GBlocks curated alignment was subject to phylogenetic analysis within MrBayes version 3.2.6 (Ronquist *et al.*, 2012), under the WAG + I + G model of amino acid substitution (Whelan and Goldman, 2001). Two parallel runs of 250,000 Metropolis Coupled Markov-Chain Monte Carlo generations were conducted with an independent gamma calculated for all lineages, each with four chains with the heat prior set to 0.02, sampled every 100

generations, with a burn-in of 25%, and all sequences bar those from *Tsukamuraella* and *Segnilliparus* set as the ingroup prior.

Ancestral sequence reconstruction was conducted with FastML (Ashkenazy *et al.*, 2012), PAML (Yang, 2007) and Ancescon (Cai *et al.*, 2004) using the manually-curated alignment and the MrBayes tree as inputs. Marginal reconstructions conducted in FastML and PAML were run with the most optimal model available previously defined by ProtTest. PAML was run with eight gamma rate categories with estimated shape parameters for α , κ and ω priors. FastML was run with optimization of branch lengths and binary maximum likelihood based indel reconstruction. Ancescon requires a polytomous root in the input tree: therefore, the MrBayes derived tree had a false polytomy introduced manually in its Newick file. Marginal reconstructions in Ancescon were run with ML based rate factors and an alignment-based PI vector. Most likely output sequences for each algorithm were aligned in Geneious using MUSCLE. Indels derived from either Ancescon or FastML were transposed to the PAML sequences, producing four final sequences: AncCAR-A, AncCAR-F, AncCAR-PA and AncCAR-PF. All sequences are available as supplementary documents.

Homology modelling of AncCARs

The ancestral CARs were modelled using YASARA v.17.8.15 (Krieger and Vriend, 2014). The models were based on the structures of the A/T domains of CARs from *Segniliparus rugosus* (SrCAR; PDB ID: 5MST) and NiCAR (PDB ID: 5MSD); and the R domains of CARs from *Mycobacterium marinum* (MmCAR; PDB ID: 5MSO) and SrCAR (PDB ID: 5MSP; Gahloth *et al.*, 2017). The alignments used for the ancestral reconstruction were used to direct the modelling. Modelling was performed using the default “hmbuild” algorithm. In each case, the preferred model was selected. Images of protein structures were prepared using PyMOL v. 2.0 (Schrödinger; DeLano, 2002). Root mean squared values for alpha carbon atom position in the modelled structures were calculated in PyMOL by calculating the best alignment without transform over 10 cycles.

Purification and storage

Sequences derived from ancestral sequence reconstruction were modified to contain a 6xHis-tag at the N-terminus. Sequences were codon optimized for *E. coli* K12, and synthesized in two sections. The first sections were synthesized into the pNic28-BSA4 expression vector (Savitsky *et al.*, 2010) by Twist Bioscience. The second sections were synthesized by Twist bioscience into their stock vector. The second sections were ligated with the first section and vector by restriction cloning, and sequence verified by Sanger sequencing (Source Bioscience). These were co-transformed into BL21(DE3) *E. coli* alongside a pCDF-Duet1 vector containing *Bacillus subtilis* phosphopantetheine transferase (Finnigan *et al.*, 2017).

Expression was carried out in LB media supplemented with 150 μ M IPTG at 20 °C overnight. Cells were harvested in 20 mM Tris-HCl pH 7.5 with 0.5 M NaCl and 10 mM imidazole and lysed by sonication. The lysate was clarified by centrifugation at 24,000 *g*. AncCARs were purified from the soluble fraction by nickel affinity using an ÄKTAXpress (GE Healthcare) using a 1 ml His-Trap FF crude column (GE Healthcare), followed by size exclusion with a Superdex 200 HiLoad 16/60 gel filtration column (GE Healthcare). The nickel affinity column was equilibrated and washed with the cell lysis buffer, and the purified proteins eluted with cell lysis buffer supplemented with 250 mM imidazole. The size exclusion column was eluted with 0.5 M NaCl in 10 mM HEPES-NaOH pH 7.5. The purified proteins were analyzed by SDS-PAGE using 4-12% precast gels run in MOPS buffer (Genscript). Protein concentration was determined using a Nanodrop N2000c nanospectrophotometer (Thermo). If required samples were concentrated to between 0.25 and 0.5 mg/ml using Vivaspin 6 mL columns with a molecular weight cut-off of 10 kDa (Generon) and stored in 20% (v/v) glycerol at -20 °C. Protein was buffer exchanged into reaction buffer using PD10 desalting columns (Generon) before enzymatic analysis.

Enzyme assays: standard conditions

All assays were performed in Grenier flat-bottomed 96 well microtitre plates. Assays were modified from those in Finnigan *et al.* (2017). Unless otherwise specified, samples were assayed in triplicate in a 200 μ l reaction containing 125 mM HEPES-NaOH (pH 7.5), 1.2 mM

ATP, 10 mM MgCl₂, 250 μM NADPH, 5 mM substrate and 5 μg enzyme. Working stocks of each assay component were dissolved in 50 mM HEPES-NaOH (pH 7.5). Their pH was modified to 7.5 to ensure consistent pH across serial dilutions, and volume was made up to 50 mM final concentration of HEPES with MilliQ water. Where necessary, substrates were dissolved in concentrations of DMSO up to 10% (v/v) final reaction in 200 mM HEPES pH 7.5. To begin the reaction, 100 μl substrate working stock in assay buffer was added to 100 μl of a master mix containing the remaining components. Each assay contained substrate buffer solution without substrate in triplicate for blank subtraction of native NADPH degradation rates. Enzyme activity was monitored at 30 °C by measuring the absorbance at 340 nm in a Tecan Infinite 200Pro plate reader in continuous cycles over the course of 10 minutes with 10 flashes per-well; or using a ThermoFisher SkanIt Pro plate reader in continuous cycles over 10 minutes. Data were processed in Microsoft Excel and Graphpad Prism v7.0. Experimental data were fit to the Michaelis-Menten equation following calculation of NADPH conversion based on an NADPH standard curve (supplementary figure 13).

Buffer optimization

HEPES and Tris were prepared to pH 7.5 at 50 mM, 75 mM, 100 mM, 125 mM, 150 mM and 275 mM. AncCARs were buffer exchanged into each buffer. AncCAR activity was tested against (*E*)-3-phenylprop-2-enoic acid. All reaction components were prepared in corresponding buffers.

Analysis of solvent stability

(*E*)-3-phenylprop-2-enoic acid dissolved in 50 mM HEPES was prepared in 50% (v/v) neat solvent, which was serially diluted in 5 mM (*E*)-3-phenylprop-2-enoic acid dissolved in 50 mM HEPES to provide a solvent gradient from 25% to 0% (v/v).

Analysis of substrate specificity

CAR activity was tested for each enzyme on seventeen aromatic carboxylic acids and four aliphatic carboxylic acids. Compounds were prepared to 0.5 M stocks in neat DMSO and

diluted to working concentration in assay buffer to a final DMSO concentration of 20%, providing a 10% (v/v) DMSO concentration on the standard assay.

pH tolerance

Buffers ranged from pH 3 to 11 in increments of 0.5, prepared at 30 °C. The buffers consisted 50 mM Na-citrate, pH 3.0 to 5.0; 50 mM MES, pH 5.5 to 6.5; 50 mM HEPES, pH 7.0 to 8.0; 50 mM Bicine pH, 8.5 to 9.0; and 50 mM CAPS, pH 9.5 to 11.0. A series of 80 µL buffer solutions containing 0.25 µg µl⁻¹ ancestral protein was constructed and incubated at 30 °C for 30 minutes. Incubated enzymes were assayed as standard on 5 mM (*E*)-3-phenylprop-2-enoic acid. Initial rates were calculated as relative activity against acquired rate values at pH 7.5 (100 %). The data were fitted to the following equation (Cornish-Bowden, 2013) to determine the limits of pH tolerance:

$$v = \frac{V_{100}}{\frac{h}{K_1} + 1 + \frac{K_2}{h}}$$

Where V_{100} is the maximum rate, K_1 and K_2 are the proton concentrations where activity drops to 50% at low and high pH respectively, and h is the proton concentration.

Thermostability following incubation

In vitro buffer system consisted standard assay buffer. *In vivo*-like *S. cerevisiae* ion buffer was based on systems described by van Eunen *et al.* (2010). Buffer consisted of 50 mM K₂HPO₄, 75 mM C₅H₉NO₄, 85 mM KCl, 10 mM Na₂SO₄, 2 mM MgCl₂, 0.5 mM CaCl₂, prepared in 50 mM HEPES and pH modified to 7.5 by adding neat KOH (45%, v/v) dropwise. Salt confirmation buffer was standard assay buffer supplemented with 500 mM NaCl.

80 µl aliquots of each AncCAR at 0.25 µg µl⁻¹ in each buffer system were incubated for 30 minutes at temperatures between 30 °C and 49 °C, and 50 °C and 70 °C in a Mastercycler nexus thermocycler (Eppendorf) set to gradient mode. The second aliquot in each gradient was reserved for 80 µl buffer for a negative control. Enzymes were then cooled to 4 °C in

the thermocycler for 5 minutes before being assayed as standard on 5 mM (*E*)-3-phenylprop-2-enoic acid.

Differential Scanning Fluorimetry (DSF)

The cleanest peaks from the size exclusion step of protein purification were buffer exchanged into each 50 mM HEPES pH 7.5. DSF running mixture was prepared by diluting enzyme to 0.1 $\mu\text{g } \mu\text{l}^{-1}$ to which 10X SYPRO orange was added (Vivoli *et al.*, 2014). DSF was run in sextuplet 20 μl volumes for each condition in a 384-well qPCR plate (Thermo) on a Life scientific QuantStudio 6 flex real-time PCR machine set to melt-curve mode, with a temperature ramp from 25 °C to 99 °C ramping at 0.17 °C s⁻¹. Data were analyzed using Protein Thermal Shift software v. 1.3.

Kinetic analysis of CARs on ATP and NADPH

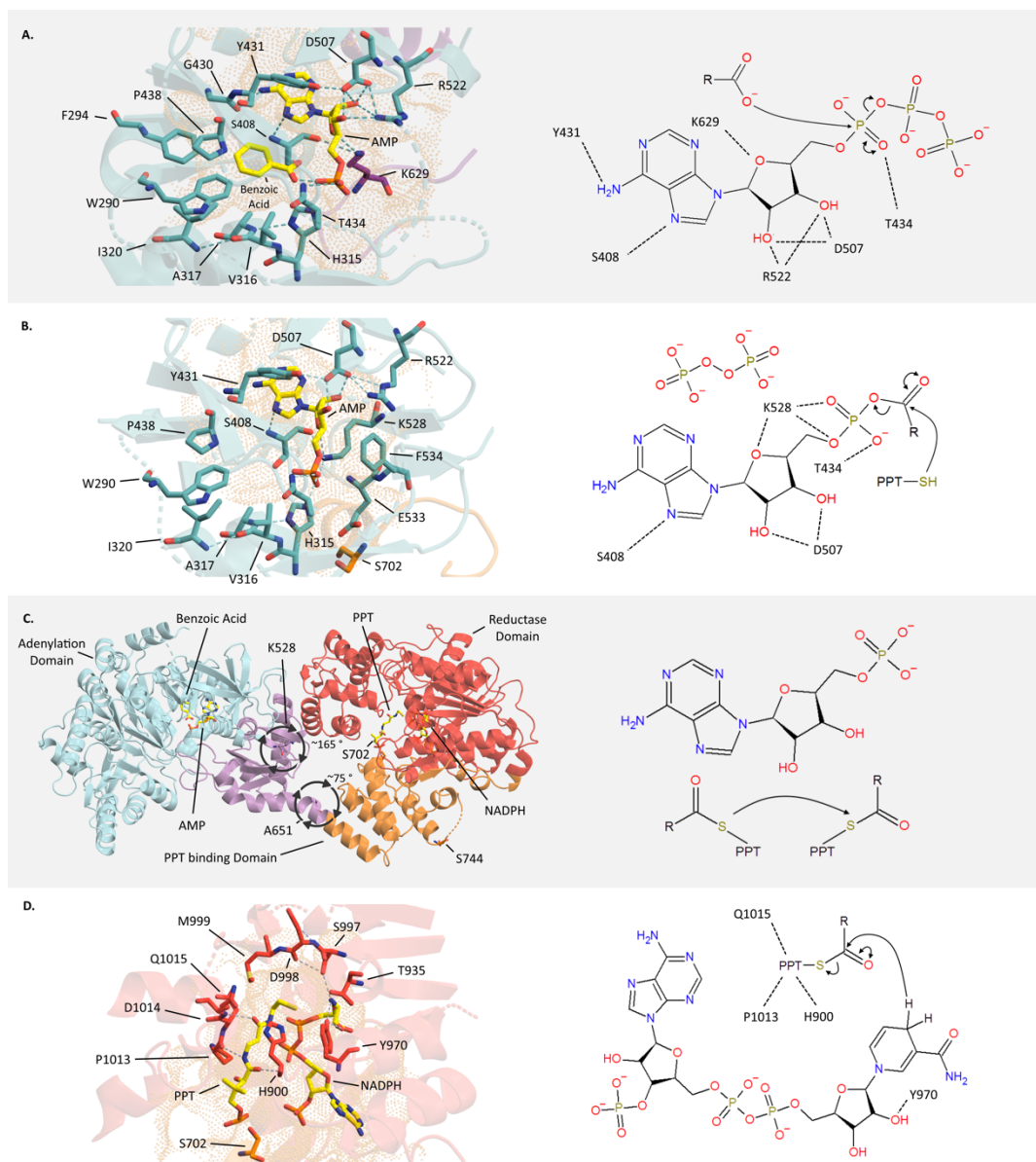
Enzyme kinetics were assessed by measuring activity of each enzyme on (*E*)-3-phenylprop-2-enoic acid in the presence of varying concentrations of ATP or NADPH. Low concentrations of NADPH or ATP caused the reaction to finish quickly meaning concentrations were represented by very few kinetic cycles, potentially skewing results. Data was therefore trimmed of concentrations showing inhibition or high signal to noise ratio. For both ATP and NADPH titrations, a 1.7x dilution series from 8 mM over 12 points was used. Points that exhibited substrate inhibition were removed from analyses. Rates were fitted to the Michaelis-Menten equation by non-linear least squares regression in GraphPad Prism v. 7.

Kinetic analysis of AncCAR substrate range

Carboxylic acids were dissolved to near saturation in assay buffer with 20% DMSO. Substrates were titrated in 1.7x dilutions over 8 points. Rates were measured continuously over 6 minutes in a ThermoFisher MultiSkan GO plate reader in precision mode. Rates were fitted to the Michaelis-Menten model by non-linear least squares regression in GraphPad Prism v7.

2.9 Supplementary Figures

Supplementary figure 1



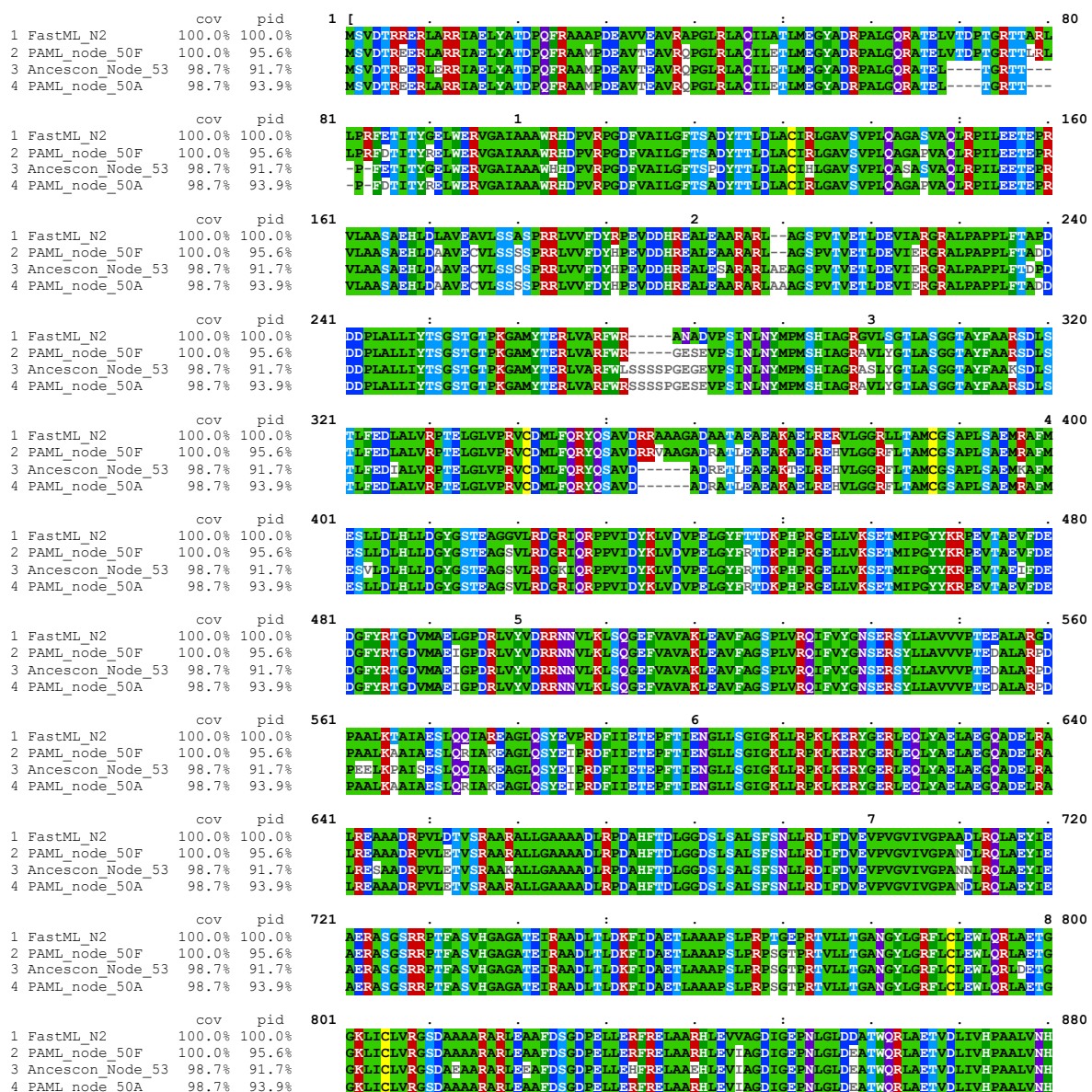
Supplementary figure 1 - Current proposed CAR reaction mechanism

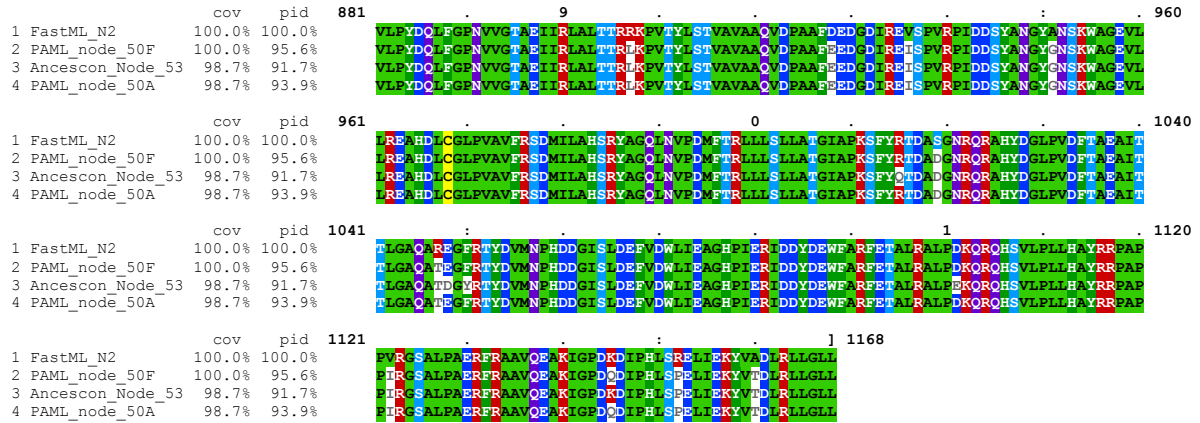
The current model of the reaction mechanism of CARs, based upon partial crystal structures of SrCAR, NiCAR and MmCAR (Gahloth *et al*, 2017): **A)** ATP and carboxylic acid enter the adenylation domain active site leading to the formation of an acyl-AMP intermediate via nucleophilic attack of the carboxylate on the α -phosphate of ATP, releasing pyrophosphate. **B)** The adenylation domain is displaced due to rotation at residues L528 ($\sim 165^\circ$) and A651 ($\sim 75^\circ$), mediating the migration of the phosphopantetheine (PPT) arm ($\sim 50 \text{ \AA}$; covalently bound to S702) into the adenylation domain active site. This movement permits nucleophilic attack by the PPT thiol on the carbonyl of the acyl-AMP intermediate, forming a thioester intermediate. **C)** The reductase domain undergoes

conformational sampling about S744 and reconciles with the relocated PPT thioester intermediate.

D) The thioester bond is reduced by NADPH releasing the aldehyde product and NADP⁺ and regenerating the PPT moiety (Crystal structures were rendered in PyMol v. 2.0; *PDB IDs: 5MSS, 5MST and, 5MSV*).

Supplementary figure 2



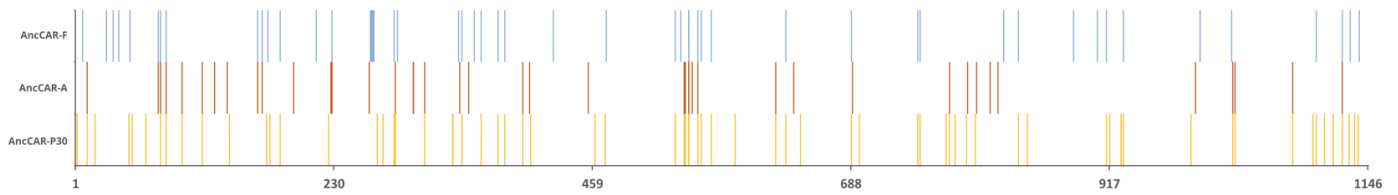


Supplementary figure 2 - Alignment of AncCAR protein sequences

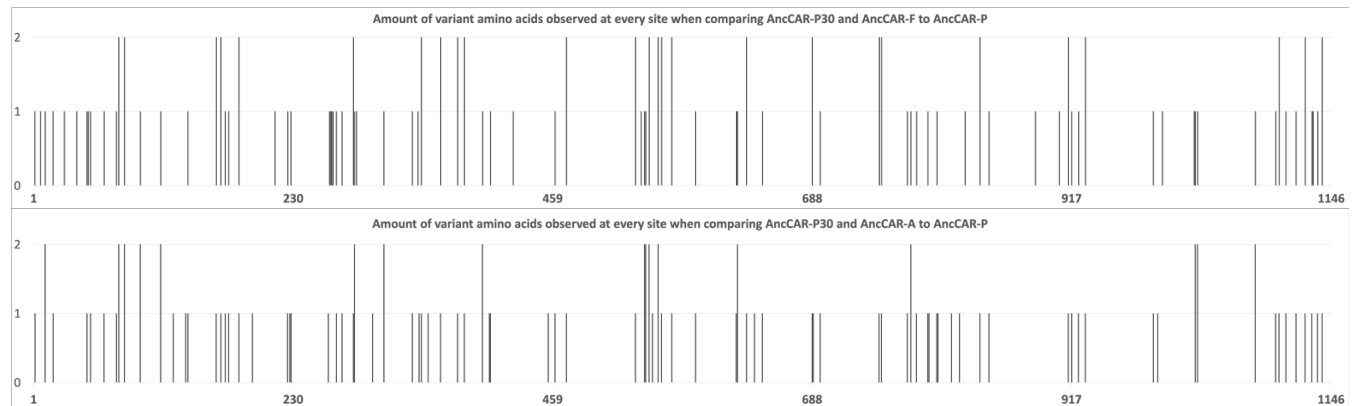
AncCAR sequences were produced with the Ancescon, PAML and FastML algorithms from the tree in figure 13A. The most likely sequence from the posterior probability distribution of the most ancestral node were taken as the ancestral state for each algorithm. Sequences were aligned with MUSCLE in the Geneious v. 10.0 software suite and visualised using MVIEW v.1.63.

Supplementary figure 3

A



B



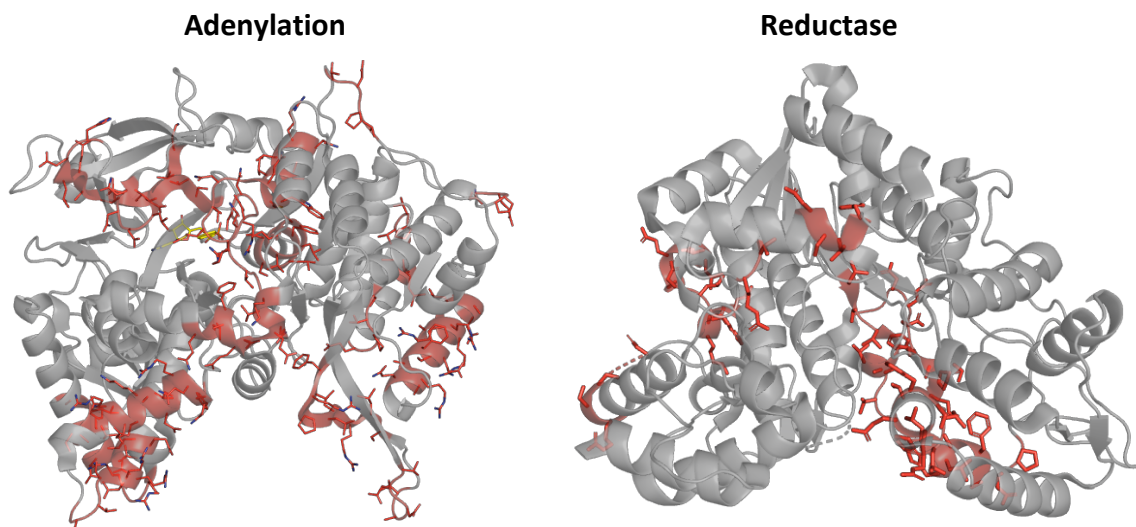
	Number of residues that vary in the dataset compared to AncCAR-P	Number of residues varying at a position in both algorithms compared to AncCAR-P	% residues with shared diversity compared to AncCAR-P identified by both algorithms
AncCAR-P30/ AncCAR-A	107	17	15.9
AncCAR-P30/ AncCAR-F	117	26	22.2

Supplementary figure 3 – Difference between ASR algorithm outputs is not alternative sampling of posterior probability tables.

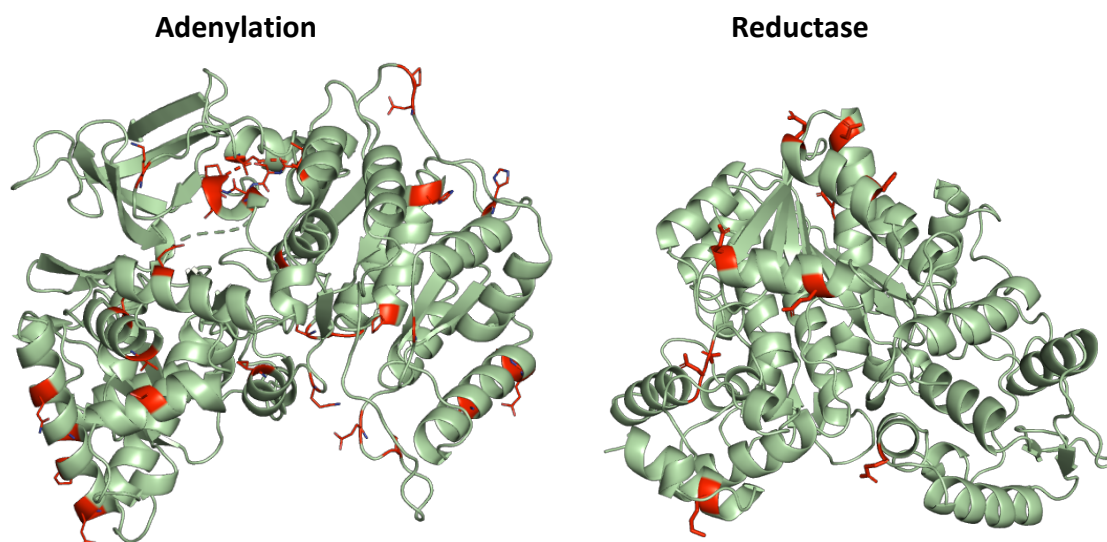
Analysis of algorithm sampling of ancestral space across the table of posterior probabilities. AncCAR-P30 was generated by selecting second or third residues with <30% likelihood from the table of posterior probabilities output by PAML using a bespoke python script. Sequences were aligned with MUSCLE in Geneious v. 10. Sites containing gaps were omitted from the analysis. Variation was identified by eye and transposed into Microsoft Excel. **A)** An identity barcode Comparing residue conservation at every position in AncCAR-A, AncCAR-F and AncCAR-P30 when compared to AncCAR-P. Coloured lines denote a site in AncCAR-F (blue; top row), AncCAR-A (burnt orange; middle row), and AncCAR-P30 (yellow; bottom row) with varying amino acid identity to the reference. **B)** Bar chart displaying instances of variation at each residue when comparing AncCAR-P to either AncCAR-A and AncCAR-P30, or AncCAR-F and AncCAR-P30. Vertical lines denote the number of residues that vary at each position in each comparison dataset (min 0, max 2) compared to AncCAR-P.

Supplementary figure 4

A



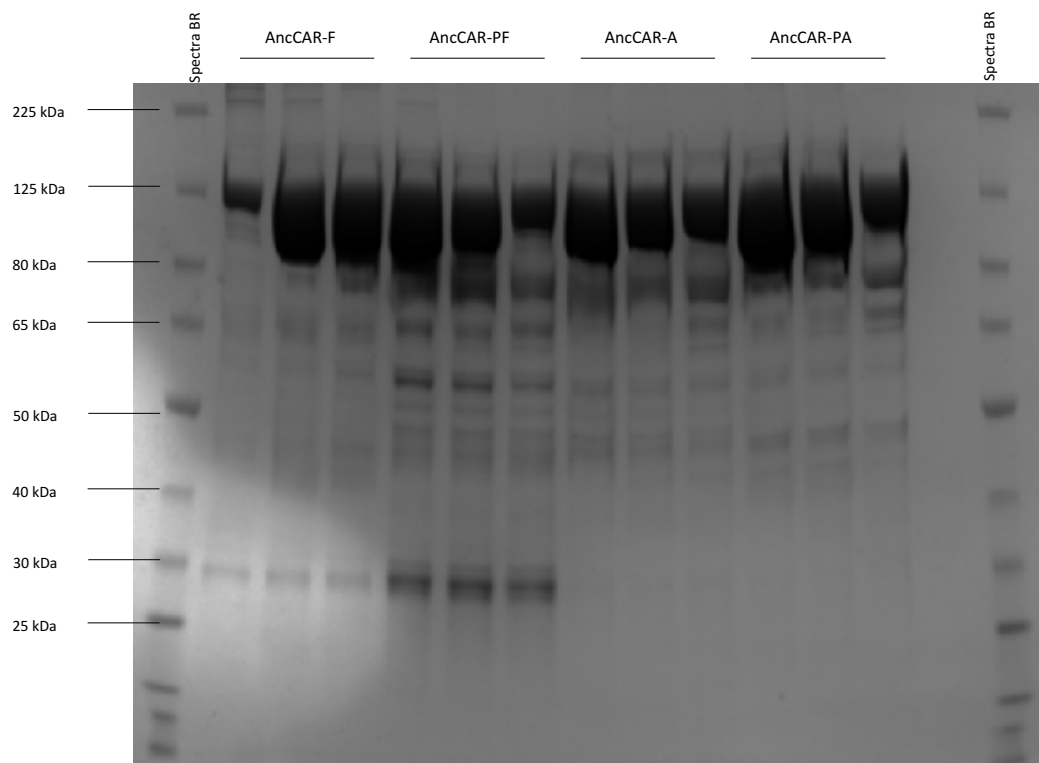
B



Supplementary figure 4 - Regions of significant sequence variation between AncCARs and 5mst/5msp

A) Sequence alignment between AncCARs and sequences of crystal structures of CAR adenylation domain (5mst) and reductase domain (5msp) was made in MUSCLE, within the Geneious software suite. Regions of considerable variation (multiple changes within 5 residues) are highlighted on the WT crystal structures in red. **B)** Sequence alignment between AncCARs was made in MUSCLE within the Geneious software suite. All regions of variation between enzymes (red) were highlighted on the modelled adenylation and reductase domains of AncCAR-A (green). Images were rendered in PyMOL v. 2.0.

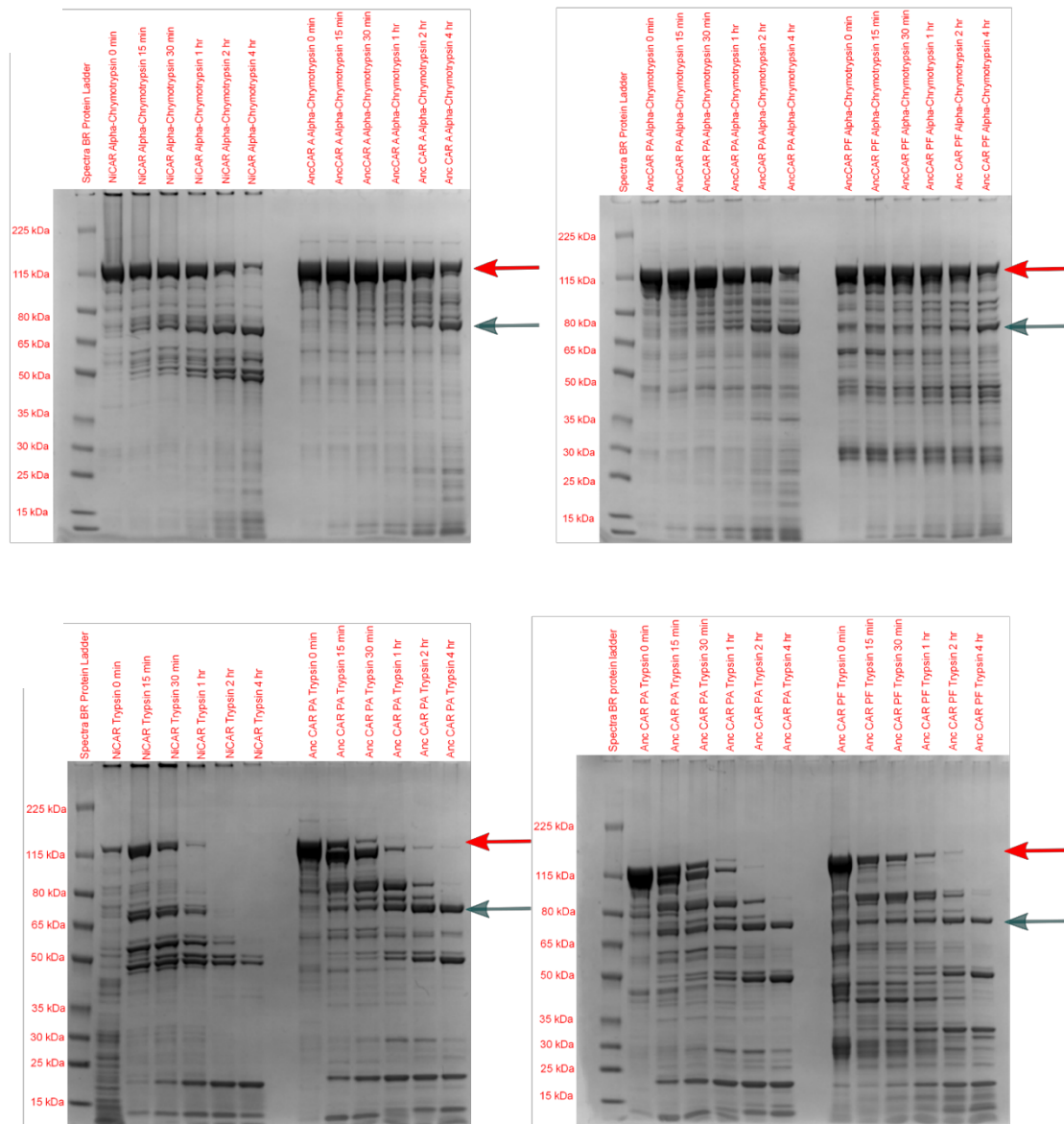
Supplementary figure 5



Supplementary figure 5 - All CAR enzymes are soluble

4-20% acrylamide SDS-PAGE gel of three fractions from the largest peaks following CAR purification by nickel-affinity followed by size exclusion chromatography. All four AncCAR proteins are soluble, producing large volumes of protein (all AncCARs are approximately 128 kDa in size). Typically, per liter bacterial culture, between 3 and 7 mg enzyme could be extracted.

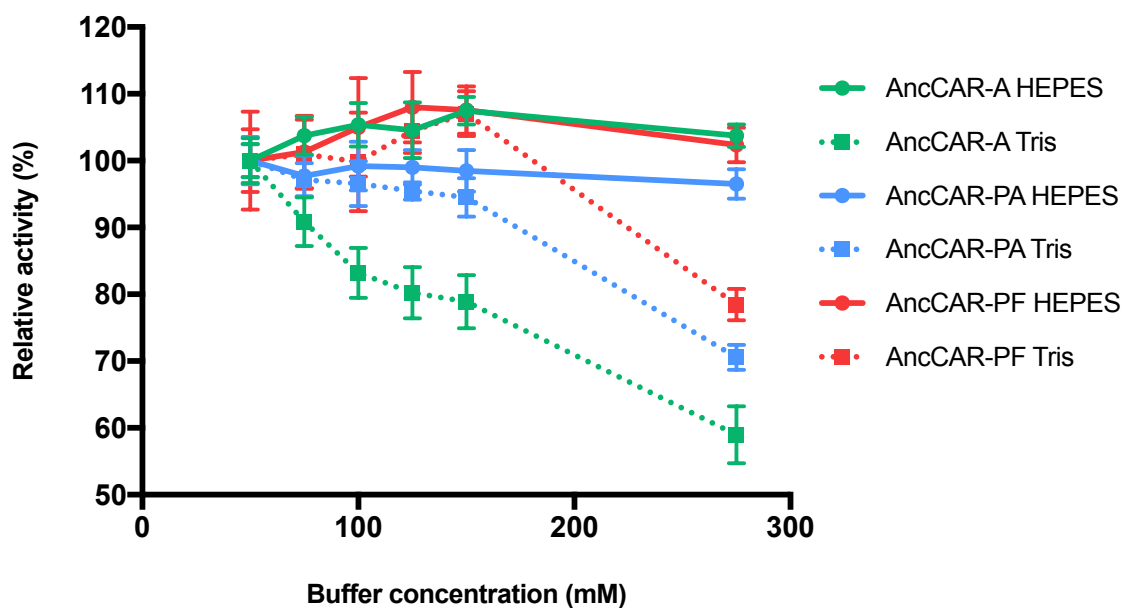
Supplementary figure 6



Supplementary figure 6 - AncCARs are less protease sensitive than ExCARs

Limited proteolysis was performed on the CAR from *N. iowensis*, AncCAR-A, AncCAR-PA and AncCAR-PF. In each case, chymotrypsin (top) or trypsin (lower) was added at 1 $\mu\text{g}/\text{mL}$ to a sample of protein at 1 mg/mL . The proteins were incubated at 37 $^{\circ}\text{C}$, samples taken at various points and quenched by boiling in SDS-PAGE sample buffer. In both cases, the progression of the proteolysis from the whole protein (red arrow) to separate domains (A domain indicated by the teal arrow) proceeds faster for NiCAR than AncCARs. With chymotrypsin, whole NiCAR is almost entirely lost by 4 hr, whilst AncCARs have considerable amounts intact. With trypsin, all proteins have at least one nick by 1 hr; the A domain of NiCAR has also received at least one nick by 4 hr, whilst it is largely intact for the AncCARs. These results show that AncCARs are less sensitive to these proteases than NiCAR.

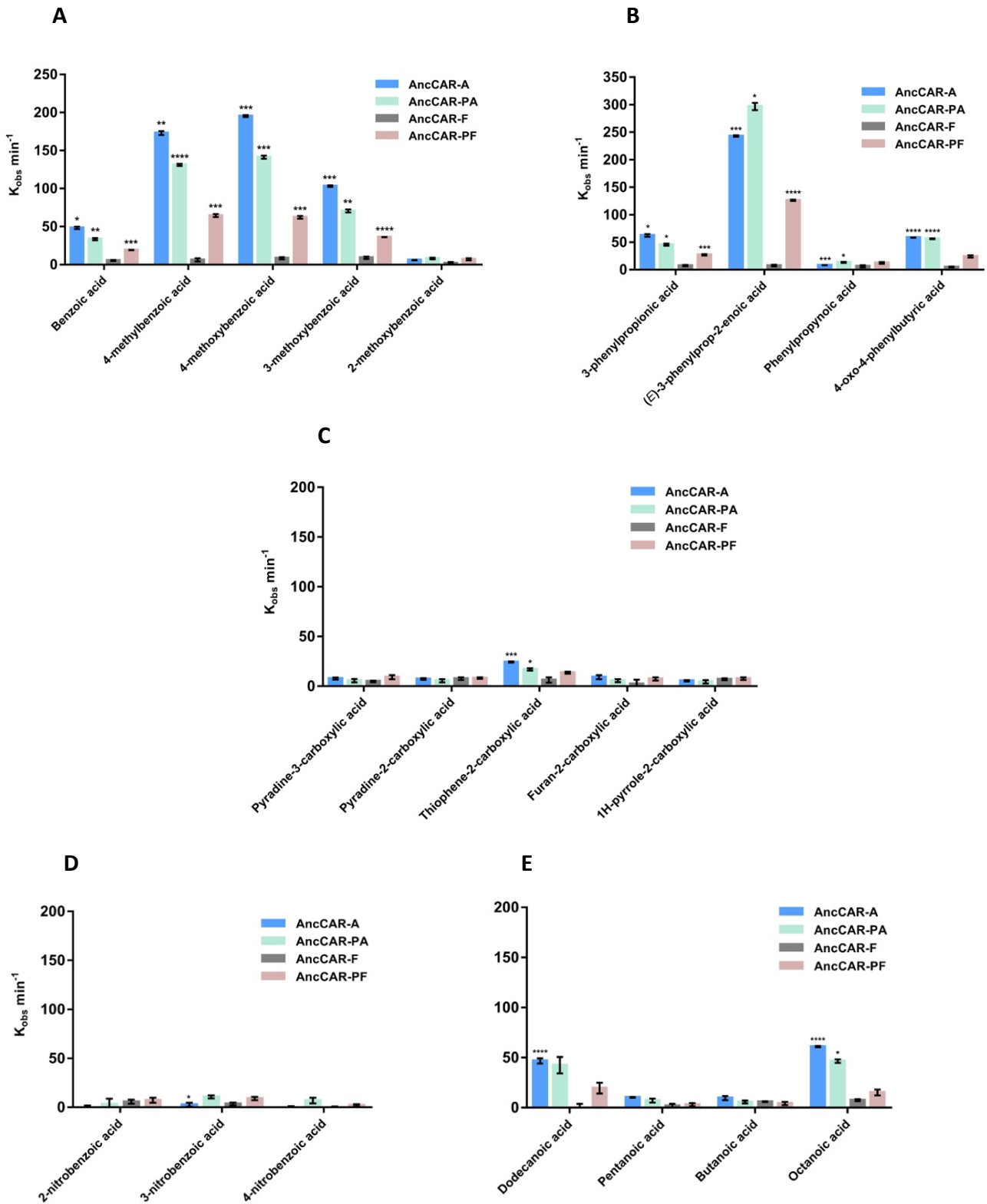
Supplementary figure 7



Supplementary figure 7 - Tris inhibits AncCAR Enzymes

Relative activity of AncCAR-A, AncCAR-PA and AncCAR-PF in the presence of (*E*)-3-phenylprop-2-enoic acid in increasing concentrations of Tris (dotted line) and HEPES to understand inhibitory effects of buffer systems. Data were visualized in Graphpad Prism v. 7.

Supplementary figure 8

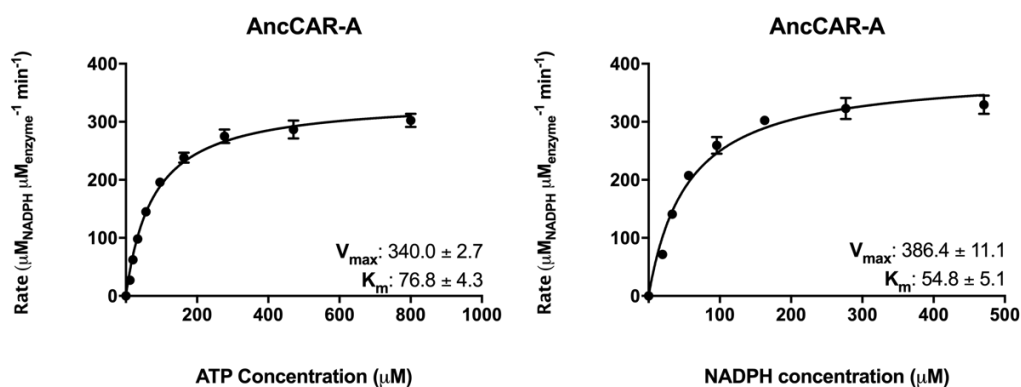


Supplementary figure 8 - AncCARs have equivalent substrate ranges

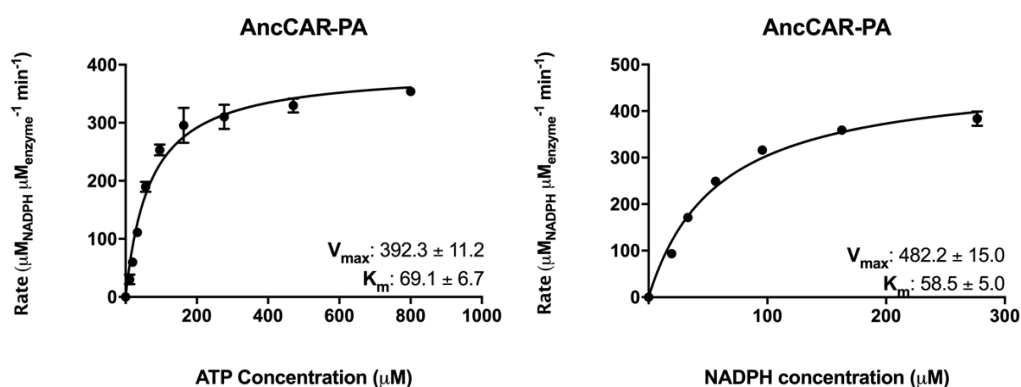
Turnover of NADPH by AncCARs was measured with 24 unique carboxylic acids. Bar charts shows activity on canonical acid substrates at 5 mM. **A)** Benzoic acid its derivatives, **B)** carboxylic acids with a conjugated carboxyl group, **C)** carboxylic acids with substitutions into the aromatic ring, **D)** carboxylic acids with nitro groups, **E)** fatty acids. Each substrate was tested in triplicate, and error bars represent standard error. Asterisks represent degrees of significance from *t*-test of triplicate verses all controls (* = $0.0001 < P \leq 0.001$; ** $0.00001 < P \leq 0.0001$; *** = $0.000001 < P \leq 0.00001$; **** = $P \leq 0.000001$). Data were visualized in Graphpad Prism v.7

Supplementary figure 9

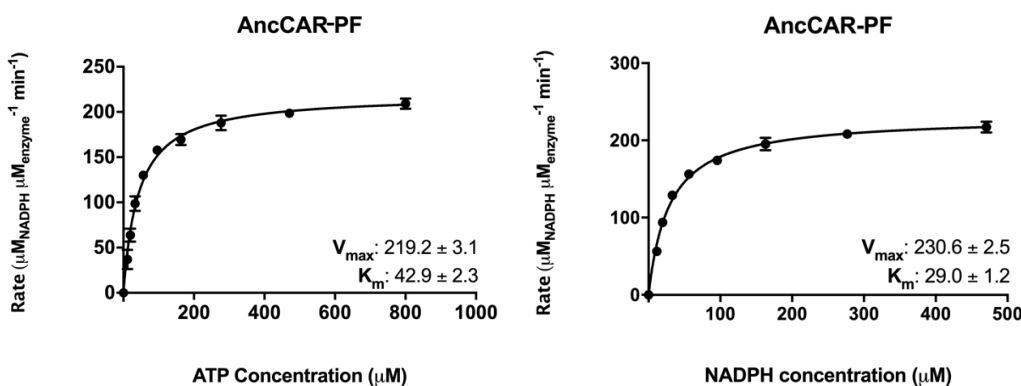
A



B



C



Supplementary figure 9 - AncCAR kinetics in ATP and NADPH

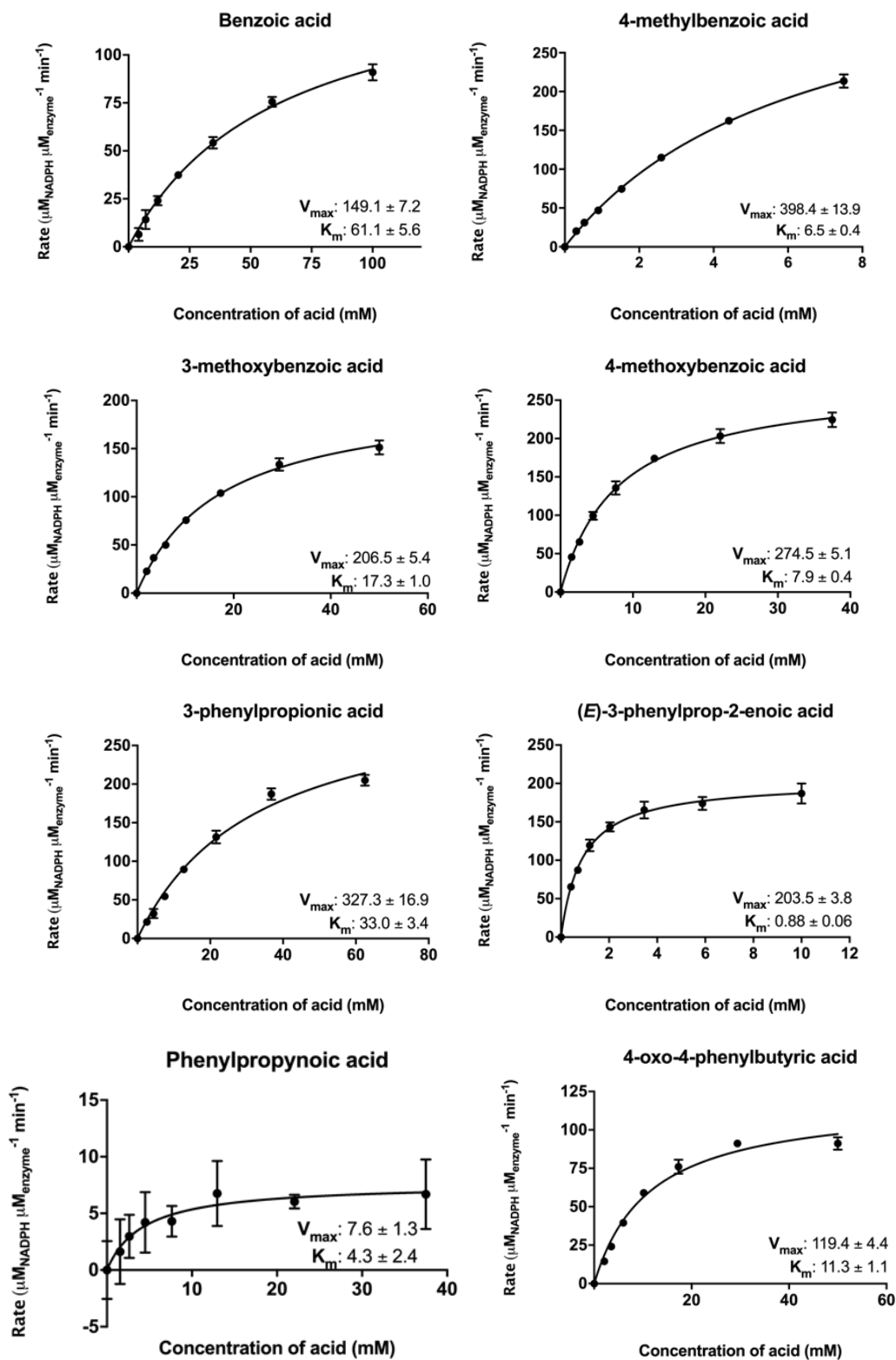
NADPH and ATP kinetics for AncCAR-A (A), AncCAR-PA (B), and AncCAR-PF (C) were obtained from NADPH turnover on (*E*)-3-phenylprop-2-enoic acid. Kinetics were determined using 12 point, 1.7x dilution series of substrate in 200 mM HEPES. Each concentration was investigated in triplicate. Data were fit to the Michaelis-Menten equation in Graphpad Prism v.7. For NADPH kinetics, values obtained for low NADPH concentrations were omitted from the curves as signal dropped below background. Additionally, instances where high concentrations of NADPH showed inhibitory effects on CAR activity were omitted. Results were time-adjusted in 20 second intervals where required.

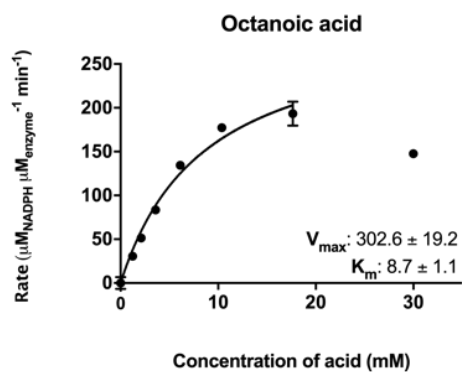
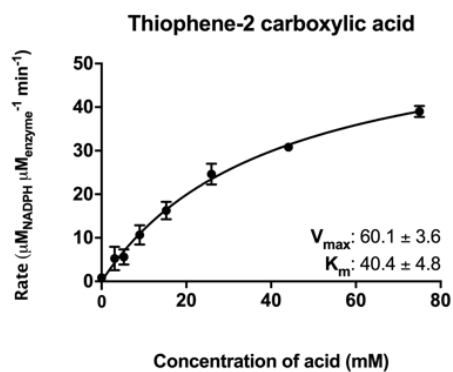
$V_{\text{max}} = k_{\text{cat}}$. Units for V_{max} : $\mu\text{M} \mu\text{M}^{-1} \text{min}^{-1}$. Units for K_M : μM

Supplementary figure 10

A

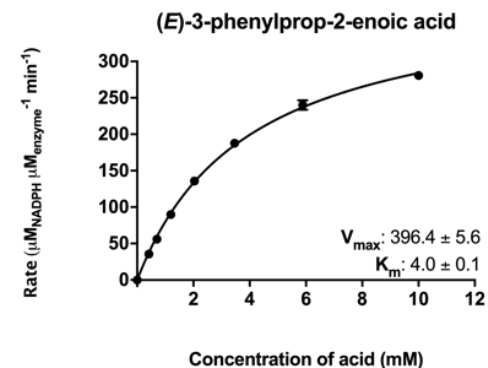
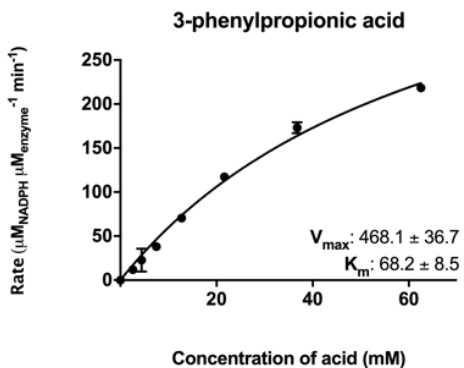
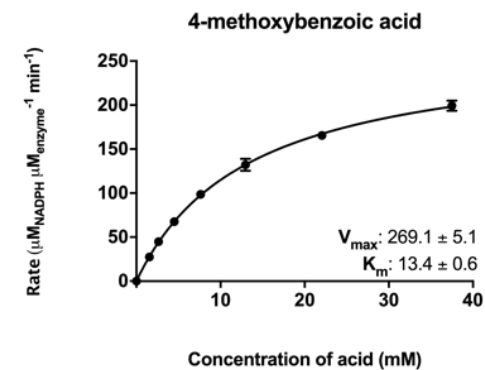
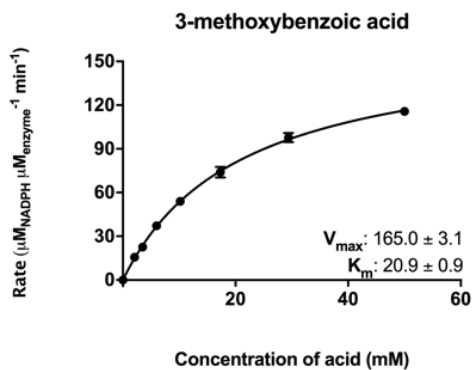
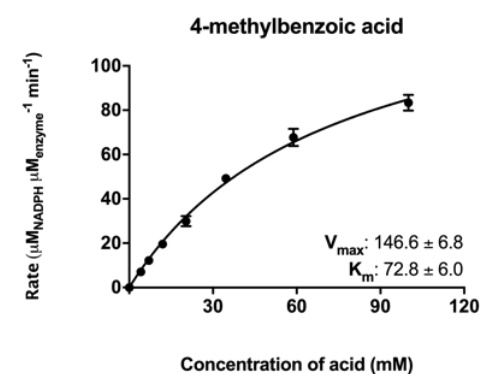
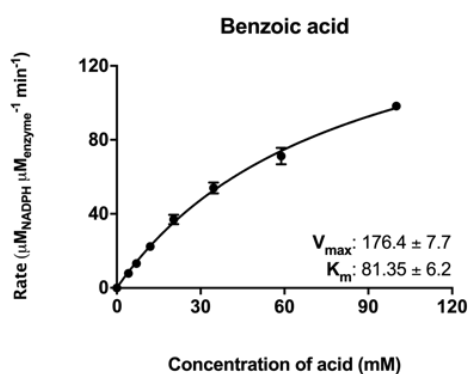
AncCAR-A

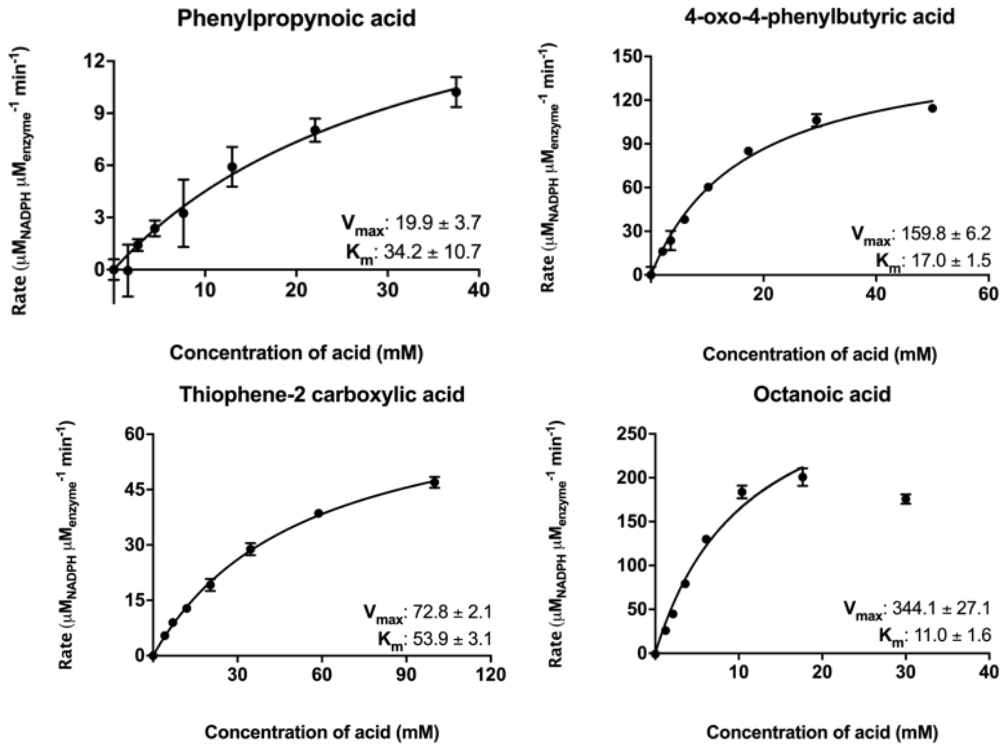




B

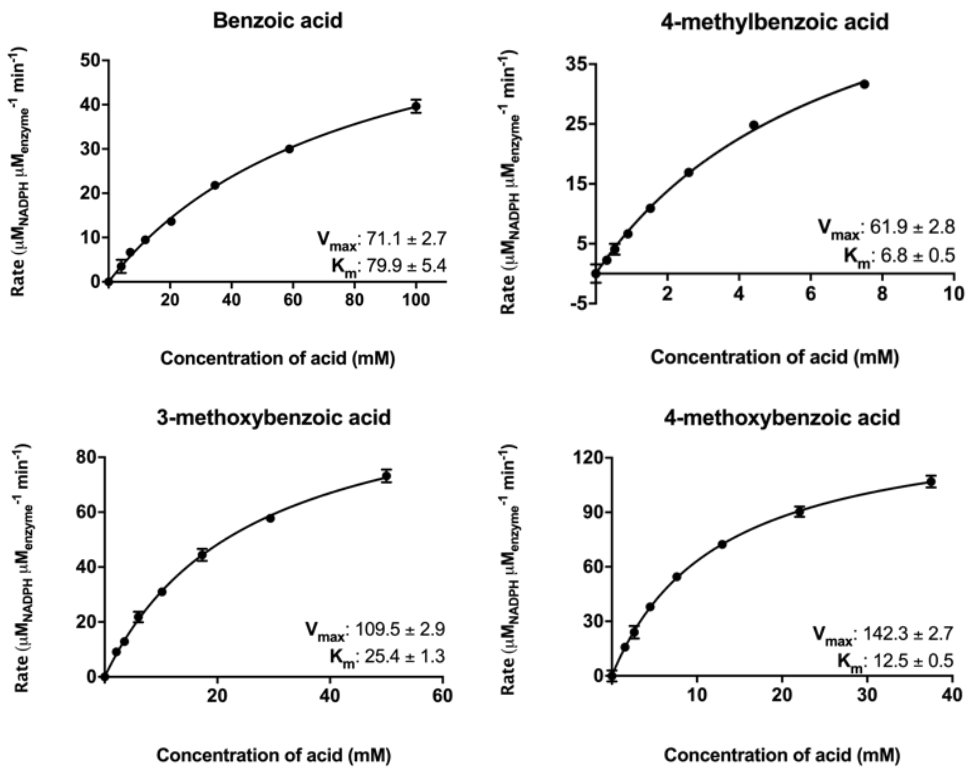
AncCAR-PA

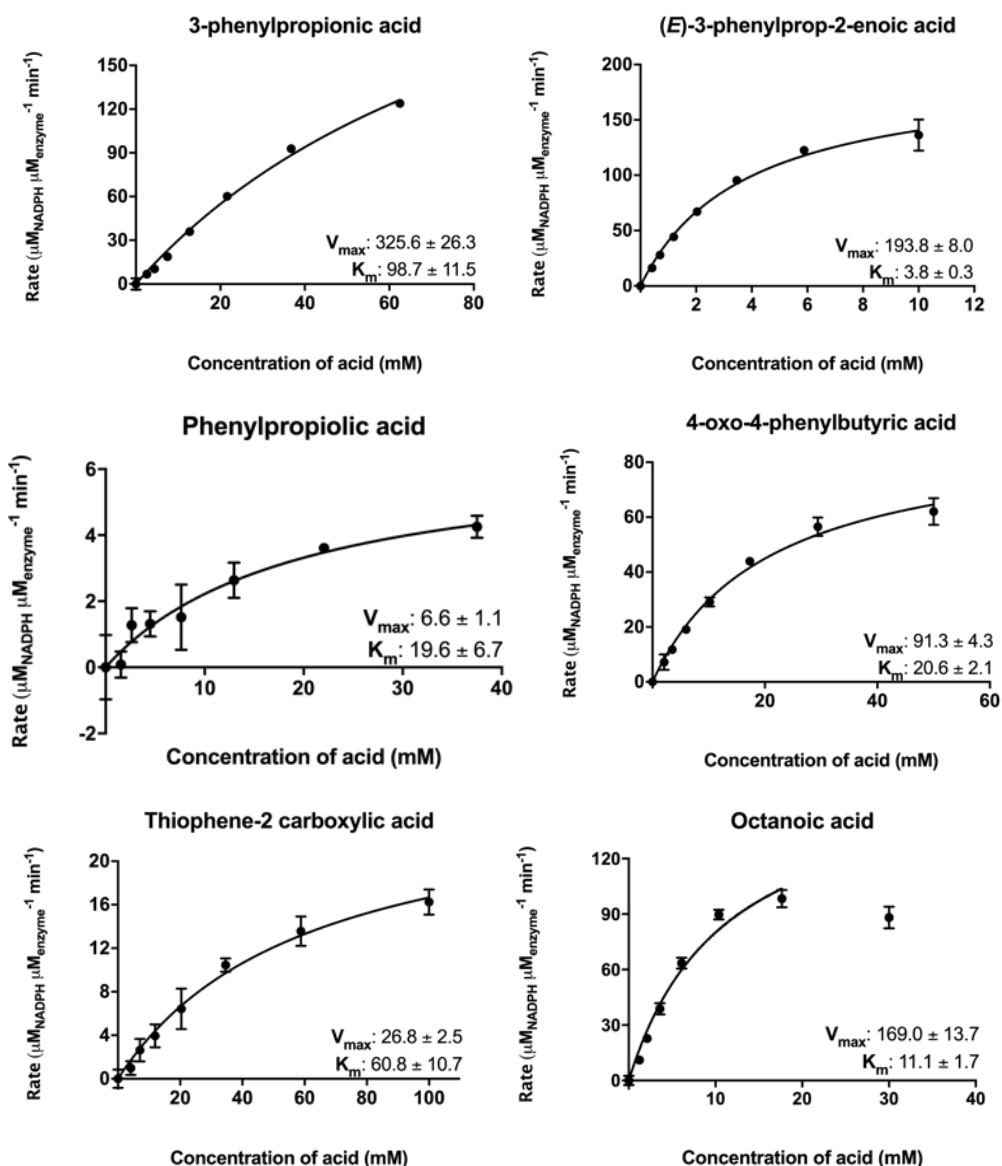




C

AncCAR-PF



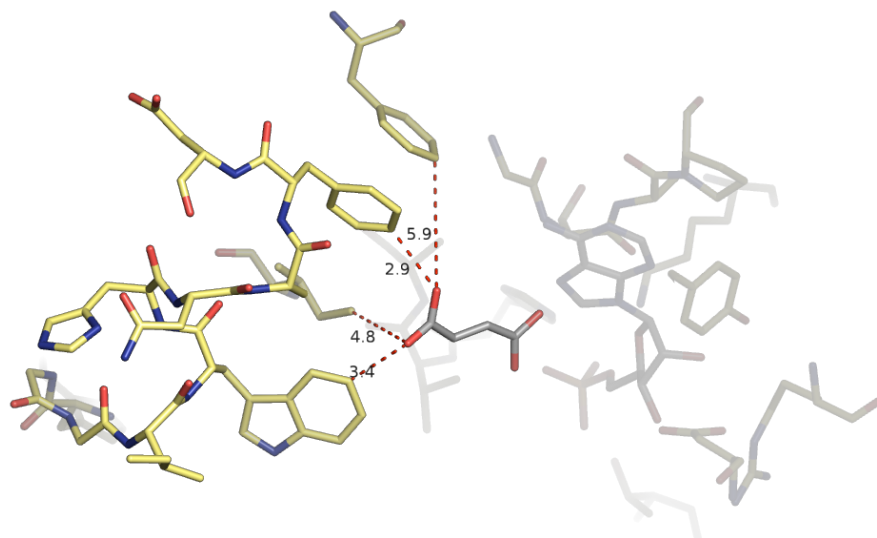


Supplementary figure 10 - AncCAR carboxylic acid substrate kinetics

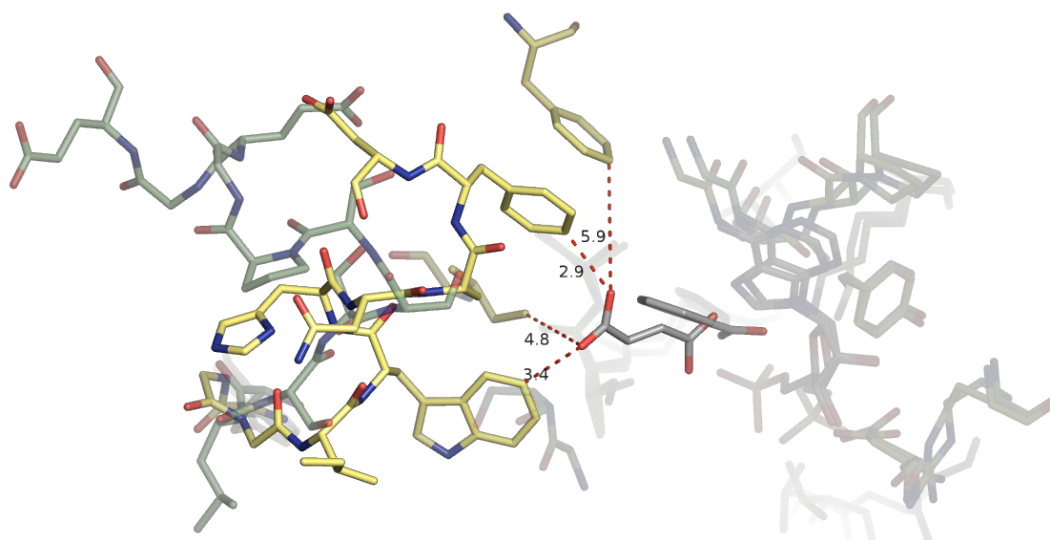
AncCAR kinetics were calculated from NADPH turnover by AncCAR-A (A), AncCAR-PA (B), and AncCAR-PF (C) in the presence of the 10 substrates exhibiting significant activity in supplementary figure 8. 10 μg enzyme were used to improve resolution of 4-methylbenzoic acid and phenylpropionic acid. Kinetics were determined using an 8 point, 1.7x dilution series of acid from near saturation in 200 mM HEPES, with concentrations starting at 800 mM. Each concentration was investigated in triplicate. Data were fit to the Michaelis-Menten equation in Graphpad Prism v. 7. $V_{\text{max}} = k_{\text{cat}}$. Units for V_{max} : $\mu\text{M} \mu\text{M}^{-1} \text{min}^{-1}$. Units for K_{M} : μM .

Supplementary figure 11

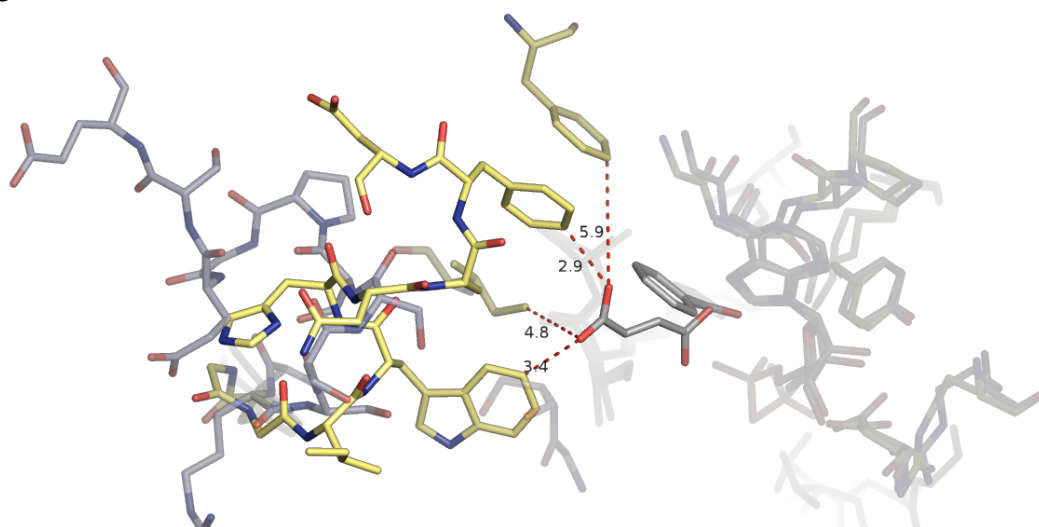
A



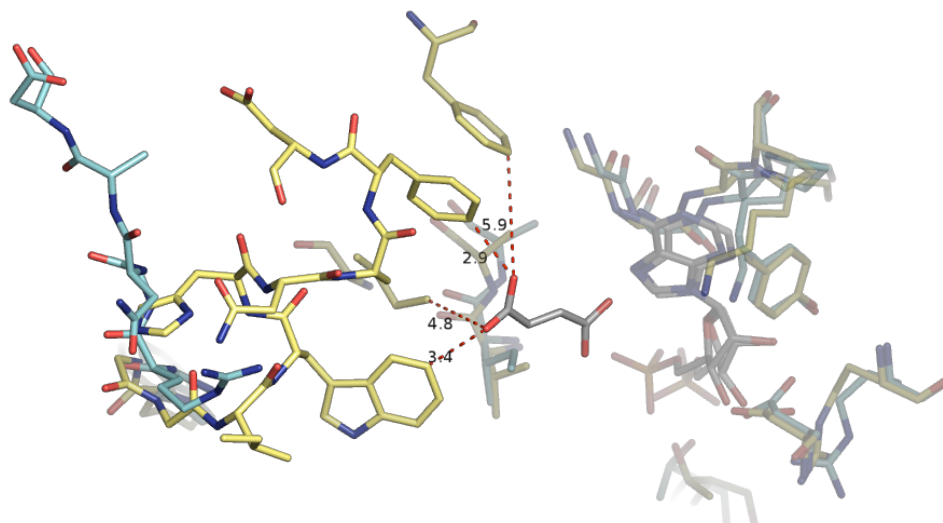
B



C



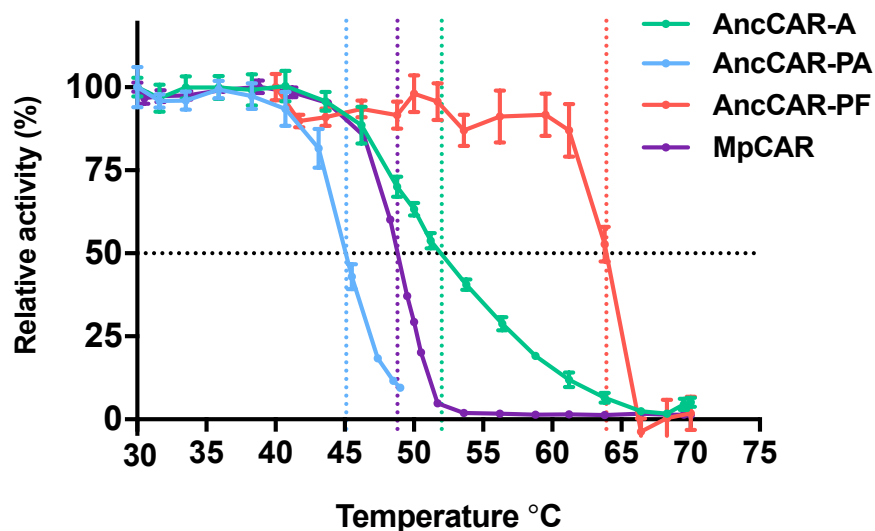
D



Supplementary figure 11 - AncCAR active site comparisons to 5MST

Close-up analysis of models of the highly variable loop in the adenylation domain of AncCARs compared to 5MST (286-302) show that AncCAR-PF (figure 15B) and AncCAR-F do not form potentially stabilizing interactions with the substrate in this region. **A)** Close-up image of 5MST active site. **B)** Modelled structure of AncCAR-A overlaid onto 5MST **C)** Modelled structure of AncCAR-PA overlaid onto 5MST. **D)** Modelled structure of AncCAR-F overlaid onto 5MST. Models were rendered in PyMOL v. 2.0

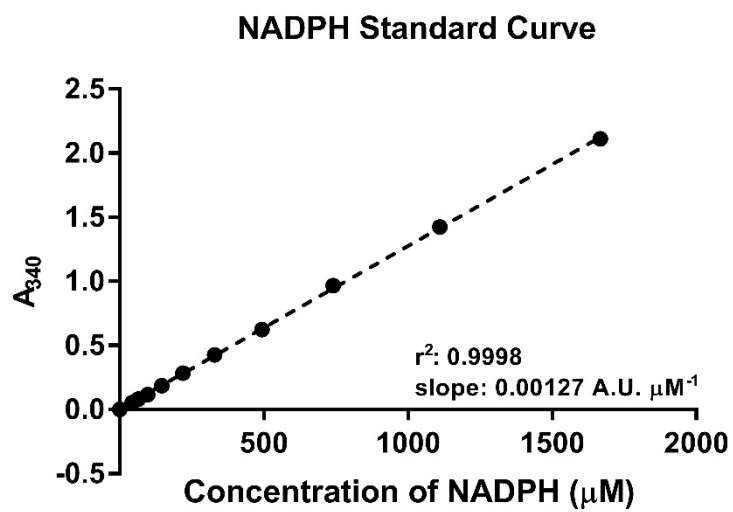
Supplementary figure 12



Supplementary figure 12 - AncCAR stability in NaCl

CARs were incubated at temperatures from 30 °C to 70 °C for 30 min in 10 mM HEPES and 500 mM NaCl. Each point represents the rate of NADPH reduction in 5 mM (*E*)-3-phenylprop-2-enoic acid relative to the rate of NADPH reduction in 5 mM (*E*)-3-phenylprop-2-enoic acid at 30 °C. Each point represents a single triplicate, with error bars representing the standard error. 50% activity was lost at 45 °C, 49 °C 52 °C and 64 °C for AncCAR-PA, MpCAR, AncCAR-A and AncCAR-PF respectively. Data were visualized in Graphpad Prism v.7

Supplementary figure 13



Supplementary figure 13 - NADPH standard curve

An NADPH standard curve was constructed for conversion of raw assay data into substrate turnover. The curve was created by titration of NADPH from 1,700 μM in a 1.5x dilution series in standard reaction buffer in triplicate. Absorbance of the solution was measured at 340 nm. Error bars are occluded by the data points. Data were visualized in Graphpad Prism v.7

Supplementary table 1

	Adenylation		Reductase	
	5MST	5MSD	5MSP	5MSO
AncCAR-A	0.987	0.372	0.898	0.944
AncCAR-PA	1.011	0.444	1.061	1.115
AncCAR-PF	1.276	0.872		
AncCAR-F	1.131	0.747	1.21	0.804
5MST		0.964		
5MSD	0.964			
5MSP				0.701
5MSO			0.701	

t-test A→R $p = 0.13$

Supplementary table 1 - Root mean squared values of alpha carbon atom displacement in AncCAR protein models - showing good fit of data

Supplementary table 2

		Benzoic acid	4-methylbenzoic acid	4-methoxybenzoic acid	3-methoxybenzoic acid	3-phenylpropionic acid
A	k_{cat} (min ⁻¹)	149.1 ± 7.1	398.4 ± 13.9	274.5 ± 5.1	206.5 ± 5.4	327.3 ± 16.9
	K_M (mM)	61.2 ± 5.6	6.5 ± 0.4	7.9 ± 0.4	17.3 ± 1.0	33.0 ± 3.4
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	2.4 ± 0.3	61.3 ± 4.3	34.7 ± 1.9	11.9 ± 0.8	9.9 ± 1.1
PA	k_{cat} (min ⁻¹)	176.4 ± 7.7	146.6 ± 6.8	269.1 ± 5.1	165 ± 3.1	468.1 ± 36.7
	K_M (mM)	81.4 ± 6.2	5.5 ± 0.5	13.4 ± 0.6	20.9 ± 0.9	68.2 ± 8.5
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	2.2 ± 0.2	26.7 ± 2.7	20.1 ± 1.0	7.9 ± 0.4	6.9 ± 1.0
PF	k_{cat} (min ⁻¹)	71.8 ± 2.7	61.9 ± 2.8	142 ± 2.7	109.5 ± 2.9	325.6 ± 26.3
	K_M (mM)	79.9 ± 5.4	7.0 ± 0.5	12.5 ± 0.6	25.4 ± 1.3	98.7 ± 11.5
	k_{cat}/K_M (min ⁻¹ mM ⁻¹)	0.9 ± 0.1	8.8 ± 0.7	11.4 ± 0.6	4.3 ± 0.2	3.3 ± 0.5

		(E)-3-phenylprop-2-enoic acid	Phenylpropionic acid	4-oxo-4-phenylbutyric acid	2-thiophene carboxylic acid	Octanoic acid
A	K_{cat} (min ⁻¹)	203.5 ± 3.8	7.6 ± 1.3	119.4 ± 4.4	60.1 ± 3.6	302.6 ± 19.2
	K_M (mM)	0.9 ± 0.06	4.3 ± 2.4	11.3 ± 1.1	40.4 ± 4.8	8.7 ± 1.1
	K_{cat}/K_M (min ⁻¹ mM ⁻¹)	226.1 ± 15.7	1.8 ± 2.4	10.6 ± 1.1	1.5 ± 0.2	34.8 ± 4.9
PA	K_{cat} (min ⁻¹)	396.4 ± 5.6	19.9 ± 3.7	159.8 ± 6.2	72.8 ± 2.1	344.1 ± 27.1
	K_M (mM)	4.0 ± 0.1	34.2 ± 10.8	17.0 ± 1.5	53.9 ± 3.1	11.0 ± 1.6
	K_{cat}/K_M (min ⁻¹ mM ⁻¹)	99.1 ± 2.8	0.6 ± 0.2	9.4 ± 0.9	1.4 ± 0.1	31.3 ± 5.2
PF	K_{cat} (min ⁻¹)	193.8 ± 8.0	6.6 ± 1.1	91.3 ± 4.3	26.8 ± 2.5	169.0 ± 13.7
	K_M (mM)	3.8 ± 0.3	19.6 ± 6.7	20.6 ± 2.1	60.8 ± 10.7	11.1 ± 1.7
	K_{cat}/K_M (min ⁻¹ mM ⁻¹)	51.0 ± 4.5	0.34 ± 0.1	4.4 ± 0.5	0.4 ± 0.1	15.2 ± 4.2

Supplementary table 2 - AncCAR kinetics on carboxylic acid substrates

Supplementary table 3

Solvent		A	PA	PF	St. Dev.	
Polar aprotic	Acetone	A ₅₀	14.3	20.4	17.9	3.1
		S ₁₀	64.4	76.3	68.8	6.0
	Acetonitrile	A ₅₀	11.8	16.7	20.2	4.2
		S ₁₀	57.3	78.3	93.4	18.1
	DMSO	A₅₀	23.9	23.7	24.3	0.3
		S₁₀	88.9	86.5	92.4	3.0
Polar protic	Ethanol	A ₅₀	10.0	14.9	17.7	3.9
		S ₁₀	50.1	66.5	72.6	11.6
	Isopropanol	A ₅₀	7.3	11.7	11.8	2.6
		S ₁₀	36.3	56.8	56.6	11.8
	Methanol	A₅₀	14.1	11.7	25.6	7.5
		S₁₀	63.3	56.8	78.6	11.2

Supplementary table 3 - AncCAR tolerance to various solvents

AncCARs activity on 5 mM (*E*)-3-phenylprop-2-enoic acid was assessed in the presence of aprotic and protic solvents by solvent titration from 25% (v/v). Inhibition curves were fit to a second order polynomial in GraphPad Prism v7, from which A₅₀ (% solvent at which 50% of activity is lost) and S₁₀ (%activity at 10% solvent relative to 0% solvent) were calculated. Emboldened text represent data for figures 17A and 17B.

3.10 Supporting information

Acknowledgements

We would like to thank the Biotechnology and Biological Research Council's South West Doctoral Training Program for funding and supporting this research. For their support with research methods, we would like to thank Professor Eric Gaucher of Georgia Technological University, and Jennifer Farrar of his research group, for their support and advice with tree building and reconstruction methods. We would finally like to thank Dr. Thomas Richards of Exeter University for support with the construction of this article, and Alice Cross, and Sumita Roy of the Harmer group for their daily support.

Author Information

The authors declare no competing interests.

Chapter 3

Survivor bias drives overestimation of stability in ancestral proteins

3.1 Authors

Adam Thomas^{1,2*}, Benjamin D. Evans^{1,3*}, Mark van der Giezen², Nicholas J. Harmer^{1,2}.

* Equal contribution.

1. Living Systems Institute, Stocker Road, Exeter EX4 4QD, U.K.
2. Department of Biosciences, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, U.K.
3. Centre for Biomedical Modelling and Analysis, Stocker Road, Exeter EX4 4QD, U.K.

3.2 Preface

This chapter consists a reformatted manuscript for an article written for submission to eLife. In this chapter, a tool is built in pure python to model changes in stability as protein populations evolves. Source code for this model is not provided in this thesis. However, the tool is open access, and source code is available at <https://github.com/bdevans/PESST>.

AT and NH conceived the hypothesis tested in this study. AT wrote the initial algorithm underlying the tool developed in this article to test the hypotheses, with input from NH. BDE wrote the public release version of the algorithm from which figures in this chapter are derived. AT wrote the article. BDE partially wrote methods, and designed the figures output from the model. All authors edited the article.

3.3 Abstract

Ancestral sequence reconstruction (ASR) has extensively probed the evolutionary history of life. Many ancestral sequences are thermostable, supporting the “hot-start” hypothesis for life’s origin. A number of recent studies have observed thermostable ancient proteins that evolved in moderate temperatures. Recent research has ascribed these effects to “consensus bias”. However, this conflicts with explanations of thermostability in consensus sequences. Here, we propose a hypothesis of “survivor bias” as an alternative rationalisation for ancestral protein stability in alignment-based methods. We propose four tenets that describe how a protein’s pressure to evolve at marginality will titrate significantly destabilizing residues from the population. Consequently, when deriving ancestral or consensus sequences from surviving sequences evolved under marginality, residues are selected from a dataset biased towards neutralizing or stabilizing mutations. We thoroughly explore the presence of marginality bias using a highly parameterizable *in silico* model of protein evolution that tracks stability at the population, protein and amino acid levels. We show that ancestors and consensus sequences derived from populations evolved at marginality throughout their history are significantly biased toward thermostability. The mechanisms underlying the stabilization of ancestral and consensus proteins have been uncovered, providing caveats for the thorough derivation of conclusions from future ASR work.

3.4 Introduction

Ancestral sequence reconstruction (ASR) allows researchers to trace changing protein sequences across evolutionary time (Merkl and Sterner, 2016; Akanuma, 2017). Recently, ASR has been used to elucidate details about the evolution of several biochemical traits. Activity-centric properties appraised include the evolution of substrate discrimination, specificity and plasticity (Wheeler *et al.*, 2018; Pawlowski *et al.*, 2018; Babkova *et al.*, 2017), trends in thermodynamic properties early in evolutionary time (Gaucher *et al.*, 2008; Hobbs *et al.*, 2012; Akanuma *et al.*, 2013; Butzin *et al.*, 2013; Hart *et al.*, 2014; Risso *et al.*, 2015; Akanuma, 2017; Okafor *et al.*, 2018), and the functional space within the alternative evolutionary histories of protein families (Starr *et al.*, 2017; Cole *et al.*, 2013). Additionally,

analyses have focused on structural characteristics of polypeptides, including the evolution of tertiary and quaternary protein structures (Lim and Marqusee, 2017, Prinston *et al.*, 2017; Hochberg and Thornton, 2017; Finnigan *et al.*, 2011), the evolution of complexity in multimeric proteins (Finnigan *et al.*, 2012), and the evolution of viral capsid intermediate structures (Gullberg *et al.*, 2010; Zinn *et al.*, 2015).

Thermostability has been a subject of particular focus in ASR studies. ASR experiments that probe the most ancient of sequences from protein families conserved across all kingdoms of life have consistently produced thermostable molecules (e.g. EF-Tu; Gaucher *et al.*, 2008; Butzin *et al.*, 2013; Hart *et al.*, 2014; Okafor *et al.*, 2018). Tracing the evolution of such proteins across their lineages has demonstrated a trend from high to low thermal stability from ancient life to modern life. These trends correlate well with estimated historical terrestrial temperatures (Gaucher *et al.*, 2008). Many studies have concluded that early organisms inhabited a warmer Earth (Gaucher *et al.*, 2008; Akamuna *et al.*, 2013; Butzin *et al.*, 2013; Hart *et al.*, 2014). This required proteins to fold and function under high temperatures. However, recent studies have uncovered thermostable proteins from lineages that are unlikely to have encountered high environmental temperatures in their evolutionary life history (Gumulya *et al.*, 2018; Trudeau *et al.*, 2016; Chapter 2). This suggests that not all thermostable ancestors are derived from the same conditions.

Illustrating this, we recently reconstructed ancient carboxylic acid reductases from the *Mycobacterium* and *Nocardia* that exhibited up to 35 °C increases in stability over their extant counterparts (Chapter 2). The stability of this ancestor does not fit into the trends laid out established by other paleotemperature studies (Gaucher *et al.*, 2008; Akamuna *et al.*, 2013; Butzin *et al.*, 2013; Hart *et al.*, 2014). There is no prevailing evidence that any ancestors of *Nocardia* and *Mycobacterium* were thermophiles. Evidence suggests this family evolved somewhere in the late Phanerozoic eon (<500 myo), when the earth was warming from a colder “snowball earth” state (Lewin *et al.*, 2016; Harland, 1964). Trudeau *et al.*, 2016 reported a similar pattern with the serum paraoxonases (PON), whose ancestor was found to be up to 30 °C more temperature resilient than their modern-day counterparts. Ancient PONs also exhibited superior folding properties when expressed in *E. coli* (Trudeau *et al.*, 2016). Furthermore, similar increases in stability were achieved by Gumulya *et al.*,

2018 in the reconstruction of CYP3 cytochrome P450 mono-oxygenases. Both PONs and CYP3 are post-Cambrian innovations of Mammalia and Vertebrata respectively. There exists no evidence that any mammalian or vertebrate ancestor thermoregulated at the temperatures suggested by PON and CYP3 ancestor stabilities (Mackness and Mackness, 2015).

In an effort to explain such stabilizing effects, Gumulya *et al.*, 2018 posited that vertebrate ancestors of CYP3 evolved in a warmer ocean environment, whose proteins subsequently approached mesophily by drift within recent evolutionary timescales. In contrast, Trudeau *et al.*, 2016 hypothesised that the consensus bias exhibited by ancestral PONs drove their stabilization. At highly divergent sites, the phylogenetic signal describing the site's history is often lost or obscured. In this case ASR algorithms have a propensity to predict the consensus sequence at these sites. ASR algorithms therefore display an inherent bias toward the consensus sequence across the protein, present in the PONs. Gumulya *et al.* (2018) do not comment on the bias toward consensus in ancestral CYP3 enzymes. However, we found that ancestral CARs also exhibit bias towards the consensus sequence (Chapter 2). Consensus sequences are a proven sequence-driven method to engineer stabilizing properties into enzyme families, hence the method has a stabilizing effect on ancestral proteins (Sternke *et al.*, 2018; Okafor *et al.*, 2018; Durani and Magliery, 2013; Kiss *et al.*, 2009). Current explanations for the thermostable properties of consensus sequences assume that common amino acids at a position contribute to thermodynamic fitness more than other possible amino acids at that position (Sternke *et al.*, 2018; Porebski and Buckle, 2016; Ye *et al.*, 2017). This suggests that the stabilizing effect arises as the consensus residues are representative of the sum of stabilizing mutations from some stable ancestor (Porebski and Buckle, 2016).

The proposed origins of stability in ancestral proteins apparently evolved from a mesophilic ancestor are therefore counterintuitive, incomplete and insufficient for describing the underlying forces driving stabilization. It cannot be excluded that recent proteins evolved in warmer environments (Gumulya *et al.*, 2018). Nevertheless, this explanation becomes less parsimonious than a ASR derived biasing effect with every discovery of a new stable ancestor from mesophilic origins. We therefore explored an alternative hypothesis that

there exists a “survivor bias” that explains the stabilization of both consensus and ancestral sequences in the absence of a stable ancestor. Briefly, the survivor bias hypothesis (Box 1) states that natural proteins incur a considerable fitness cost if their maximum folding temperature is below that of their immediate environment. As present-day proteins typically display marginal stability, significantly destabilizing mutations are selected against, and are therefore underrepresented in extant protein datasets. This effect over-represents stabilizing residues that are then selected in both consensus and ASR derived sequences (described in detail later).

To test the marginality bias hypothesis, we have developed a freely available Python-based *in silico* model of sequence evolution called “PESST” (Protein Evolution Simulations with Stability Tracking). PESST evolves a population of protein sequences according to an accepted model of amino acid replacement, and tracks the changing stability of these sequences defined at the amino acid level. PESST was designed as a sequence evolver that follows standard amino acid evolution, generates phylogenies *de novo*, and focuses on the integration of environmental constraints on the evolving population fitness. By observing the outcomes of simulated evolution, we identified that simultaneous effects from both the destabilizing force of drift and the stabilizing force of a stability threshold are driving bias in ASR. Generally, under such bidirectional pressure, the most ancient nodes were more stable than contemporary nodes. There is a significant correlation of stability with node age. The simulated populations produced consensus proteins that were significantly stabilized. PESST provides a toolbox to test evolutionary hypotheses, and provides strong support that marginality bias underlies protein stabilization in sequence alignment-driven protein engineering tools. Furthermore, these data suggest that ASR is a powerful engineering tool for the biasing of sequences towards stability irrespective of a protein’s evolutionary history.

3.5 Hypotheses – tenets of survivor bias

Four tenets (Box 1) provide the theoretical basis for the survivor bias hypothesis. Tenets 1 and 2 are derived from the literature, and Tenets 3 and 4 are proposed as a logical conclusion of these.

Box 1: Tenets of the Survivor Bias Hypothesis

Tenet 1: A mutation's contribution to protein stability is derived from a normal distribution with a negative (destabilizing) mean.

Tenet 2: The majority of proteins are marginally stable.

Tenet 3: Contemporary proteins contain fewer significantly destabilizing amino acids than the global distribution of possible mutations.

Tenet 4: The sequence space from which ancestral proteins are derived is positively biased for stabilizing mutations.

3.5.1 Tenet 1: A mutation's contribution to protein stability is derived from a normal distribution with a negative mean

In this instance, protein stability is defined as the temperature at which a protein unfolds to lose its native function. For any given protein, its sequence space contains considerably more amino acids that confer a negative change to stability than amino acids that confer a positive change (Taverna and Goldstein, 2002). Surface residues generally show a tight, effectively neutral distribution; whilst core residues show a wide, generally Gaussian distribution with a strictly negative mean (Tokuriki *et al.*, 2007; Faure and Koonin, 2015). The mean change to stability for all mutations is approximately -5 °C (Pucci and Rومان, 2016).

3.5.2 Tenet 2: The majority of proteins are marginally stable

For most globular proteins, the native thermodynamic state exists close to a threshold between a folded and unfolded state (Goldstein, 2011; Bershtein *et al.*, 2006; Williams *et al.*, 2007) that tracks environmental temperatures. Substantially more stable proteins are possible. However, due to the “neutral ratchet”, protein thermal stability tends towards a stability threshold around its environmental temperature, at which there is a selective

pressure to maintain “marginality” (Williams *et al.*, 2007; Harms and Thornton, 2013; Khersonsky *et al.*, 2018). Following Tenet 1, the majority of attempted mutations are destabilizing (Goldstein *et al.*, 2011), and sequences with increased thermodynamic stability become less common with increasing stability (Taverna and Goldstein, 2002; Williams *et al.*, 2006). In a simplified model, protein functionality is directly linked to an organism fitness. Once marginality is reached, accruing additional destabilizing mutations incurs a considerable fitness cost as the protein fails to fold at ambient temperature (Bershtein *et al.*, 2006; Tokuriki and Tawfik, 2009A). Tenet 2 dictates that the majority of proteins exist around marginal stability due to the counteracting forces of mutations on average reducing stability, and selective pressure to maintain folding at ambient temperature. Proteins that do not maintain marginality are more likely to be lost from a population.

3.5.3 Tenet 3: Contemporary proteins contain fewer significantly destabilizing amino acids

From Tenets 1 and 2, it follows that to maintain stability at a marginal threshold of folded and unfolded states, contemporary proteins cannot accept significantly destabilizing amino acids due to their immediate and extensive fitness cost (Bloom *et al.*, 2006). Therefore, destabilizing amino acids only become fixed in the population by provision of beneficial function (e.g. serine proteases; Hedstrom, 2002; Kramer *et al.*, 2014). Due to the effects explained in Tenets 1 and 2, such destabilizing effects will be mitigated by incorporation of stabilizing mutations elsewhere in the sequence. The maintenance of marginality (Tenet 2) will deliver a selective pressure against destabilizing residues at non-catalytic sites. Destabilizing residues will be under-populated datasets evolving at marginality.

3.5.4 Tenet 4: Ancestors sample from a stabilizing mutation space, despite a destabilizing global mutational landscape

Tenet 3 dictates that the sequence landscape of extant proteins will under-populate significantly destabilizing residues despite being derived from a densely destabilizing sequence space (Tenet 1). Conversely, extant sequence space will have a greater population of neutral or stabilizing residues than would be expected by chance, allowing proteins to survive at a marginal stability (Tenet 2). Both ASR and consensus protein design critically depend upon the sampling space provided within alignments of extant proteins. As a result,

despite the overall sequence space being dense with potentially destabilizing mutations, the sequence space sampled by the algorithms significantly over-samples stabilizing and neutral residues. It follows that the generation of stable ancestral and consensus proteins can occur even when no stable ancestor is predicted to have existed (i.e. Chapter 2; Gumulya *et al.*, 2018; Lewin *et al.*, 2016).

3.6 Methods

Model description

PESST (Algorithm 1) simulates a fixed population of N proteins, (Φ) , evolving over G generations. Each protein, η , is of fixed length, R , with a defined proportion of invariant sites, $p_{invariant}$. Each protein has an associated thermal stability (denoted T), defined as: $T = \sum_{r=1}^R \Delta_{r,a}$, where $\Delta_{r,a}$ is the change in thermal stability conferred by amino acid a at location r (supplementary figure 14). The global set of $\Delta_{r,a}$ stability changes are randomly drawn from a Gaussian distribution of defined mean, variance and optionally skew, $\sim \mathcal{N}(\mu, \sigma^2, skew)$, in accordance with *Tenet 1*. As it is understood that for the majority of proteins stability contributions are approximately additive, and a key requirement of PESST is simplification; epistasis was not modelled (Bloom *et al.*, 2005).

During the course of simulated evolution, the population bifurcates every g_B generations into independent subpopulations (branches) which undergo sequence replacement *in populo* (supplementary figure 15). The bifurcation interval is defined as $g_B = \left\lfloor \frac{G}{\lfloor \log_2(N - n_{roots}) - \log_2(3) + 1 \rfloor} \right\rfloor$ such that there are always 3, 4 or 5 proteins left in each branch at the end of the simulation.

Proteins evolve according to a simple uniform clock, with a fixed probability of mutation (p_m) for each amino acid in every generation. Mutation follows the LG model of amino acid substitution (Le and Gascuel 2008; supplementary figure 18), with transition probabilities defined by the matrix \mathbf{L} , where $L_{a,a'}$ is the probability that amino acid a transitions to amino acid a' , with $a \neq a'$. Mutation rates vary across sites in the protein (defined by the vector

\mathbf{m}), and are drawn from a gamma distribution, $\Gamma(\kappa, \theta)$ with four rate-categories (supplementary figure 15).

During evolution, PESST continually tracks changes in stability at the amino acid ($\Delta_{r,a}$), protein (T) and population (\bar{T}_Φ) levels. Proteins falling below the stability threshold ($T < \Omega$) and those randomly selected according to p_{death} within each generation are killed and randomly replaced *in populo* by a stable protein.

The model has a set of default parameters (table 10), which are varied throughout the work presented in order to explore their impact upon the survivor bias hypothesis. From the tracked stabilities, PESST automatically produces animated figures in addition to statistical analyses and fasta files for ancestral reconstruction. The model is summarised in figure 18 and described in detail in supplementary methods 1. Additional symbols can be found in supplementary table 4. The code is open source and freely available for download*.

Precondition: Define \mathbf{L} , a matrix of amino acid transition probabilities, where $L_{a,a'}$ is the probability of amino acid a transitioning to amino acid a'

$\Delta \leftarrow [\sim \mathcal{N}(\mu, \sigma^2, skew)]_{r,a}$ ▷ Generate stabilities, $\Delta_{r,a}$

$\mathbf{m} \leftarrow [\sim \Gamma(\kappa, \theta)]_r$ ▷ Generate site mutation probabilities

$\eta \leftarrow \text{CREATEPROTEIN}(R, T_0)$ ▷ Create protein of length R , stability T_0

$\Phi \leftarrow \text{CREATEPOPULATION}(\eta, N)$ ▷ Clone initial protein N times

5: $\mathcal{T} \leftarrow \text{CREATEPARTITIONS}(\Phi, n_{roots})$ ▷ $\mathcal{T} := \{\{\Phi_{roots}\}, \{\Phi_{branches}\}\}$

$h_\Phi \leftarrow \text{EVOLVE}(\Phi, \mathcal{T}, g_B, \mathbf{L}, \Delta, p_m, \mathbf{m}, p_{death}, \Omega)$

function $\text{EVOLVE}(\Phi, \mathcal{T}, g_B, \mathbf{L}, \Delta, p_m, \mathbf{m}, p_{death}, \Omega)$

$h_\Phi[0] \leftarrow \Phi$ ▷ Record initial population

10: **for** $g \leftarrow 1$ to G **do** ▷ Loop over G generations

if $g \% g_B = 0$ **then**

$\mathcal{T} \leftarrow \text{BIFURCATE}(\Phi, \mathcal{T})$ ▷ Bifurcate every g_B generations

end if

$\Phi \leftarrow \text{MUTATE}(\Phi, \mathbf{L}, p_m, \mathbf{m})$ ▷ Mutate proteins by substitution

15: $\Phi \leftarrow \text{REPLACE}(\Phi, \mathcal{T}, \Delta, \Omega)$ ▷ Replace unfit proteins

$\Phi \leftarrow \text{KILL}(\Phi, \mathcal{T}, p_{death})$ ▷ Randomly kill proteins

$h_\Phi[g] \leftarrow \Phi$ ▷ Record generation history

end for

return h_Φ ▷ Return evolutionary history

20: **end function**

Figure 18 - PESST evolutionary algorithm pseudocode

* <https://github.com/bdevans/PESST>

Parameter	Symbol	Value(s)	Notes
Number of generations	G	2,000	Typically $G \geq 1,500$
Initial protein stability	T_0	{“low”, (lower, upper), “high”}	Typically $(\Omega + 5, \Omega + 25)$
Stability threshold	Ω	0	$T \geq \Omega := \text{fit}; T < \Omega := \text{unfit}$
Mean stability change ($\overline{\Delta}$)	μ	-2	$\mu \approx \frac{1}{R \cdot N} \cdot \sum_r \sum_a \Delta_{r,a}$
Standard deviation of Δ	σ	2.5	$\sigma \approx \text{var}(\Delta)^{\frac{1}{2}}$
Skew of Δ	k	0	
Population size	N	52	
Number of roots	n_{roots}	4	
Protein length	R	100	
Proportion of invariant sites	$p_{invariant}$	0.1	
Probability of mutation	p_m	0.001	Total mutations := $p_m \cdot R \cdot N$
Gamma distribution shape	κ	1.90	
Gamma distribution scale	θ	0.53	$\frac{1}{\kappa}$
Probability of Death	p_{death}	0.02	

Table 10 - Default parameters in PESST

Statistical tests

Equality comparisons of $\Delta_{r,a}$ distributions in the evolving dataset and the global stability matrix Δ were performed by the Kolmogorov-Smirnov test, computed by PESST. Equality meta-analyses were manually computed with Fisher’s combined probability analysis ($1 - X_{2k}^2$ where $X_{2k}^2 = -2 \sum_{i=1}^k \ln(p_i)$ with $2k$ degrees of freedom where k is the number of p -values from the 2-sided Kolmogorov-Smirnov tests in the meta-analysis; Winkler *et al.*, 2016). Analyses of the difference between the populations of ancestral and extant stabilities in each simulation were performed with both the Mann-Whitney U test and Welch’s t-test in Graphpad PRISM v7, as we cannot assume whether the distributions are normal or non-normal in every case. Analyses of correlations between node age and stability were computed all replicate simulations in a given condition with Spearman’s method in Graphpad PRISM v7. Trees used for the analyses were cladograms output from CodeML, where age was defined as the number of nodes preceding a node of interest until the root is reached. Stabilities were the averaged the normalized stability of ancestors in the longest subtrees containing at least three representative nodes of a given maximum subtree length. Equality comparisons between the stabilities of ancestors or consensus sequences of

separate simulation conditions were performed with the Mann-Whitney U test in Graphpad PRISM v7.

Ancestral Sequence Reconstruction

Fasta files were imported into the Geneious sequence analysis suite (ver. 10; Kearse *et al.*, 2012). Phylogenies were produced with PhyML (Guindon *et al.*, 2010) under standard settings with the LG+I+G model of amino acid substitution with estimated rates (Le and Gascuel, 2008). SH-like branch supports were computed for phylogenies. Resulting phylogenies were manually rooted on the root sub-population defined by PESST. Marginal reconstruction of ancestors within the dataset was performed with CodeML of the PAML software suite (Yang, 2007). Reconstructions of these data were performed under standard settings implementing the LG model of amino acid substitution with an estimation of gamma and of invariant sites (Le and Gascuel, 2008). Both reconstructed sequences and PAML generated cladograms were extracted from PAML outputs. Reconstructed sequence's stability values were calculated with PESST.extras.stability. PESST.extras.stability cross references a user defined stability matrix with a FASTA formatted list of sequences of equal length. Output stabilities were analysed manually.

Consensus sequences

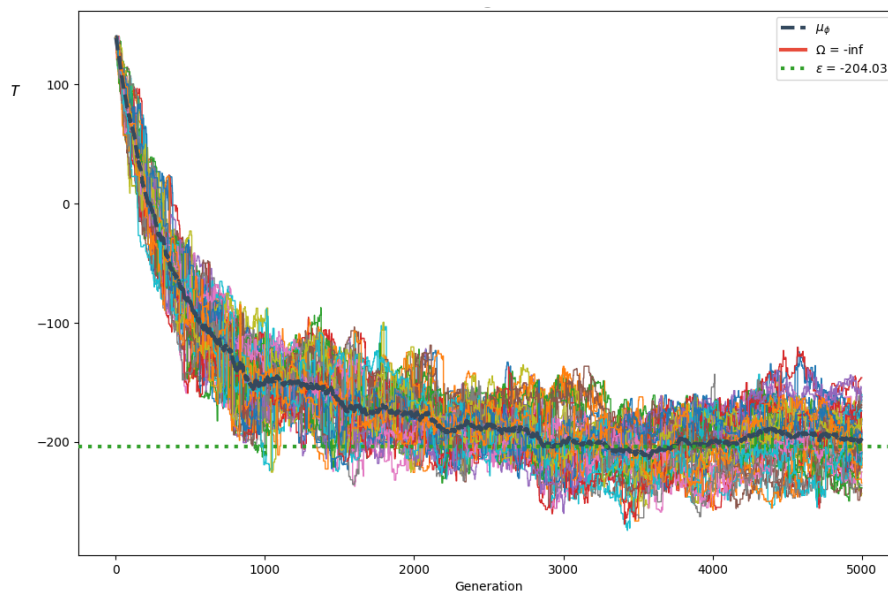
Consensus sequences of FASTA formatted alignments were generated with PESST.extras.consensus. This is a low powered consensus sequence builder that generates up to 3 consensus sequences from a user input alignment by outputting the most common amino acid at every site. For alignments with ambiguous sites (multiple amino acids are equally common), the algorithm outputs up to 3 sequences built from the second and if present, third equally likely amino acids. Consensus sequence stability was calculated with PESST.extras.stability as before. Output stabilities were analysed manually.

3.7 Results and discussion

3.7.1 Stability can be tracked as the protein evolves within PESST.

We firstly validated that our model (PESST v1.0) faithfully reconstitutes and tracks the expected population behaviour over time. We firstly established that, in the absence of any selective pressure, the mean stability of a population of sequences will neutrally evolve to the equilibration (ϵ) of the stability space. ϵ should approximate μR as a sequence's stability is the sum of its constituent amino acid stability contributions (which follow Tenet 1 above). We performed three sets of simulations, where sequences evolved from a "high", "medium" and "low" starting stability with respect to ϵ (figure 19; seeds in supplementary table 5). We simulated five repeats of each scenario (supplementary files). For both of the simulations initiated at T_0^{high} and T_0^{low} starting stabilities, the mean stability of the population converged on the expected value of ϵ (-200) within the 5,000 simulated generations. The simulations initiated at T_0^{med} had a mean stability that fluctuated around ϵ .

A



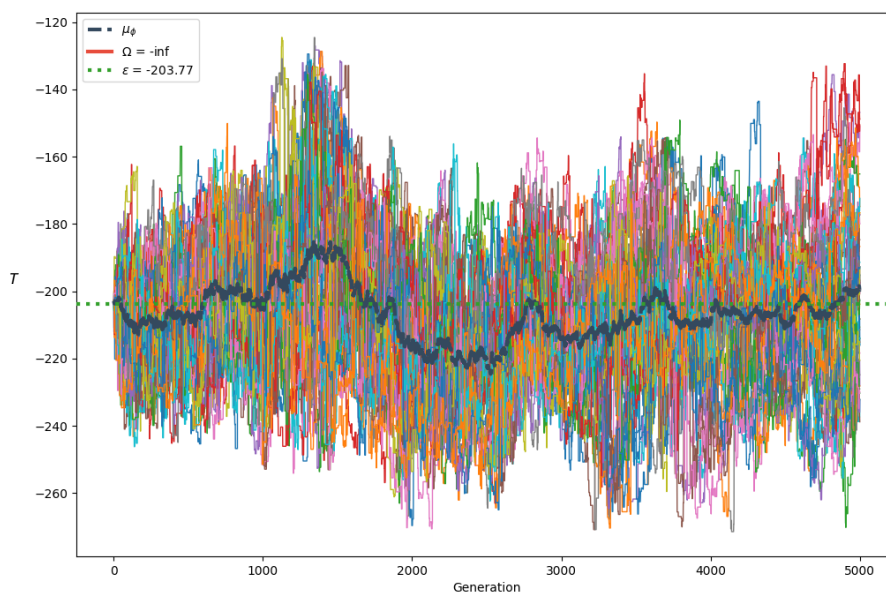
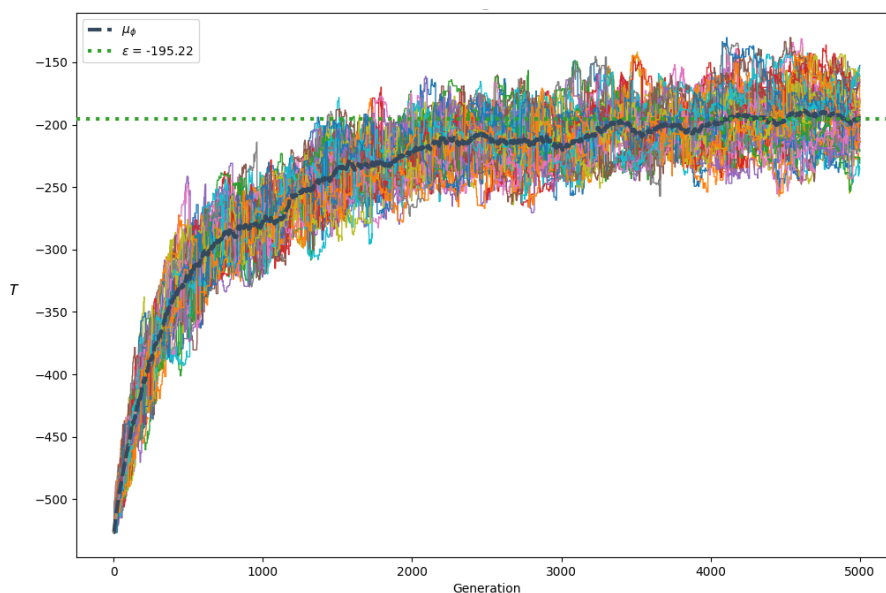
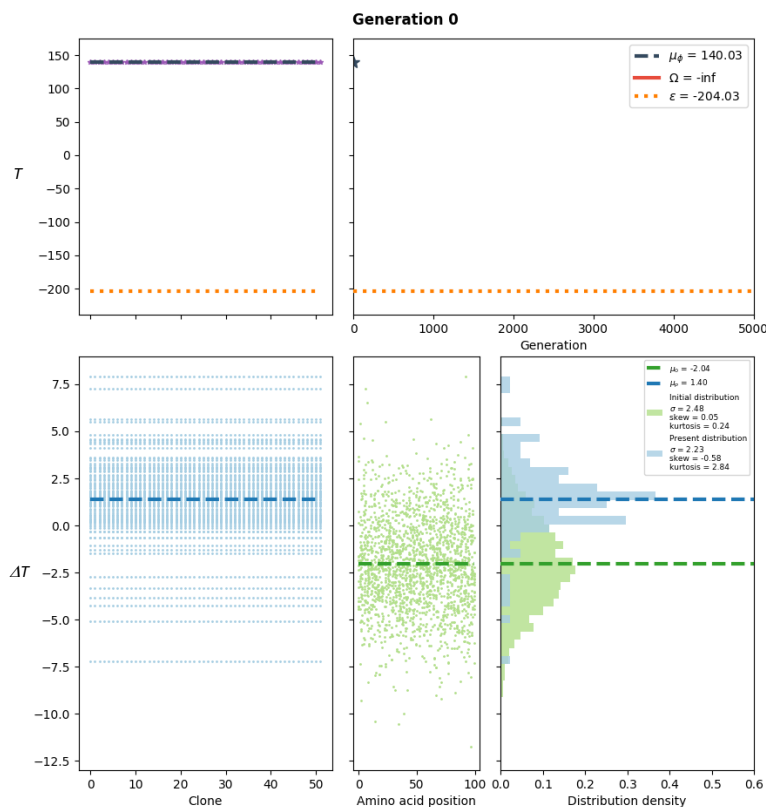
B**C**

Figure 19 - Mean stability of the population simulated in PESST tends toward ϵ during evolution

Representative stability traces for PESST simulations of 5,000 generations where $\mu = -2$ and $\delta = 0.002$, showing that the stability of PESST simulated protein populations approaches the predicted value of $\epsilon \approx -200$ for the given settings. Simulations were initialised at T_0^{high} (A), $T_0^{\epsilon+20}$ (B), or T_0^{low} (C). In each graph, each coloured line represents the stability of one of 52 clones in the dataset, which are each tracked independently and simultaneously by PESST. The green horizontal dashed line represents the predicted value of ϵ . The tight dashed black bold line represents the average T of the population.

To confirm whether variation was being captured within PESST, we tracked changes to the distribution of $\Delta_{r,a}$ values in the evolving dataset (figure 20; supplementary figure 18). As the simulation reaches equilibrium, it is expected that the distribution of $\Delta_{r,a}$ values in the population (Δ_Φ) will approximate the initially defined matrix of stability change values (Δ). We observed in all simulations that the mean population stability change $\bar{\Delta}_\Phi$ rapidly approaches the mean of the global stability matrix, $\bar{\Delta}$ (figure 21; supplementary files). After 5,000 generations the evolving Δ_Φ distribution and the global distribution of possible $\Delta_{r,a}$ values converge with significance regardless of the initial population starting stability (supplementary figure 19; $p = 0.48, 0.09$ and 0.79 for T_0^{high}, T_0^{low} and T_0^{med} respectively[†]). These results provide confidence that the evolution of sequences within PESST occurs according to pre-defined rules, and so produces predictable evolutionary outputs upon implementation of the first tenet of the survivor bias hypothesis.



[†] p -values from Kolmogorov-Smirnov tests ranged from 0.001 to 0.88

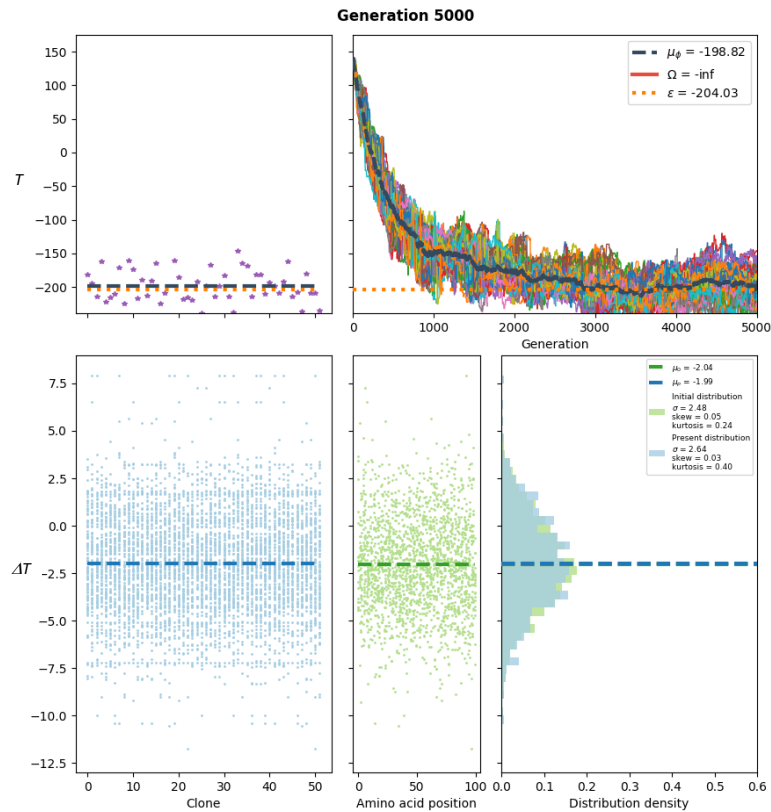


Figure 20 - Mean stability change of the evolving dataset tends toward the mean of the initial matrix (Δ) from which the protein is defined

PESST allow users to compare $\Delta_{r,a}$ values within the global stability matrix (Δ), to the $\Delta_{r,a}$ values in an evolving dataset (Δ_Φ) generated from the given matrix with no fitness threshold. Data represents the evolutionary simulation of figure 19A. Equivalent representations for figures 19B and 19C are shown in supplementary figure 18. Figures representing the five simulations for each starting parameter are available in supplementary files. Animations of simulations from generation 0-5000 in each case are also available in the supplementary files. Figures are representative simulations comparing Δ to Δ_Φ at generation 0 and at generation 5000. In each figure, boxes show: **Top-Left:** mean T of each clone in the dataset at generation. **Top-right:** progression of mean T until a given generation. **Bottom-left:** the distribution of every $\Delta_{r,a}$ value represented in every clone (Δ_Φ). **Bottom-centre:** the global distribution of $\Delta_{r,a}$ values showing every possible $\Delta_{r,a}$ value at every position in the protein (Δ). **Bottom-right:** Histograms of the $\Delta_{r,a}$ distributions produced by both matrices. In the top boxes, orange dotted lines represent the value of ϵ derived from $\bar{\Delta}$. In the bottom boxes, the dashed coloured lines represent the distribution averages ($\bar{\Delta}_\Phi$ and $\bar{\Delta}$). At generation 5000, both distributions and their means have converged.

3.7.2 A stability threshold biases the distribution of $\Delta_{r,a}$ values in the evolving dataset

Tenet 2 of the marginality bias hypothesis proposes that protein stabilities ratchet toward a stability threshold (Ω) below which the protein cannot fold, causing a fitness cost to the parent organism (Williams *et al.*, 2007; Harms and Thornton, 2013; Khersonsky *et al.*, 2018). We implemented this within PESST by imposing an Ω of 0. We then simulated evolution of sequences starting with either moderate (T_0^{med}) or high (T_0^{high}) starting fitness (figure 21; supplementary figure 20; supplementary table 6; supplementary files; T_0^{med} is defined as $\Omega + 5 < T < \Omega + 25$). In these simulations, the population was observed to evolve as before for between 250 and 500 generations. Once the first proteins violate the stability threshold and are eliminated from the population, the mean stability (\bar{T}_Φ) tends toward an equilibrium. The population mean stability equilibrium is maintained slightly above Ω due to a bidirectional pressure: neutral evolution following the destabilizing nature of Δ causes \bar{T}_Φ to tend lower towards ϵ , whilst the stability threshold Ω imposes a stabilizing pressure tending towards $\sim\Omega + 5$. These observations confirmed that PESST successfully simulated the population behaviour expected, following the previously reported observations of protein evolution that underpin the survivor bias hypothesis (Faure and Koonin, 2015; Pucci and Rومان, 2016; Goldstein, 2011; Tokuriki and Tawfik, 2009A).

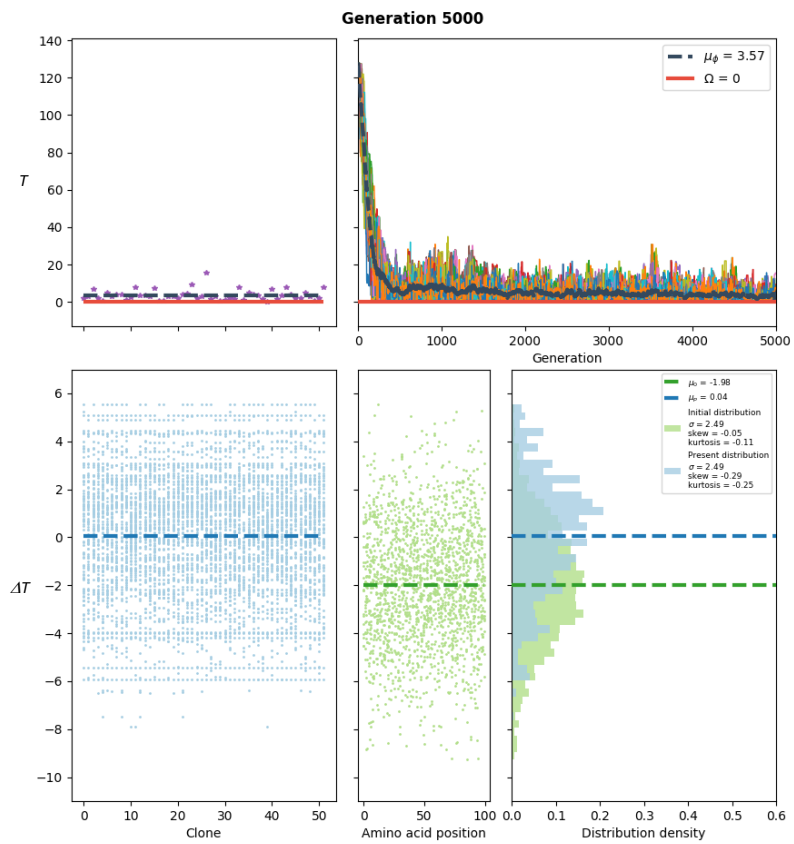
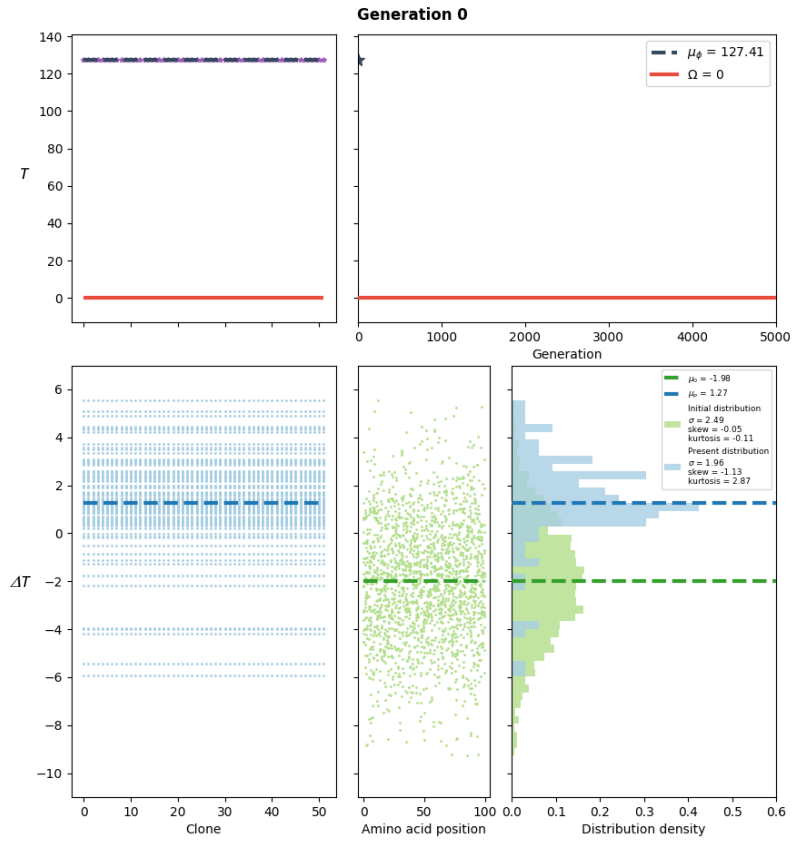


Figure 21 - Imposition of a stability threshold leads to equilibration of mean protein stability, and positive-bias in the population's $\Delta_{r,a}$ distribution

Data represents the simulation of 5,000 generations with an imposed Ω of 0. Sequences with a T that falls below Ω are immediately removed from the population and replaced *in populo*. Five simulations were initiated at T_0^{high} with $\mu = -2$ for the global stability distribution. All figures representing five simulations under these parameters are available in supplementary files. Animations of simulations from generation 0-5000 are also available for all cases in the supplementary files. Figures representing equivalent simulations initiated at T_0^{med} are available in supplementary figure 25 and supplementary files. Representative simulation comparing Δ to Δ_ϕ at generation 0 and generation 5,000. Boxes represent the same datatypes as figure 20. In top boxes, the red horizontal line represents Ω . Average stability across the population is maintained at equilibrium at $\sim\Omega + 5$. At generation 5,000, $\bar{\Delta}_\phi$ does not converge to $\bar{\Delta}$ ($P < 0.0005$)[‡]. High $\Delta_{r,a}$ values are significantly overrepresented, and low $\Delta_{r,a}$ values significantly underrepresented (supplementary figures 23 and 24).

We propose (*Tenets 3 and 4*) that under conditions of marginal stability, the competing pressures on protein stability cause an overrepresentation of high $\Delta_{r,a}$ amino acids, and an underrepresentation of low $\Delta_{r,a}$ amino acids within the evolving population. Our simulations with a stability threshold show positive-shifting of the distribution of evolving $\Delta_{r,a}$ values (Δ_ϕ) away from Δ (figure 21). These effects are observed when equilibrium is realised (figure 21; supplementary files). Across all of our simulations we observed that many of the most destabilizing residues are titrated out of the population, likely due to their large stability penalty when introduced (supplementary figure 21; supplementary files). It would be expected that such penalties would often push a sequence below Ω , causing such residues to be readily removed from the evolving dataset, as observed in analyses of factors underlying marginal stability in proteins (Tokuriki *et al.*, 2007; Goldstein *et al.*, 2011). Heat maps of $\Delta_{r,a}$ values from simulations evolved in the presence of a stability threshold over-represent stabilizing $\Delta_{r,a}$ values compared to $\Delta_{r,a}$ values from simulations without a threshold (supplementary figure 22; supplementary files). This disconnect between Δ_ϕ and Δ was highly, and globally, significant ($p < 0.0001$ for simulations initiated at both T_0^{high}

[‡] p -values from Kolmogorov-Smirnov tests ranged from 3.58×10^{-128} to 2.73×10^{-218}

and T_0^{med}). Therefore, whilst the evolving protein population is destabilized on average, we observe that the stability threshold limits the evolutionary space that the protein can sample. These results are in line with work by Goldstein, 2011, who also showed that a protein evolving at marginality in the absence of a fitness benefit for maintaining marginality is more likely to fix rare stabilizing residues than common destabilizing residues. Additionally, the same shift in the distribution of stability contributions compared to the global distribution of possible stability contributions was observed for all mutations in existing proteins by Tokuriki *et al.*, 2007. This work provides further evidence that for a protein evolving under the opposing evolutionary pressures that define marginality, positive biases are induced in the distribution of $\Delta_{r,a}$ values sampled by the protein population even once it has reached equilibrium.

3.7.3 Marginality causes overestimation of stability in ancestral sequences

We propose (*Tenet 4*) that the positive biases in the distribution of $\Delta_{r,a}$ values sampled by a population will cause reconstruction of ancestral proteins to bias towards stability. To test this, we reconstructed “ancestors” of PESST simulated evolution using the commonly used CodeML reconstruction algorithm in PAML (Yang, 2007). If there is no induced bias in the ancestors (null hypothesis), ancestors are expected to share an equivalent stability space as the evolutionary history of the evolved dataset. To test this, we simulated evolution with a stability threshold $\Omega = 0$ imposed, where $\bar{\Delta}$ is negative (where neutral evolution is destabilising, and a bias effect is expected), 0 (neutral evolution is non-destabilising; no bias effect expected) and 1 (neutral evolution is stabilising; no bias expected) in quintuplet (supplementary table 7). Simulations were run under standard parameters for 2,000 generations initiated at T_0^{med} starting fitness, defined by $\Omega + 5 < T < \Omega + 25$,

Throughout their evolutionary history simulations where global stability distributions had a negative mean (i.e. $\bar{\Delta} = -1, -2$) converged to an equilibrium slightly above Ω (supplementary files) as was previously observed. In these instances, the distribution of Δ_{Φ} in the evolving dataset did not converge on Δ (the global stability matrix). The evolving $\Delta_{r,a}$ distributions are equivalent when $\mu = -1$ and $\mu = -2$ (supplementary figure 23; supplementary files). When evolution was simulated with $\mu = 0$, equilibrium was not well

maintained. As $\bar{\Delta}_\Phi \approx \Omega$, the selective pressures exerted by both Δ and Ω are weak. The broad stability spaces explored by simulations with $\mu = 0$ is reflective of the release in selective pressure (supplementary figure 24A). This effect was even more pronounced in simulations where $\mu = 1$, where destabilizing selective pressures are released (supplementary figure 24B). Importantly, such large varieties in evolutionary history can be created in PESST by the change of a single parameter, allowing us to mitigate extraneous factors when analysing the effects of these histories on ancestral stability.

Resultant alignments at generation 2,000 had between 30 and 50% pairwise sequence identity. Phylogenies generated from each alignment of sequences at generation 2,000 (Guindon *et al.*, 2010) were generally well supported for the majority of nodes (supplementary figure 25; supplementary files). We then reconstructed ancestors of the derived phylogenies (Yang *et al.*, 2007). “Ancestor” stabilities were calculated against from their derivative global $\Delta_{r,a}$ matrices using PESST. To measure the impact of the bi-directional selective pressure, we analysed the global stability space that is represented by each evolved dataset. Global stability space can be defined as the combined normalized stabilities of both ancestor and extant (generation 2,000) populations in a given tree. Our null hypothesis states that proteins that have evolved at marginality throughout their history should produce a global stability space that is equally proportioned between the ancestral and extant sequences. As sampled stability space has been consistently maintained throughout the simulation ancestors should not show derivation in stability space from the extant sequences. However, for simulations that maintained marginality, the average proportion of normalized stability space represented by extant:ancestor sequences was unevenly distributed 28:72 and 34:66 for global stability matrices satisfying $\mu = -2$ and $\mu = -1$ respectively. In contrast, simulations that escape marginality showed almost evenly distributed average normalized stability spaces, with average ratios of 47:53 and 52:48 for extant:ancestor sequences with global stability matrices satisfying $\mu = 0$ and $\mu = 1$ respectively (figure 22A). These data are in support of previous simulations of protein evolution by Williams *et al.*, 2006, who found that maximum likelihood based reconstructions of simulated datasets based on evolving lattice proteins are considerably overestimated. Overestimation of stability in our reconstructed simulations suggest that the combined pressures imposed from both the destabilizing $\bar{\Delta}$ and a stability threshold on

evolving proteins is sufficient to positively bias the stabilities represented by predicted ancestors.

To further understand how this bias manifests in the global stability space, we compared the distributions of stabilities attained in the extant and ancestral sequence populations on a simulation-by-simulation basis. Disparity between stabilities sampled by extant and ancestral proteins is consistently observed when a stability threshold is imposed and μ is negative (figure 22B). The range of stabilities sampled by the ancestors are between 1.8 and 5.1 times wider than the ranges sampled in the extant sequences. All simulations showed a significant difference between the ancestral and extant stability distributions ($p < 0.05$; figure 22B; supplementary table 8). Distributions of ancestral stabilities show high density at low values, with a long tail of high stabilities (figure 22B). The majority of the most stable ancestors represented stabilities higher than any stability value sampled by their derivative population throughout their evolutionary history (supplementary figure 26; supplementary files). In contrast, these patterns are not observed for simulations where the global stability matrix satisfies $\mu = 0$ and $\mu = 1$, as the selection bias is considerably reduced or non-existent. In these cases, ancestral stabilities are not apparently overestimated (figure 22B), and the ranges of stabilities sampled in the extant and ancestor populations are equivalent ($p > 0.05$ for eight of ten, and ten of ten simulations in the statistical tests used). These data are intuitive when considering the nature of the positive-biased $\Delta_{r,a}$ distribution in the evolving population. Simply because ASR can only act on present-day data to generate hypotheses about ancient sequence space. In a single protein, $\Delta_{r,a}$ values combine to produce marginal stability. However, when calculating across all proteins, marginality increases the likelihood that ASR selects neutral or stabilizing residues. These data lead us to reject our null hypothesis, as the maintenance of marginality causes the stability space across a given phylogeny to become positively biased towards the reconstructed sequences.

Many past ASR studies have shown that sequences trend toward increased stability as they are reconstructed from more ancient nodes (Gaucher *et al.*, 2008, Hobbs *et al.*, 2012; Akanuma *et al.*, 2013; Butzin *et al.*, 2013; Hart *et al.*, 2014; Risso *et al.*, 2015; Okafor *et al.*, 2018). To understand whether these properties can manifest in simulated datasets, we analysed correlations between space-normalized ancestor stability versus node distance

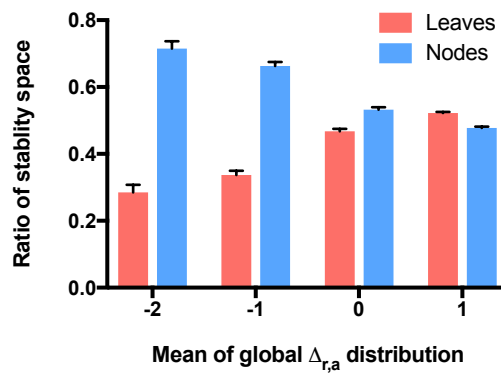
from the root of the tree (figure 22C). Significant ($p < 0.001$), strong negative correlations were observed for simulations under opposing pressure from the stability threshold and the global stability distribution ($r = -0.94$ and -0.88 for global stability matrices satisfying $\mu = -2$ and $\mu = -1$ respectively). When the global stability matrix satisfied $\mu = -2$, the linear regression gradient ($r = -0.19$) was almost twice that where global stability matrix satisfied $\mu = -1$ ($r = -0.099$), counterintuitively suggesting a stronger stabilizing force is exhibited under stronger destabilizing pressures. This was further evidenced when comparing the fold-difference in stability ranges attained between extant versus ancestral sequences in both simulation scenarios ($\mu = -2$: 3.71x; $\mu = -1$: 2.47x; figure 22B). On the other hand, simulations where the global stability matrix satisfies $\mu = 0$ a slight, significant negative correlation ($r = -0.39$; $p = 0.015$) was observed, whereas when $\mu = 1$ showed a slight, significant positive correlation ($r = 0.33$; $p = 0.041$). Therefore it is evident that evolution under marginality is a sufficient pressure in the protein's evolutionary history to produce ancestral reconstructions that significantly increase in stability with age. This raises the possibility that in some cases, the thermostability of ancient proteins is not indicative of their environmental temperature, but instead of marginality bias. We therefore urge caution when drawing such conclusions from ASR studies.

To better illustrate the need for caution, we also performed reconstructions of simulations where the global stability distribution satisfied $\mu = -2$, but begin their evolution at high stability (supplementary figure 27). Correlations in ancestor stability over time were calculated as before. A significant negative correlation was observed ($r = -0.94$, $p < 0.001$; figure 22D) that was statistically indistinguishable from correlations observed for the same stability distribution Δ_{Φ} evolving continually at marginality ($Z_{obs} = 0$). However, when comparing the five highest stability ancestors for each simulation in both datasets, it can be seen that initiation of simulations from high stability leads to the reconstruction of proteins with significantly higher stabilities ($U = 88^{\S}$; $p < 0.0001$; figure 22E). Therefore, ASR is still able to reconstruct accurate models of a protein's evolutionary history. However, confidently distinguishing whether trends of increasing stability are caused by survivor bias or a true evolutionary history requires additional evidence about the evolutionary history of

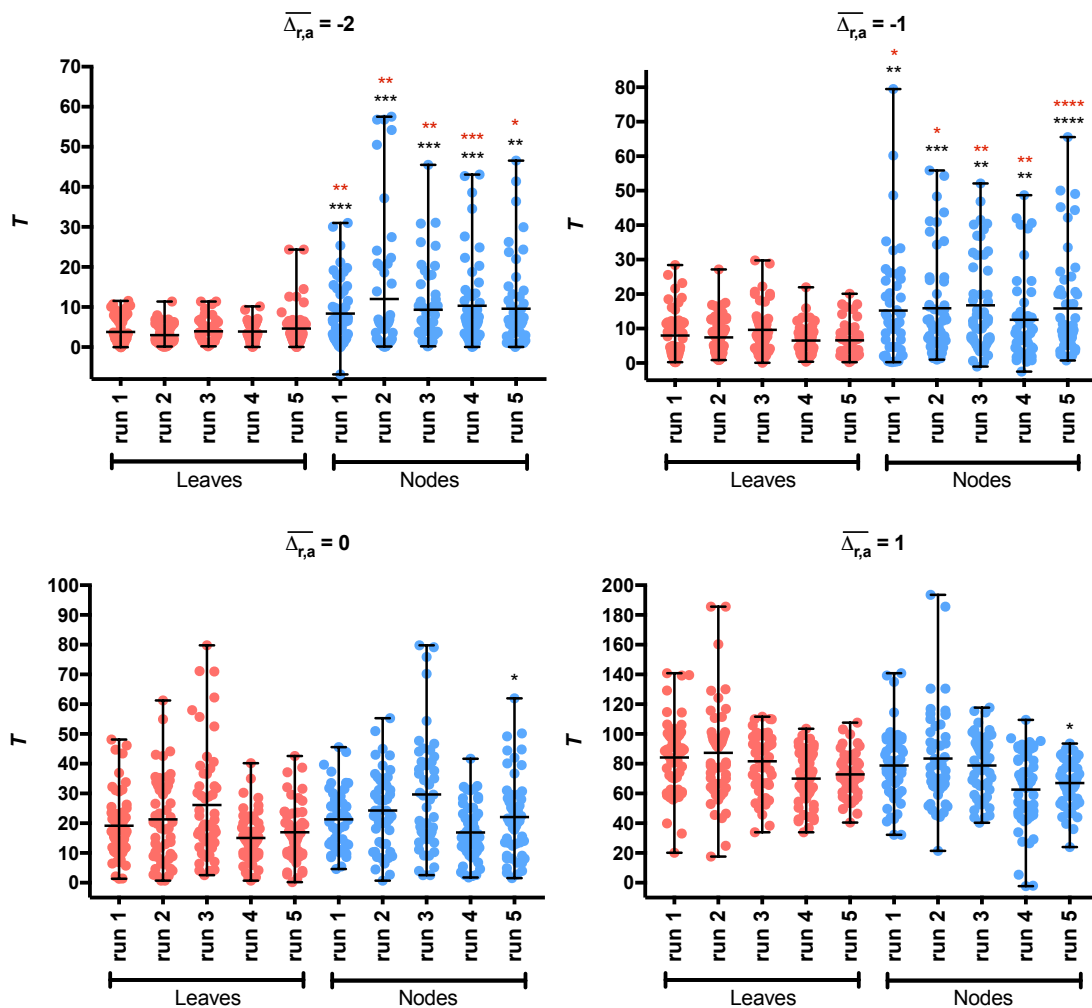
[§] $U_{Max} = 625$. U_{Max} is defined as $n_1 n_2$

the protein in question (i.e. correlation with evidence from isotope data in Gaucher *et al.*, 2008, Garcia *et al.*, 2017 and Akanuma, 2017). Others have suggested that only tracking single proteins is an unreliable way to infer conclusions of the ancient biosphere (Hart *et al.*, 2014). Methods and guidelines for performing such reconstruction experiments with apt rigor are outlined in Kacar *et al.*, 2017.

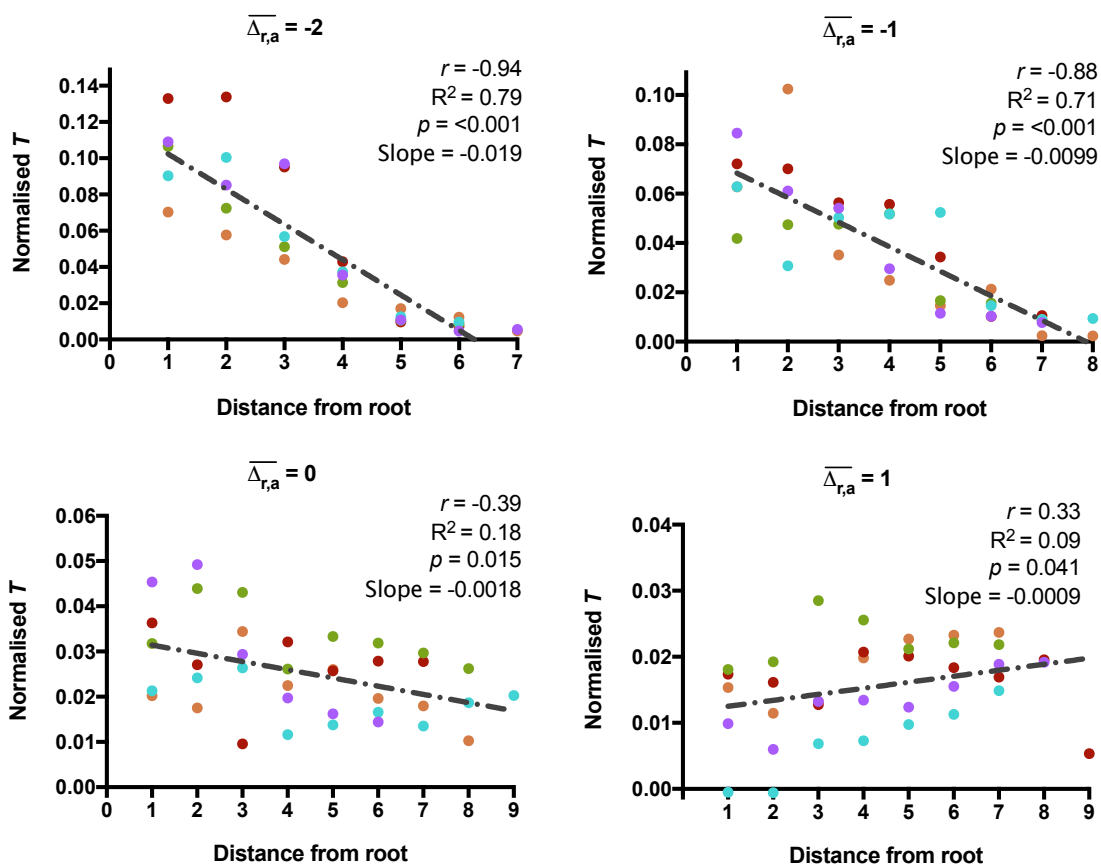
A



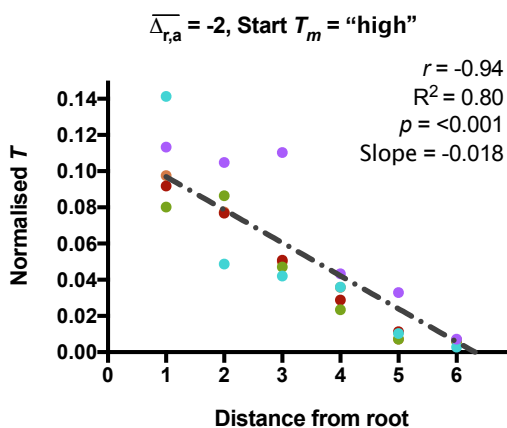
B



C



D



E)

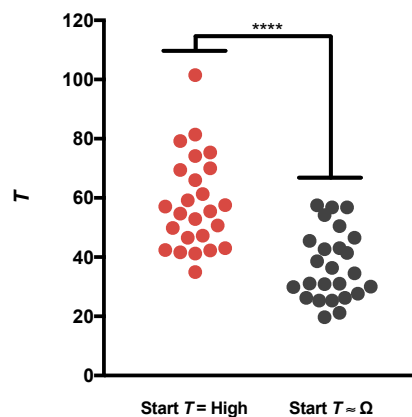


Figure 22 - Survivor bias drives increasing stability in ancestral proteins

Unless stated, figures represent analyses of stability space across ancestral and extant sequences, following reconstructions of quintuplet simulations with global stability distribution μ values of -2 , -1 , 0 and 1 under a stability threshold, initiated at T_0^{med} and allowed to evolve for 2,000 generations. Asterisks represent the degrees of significance (**** = $p < 0.0001$; *** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.05$) **A)** Histogram showing that global stability space is biased toward nodes when the global stability distribution has a negative mean (exerting pressure). Tree stability space for each run was

normalized, and proportioned between node and leaf stability. Average proportion values were taken of 5 runs. Error bars represent standard error of the mean. **B)** Columnar scatter plots comparing the distribution of leaf stabilities (Red) to node stabilities (Blue), showing that when the mean of the global stability matrix is negative, node stabilities are overestimated. Comparisons of the distributions were undertaken on a run-by-run bases with both Welch's *t*-test (significance shown by black asterisks) and Mann-Whitney U test (significance shown by red asterisks). Black horizontal bars represent the mean and range of the datasets. **C)** Scatter plots of normalized stability with respect to distance from the root of the tree, showing that a global stability distribution with negative μ leads to more ancient nodes having increased stability. Node distance from the root was calculated from cladograms output from PAML as the number of nodes preceding the node of interest. Node stabilities were normalised to 1. Spearman's correlation and goodness of fit were calculated independently for each dataset. **D)** Scatter plot analysed as above for a simulation of protein evolution with a global stability matrix satisfying $\mu = -2$, initiated with T_0^{med} . **E)** Grouped min-max scatter of the five highest *T* values from the five simulations of both scenarios where the global stability matrix satisfies $\mu = -2$, comparing the difference in stabilization when sequences evolve from a starting point of high stability (red points), or evolve at marginality their entire history (grey points). Mann Whitney U test was used to calculate whether the populations were significantly dissimilar.

3.7.4 Marginality causes overestimation of stability in consensus sequences

Importantly, consensus proteins present another alignment-based tool for the engineering of protein stability (Okafor *et al.*, 2018; Durani and Magliery, 2013; Kiss *et al.*, 2009). However, the driving forces behind stabilization are poorly understood. As discussed, marginality bias titrates out destabilizing residues, populating the global extant protein population with stabilizing or neutral residues. As consensus proteins are effectively averages of extant sequence space, it is expected that marginality induces biases in consensus sequences towards increased stabilities. To test this hypothesis, we used PESST to generate consensus sequences of generation 2,000 of each simulation used for the previous analysis (figure 23). Again, the pressure to maintain marginality was a sufficient force to generate consensus sequences with considerably higher stability than the stabilities sampled across the population's evolution. When marginality could be escaped, consensus sequences were comparatively stable to the extant sequences. Interestingly, when we

simulated a strong destabilising neutral drift (global stability matrix satisfies $\mu = -2$) starting from a highly stable sequence, the simulation produced consensus sequences with markedly high stabilities. The stability space sampled by these consensus sequences was significantly higher than the space sampled by the counterpart simulation that maintained marginality throughout its history ($U = 0^{**}$; $p < 0.0001$; supplementary figure 28). These data support the hypothesis that consensus sequences are stabilized by introduction of ancestral residues, albeit in an evolutionary history dependent manner (Porebski and Buckle, 2016; Ye *et al.*, 2017).

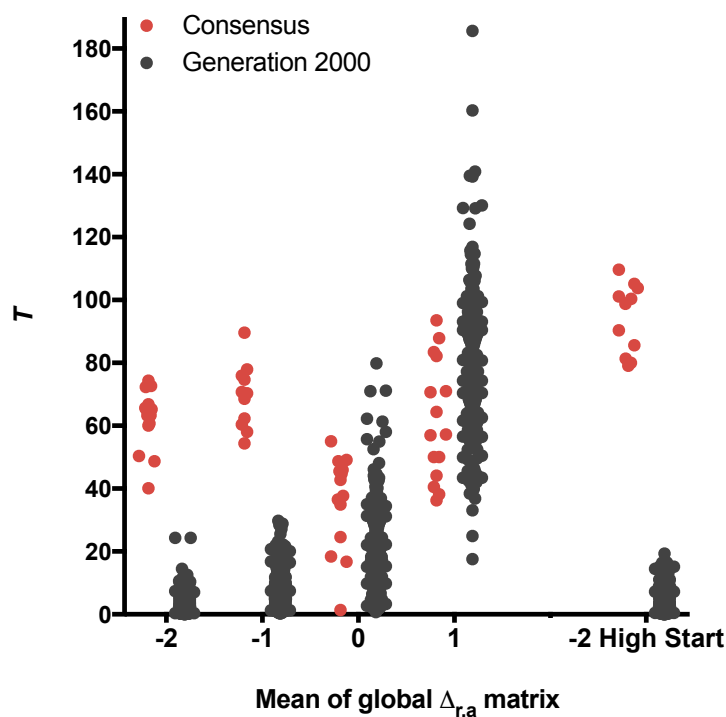


Figure 23 - Survivor bias drives the stabilization of consensus sequences

Grouped min-max scatter plots representing the stability of consensus sequences derived from generation 2000 of the protein simulations described in figure 22. From each simulation scenario, a combined dataset of consensus sequences from each simulation replicate is compared to a combined dataset of extant sequences from each simulation replicate.

** $U_{Max} = 154$. U_{Max} is defined as $n_1 n_2$

Overall, these data suggest that survivor bias is an inherent property of sequences derived from a multiple sequence alignment. Therefore, we posit that stabilization bias will be a common property of sequences derived from such datasets. This can be expected as the overrepresentation of stabilizing residues is predicted to be a trait shared across all marginally evolving protein families (Tokuriki *et al.*, 2007). Stabilizing residues are considerably more rare than destabilizing residues when randomly mutation real proteins (Bloom and Glassman, 2009). Selecting residues from alignment datasets presents a method to “game” these probabilities by offering a biased selection space. In accordance, a recent study engineering sequences with the consensus method found that stabilization was typical, with 75% of families tested exhibiting higher stability than any modern counterpart (Sternke *et al.*, 2018). As the stabilization of consensus sequences and ancestral sequences are both driven by survivor bias, ASR-based stabilization should also be robust for a broad range of protein families, as reconstructions of the CAR, PON and CYP3 protein families already evidence (Thomas *et al.*, 2018; Trudeau *et al.*, 2016; Gumulya *et al.*, 2018). While this may add a layer of complexity to future ASR studies probing relationships between evolutionary history and stability in ancient proteins, it provides further evidence that ASR should have considerable and consistent utility in protein engineering fields (Wilding *et al.*, 2017; Clifton *et al.*, 2017; Zakas *et al.*, 2016; Gumulya *et al.*, 2018). These results add credence to alignment based stability engineering methods, as until now the underlying mechanisms behind alignment-based stabilization were poorly understood (Sternke *et al.*, 2018; Gumulya *et al.*, 2018; Trudeau *et al.*, 2016).

3.8 Conclusion and perspective

Here we have developed the survivor bias hypothesis, describing the origin of stability in ancestral proteins derived from protein families with a mesostable evolutionary history. We hypothesised that a bi-directional pressure exists on the majority of evolving protein populations. A constant destabilizing pressure is exhibited as the majority of potential mutations across the protein will lower its stability. This is counteracted by a constant stabilizing pressure caused by a stability threshold, whereby if the protein’s denaturation temperature falls below its typical ambient temperature, a significant fitness cost is imposed onto its parent organism. We further hypothesised that exertion of opposing forces from a

stability threshold and destabilizing mutations leads to an overrepresentation of stabilizing residues in the evolving dataset. As most proteins exist at marginality due to the bidirectional selective pressure, significantly destabilizing residues can rarely be permitted as they are highly likely to force the protein below the stability threshold. Therefore, any method that derives protein sequences from alignment data will overestimate stability values with respect to their evolutionary history as the modern population is saturated with stabilizing and neutral residues. To test this hypothesis, we built a simple, highly parameterizable model of protein evolution that follows pre-established evolutionary models, allows us to track simulated protein stability across evolution, and provides the ability to exert both selective pressures described.

The data presented in this study provide strong evidence for the existence of a “marginality bias” that can explain the thermostability of ancestral proteins derived from extant sequences. We developed a novel algorithm, PESST, to simulate evolution whilst tracking protein stability. This allowed us to postulate that when proteins evolve under a bidirectional selective pressure, reconstructed ancestral proteins have overestimated stability. Additionally, our simulations show that this marginality bias produces strong, significant correlations between stability and age, showing that the older a node is, the higher the reconstructed T . Furthermore, we find that consensus sequences are also stabilized by these same forces. These observations provide an explanation for why reconstructions of real data without a high temperature evolutionary history still produce thermostable ancestors (Trudeau *et al.*, 2016; Chapter 2). The data presented in this study therefore provide a broader understanding about the mechanics underlying stabilization in the burgeoning, but still young protein engineering tool of ASR. We provide evidence that ASR is a ubiquitous engineering tool, enabling the engineering of thermostability regardless of a protein’s evolutionary history.

3.9 Supplementary methods

A detailed description of the PESST algorithm

Initiating a starting sequence

- Additional symbols can be found in supplementary table 4
- A protein ($\eta_{initial}$) of user definable length R is formed, containing a start methionine followed by randomly generated amino acids.
- Each position (r) can contain one of 20 amino acids (a). From a user definable normal distribution \mathcal{N} of mean μ and shape σ^2 , the model randomly generates a 2D matrix Δ , where $\Delta_{r,a}$ describes a ΔT value of a given amino acid at a given position. The stability (T) of η is given by $\sum_{r=1}^R \Delta_{r,a_r}$ (supplementary figure 14).
- In nature, sites become fixed in a population if they are essential for function despite possible detrimental ΔT values. Therefore, to account for this behaviour, the model defines invariant sites to a proportion of the amino acids in the protein ($p_{invariant}$)
- The user sets a stability threshold (Ω), where Ω satisfies $-\infty < \Omega < T^{initial}$.
- Natural sequences exhibit rate variation across sites. Rate variation can be modelled to a gamma distribution (Γ) with four independent rate categories (Yang et al., 1994).
- Independent rate categories are generated each run by taking the median value of four quartiles of 10,000 samples from a gamma distribution of a user defined shape (κ) and scale. Typically a scale of $\frac{1}{\kappa}$ is used. Each variant position is randomly assigned to one of four rate categories, defining a matrix of site-wise mutation probabilities \mathbf{m} (where $m_r \sim \Gamma(\kappa, \theta)$; $\sum_{r=1}^R m_r = 1$), which remains constant throughout the simulation (supplementary figure 15).
- The user sets one of three possible initial T values (low, medium [bounded range], high) that modifies P into the sequence that will be used for evolution (Q).
 - T_0^{low} and T_0^{high} are treated in the following manner: every site where residues are not fixed is swapped for another amino acid chosen randomly from a pool of the three largest or smallest values of $\Delta_{r,a}$.

- T_0^{med} requires the user to input a T range where $T_{min} > \Omega$ and T_{min} and T_{max} are between the minimum and maximum bounds of $\sum_{r=1}^R \Delta_{r,a_r}$. The model then modifies non-fixed residues until the first protein sequence is discovered that satisfies a value in the range by hill-climbing.

Evolving a sequence

- Once a starting sequence η of length R , with site-wise mutation probability \mathbf{m} , and a global fitness of $T_{max} \geq T > \Omega$ has been generated by the model, the sequence is cloned to generate a starting population (Φ) of a user-defined size (N).
- The population evolves according to a uniform clock over a user-defined number of generations (G). At every generation, each amino acid undergoes mutation with a constant probability p_m , where $p_m \cdot R \cdot N$ defines the total number of mutations per generation. $p_m \cdot R \cdot N$ sites are selected to mutate at rates according to \mathbf{m} . A site with amino acid a transition to a new amino acid a' based on the Le and Gascuel (LG) amino acid replacement matrix (\mathbf{L} ; Le and Gascuel, 2008) that is modified so $a \neq a'$ (supplementary figure 16).
- A protein's fitness is considered binary (*fit|unfit*). Proteins are considered unfit when $T < \Omega$. Before each generation, the model checks for unfit sequences in Φ . If this is satisfied, η_{unfit} is deleted and replaced with another sequence in the population that satisfies $T > \Omega$.
- Evolution is simulated with population isolation to mimic bifurcations. In this instance, the model divides the global population into even sub-populations Φ_{roots} and $\Phi_{branches}$ where $\Phi_{branches}$ split at a bifurcation interval g_B where $g_B = \left\lfloor \frac{G}{\lfloor \log_2(N - n_{roots}) - \log_2(3) + 1 \rfloor} \right\rfloor$ (supplementary figure 17). Isolation events occur at equal time-points such that every final population at the end of the run contains 3, 4 or 5 individuals. When an individual in a sub-population dies, it can then only be replaced *in populo*, generating independent lineages. An edge case in this factor required a feature in the model that diverges significantly from nature. If every sequence in a subpopulation of Φ satisfies $T < \Omega$ the entire subpopulation goes extinct.

Therefore, the simulation reverts to the prior generation to re-attempt mutating sequences to avoid complete branch extinction.

- If the user desires, evolution can be run assuming death happens naturally in the population, aside from being outcompeted due to fitness. At every generation, each member of Φ has a user defined probability of dying (p_{death}). As before, dead individuals are immediately replaced by other individuals *in populo*. This allows for evolution that occurs without replacement caused by fitness to produce a phylogeny that is not a star-phylogeny.

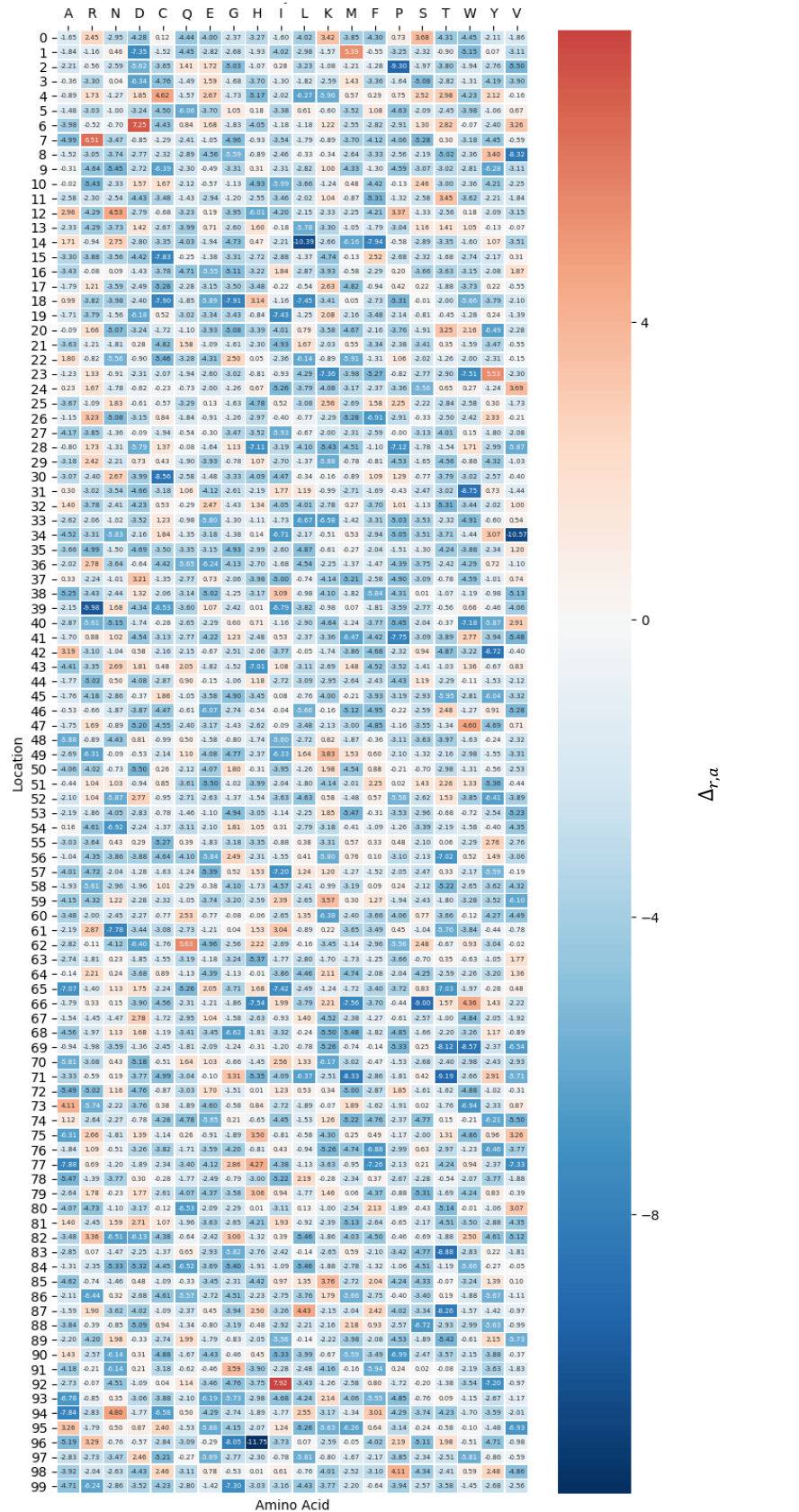
Outputs

The model is able to track and output a variety of useful data about the population's evolution:

- At a user defined generation rate, the model can output FASTA files describing the sequences of Φ .
- A scatter plot describing the change in T of every sequence in Φ over time.
- At a user defined generation rate, the model will output data, graphs and animations describing every $\Delta_{r,a}$ of each amino acid within Φ at a given generation compared to $\Delta_{r,a}$ values stored in Δ .
- At a user defined generation rate, the model will output data on the distribution of ΔT values within Φ (Δ_{Φ}), including the Anderson-Darling, Skewness-Kurtosis all, and 2-sided Kolmogorov-Smirnoff statistical tests for normality of the data.

3.10 Supplementary figures

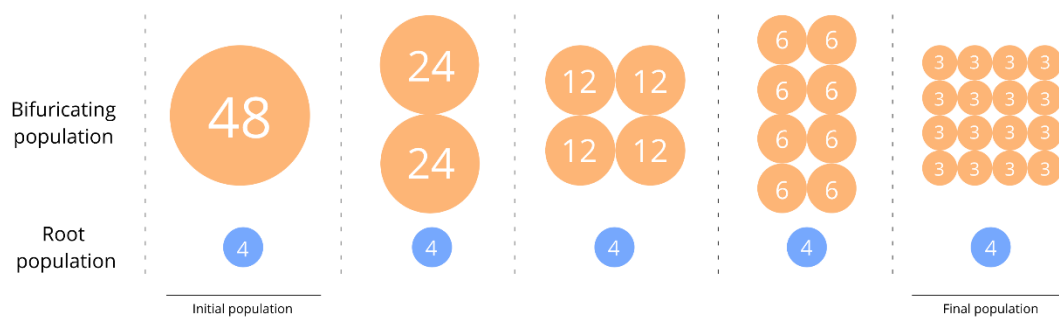
Supplementary figure 14



Supplementary figure 14 - Representative Δ matrix generated in PESST simulations

Heat map of a representative Δ matrix generated from a Gaussian distribution at the initiation of a PESST simulation. Each possible amino acid at each position of the protein is assigned a $\Delta_{r,a}$ value. The matrix remains consistent for the entire simulation, and can be used to back-calculate the stabilities of protein sequences derived from the representative simulation. Given matrix is derived from a Gaussian distribution of $\mu = -2$ and $\sigma = 2.5$.

Supplementary figure 15

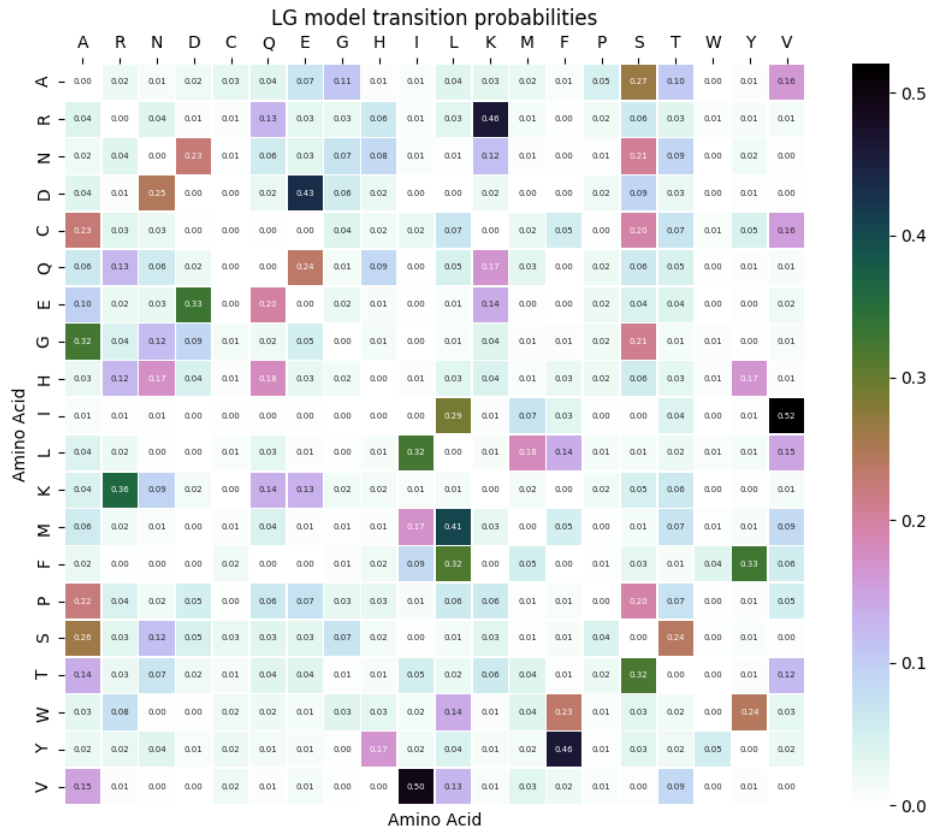


Supplementary figure 15 - Implementation scheme of bifurcations in PESST

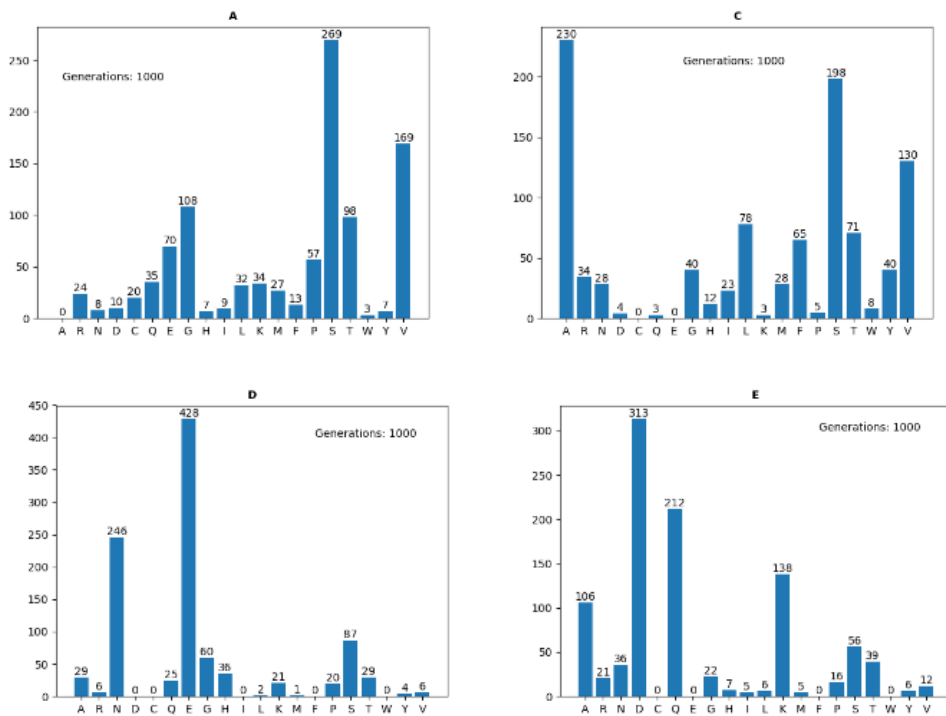
Scheme representing how bifurcations are implemented in PESST. Bifurcating the population “geographically separates” the data into subpopulations, meaning as the protein population evolves, Each sub-population can only replicate *in populo*. This is intended to be a rough mimic for natural bifurcating data. At even generation intervals, the population will bifurcate completely once. The amount of bifurcations is dependent on the population size. Bifurcations occur until all final sub-population sizes are 3, 4 or 5. Additionally a root sub-population is isolated from the data at generation 0. Implementation of bifurcation into PESST allows for increased phylogenetic signal within simulations, and the generation of clear well supported phylogenies.

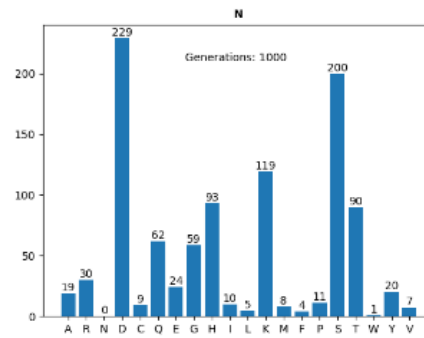
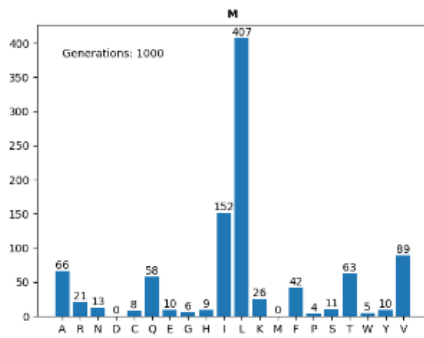
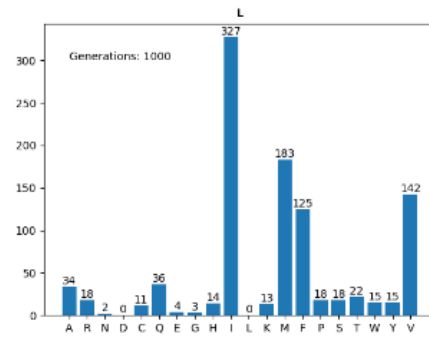
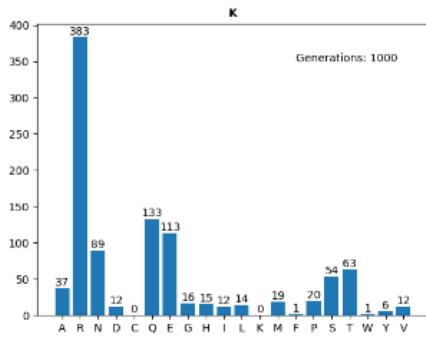
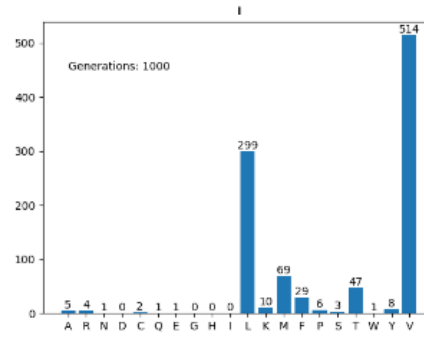
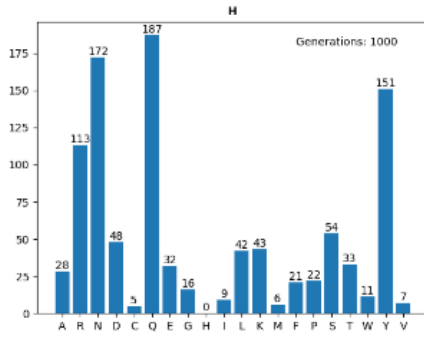
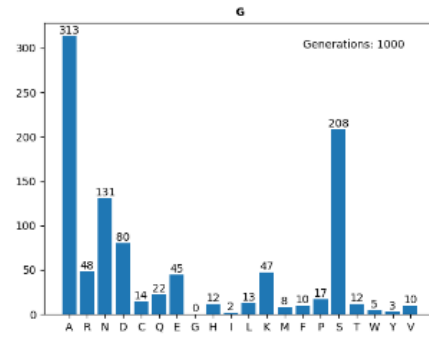
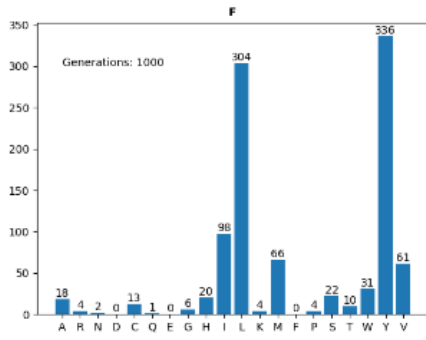
Supplementary figure 16

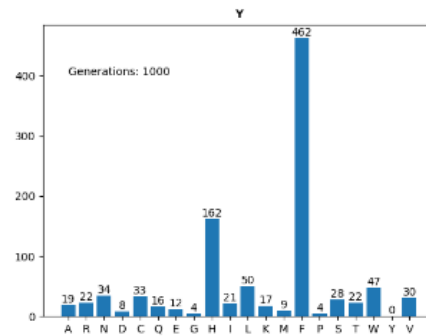
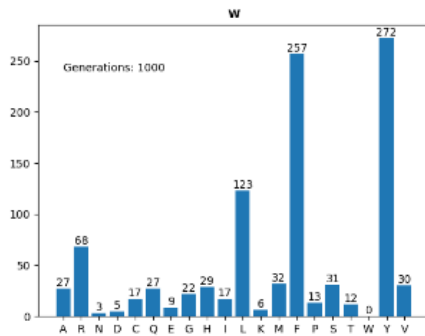
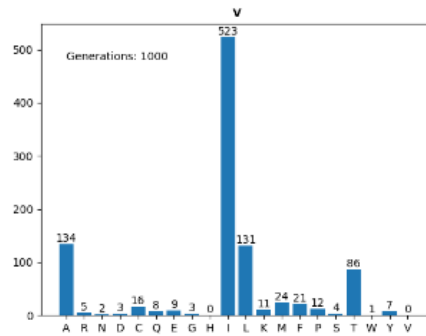
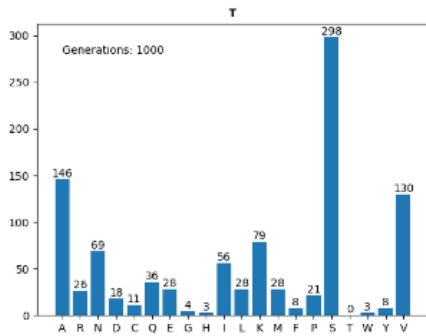
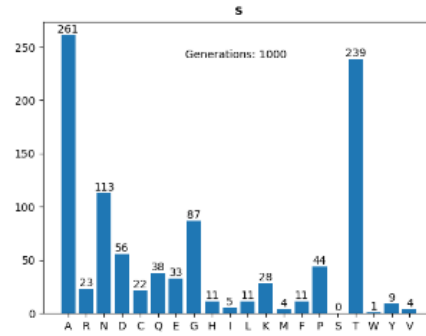
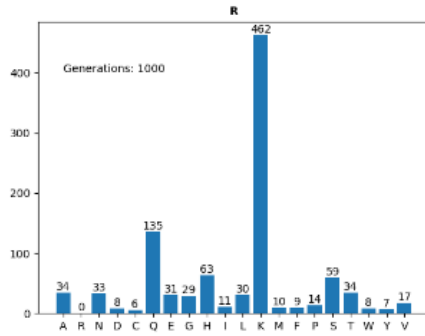
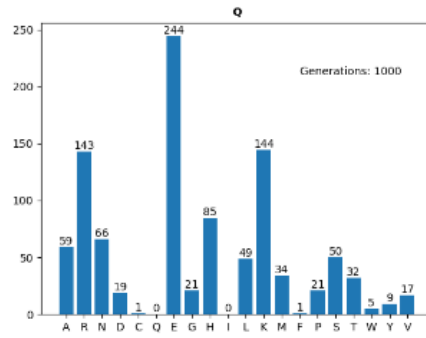
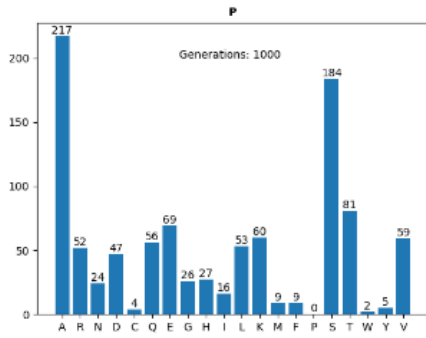
A



B





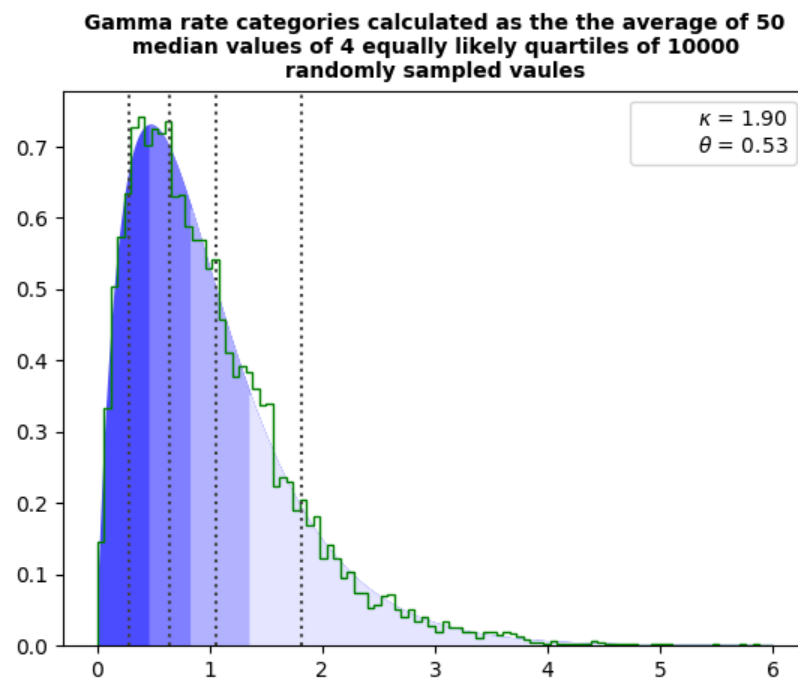


Supplementary figure 16 - Transition rates at each site derived from a modified LG model implemented into PESST

The LG model is a matrix of substitution likelihoods used to compute the likelihood and nature of mutations along a phylogenetic branch. Diagonals in the matrix describe the rate at which an amino acid does not change. By modelling evolution to a uniform clock, and dictating that a set number of

amino acid conversions must happen at each generation, the diagonal rates in the original LG model become moot. **A)** The LG matrix with, diagonal values set to 0 and the remaining values normalised to 1 providing a means to calculate the substitution likelihoods when an amino acid change is forced. **B)** The distribution of substitutions that occur according to the model from 1,000 generations of mutation at each possible amino acid state, suggesting the LG model is implemented as expected.

Supplementary figure 17

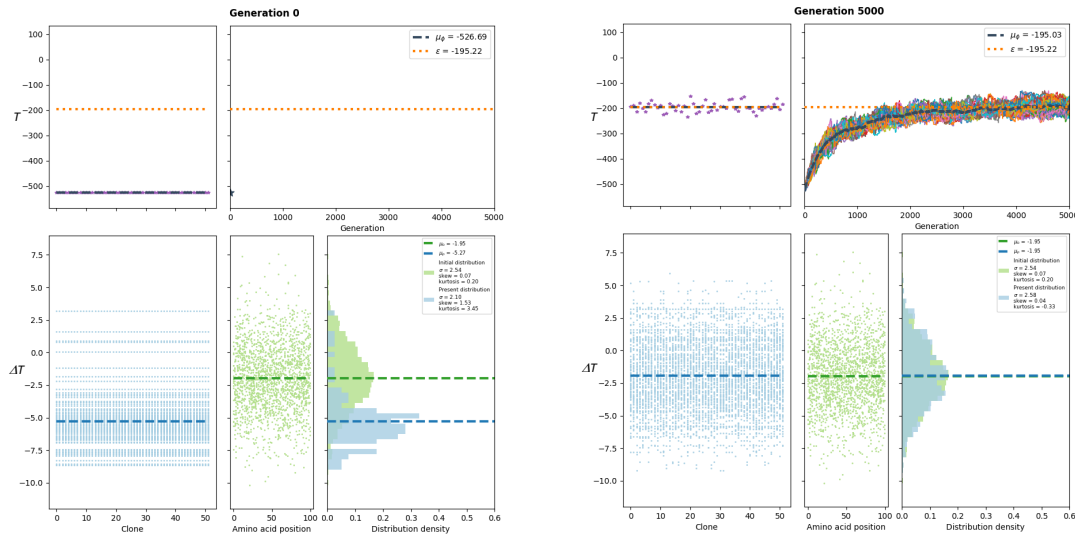


Supplementary figure 17 - Four independent rate categories defined by the median values of four quartiles from 10,000 samples of the gamma distribution

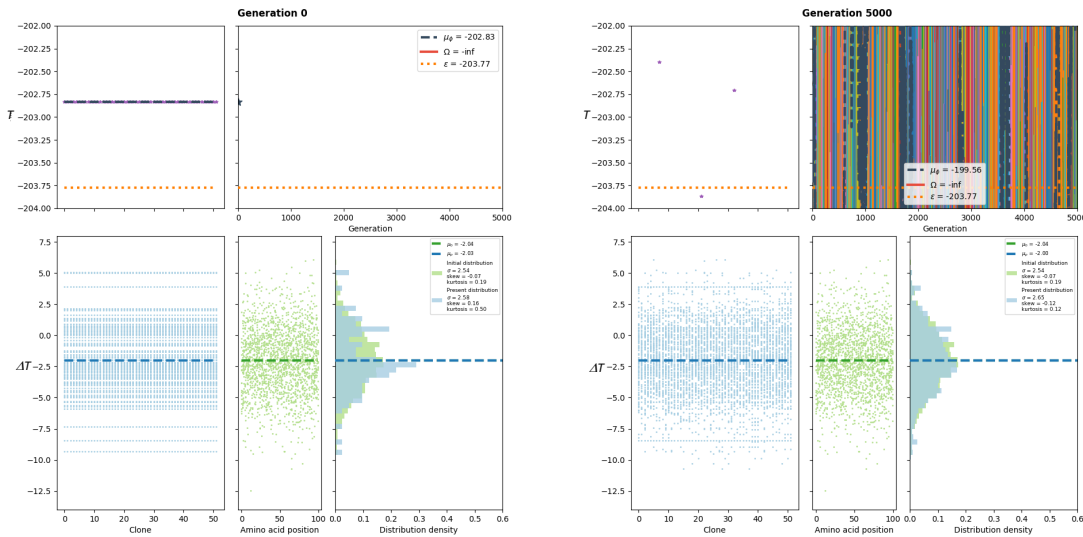
Gamma distribution sampling implemented into PESST. Across a protein, mutation rates are found to fit a gamma distribution, which can be simplified for computational purposes without cost to accuracy by assigning residues a mutation rate from one of four independent gamma rate categories. Rate categories are calculated as the median value of quartiles from 10,000 random samples of a gamma distribution.

Supplementary figure 18

A



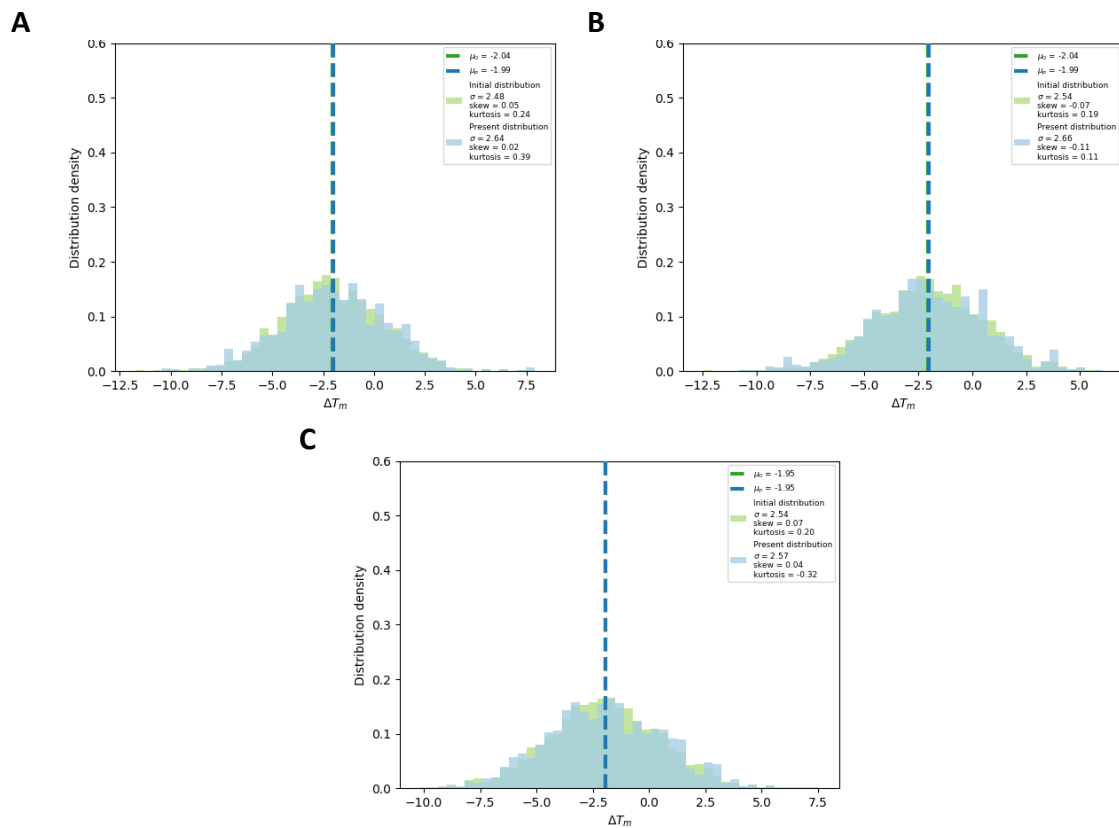
B



Supplementary 18 - Summary data for runs presented in figure 1B and C

Data represents the evolutionary simulation of figures 19B and 19C, where runs were initiated at T_0^{low} (A), $T_0^{\epsilon+20}$ (B). Figures are representative simulations showing comparison of the $\Delta_{r,a}$ states defined by the Δ matrix, to the $\Delta_{r,a}$ state of individual sites at generation 0 and at generation 5000. Data for all simulations is in supplementary files. In each figure, boxes show: **Top-Left:** mean stability of each clone in the dataset at generation. **Top-right:** progression of mean stability until a given generation. **Bottom-left:** the distribution of every $\Delta_{r,a}$ value represented in every clone. **Bottom-centre:** the global distribution of $\Delta_{r,a}$ values showing every possible $\Delta_{r,a}$ value at every position in the protein. **Bottom-left:** Overlay of the $\Delta_{r,a}$ distributions produced by both matrices. In the top boxes, orange dotted lines represent the value of ϵ derived from the global $\Delta_{r,a}$ matrix. In the bottom boxes, the dashed coloured lines represent the distribution average. At generation 5000, both $\Delta_{r,a}$ distributions and their means have converged. Note: a bug in PESST V1.0 leads to a zoomed in rendering of the stability trace based on issues with coding “catch-all” white-space buffering.

Supplementary figure 19

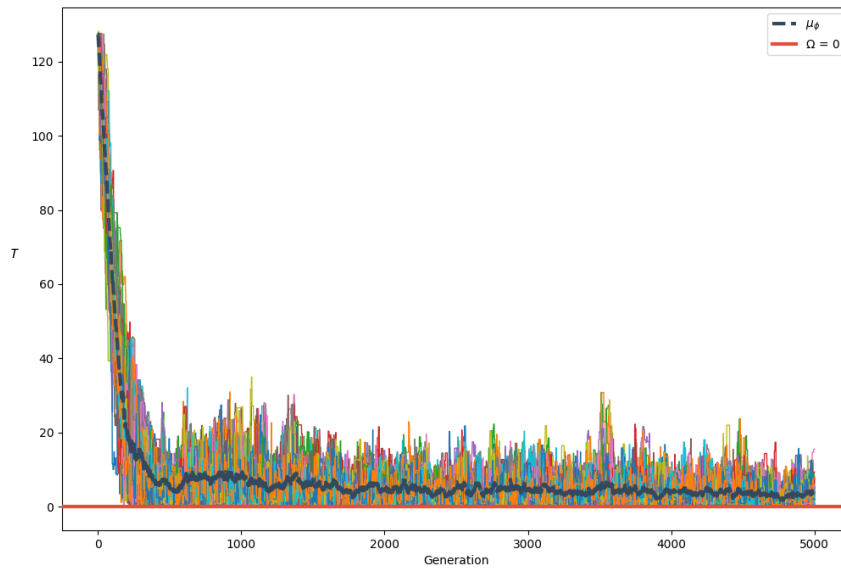


Supplementary figure 19 - During simulated evolution, the distribution of $\Delta_{r,a}$ values in evolving data approaches the distribution of $\Delta_{r,a}$ values in the stability matrix

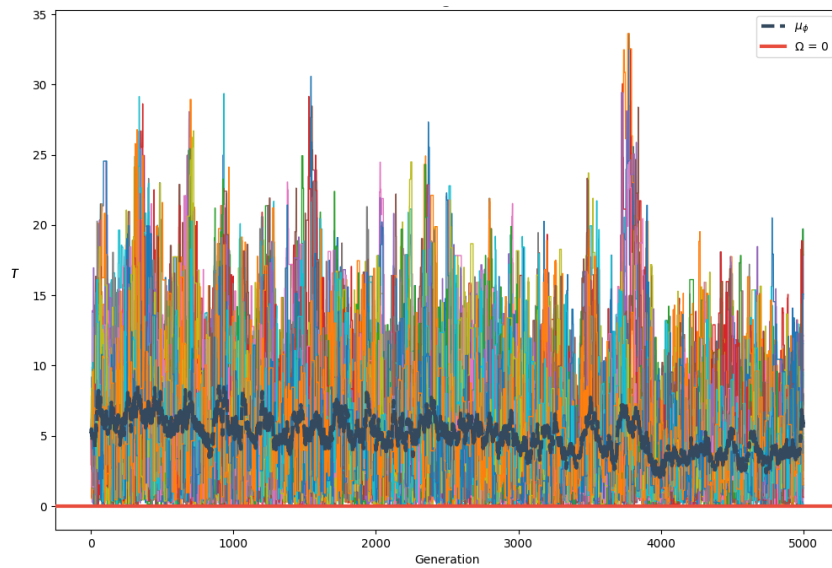
Data represents comparison of the distribution of evolving $\Delta_{r,a}$ values with the distribution of $\Delta_{r,a}$ values defined by the global stability matrix. Figures are representative the state at generation 5000 from simulations initiated at T_0^{high} (A), $T_0^{\epsilon+20}$ (B), or T_0^{low} (C). The mean and shape of the evolving distribution (Blue), approximates the mean and shape of the stability matrix (Green). Data for all other simulations with such parameters are in supplementary files.

Supplementary figure 20

A



B

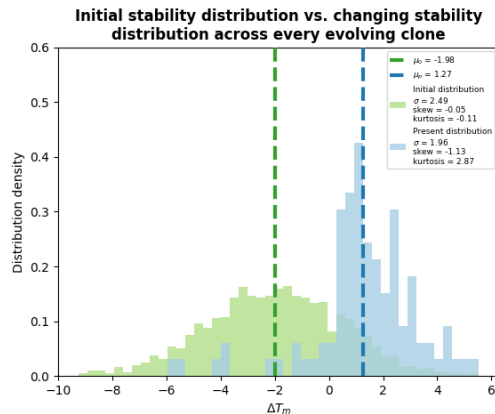


Supplementary figure 20 - Simulations with imposed Ω reach equilibrium slightly above Ω

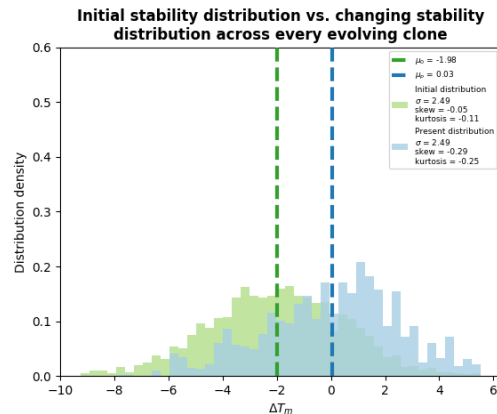
Representative stability traces for PESST simulations of 5000 generations where $\overline{\Delta_{r,a}} = -2$, $p_m = 0.002$, and $\Omega = 0$, showing that the stability of PESST simulated protein populations reach and maintain an equilibrium that approximates a value slightly above Ω . Simulations were initialised at T^{high} (A), or $T^{\Omega+5;\Omega+25}$ (B). In each graph, coloured lines represent the stability of one of 52 clones in the dataset, which are each tracked independently and simultaneously by PESST. The solid red line represents Ω . The tight dashed black bold line represents the average stability of the population. Data for all other simulations under such parameters are in supplementary files.

Supplementary figure 21

A



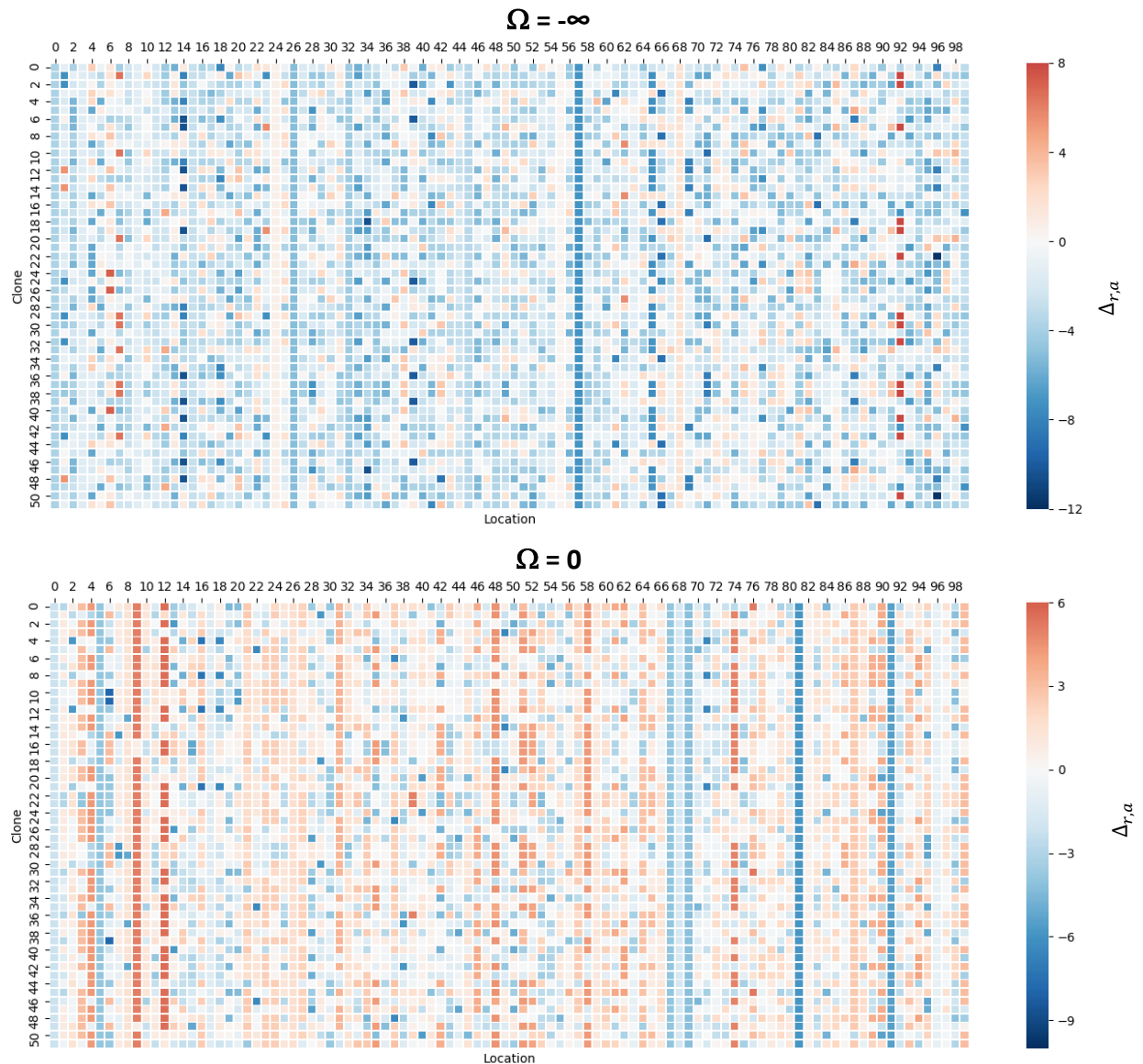
B



Supplementary figure 21 - During simulated evolution where $\Omega > \epsilon$, the distribution of $\Delta_{r,a}$ values in evolving data is positively biased compared to the distribution of $\Delta_{r,a}$ values in the global stability matrix Δ

Data is a representative comparison of the distribution of evolving $\Delta_{r,a}$ values with the distribution of $\Delta_{r,a}$ values defined by the global stability matrix. Figures are representative the state at generation 0 (A), and generation 5000 (B) from simulations initiated at T^{high} with a $\overline{\Delta_{r,a}} = -2$, $p_m = 0.002$, and $\Omega = 0$. The mean of the evolving distribution (Blue), is positively biased compared to the mean of the stability matrix (Green), showing that $\Delta_{r,a}$ values providing high stability increases to an evolving dataset are overrepresented with respect to their derivative matrix, whereas $\Delta_{r,a}$ values providing stability decreases are underrepresented. Figure generated using the Matplotlib library in Python. Data for all other simulations with such parameters are in supplementary files.

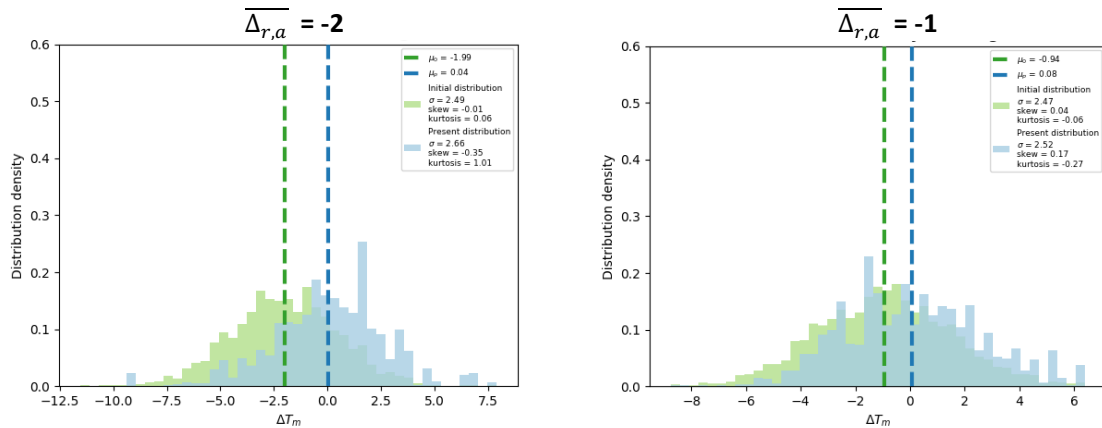
Supplementary figure 22



Supplementary figure 22 - Comparing heat maps of $\Delta_{r,a}$ values in simulations that have no Ω imposed to simulations with imposed Ω clearly show the positive shifted bias toward stability under Ω

Heat map is representative comparison of the effects of Ω on the distribution of evolving $\Delta_{r,a}$ values from simulations initiated at T_0^{high} with $\overline{\Delta_{r,a}} = -2$, $p_m = 0.002$ at generation 5000. Populations that have evolved under a selective pressure imposed by Ω have considerably more stabilizing residues than populations released from the selective pressure imposed by Ω (where $\Omega = -\infty$). Stabilizing residues are therefore overrepresented in populations evolving under marginality. Data for all other simulations with such parameters are in supplementary files.

Supplementary figure 23

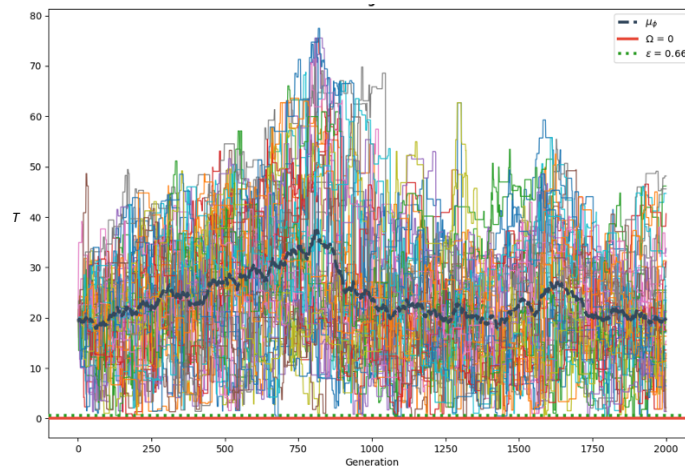


Supplementary figure 23 - Distribution of $\Delta_{r,a}$ values in evolving data does not converge with $\Delta_{r,a}$ values in global stability matrix Δ when evolved at marginality

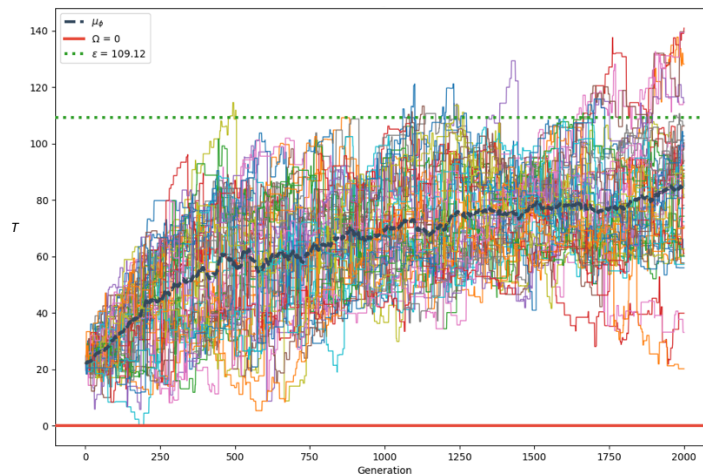
Data is a representative comparison of the effects differing global $\Delta_{r,a}$ matrix means (green) have on the positive bias in the distribution of evolving $\Delta_{r,a}$ values (blue) from simulations initiated at $T_0^{\Omega+5; \Omega+25}$, where $p_m = 0.002$, and $\Omega = 0$. In both instances, the mean of the evolving distribution (Blue), is positively biased compared to the mean of the stability matrix (Green), showing that $\Delta_{r,a}$ values providing high stability increases to an evolving dataset are overrepresented with respect to their derivative matrix, whereas $\Delta_{r,a}$ values providing stability decreases are underrepresented. However, where $\overline{\Delta_{r,a}} = -2$, this bias is more extreme compared to where $\overline{\Delta_{r,a}} = -1$, with a positive shifting of the $\Delta_{r,a}$ distribution in the evolving dataset. Figures were generated using the Matplotlib library in Python. Data for all other simulations with such parameters are in supplementary files.

Supplementary figure 24

A)



B)



Supplementary figure 24 - Simulations where global stability matrix $\overline{\Delta_{r,a}} = 0$ and $\overline{\Delta_{r,a}} = 1$ are able to explore a wide stability space due to release of selective pressure

Representative stability traces for PESST simulations of 2000 generations initiated at

$T_0^{\Omega+5; \Omega+25}$, where $p_m = 0.002$, and $\Omega = 0$, showing that the stability of PESST simulated protein

populations where $\overline{\Delta_{r,a}} = 0$ (A) and $\overline{\Delta_{r,a}} = 1$ (B) are released from the bidirectional selective

pressure as $\epsilon > \Omega$. The population is able to broadly sample stability space in these instances. In

each graph, coloured lines represent the stability of one of 52 clones in the dataset, which are each

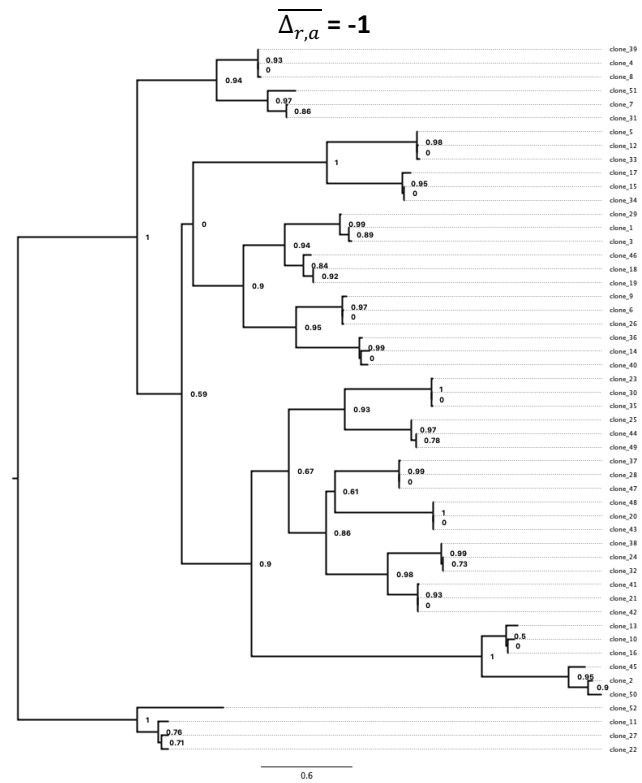
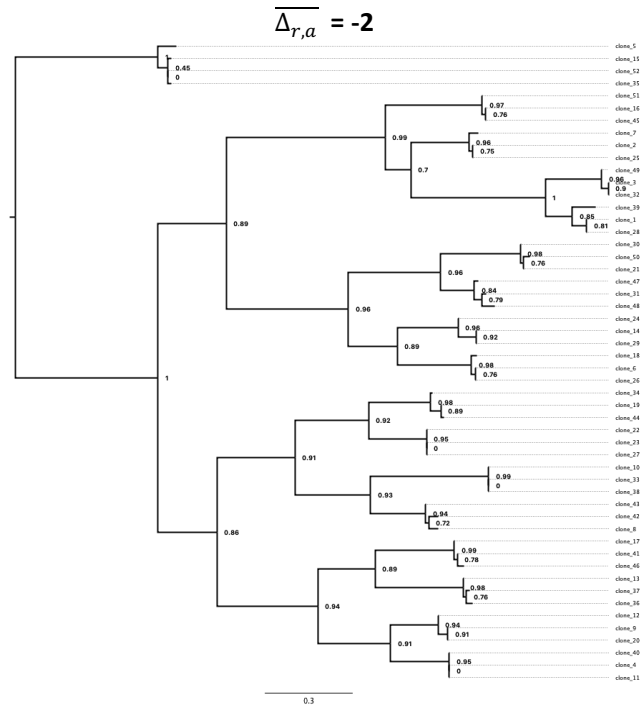
tracked independently and simultaneously by PESST. The solid red line represents Ω . The tight

dashed black bold line represents the average stability of the population. Figures generated using

the Matplotlib library for Python. Data for all other simulations under such parameters are in

supplementary files.

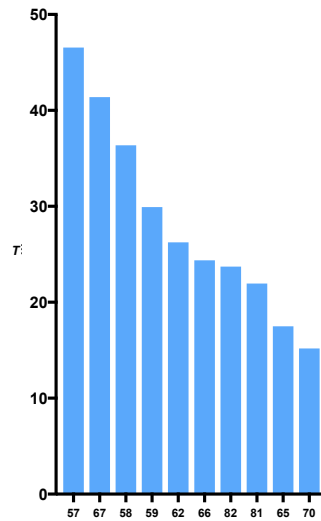
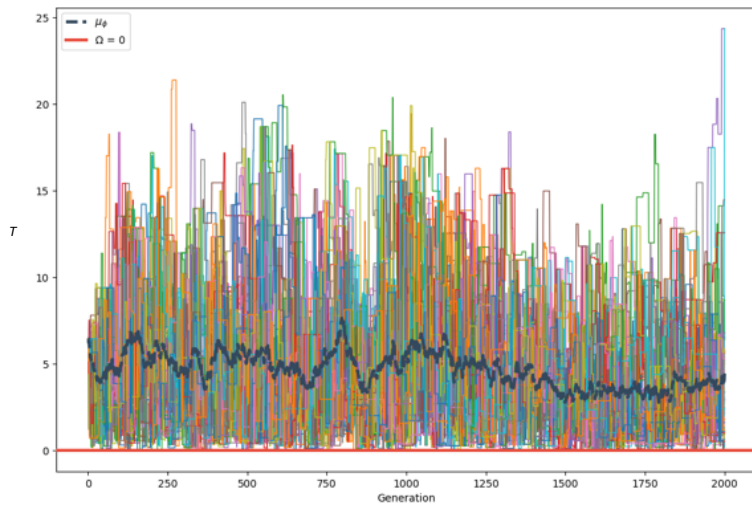
Supplementary figure 25



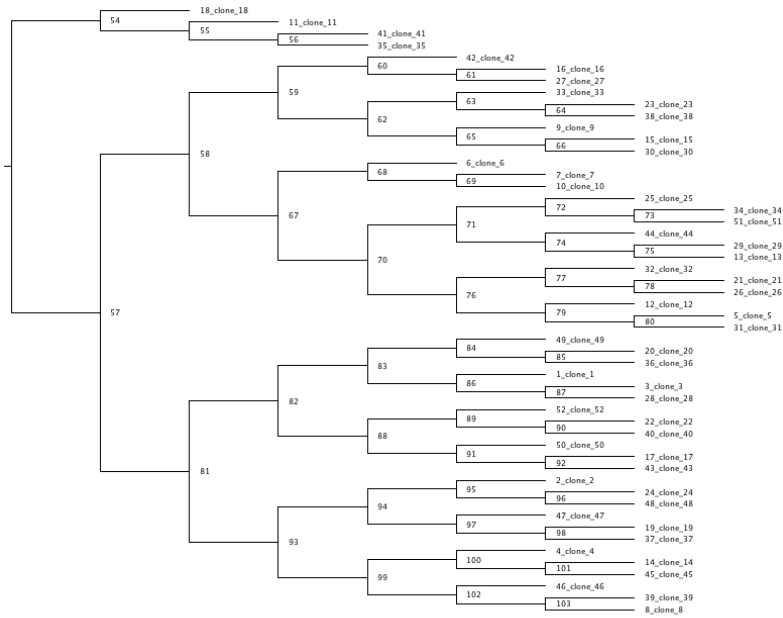
Supplementary figure 26

A

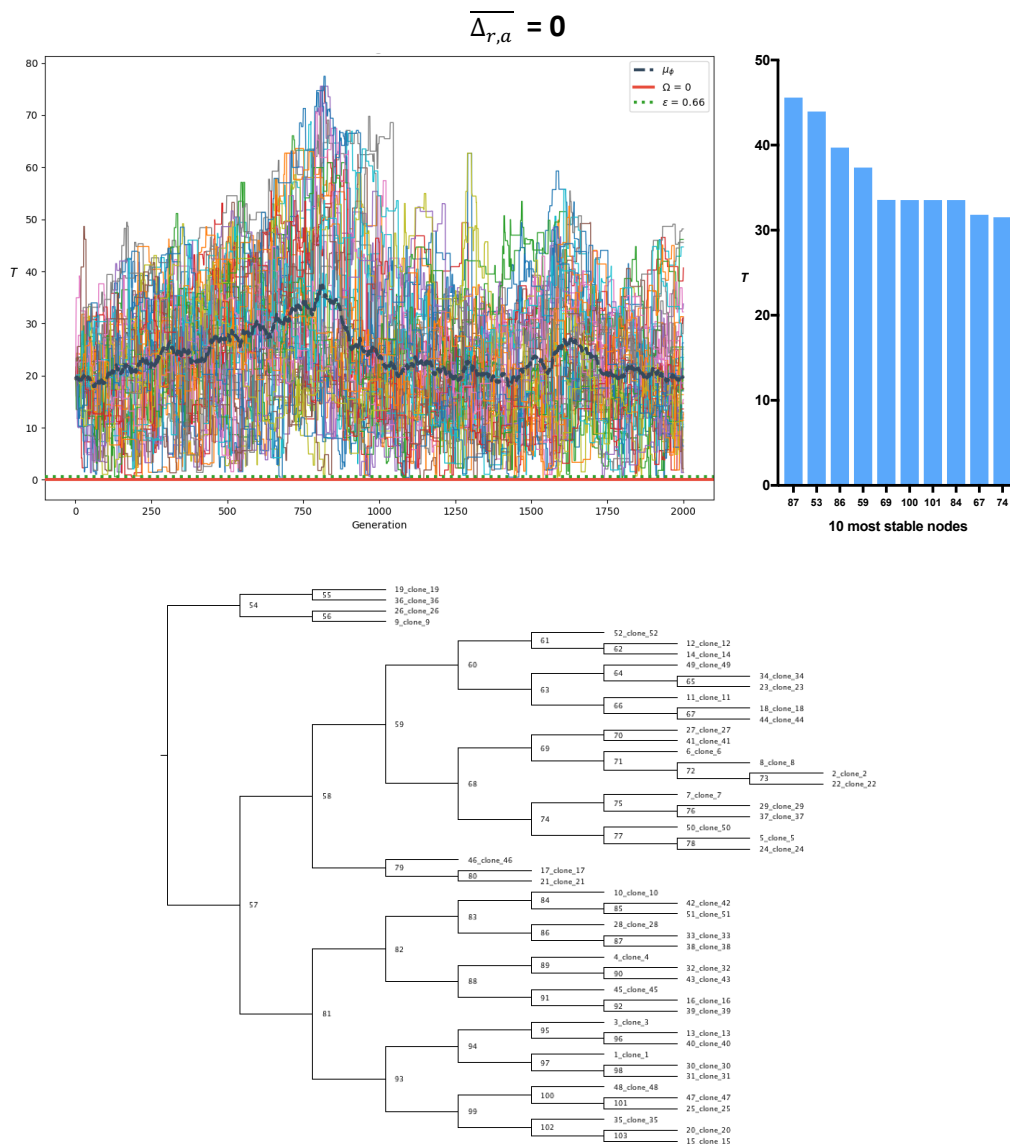
$$\overline{\Delta_{r,a}} = -2$$



10 most stable nodes



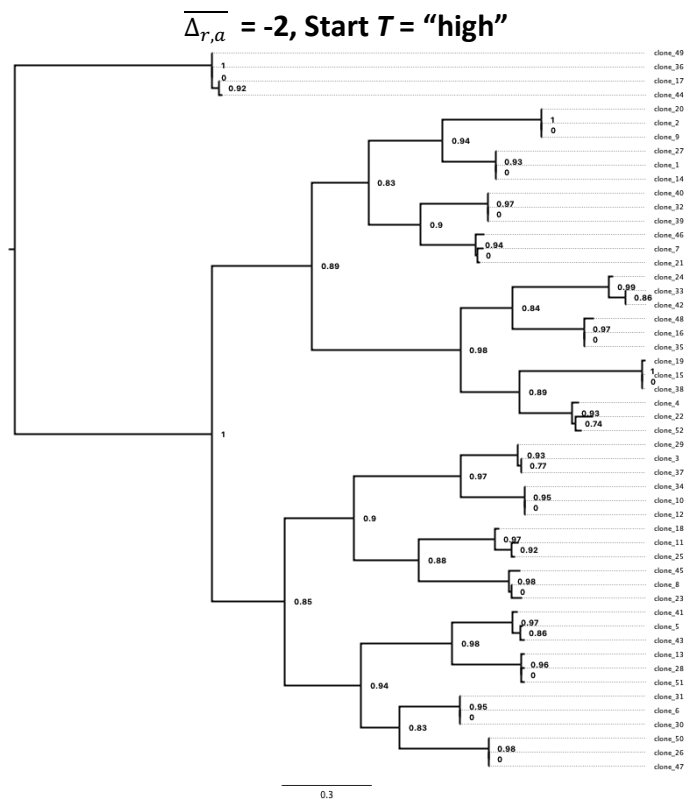
B



Supplementary figure 26 - Representative simulations showing that reconstructed ancestors explore a higher stability space than that of their evolutionary history when a bidirectional selective pressure is present

Data representative of ancestral sequences generated from PESST simulations of 2000 generations initiated at $T_0^{\Omega+5;\Omega+25}$, where $p_m = 0.002$, $\Omega = 0$, and $\overline{\Delta_{r,a}} = -2$, (A) or $\overline{\Delta_{r,a}} = 0$ (B). **Top left:** Stability trace from the representative simulation. **Top right:** Stability of the 10 most stable ancestral nodes derived from a given simulation calculated with CodeML in PAML (Yang, 2007). **Bottom:** Cladogram output by PAML visualised in FigTree. Node labels represent the ancestor sequence identifier produced by PAML. Ancestral reconstruction of simulations evolved at marginality caused by a bidirectional selective pressure leads proteins with stabilities that are higher than any stability sampled by the evolving population. This effect does not occur when the bidirectional selective pressure is released.

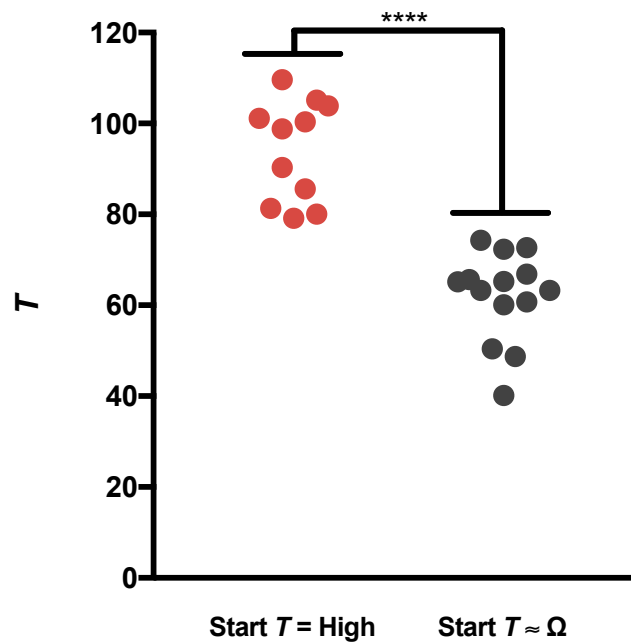
Supplementary figure 27



Supplementary figure 27 - Phylogeny for simulation that evolves from high stability to a threshold

Alignments generated by PESST simulations initiated at T_0^{high} , where $p_m = 0.002$, $\Omega = 0$, and $\overline{\Delta_{r,a}} = -2$ were exported into Geneious ver. 10 (Kearse *et al.*, 2012). Phylogeny of simulated evolution was generated with PhyML (Guindon *et al.*, 2010). Data presented is representative of a single simulation under the given conditions. Data for all simulations is available in supplementary files. Phylogeny was modified in FigTree. Node values represent SH-like support values calculated in PhyML. Scale represents mutations per site.

Supplementary figure 28



Supplementary figure 28 - Consensus sequences from simulations initiated at T_0^{high} are significantly more stable than those from $T_0^{\Omega+5; \Omega+25}$

Grouped min-max scatter of the stabilities of consensus values derived from the five simulations of both scenarios where $\overline{\Delta_{r,a}} = -2$, comparing the difference in stabilization when sequences evolve from a starting point of high stability (red points), or evolve at marginality their entire history (grey points). Mann Whitney U test was used to calculate whether the populations were significantly dissimilar ($p < 0.05$). Asterisks represent the degrees of significance (**** = $p < 0.00005$; *** = $p < 0.0005$; ** = $p < 0.005$; * = $p < 0.05$).

Supplementary table 4

Parameter	Symbol	Notes
Population of proteins	Φ	$\{\eta_1, \eta_2, \dots, \eta_N\}$
Protein sequence	η	$[a_1, a_2, \dots, a_R]$
Matrix of transition probabilities	\mathbf{L}	$L_{a,a'} := p(a \rightarrow a'); a \neq a'$
Matrix of amino acid stabilities	Δ	
Amino acid stability contribution	$\Delta_{r,a}$	
Protein stability	T	$:= \sum_{r=1}^R \Delta_{r,a_r}$
Mean protein stability	\bar{T}_Φ	$:= \frac{1}{N} \sum_{n=1}^N T_n$
Vector of mutation probabilities	\mathbf{m}	$m_r \sim \Gamma(\kappa, \theta); \sum_{r=1}^R m_r = 1$
Evolutionary history	h_Φ	$[\Phi_1, \Phi_2, \dots, \Phi_G]$
Expected stability convergence	ϵ	$\approx \mu \cdot R$
Population of roots	Φ_{roots}	$\{\eta_1, \eta_2, \dots, \eta_{n_{roots}}\}$
Population of branches	$\Phi_{branches}$	$\{\{\eta_i\}_1, \{\eta_j\}_2, \dots, \{\eta_k\}_{n_{branches}}\}$

Supplementary table 4 - Additional parameters handled by PESST

Supplementary table 5

Starting parameter	Simulation	Seed
T^{high}	1	1066467549
	2	4067081442
	3	3988311094
	4	2564259790
	5	1378044692
T^{low}	1	2996085632
	2	4193085486
	3	1919944244
	4	3809257325
	5	3123945229
$T^{\epsilon \pm 20}$	1	1570260454
	2	657585481
	3	2411380004
	4	1411240584
	5	2496233140

Supplementary table 5 - Seeds for the simulations used for figures 19 and 20

PESST implements a random number generator to define various parameters so each simulation is a unique evolutionary scenario. To ensure that PESST simulations are replicable, the random number generator can be seeded. The seeds used for the simulations that make up the results in figures 19 and 20 are presented.

Supplementary table 6

Starting parameter	Simulation	Seed
$T^{high}; \Omega = 0$	1	1137221125
	2	3538885643
	3	1741741204
	4	2846674525
	5	3126636696
$T^{\Omega+5;\Omega+25}; \Omega = 0$	1	1773503249
	2	3170351635
	3	4209518524
	4	162434345
	5	3890946287

Supplementary table 6 - Seeds for the simulations used for figure 21

Supplementary table 7

Starting parameter	Simulation	Seed
$\overline{\Delta_{r,a}} = -2$	1	1585917709
	2	4027632114
	3	526607129
	4	1222718357
	5	2606164749
$\overline{\Delta_{r,a}} = -1$	1	2887861297
	2	3274037544
	3	2779731581
	4	743133157
	5	3977220919
$\overline{\Delta_{r,a}} = 0$	1	110209255
	2	4157235131
	3	1982327043
	4	1989687036
	5	1168569330
$\overline{\Delta_{r,a}} = 1$	1	3813657984
	2	785464822
	3	2631411974
	4	4151143046
	5	2049625170
$\overline{\Delta_{r,a}} = -2;$ T^{high}	1	2229567065
	2	2510842188
	3	2535624228
	4	1401832132
	5	2624859900

Supplementary table 7 - Seeds for figures 22 and 23

Supplementary table 8

$\overline{\Delta}_{r,a}$	-2		-1		0		1	
Test	Mann-Whitney U (U, p)	Welch's t-test (t, p)	Mann-Whitney U (U, p)	Welch's t-test (t, p)	Mann-Whitney U (p)	Welch's t-test (p)	Mann-Whitney U (p)	Welch's t-test (p)
Simulation 1	879, 0.0029	3.67, 0.0005	950, 0.0126	2.99, 0.0039	1145, 0.2329	0.98, 0.3285	1133, 0.2045	1.13, 0.2596
Simulation 2	888, 0.0036	3.82, 0.0004	962, 0.0158	3.82, 0.0003	1133, 0.2033	1.11, 0.2664	1200, 0.4065	0.61, 0.5454
Simulation 3	837, 0.0011	3.89, 0.0003	911, 0.0059	3.31, 0.0015	1155, 0.2612	0.96, 0.3419	1203, 0.4197	0.49, 0.4865
Simulation 4	789, 0.0003	4.16, 0.0001	855, 0.0017	3.38, 0.0013	1179, 0.3328	1.06, 0.2907	1138, 0.2153	1.68, 0.0973
Simulation 5	948, 0.0121	2.94, 0.0044	740, <0.0001	4.25, <0.0001	1056, 0.0751	2.18, 0.0319	1075, 0.0976	2.01, 0.0471

Supplementary table 8 - Simulation p-values in figure 22

Table represents statistics and confidence values (p -values) obtained with the Mann-Whitney U test^{††} or Welch's t-test when comparing the similarity of the distribution of comparing the stability values of proteins generation 2000 to PAML calculated ancestors of the simulation. PESST simulations were initiated at $T_0^{\Omega+5; \Omega+25}$, where $p_m = 0.002$, $\Omega = 0$, and $\overline{\Delta}_{r,a} = -2, -1, 0$ and 1 . Distributions of stability values that are not significantly different ($p > 0.05$) are coloured red.

3.11 Supporting information

Acknowledgements

AT and NH acknowledge the generous support from the BBSRC SWBio Doctoral Training Program. BDE acknowledges that this work was generously supported by the Wellcome Trust Institutional Strategic Support Award (204909/Z/16/Z). We would also like to thank Dr. Guy Leonard for many early discussions about required PESST features, and methods to implement bifurcation. Finally, we would also like to thank the Harmer group for their constant support.

Author Information

The authors declare no competing interests.

^{††} For all tests, $U_{Max} = 2652$. U_{max} is defined as n_1n_2

Chapter 4

*Simplified ancestral sequence
reconstruction – an accessible tool for
engineering protein stability.*

4.1 Authors

Adam Thomas^{1,2}, Benjamin D. Evans^{1,3}, Mark van der Giezen², Nicholas J. Harmer^{1,2}.

1. Living Systems Institute, Stocker Road, Exeter EX4 4QD, U.K.
2. Department of Biosciences, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, U.K.
3. Centre for Biomedical Modelling and Analysis, Stocker Road, Exeter EX4 4QD, U.K.

4.2 Preface

The following chapter consists a reformatted manuscript for an article written for submission to eLife. This article will be submitted simultaneously with the manuscript in chapter 4, as it shows direct application of PESST. In this chapter, we design and undertake a simplified version of ASR developed to be a broadly accessible engineering tool. We use PESST to directly inform decisions in the design process. PESST generates large volumes of data, therefore not all of the data generated to inform this study is included. However, This data can be accessed in the supplementary files on the flash drive attached to this thesis, and from Nicholas Harmer on request.

AT and NH conceived the study. AT wrote the manuscript, with input from NH. All authors edited the manuscript. BDE wrote the version of PESST with which the data in this chapter is derived. AT performed the simulations of protein evolution and wet lab experiments with input from NH.

4.3 Abstract

Ancestral Sequence Reconstruction (ASR) promises simple, low cost engineering of protein stability. However, ASR still has barriers for access that limit its use amongst scientific communities, largely the requirement of expertise in phylogenetics. Recent studies have used models of protein engineering to uncover the underlying forces driving stabilization in ASR experiments. Here, we utilize these models to develop a highly-simplified ASR (sASR) methodology based on the Ancescon algorithm that allows for the rapid engineering of stable protein sequences in ancestral space without the need for an input phylogeny. Using

models simulating protein evolution, we developed criteria for sampling sequences from Ancestron reconstructions based on Ancestron-derived phylogenetic outputs. Ancestron produced dramatically stabilized nodes scattered throughout output trees. We subsequently developed criteria for predicting stable nodes from the consortium, and validated sASR by reconstructing variants of the dynamically complex carboxylic acid reductase (CAR) enzyme. Two sequences from CAR's ancestral space were obtained, both of which could be expressed, and were functional and thermostable. One of the two enzymes represented the most thermostable CAR observed to date, with a T_m of 74 °C, an up to 9 °C increase over previously reconstructed thermostable CAR ancestors. This work provides a straightforward method for constructing proteins with novel properties, allowing non-experts to access thermostable proteins through a truly democratized engineering tool.

We're now thirty years into biotechnology. Are we ever going to get to the point where it's not an exclusive technology, it's not a technology that requires experts?

-Drew Endy (*Edge*, 2008)

4.4 Introduction

There is a strong community desire, especially in biocatalysis and biotechnology, to have access to protein variants with increased stability (Rigoldi *et al.*, 2018; Elleuche *et al.*, 2014). Stable enzymes provide several advantages for the biotechnologist's toolbox. They reduce the risk of biocontamination in energy-rich fermentation environments (Akram *et al.*, 2018). Reactions with stable enzymes can have increased rates over their mesophilic counterparts in accordance with the Arrhenius equation (Lin and Xu, 2013). Additionally, high temperature reaction conditions allow for improved substrate and product solubility, and maintain reactivity in the challenging environments imposed by increased solute concentrations (Elleuche *et al.*, 2014; Tavanti *et al.*, 2017). Furthermore, stable proteins typically possess increased lifetime productivity per enzyme (Akram *et al.*, 2018). For biopharmaceuticals, longer *in corpus* half-lives facilitate optimization of viable and effective dose-response relationships (Zakas *et al.*, 2015; Zakas *et al.*, 2017).

Engineering proteins for stability is an imprecise process that has been described as “one of the most challenging problems in protein science” (Suplatov *et al.*, 2015). Contemporary tools available for engineering protein stability include high throughput directed evolution, focused *in silico* design, and machine learning based design (Arnold, 2018; Fürst *et al.*, 2018; Wijma *et al.*, 2018; Rigoldi *et al.*, 2018; Bendl *et al.*, 2016). Such technologies are a long way from the vision of democratized biotechnology, requiring considerable expertise, high expenditure and/or access to high throughput screening tools to achieve success. Even when stabilization is achieved, results are still often unsatisfactory, with increases of only a few degrees achieved (Rigoldi *et al.*, 2018; Yu *et al.*, 2017).

Existing methods to engineer stability strive against the bidirectional selective pressure that forced natural proteins toward marginal stability (Suplatov *et al.*, 2015; Chapter 3). There are currently no generic rules to guide the rational engineering of stability. Attempts to define applicable global first principles of protein stability based on natural thermostable enzymes have also failed (Okafor *et al.*, 2018; Rigoldi *et al.*, 2018). As a result, each stability engineering attempt is resource intensive, without a guarantee of success. Clearly there is a strong community requirement for accessible, democratized technologies that make the

engineering of enzymes routine (Hughes and Ellington, 2017; Tachioka *et al.*, 2016).

Democratized protein engineering tools would require only a modest laboratory set up and investment to deliver proteins with desirable properties. Such enabling technologies should have wide accessibility to facilitate broad-scale and open innovation (Endy, 2005; Jefferson *et al.*, 2014; Frow, 2015; Frow, 2017).

In recent years, a body of research has emerged exploring ancestral protein reconstruction, a “bias”-based protein engineering method, for imparting beneficial stabilizing properties into proteins (Gumulya *et al.*, 2018; Okafor *et al.*, 2018; Durani and Magliery, 2013; Kiss *et al.*, 2009; Chapter 2; Chapter 3). ASR is a computational tool designed to predict a protein family’s evolutionary history (Hochberg and Thornton, 2017; Akanuma, 2017). The core prerequisites are a multiple sequence alignment, a phylogeny of this alignment, and a model of amino acid substitution. Ancestral sequences are calculated for each node of the phylogenetic tree using a Bayesian approach that maximises the likelihood of the sequence of each node across sequence space based on a model of amino acid substitution and the given alignment (Yang, 2007; Joy *et al.*, 2016). These predictions provide considerable insight into how a protein’s form, function and specificity might have evolved (Siddiq *et al.*, 2017; Hochberg and Thornton, 2017).

Importantly, ASR studies consistently report that ancient enzymes exhibit thermostability, a trait often hypothesised to derive directly from the protein’s evolutionary history (Akanuma, 2017). This thermostable-biasing property has been successfully co-opted as an engineering tool. Numerous studies have been published in over recent years showing increases in the stability of commercially important enzymes (table 11). Recently, we reported that ASR is a diverse engineering tool, enabling the shotgun-like engineering of large, highly complex proteins towards stable properties and novel substrate ranges (Chapter 2). This work also corroborated research on EF-Tu proteins by Okafor *et al.* (2018), showing ASR is suitable for engineering multi-domain enzymes with high dynamic complexity. To our knowledge this is one of few low-cost engineering options available for such enzymes. We showcased this method by resurrecting an early ancestor of the bacterial CAR1 subfamily of carboxylic acid reductases (CARs; E.C. 1.2.1.30; Chapter 2; Stolterfoht *et al.*, 2017). CARs are large (>1,200 amino acid) and complex (three domain, three reaction) enzymes, and so provide a

challenging test for ASR. Ancestral CARs showed an up to 34 °C increase in T_m compared to extant enzymes, with up to quintupled half-lives at bioindustrially relevant temperatures compared to extant CARs (Finnigan *et al.*, 2017). Other industrially useful, and novel CAR properties were also reported in the property-rich ancestral space of the CARs, including solvent, pH, and salt tolerance, as well as improved substrate turnovers.

Protein	Industrial importance	ASR method	Reported stabilization	Reference
Arginine binding protein	Biosensor used in FRET experiments	PAML	T_m increased 30 °C in <i>in vitro</i> conditions	Whitfield <i>et al.</i> , 2015
Carboxylic acid reductase	Conversion of over 100 carboxylic acid compounds to aldehydes. Production of pharmaceutical, flavour and scent compounds	Ancescon, FastML, PAML	T_m increased 16-34 °C in <i>in vitro</i> conditions; 5-29 °C in model <i>in vivo</i> conditions. Half-life doubled*. pH range of 5 to 10.	Chapter 2
Coagulation factor VIII	Blood clotting protein drug for haemophilia	PAML	Half-life doubled	Zakas <i>et al.</i> , 2017
Haloalkane dehalogenase	Cleavage of carbon-halogen bond in halogenated aliphatic hydrocarbons. Pollutant remediation and hydrocarbon biosynthesis	PAML (Lazarus)	T_m increased 8-24 °C in <i>in vitro</i> conditions	Babkova <i>et al.</i> , 2017
ω -Transaminase	Transamination of various ω -amino acids and α,ω -diamines. Production of nylon-12	FastML	T_m increased 10 °C in <i>in vitro</i> conditions.	Wilding <i>et al.</i> , 2017
Cytochrome P450 monooxygenase (CYP3)	Regio and stereoselective oxidations of C-H bonds by reductive scission of molecular oxygen. Wide application in molecular functionalization.	FastML	~20-30 °C increase in T_{50} after 60 minute incubation.	Gumulya <i>et al.</i> , 2018
Class II ketol-acid reductoisomerase	Oxidoreductase acting on CH-OH donor groups. Synthesis of amino acids.	Unpublished Bayesian Network	15-17 °C increase in T_m in <i>in vitro</i> conditions.	Gumulya <i>et al.</i> , 2018

*Half-life at 37 °C compared to extant half-life calculated at 30 °C

Table 11 - Stability increases in commercially important proteins engineered by ASR, and their potential utilization in current bioindustrial workflows

To better understand the stabilizing effects of ASR, we recently developed a stochastic, constrainable model of sequence evolution called Protein Evolution Simulations with Stability Tracking (PESST; Chapter 3). Using this model, we presented strong evidence for the existence of “marginality bias” driving stabilization in ASR. Bi-directional selective pressure is imposed on a protein as most amino acid choices at each position are destabilizing, yet the protein must be stable at the organism’s operating temperature (Taverna and Goldstein, 2002; Tokuriki *et al.*, 2008; Goldstein, 2011). These pressures significantly titrate destabilizing residues from the evolving population, over-representing stabilizing residues. These effects cause the overestimation of ancestral thermostability even if the protein’s evolutionary history has never innovated such traits (Chapter 3). PESST also offers a tool with which to discover and test alternative methods to introduce biasing effects into a protein. This makes ASR a powerful protein engineering tool, especially as it requires far fewer resources than directed evolution or sequence guided mutagenesis. Nevertheless, it has a steep learning curve, and requires considerable time investment (Vialle *et al.*, 2018; Gumulya and Gillam, 2016).

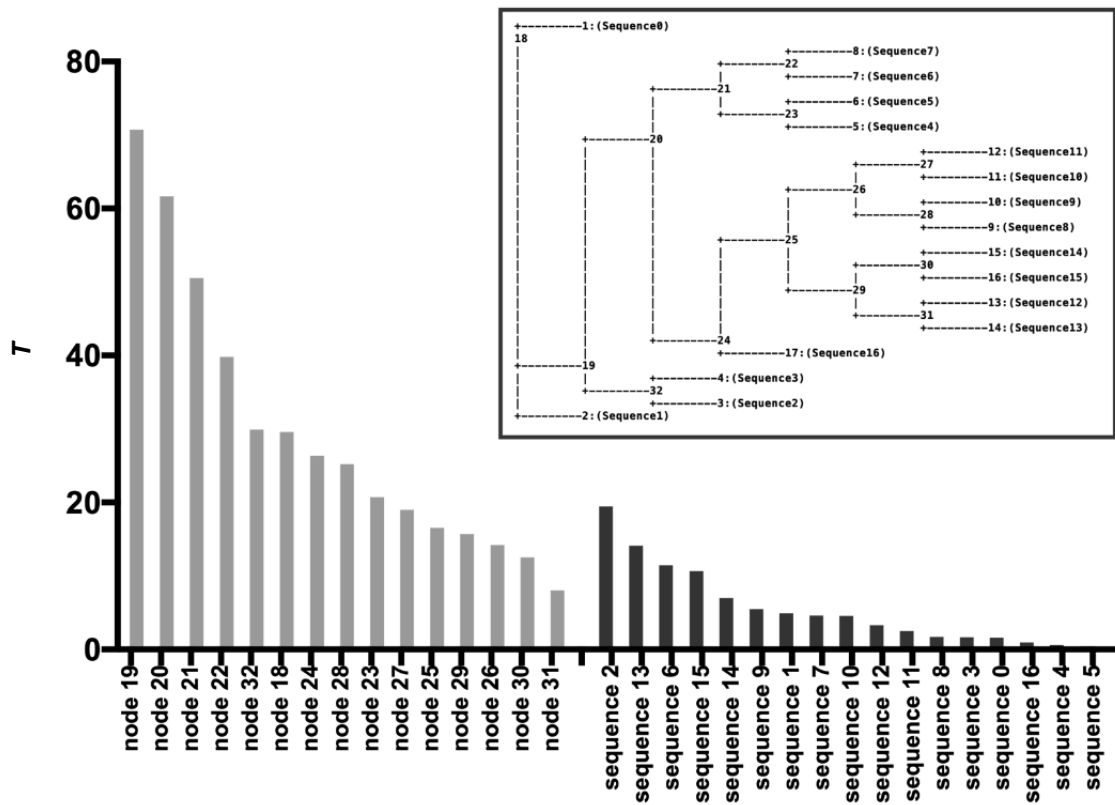
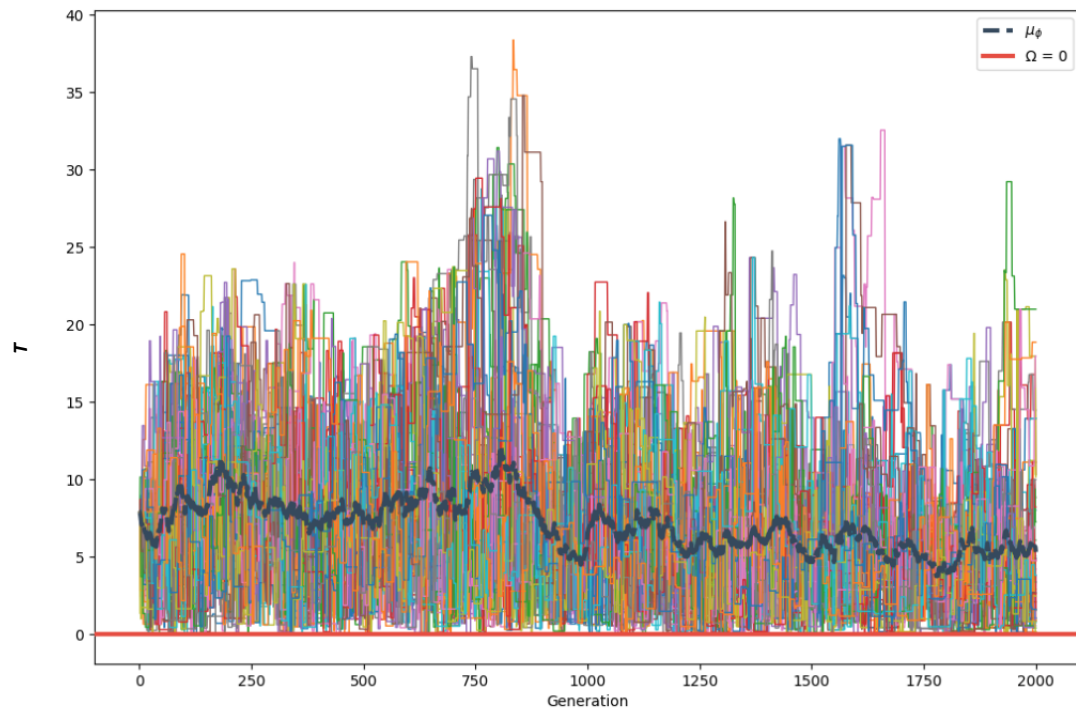
Here, we developed a simplified and accessible method for obtaining thermostable proteins from ancestral space. We based this method on Ancescon, an algorithm in the ASR suite with simple input requirements, requiring only a multiple sequence alignment to generate ancestral sequences (Cai *et al.*, 2004). Using PESST, we generated ten independent *in silico* phylogenies, and probed the stabilizing potential of the minimal Ancescon algorithm (hereafter dubbed simplified ASR [sASR]). 100% of sASR reconstructions engendered stabilized ancestors from simulations; however stable nodes were found scattered throughout each phylogeny. Using correlative analyses of stable ancestor positions within the Ancescon output trees, we derived rules for predicting which nodes in a phylogeny are likely to harbor stabilized proteins. To validate this strategy, we conducted sASR on a dataset of 42 CAR enzymes. Targeting the two most likely nodes to produce a stable ancestor generated two CAR-like enzymes with T_m values of approximately 57 °C and 74 °C in both *in vitro* and *in vivo*-analogue conditions. These enzymes show increases in stability between 10-40 °C in comparison to well-studied extant enzymes. Our results show that sASR can produce stable enzymes from a simple sequence alignment input. This work provides a new, accessible engineering option in the protein engineer’s toolbox.

4.5 Results

4.5.1 *Ancescon produces ancestors with high stability in simulations*

We previously showed that the predominantly-used ASR algorithm PAML will overestimate ancestor stability when a family evolves at marginality (Chapter 3). To assess whether sASR could be used for the engineering of protein thermostability when a protein family's evolutionary history had not explored stable sequences we simulated ten independent evolutionary histories with identical starting parameters with PESST (supplementary table 9). Simulations featured bifurcating populations of 52 proteins each with 100 amino acids, evolved under a constant bidirectional selective pressure (Stability traces of each simulation in supplementary figure 29). To simplify downstream analysis, PESST pruned leaves (sequences of the final generation) from each output alignment to a single sequence per bifurcative population, reducing the number of sequences per analysis from 52 to 17 in all simulations. Pruned alignments from the final generation in each simulation were then used to generate ancestral sequences using Ancescon (Zimmerman *et al.*, 2018). The stability of all ancestor nodes for each simulation was then calculated in PESST. Across all simulations, the most stable ancestors were between 1.4-fold and 2.1-fold more stable than the most stable state achieved by any corresponding clone throughout its evolutionary history (figure 24A, supplementary figure 30). If there was no biasing effect, nodes and leaves would sample an equivalent sequence space, and share an even representation of normalized global stability space. However, the division of normalized global stability space between the nodes and leaves of each simulation was significantly unevenly 84:16 between nodes:leaves respectively (binomial probability $p < 0.0001$; figure 24B). Ancescon therefore introduces considerable stabilizing bias into node sequences when deriving sequences from a population evolved under marginality.

A



B

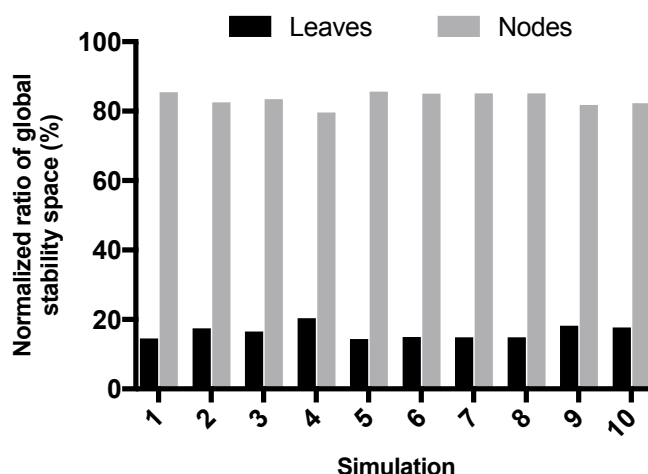


Figure 24 - Ancescon introduces stability bias in PESST simulations of protein evolution

A) Reconstructions of PESST simulations generate sequences with stability values higher than any value sampled by the population. **Upper:** trace showing the changing stability of each protein in the population over time. Each colored line represents the stability of a single protein as it evolves. Grey tight dashed line represents the average stability of all proteins in the population. Red line represents the stability threshold. Figure was rendered with Matplotlib in Python. **Lower:** The calculated stability of nodes reconstructed from the population with Ancescon, with the corresponding ASCII tree output by Ancescon produced by the weighbour method (Bruno *et al.*, 2000). **B)** The ratio of global stability space across the tree is largely shifted in favour of the nodes. Node stability in every simulation, was calculated by PESST. The global stability of all sequences in each simulation were normalized to 1. The ratio of space derived from each population were then reported. Data was analysed and visualized in Graphpad Prism v7.

4.5.2 High weighted node balance is a strong indicator of stability in simulations

For sASR to be a successful and accessible protein engineering tool, it is important that stable nodes can be consistently predicted within the output tree to minimize protein screening. The weighted neighbor-joining tree-building algorithm employed by Ancescon (Weighbour; Cai *et al.*, 2004; Bruno *et al.*, 2000) is an arguably poor measure of phylogeny, considering neither node age nor accuracy. We therefore assessed the distribution of highly stable nodes across each weighbour tree in the Ancescon output. Each ancestor's stability value was normalized to the most stable sequence in the population. A non-significant negative correlation ($r = -0.05$; $p = 0.94$) was observed when comparing stability and age,

suggesting that stable sequences are slightly more common toward the base of the tree (figure 25A). However, no simulation produced the most stable ancestor at the base of the tree, and only two of the ten simulations produced basal ancestors in the upper quartile of stabilities obtained.

Therefore, we sought improved strategies for stable node selection. As leaves are generally marginally stable, and posterior node probability is calculated from its set of leaves in Ancestron (Cai *et al.*, 2004), we hypothesized that single leaves could have a destabilizing effect on immediate parent nodes. We observed that nodes with only multi-leaved subtrees have significantly higher average stability than nodes with at least one single leaved subtree for seven out of the ten simulations (figure 25B; Welch's *t*-test; $p < 0.05$). For all simulations, the average stability of nodes that parent multi-leaf subtrees was above the 50th percentile. Comparatively, nodes with single leaf subtrees presented significantly fewer (10%) nodes with stabilities in the 50th percentile (binomial probability $p < 0.0001$). This strongly suggests that single leaves have a destabilizing effect on parent nodes in this approach. While selecting multi-leaf nodes may be a favorable strategy for most trees, it must be noted that across all simulations only 58% of nodes with multi-leaved subtrees still produced sequences with stabilities above the 75th percentile. Therefore, the synthesis of numerous candidate nodes would be required to ensure the desired stability is captured.

We then considered whether node stability could be predicted by a node's "weighted balance". Weighted balance describes how evenly leaves are distributed around the node. We hypothesized that the weight of the leaves around a node would impact its stability (i.e. more leaves provide more sequences from which to select stabilizing ancestral residues from; see methods for formula). We observed a strong significant correlation between weighted balance and node stability (figure 25C; $r = 0.42$; $p > 0.0001$). Additionally, when the two nodes with the highest weighted balance were selected for each simulation, nineteen of the twenty-one nodes had stabilities above the 75th percentile of all nodes; and at least one node per simulation was above the 75th percentile. These nodes were all more stable than any stability sampled throughout the simulation's histories. Given these data, weighted node balance provides a significantly strong predictive tool to identify highly stable proteins from Ancestron derived phylogenies (binomial probability: $p < 0.0001$).

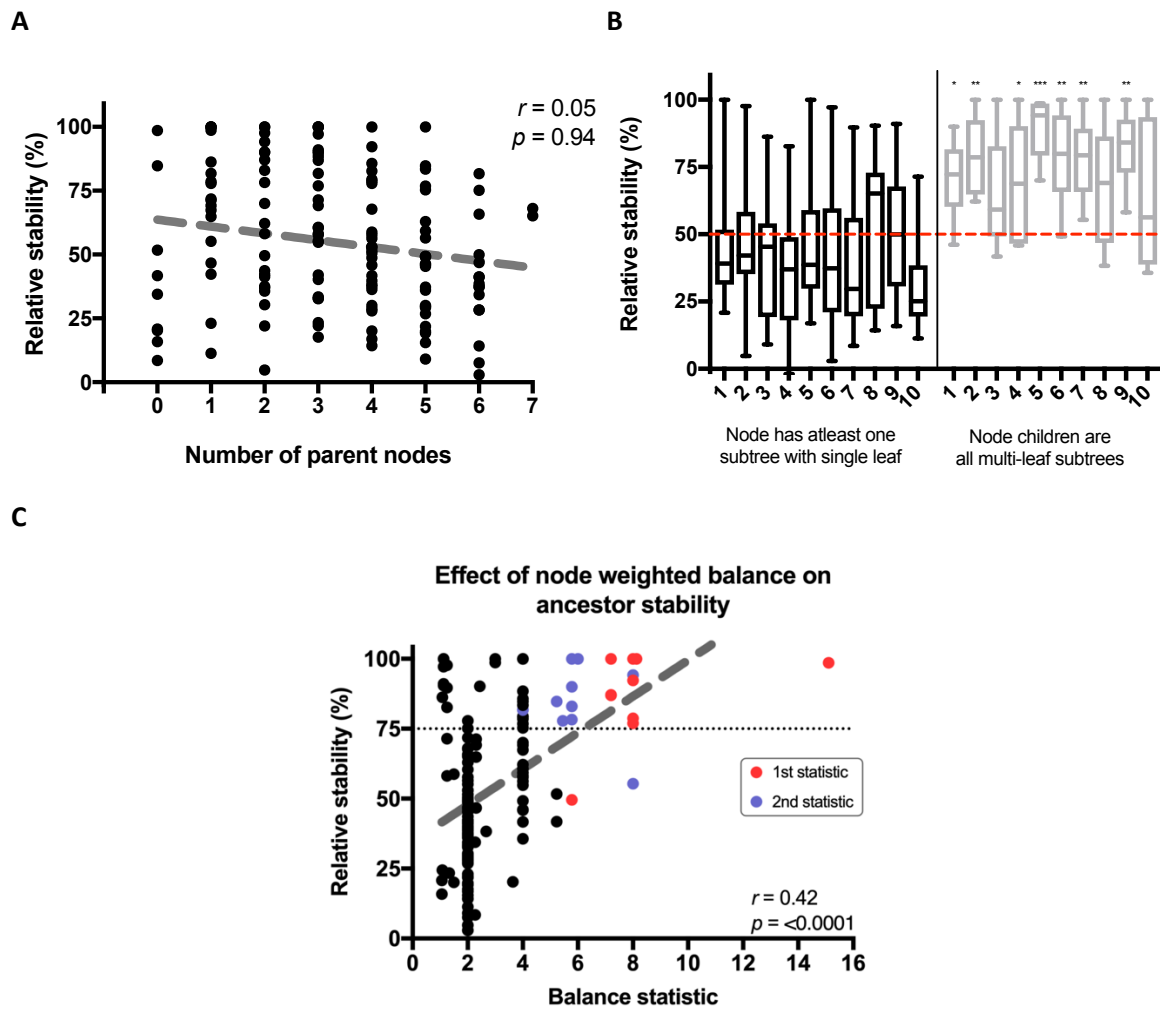


Figure 25 - Node weighted balance is an accurate predictor of high ancestor stability

Node stabilities in simulated phylogenies reconstructed by Ancestron were calculated by PESST, analysed in Microsoft excel and Graphpad PRISM 7, and visualized in Graphpad PRISM 7. Stabilities were normalized to the most stable sequence in the population. **A)** Node stability is negatively correlated with its distance from the base of the tree. Node stabilities were normalized to the most stable node in the population. Correlation was calculated with Spearman’s *r*. Red line represents the linear regression of all data points **B)** Box and whisker plots showing that in the majority of cases, nodes that parent at least one single leaf are less stable than nodes parenting only multi-leaf subtrees. Significant difference between populations was calculated on a per-run basis using Welch’s *t*-test ($p < 0.05$). Red dashed line represents 50th percentile cut-off. **C)** A high weighted balance is a good predictor of high stability. Red points represent the node with the highest weighted balance in each run. Blue points represent the nodes with the second highest weighted balance in each run. Grey dashed line represents the linear regression of all data points. Correlation was calculated according to Spearman’s *r*.

4.5.3 sASR produces functional CAR enzymes

In order to validate whether sASR sampling of ancestral space is a valid engineering tool, we targeted the challenging CAR family of enzymes for stabilization. 42 actinomyete CAR1 sequences were semi-randomly collected, ensuring that they broadly represent both *Mycobacterium* and *Nocardia* CAR clades. We included canonical outgroup sequences from our previously reported dataset of *Actinomycete* CAR enzymes (Finnigan *et al.*, 2017). We aligned these sequences and performed sASR with Ancescon (Zimmerman *et al.*, 2018). The phylogeny produced in Ancescon by the weighbour method provides no support values or branch lengths for nodes (figure 26; supplementary figure 31). Within the phylogeny, node 43 and node 50 possessed the greatest weighted balance (figure 26; 28.5 and 11.5 respectively). These nodes were selected for synthesis (hereafter AspCAR-A43 and AspCAR-A50 respectively; supplementary figure 32). AspCAR-A50 was not resolved with a start methionine so was trimmed of five N-terminal residues to a methionine shared with the start of AspCAR43. AspCAR-A43 and AspCAR-A50 share 78% sequence identity with one another; and 60% and 59% average pairwise identity respectively with the extant CAR dataset from which they were derived. AspCARs were expressed recombinantly in *E. coli*. Both AspCARs were soluble, showing low to moderate expression under previously described CAR expression conditions (Chapter 2), obtaining approximately 2 to 5 mg of enzyme per liter of culture (supplementary figure 33). AspCAR-A50 appears to be sensitive to proteases, generating a number of degradation products, which could affect expression.

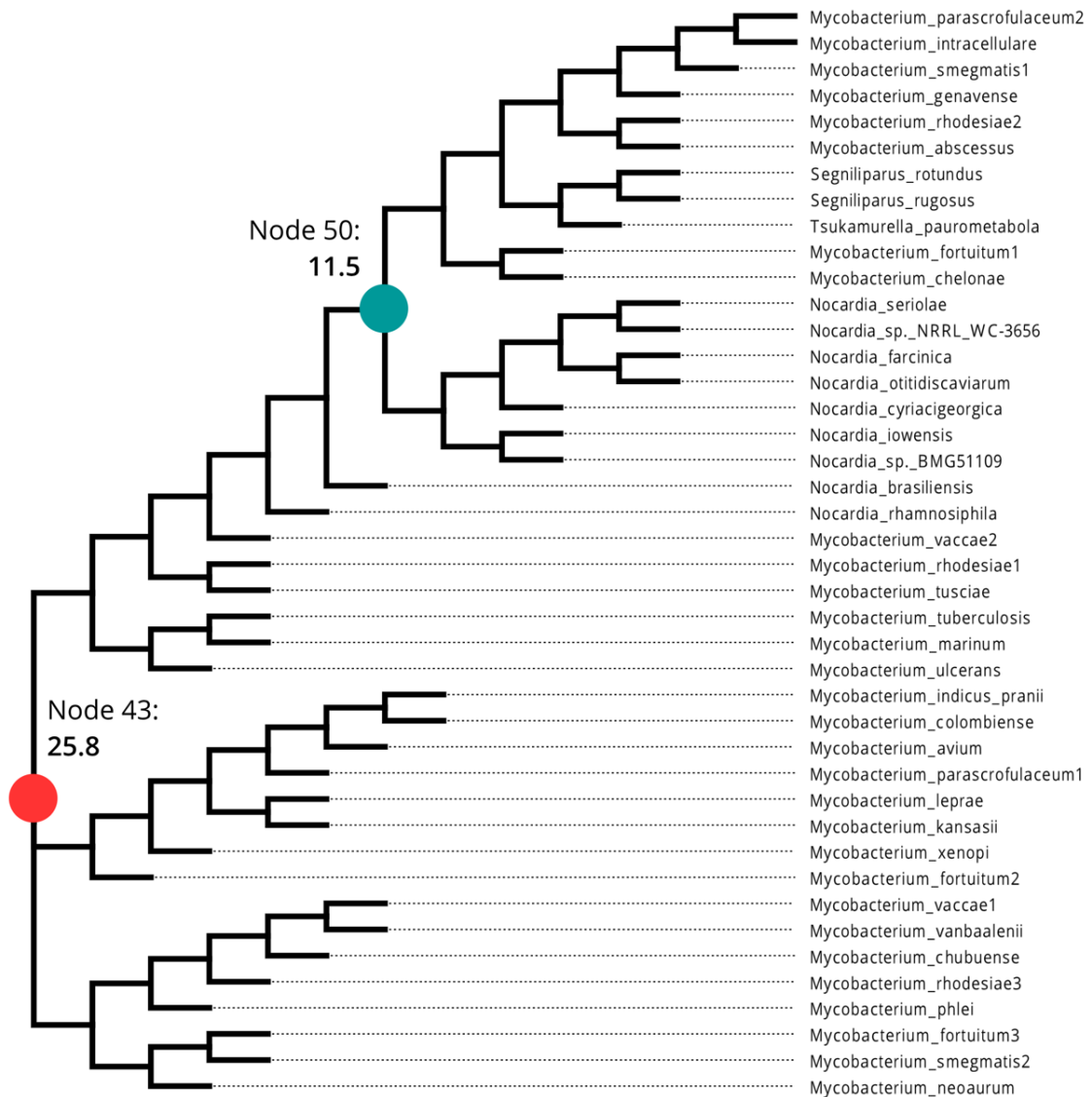


Figure 26 - Tree output from Ancecon built with the weighbour method

Ancecon (Cai *et al.*, 2004) constructs a phylogeny from the user inputted multiple sequence alignment with the weighted neighbor joining (weighbour; Bruno *et al.*, 2000) method when no phylogeny prior is defined by the user. This tree was hand-drawn in Newick format from the information provided in the Ancecon output file (supplementary figure 31), and visualized and manipulated in FigTree v1.4.3 and Gravit.io. Nodes that were taken forward for further experiments, and their weighted balance scores are highlighted.

AspCAR activity was assayed by measuring NADPH absorbance at 340 nm in the presence of ATP and one of 20 carboxylic acids (seventeen aromatic and three fatty acids; figure 27). AspCAR-A43 showed a preference for benzoic acid derivative substrates, whereas AspCAR-A50 showed activity on a broad range of substrates. AspCAR-A50, unlike most extant and ancestral CARs, was able to reduce aromatics with an electron withdrawing 3-nitro group (Winkler, 2018). Kinetic analysis of AspCARs was performed on benzoic acid and its derivatives (figure 27). Both enzymes showed similar substrate turnovers. AspA50 showed between 5 and 10-fold higher substrate affinity, and between 10 and 20-fold improved catalytic efficiency than AspCAR-A43. For AspCAR-A50 compared to AspCAR-A43, catalytic efficiency on NADPH and ATP is again higher (table 12), supplementary figure 34; K_M for ATP was $218 \pm 13 \mu\text{M}$ and $102 \pm 8 \mu\text{M}$ for AspCAR-A43 and AspCAR-A50 respectively; and K_M for NADPH was $115 \pm 5 \mu\text{M}$ and $64 \pm 4 \mu\text{M}$ respectively). These data show that AspCARs had lower turnover rates compared to previously derived ancestral carboxylic acid reductase enzymes (Chapter 2), but generally had better substrate efficiency. Low turnover rates are likely indicative of the enzyme's non-natural nature. Furthermore, AspCAR-A50 had considerably improved catalytic efficiencies over AspCAR-A43 for the majority of substrates, with upwards of a 50-fold improvement on 3-methoxybenzoic acid.

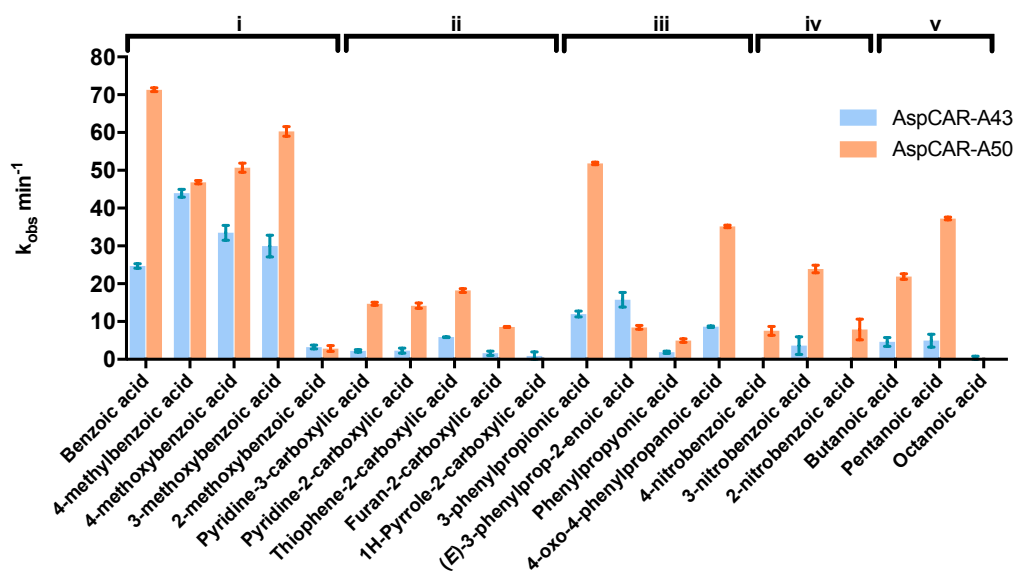


Figure 27 - AspCARs are CAR enzymes with activity on canonical substrates

AspCAR activity was tested on 20 canonical CAR substrates. **i)** Carboxylic acids that are derivatives of benzene; **ii)** Carboxylic acids with aromatic ring structures other than benzene; **iii)** Carboxylic acids with the carboxyl group conjugated from the aromatic ring; **iv)** Carboxylic acids with electron withdrawing nitro groups; **v)** fatty acids. k_{obs} was calculated for 10 g enzyme in the presence of 5 mM substrate. Error bars represent the standard error.

Substrate		k_{cat} (min^{-1})	K_M (mM)	$\frac{k_{cat}}{K_M}$ ($\text{min}^{-1} \text{mM}^{-1}$)
Benzoic acid	AspA43	63.7 ± 1.4	11.6 ± 0.7	5.5 ± 0.5
	AspA50	84 ± 1	1.1 ± 0.1	76 ± 8
4-methylbenzoic acid	AspA43	28.2 ± 0.9	3.9 ± 0.2	7.2 ± 0.6
	AspA50	53 ± 2	0.37 ± 0.03	143 ± 17
4-methoxybenzoic acid	AspA43	52.0 ± 1.1	3.7 ± 0.3	14.1 ± 1.4
	AspA50	38 ± 1	0.26 ± 0.03	146 ± 17
3-methoxybenzoic acid	AspA43	72.3 ± 1.2	6.7 ± 0.4	10.8 ± 0.8
	AspA50	41 ± 1	0.17 ± 0.02	241 ± 28.2
ATP	AspA43	74.6 ± 1.8	0.218 ± 0.013	342 ± 29
	AspA50	162 ± 4	0.064 ± 0.004	$2,530 \pm 220$
NADPH	AspA43	75.2 ± 1.3	0.115 ± 0.005	654 ± 39
	AspA50	164 ± 3	0.10 ± 0.01	$1,640 \pm 200$

Table 12 - AspCAR kinetics

AspCAR kinetics were calculated for benzoic acid derivatives, ATP and NADPH. All kinetics were performed with 1.7x titrations for three experimental replicates. Data were fit to the Michaelis-Menten equation in GraphPad PRISM v. 7. Errors shown as standard errors. Saturation curves are shown in supplementary figure 34.

4.5.4 AspCARs are thermostable CAR variants

AspCAR thermostability was assessed by incubating AspCARs at incremental temperatures from 30 °C to 80 °C for 30 minutes, before testing their activity as before on 4-methoxybenzoic acid. We have previously shown that the salt concentration of the incubation buffer can considerably effect the point at which ancient CAR enzymes lose 50% of their activity (A_{50} ; Chapter 2). Therefore, we assayed A_{50} for both AspCARs in both HEPES and a buffer modelling *in vivo*-like *S. cerevisiae* ionic concentrations (ivSc buffer; van Eunen *et al.*, 2010; Chapter 2). Both AspCAR-A43 and AspCAR-A50 showed increased thermostability compared to extant CARs, and were resistant to salt at *in vivo*-like concentrations (figure 28A). AspCAR-A43 showed A_{50} values of 58 °C and 58.9 °C in HEPES and ivSc respectively. The A_{50} of AspCAR-A50 was 71.1 °C and 70 °C in HEPES and ivSc respectively. We then used differential scanning fluorimetry to corroborate these stability data and calculate enzyme T_m values. T_m values of 57.1 °C and 73.7 °C were obtained for AspCAR-A43 and AspCAR-A50 respectively, closely reflecting the results from the enzyme assays (figure 28B). AspCAR-A50 represents an approximately 25 °C increase in stability over the most thermostable natural carboxylic acid reductase observed today. Moreover, AspCAR-A50 represents up to a 7 °C increase in T_m , and a 9 °C increase in A_{50} over previously reconstructed ancestral CARs (Chapter 2) and represents the most thermostable CAR variant discovered to date.

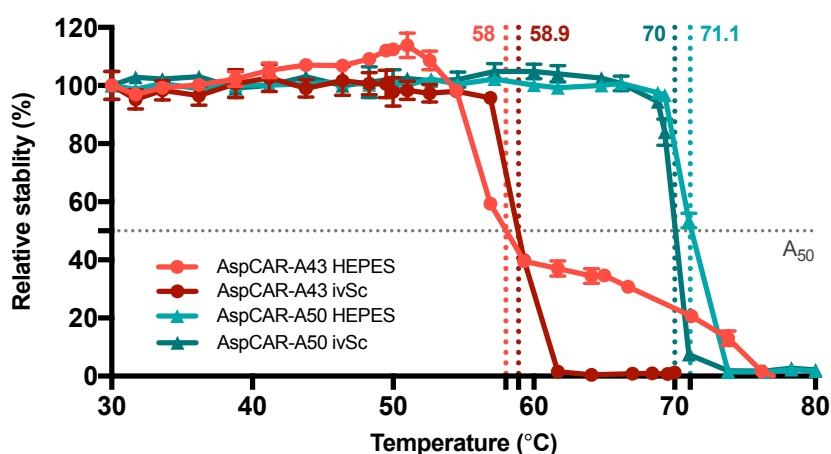
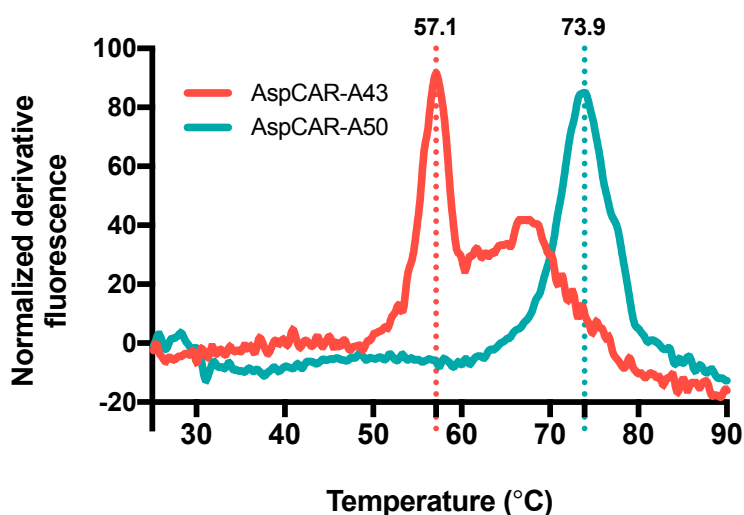
A**B**

Figure 28 - AspCARs are highly thermostable enzymes

The thermostability of AspCAR enzymes. **A)** The temperature at which AspCARs lose 50% activity (A_{50} ; grey broken horizontal line) following incubation for 30 minutes at increasing temperatures from 30 to 80 °C measured by reduction of 20 mM 4-methoxybenzoic acid and 10 mM benzoic acid for AspA43 (red lines; circular points) and AspA50 (blue lines; triangular points) respectively. Activity is displayed as the rate of NADPH reduction in the reaction, relative to the rate of NADPH reduction following incubation at 30 °C. Both AspCAR stabilities were tested in HEPES (light lines) and ivSc buffer (dark lines). Error bars represent the standard error. **B)** Differential scanning fluorimetry to assess AspCAR critical unfolding temperature (Vivoli *et al.*, 2014). Enzyme was incubated under temperature ramping from 25 °C to 100 °C in the presence of SYPRO orange. Critical unfolding was defined as point at which the highest rate of change in fluorescence. Critical unfolding points were calculated in Protein Thermal Shift software v1.3. Images were created using GraphPad v7.

4.6 Discussion

This study was inspired by the need to provide highly accessible protein engineering tools for synthetic biology. Protein engineering is an essential technology for the long-term success of synthetic biology. It permits systems to escape the constraints upon function introduced by evolution (Lin and Xu, 2013; Currin *et al.*, 2015), a key challenge for enabling synthetic biology (Endy, 2005). “Point and click” tools that introduce beneficial properties into proteins without limited costs in time, expertise, and infrastructure offer expedient access to proteins with a wide variety of improved properties. Here, we present a simplified version of ASR that offers such a democratized protein engineering tool. As ASR is a purely phylogenetic method, there is no requirement to undertake involved methods (e.g. protein structure analysis or high throughput screening; Gumulya and Gillam, 2017). ASR derived sequence variants can be derived using entirely open source software, and requires only a multiple sequence alignment as prior knowledge. Additionally, our simplified ASR requires no prior expertise in phylogenetics to undertake. While other automated ASR tools exist (i.e. PhyloBot; Hanson-Smith and Johnson, 2016), sASR is the first simplified tool that is engineering centric.

With its alignment only mode, Ancescon (Cai *et al.*, 2004) is able to generate a crude phylogeny using the weighted neighbor joining method. By reconstructing simulated alignments (supplementary figure 29), we observed that nodes considerably biased toward thermostability despite their evolutionary histories being mesophilic (figure 24). In previous analyses of stability bias in ASR, a positive correlation was observed between node age and its stability, suggesting that more ancient sequences undergo increased stabilizing bias (Chapter 3). On the contrary, sASR produced stability spaces that were only weakly correlated with node age, with stabilized sequences being scattered throughout the phylogenies (figure 25A).

Selecting nodes based on their weighted balance was an effective solution to the node selection problem. In every simulation, At least one of the nodes with two highest weighted balances exhibited stabilities in the upper quartile (figure 25C). Weighted balance proves an efficient method to search the tree-space for nodes that simultaneously optimizes for both

evenly distributed leaf density and high leaf weight. We hypothesise that survivor bias, causing stabilising residues to be overrepresented in population (Bloom *et al.*, 2006; Chapter 3), leads to an increased stabilizing bias acting on nodes that parent multiple leaves (e.g. node 50 in figure 26). Selecting residues in ancestral space from a heavily weighted node leads to an increased opportunity to select stabilizing residues. We also observe significant destabilization of nodes daughtering subtrees represented by a solitary leaf. Therefore, it can be hypothesised that single leaves dilute the biasing effect. In marginal reconstruction, the sequence of the corresponding node and the solitary leaf are priors for calculating the immediate ancestral node. Leaf instability may therefore have a significant effect on immediate parent node stability (Bar-Rogovski *et al.*, 2015; Pupko, 2000; Koshi and Goldstein, 1996). Therefore, we suggest that nodes with a high weighted balance have the optimal probability of selecting (biased) stabilizing residues as they sample an space evenly throughout the marginal reconstruction.

We tested the hypothesis that selecting high weighted balance nodes from sASR makes for a simplified enzyme engineering tool by engineering CAR enzymes (Winkler, 2018). We collected and subsequently aligned homologues to broadly represent the family (Finnigan *et al.*, 2017). We cannot determine from this work whether randomly selected sequences would also produce a similar stabilizing effect. From the alignment-only reconstruction, two functional CAR enzymes were obtained from ancestral space that are between 8-38 °C more stable than the best studied extant CARs (Finnigan *et al.*, 2017; Kramer *et al.*, 2018). Surprisingly, AspCAR-A50 is also up to 9 °C more stable than true CAR ancestors (Chapter 2; figure 28). The enzymes that are reconstructed here did not achieve as high turnovers of their most optimal substrates at standard assay temperatures, compared to both extant proteins and ASR reconstructed CAR ancestors. It may be the case that poor activities are driven by the lack of alignment refinement or the nature of the reconstruction by sASR.

Despite being based on ASR, sASR is a distinct methodology as it is unlikely to accurately factor in evolutionary history as the reconstruction uses a phylogeny derived by weighted neighbor joining in Ancestron (Cai *et al.*, 2004). It should therefore be expected that Ancestron generates “ancestors” that are not reflective of real-world protein ancestry, but instead the ancestry of the inaccurate tree space. While the sequences can be considered

ancestral, they are derived from potential points in ancestral space. Indeed, the positioning of deeper nodes in the neighbour phylogeny of CARs is a poor match with the previously reported well-supported phylogeny of CAR enzymes (Finnigan *et al.*, 2017). Nevertheless, as the method selects optimal residues from the sequence consortium, it is still likely to over-represent stabilizing residues in the reconstructed sequence regardless of the phylogeny. If the quality of the alignment is sufficient to produce an optimal opportunity for the generation of functional proteins from their consensus (i.e. functional domains are conserved), it is likely that sASR will induce a stabilizing effect. sASR therefore offers a novel method for the engineering of thermostability exploiting the bias towards stability, alongside ASR and consensus sequences. An sASR reconstruction can be performed in a matter of hours, meaning that the experimental design process is not a bottleneck in protein engineering using sASR. Combined with current generation DNA synthesis technologies that provide ready-cloned sequences, high volumes of engineering experiments can be performed with sASR in a short space of time, without a large overhead. The low risk and low cost make it an ideal "proof-of-principal" testbed for optimization, and opens enzyme engineering to a wider community.

We envisage that sASR will complement other protein engineering tools. In this case it is not necessary for the sASR derived protein to have high activity, as protein stability is the premium property. Conceptually, protein engineering can be considered as a penalty matrix where penalties (or rewards) for mutation at a given position are influenced by all relevant protein properties. Any engineering process aims to optimize over this penalty matrix. A mutation desirable for one property that imposes penalties in another dimension of the matrix may be unfavorable overall (Romero and Arnold, 2009). For mesophilic proteins, stability penalties are accentuated: a stability decrease may render the protein unable to fold into a functional molecule at a given temperature (Huang *et al.*, 2016). In contrast, stable enzymes have their functional landscape smoothed, so penalties across the property matrix become dampened during subsequent engineering (Wagner, 2008; Tavanti *et al.*, 2017). Consequently, the *de novo* protein design community has aimed to obtain stable scaffolds as a starting point onto which to impart novel functionalities (Bozkurt *et al.*, 2018; Huang *et al.*, 2016). This effort is translatable to any protein engineering pursuit, using well-described protein engineering methods (Gumulya and Gillam, 2017; Romero and Arnold,

2009). We therefore see great potential for sASR as an initial protein engineering step, providing a stable scaffold for other, more involved screening-based technologies to more effectively optimize proteins across their property matrices.

4.7 Conclusion

In this study, we have developed and validated a novel protein engineering tool based on a simplified ASR workflow. The only required input for sASR is a multiple sequence alignment, which is then used by Ancescon to produce an estimation of ancestral sequence space based on a simple phylogeny produced by the algorithm. Our approach can be performed using free software in a short space of time. Using PESST, we show that sASR provides a high chance of obtaining thermostable sequences if the two nodes with the highest weighted balance are synthesised. We validated this strategy, and sASR by reconstructing two highly thermostable CAR enzymes. The sASR method has low requirements of cost, prior knowledge and expertise, and provides rapid access to protein stabilization to a broad community of would-be protein engineers.

4.8 Methods

PESST priors

Ten PESST V1.0 (Chapter 3) simulations were run under identical priors, evolving 52 proteins of 100 amino acids for 2,000 generations. Amino acid stability contributions ($\Delta_{r,a}$ values) were modelled to a Gaussian distribution defined by $\mathcal{N}(-1.5, 2.5)$, under a stability threshold of 0. Simulations were initiated at a value between $\Omega + 5 \geq T_0^{med} \geq \Omega + 25$ to ensure the average stability of the simulated dataset sustained marginality throughout evolution. All other settings were set to default.

Sequence handling

Alignments output from the model were input into Ancescon on the MPI bioinformatics tool for ancestral reconstruction (<https://toolkit.tuebingen.mpg.de/>; Cai *et al.*, 2004; Zimmerman *et al.*, 2018). Output ancestral sequences were manually converted to FASTA

format. Ancestral sequence stability was calculated with the “stability calculator” module within PESST. Analyses of fitness were performed in Microsoft Excel and Graphpad PRISM v7.

Calculation of weighted balance

A node’s weighted balance $N_{w.bal}$ is calculated as: $N_{w.bal} = \frac{a}{b} \times (a + b)$ (figure 29)

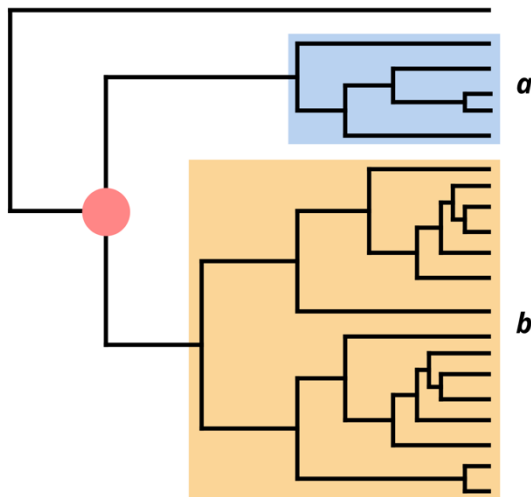


Figure 29 - Calculating weighted balance

Weighted node balance of the node of interest (red circle) is calculated with the displayed formula where a = the amount of leaves in the smallest subtree immediately daughtered by the node of interest (blue square), and b = the amount of leaves in the largest subtree immediately daughtered by the node of interest (orange square). Therefore, in this instance $N_{w.bal} = \left(\frac{5}{15}\right) * 20 = 6.67$.

Polytomies represent a branch of length 0 connecting subtrees (Coddington and Schraff, 1995). Polytomous subtrees that clustered to one side of a root branch were treated as a single subtree for the sake of the calculation.

Sequence handling of CARs

42 CAR sequences were semi-randomly selected from the dataset of CARs previously curated in Finnigan *et al.*, 2017. CARs were aligned with MUSCLE in the phylogeny.fr online phylogenetics tool (<http://www.phylogeny.fr/>; Edgar, 2004; Dereeper *et al.*, 2008). Aligned sequences were input into the Ancestron algorithm as described before. Ancestors were confirmed to contain essential catalytic residues (Finnigan *et al.*, 2017) by eye following sequence alignment in the Geneious v10 phylogenetics suite (Kearse *et al.*, 2012).

Molecular biology

AspCARs A43 and A50 were codon optimized for expression in *E. coli* K12 using an internal codon optimization script. AspCAR-A43 was purchased as two gBlocks (IDT) adhering to the Biobricks standard, sharing an internal *Xma*I restriction site. Both parts were cloned into

pNIC28-Bsa4 (Savitsky *et al.*, 2010) expression vector by one-pot restriction cloning. Plasmid was digested with *Xba*I and *Hind*III. Part-1 was digested with *Xba*I and *Xma*I. Part-2 was digested with *Xma*I and *Hind*III. Successful cloning was screened through removal of the *Sac*B counter-selectable marker by growth on kanamycin-agar plates in the presence of 5% v/v sucrose. The final AspCAR-A43 sequence was confirmed by Sanger sequencing (Source Bioscience). AspCAR-A50 was synthesized as a complete cloned construct in pNIC28-Bsa4 by Dundee Cell Products (supplementary files).

Plasmids were co-transformed into *E. coli* BL21(DE3) alongside a pCDF-Duet1 vector containing *B. subtilis* phosphopantetheine transferase (Finnigan *et al.*, 2017). AspCAR expression was induced under 150 μ M IPTG at log phase, and cells were grown overnight at 20 °C. Cells were harvested, resuspended in 20 mM Tris-HCl, 0.5 M NaCl, 20 mM imidazole pH 8.0 and lysed by sonication. Purification was performed using an ÄKTApur automated chromatography system (GE Healthcare). Nickel affinity chromatography was performed using a 1 mL HisTrap crude column (GE Healthcare), eluting with a step gradient to resuspension buffer supplemented with 250 mM imidazole. This was followed by size exclusion chromatography using a Superdex 200 16/600 preparative column (GE Healthcare), eluting isocratically with 10 mM Hepes pH 7.0, 0.5 M NaCl. Chromatography peaks were analyzed by SDS-PAGE with ExpressPAGE pre-cast gels (Genscript). Peaks containing AspCAR protein were concentrated with Amicon Ultra 50,000 mwco centrifugal filters, and where necessary stored in 20% glycerol at -20 °C. Before assays, ancestral proteins were buffer exchanged into appropriate buffer with P10 desalting columns (Generon).

Enzymatic assay

The standard enzymatic assay for AspCAR activity was modified from Finnigan *et al.*, 2017. Unless otherwise specified, protein was assayed in triplicate in 200 μ l reactions containing 125 mM HEPES-NaOH (pH 7.5), 1.2 mM ATP, 10 mM MgCl₂, 500 μ M NADPH, 5 mM substrate and 10 μ g enzyme. Working stocks of each assay component were dissolved in 50 mM HEPES-NaOH (pH 7.5). Substrates were dissolved in 200 mM HEPES-NaOH (pH 7.5). Where necessary substrates were dissolved in 20% (v/v) DMSO.

Assays were performed in flat-bottomed 96 well microtiter plates (Greiner). 100 μ l volumes each of reaction mix and dissolved substrate were incubated at 30 °C for 5 minutes before being combined to start the reaction. Enzyme activity was measured at 30 °C by absorbance at 340 nm in a Tecan Infinite 200Pro plate reader in continuous cycles over the course of 10 minutes with 10 flashes per-well. Data were processed using the standard curve presented in Thomas et al., 2018 in Microsoft Excel and GraphPad PRISM v7.0.2.

Substrate range

Both AspCAR substrate ranges were tested using the standard assay on 20 carboxylic acid substrates, including three fatty acids, and seventeen aromatic carboxylic acids. Compounds were prepared to 0.5 M stocks in DMSO and diluted to working concentration in assay buffer.

Kinetics analysis of AspCARs

Kinetic analyses of AspCAR activity on ATP and NADPH were performed in the presence of 20 mM 4-methoxybenzoic acid for AspCAR43, and 10 mM benzoic acid for AspCAR50. NADPH and ATP were titrated in a 1.7x dilution series over 12 points. Points were omitted at low concentrations where signal was obscured by background noise. Kinetic analyses of AspCAR activity on benzoic acid derivatives were performed in at least 5x saturating concentrations of ATP and NADPH. Substrates were titrated in 1.7x dilutions over 8 points in assay buffer. All rates were fitted to the Michaelis-Menten equation by in GraphPad Prism v7.0.2.

Analysis of thermal tolerance in AspCARs.

Stability analyses were performed in both in vitro conditions (assay buffer) and in vivo conditions (Buffer modelling in vivo-like *S. cerevisiae* ionic concentrations (FSC); described in Thomas et al., 2018). 80 μ l aliquots of each AspCAR at 0.5 μ g μ l⁻¹ were incubated in each buffer system for 30 minutes at temperatures between 30 °C and 85 °C in an Mastercycler nexus thermocycler (Eppendorf) set to gradient mode. The second and penultimate aliquot in each gradient segment was reserved for 80 μ l buffer for a negative control. Enzymes were then cooled to 4 °C in the thermocycler for 5 minutes before being assayed as standard for

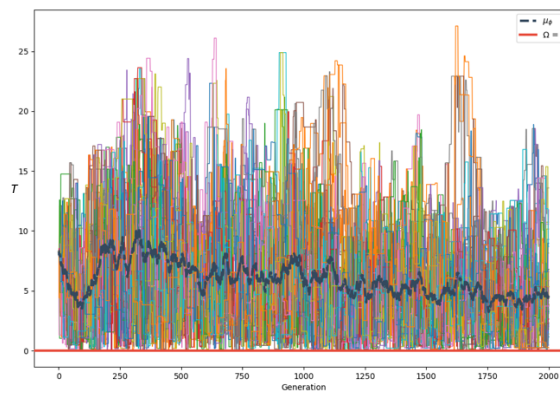
their activity. AspCAR43 activity was assessed in the presence of 20 mM 4-methoxybenzoic acid. AspCAR50 activity was assessed in the presence of 10 mM benzoic acid.

For differential scanning fluorimetry, enzyme was diluted to 0.25 $\mu\text{g ml}^{-1}$ in standard assay buffer, and 10X SYPRO orange (Invitrogen). DSF on both enzymes was run in triplicate 20 μl volumes with a triplicate no enzyme control in a 384 qPCR plate (Thermo) on a QuantStudio 6 flex real-time PCR machine (Life Technologies) set to melt curve mode, with a temperature ramp from 25 $^{\circ}\text{C}$ to 99 $^{\circ}\text{C}$ at a rate of 0.17 $^{\circ}\text{C s}^{-1}$. Data were analyzed using Protein thermal shift software (Life Technologies) v 1.3.

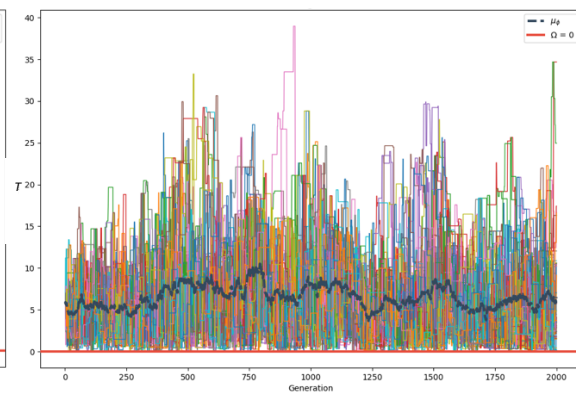
4.9 Supplementary figures

Supplementary figure 29

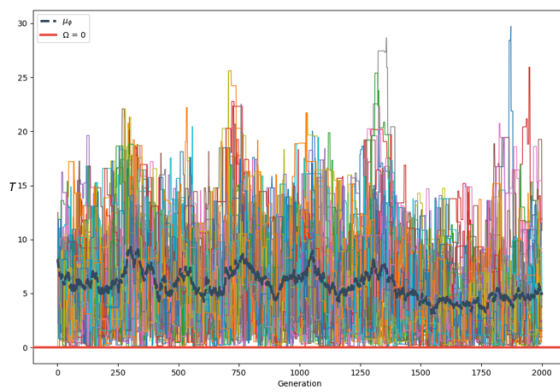
Simulation 2



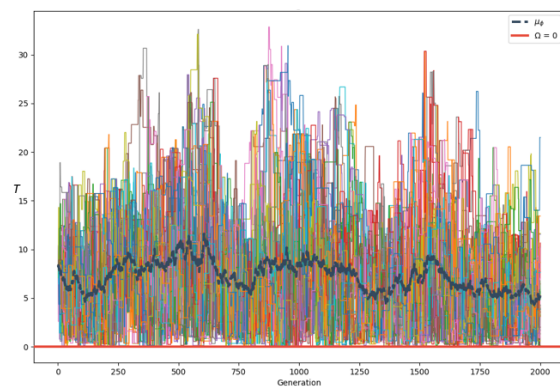
Simulation 3



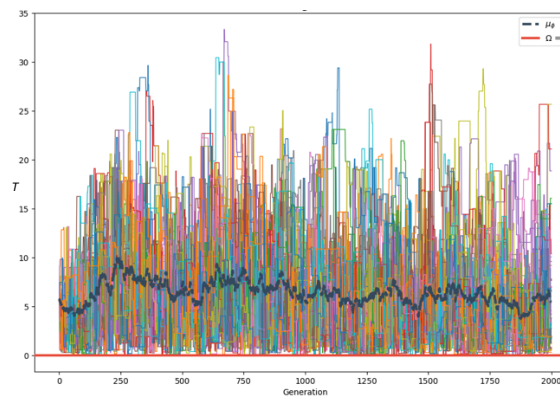
Simulation 4



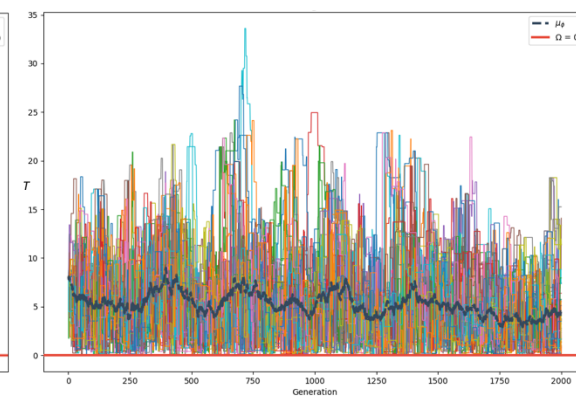
Simulation 5

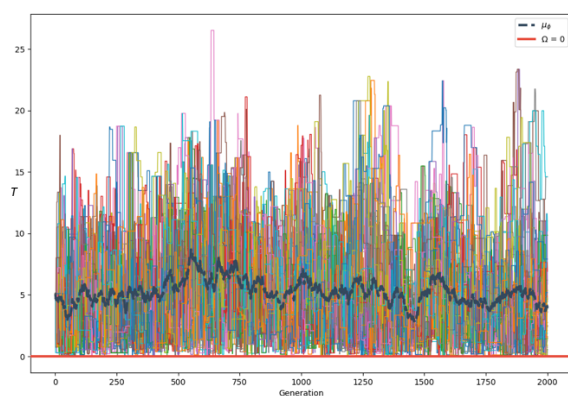
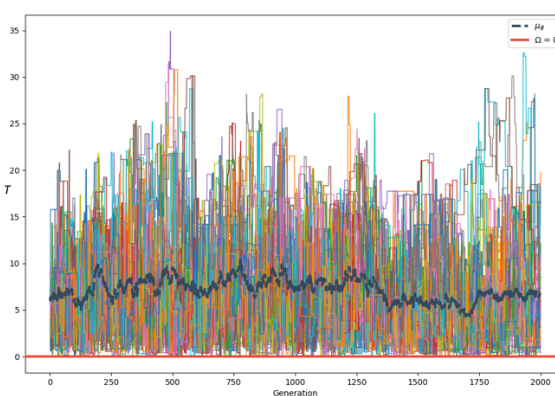
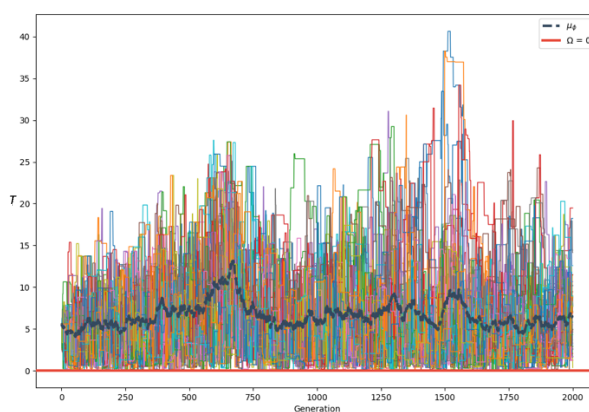


Simulation 6



Simulation 7

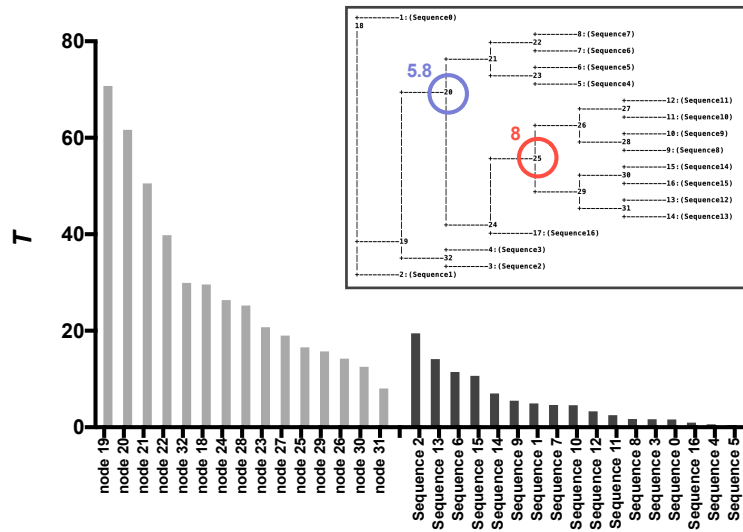


Simulation 8**Simulation 9****Simulation 10****Supplementary figure 29 - PESST simulations used in analysis of Ancescon reconstructions**

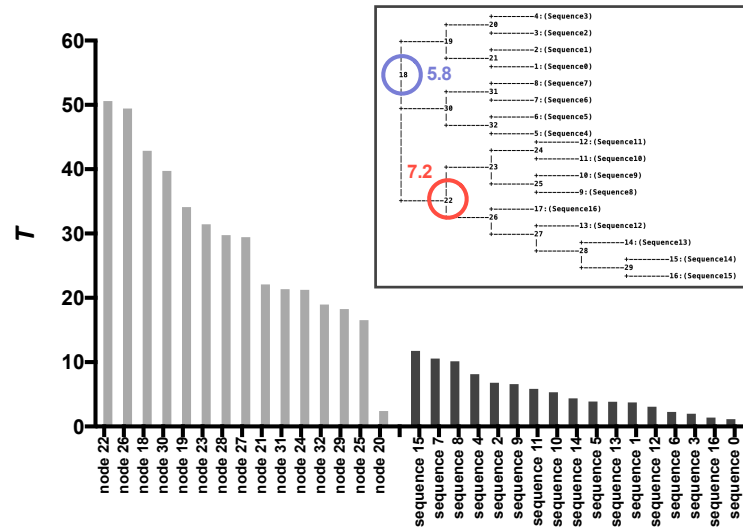
Set of stability traces of simulations 2-10 (simulation 1: figure 24A) showing the changing stability of independent simulations of evolving protein populations over time. Each coloured line represents the stability of a single protein as its population evolves. Grey tight dashed line represents the average stability of all proteins in the population. Red line represents the stability threshold. Figure was rendered with Matplotlib in Python. Seeds for simulations are in supplementary table 9.

Supplementary figure 30

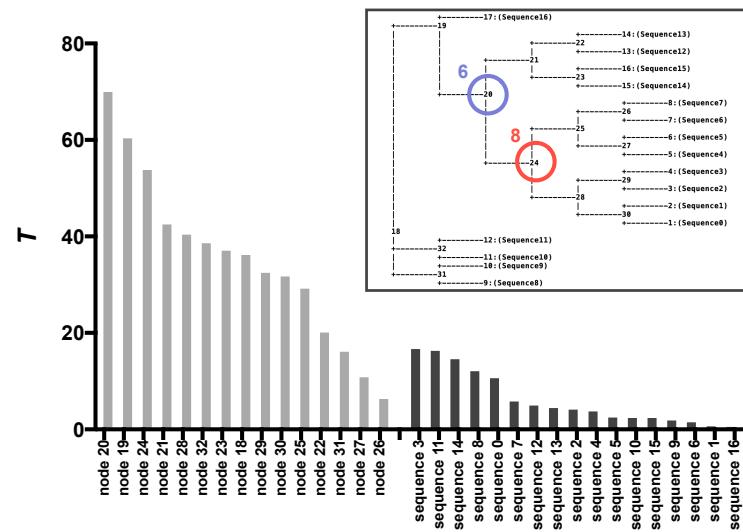
Simulation 1



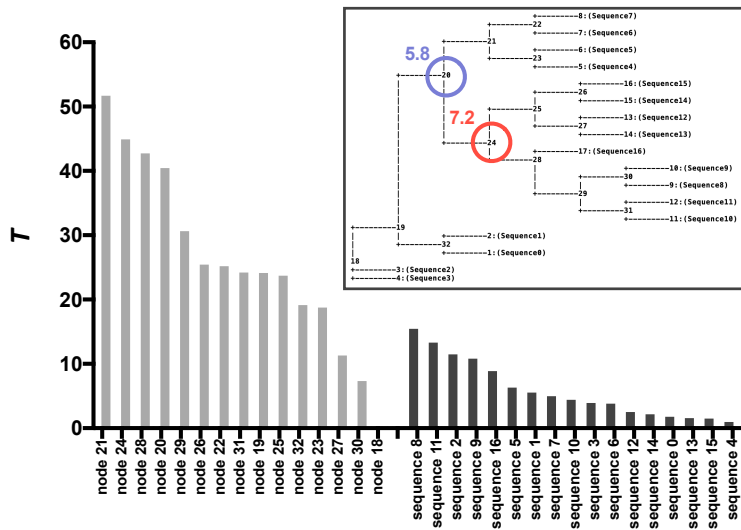
Simulation 2



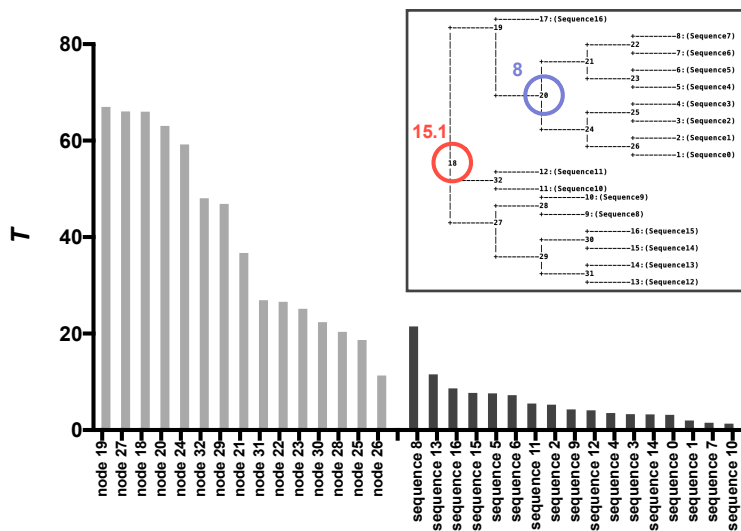
Simulation 3



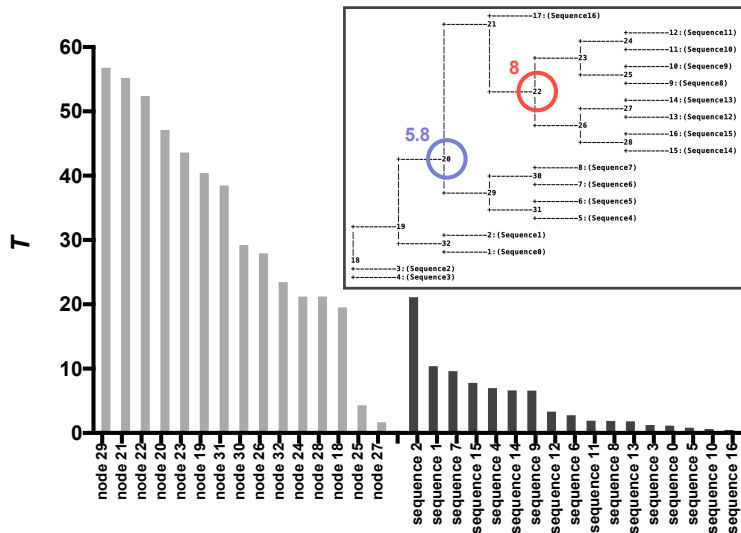
Simulation 4



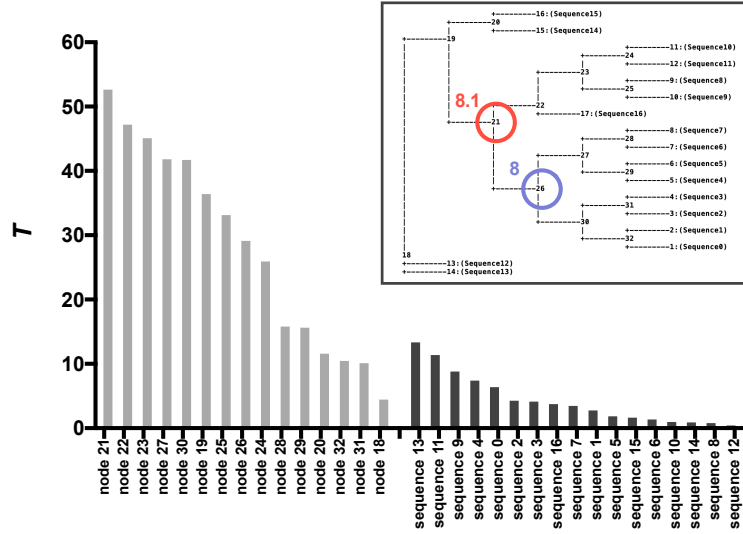
Simulation 5



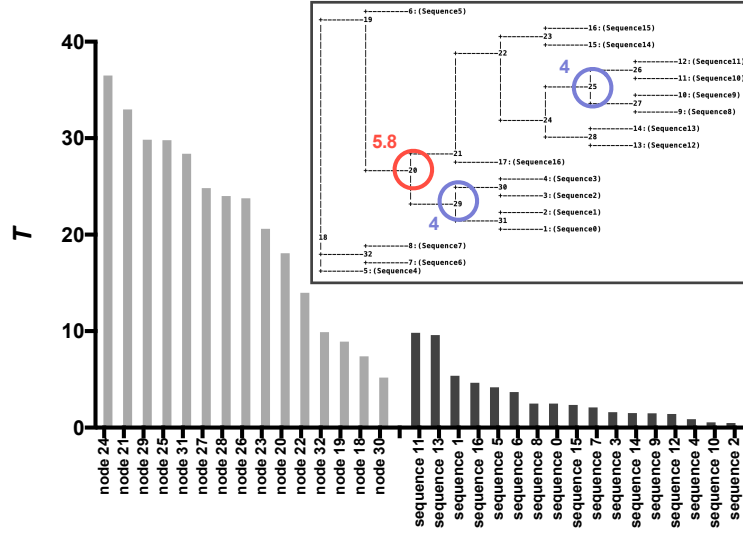
Simulation 6



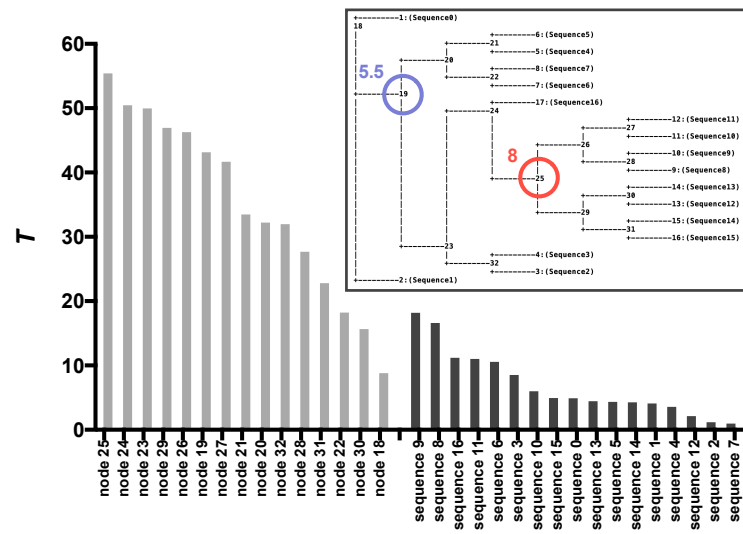
Simulation 7



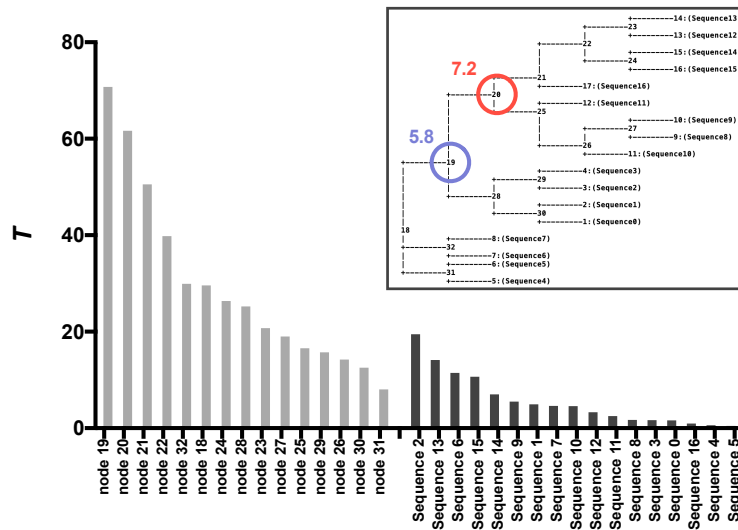
Simulation 8



Simulation 9



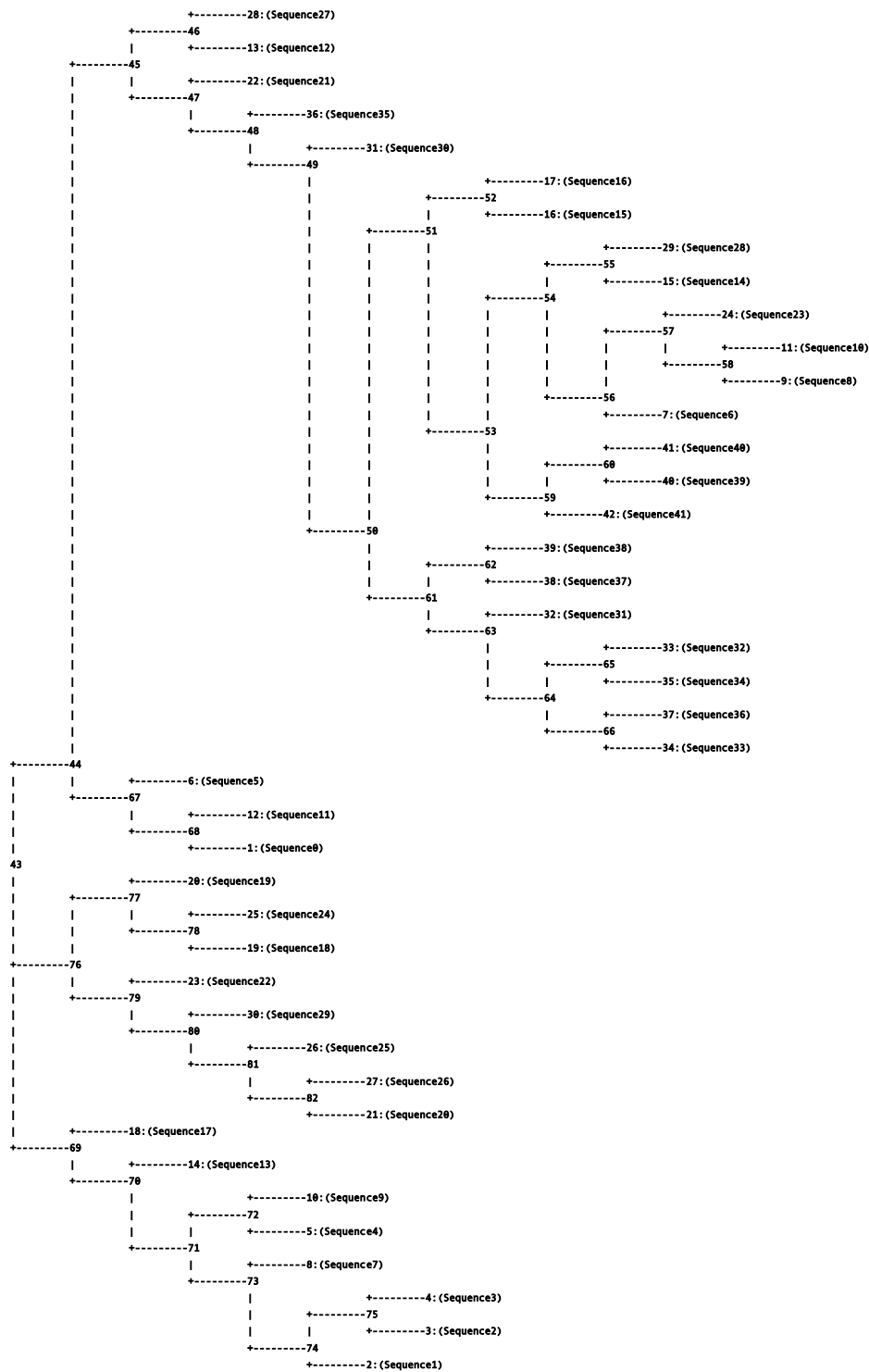
Simulation 10



Supplementary figure 30 - Nodes predicted by sASR are considerably biased toward stability. High stability nodes can be predicted with node balance

Graphs represent the stability space of ten PESST simulations across their reconstructed Weighbour phylogenies calculated by Ancestron, presented in-picture. Reconstructed node stability is based on node sequences predicted by Ancestron, back-calculated in PESST (Chapter 3). Circled nodes on each phylogeny represent the nodes with the highest (red) and second highest (blue) weighted balance (figure 25C). Data was analysed in Microsoft Excel and visualised in GraphPad PRISM ver. 7.

Supplementary figure 31



Supplementary figure 31 - Raw tree output from Ancecon

Ancecon ASCII tree from the Raw ancecon output, produced with the Weighbour method (Bruno *et al.*, 2000). Node labels correspond to the predicted ancestral sequences.

Supplementary figure 32

>AspCAR-A50

```
MSTDTRREERLARRIAELFATDEQFRAARPDPAVSEAVSQPGLRLAQIIATVMEGYADRPALGQRAVELVTDAAATGRTRARLLPRFETITYGEL
WSRVGAIAAAQWHPENPVVRAGDFVATLGFTSVDYTVVDLACTRLGAVSVPLQASAPVAQLTPI LAETEPVLAASA EHLDAAVECVLAGPSP
RRLVVFYDHPVEVDHREAL EAARERLAEAGSPVTVELDEVIARGRALPAAPLYTPDDDDPLALLIYTSGSTGTPKGAMYTERLVARMWLRA
SKLASGSQVPSINLNFMPMSHVMGRASLYGTLARGGTAYFAAKSDMSTLFEDIALVRPTELA FVPRVCDMLFQRYQSEVDRRMAAGADRETAE
AEVKAELRENLLGGRFLSAMCGSAPLSAEMKAFMESCLDLHLVDGYGSTEAGMVLVDGQIQRPPVIDYKLVDPPELGYFSTDKPHPRGELLVK
TETMIPGYKRPEVTAEVFDADGFYRTGDIVAELEPDHLVYVDRRNNVLKLSQGEFVTVAKLEAVFANSPLVRQIFVYVYNSERSYLLAVVVP
EEALAAAGDTEELKAAIAESLQIQAKDAGLQSYEIPRDFLIETEPFTIENGLLSGIGKLLRPKPKERYGERLEQLYAELAEQADELRALRR
AAADRPVLETVTRAAAALLGVAAADVSPDAHFTDLGGDSL SALSFNLLQEIFGVEVPVGVIVSPANDLRGIAEYIEAERESGSKRPTFASVH
GAGATEIRAADLTLDKFI DAETLAAAPSLPAATATPRTVLLTGANGYLGRFLALEWLERLDKTTGGKLCIVRGKDAARRRLEAFDSDGPE
LLARYRELAERHLEVLGDI GEPNLGLDEATWQRLAETVDLIVHPAALVNHVLPYSQLFGPNVVGTA EIIRLAITTKIKPVTYLSTVAVAAQV
DPAVFTEDGDIREISPVRAIDDSYANGYGNSKWAGEVLLREAHDL CGLPVAVFRSDMILAH SRYAGQLNVPDMFTRLILSLLATGIAPKSFYQ
ADADGNRQRAHYDGLPVDF TAEAITTLGAQVAEGFETYDVMNPHDDGISLDEFVDWLEIAGHP IERIDDYAEWFRFETALRALPEKQRQHSV
LPLLHAYRHPQPPVRGSLPTKRFRAAVQEAKIGPDGDI PHLSRELIEKYVSDLKLGLLSSG
```

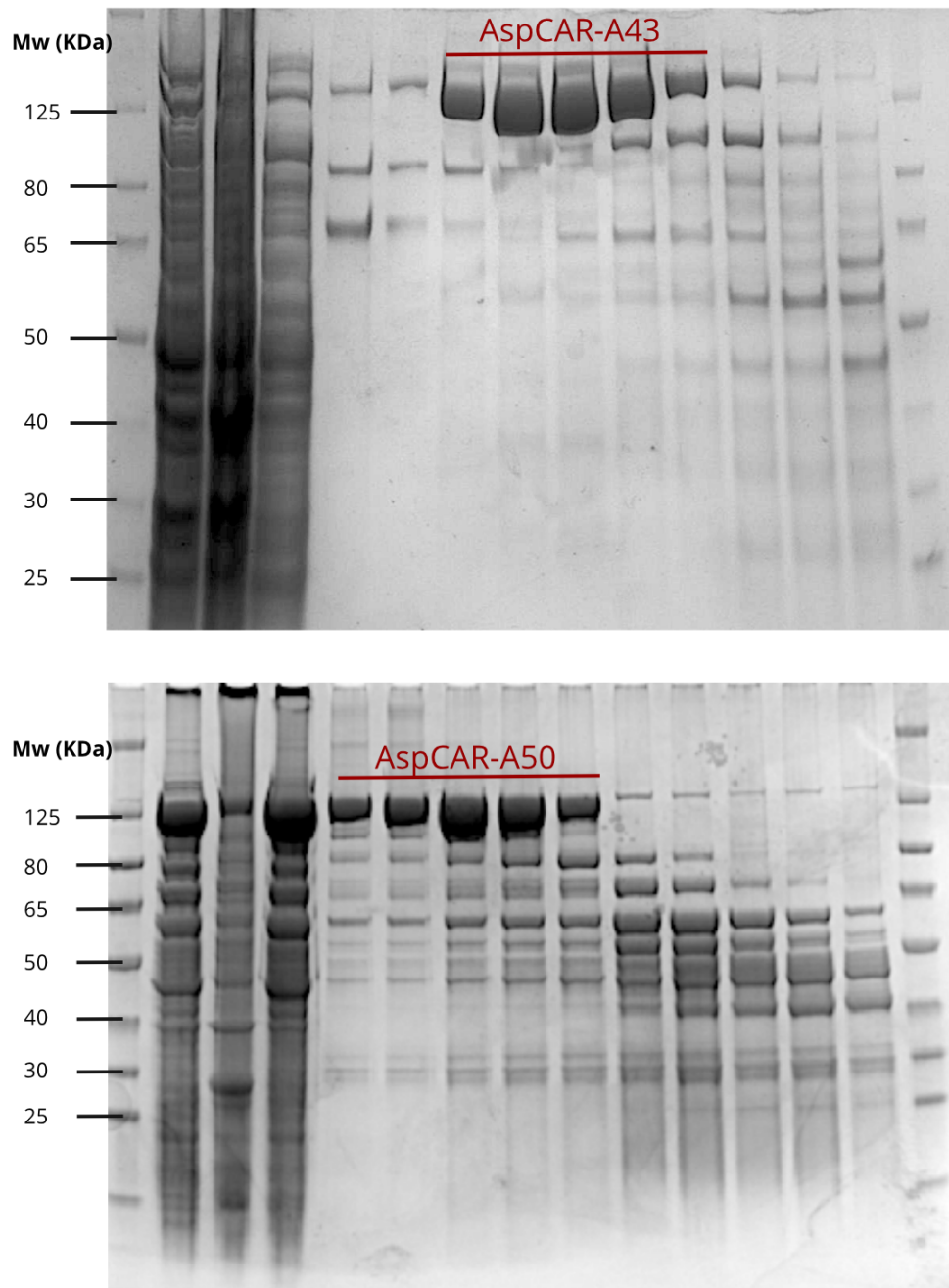
>AspCAR-A43

```
MSTDTRREERLERRIADLYATDPQFAAARPDPAITAAVSVQGLRLPEIIQTVLEGYADRPALGQRAVEFVTD PATGRRTAQLLPRFETITYREL
WDRVGALANAWSNDVVRPGDRVCILGFTSVDYTTIDMALIRLGAVSVPLQTSAPVTQLRP IVAETEPTVIASSVDHLADAVELVLSGHAPARL
VVFYDHPVEVDHREAL EAARARLAEAGTAVTVETLAEVIARGSLPAAAPAPTDDSDPLALLIYTSGSTGAPKGAMY PESKVANMMWRASKA
WFGPAAPSI TLNFMPMSHVMGRGILYGTLANGGTAYFAARS DLSTLLEDLALVRPTQLNFVPR IWDMLFQEQSEVDRRLADGADRAAAAEV
LAELRQNLGGRFVSAMTGSAPISPEMKAWVESLDMHLVDGYGSTEAGMVLVDGQVQRPVIDYKLVDPPELGYFSTDRPHPRGELLVKTEN
MFPGYKRPEVTAEVFDEDGYYRTGDIVAEVGPDLVYVDRRNNVLKLSQGEFVTVSKLEAVFGNSPLVRQIYVYVYNSARPYLLAVVVPTEEA
LARHDVEELKPAISESLQEVAKAAGLQSYEIPRDFI IETTPFTLENGLLTGIRKLARPKLKEHYGERLEQLYTELAEQADELRRLRSGADA
PVLETVSRAGALLGAAASDLQPD AHFTDLGGDSL SALTFGNLHEIFDVDPVGVIVSPANDLQAIADYIEAQRQGSKRPTFASVHGRDATE
VHAGDLTLDKFI DAATLAAAPSLPGPSSEVRTVLLTGATGFLGRYLALEWLERMDLVGGKVICLVRAKSDAEARARLDATFDSGDPKLLAHYR
ELAADHLEVIAGDKGEADLGLDRQTWQLADTVDLIVDPAALVNHVLPYSELFGPNALGTAE LIRIALTTIKPYTYVSTIGVGDQIEPGKFT
EDADIRQISATRKIDDSYANGYGNSKWAGEVLLREAHDL CGLPVAVFRCDMILADTTYAGQLNLPDMFTRMMLSLVATGIAPKSFYELDADGN
RQRAHYDGLPVEFIAEAI STLGAQVAEDEGFETYHVMNPHYDDGIGLDEFVDWLEIAGYPIQRIDDYGEWLQRFETALRALPDRQRQASLLPLL
HNYQQPEKPIRGSMAPTDRFRAAVQEAKIGPDKDIPHSPEIIVKYITDLQLLGLLDAKR
```

Supplementary figure 32 - AspCAR sequences reconstructed with Ancestron

Based on weighted balance calculations, two AspCAR sequences were isolated from the phylogeny presented in figure 26 – AspA43 and AspA50. Amino acid sequences of both AspCAR enzymes that were synthesised are presented.

Supplementary figure 33

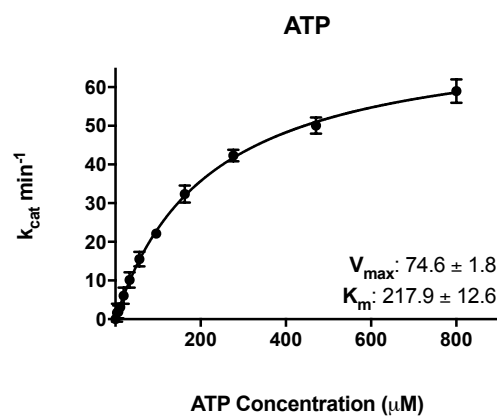
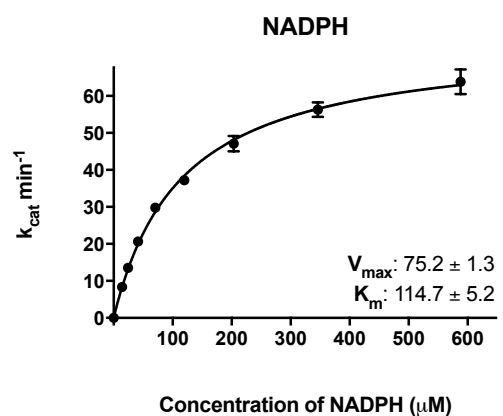
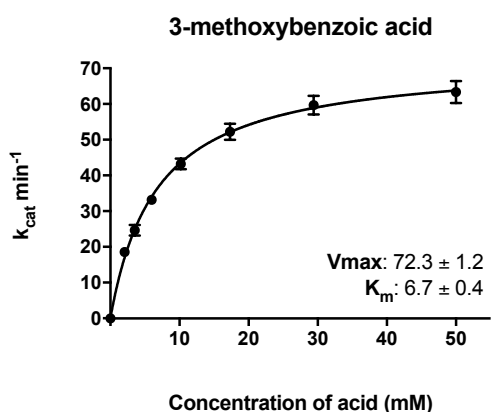
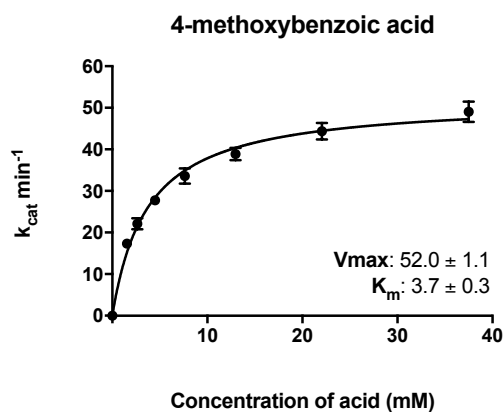
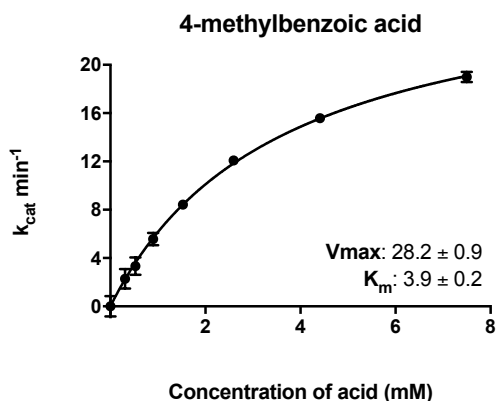
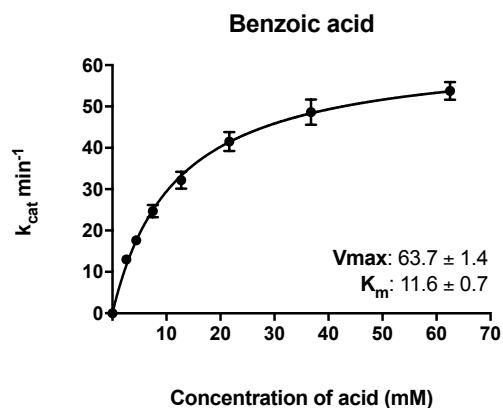


Supplementary figure 33 - SDS-PAGE gels of AspCAR-A43 and AspCAR-A50

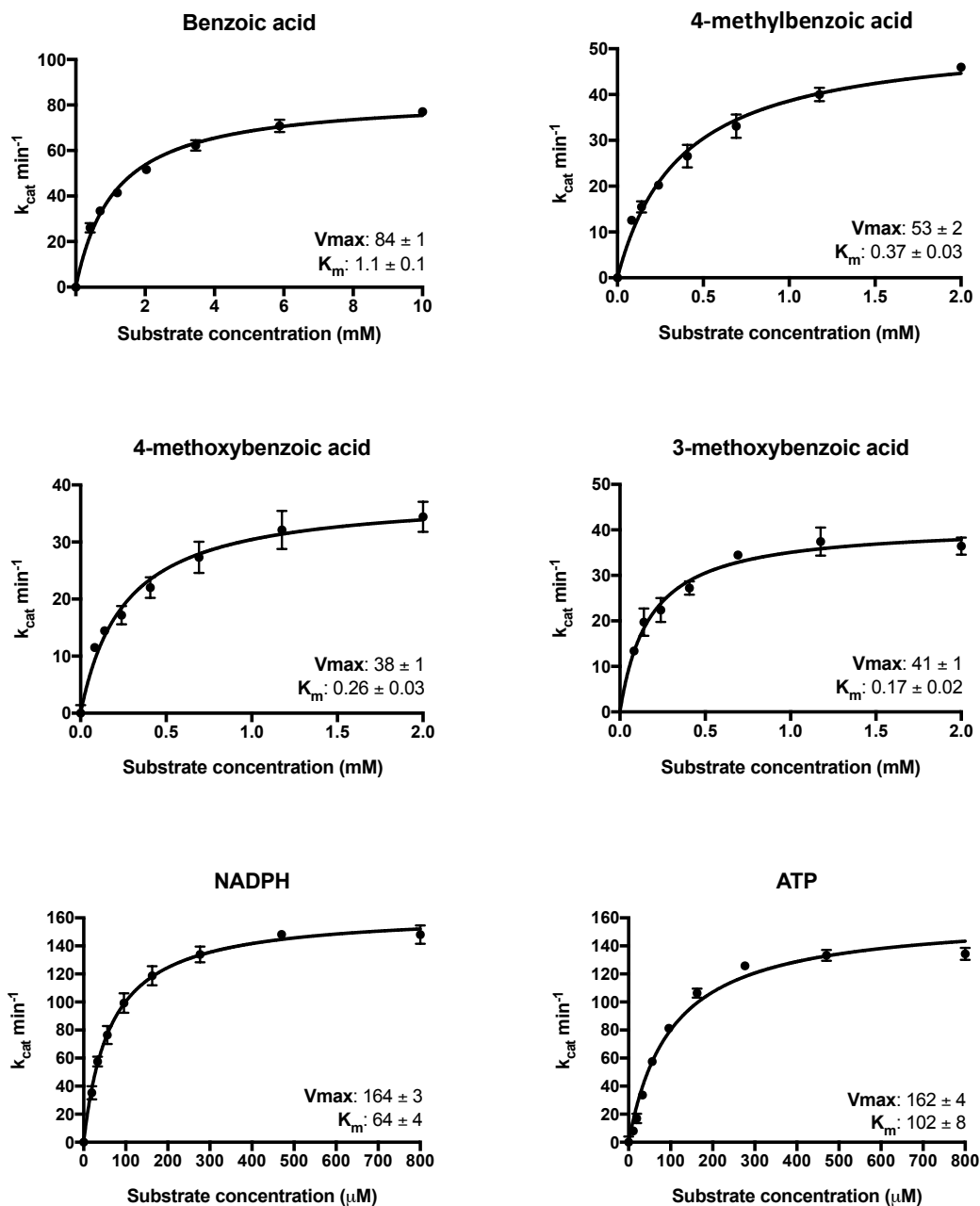
AspCAR-A43 and AspCAR-A50 were purified by nickel affinity chromatography with subsequent size exclusion chromatography. AspCAR-A sizes were analysed by SDS-PAGE on 4-20% precast gels against Spectra BR ladder. The expected size of both AspCAR enzymes is approximately 129 kDa. Bands corresponding to AspCAR-A enzymes are highlighted.

Supplementary figure 34

AspA43



AspA50



Supplementary figure 34 - Saturation curves for AspCAR kinetics

AspCAR kinetics were calculated for benzoic acid derivatives, ATP and NADPH. All kinetics were performed with 1.7x titrations for three experimental replicates. Some values for low concentrations of ATP and NADPH were omitted as their signal dropped below background noise. Data were fitted to the Michaelis-Menten equation in GraphPad PRISM v7. Error bars represent standard error. $V_{max} = k_{cat}$. Units for V_{max} : $\mu\text{M} \mu\text{M}^{-1} \text{ min}^{-1}$. Units for K_m : μM

Supplementary table 9

Simulation	1	2	3	4	5	6	7	8	9	10
Seed	1110687841	3230332449	1979310082	1497264872	3966016719	2976148564	4060848892	4130607136	1869182156	2864906632

Supplementary table 9 – Seeds for PESST simulations

PESST simulations rely on random number generators to ensure every simulation is a unique evolutionary scenario. Seeding allows for runs to be replicable and repeatable. Seeds generated for the 10 PESST simulations utilized in this study are presented.

4.10 Supporting information

Acknowledgements

AT and NH acknowledge the generous support from the BBSRC SWBio Doctoral Training Program. BDE acknowledges that this work was generously supported by the Wellcome Trust Institutional Strategic Support Award (204909/Z/16/Z). We would like to specifically thank Rhys Cutlan of the Harmer group for his occasional wet lab help throughout the study. Finally, we would also like to thank the Harmer group for their constant support.

Author Information

The authors declare no competing interests.

Chapter 5

General discussion

5.1 ASR's place as an engineering tool

As discussed in the introduction, synthetic biology suffers bottlenecks based on treacherous paths to scale-up (Boehm and Bock, 2019). A large part of the scale-up issue comes from the difficulty in optimizing life as a system (Bommarius, 2015). It was discussed how the optimization of life for commercial use is a multi-faceted engineering challenge (Cardinale and Arkin, 2012). Dramatic improvements to the reading and writing of DNA has led to broad access to the tools needed for the rapid prototyping of pathways and strains (Smolke *et al.*, 2018). However, tools to optimize the proteins encoded by DNA are considerably lacking, being both slow and costly – somewhat mirroring synthetic DNA in the 80s (Bommarius, 2015). In protein engineering experiments, it is seen that engineering and optimizing proteins from a stable initiation point allows for the “flattening” of minima in the fitness landscape of a protein, and enables access to broader functional space (Porebski and Buckle, 2016; Bloom *et al.*, 2006; Goldsmith *et al.*, 2017). Additionally, it is argued that the utilization of high temperature biotransformations will allow for several functional optimizations to the synthetic system, including better substrate dilution, increased reaction rates, resistance to adverse conditions, avoidance of contamination, and maximization of yield-per-enzyme based on extended half-lives (Long *et al.*, 2018; Noordam *et al.*, 2018; Yeoman *et al.*, 2010; Gumulya *et al.*, 2018). Stable proteins may therefore become a cornerstone of future synthetic biology processes. However, the requirement of stable proteins merely shifts the bottleneck as the generation of stable proteins is still an engineering challenge.

In light of these challenges, this thesis explored the possibility of generating thermostable proteins at minimal expense, by designing methods to rapidly generate stable enzymes for the cost of their encoded synthetic DNA. According to accounting organization Deloitte, democratization of a technology involves its simplification, allowing users with diverse backgrounds to access the innovation with equal capabilities following a manageable learning curve (Schatsky *et al.*, 2018). Thus, democratization pushes technologies toward wide dissemination and ultimately leads to the acceleration of innovation (Frow, 2015). A democratized tool for the engineering of protein stability should to fulfil a number of requirements: it should generate small numbers of candidate sequences, it should be expert

agnostic, it should be easily accessible, the majority of use-cases should produce stable proteins, and each experiment should have a rapid turnaround (Endy *et al.*, 2005; Sun *et al.*, 2014). As discussed in the introduction, the use of the evolutionary biology tool ASR has the potential to fulfil these requirements.

On the conception of this PhD in 2014, the potential utilization of ASR as a direct and simplified protein engineering tool had only been discussed as a possibility in the literature (Wijma, 2013). At the time of writing this thesis, there is now a small but growing interest in the utilization of the tool, with a number of successful engineering attempts reported (Whitfield *et al.*, 2015; Babkova *et al.*, 2017; Gumulya *et al.*, 2018; Wilding *et al.*, 2017; Blanchet *et al.*, 2017; Zakas *et al.*, 2017). ASR is attractive, as it relies on free to access computational tools, and the input requirements are simply an alignment of homologous sequences and a phylogeny (Gumulya and Gillam 2017). Importantly, with the publication of works by Gumulya *et al.* (2018) and Trudeau *et al.* (2016), it has recently become apparent that ASR's stabilizing effect is not bespoke to the most ancient of proteins that inhabited higher temperature environments, as previously thought (Wijma, 2013). Here, stable proteins were generated from families with mesophilic ancestry.

5.2 Protein stabilization with ASR

In chapter 2 and chapter 4, we explored the use of ASR for the stabilization of carboxylic acid reductases (CARs). CARs are an important test case for ASR engineering, as their high dynamism, large structure, and multi-step reaction confer multiple points-of-failure to the engineering experiment. For ASR, CARs are also significant as they are only estimated to have emerged around 500 myo, from other members of the ANL superfamily (Finnigan *et al.*, 2017). and therefore it is most likely that they evolved in relatively mild ambient conditions. Despite such challenges, five of the six constructed ancestors were functional, and all of the functional sequences were more stable than the most stable modern protein (*M. avium*: 49°C). Of the functional sequences, four of the five attained increases in stability that are considered exceptional in the field (>15 °C). Therefore, this thesis presents strong evidence that ancestral reconstruction is a powerful tool for the engineering of protein stability without the need for rational or iterative workflows.

Gumulya *et al.*, 2018, hypothesised that the stability observed in the CYP3 cytochrome P450 monooxygenase ancestors was a result of the enzyme family evolving from a warmer ocean inhabiting ancestor. This hypothesis is also applicable to the CARs, yet without the ability to directly observe the LUCA of *Mycobacterium* and *Nocardia*, we cannot know whether it inhabited warmer environments. However, as more protein families with a mesophilic history are discovered to produce stable ancestors, the parsimony of the stable ancestor hypothesis decreases, and it becomes more evident that stabilization is an inherent feature of ASR. If so, it may be necessary to re-evaluate existing literature that draws hypotheses about the nature of ancient life from the stability of ancestral proteins. This is especially true when the observations are not backed up with real world data (i.e. Butzin *et al.*, 2013).

Chapter 3 provides evidence that survivor bias is the driving force behind protein stabilization in ancestral proteins derived from a mesophilic history. Survivor bias provides evidence for the broad spectrum nature of stabilization by ASR. It is understood that unless a selective pressure mandates significant stabilization of a protein family for its function (for example CutA1; Tanaka *et al.*, 2006), then the protein will evolve at marginality. As is shown by PESST simulations, the maintenance of marginality causes a disconnect between the distribution of stability contribution of mutations in the evolving dataset and the distribution of possible stability contributions in global sequence space. This disconnect causes ancestral sequence reconstruction (and consensus sequences) to sample from a net-stabilizing dataset, leading to the overestimation of stability. It can therefore be hypothesised any proteins that has evolved at marginality will exhibit overrepresented stabilizing mutations at a familial level.

By extension, the majority of protein families should generate a stabilized ancestor, as either a truly stable ancestor or an ancestor with biased stability will be derived from an ASR experiment if performed adequately. ASR engineering should therefore work for any dataset that allows for the bias to be incorporated into the ancestor. This prospect raises the question of whether other proteins in the reconstruction literature are also thermostable despite their stability not being reported (i.e. Finnigan *et al.*, 2011; Randall *et al.*, 2016; Shih *et al.*, 2016; Risso *et al.*, 2015). To better understand the requirements for

successful ASR engineering, it will be important to ask how many sequences should consist an alignment, how diverse the alignment should be, and how far back in time should sequences be resurrected? Gumulya *et al.*, 2018 present an ancestor of a 1,244 member dataset of KARI sequences reconstructed with a proprietary Bayesian reconstruction algorithm, showing stabilization of 15 °C compared to *E. coli* and *O. sativa* KARI. Comparatively, AspCAR-A50 generated from sASR is the most thermostable CAR variant observed to date, with an improvement of 25 °C over the most stable extant protein, and a stabilization of up to 38 °C compared to well-studied homologues. Yet it was generated from a dataset of 18 sequences, suggesting the number of sequences required to observe stabilization is relatively small. Even though it may be the case that the requirements for the most optimal ASR engineering experiment are unique to each protein family, the requirements for successful engineering appear to be surprisingly simple.

A fascinating future experiment to explore the question of reliability would involve the translation of the PESST model to the lab bench, where ASR would probe the stability of the ancestors of extensive experimental phylogenies generated by directed evolution under marginality (Randall *et al.*, 2016). The derivation of a stable ancestor in place of a known mesophilic ancestor of a real dataset would provide considerable evidence for both the survivor bias hypothesis and the use of ASR as a broad spectrum protein engineering tool. Additionally, subjecting multiple protein families to ASR based engineering efforts will be important, as a portfolio of evidence highlighting both success and failure should provide considerable insight into the tool's reliability. Such a study has recently been performed on consensus sequences, where 75% of sequences generated were shown to be more stable than their constituents (Sternke *et al.*, 2018). By generating consensus sequences of PESST simulations evolving at marginality in chapter 3, we observed that marginality also drives the stabilization of consensus sequences. As consensus and ancestral sequence stability is driven by the same underlying force, it can therefore be expected that at least equal success rates are possible with ASR.

It is important to reiterate that ASR is not necessarily the panacea of stability engineering, nor is it intended to be. For example, a key drawback that is observed in this thesis is loss of total turnover rate compared to extant counterparts under standard conditions following

the engineering process (somewhat in line with the trade-off hypothesis). This is especially the case with the most robust of proteins generated in this thesis by thorough ASR (AncCAR-PF), and both proteins generated with sASR. A similar observation is reported in Wilding *et al.* (2017), where decreased activity on the majority of substrates was observed in ancestral ω -transaminases. On the other hand, Gumulya *et al.*, 2018 and Babkova *et al.*, 2016 report improved activities of ancestral enzymes on a number of substrates, and AncCAR-A and AncCAR-PA presented in this theses represent the highest observed turnover of cinammic acid derivatives. Notwithstanding, considering the aim of an ASR engineering experiment is the engineering of stability, it should be expected that derived enzymes display improved activity when reactions are performed at high temperatures compared their extant mesostable counterparts.

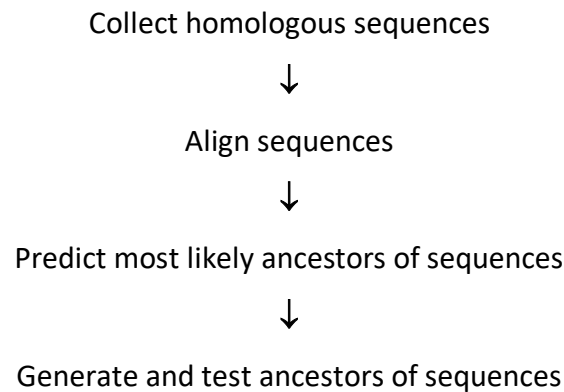
5.3 Accessible protein engineering with ASR

Considering technological adoption, Moore (2014) describes “a vast chasm” between early adopters and the early majority users of technology, where poor accessibility and lack of proven reliability halt wide adoption. In chapter 4 we aimed to develop a tool that potentially enables stability engineering and subsequently protein engineering to bridge this chasm in synthetic biology. While ASR is powerful in its own right, its maximum accessibility is stymied by the apparent need to generate accurate alignments and phylogenies to generate stable sequences (Vialle *et al.*, 2018). sASR solves this key accessibility issue by utilizing Ancescon, a reconstruction algorithm that generates its own rate matrix for amino acid substitution, and can be run without a phylogeny. While the proteins that are generated by sASR are unlikely to bare semblance to the true ancestor, we have shown that sASR generates both functional and thermostable protein variants.

In aid of accessibility, a set of simple criteria are defined that allow users to select valuable sequences from the generated phylogeny based on the weighted balance of nodes. Using PESST simulations, we identified that the generation of just two ancestral sequences from the Ancescon tree can reliably generate exceptionally thermostable proteins.

Comparatively, the typical generation of hundreds of sequences over numerous iterations was required to identify outstandingly stable variants with semi-rational protein engineering

approaches, as discussed in the introduction. Even compared to the ASR workflow discussed in the introduction, the workflow for the engineering of stable proteins is now considerably reduced with sASR, to:



Such a workflow is therefore only inhibited by the speed and cost of DNA synthesis, and the lack of significant data supporting its continued successful application. As with ASR, sASR will benefit significantly from a comparative study that targets numerous enzyme families for the engineering of stability.

Additionally, in the introduction, the importance of accessible stability engineering to general protein engineering workflows was outlined. Synthetic biology is already beginning to see the front-loading of engineering experiments with stable enzymes (Trudeau *et al.*, 2018; Goldsmith *et al.*, 2017). Gumulya *et al.*, 2018 presented the first example of an ASR product being used for the starting point of subsequent protein engineering. Thermostable ancestral CYP3 was subject to further directed evolution based on ambiguous sites in the posterior probability of the reconstruction. Approximately 20% of the 1023 mutants screened were more stable than the initiation point, and the library showed functional diversity throughout. Further experiments assessing the value of sASR should therefore focus on the front-loading of synthetic biology workflows with thermostable proteins from sASR. Furthermore, the Ancescon output does provide a table of posterior probability, allowing for the ASR output to guide library design required for subsequent engineering experiments, again lowering the barrier to access functional diversity in synthetic biology.

5.4 Developments in the modelling of protein evolution

In chapter 3 we produced PESST, a highly parameterizable evolution modelling toolbox that allows for the testing of hypotheses around thresholds in protein evolution. With PESST we were able to provide considerable evidence that stabilization doesn't always derive from the protein's ancestral history or a consensus effect as was previously predicted, and instead stabilization is driven by survivor bias. Conception and evidence for survivor bias provides considerable validity to the broad spectrum applicability of protein engineering. As PESST is designed for the testing of the outcomes of various evolutionary hypotheses, the application of PESST to the development of sASR was simple, as the standard set of parameters are easily modifiable to fit another stability-centric experiment.

Importantly, in the design of PESST, we opted to not define stability based on real world data on Gibbs free energy contributions, and instead chose to model these data through a matrix of stability contributions that can be modelled to a gaussian distribution. As a result, we are able to generate an adequate model of real Gibbs free energy data (Tokuriki *et al.*, 2007; Faure and Koonin, 2015), but are also able to adapt the distribution describing stability effects as we please. Conceptually, this distribution, can be abstracted from stability, and be described to model "trait contribution", where in theory any trait described by a gaussian distribution across sites can be modelled. A key future direction of PESST will be the development of the tool to allow for multi-variate analyses, including the constraint of multiple traits in tandem. This will include the definition of multiple thresholds, multiple contribution matrices, and the provision of tools for the manual definition of various trait distribution shapes. Finally, every amino acid is considered independently within PESST. In nature, protein properties can manifest from the co-evolution of sites (Sandler *et al.*, 2014; Bloom *et al.*, 2005). Therefore, addition of the ability to model co-operativity will dramatically increase the functionality and applicability of PESST to a broad spectrum of evolutionary queries.

5.5 Conclusions

Considering the immediate requirement of thermostable proteins for improved synthetic biology workflows, and the unattractive state of protein engineering based on its inherent

complexity, the need for accessible engineering tools is only set to increase in the future. ASR and sASR represent the first generation of tools that may be able to cross the chasm toward broadly accessible protein engineering tools. Therefore, the democratization of protein engineering is set to provide smoother roadmaps to market for synthetic biology technologies, and allow more researchers to develop synthetic biology applications beyond a proof of concept. Such alignment based engineering tools could enable the engineering of a broad spectrum of proteins for stabilization, based on their bias driven functionality, and their ability to engineer large, complex enzymes that lack a complete crystal structure like the CARs. Additionally, as ASR and sASR are driven by the evolution of proteins, tools to model protein evolution like PESST will allow for the improved development of such tools going forward. This thesis therefore provides the groundwork for broad adoption of ancestral reconstruction as an engineering tool.

Chapter 6

Bibliography

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. <https://doi.org/10.1093/bioinformatics/bti263>

Abdel-Banat, B.M.A., Hoshida, H., Ano, A., Nonklang, S., Akada, R., 2010. High-temperature fermentation: how can processes for ethanol production at high temperatures become superior to the traditional process using mesophilic yeast? *Appl Microbiol Biotechnol* 85, 861–867. <https://doi.org/10.1007/s00253-009-2248-5>

Adams, B.L., 2016. The Next Generation of Synthetic Biology Chassis: Moving Synthetic Biology from the Laboratory to the Field. *ACS Synth. Biol.* 5, 1328–1330. <https://doi.org/10.1021/acssynbio.6b00256>

Agapakis, C.M., 2014. Designing synthetic biology. *ACS Synth Biol* 3, 121–128. <https://doi.org/10.1021/sb4001068>

Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.-C., Marquez, M., Klibanov, A.M., Griffiths, A.D., Weitz, D.A., 2010. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *PNAS* 107, 4004–4009. <https://doi.org/10.1073/pnas.0910781107>

Akanuma, S., 2017. Characterization of Reconstructed Ancestral Proteins Suggests a Change in Temperature of the Ancient Biosphere. *Life (Basel)* 7. <https://doi.org/10.3390/life7030033>

Akanuma, S., Nakajima, Y., Yokobori, S., Kimura, M., Nemoto, N., Mase, T., Miyazono, K., Tanokura, M., Yamagishi, A., 2013. Experimental evidence for the thermophilicity of ancestral life. *PNAS* 110, 11067–11072. <https://doi.org/10.1073/pnas.1308215110>

Akhtar, M.K., Turner, N.J., Jones, P.R., 2013. Carboxylic acid reductase is a versatile enzyme for the conversion of fatty acids into fuels and chemical commodities. *Proc. Natl. Acad. Sci. U.S.A.* 110, 87–92. <https://doi.org/10.1073/pnas.1216516110>

Akram, F., Haq, I. ul, Imran, W., Mukhtar, H., 2018. Insight perspectives of thermostable endoglucanases for bioethanol production: A review. *Renewable Energy* 122, 225–238. <https://doi.org/10.1016/j.renene.2018.01.095>

Alcolombri, U., Elias, M., Tawfik, D.S., 2011. Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. *J. Mol. Biol.* 411, 837–853. <https://doi.org/10.1016/j.jmb.2011.06.037>

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., Ronquist, F., 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20, 407–415. <https://doi.org/10.1093/bioinformatics/btg427>

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

Amos, M., 2014. Population-based microbial computing: a third wave of synthetic biology? *International Journal of General Systems* 43, 770–782. <https://doi.org/10.1080/03081079.2014.921001>

Amyris, 2018. Amyris Reports Third Quarter 2018 Financial Results [WWW Document]. Amyris. URL <http://investors.amyris.com/news-releases/news-release-details/amyris-reports-third-quarter-2018-financial-results> (accessed 11.22.18).

Anastas, P., Eghbali, N., 2010. Green Chemistry: Principles and Practice. *Chemical Society Reviews* 39, 301–312. <https://doi.org/10.1039/B918763B>

Angelastro, A., Dawson, W.M., Luk, L.Y.P., Allemann, R.K., 2016. A Versatile Disulfide-Driven Recycling System for NADP⁺ with High Cofactor Turnover Number. *ACS Catal.* 1025–1029. <https://doi.org/10.1021/acscatal.6b03061>

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., Gascuel, O., 2011. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Syst Biol* 60, 685–699. <https://doi.org/10.1093/sysbio/syr041>

Anselme Payen, Jean-François Persoz, 1833. Mémoire sur la Diastase, les principaux Produits de ses Réactions, et leurs applications aux arts industriels, 53rd ed, *Annales de Chimie et de Physique*.

Arabnejad, H., Dal Lago, M., Jekel, P.A., Floor, R.J., Thunnissen, A.-M.W.H., Terwisscha van Scheltinga, A.C., Wijma, H.J., Janssen, D.B., 2017. A robust cosolvent-compatible halohydrin dehalogenase by computational library design. *Protein Eng. Des. Sel.* 30, 173–187. <https://doi.org/10.1093/protein/gzw068>

Armstrong, E.F., 1933. Enzymes: A Discovery and its Consequences. *Nature* 131, 535–537. <https://doi.org/10.1038/131535a0>

Arnold, F.H., 2018. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed. Engl.* 57, 4143–4148. <https://doi.org/10.1002/anie.201708408>

Arnold, F.H., Volkov, A.A., 1999. Directed evolution of biocatalysts. *Curr Opin Chem Biol* 3, 54–59.

- Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., Pupko, T., 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40, W580-584. <https://doi.org/10.1093/nar/gks498>
- Asial, I., Cheng, Y.X., Engman, H., Dollhopf, M., Wu, B., Nordlund, P., Cornvik, T., 2013. Engineering protein thermostability using a generic activity-independent biophysical screen inside the cell. *Nat Commun* 4, 2901. <https://doi.org/10.1038/ncomms3901>
- Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R., Damborsky, J., 2017. Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem* 18, 1448–1456. <https://doi.org/10.1002/cbic.201700197>
- Balasco, N., Esposito, L., Simone, A.D., Vitagliano, L., 2013. Role of loops connecting secondary structure elements in the stabilization of proteins isolated from thermophilic organisms. *Protein Sci* 22, 1016–1023. <https://doi.org/10.1002/pro.2279>
- Baltes, N.J., Voytas, D.F., 2015. Enabling plant synthetic biology through genome engineering. *Trends in Biotechnology, Special Issue: Manifesting Synthetic Biology* 33, 120–131. <https://doi.org/10.1016/j.tibtech.2014.11.008>
- Bar-Rogovsky, H., Stern, A., Penn, O., Kobl, I., Pupko, T., Tawfik, D.S., 2015. Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Engineering, Design and Selection* 28, 507–518. <https://doi.org/10.1093/protein/gzv038>
- Baum, D.A., Smith, S.D., 2012. *Tree Thinking: An Introduction to Phylogenetic Biology*. Roberts, Greenwood Village, Colo.
- Bednar, D., Beerens, K., Sebestova, E., Bendl, J., Khare, S., Chaloupkova, R., Prokop, Z., Brezovsky, J., Baker, D., Damborsky, J., 2015. FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants. *PLOS Computational Biology* 11, e1004556. <https://doi.org/10.1371/journal.pcbi.1004556>
- Bendl, J., Stourac, J., Sebestova, E., Vavra, O., Musil, M., Brezovsky, J., Damborsky, J., 2016. HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res* 44, W479–W487. <https://doi.org/10.1093/nar/gkw416>
- Benkovic, S.J., Hammes-Schiffer, S., 2003. A Perspective on Enzyme Catalysis. *Science* 301, 1196–1202. <https://doi.org/10.1126/science.1085515>
- Benn, F., Haley, N.E.C., Lucas, A.E., Silvester, E., Helmi, S., Schreiber, R., Bath, J., Turberfield, A.J., 2018. Chiral DNA Origami Nanotubes with Well-Defined and Addressable Inside and Outside Surfaces. *Angewandte Chemie* 130, 7813–7816. <https://doi.org/10.1002/ange.201800275>

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. *Nucleic Acids Res.* 41, D36-42. <https://doi.org/10.1093/nar/gks1195>

Bershtein, S., Goldin, K., Tawfik, D.S., 2008. Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins. *Journal of Molecular Biology* 379, 1029–1044. <https://doi.org/10.1016/j.jmb.2008.04.024>

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., Tawfik, D.S., 2006. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444, 929–932. <https://doi.org/10.1038/nature05385>

Betts, M.J., Russell, R.B., 2003. Amino Acid Properties and Consequences of Substitutions, in: *Bioinformatics for Geneticists*. John Wiley & Sons, Ltd, pp. 289–316. <https://doi.org/10.1002/0470867302.ch14>

Betz, S.F., 1993. Disulfide bonds and the stability of globular proteins. *Protein Sci* 2, 1551–1558.

Blanchet, G., Alili, D., Protte, A., Upert, G., Gilles, N., Tepshi, L., Stura, E.A., Mourier, G., Servent, D., 2017. Ancestral protein resurrection and engineering opportunities of the mamba aminergic toxins. *Scientific Reports* 7, 2701. <https://doi.org/10.1038/s41598-017-02953-0>

Bloom, J.D., Glassman, M.J., 2009. Inferring Stabilizing Mutations from Protein Phylogenies: Application to Influenza Hemagglutinin. *PLOS Computational Biology* 5, e1000349. <https://doi.org/10.1371/journal.pcbi.1000349>

Bloom, J.D., Labthavikul, S.T., Otey, C.R., Arnold, F.H., 2006. Protein stability promotes evolvability. *PNAS* 103, 5869–5874. <https://doi.org/10.1073/pnas.0510098103>

Bloom, J.D., Silberg, J.J., Wilke, C.O., Drummond, D.A., Adami, C., Arnold, F.H., 2005. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci U S A* 102, 606–611. <https://doi.org/10.1073/pnas.0406744102>

Boehm, C.R., Bock, R., 2019. Recent advances and current challenges in synthetic biology of the plastid genetic system and metabolism. *Plant Physiology* pp.00767.2018. <https://doi.org/10.1104/pp.18.00767>

Bommarius, A.S., 2015. Biocatalysis: A Status Report. *Annual Review of Chemical and Biomolecular Engineering* 6, 319–345. <https://doi.org/10.1146/annurev-chembioeng-061114-123415>

Bommarius, A.S., Blum, J.K., Abrahamson, M.J., 2011. Status of protein engineering for biocatalysts: how to design an industrially useful biocatalyst. *Curr Opin Chem Biol* 15, 194–200. <https://doi.org/10.1016/j.cbpa.2010.11.011>

- Born, B., Kim, S.J., Ebbinghaus, S., Gruebele, M., Havenith, M., 2008. The terahertz dance of water with the proteins: the effect of protein flexibility on the dynamical hydration shell of ubiquitin. *Faraday Discuss.* 141, 161–173. <https://doi.org/10.1039/B804734K>
- Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., Robins, K., 2012. Engineering the third wave of biocatalysis. *Nature* 485, 185–194. <https://doi.org/10.1038/nature11117>
- Bosshard, H.R., Marti, D.N., Jelesarov, I., 2004. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J. Mol. Recognit.* 17, 1–16. <https://doi.org/10.1002/jmr.657>
- Boucher, J.I., Cote, P., Flynn, J., Jiang, L., Laban, A., Mishra, P., Roscoe, B.P., Bolon, D.N.A., 2014. Viewing Protein Fitness Landscapes Through a Next-Gen Lens. *Genetics* 198, 461–471. <https://doi.org/10.1534/genetics.114.168351>
- Bozkurt, E., Perez, M.A.S., Hovius, R., Browning, N.J., Rothlisberger, U., 2018. Genetic Algorithm Based Design and Experimental Characterization of a Highly Thermostable Metalloprotein. *J. Am. Chem. Soc.* 140, 4517–4521. <https://doi.org/10.1021/jacs.7b10660>
- Brown, K.M., Costanzo, M.S., Xu, W., Roy, S., Lozovsky, E.R., Hartl, D.L., 2010. Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase. *Mol Biol Evol* 27, 2682–2690. <https://doi.org/10.1093/molbev/msq160>
- Bruno, W.J., Socci, N.D., Halpern, A.L., 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* 17, 189–197. <https://doi.org/10.1093/oxfordjournals.molbev.a026231>
- Buchholz, P.C.F., Zeil, C., Pleiss, J., 2018. The scale-free nature of protein sequence space. *PLOS ONE* 13, e0200815. <https://doi.org/10.1371/journal.pone.0200815>
- Buchko, G.W., Abendroth, J., Clifton, M.C., Robinson, H., Zhang, Y., Hewitt, S.N., Staker, B.L., Edwards, T.E., Van Voorhis, W.C., Myler, P.J., 2015. Structure of a CutA1 divalent-cation tolerance protein from *Cryptosporidium parvum*, the protozoal parasite responsible for cryptosporidiosis. *Acta Crystallogr F Struct Biol Commun* 71, 522–530. <https://doi.org/10.1107/S2053230X14028210>
- Butzin, N.C., Lapierre, P., Green, A.G., Swithers, K.S., Gogarten, J.P., Noll, K.M., 2013. Reconstructed Ancestral Myo-Inositol-3-Phosphate Synthases Indicate That Ancestors of the Thermococcales and Thermotoga Species Were More Thermophilic than Their Descendants. *PLoS One* 8. <https://doi.org/10.1371/journal.pone.0084300>

- Cai, W., Pei, J., Grishin, N.V., 2004. Reconstruction of ancestral protein sequences and its applications. *BMC Evolutionary Biology* 4, 33. <https://doi.org/10.1186/1471-2148-4-33>
- Cameron, D.E., Bashor, C.J., Collins, J.J., 2014. A brief history of synthetic biology. *Nat Rev Micro* 12, 381–390. <https://doi.org/10.1038/nrmicro3239>
- Canton, B., Labno, A., Endy, D., 2008. Refinement and standardization of synthetic biological parts and devices. *Nature Biotechnology* 26, 787–793. <https://doi.org/10.1038/nbt1413>
- Carbonell, P., Currin, A., Dunstan, M., Fellows, D., Jervis, A., Rattray, N.J.W., Robinson, C.J., Swainston, N., Vinaixa, M., Williams, A., Yan, C., Barran, P., Breitling, R., Chen, G.G.-Q., Faulon, J.-L., Goble, C., Goodacre, R., Kell, D.B., Feuvre, R.L., Micklefield, J., Scrutton, N.S., Shapira, P., Takano, E., Turner, N.J., 2016. SYNBIOCHEM—a SynBio foundry for the biosynthesis and sustainable production of fine and speciality chemicals. *Biochem. Soc. Trans.* 44, 675–677. <https://doi.org/10.1042/BST20160009>
- Cardinale, S., Arkin, A.P., 2012. Contextualizing context for synthetic biology – identifying causes of failure of synthetic biological systems. *Biotechnology Journal* 7, 856–866. <https://doi.org/10.1002/biot.201200085>
- Casini, A., Storch, M., Baldwin, G.S., Ellis, T., 2015. Bricks and blueprints: methods and standards for DNA assembly. *Nature Reviews Molecular Cell Biology* 16, 568–576. <https://doi.org/10.1038/nrm4014>
- Chaliotis, A., Vlastaridis, P., Mossialos, D., Ibba, M., Becker, H.D., Stathopoulos, C., Amoutzias, G.D., 2017. The complex evolutionary history of aminoacyl-tRNA synthetases. *Nucleic Acids Res* 45, 1059–1068. <https://doi.org/10.1093/nar/gkw1182>
- Chandran, D., Copeland, W.B., Sleight, S.C., Sauro, H.M., 2008. Mathematical modeling and synthetic biology. *Drug Discovery Today: Disease Models, Animal models of addiction / Kinetic models* 5, 299–309. <https://doi.org/10.1016/j.ddmod.2009.07.002>
- Chapman, J., Ismail, A.E., Dinu, C.Z., 2018. Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks. *Catalysts* 8, 238. <https://doi.org/10.3390/catal8060238>
- Chatterjee, C., Pong, F., Sen, A., 2014. Chemical conversion pathways for carbohydrates. *Green Chem.* 17, 40–71. <https://doi.org/10.1039/C4GC01062K>
- Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., Mu, R., 2013. The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Genomics, Proteomics & Bioinformatics* 11, 34–40. <https://doi.org/10.1016/j.gpb.2013.01.003>

Chen, F., Gaucher, E.A., Leal, N.A., Hutter, D., Havemann, S.A., Govindarajan, S., Ortlund, E.A., Benner, S.A., 2010. Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. *PNAS* 107, 1948–1953. <https://doi.org/10.1073/pnas.0908463107>

Chen, M.M.Y., Snow, C.D., Vizcarra, C.L., Mayo, S.L., Arnold, F.H., 2012. Comparison of random mutagenesis and semi-rational designed libraries for improved cytochrome P450 BM3-catalyzed hydroxylation of small alkanes. *Protein Eng Des Sel* 25, 171–178. <https://doi.org/10.1093/protein/gzs004>

Cheng, G., Varanasi, P., Li, C., Liu, H., Melnichenko, Y.B., Simmons, B.A., Kent, M.S., Singh, S., 2011. Transition of Cellulose Crystalline Structure and Surface Morphology of Biomass as a Function of Ionic Liquid Pretreatment and Its Relation to Enzymatic Hydrolysis. *Biomacromolecules* 12, 933–941. <https://doi.org/10.1021/bm101240z>

Cherny, I., Greisen, P., Ashani, Y., Khare, S.D., Oberdorfer, G., Leader, H., Baker, D., Tawfik, D.S., 2013. Engineering V-type nerve agents detoxifying enzymes using computationally focused libraries. *ACS Chem. Biol.* 8, 2394–2403. <https://doi.org/10.1021/cb4004892>

Cherry, J.R., Fidantsef, A.L., 2003. Directed evolution of industrial enzymes: an update. *Current Opinion in Biotechnology* 14, 438–443. [https://doi.org/10.1016/S0958-1669\(03\)00099-5](https://doi.org/10.1016/S0958-1669(03)00099-5)

Christensen, N.J., Kepp, K.P., 2012. Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol. *J. Chem. Inf. Model.* 52, 3028–3042. <https://doi.org/10.1021/ci300398z>

Chubukov, V., Mukhopadhyay, A., Petzold, C.J., Keasling, J.D., Martín, H.G., 2016. Synthetic and systems biology for microbial production of commodity chemicals. *npj Systems Biology and Applications* 2, 16009. <https://doi.org/10.1038/npjbsa.2016.9>

Cirino, P.C., Mayer, K.M., Umeno, D., 2003. Generating Mutant Libraries Using Error-Prone PCR, in: Arnold, F.H., Georgiou, G. (Eds.), *Directed Evolution Library Creation: Methods and Protocols*, *Methods in Molecular Biology™*. Humana Press, Totowa, NJ, pp. 3–9. <https://doi.org/10.1385/1-59259-395-X:3>

Clarke, L., Adams, J., Sutton, P., Bainbridge, J.W., Birney, E., Calvert, J., Collis, A., Kitney, R., Freemont, P., Mason, P., Pandya, K., Ghaffar, T., Rose, N., Marris, C., Woolfson, D., Boyce, A., 2012. A Synthetic Biology Roadmap for the UK [WWW Document]. URL <https://connect.innovateuk.org/web/synthetic-biology-special-interest-group/roadmap-for-synthetic-biology> (accessed 11.23.18).

Clifton, B.E., Whitfield, J.H., Sanchez-Romero, I., Herde, M.K., Henneberger, C., Janovjak, H., Jackson, C.J., 2017. Ancestral Protein Reconstruction and Circular Permutation for Improving the Stability and

Dynamic Range of FRET Sensors. *Methods Mol. Biol.* 1596, 71–87. https://doi.org/10.1007/978-1-4939-6940-1_5

Coddington, J., Scharff, N., 1995. Problems with Zero-Length Branches. *Cladistics* 10, 415–423. <https://doi.org/10.1111/j.1096-0031.1994.tb00187.x>

Cole, M.F., Cox, V.E., Gratton, K.L., Gaucher, E.A., 2013. Reconstructing evolutionary adaptive paths for protein engineering. *Methods Mol. Biol.* 978, 115–125. https://doi.org/10.1007/978-1-62703-293-3_8

Cole, M.F., Gaucher, E.A., 2011. Exploiting Models of Molecular Evolution to Efficiently Direct Protein Engineering. *J Mol Evol* 72, 193–203. <https://doi.org/10.1007/s00239-010-9415-2>

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L., Zhang, F., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 1231143. <https://doi.org/10.1126/science.1231143>

Conti, G., Pollegioni, L., Molla, G., Rosini, E., 2014. Strategic manipulation of an industrial biocatalyst-evolution of a cephalosporin C acylase. *FEBS J.* 281, 2443–2455. <https://doi.org/10.1111/febs.12798>

Cooper, G.M., 2000. *The Central Role of Enzymes as Biological Catalysts. The Cell: A Molecular Approach.* 2nd edition.

Corbion, 2017. Corbion signs agreement to bid for substantially all of the assets of innovative microalgae specialist TerraVia [WWW Document]. GlobeNewswire News Room. URL <http://globenewswire.com/news-release/2017/08/02/1070712/0/en/Corbion-signs-agreement-to-bid-for-substantially-all-of-the-assets-of-innovative-microalgae-specialist-TerraVia.html> (accessed 11.22.18).

Cornish-Bowden, A., 2013. *Fundamentals of Enzyme Kinetics.* John Wiley & Sons.

Cox, A., Chen, S., 2018. De Novo Synthesized Nucleic Acid Libraries. 20180051278.

Currin, A., Swainston, N., Day, P.J., Kell, D.B., 2015. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.* 44, 1172–1239. <https://doi.org/10.1039/C4CS00351A>

Dagan, S., Hagai, T., Gavrilov, Y., Kapon, R., Levy, Y., Reich, Z., 2013. Stabilization of a protein conferred by an increase in folded state entropy. *Proc Natl Acad Sci U S A* 110, 10628–10633. <https://doi.org/10.1073/pnas.1302284110>

Dahanayake, J.N., Mitchell-Koch, K.R., 2018. How Does Solvation Layer Mobility Affect Protein Structural Dynamics? *Front Mol Biosci* 5, 65. <https://doi.org/10.3389/fmolb.2018.00065>

Dahl, R.H., Zhang, F., Alonso-Gutierrez, J., Baidoo, E., Batth, T.S., Redding-Johanson, A.M., Petzold, C.J., Mukhopadhyay, A., Lee, T.S., Adams, P.D., Keasling, J.D., 2013. Engineering dynamic pathway regulation using stress-response promoters. *Nat Biotech* 31, 1039–1046. <https://doi.org/10.1038/nbt.2689>

Dai, M., Fisher, H.E., Temirov, J., Kiss, C., Phipps, M.E., Pavlik, P., Werner, J.H., Bradbury, A.R.M., 2007. The creation of a novel fluorescent protein by guided consensus engineering. *Protein Eng. Des. Sel.* 20, 69–79. <https://doi.org/10.1093/protein/gzl056>

Daniel, R.M., Danson, M.J., 2013. Temperature and the catalytic activity of enzymes: A fresh understanding. *FEBS Letters, A century of Michaelis - Menten kinetics* 587, 2738–2743. <https://doi.org/10.1016/j.febslet.2013.06.027>

Daniel, R.M., Peterson, M.E., Danson, M.J., Price, N.C., Kelly, S.M., Monk, C.R., Weinberg, C.S., Oudshoorn, M.L., Lee, C.K., 2010. The molecular basis of the effect of temperature on enzyme activity. *Biochemical Journal* 425, 353–360. <https://doi.org/10.1042/BJ20091254>

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165. <https://doi.org/10.1093/bioinformatics/btr088>

Daudé, D., Topham, C.M., Remaud-Siméon, M., André, I., 2013. Probing impact of active site residue mutations on stability and activity of *Neisseria polysaccharea* amylosucrase. *Protein Sci* 22, 1754–1765. <https://doi.org/10.1002/pro.2375>

David Westman, 2017. Cost and impact of a bioburden incident [WWW Document]. GE Healthcare Life Sciences. URL <https://www.gelifesciences.com/en/gb/news-center/cost-and-impact-of-a-bioburden-incident-10001> (accessed 11.24.18).

de Visser, J.A.G.M., Krug, J., 2014. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* 15, 480–490. <https://doi.org/10.1038/nrg3744>

DeFrancesco, L., 2017. Hanging on a thread. *Nature Biotechnology* 35, 496–499. <https://doi.org/10.1038/nbt.3894>

DeLano, W.L., 2002. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA. <http://www.pymol.org>.

Denard, C.A., Ren, H., Zhao, H., 2015. Improving and repurposing biocatalysts via directed evolution. *Current Opinion in Chemical Biology, Biocatalysis and biotransformation • Bioinorganic chemistry* 25, 55–64. <https://doi.org/10.1016/j.cbpa.2014.12.036>

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., Gascuel, O., 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36, W465-469. <https://doi.org/10.1093/nar/gkn180>

Dhar, A., Minin, V.N., 2015. *Maximum Likelihood Methods for Phylogenetic Inference.*

Dickson, R.J., Wahl, L.M., Fernandes, A.D., Gloor, G.B., 2010. Identifying and Seeing beyond Multiple Sequence Alignment Errors Using Intra-Molecular Protein Covariation. *PLOS ONE* 5, e11082. <https://doi.org/10.1371/journal.pone.0011082>

Dill, K.A., 1990. Dominant forces in protein folding. *Biochemistry* 29, 7133–7155. <https://doi.org/10.1021/bi00483a001>

Directorate-General for Research European Commission, 2005. *Synthetic Biology: Applying Engineering to Biology : Report of a Nest High-Level Expert Group.* Office for Official Publications of the European Communities, Luxembourg.

Dominy, B.N., Perl, D., Schmid, F.X., Brooks, C.L., 2002. The effects of ionic strength on protein stability: the cold shock protein family. *J. Mol. Biol.* 319, 541–554. [https://doi.org/10.1016/S0022-2836\(02\)00259-0](https://doi.org/10.1016/S0022-2836(02)00259-0)

Doudna, J.A., Charpentier, E., 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346, 1258096. <https://doi.org/10.1126/science.1258096>

Dryden, D.T.F., Thomson, A.R., White, J.H., 2008. How much of protein sequence space has been explored by life on Earth? *Journal of The Royal Society Interface* 5, 953–956. <https://doi.org/10.1098/rsif.2008.0085>

Duan, Y., Kollman, P.A., 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282, 740–744.

Durani, V., Magliery, T.J., 2013. Protein engineering and stabilization from sequence statistics: variation and covariation analysis. *Meth. Enzymol.* 523, 237–256. <https://doi.org/10.1016/B978-0-12-394292-0.00011-4>

Dvorak, P., Bidmanova, S., Damborsky, J., Prokop, Z., 2014. Immobilized synthetic pathway for biodegradation of toxic recalcitrant pollutant 1,2,3-trichloropropane. *Environ. Sci. Technol.* 48, 6859–6866. <https://doi.org/10.1021/es500396r>

Ebbinghaus, S., Kim, S.J., Heyden, M., Yu, X., Gruebele, M., Leitner, D.M., Havenith, M., 2008. Protein Sequence- and pH-Dependent Hydration Probed by Terahertz Spectroscopy. *J. Am. Chem. Soc.* 130, 2374–2375. <https://doi.org/10.1021/ja0746520>

Ebbinghaus, S., Kim, S.J., Heyden, M., Yu, X., Heugen, U., Gruebele, M., Leitner, D.M., Havenith, M., 2007. An extended dynamical hydration shell around proteins. *Proc Natl Acad Sci U S A* 104, 20749–20752. <https://doi.org/10.1073/pnas.0709207104>

Eberhardt, F., Aguirre, A., Paoletti, L., Hails, G., Braia, M., Ravasi, P., Peiru, S., Menzella, H.G., 2018. Pilot-scale process development for low-cost production of a thermostable biodiesel refining enzyme in *Escherichia coli*. *Bioprocess Biosyst Eng* 41, 555–564. <https://doi.org/10.1007/s00449-018-1890-7>

Ebert, M.C., Pelletier, J.N., 2017. Computational tools for enzyme improvement: why everyone can – and should – use them. *Current Opinion in Chemical Biology, Biocatalysis & biotransformation* * *Bioinorganic Chemistry* 37, 89–96. <https://doi.org/10.1016/j.cbpa.2017.01.021>

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792–1797. <https://doi.org/10.1093/nar/gkh340>

Edge, 2008. ENGINEERING BIOLOGY - An interview with Drew Endy | Edge.org [WWW Document]. URL https://www.edge.org/conversation/drew_endy-engineering-biology (accessed 12.20.18).

Eick, G.N., Bridgham, J.T., Anderson, D.P., Harms, M.J., Thornton, J.W., 2017. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Mol Biol Evol* 34, 247–261. <https://doi.org/10.1093/molbev/msw223>

Eijsink, V.G.H., Gåseidnes, S., Borchert, T.V., van den Burg, B., 2005. Directed evolution of enzyme stability. *Biomolecular Engineering, Directed Enzyme Evolution* 22, 21–30. <https://doi.org/10.1016/j.bioeng.2004.12.003>

Elleuche, S., Schröder, C., Sahm, K., Antranikian, G., 2014. Extremozymes—biocatalysts with unique properties from extremophilic microorganisms. *Current Opinion in Biotechnology, Cell and Pathway Engineering* 29, 116–123. <https://doi.org/10.1016/j.copbio.2014.04.003>

Endy, D., 2005. Foundations for engineering biology. *Nature* 438, 449–453. <https://doi.org/10.1038/nature04342>

Endy, D., 2003. 2003 Synthetic Biology study (Working Paper).

Erb, T.J., Jones, P.R., Bar-Even, A., 2017. Synthetic metabolism: metabolic engineering meets enzyme design. *Current Opinion in Chemical Biology, Biocatalysis & biotransformation * Bioinorganic Chemistry* 37, 56–62. <https://doi.org/10.1016/j.cbpa.2016.12.023>

Fang, Y., 2014. A Gibbs free energy formula for protein folding derived from quantum statistics. *Sci. China Phys. Mech. Astron.* 57, 1547–1551. <https://doi.org/10.1007/s11433-013-5288-x>

Fasan, R., Chen, M.M., Crook, N.C., Arnold, F.H., 2007. Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting natively-like catalytic properties. *Angew. Chem. Int. Ed. Engl.* 46, 8414–8418. <https://doi.org/10.1002/anie.200702616>

Faure, G., Koonin, E.V., 2015. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys Biol* 12, 035001. <https://doi.org/10.1088/1478-3975/12/3/035001>

Feeney, M., Punja, Z.K., 2017. The Role of Agrobacterium-Mediated and Other Gene-Transfer Technologies in Cannabis Research and Product Development, in: Chandra, S., Lata, H., ElSohly, M.A. (Eds.), *Cannabis Sativa L. - Botany and Biotechnology*. Springer International Publishing, Cham, pp. 343–363. https://doi.org/10.1007/978-3-319-54564-6_16

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791. <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x>

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17, 368–376. <https://doi.org/10.1007/BF01734359>

Feng, D.F., Doolittle, R.F., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25, 351–360.

Fenila, F., Shastri, Y., 2016. Optimal control of enzymatic hydrolysis of lignocellulosic biomass. *Resource-Efficient Technologies, Special Issue on Technoscape-2016* 2, S96–S104. <https://doi.org/10.1016/j.reffit.2016.11.006>

Ferrandi, E.E., Sayer, C., Isupov, M.N., Annovazzi, C., Marchesi, C., Iacobone, G., Peng, X., Bonch-Osmolovskaya, E., Wohlgemuth, R., Littlechild, J.A., Monti, D., 2015. Discovery and characterization of thermophilic limonene-1,2-epoxide hydrolases from hot spring metagenomic libraries. *FEBS J.* 282, 2879–2894. <https://doi.org/10.1111/febs.13328>

Fersht, A., 2017. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. World Scientific.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. *Nucleic Acids Res.* 42, D222-230. <https://doi.org/10.1093/nar/gkt1223>

Finnigan, G.C., Hanson-Smith, V., Houser, B.D., Park, H.J., Stevens, T.H., 2011. The reconstructed ancestral subunit a functions as both V-ATPase isoforms Vph1p and Stv1p in *Saccharomyces cerevisiae*. *Mol. Biol. Cell* 22, 3176–3191. <https://doi.org/10.1091/mbc.E11-03-0244>

Finnigan, W., Thomas, A., Cromar, H., Gough, B., Snajdrova, R., Adams, J.P., Littlechild, J.A., Harmer, N.J., 2017a. Characterization of Carboxylic Acid Reductases as Enzymes in the Toolbox for Synthetic Chemistry. *ChemCatChem* 9, 1005–1017. <https://doi.org/10.1002/cctc.201601249>

Finnigan, W., Thomas, A., Cromar, H., Gough, B., Snajdrova, R., Adams, J.P., Littlechild, J.A., Harmer, N.J., 2017b. Characterization of Carboxylic Acid Reductases as Enzymes in the Toolbox for Synthetic Chemistry. *ChemCatChem* 9, 1005–1017. <https://doi.org/10.1002/cctc.201601249>

France, S.P., Hussain, S., Hill, A.M., Hepworth, L.J., Howard, R.M., Mulholland, K.R., Flitsch, S.L., Turner, N.J., 2016. One-Pot Cascade Synthesis of Mono- and Disubstituted Piperidines and Pyrrolidines using Carboxylic Acid Reductase (CAR), ω -Transaminase (ω -TA), and Imine Reductase (IRED) Biocatalysts. *ACS Catal.* 6, 3753–3759. <https://doi.org/10.1021/acscatal.6b00855>

Frow, E., 2017. From “Experiments of Concern” to “Groups of Concern”: Constructing and Containing Citizens in Synthetic Biology. *Science, Technology, & Human Values* 0162243917735382. <https://doi.org/10.1177/0162243917735382>

Frow, E., 2015. Knowing New Biotechnologies: Social Aspects of Technological Convergence - Chapter 12: Rhetorics and Practices of Democratization in Synthetic Biology. Routledge.

Fürst, M.J.L.J., Martin, C., Lončar, N., Fraaije, M.W., 2018. Experimental Protocols for Generating Focused Mutant Libraries and Screening for Thermostable Proteins. *Meth. Enzymol.* 608, 151–187. <https://doi.org/10.1016/bs.mie.2018.04.007>

Gahlth, D., Dunstan, M.S., Quaglia, D., Klumbys, E., Lockhart-Cairns, M.P., Hill, A.M., Derrington, S.R., Scrutton, N.S., Turner, N.J., Leys, D., 2017. Structures of carboxylic acid reductase reveal domain dynamics underlying catalysis. *Nat. Chem. Biol.* <https://doi.org/10.1038/nchembio.2434>

Galtier, N., Tourasse, N., Gouy, M., 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283, 220–221.

Gao, S., Zhu, S., Huang, R., Li, H., Wang, H., Zheng, G., 2018. Engineering the Enantioselectivity and Thermostability of a (+)- γ -Lactamase from *Microbacterium hydrocarbonoxydans* for Kinetic

Resolution of Vince Lactam (2-Azabicyclo[2.2.1]hept-5-en-3-one). *Appl. Environ. Microbiol.* 84, e01780-17. <https://doi.org/10.1128/AEM.01780-17>

Garcia, A.K., Schopf, J.W., Yokobori, S., Akanuma, S., Yamagishi, A., 2017. Reconstructed ancestral enzymes suggest long-term cooling of Earth's photic zone since the Archean. *PNAS* 114, 4619–4624. <https://doi.org/10.1073/pnas.1702729114>

Gaucher, E.A., 2007. Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing and extending the vision of Zuckerkandl and Pauling. Oxford University Press.

Gaucher, E.A., Govindarajan, S., Ganesh, O.K., 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451, 704–707. <https://doi.org/10.1038/nature06510>

Gaucher, E.A., Thomson, J.M., Burgan, M.F., Benner, S.A., 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425, 285–288. <https://doi.org/10.1038/nature01977>

Gaylord, N.G., 1957. Reduction with complex metal hydrides. *J. Chem. Educ.* 34, 367. <https://doi.org/10.1021/ed034p367>

George Church, n.d. iGEM Giant Jamboree 2018, keynote speech.

Giver, L., Gershenson, A., Freskgard, P.-O., Arnold, F.H., 1998. Directed evolution of a thermostable esterase. *PNAS* 95, 12809–12813. <https://doi.org/10.1073/pnas.95.22.12809>

Goldenzweig, A., Fleishman, S.J., 2018. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* 87, 105–129. <https://doi.org/10.1146/annurev-biochem-062917-012102>

Goldenzweig, A., Goldsmith, M., Hill, S.E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R.L., Aharoni, A., Silman, I., Sussman, J.L., Tawfik, D.S., Fleishman, S.J., 2016. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell* 63, 337–346. <https://doi.org/10.1016/j.molcel.2016.06.012>

Golding, G.B., 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.* 1, 125–142. <https://doi.org/10.1093/oxfordjournals.molbev.a040303>

Goldsmith, M., Aggarwal, N., Ashani, Y., Jubran, H., Greisen, P.J., Ovchinnikov, S., Leader, H., Baker, D., Sussman, J.L., Goldenzweig, A., Fleishman, S.J., Tawfik, D.S., 2017. Overcoming an optimization plateau in the directed evolution of highly efficient nerve agent bioscavengers. *Protein Eng. Des. Sel.* 30, 333–345. <https://doi.org/10.1093/protein/gzx003>

- Goldsmith, M., Ashani, Y., Simo, Y., Ben-David, M., Leader, H., Silman, I., Sussman, J.L., Tawfik, D.S., 2012. Evolved Stereoselective Hydrolases for Broad-Spectrum G-Type Nerve Agent Detoxification. *Chemistry & Biology* 19, 456–466. <https://doi.org/10.1016/j.chembiol.2012.01.017>
- Goldstein, R.A., 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins: Structure, Function, and Bioinformatics* 79, 1396–1407. <https://doi.org/10.1002/prot.22964>
- Gonzalez, D., Hiblot, J., Darbinian, N., Miller, J.C., Gotthard, G., Amini, S., Chabriere, E., Elias, M., 2014. Ancestral mutations as a tool for solubilizing proteins: The case of a hydrophobic phosphate-binding protein. *FEBS Open Bio* 4, 121–127. <https://doi.org/10.1016/j.fob.2013.12.006>
- Gottardi, M., Grün, P., Bode, H.B., Hoffmann, T., Schwab, W., Oreb, M., Boles, E., 2017. Optimisation of trans-cinnamic acid and hydrocinnamyl alcohol production with recombinant *Saccharomyces cerevisiae* and identification of cinnamyl methyl ketone as a by-product. *FEMS Yeast Res.* <https://doi.org/10.1093/femsyr/fox091>
- Goundry, W.R.F., Adams, B., Benson, H., Demeritt, J., McKown, S., Mulholland, K., Robertson, A., Siedlecki, P., Tomlin, P., Vare, K., 2017. Development and Scale-up of a Biocatalytic Process To Form a Chiral Sulfoxide. *Org. Process Res. Dev.* <https://doi.org/10.1021/acs.oprd.6b00391>
- Gromiha, M.M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P., Sarai, A., 2002. ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0. *Nucleic Acids Res.* 30, 301–302.
- Groot, C.C.M., Bakker, H.J., 2016. Proteins Take up Water Before Unfolding. *J. Phys. Chem. Lett.* 7, 1800–1804. <https://doi.org/10.1021/acs.jpcllett.6b00708>
- Gu, X., Fu, Y.X., Li, W.H., 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol* 12, 546–557. <https://doi.org/10.1093/oxfordjournals.molbev.a040235>
- Guazzaroni, M.-E., Silva-Rocha, R., Ward, R.J., 2015. Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. *Microb Biotechnol* 8, 52–64. <https://doi.org/10.1111/1751-7915.12146>
- Guerriero, G., Hausman, J.-F., Strauss, J., Ertan, H., Siddiqui, K.S., 2016. Lignocellulosic biomass: Biosynthesis, degradation, and industrial utilization. *Engineering in Life Sciences* 16, 1–16. <https://doi.org/10.1002/elsc.201400196>

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010a. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. <https://doi.org/10.1093/sysbio/syq010>

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010b. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* 59, 307–321. <https://doi.org/10.1093/sysbio/syq010>

Guindon, S., Gascuel, O., 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* 52, 696–704. <https://doi.org/10.1080/10635150390235520>

Gullberg, M., Tolf, C., Jonsson, N., Mulders, M.N., Savolainen-Kopra, C., Hovi, T., Van Ranst, M., Lemey, P., Hafenstein, S., Lindberg, A.M., 2010. Characterization of a putative ancestor of coxsackievirus B5. *J. Virol.* 84, 9695–9708. <https://doi.org/10.1128/JVI.00071-10>

Gumulya, Y., Baek, J.-M., Wun, S.-J., Thomson, R.E.S., Harris, K.L., Hunter, D.J.B., Behrendorff, J.B.Y.H., Kulig, J., Zheng, S., Wu, X., Wu, B., Stok, J.E., Voss, J.J.D., Schenk, G., Jurva, U., Andersson, S., Isin, E.M., Bodén, M., Guddat, L., Gillam, E.M.J., 2018. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nature Catalysis* 1. <https://doi.org/10.1038/s41929-018-0159-5>

Gumulya, Y., Gillam, E.M.J., 2017. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the “retro” approach to protein engineering. *Biochem. J.* 474, 1–19. <https://doi.org/10.1042/BCJ20160507>

Gurung, N., Ray, S., Bose, S., Rai, V., 2013. A Broader View: Microbial Enzymes and Their Relevance in Industries, Medicine, and Beyond. *Biomed Res Int* 2013. <https://doi.org/10.1155/2013/329121>

Hansen, J., Hansen, E.H., SOMPALLI, H.P., Sheridan, J.M., Heal, J.R., Hamilton, W.D.O., 2013. Compositions and methods for the biosynthesis of vanillin or vanillin beta-d-glucoside. WO2013022881 A8.

Hanson-Smith, V., Johnson, A., 2016. PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories. *PLOS Computational Biology* 12, e1004976. <https://doi.org/10.1371/journal.pcbi.1004976>

Hanson-Smith, V., Kolaczkowski, B., Thornton, J.W., 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Mol Biol Evol* 27, 1988–1999. <https://doi.org/10.1093/molbev/msq081>

Hao, J., Berry, A., 2004. A thermostable variant of fructose biphosphate aldolase constructed by directed evolution also shows increased stability in organic solvents. *Protein Engineering, Design and Selection* 17, 689–697. <https://doi.org/10.1093/protein/gzh081>

Harland, W.B., 1964. Critical evidence for a great infra-Cambrian glaciation. *Geol Rundsch* 54, 45–61. <https://doi.org/10.1007/BF01821169>

Harms, M.J., Thornton, J.W., 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics* 14, 559–571. <https://doi.org/10.1038/nrg3540>

Harms, M.J., Thornton, J.W., 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* 20, 360–366. <https://doi.org/10.1016/j.sbi.2010.03.005>

Hart, K.M., Harms, M.J., Schmidt, B.H., Elya, C., Thornton, J.W., Marqusee, S., 2014. Thermodynamic System Drift in Protein Evolution. *PLOS Biology* 12, e1001994. <https://doi.org/10.1371/journal.pbio.1001994>

Hartl, D.L., 2014. What can we learn from fitness landscapes? *Current Opinion in Microbiology, Antimicrobials* 21, 51–57. <https://doi.org/10.1016/j.mib.2014.08.001>

Hecky, J., Müller, K.M., 2005. Structural perturbation and compensation by directed evolution at physiological temperature leads to thermostabilization of beta-lactamase. *Biochemistry* 44, 12640–12654. <https://doi.org/10.1021/bi0501885>

Hedstrom, L., 2002. Serine Protease Mechanism and Specificity. *Chem. Rev.* 102, 4501–4524. <https://doi.org/10.1021/cr000033x>

Helm, E. van der, Genee, H.J., Sommer, M.O.A., 2018. The evolving interface between synthetic biology and functional metagenomics. *Nature Chemical Biology* 14, 752. <https://doi.org/10.1038/s41589-018-0100-x>

Hendriks, A.T.W.M., Zeeman, G., 2009. Pretreatments to enhance the digestibility of lignocellulosic biomass. *Bioresource Technology* 100, 10–18. <https://doi.org/10.1016/j.biortech.2008.05.027>

Hewitt, C.J., Nienow, A.W., 2007. The Scale-Up of Microbial Batch and Fed-Batch Fermentation Processes, in: *Advances in Applied Microbiology*. Academic Press, pp. 105–135. [https://doi.org/10.1016/S0065-2164\(07\)62005-X](https://doi.org/10.1016/S0065-2164(07)62005-X)

Hickey, A.M., Ngamsom, B., Wiles, C., Greenway, G.M., Watts, P., Littlechild, J.A., 2009. A microreactor for the study of biotransformations by a cross-linked gamma-lactamase enzyme. *Biotechnol J* 4, 510–516. <https://doi.org/10.1002/biot.200800302>

- Hilser, V.J., Gómez, J., Freire, E., 1996. The enthalpy change in protein folding and binding: refinement of parameters for structure-based calculations. *Proteins* 26, 123–133. [https://doi.org/10.1002/\(SICI\)1097-0134\(199610\)26:2<123::AID-PROT2>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1097-0134(199610)26:2<123::AID-PROT2>3.0.CO;2-H)
- Hirata, A., Sato, A., Tadokoro, T., Koga, Y., Kanaya, S., Takano, K., 2012. A Stable Protein - CutA1. *Protein Structure*. <https://doi.org/10.5772/37042>
- Ho, Y.-K., Karavaggelis, M., Datta, P., Hallinan, J., 2017. Synthetic Biology Start-ups in the UK and Worldwide 29.
- Hobbs, J.K., Shepherd, C., Saul, D.J., Demetras, N.J., Haaning, S., Monk, C.R., Daniel, R.M., Arcus, V.L., 2012. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of *Bacillus*. *Mol Biol Evol* 29, 825–835. <https://doi.org/10.1093/molbev/msr253>
- Hochberg, G.K.A., Thornton, J.W., 2017. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys*. <https://doi.org/10.1146/annurev-biophys-070816-033631>
- Hodgman, C.E., Jewett, M.C., 2012. Cell-free synthetic biology: Thinking outside the cell. *Metabolic Engineering, Synthetic Biology: New Methodologies and Applications for Metabolic Engineering* 14, 261–269. <https://doi.org/10.1016/j.ymben.2011.09.002>
- Hollinshead, W., He, L., Tang, Y.J., 2014. Biofuel production: an odyssey from metabolic engineering to fermentation scale-up. *Front Microbiol* 5. <https://doi.org/10.3389/fmicb.2014.00344>
- Holmes, I.H., 2017. Solving the master equation for Indels. *BMC Bioinformatics* 18, 255. <https://doi.org/10.1186/s12859-017-1665-1>
- Huang, P.-S., Boyken, S.E., Baker, D., 2016. The coming of age of de novo protein design. *Nature* 537, 320–327. <https://doi.org/10.1038/nature19946>
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. <https://doi.org/10.1093/bioinformatics/17.8.754>
- Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294, 2310–2314. <https://doi.org/10.1126/science.1065889>
- Hughes, R.A., Ellington, A.D., 2017. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harb Perspect Biol* 9, a023812. <https://doi.org/10.1101/cshperspect.a023812>

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., Schuster, S.C., 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. <https://doi.org/10.1101/gr.120618.111>

Illumina, 2017. Illumina Introduces the NovaSeq Series—a New Architecture Designed to Usher in the \$100 Genome [WWW Document]. URL <https://www.businesswire.com/news/home/20170109006363/en/Illumina-Introduces-NovaSeq-Series%E2%80%94a-New-Architecture-Designed> (accessed 11.18.18).

Ilmberger, N., Meske, D., Juergensen, J., Schulte, M., Barthen, P., Rabausch, U., Angelov, A., Mientus, M., Liebl, W., Schmitz, R.A., Streit, W.R., 2012. Metagenomic cellulases highly tolerant towards the presence of ionic liquids—linking thermostability and halotolerance. *Appl Microbiol Biotechnol* 95, 135–146. <https://doi.org/10.1007/s00253-011-3732-2>

Jacob, F., Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* 3, 318–356. [https://doi.org/10.1016/S0022-2836\(61\)80072-7](https://doi.org/10.1016/S0022-2836(61)80072-7)

Jaenicke, R., 1992. Protein stability and molecular adaptation to extreme conditions, in: Christen, P., Hofmann, E. (Eds.), *EJB Reviews 1991*, EJB Reviews 1991. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 291–304. https://doi.org/10.1007/978-3-642-77200-9_22

Jefferson, C., Lentzos, F., Marris, C., 2014. Synthetic Biology and Biosecurity: Challenging the “Myths.” *Front. Public Health* 2. <https://doi.org/10.3389/fpubh.2014.00115>

Jemli, S., Ayadi-Zouari, D., Hlima, H.B., Bejar, S., 2016. Biocatalysts: application and engineering for industrial purposes. *Critical Reviews in Biotechnology* 36, 246–258. <https://doi.org/10.3109/07388551.2014.950550>

Jermann, T.M., Opitz, J.G., Stackhouse, J., Benner, S.A., 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374, 57–59. <https://doi.org/10.1038/374057a0>

Jim Lane, 2016. “Call me TerraVia”: Solazyme transforms, shakes up and refines : *Biofuels Digest*.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8, 275–282. <https://doi.org/10.1093/bioinformatics/8.3.275>

Joseph Felsenstein, 1995. PHYLIP (Phylogeny Inference Package) Version 3.57c [WWW Document]. URL <http://www.dbbm.fiocruz.br/molbiol/main.html> (accessed 12.5.18).

Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T., Poon, A.F.Y., 2016a. Ancestral Reconstruction. *PLOS Computational Biology* 12, e1004763. <https://doi.org/10.1371/journal.pcbi.1004763>

Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T., Poon, A.F.Y., 2016b. Ancestral Reconstruction. *PLOS Computational Biology* 12, e1004763. <https://doi.org/10.1371/journal.pcbi.1004763>

Jullesson, D., David, F., Pflieger, B., Nielsen, J., 2015. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnology Advances, Industrial Biotechnology: Tools and Applications* 33, 1395–1402. <https://doi.org/10.1016/j.biotechadv.2015.02.011>

Jungwirth, P., Cremer, P.S., 2014. Beyond Hofmeister. *Nature Chemistry* 6, 261–263. <https://doi.org/10.1038/nchem.1899>

K. Survana, A. Lolas, P. Hughes, R.E. Feiedman, 2011. Case studies of microbial contamination in biologic product manufacturing [WWW Document]. *American Pharmaceutical Review*. URL <https://www.americanpharmaceuticalreview.com/Featured-Articles/36755-Case-Studies-of-Microbial-Contamination-in-Biologic-Product-Manufacturing/> (accessed 11.24.18).

Kacar, B., Guy, L., Smith, E., Baross, J., 2017. Resurrecting ancestral genes in bacteria to interpret ancient biosignatures. *Philos Trans A Math Phys Eng Sci* 375. <https://doi.org/10.1098/rsta.2016.0352>

Kallio, P., Pásztor, A., Thiel, K., Akhtar, M.K., Jones, P.R., 2014. An engineered pathway for the biosynthesis of renewable propane. *Nature Communications* 5, 4731. <https://doi.org/10.1038/ncomms5731>

Kallioinen, A., Puranen, T., Siika-aho, M., 2014. Mixtures of thermostable enzymes show high performance in biomass saccharification. *Appl. Biochem. Biotechnol.* 173, 1038–1056. <https://doi.org/10.1007/s12010-014-0893-3>

Kaper, T., van der Maarel, M.J.E.C., Euverink, G.J.W., Dijkhuizen, L., 2004. Exploring and exploiting starch-modifying amyloamylases from thermophiles. *Biochem. Soc. Trans.* 32, 279–282. <https://doi.org/10.1042/>

Karshikoff, A., Nilsson Lennart, Ladenstein Rudolf, 2015. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *The FEBS Journal* 282, 3899–3917. <https://doi.org/10.1111/febs.13343>

Katie Fehrenbacher, 2016. Solazyme Ditches Biofuels (& Name) in a World of Cheap Oil [WWW Document]. *Fortune*. URL <http://fortune.com/2016/03/16/solazyme-terravia-ditches-biofuels/> (accessed 11.22.18).

Katie Fehrenbacher, 2015. How A Tech Billionaire’s Biofuel Dream Went Bad. *Fortune*.

- Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30, 3059–3066. <https://doi.org/10.1093/nar/gkf436>
- Katz, L., Chen, Y.Y., Gonzalez, R., Peterson, T.C., Zhao, H., Baltz, R.H., 2018. Synthetic biology advances and applications in the biotechnology industry: a perspective. *J Ind Microbiol Biotechnol* 45, 449–461. <https://doi.org/10.1007/s10295-018-2056-y>
- Kaufmann, K.W., Lemmon, G.H., DeLuca, S.L., Sheehan, J.H., Meiler, J., 2010. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry* 49, 2987–2998. <https://doi.org/10.1021/bi902153g>
- Kaushik, M., Sinha, P., Jaiswal, P., Mahendru, S., Roy, K., Kukreti, S., 2016. Protein engineering and de novo designing of a biocatalyst. *J. Mol. Recognit.* 29, 499–503. <https://doi.org/10.1002/jmr.2546>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Keasling, J.D., 2012. Synthetic biology and the development of tools for metabolic engineering. *Metabolic Engineering, Synthetic Biology: New Methodologies and Applications for Metabolic Engineering* 14, 189–195. <https://doi.org/10.1016/j.ymben.2012.01.004>
- Keefe, A.D., Szostak, J.W., 2001. Functional proteins from a random-sequence library. *Nature* 410, 715–718. <https://doi.org/10.1038/35070613>
- Kelley, N.J., Whelan, D.J., Kerr, E., Apel, A., Beliveau, R., Scanlon, R., 2014. Engineering Biology to Address Global Problems: Synthetic Biology Markets, Needs, and Applications. *Industrial Biotechnology* 10, 140–149. <https://doi.org/10.1089/ind.2014.1515>
- Kempton, C.L., White, G.C., 2009. How we treat a hemophilia A patient with a factor VIII inhibitor. *Blood* 113, 11–17. <https://doi.org/10.1182/blood-2008-06-160432>
- Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D.L., Prilusky, J., Amengual-Rigo, P., Guallar, V., Tawfik, D.S., Fleishman, S.J., 2018. Automated design of efficient and functionally diverse enzyme repertoires. *Mol Cell* 72, 178-186.e5. <https://doi.org/10.1016/j.molcel.2018.08.033>
- Khusnutdinova, A.N., Flick, R., Popovic, A., Brown, G., Tchigvintsev, A., Nocek, B., Correia, K., Joo, J.C., Mahadevan, R., Yakunin, A.F., 2017. Exploring Bacterial Carboxylate Reductases for the

Reduction of Bifunctional Carboxylic Acids. *Biotechnol. J.* 12, n/a-n/a.
<https://doi.org/10.1002/biot.201600751>

Kimura, M., 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature* 217, 624–626.
<https://doi.org/10.1038/217624a0>

King, J.L., Jukes, T.H., 1969. Non-Darwinian Evolution. *Science* 164, 788–798.
<https://doi.org/10.1126/science.164.3881.788>

Kingsley, L.J., Lill, M.A., 2015. Substrate Tunnels in Enzymes: Structure-Function Relationships and Computational Methodology. *Proteins* 83, 599–611. <https://doi.org/10.1002/prot.24772>

Kiss, C., Temirov, J., Chasteen, L., Waldo, G.S., Bradbury, A.R.M., 2009. Directed evolution of an extremely stable fluorescent protein. *Protein Eng. Des. Sel.* 22, 313–323.
<https://doi.org/10.1093/protein/gzp006>

Kitney, R., Freemont, P., 2012. Synthetic biology – the state of play. *FEBS Letters* 586, 2029–2036.
<https://doi.org/10.1016/j.febslet.2012.06.002>

Klein-Marcuschamer, D., Oleskowicz-Popiel, P., Simmons, B.A., Blanch, H.W., 2012. The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnology and Bioengineering* 109, 1083–1087. <https://doi.org/10.1002/bit.24370>

Koizumi, M., Hirai, H., Onai, T., Inoue, K., Hirai, M., 2007. Collapse of the hydration shell of a protein prior to thermal unfolding. *J Appl Cryst* 40, s175–s178. <https://doi.org/10.1107/S0021889807003354>

Kondrashov, D.A., Kondrashov, F.A., 2015. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* 31, 24–33. <https://doi.org/10.1016/j.tig.2014.09.009>

Koshi, J.M., Goldstein, R.A., 1996. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* 42, 313–320.

Koshland, D.E., 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *PNAS* 44, 98–104. <https://doi.org/10.1073/pnas.44.2.98>

Koudelakova, T., Bidmanova, S., Dvorak, P., Pavelka, A., Chaloupkova, R., Prokop, Z., Damborsky, J., 2013. Haloalkane dehalogenases: Biotechnological applications. *Biotechnology Journal* 8, 32–45.
<https://doi.org/10.1002/biot.201100486>

Kramer, L., Hankore, E., Liu, Y., Liu, K., Jimenez, E., Guo, J., Niu, W., 2018. Characterization of Carboxylic Acid Reductases for Biocatalytic Synthesis of Industrial Chemicals. *Chembiochem*. <https://doi.org/10.1002/cbic.201800157>

Kramer, M., Halleran, D., Rahman, M., Iqbal, M., Anwar, M.I., Sabet, S., Ackad, E., Yousef, M., 2014. Comparative Molecular Dynamics Simulation of Hepatitis C Virus NS3/4A Protease (Genotypes 1b, 3a and 4a) Predicts Conformational Instability of the Catalytic Triad in Drug Resistant Strains. *PLOS ONE* 9, e104425. <https://doi.org/10.1371/journal.pone.0104425>

Krieger, E., Vriend, G., 2014. YASARA View - molecular graphics for all devices - from smartphones to workstations. *Bioinformatics* 30, 2981–2982. <https://doi.org/10.1093/bioinformatics/btu426>

Kuriyan, J., Konforti, B., Wemmer, D., 2012. *The Molecules of Life: Physical and Chemical Principles*. Garland Science.

Kwok, R., 2010. Five hard truths for synthetic biology. *Nature* 463, 288–290. <https://doi.org/10.1038/463288a>

Langerak, F., Hultink, E.J., 2006. The Impact of Product Innovativeness on the Link between Development Speed and New Product Profitability*. *Journal of Product Innovation Management* 23, 203–214. <https://doi.org/10.1111/j.1540-5885.2006.00194.x>

Larget, B., Simon, D.L., 1999. Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Mol Biol Evol* 16, 750–750. <https://doi.org/10.1093/oxfordjournals.molbev.a026160>

Le Feuvre, R.A., Scrutton, N.S., 2018. A living foundry for Synthetic Biological Materials: A synthetic biology roadmap to new advanced materials. *Synthetic and Systems Biotechnology* 3, 105–112. <https://doi.org/10.1016/j.synbio.2018.04.002>

Le, Q., Sievers, F., Higgins, D.G., 2017. Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics* 33, 1331–1337. <https://doi.org/10.1093/bioinformatics/btw840>

Le, S.Q., Gascuel, O., 2010. Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial. *Syst Biol* 59, 277–287. <https://doi.org/10.1093/sysbio/syq002>

Le, S.Q., Gascuel, O., 2008. An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25, 1307–1320. <https://doi.org/10.1093/molbev/msn067>

Lechner, A., Brunk, E., Keasling, J.D., 2016. The Need for Integrated Approaches in Metabolic Engineering. *Cold Spring Harb Perspect Biol* 8. <https://doi.org/10.1101/cshperspect.a023903>

Lee, J.W., Na, D., Park, J.M., Lee, J., Choi, S., Lee, S.Y., 2012. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature Chemical Biology* 8, 536–546. <https://doi.org/10.1038/nchembio.970>

Lee, S.K., Chou, H., Ham, T.S., Lee, T.S., Keasling, J.D., 2008. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology, Chemical biotechnology / Pharmaceutical biotechnology* 19, 556–563. <https://doi.org/10.1016/j.copbio.2008.10.014>

Lehmann, M., Pasamontes, L., Lassen, S.F., Wyss, M., 2000. The consensus concept for thermostability engineering of proteins. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology, Protein engineering of enzymes* 1543, 408–415. [https://doi.org/10.1016/S0167-4838\(00\)00238-7](https://doi.org/10.1016/S0167-4838(00)00238-7)

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. *Syst Biol* 58, 130–145. <https://doi.org/10.1093/sysbio/syp017>

Lewin, G.R., Carlos, C., Chevrette, M.G., Horn, H.A., McDonald, B.R., Stankey, R.J., Fox, B.G., Currie, C.R., 2016. Evolution and Ecology of Actinobacteria and Their Bioenergy Applications. *Annu Rev Microbiol* 70, 235–254. <https://doi.org/10.1146/annurev-micro-102215-095748>

Li, A., Acevedo-Rocha, C.G., Sun, Z., Cox, T., Xu, J.L., Reetz, M.T., 2018a. Beating Bias in the Directed Evolution of Proteins: Combining High-Fidelity on-Chip Solid-Phase Gene Synthesis with Efficient Gene Assembly for Combinatorial Library Construction. *Chembiochem* 19, 221–228. <https://doi.org/10.1002/cbic.201700540>

Li, A., Sun, Z., Reetz, M.T., 2018b. Solid-Phase Gene Synthesis for Mutant Library Construction: The Future of Directed Evolution? *ChemBioChem* 19, 2023–2032. <https://doi.org/10.1002/cbic.201800339>

Li, Y., Drummond, D.A., Sawayama, A.M., Snow, C.D., Bloom, J.D., Arnold, F.H., 2007. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* 25, 1051–1056. <https://doi.org/10.1038/nbt1333>

Lim, S.A., Marqusee, S., 2017. The burst-phase folding intermediate of ribonuclease H changes conformation over evolutionary history. *Biopolymers*. <https://doi.org/10.1002/bip.23086>

Lin, L., Xu, J., 2013. Dissecting and engineering metabolic and regulatory networks of thermophilic bacteria for biofuel production. *Biotechnology Advances, “Bioenergy and Biorefinery from Biomass” through innovative technology development* 31, 827–837. <https://doi.org/10.1016/j.biotechadv.2013.03.003>

Lin, P.P., Rabe, K.S., Takasumi, J.L., Kadisch, M., Arnold, F.H., Liao, J.C., 2014. Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*. *Metab. Eng.* 24, 1–8. <https://doi.org/10.1016/j.ymben.2014.03.006>

Lips, D., M. Schuurmans, J., Santos, F.B. dos, J. Hellingwerf, K., 2018. Many ways towards ‘solar fuel’: quantitative analysis of the most promising strategies and the main challenges during scale-up. *Energy & Environmental Science* 11, 10–22. <https://doi.org/10.1039/C7EE02212C>

Littlechild, J.A., 2017. Improving the ‘tool box’ for robust industrial enzymes. *J Ind Microbiol Biotechnol* 44, 711–720. <https://doi.org/10.1007/s10295-017-1920-5>

Littlechild, J.A., 2015. Enzymes from Extreme Environments and Their Industrial Applications. *Front Bioeng Biotechnol* 3. <https://doi.org/10.3389/fbioe.2015.00161>

Liu, Z., Si, B., Li, J., He, J., Zhang, C., Lu, Y., Zhang, Y., Xing, X.-H., 2018. Bioprocess engineering for biohythane production from low-grade waste biomass: technical challenges towards scale up. *Current Opinion in Biotechnology, Energy biotechnology • Environmental biotechnology* 50, 25–31. <https://doi.org/10.1016/j.copbio.2017.08.014>

Lo Surdo, P., Walsh, M.A., Sollazzo, M., 2004. A novel ADP- and zinc-binding fold from function-directed in vitro evolution. *Nat. Struct. Mol. Biol.* 11, 382–383. <https://doi.org/10.1038/nsmb745>

Long, L., Tian, D., Zhai, R., Li, X., Zhang, Y., Hu, J., Wang, F., Saddler, J., 2018. Thermostable xylanase-aided two-stage hydrolysis approach enhances sugar release of pretreated lignocellulosic biomass. *Bioresour. Technol.* 257, 334–338. <https://doi.org/10.1016/j.biortech.2018.02.104>

Lu, Y., 2017. Cell-free synthetic biology: Engineering in an open world. *Synthetic and Systems Biotechnology, A tribute to Arny Demain, for his lifelong pioneering contributions to biochemical engineering* 2, 23–27. <https://doi.org/10.1016/j.synbio.2017.02.003>

Luo, Y., Lee, J.-K., Zhao, H., 2013. Challenges and opportunities in synthetic biology for chemical engineers. *Chem Eng Sci* 103. <https://doi.org/10.1016/j.ces.2012.06.013>

Lutz, S., 2010. Beyond directed evolution—semi-rational protein engineering and design. *Current Opinion in Biotechnology, Chemical biotechnology – Pharmaceutical biotechnology* 21, 734–743. <https://doi.org/10.1016/j.copbio.2010.08.011>

Mackness, M., Mackness, B., 2015. Human paraoxonase-1 (PON1): Gene structure and expression, promiscuous activities and multiple physiological roles. *Gene* 567, 12–21. <https://doi.org/10.1016/j.gene.2015.04.088>

- Madhusudan Makwana, K., Mahalakshmi, R., 2015. Implications of aromatic–aromatic interactions: From protein structures to peptide models. *Protein Sci* 24, 1920–1933. <https://doi.org/10.1002/pro.2814>
- Magliery, T.J., 2015a. Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol* 33, 161–168. <https://doi.org/10.1016/j.sbi.2015.09.002>
- Magliery, T.J., 2015b. Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol* 33, 161–168. <https://doi.org/10.1016/j.sbi.2015.09.002>
- Marshall, D.C., 2010. Cryptic Failure of Partitioned Bayesian Phylogenetic Analyses: Lost in the Land of Long Trees. *Syst Biol* 59, 108–117. <https://doi.org/10.1093/sysbio/syp080>
- Martin, C., Ovalle Maqueo, A., Wijma, H.J., Fraaije, M.W., 2018. Creating a more robust 5-hydroxymethylfurfural oxidase by combining computational predictions with a novel effective library design. *Biotechnol Biofuels* 11. <https://doi.org/10.1186/s13068-018-1051-x>
- Martin, C.H., Nielsen, D.R., Solomon, K.V., Prather, K.L.J., 2009. Synthetic Metabolism: Engineering Biology at the Protein and Pathway Scales. *Chemistry & Biology* 16, 277–286. <https://doi.org/10.1016/j.chembiol.2009.01.010>
- Martin LaMonica, 2014. Why the Promise of Cheap Fuel from Super Bugs Fell Short - MIT Technology Review [WWW Document]. MIT Technology Review. URL <https://www.technologyreview.com/s/524011/why-the-promise-of-cheap-fuel-from-super-bugs-fell-short/> (accessed 11.22.18).
- Martínez, R., Schwaneberg, U., 2013. A roadmap to directed enzyme evolution and screening systems for biotechnological applications. *Biological Research* 46, 395–405. <https://doi.org/10.4067/S0716-97602013000400011>
- Matsuura, Y., Takehira, M., Joti, Y., Ogasahara, K., Tanaka, T., Ono, N., Kunishima, N., Yutani, K., 2015. Thermodynamics of protein denaturation at temperatures over 100 °C: CutA1 mutant proteins substituted with hydrophobic and charged residues. *Scientific Reports* 5, 15545. <https://doi.org/10.1038/srep15545>
- Mattos, C., 2002. Protein–water interactions in a dynamic world. *Trends in Biochemical Sciences* 27, 203–208. [https://doi.org/10.1016/S0968-0004\(02\)02067-4](https://doi.org/10.1016/S0968-0004(02)02067-4)
- McDonald, A.G., Tipton, K.F., 2014. Fifty-five years of enzyme classification: advances and difficulties. *FEBS J.* 281, 583–592. <https://doi.org/10.1111/febs.12530>

Melody M. Bomgardner, 2017. Amyris will sell farnesene plant to DSM [WWW Document]. Chemical & Engineering News. URL <https://cen.acs.org/articles/95/i47/Amyris-sell-farnesene-plant-DSM.html> (accessed 11.22.18).

Merkl, R., Sterner, R., 2016. Reconstruction of ancestral enzymes. Perspectives in Science, Proceedings of the Beilstein ESCEC Symposium 2015 9, 17–23. <https://doi.org/10.1016/j.pisc.2016.08.002>

Merkl, R., Sterner, R., 2015. Ancestral protein reconstruction: techniques and applications. Biological Chemistry 397, 1–21. <https://doi.org/10.1515/hsz-2015-0158>

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. 21, 1087–1092. <https://doi.org/10.1063/1.1699114>

Miller, S.L., Lazcano, A., 1995. The origin of life--did it occur at high temperatures? J. Mol. Evol. 41, 689–692.

Miyazaki, J., Nakaya, S., Suzuki, T., Tamakoshi, M., Oshima, T., Yamagishi, A., 2001. Ancestral Residues Stabilizing 3-Isopropylmalate Dehydrogenase of an Extreme Thermophile: Experimental Evidence Supporting the Thermophilic Common Ancestor Hypothesis. J Biochem 129, 777–782. <https://doi.org/10.1093/oxfordjournals.jbchem.a002919>

Moore, G.A., 2014. Crossing the Chasm, 3rd Edition: Marketing and Selling Disruptive Products to Mainstream Customers. HarperCollins.

Moratorio, G., Vignuzzi, M., 2018. Monitoring and redirecting virus evolution. PLOS Pathogens 14, e1006979. <https://doi.org/10.1371/journal.ppat.1006979>

Moser, F., Broers, N.J., Hartmans, S., Tamsir, A., Kerkman, R., Roubos, J.A., Bovenberg, R., Voigt, C.A., 2012. Genetic Circuit Performance under Conditions Relevant for Industrial Bioreactors. ACS Synth Biol 1, 555–564. <https://doi.org/10.1021/sb3000832>

Moura, M., Pertusi, D., Lenzini, S., Bhan, N., Broadbelt, L.J., Tyo, K.E.J., 2016. Characterizing and predicting carboxylic acid reductase activity for diversifying bioaldehyde production. Biotechnol. Bioeng. 113, 944–952. <https://doi.org/10.1002/bit.25860>

Na, D., Kim, T.Y., Lee, S.Y., 2010. Construction and optimization of synthetic pathways in metabolic engineering. Current Opinion in Microbiology, Ecology and industrial microbiology • Special section: Systems biology 13, 363–370. <https://doi.org/10.1016/j.mib.2010.02.004>

- Napora-Wijata, K., Strohmeier, G.A., Winkler, M., 2014. Biocatalytic reduction of carboxylic acids. *Biotechnology Journal* 9, 822–843. <https://doi.org/10.1002/biot.201400012>
- Narancic, T., O'Connor, K.E., 2017. Microbial biotechnology addressing the plastic waste disaster. *Microb Biotechnol* 10, 1232–1235. <https://doi.org/10.1111/1751-7915.12775>
- Nardo, A.E., Añón, M.C., Parisi, G., 2018. Large-scale mapping of bioactive peptides in structural and sequence space. *PLoS One* 13. <https://doi.org/10.1371/journal.pone.0191063>
- Nascimento, F.F., Reis, M. dos, Yang, Z., 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution* 1, 1446. <https://doi.org/10.1038/s41559-017-0280-x>
- Nelson David R., Goldstone Jared V., Stegeman John J., 2013. The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368, 20120474. <https://doi.org/10.1098/rstb.2012.0474>
- Nestl, B.M., Hauer, B., 2014. Engineering of Flexible Loops in Enzymes. *ACS Catal.* 4, 3201–3211. <https://doi.org/10.1021/cs500325p>
- Nevozhay, D., Adams, R.M., Itallie, E.V., Bennett, M.R., Balázsi, G., 2012. Mapping the Environmental Fitness Landscape of a Synthetic Gene Circuit. *PLOS Computational Biology* 8, e1002480. <https://doi.org/10.1371/journal.pcbi.1002480>
- Neylon, C., 2004. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res* 32, 1448–1459. <https://doi.org/10.1093/nar/gkh315>
- Nguyen, V., Wilson, C., Hoemberger, M., Stiller, J.B., Agafonov, R.V., Kutter, S., English, J., Theobald, D.L., Kern, D., 2017. Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* 355, 289–294. <https://doi.org/10.1126/science.aah3717>
- Nielsen, D.R., Moon, T.S., 2013. From promise to practice. The role of synthetic biology in green chemistry. *EMBO Rep.* 14, 1034–1038. <https://doi.org/10.1038/embor.2013.178>
- Nisbet, E.G., Sleep, N.H., 2001. The habitat and nature of early life. *Nature* 409, 1083–1091. <https://doi.org/10.1038/35059210>
- Noordam, B., Berkhout, M.P.J., HOFMEESTER, J.J.M., 2018. Process and apparatus for enzymatic hydrolysis of lignocellulosic material and fermentation of sugars. US10087475B2.
- Notredame, C., 2002. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* 3, 131–144. <https://doi.org/10.1517/14622416.3.1.131>

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment¹¹Edited by J. Thornton. *Journal of Molecular Biology* 302, 205–217. <https://doi.org/10.1006/jmbi.2000.4042>

Nyffenegger, C., Carvalho, Â., Hansen, K.R., Hansen, E.H., Hansen, J., Amick, J., Salerno, G., 2017. Metabolic engineering of *Saccharomyces cerevisiae* to harness nature's valuable compounds. *Enzyme Engineering XXIV*.

Okafor, C.D., Pathak, M.C., Fagan, C.E., Bauer, N.C., Cole, M.F., Gaucher, E.A., Ortlund, E.A., 2018. Structural and Dynamics Comparison of Thermostability in Ancient, Modern, and Consensus Elongation Factor Tus. *Structure* 26, 118-129.e3. <https://doi.org/10.1016/j.str.2017.11.018>

O'Reilly, E., Turner, N.J., 2015. Enzymatic cascades for the regio- and stereoselective synthesis of chiral amines. *Perspectives in Science, Proceedings of the Beilstein ESCEC Symposium - Celebrating the 100th Anniversary of Michaelis-Menten-Kinetics* 4, 55–61. <https://doi.org/10.1016/j.pisc.2014.12.009>

Paatero, A., Rosti, K., Shkumatov, A.V., Sele, C., Brunello, C., Kysenius, K., Singha, P., Jokinen, V., Huttunen, H., Kajander, T., 2016. Crystal Structure of an Engineered LRRTM2 Synaptic Adhesion Molecule and a Model for Neurexin Binding. *Biochemistry* 55, 914–926. <https://doi.org/10.1021/acs.biochem.5b00971>

Pabis, A., Risso, V.A., Sanchez-Ruiz, J.M., Kamerlin, S.C., 2018. Cooperativity and flexibility in enzyme evolution. *Current Opinion in Structural Biology, Folding and binding in silico, in vitro and in cellula • Proteins: An Evolutionary Perspective* 48, 83–92. <https://doi.org/10.1016/j.sbi.2017.10.020>

Pace, C.N., 1990. Conformational stability of globular proteins. *Trends in Biochemical Sciences* 15, 14–17. [https://doi.org/10.1016/0968-0004\(90\)90124-T](https://doi.org/10.1016/0968-0004(90)90124-T)

Pace, C.N., Shirley, B.A., McNutt, M., Gajiwala, K., 1996. Forces contributing to the conformational stability of proteins. *The FASEB Journal* 10, 75–83. <https://doi.org/10.1096/fasebj.10.1.8566551>

Packer, M.S., Liu, D.R., 2015. Methods for the directed evolution of proteins. *Nature Reviews Genetics* 16, 379. <https://doi.org/10.1038/nrg3927>

Paddon, C.J., Keasling, J.D., 2014. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat Rev Micro* 12, 355–367. <https://doi.org/10.1038/nrmicro3240>

Pál, C., Papp, B., Lercher, M.J., 2006. An integrated view of protein evolution. *Nature Reviews Genetics* 7, 337–348. <https://doi.org/10.1038/nrg1838>

- Pal, S.K., Peon, J., Zewail, A.H., 2002. Biological water at the protein surface: Dynamical solvation probed directly with femtosecond resolution. *Proc Natl Acad Sci U S A* 99, 1763–1768. <https://doi.org/10.1073/pnas.042697899>
- Papaleo, E., Saladino, G., Lambrugh, M., Lindorff-Larsen, K., Gervasio, F.L., Nussinov, R., 2016. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chem. Rev.* 116, 6391–6423. <https://doi.org/10.1021/acs.chemrev.5b00623>
- Pardo, I., Rodríguez-Escribano, D., Aza, P., de Salas, F., Martínez, A.T., Camarero, S., 2018. A highly stable laccase obtained by swapping the second cupredoxin domain. *Sci Rep* 8. <https://doi.org/10.1038/s41598-018-34008-3>
- Park, J.H., Lee, S.Y., Kim, T.Y., Kim, H.U., 2008. Application of systems biology for bioprocess development. *Trends in Biotechnology* 26, 404–412. <https://doi.org/10.1016/j.tibtech.2008.05.001>
- Parthasarathy, S., Murthy, M.R.N., 2000. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng Des Sel* 13, 9–13. <https://doi.org/10.1093/protein/13.1.9>
- Pauling, L., Zuckerkandl, E., 1963. Chemical Paleogenetics. Molecular “Restoration Studies” of Extinct Forms of Life. *Acta Chemica Scandinavica* 17 suppl., 9–16. <https://doi.org/10.3891/acta.chem.scand.17s-0009>
- Pawlowski, A.C., Stogios, P.J., Koteva, K., Skarina, T., Evdokimova, E., Savchenko, A., Wright, G.D., 2018. The evolution of substrate discrimination in macrolide antibiotic resistance enzymes. *Nat Commun* 9, 112. <https://doi.org/10.1038/s41467-017-02680-0>
- Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T.C., Waldo, G.S., 2006. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotech* 24, 79–88. <https://doi.org/10.1038/nbt1172>
- Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z.-M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T.J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J.M., Gaucher, E.A., Fernandez, J.M., 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nature Structural & Molecular Biology* 18, 592–596. <https://doi.org/10.1038/nsmb.2020>
- Peterson, M.E., Eisenthal, R., Danson, M.J., Spence, A., Daniel, R.M., 2004. A New Intrinsic Thermal Parameter for Enzymes Reveals True Temperature Optima. *J. Biol. Chem.* 279, 20717–20722. <https://doi.org/10.1074/jbc.M309143200>
- Pham, V.H.T., Kim, J., 2012. Cultivation of unculturable soil bacteria. *Trends Biotechnol.* 30, 475–484. <https://doi.org/10.1016/j.tibtech.2012.05.007>

Phillips, A., Janies, D., Wheeler, W., 2000. Multiple Sequence Alignment in Phylogenetic Analysis. *Molecular Phylogenetics and Evolution* 16, 317–330. <https://doi.org/10.1006/mpev.2000.0785>

Planck Collaboration, et al., 2016. Planck 2015 results. XIII. Cosmological parameters. *Astronomy and Astrophysics* 594, A13. <https://doi.org/10.1051/0004-6361/201525830>

Polizzi, K.M., Bommarius, A.S., Broering, J.M., Chaparro-Riggers, J.F., 2007. Stability of biocatalysts. *Current Opinion in Chemical Biology, Bioinorganic chemistry / Biocatalysis and biotransformation* 11, 220–225. <https://doi.org/10.1016/j.cbpa.2007.01.685>

Porebski, B.T., Nickson, A.A., Hoke, D.E., Hunter, M.R., Zhu, L., McGowan, S., Webb, G.I., Buckle, A.M., 2015. Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain. *Protein Eng. Des. Sel.* 28, 67–78. <https://doi.org/10.1093/protein/gzv002>

Povolotskaya, I.S., Kondrashov, F.A., 2010. Sequence space and the ongoing expansion of the protein universe. *Nature* 465, 922–926. <https://doi.org/10.1038/nature09105>

Prajapati, R.S., Sirajuddin, M., Durani, V., Sreeramulu, S., Varadarajan, R., 2006. Contribution of cation- π interactions to protein stability. *Biochemistry* 45, 15000–15010. <https://doi.org/10.1021/bi061275f>

Prinston, J.E., Emlaw, J.R., Dextraze, M.F., Tessier, C.J.G., Pérez-Areales, F.J., McNulty, M.S., daCosta, C.J.B., 2017. Ancestral Reconstruction Approach to Acetylcholine Receptor Structure and Function. *Structure* 25, 1295-1302.e3. <https://doi.org/10.1016/j.str.2017.06.005>

Privalov, P.L., Khechinashvili, N.N., 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *Journal of Molecular Biology* 86, 665–684. [https://doi.org/10.1016/0022-2836\(74\)90188-0](https://doi.org/10.1016/0022-2836(74)90188-0)

Prokop, Z., Damborsky, J., Janssen, D.B., Nagata, Y., 2009. Method of production of optically active halohydrocarbons and alcohols using hydrolytic dehalogenation catalysed by haloalkane dehalogenases. US7632666B2.

Pucci, F., Rومان, M., 2017. Physical and molecular bases of protein thermal stability and cold adaptation. *Current Opinion in Structural Biology, Folding and binding • Proteins: Bridging theory and experiment* 42, 117–128. <https://doi.org/10.1016/j.sbi.2016.12.007>

Pucci, F., Rومان, M., 2016. Improved insights into protein thermal stability: from the molecular to the structurome scale. *Philos Trans A Math Phys Eng Sci* 374. <https://doi.org/10.1098/rsta.2016.0141>

Pupko, T., Pe, I., Shamir, R., Graur, D., 2000. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Mol Biol Evol* 17, 890–896.
<https://doi.org/10.1093/oxfordjournals.molbev.a026369>

Purnick, P.E.M., Weiss, R., 2009. The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology* 10, 410–422. <https://doi.org/10.1038/nrm2698>

Randall, R.N., Radford, C.E., Roof, K.A., Natarajan, D.K., Gaucher, E.A., 2016. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun* 7.
<https://doi.org/10.1038/ncomms12847>

Raveendran, S., Parameswaran, B., Ummalyma, S.B., Abraham, A., Mathew, A.K., Madhavan, A., Rebello, S., Pandey, A., 2018. Applications of Microbial Enzymes in Food Industry. *Food Technol Biotechnol* 56, 16–30. <https://doi.org/10.17113/ftb.56.01.18.5491>

Ravikumar, A., Arzumanyan, G.A., Obadi, M.K.A., Javanpour, A.A., Liu, C.C., 2018. Scalable, Continuous Evolution of Genes at Mutation Rates above Genomic Error Thresholds. *Cell* 175, 1946–1957.e13. <https://doi.org/10.1016/j.cell.2018.10.021>

Razvi, A., Scholtz, J.M., 2006. Lessons in stability from thermophilic proteins. *Protein Science* 15, 1569–1578. <https://doi.org/10.1110/ps.062130306>

Reetz, M.T., Carballeira, J.D., 2007. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nature Protocols* 2, 891–903.
<https://doi.org/10.1038/nprot.2007.72>

Reetz, M.T., Carballeira, J.D., Vogel, A., 2006. Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability. *Angewandte Chemie International Edition* 45, 7745–7751. <https://doi.org/10.1002/anie.200602795>

Reetz, M.T., Kahakeaw, D., Lohmer, R., 2008. Addressing the Numbers Problem in Directed Evolution. *ChemBioChem* 9, 1797–1804. <https://doi.org/10.1002/cbic.200800298>

Reetz, M.T., Soni, P., Fernández, L., Gumulya, Y., Carballeira, J.D., 2010. Increasing the stability of an enzyme toward hostile organic solvents by directed evolution based on iterative saturation mutagenesis using the B-FIT method. *Chem. Commun.* 46, 8657–8658.
<https://doi.org/10.1039/C0CC02657C>

Renata, H., Wang, Z.J., Arnold, F.H., 2015. Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution. *Angewandte Chemie International Edition* 54, 3351–3367. <https://doi.org/10.1002/anie.201409470>

Rhee, J.-K., Ahn, D.-G., Kim, Y.-G., Oh, J.-W., 2005. New Thermophilic and Thermostable Esterase with Sequence Similarity to the Hormone-Sensitive Lipase Family, Cloned from a Metagenomic Library. *Appl. Environ. Microbiol.* 71, 817–825. <https://doi.org/10.1128/AEM.71.2.817-825.2005>

Rigoldi, F., Donini, S., Redaelli, A., Parisini, E., Gautieri, A., 2018. Review: Engineering of thermostable enzymes for industrial applications. *APL Bioengineering* 2, 011501. <https://doi.org/10.1063/1.4997367>

Risso, V.A., Gavira, J.A., Mejia-Carmona, D.F., Gaucher, E.A., Sanchez-Ruiz, J.M., 2013. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β -Lactamases. *J. Am. Chem. Soc.* 135, 2899–2902. <https://doi.org/10.1021/ja311630a>

Risso, V.A., Manssour-Triedo, F., Delgado-Delgado, A., Arco, R., Barroso-delJesus, A., Ingles-Prieto, A., Godoy-Ruiz, R., Gavira, J.A., Gaucher, E.A., Ibarra-Molero, B., Sanchez-Ruiz, J.M., 2015. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol. Biol. Evol.* 32, 440–455. <https://doi.org/10.1093/molbev/msu312>

Risso, V.A., Sanchez-Ruiz, J.M., Ozkan, S.B., 2018. Biotechnological and protein-engineering implications of ancestral protein resurrection. *Current Opinion in Structural Biology, Engineering and design: New applications • Membranes* 51, 106–115. <https://doi.org/10.1016/j.sbi.2018.02.007>

Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C.Y., Withers, S.T., Shiba, Y., Sarpong, R., Keasling, J.D., 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440, 940–943. <https://doi.org/10.1038/nature04640>

Robert, F., Chaussidon, M., 2006. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature* 443, 969–972. <https://doi.org/10.1038/nature05239>

Rodrigue, N., Philippe, H., Lartillot, N., 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *PNAS* 200910915. <https://doi.org/10.1073/pnas.0910915107>

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D., 2004. Protein structure prediction using Rosetta. *Meth. Enzymol.* 383, 66–93. [https://doi.org/10.1016/S0076-6879\(04\)83004-0](https://doi.org/10.1016/S0076-6879(04)83004-0)

Romero, P.A., Arnold, F.H., 2009. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10, 866–876. <https://doi.org/10.1038/nrm2805>

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and

Model Choice Across a Large Model Space. *Syst Biol* 61, 539–542.

<https://doi.org/10.1093/sysbio/sys029>

Salque, M., Bogucki, P.I., Pyzel, J., Sobkowiak-Tabaka, I., Grygiel, R., Szmyt, M., Evershed, R.P., 2013. Earliest evidence for cheese making in the sixth millennium bc in northern Europe. *Nature* 493, 522–525. <https://doi.org/10.1038/nature11698>

Sammond, D.W., Kastelowitz, N., Himmel, M.E., Yin, H., Crowley, M.F., Bomble, Y.J., 2016. Comparing Residue Clusters from Thermophilic and Mesophilic Enzymes Reveals Adaptive Mechanisms. *PLOS ONE* 11, e0145848. <https://doi.org/10.1371/journal.pone.0145848>

Sandler, I., Zigdon, N., Levy, E., Aharoni, A., 2014. The functional importance of co-evolving residues in proteins. *Cell. Mol. Life Sci.* 71, 673–682. <https://doi.org/10.1007/s00018-013-1458-2>

Sanford, K., Chotani, G., Danielson, N., Zahn, J.A., 2016. Scaling up of renewable chemicals. *Current Opinion in Biotechnology, Energy biotechnology • Environmental biotechnology* 38, 112–122. <https://doi.org/10.1016/j.copbio.2016.01.008>

Sato, A., Yokotani, S., Tadokoro, T., Tanaka, S., Angkawidjaja, C., Koga, Y., Takano, K., Kanaya, S., 2011. Crystal structure of stable protein CutA1 from psychrotrophic bacterium *Shewanella* sp. SIB1. *J Synchrotron Rad* 18, 6–10. <https://doi.org/10.1107/S0909049510028669>

Savitsky, P., Bray, J., Cooper, C.D.O., Marsden, B.D., Mahajan, P., Burgess-Brown, N.A., Gileadi, O., 2010. High-throughput production of human proteins for crystallization: the SGC experience. *J. Struct. Biol.* 172, 3–13. <https://doi.org/10.1016/j.jsb.2010.06.008>

Schatsky, D., Muraskin, C., Chauhan, R., 2018. Democratizing data science to bridge the talent gap | Deloitte Insights [WWW Document]. URL <https://www2.deloitte.com/insights/us/en/focus/signals-for-strategists/democratization-of-data-science-talent-gap.html> (accessed 12.18.18).

Schellman, J.A., 2003. Protein stability in mixed solvents: a balance of contact interaction and excluded volume. *Biophys. J.* 85, 108–125. [https://doi.org/10.1016/S0006-3495\(03\)74459-2](https://doi.org/10.1016/S0006-3495(03)74459-2)

Schmidt-Dannert, C., Lopez-Gallego, F., 2016. A roadmap for biocatalysis - functional and spatial orchestration of enzyme cascades. *Microb Biotechnol* 9, 601–609. <https://doi.org/10.1111/1751-7915.12386>

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., 2005. The FoldX web server: an online force field. *Nucleic Acids Res* 33, W382–W388. <https://doi.org/10.1093/nar/gki387>

Senisterra, G., Chau, I., Vedadi, M., 2011. Thermal Denaturation Assays in Chemical Biology. *ASSAY and Drug Development Technologies* 10, 128–136. <https://doi.org/10.1089/adt.2011.0390>

Serrano, L., 2007. Synthetic biology: promises and challenges. *Molecular Systems Biology* 3, 158. <https://doi.org/10.1038/msb4100202>

Shahid, S., Ahmad, F., Hassan, M.I., Islam, A., 2015. Relationship between protein stability and functional activity in the presence of macromolecular crowding agents alone and in mixture: An insight into stability-activity trade-off. *Arch. Biochem. Biophys.* 584, 42–50. <https://doi.org/10.1016/j.abb.2015.08.015>

Shakhnovich, B.E., Deeds, E., Delisi, C., Shakhnovich, E., 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Res.* 15, 385–392. <https://doi.org/10.1101/gr.3133605>

Sharma, P.K., Kumar, R., Garg, P., Kaur, J., 2014. Insights into controlling role of substitution mutation, E315G on thermostability of a lipase cloned from metagenome of hot spring soil. *3 Biotech* 4, 189–196. <https://doi.org/10.1007/s13205-013-0142-4>

Sheldon, R.A., 2016. Biocatalysis and Green Chemistry, in: Patel, R.N. (Ed.), *Green Biocatalysis*. John Wiley & Sons, Inc, pp. 1–15. <https://doi.org/10.1002/9781118828083.ch1>

Shetty, R., 2016. Developing plant-inspired cultured aromas using a foundry for organism engineering. Presented at the 2016 SIMB Annual Meeting and Exhibition, Simb.

Shi, T., Han, P., You, C., Zhang, Y.-H.P.J., 2018. An in vitro synthetic biology platform for emerging industrial biomanufacturing: Bottom-up pathway design. *Synthetic and Systems Biotechnology* 3, 186–195. <https://doi.org/10.1016/j.synbio.2018.05.002>

Shih, P.M., Occhialini, A., Cameron, J.C., Andralojc, P.J., Parry, M.A.J., Kerfeld, C.A., 2016. Biochemical characterization of predicted Precambrian RuBisCO. *Nat Commun* 7, 10382. <https://doi.org/10.1038/ncomms10382>

Shimodaira, H., Hasegawa, M., 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* 16, 1114–1114. <https://doi.org/10.1093/oxfordjournals.molbev.a026201>

Shivange, A.V., Roccatano, D., Schwaneberg, U., 2016. Iterative key-residues interrogation of a phytase with thermostability increasing substitutions identified in directed evolution. *Appl Microbiol Biotechnol* 100, 227–242. <https://doi.org/10.1007/s00253-015-6959-5>

Shoichet, B.K., Baase, W.A., Kuroki, R., Matthews, B.W., 1995. A relationship between protein stability and protein function. *PNAS* 92, 452–456. <https://doi.org/10.1073/pnas.92.2.452>

Shortle, D., 1996. The denatured state (the other half of the folding equation) and its role in protein stability. *The FASEB Journal* 10, 27–34. <https://doi.org/10.1096/fasebj.10.1.8566543>

Siddiq, M.A., Hochberg, G.K., Thornton, J.W., 2017. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr. Opin. Struct. Biol.* 47, 113–122. <https://doi.org/10.1016/j.sbi.2017.07.003>

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. <https://doi.org/10.1038/msb.2011.75>

Singh, R., Kumar, M., Mittal, A., Mehta, P.K., 2016. Microbial enzymes: industrial progress in 21st century. *3 Biotech* 6. <https://doi.org/10.1007/s13205-016-0485-8>

Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., Krogh, A., 2001. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* 17, 425–428.

Smolke, C., Lee, S.Y., Nielsen, J., Stephanopoulos, G., 2018. *Synthetic Biology: Parts, Devices and Applications*. John Wiley & Sons.

Srinivasan, V., Lovejoy, W.S., Beach, D., 1997. Integrated Product Design for Marketability and Manufacturing. *Journal of Marketing Research* 34, 154–163. <https://doi.org/10.2307/3152072>

Stackhouse, J., Presnell, S.R., McGeehan, G.M., Nambiar, K.P., Benner, S.A., 1990. The ribonuclease from an extinct bovid ruminant. *FEBS Letters* 262, 104–106. [https://doi.org/10.1016/0014-5793\(90\)80164-E](https://doi.org/10.1016/0014-5793(90)80164-E)

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* 57, 758–771. <https://doi.org/10.1080/10635150802429642>

Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21, 456–463. <https://doi.org/10.1093/bioinformatics/bti191>

Starr, T.N., Picton, L.K., Thornton, J.W., 2017. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* 549, 409. <https://doi.org/10.1038/nature23902>

Steel, M., Penny, D., 2000. Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. *Mol Biol Evol* 17, 839–850. <https://doi.org/10.1093/oxfordjournals.molbev.a026364>

Steiner, K., Schwab, H., 2012. Recent advances in rational approaches for enzyme engineering. *Comput Struct Biotechnol J* 2. <https://doi.org/10.5936/csbj.201209010>

Steiner, T., Hess, P., Bae, J.H., Wiltschi, B., Moroder, L., Budisa, N., 2008. Synthetic Biology of Proteins: Tuning GFPs Folding and Stability with Fluoroproline. *PLOS ONE* 3, e1680. <https://doi.org/10.1371/journal.pone.0001680>

Sternke, M., Tripp, K.W., Barrick, D., 2018. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *bioRxiv* 466391. <https://doi.org/10.1101/466391>

Stolterfoht, H., Schwendenwein, D., Sensen, C.W., Rudroff, F., Winkler, M., 2017. Four distinct types of E.C. 1.2.1.30 enzymes can catalyze the reduction of carboxylic acids to aldehydes. *Journal of Biotechnology, Dedicated to Prof. Dr. Alfred Pühler on the occasion of his 75th birthday* 257, 222–232. <https://doi.org/10.1016/j.jbiotec.2017.02.014>

Strucksberg, K.H., Rosenkranz, T., Fitter, J., 2007. Reversible and irreversible unfolding of multi-domain proteins. *Biochim. Biophys. Acta* 1774, 1591–1603. <https://doi.org/10.1016/j.bbapap.2007.09.005>

Sugita, Y., Kitao, A., 1998. Dependence of Protein Stability on the Structure of the Denatured State: Free Energy Calculations of I56V Mutation in Human Lysozyme. *Biophysical Journal* 75, 2178–2187. [https://doi.org/10.1016/S0006-3495\(98\)77661-1](https://doi.org/10.1016/S0006-3495(98)77661-1)

Sullivan, B.J., Durani, V., Magliery, T.J., 2011. Triosephosphate isomerase by consensus design: dramatic differences in physical properties and activity of related variants. *J. Mol. Biol.* 413, 195–208. <https://doi.org/10.1016/j.jmb.2011.08.001>

Sullivan, B.J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., Syu, T., Magliery, T.J., 2012. Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.* 420, 384–399. <https://doi.org/10.1016/j.jmb.2012.04.025>

Sumbalova, L., Stourac, J., Martinek, T., Bednar, D., Damborsky, J., 2018. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res* 46, W356–W362. <https://doi.org/10.1093/nar/gky417>

Sun, Z.Z., Yeung, E., Hayes, C.A., Noireaux, V., Murray, R.M., 2014. Linear DNA for Rapid Prototyping of Synthetic Biological Circuits in an Escherichia coli Based TX-TL Cell-Free System. *ACS Synth. Biol.* 3, 387–397. <https://doi.org/10.1021/sb400131a>

Suplatov, D., Voevodin, V., Švedas, V., 2015. Robust enzyme design: Bioinformatic tools for improved protein stability. *Biotechnology Journal* 10, 344–355. <https://doi.org/10.1002/biot.201400150>

Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis Using Parsimony (and other methods) 4.0.b5.

Synthetic Biology Leadership Council, 2016. UK Synthetic Biology Strategic Plan 2016.

Szijártó, N., Siika-Aho, M., Tenkanen, M., Alapuranen, M., Vehmaanperä, J., Réczey, K., Viikari, L., 2008. Hydrolysis of amorphous and crystalline cellulose by heterologously produced cellulases of *Melanocarpus albomyces*. *J. Biotechnol.* 136, 140–147.

<https://doi.org/10.1016/j.jbiotec.2008.05.010>

Tachioka, M., Sugimoto, N., Nakamura, A., Sunagawa, N., Ishida, T., Uchiyama, T., Igarashi, K., Samejima, M., 2016. Development of simple random mutagenesis protocol for the protein expression system in *Pichia pastoris*. *Biotechnol Biofuels* 9, 199. <https://doi.org/10.1186/s13068-016-0613-z>

Talavera, G., Castresana, J., 2007a. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577.

<https://doi.org/10.1080/10635150701472164>

Talavera, G., Castresana, J., 2007b. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577.

<https://doi.org/10.1080/10635150701472164>

Tanaka, T., Sawano, M., Ogasahara, K., Sakaguchi, Y., Bagautdinov, B., Katoh, E., Kuroishi, C., Shinkai, A., Yokoyama, S., Yutani, K., 2006. Hyper-thermostability of CutA1 protein, with a denaturation temperature of nearly 150 degrees C. *FEBS Lett.* 580, 4224–4230.

<https://doi.org/10.1016/j.febslet.2006.06.084>

Tanimoto, K., Higashi, N., Nishioka, M., Ishikawa, K., Taya, M., 2008. Characterization of thermostable aminoacylase from hyperthermophilic archaeon *Pyrococcus horikoshii*. *FEBS J.* 275, 1140–1149. <https://doi.org/10.1111/j.1742-4658.2008.06274.x>

Tavanti, M., Porter, J.L., Sabatini, S., Turner, N.J., Flitsch, S.L., 2018. Panel of New Thermostable CYP116B Self-Sufficient Cytochrome P450 Monooxygenases that Catalyze C–H Activation with a Diverse Substrate Scope. *ChemCatChem* 10, 1042–1051. <https://doi.org/10.1002/cctc.201701510>

Taverna, D.M., Goldstein, R.A., 2002. Why are proteins marginally stable? *Proteins* 46, 105–109.

Taylor, S.V., Walter, K.U., Kast, P., Hilvert, D., 2001. Searching sequence space for protein catalysts. *PNAS* 98, 10596–10601. <https://doi.org/10.1073/pnas.191159298>

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap

penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673–4680.
<https://doi.org/10.1093/nar/22.22.4673>

Timasheff, S.N., 2002. Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9721–9726. <https://doi.org/10.1073/pnas.122225399>

Tizei, P.A.G., Csibra, E., Torres, L., Pinheiro, V.B., 2016. Selection platforms for directed evolution in synthetic biology. *Biochem Soc Trans* 44, 1165–1175. <https://doi.org/10.1042/BST20160076>

Todd, A.E., Orengo, C.A., Thornton, J.M., 2002. Plasticity of enzyme active sites. *Trends Biochem. Sci.* 27, 419–426.

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., Tawfik, D.S., 2007. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *Journal of Molecular Biology* 369, 1318–1332. <https://doi.org/10.1016/j.jmb.2007.03.069>

Tokuriki, N., Stricher, F., Serrano, L., Tawfik, D.S., 2008a. How Protein Stability and New Functions Trade Off. *PLOS Computational Biology* 4, e1000002. <https://doi.org/10.1371/journal.pcbi.1000002>

Tokuriki, N., Stricher, F., Serrano, L., Tawfik, D.S., 2008b. How Protein Stability and New Functions Trade Off. *PLOS Computational Biology* 4, e1000002. <https://doi.org/10.1371/journal.pcbi.1000002>

Tokuriki, N., Tawfik, D.S., 2009a. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology, Carbohydrates and glycoconjugates / Biophysical methods* 19, 596–604. <https://doi.org/10.1016/j.sbi.2009.08.003>

Tokuriki, N., Tawfik, D.S., 2009b. Protein dynamism and evolvability. *Science* 324, 203–207.
<https://doi.org/10.1126/science.1169375>

Toogood, H.S., Hollingsworth, E.J., Brown, R.C., Taylor, I.N., Taylor, S.J.C., McCague, R., Littlechild, J.A., 2002. A thermostable L-aminoacylase from *Thermococcus litoralis*: cloning, overexpression, characterization, and applications in biotransformations. *Extremophiles* 6, 111–122.

Tringe, S.G., Mering, C. von, Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., Bork, P., Hugenholtz, P., Rubin, E.M., 2005. Comparative Metagenomics of Microbial Communities. *Science* 308, 554–557.
<https://doi.org/10.1126/science.1107851>

Trudeau, D.L., Edlich-Muth, C., Zarzycki, J., Scheffen, M., Goldsmith, M., Khersonsky, O., Avizemer, Z., Fleishman, S.J., Cotton, C.A.R., Erb, T.J., Tawfik, D.S., Bar-Even, A., 2018. Design and in vitro

realization of carbon-conserving photorespiration. PNAS 201812605.
<https://doi.org/10.1073/pnas.1812605115>

Trudeau, D.L., Kaltenbach, M., Tawfik, D.S., 2016. On the potential origins of the high stability of reconstructed ancestral proteins. *Mol Biol Evol* msw138. <https://doi.org/10.1093/molbev/msw138>

Tsou, C.L., 1998. The role of active site flexibility in enzyme catalysis. *Biochemistry Mosc.* 63, 253–258.

Turner, N.J., 2009. Directed evolution drives the next generation of biocatalysts. *Nature Chemical Biology* 5, 567–573. <https://doi.org/10.1038/nchembio.203>

Turner, P., Mamo, G., Karlsson, E.N., 2007. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microb Cell Fact* 6, 9. <https://doi.org/10.1186/1475-2859-6-9>

Twist Bioscience, n.d. Twist Bioscience Technology | Silicon-Based Gene Synthesis [WWW Document]. URL <https://twistbioscience.com/technology> (accessed 11.18.18).

Uniprot, n.d. Current Release Statistics < Uniprot < EMBL-EBI [WWW Document]. URL <https://www.ebi.ac.uk/uniprot/TrEMBLstats> (accessed 11.16.18).

van Eunen, K., Bouwman, J., Daran-Lapujade, P., Postmus, J., Canelas, A.B., Mensonides, F.I.C., Orij, R., Tuzun, I., van den Brink, J., Smits, G.J., Gulik, V., M, W., Brul, S., Heijnen, J.J., Winde, D., H, J., Mattos, T. de, Joost, M., Kettner, C., Nielsen, J., Westerhoff, H.V., Bakker, B.M., 2010. Measuring enzyme activities under standardized in vivo-like conditions for systems biology. *The FEBS Journal* 277, 749–760. <https://doi.org/10.1111/j.1742-4658.2009.07524.x>

Verma, D., Kawarabayasi, Y., Miyazaki, K., Satyanarayana, T., 2013. Cloning, Expression and Characteristics of a Novel Alkalistable and Thermostable Xylanase Encoding Gene (Mxyl) Retrieved from Compost-Soil Metagenome. *PLOS ONE* 8, e52459.
<https://doi.org/10.1371/journal.pone.0052459>

Vialle, R.A., Tamuri, A.U., Goldman, N., Thorne, J., 2018. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol Biol Evol* 35, 1783–1797. <https://doi.org/10.1093/molbev/msy055>

Viihari, L., Alapuranen, M., Puranen, T., Vehmaanperä, J., Siika-Aho, M., 2007. Thermostable enzymes in lignocellulose hydrolysis. *Adv. Biochem. Eng. Biotechnol.* 108, 121–145.
https://doi.org/10.1007/10_2007_065

Vinay-Lara, E., Wang, S., Bai, L., Phrommao, E., Broadbent, J.R., Steele, J.L., 2016. *Lactobacillus casei* as a biocatalyst for biofuel production. *J. Ind. Microbiol. Biotechnol.* 43, 1205–1213.
<https://doi.org/10.1007/s10295-016-1797-8>

Vingron, M., Waterman, M.S., 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* 235, 1–12.

Vivoli, M., Novak, H.R., Littlechild, J.A., Harmer, N.J., 2014. Determination of protein-ligand interactions using differential scanning fluorimetry. *J Vis Exp* 51809. <https://doi.org/10.3791/51809>

Voordeckers, K., Brown, C.A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., Verstrepen, K.J., 2012. Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biol* 10. <https://doi.org/10.1371/journal.pbio.1001446>

Wagner, A., 2008. Robustness and evolvability: a paradox resolved. *Proc. Biol. Sci.* 275, 91–100. <https://doi.org/10.1098/rspb.2007.1137>

Wallace, S., Balskus, E.P., 2014. Opportunities for Merging Chemical and Biological Synthesis. *Curr Opin Biotechnol* 30, 1–8. <https://doi.org/10.1016/j.copbio.2014.03.006>

Wang, G., Tang, W., Xia, J., Chu, J., Noorman, H., Gulik, W.M. van, 2015. Integration of microbial kinetics and fluid dynamics toward model-driven scale-up of industrial bioprocesses. *Engineering in Life Sciences* 15, 20–29. <https://doi.org/10.1002/elsc.201400172>

Warshel, A., Sharma, P.K., Kato, M., Xiang, Y., Liu, H., Olsson, M.H.M., 2006. Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* 106, 3210–3235. <https://doi.org/10.1021/cr0503106>

Watanabe, K., Yamagishi, A., 2006. The effects of multiple ancestral residues on the *Thermus thermophilus* 3-isopropylmalate dehydrogenase. *FEBS Lett.* 580, 3867–3871. <https://doi.org/10.1016/j.febslet.2006.06.012>

Wedge, D.C., Rowe, W., Kell, D.B., Knowles, J., 2009. In silico modelling of directed evolution: Implications for experimental design and stepwise evolution. *J. Theor. Biol.* 257, 131–141. <https://doi.org/10.1016/j.jtbi.2008.11.005>

Weng, Y.-Z., Chang, D.T.-H., Huang, Y.-F., Lin, C.-W., 2011. A study on the flexibility of enzyme active sites. *BMC Bioinformatics* 12, S32. <https://doi.org/10.1186/1471-2105-12-S1-S32>

Westesson, O., Lunter, G., Paten, B., Holmes, I., 2012. Accurate Reconstruction of Insertion-Deletion Histories by Statistical Phylogenetics. *PLOS ONE* 7, e34572. <https://doi.org/10.1371/journal.pone.0034572>

Wheeler, L.C., Anderson, J.A., Morrison, A.J., Wong, C.E., Harms, M.J., 2018. Conservation of Specificity in Two Low-Specificity Proteins. *Biochemistry* 57, 684–695. <https://doi.org/10.1021/acs.biochem.7b01086>

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699.

Whitfield, J.H., Zhang, W.H., Herde, M.K., Clifton, B.E., Radziejewski, J., Janovjak, H., Henneberger, C., Jackson, C.J., 2015. Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci.* 24, 1412–1422. <https://doi.org/10.1002/pro.2721>

Wierckx, N., Prieto, M.A., Pomposiello, P., de Lorenzo, V., O'Connor, K., Blank, L.M., 2015. Plastic waste as a novel substrate for industrial biotechnology. *Microb Biotechnol* 8, 900–903. <https://doi.org/10.1111/1751-7915.12312>

Wijma, H.J., Floor, R.J., Janssen, D.B., 2013. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Current Opinion in Structural Biology* 23, 588–594. <https://doi.org/10.1016/j.sbi.2013.04.008>

Wijma, H.J., Floor, R.J., Jekel, P.A., Baker, D., Marrink, S.J., Janssen, D.B., 2014. Computationally designed libraries for rapid enzyme stabilization. *Protein Eng Des Sel* 27, 49–58. <https://doi.org/10.1093/protein/gzt061>

Wijma, H.J., Fürst, M.J.L.J., Janssen, D.B., 2018. A Computational Library Design Protocol for Rapid Improvement of Protein Stability: FRESCO. *Methods Mol. Biol.* 1685, 69–85. https://doi.org/10.1007/978-1-4939-7366-8_5

Wilding, M., Peat, T.S., Kalyaanamoorthy, S., Newman, J., Scott, C., Jermiin, L.S., 2017. Reverse engineering: transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chem.* 19, 5375–5380. <https://doi.org/10.1039/C7GC02343J>

Wiley, E.O., Lieberman, B.S., 2011. *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons.

Williams, P.D., Pollock, D.D., Blackburne, B.P., Goldstein, R.A., 2006. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. *PLOS Comput Biol* 2, e69. <https://doi.org/10.1371/journal.pcbi.0020069>

Williams, P.D., Pollock, D.D., Goldstein, R.A., 2007. Functionality and the evolution of marginal stability in proteins: Inferences from lattice simulations. *Evol Bioinform Online* 2, 91–101.

Willies, S., Isupov, M., Littlechild, J., 2010. Thermophilic enzymes and their applications in biocatalysis: a robust aldo-keto reductase. *Environmental Technology* 31, 1159–1167. <https://doi.org/10.1080/09593330.2010.490857>

Wilson, C., Agafonov, R.V., Hoemberger, M., Kutter, S., Zorba, A., Halpin, J., Buosi, V., Otten, R., Waterman, D., Theobald, D.L., Kern, D., 2015. Kinase dynamics. Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* 347, 882–886.
<https://doi.org/10.1126/science.aaa1823>

Winkler, A.M., Webster, M.A., Brooks, J.C., Tracey, I., Smith, S.M., Nichols, T.E., 2016. Non-parametric combination and related permutation tests for neuroimaging. *Hum Brain Mapp* 37, 1486–1511. <https://doi.org/10.1002/hbm.23115>

Winkler, J.D., Kao, K.C., 2014. Recent advances in the evolutionary engineering of industrial biocatalysts. *Genomics, Experimental evolution and the use of genomics* 104, 406–411.
<https://doi.org/10.1016/j.ygeno.2014.09.006>

Winkler, M., 2018. Carboxylic acid reductase enzymes (CARs). *Curr Opin Chem Biol* 43, 23–29.
<https://doi.org/10.1016/j.cbpa.2017.10.006>

Woese, C.R., 1987. Bacterial evolution. *Microbiol Rev* 51, 221–271.

Woese, C.R., Olsen, G.J., Ibba, M., Söll, D., 2000. Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process. *Microbiol Mol Biol Rev* 64, 202–236.

Wójcik, M., Telzerow, A., Quax, W.J., Boersma, Y.L., 2015. High-Throughput Screening in Protein Engineering: Recent Advances and Future Perspectives. *Int J Mol Sci* 16, 24918–24945.
<https://doi.org/10.3390/ijms161024918>

Wu, I., Arnold, F.H., 2013. Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. *Biotechnology and Bioengineering* 110, 1874–1883. <https://doi.org/10.1002/bit.24864>

Xia, Y., Ju, F., Fang, H.H.P., Zhang, T., 2013. Mining of Novel Thermo-Stable Cellulolytic Genes from a Thermophilic Cellulose-Degrading Consortium by Metagenomics. *PLoS One* 8.
<https://doi.org/10.1371/journal.pone.0053779>

Xiao, H., Bao, Z., Zhao, H., 2015. High Throughput Screening and Selection Methods for Directed Enzyme Evolution. *Ind. Eng. Chem. Res.* 54, 4011–4020. <https://doi.org/10.1021/ie503060a>

Yang, C.-Y., Wang, S., 2010. Computational Analysis of Protein Hotspots. *ACS Med. Chem. Lett.* 1, 125–129. <https://doi.org/10.1021/ml100026a>

Yang, H., Liu, L., Li, J., Chen, J., Du, G., 2015. Rational Design to Improve Protein Thermostability: Recent Advances and Prospects. *ChemBioEng Reviews* 2, 87–94.
<https://doi.org/10.1002/cben.201400032>

- Yang, L.-Q., Ji, X.-L., Liu, S.-Q., 2013. The free energy landscape of protein folding and dynamics: a global view. *Journal of Biomolecular Structure and Dynamics* 31, 982–992. <https://doi.org/10.1080/07391102.2012.748536>
- Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24, 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yang, Z., 2005. Phylogenetic Analysis by Maximum Likelihood (PAML; PAML FAQs) [WWW Document]. URL <http://abacus.gene.ucl.ac.uk/software/paml.html> (accessed 4.16.18).
- Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* 11, 367–372. [https://doi.org/10.1016/0169-5347\(96\)10041-0](https://doi.org/10.1016/0169-5347(96)10041-0)
- Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39, 306–314. <https://doi.org/10.1007/BF00160154>
- Yang, Z., Kumar, S., Nei, M., 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650.
- Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.-M.K., 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics* 155, 431–449.
- Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. *Nat Rev Genet* 13, 303–314. <https://doi.org/10.1038/nrg3186>
- Ye, L., Yang, C., Yu, H., 2017. From molecular engineering to process engineering: development of high-throughput screening methods in enzyme directed evolution. *Appl. Microbiol. Biotechnol.* <https://doi.org/10.1007/s00253-017-8568-y>
- Ye, Q., Bao, J., Zhong, J.-J., 2016. *Bioreactor Engineering Research and Industrial Applications I: Cell Factories*. Springer.
- Yeoman, C.J., Han, Y., Dodd, D., Schroeder, C.M., Mackie, R.I., Cann, I.K.O., 2010. Chapter 1 - Thermostable Enzymes as Biocatalysts in the Biofuel Industry, in: *Advances in Applied Microbiology*, *Advances in Applied Microbiology*. Academic Press, pp. 1–55. [https://doi.org/10.1016/S0065-2164\(10\)70001-0](https://doi.org/10.1016/S0065-2164(10)70001-0)
- You, C., Chen, H., Myung, S., Sathitsuksanoh, N., Ma, H., Zhang, X.-Z., Li, J., Zhang, Y.-H.P., 2013. Enzymatic transformation of nonfood biomass to starch. *Proc Natl Acad Sci U S A* 110, 7182–7187. <https://doi.org/10.1073/pnas.1302420110>

Yu, H., Huang, H., 2014. Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnology Advances* 32, 308–315. <https://doi.org/10.1016/j.biotechadv.2013.10.012>

Yu, H., Yan, Y., Zhang, C., Dalby, P.A., 2017. Two strategies to engineer flexible loops for improved enzyme thermostability. *Scientific Reports* 7, 41212. <https://doi.org/10.1038/srep41212>

Zakas, P., Knight, K., Parker, E.T., Spencer, H.T., Gaucher, E., Doering, C.B., 2015. Bioengineering Coagulation Factor VIII through Ancestral Protein Reconstruction. *Blood* 126, 123–123.

Zakas, P.M., Brown, H.C., Knight, K., Meeks, S.L., Spencer, H.T., Gaucher, E.A., Doering, C.B., 2017. Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nature Biotechnology* 35, 35. <https://doi.org/10.1038/nbt.3677>

Zanghellini, A., Jiang, L., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., Röthlisberger, D., Baker, D., 2006. New algorithms and an in silico benchmark for computational enzyme design. *Protein Science* 15, 2785–2794. <https://doi.org/10.1110/ps.062353106>

Závodszy, P., Kardos, J., Svingor, Á., Petsko, G.A., 1998. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci U S A* 95, 7406–7411.

Zhang, S.-W., Jin, X.-Y., Zhang, T., 2017. Gene Prediction in Metagenomic Fragments with Deep Learning [WWW Document]. *BioMed Research International*. <https://doi.org/10.1155/2017/4740354>

Zhao, X., Zhang, L., Liu, D., 2012. Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. *Biofuels, Bioproducts and Biorefining* 6, 465–482. <https://doi.org/10.1002/bbb.1331>

Zhou, R., Huang, X., Margulis, C.J., Berne, B.J., 2004. Hydrophobic Collapse in Multidomain Protein Folding. *Science* 305, 1605–1609. <https://doi.org/10.1126/science.1101176>

Zhou, T., Drummond, D.A., Wilke, C.O., 2008. Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Evol.* 66, 395–404. <https://doi.org/10.1007/s00239-008-9094-4>

Ziheng Yang, 2016. codeml gaps - Google Groups [WWW Document]. URL <https://groups.google.com/forum/#!topic/pamlsoftware/8F3msLDjczY> (accessed 12.6.18).

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology, Computation Resources for Molecular Biology* 430, 2237–2243. <https://doi.org/10.1016/j.jmb.2017.12.007>

Zinn, E., Pacouret, S., Khaychuk, V., Turunen, H.T., Carvalho, L.S., Andres-Mateos, E., Shah, S., Shelke, R., Maurer, A.C., Plovie, E., Xiao, R., Vandenberghe, L.H., 2015. In Silico Reconstruction of the Viral Evolutionary Lineage Yields a Potent Gene Therapy Vector. *Cell Rep* 12, 1056–1068.
<https://doi.org/10.1016/j.celrep.2015.07.019>

Zwanzig, R., Szabo, A., Bagchi, B., 1992. Levinthal's paradox. *Proc Natl Acad Sci U S A* 89, 20–22.


Chapter 7

Appendices

7.1 Permissions for figure 5

Nature Reviews Genetics


Order detail ID: 71699197
Order License Id: 4482450374762
ISSN: 1471-0064
Publication Type: e-Journal
Volume:
Issue:
Start page:
Publisher: NATURE PUBLISHING GROUP

Permission Status:  **Granted**
Permission type: Republish or display content
Type of use: Thesis/Dissertation
[View details](#)

7.2 Permissions for figure 6

Journal of molecular biology

Order detail ID: 71726457
Order License Id: 4492151006635
ISSN: 0022-2836
Publication Type: Journal
Volume:
Issue:
Start page:
Publisher: ACADEMIC PRESS

Permission Status:  **Granted**
Permission type: Republish or display content
Type of use: Thesis/Dissertation
[View details](#)

7.3 Characterization of carboxylic acid reductases in the toolbox of synthetic chemistry

Finnigan et al., 2017 is presented overleaf.