Functional evolution of solute-binding proteins

Ben E. Clifton

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

May 2017

Statement of authorship

The research described in this thesis was carried out between March 2013 and August 2016 under the supervision of Associate Professor Colin Jackson at the Australian National University.

Part of Section 1.3.2 is adapted from a publication; the adapted section was written by me. Chapter 2 is adapted from a publication to which I contributed all experimental and computational work and which I drafted. Section 3.2.2 is a summary of experimental work completed by Joe Kaczmarski under my supervision, which is provided primarily for context. Section 4.2.2 describes experimental work that I initiated at the University of British Columbia, under the supervision of Assistant Professor Nobuhiko Tokuriki, as part of a six-week exchange during my PhD program, and completed at the Australian National University. Section 5.2.3 describes simulations performed by me, Joe Kaczmarski, and Elaaf Mohamed after the initial submission of this thesis; this work was included in response to examiners' comments. The remainder of this thesis describes my own work except where due acknowledgment is made, and has not been previously submitted for a degree at any university.

The word count of this thesis is 55,569 words.

Ben Clifton

May 2017

Acknowledgements

First of all, I would like to thank my supervisor, Assoc. Prof. Colin Jackson, for his support and expertise. I have appreciated his optimism, generosity and savvy over the past five years, and I am grateful for the many opportunities he has given me to develop as a scientist.

I would like to thank Assistant Prof. Nobuhiko Tokuriki and the members of his group for welcoming me into the Tokuriki lab and sharing their knowledge. Visiting the Tokuriki lab was a thoroughly enjoyable and motivating experience, and invaluable for getting a fresh perspective on protein evolution.

I would like to thank Dr Paul Carr for his patient assistance with collecting and analysing X-ray diffraction data.

It has been a real pleasure to work with everyone in the Jackson group (past and present) – thank you all for your banter and your friendship. I was particularly fortunate to have the company of Joe Kaczmarski in the lab; his work on the evolution of CDT was crucial for this thesis.

I would like to thank the Research School of Chemistry and the Rickards family for financial support.

Finally, special thanks to my family and friends. Needless to say, I would not have reached this point without them.

Abstract

Solute-binding proteins (SBPs) comprise an abundant and adaptable superfamily of extracytoplasmic solute receptors involved in nutrient uptake and chemotaxis, and constitute an important component of the nutrient-scavenging arsenal in bacteria. The SBP superfamily exemplifies the power of evolution to generate functional diversity by tinkering with an existing protein fold; SBPs have evolved to recognise a wide variety of solutes with high affinity and specificity, and have also been co-opted into roles in signal transduction, transcriptional regulation and catalysis. However, the historical sequence-structure-function relationships that explain how this functional diversity could have evolved are not well understood. This thesis describes the use of ancestral protein reconstruction, a technique that leverages phylogenetic information to enable experimental characterisation of extinct proteins, to investigate two case studies of functional evolution in the SBP superfamily: the evolution of new binding specificities in the amino acid-binding protein (AABP) family, and the emergence of the enzyme cyclohexadienyl dehydratase (CDT) from a non-catalytic ancestor that belonged to the SBP superfamily.

The evolution of binding specificity in the AABP family was explored by reconstruction and functional characterisation of ancestral AABPs that predated the divergence of modern AABP subfamilies. The binding specificities of these ancestral proteins were comparable with modern AABPs, contradicting the prevailing view that ancient proteins had lower specificity than modern proteins. X-ray crystallography and isothermal titration calorimetry experiments showed that specialised glutamine-binding proteins originated from ancestral arginine-binding proteins that bound glutamine promiscuously, and that the promiscuous binding of glutamine was enabled by multi-scale conformational plasticity, water-mediated hydrogen bonding interactions and

co-option of an alternative low energy conformational sub-state productive for glutamine binding. This promiscuous binding mode was enthalpically favourable and entropically unfavourable; evolution of high-affinity glutamine-binding proteins was achieved by reduction of this entropic penalty to binding.

CDT catalyses the decarboxylative aromatisation of prephenate and arogenate; these reactions are involved in phenylalanine biosynthesis. Because CDT is closely related to non-catalytic SBPs, this enzyme provides a useful model system for understanding the emergence of catalytic activity *de novo*. The evolution of CDT from a SBP was investigated by functional characterisation of reconstructed ancestors and extant homologues of the enzyme, which showed that CDT evolved from cationic amino acid-binding proteins. Directed evolution, X-ray crystallography and molecular dynamics simulations were used to determine the genetic, structural and dynamic bases for this functional transition. These experiments showed how individual substitutions contributed to activation of the ancestral SBP scaffold for decarboxylative aromatisation of cyclohexadienols by remodelling, functionalisation and refinement of the active site. These case studies of functional evolution in the SBP superfamily provide insight into two important evolutionary processes: the evolution of protein-ligand interactions with high affinity and specificity by adaptive improvement of promiscuous interactions, and the *de novo* evolution of enzymes from non-catalytic ancestors.

Publications arising

Research articles

Clifton, B.E., Jackson, C.J. (2016). Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem. Biol. 23*, 236-245.

Book chapters

Clifton, B.E.*, Whitfield, J.H.*, Sanchez-Romero, I., Herde, M.K., Henneberger, C., Janovjak, H., Jackson, C.J. (2016) Ancestral protein reconstruction and circular permutation for improving the stability and dynamic range of FRET sensors. *Methods Mol. Biol. 1596*, 71-87. *contributed equally.

Poster presentations

Clifton, B.E., Kaczmarski, J.A., Carr, P.D., Jackson C.J. (2016) Enzyme design in the wild: evolution of an enzyme from a non-catalytic protein. Poster presented at the Synthetic Biology Cutting Edge Symposium, Canberra, ACT, Australia.

Clifton, B.E., Kaczmarski, J.A., Jackson, C. J. (2015) Functional evolution of solute-binding proteins: Diversification of binding specificity and emergence of catalytic activity. Poster presented at the Molecular Mechanisms in Evolution Gordon Research Conference, Easton, MA, U.S.A.

Clifton, B.E., Jackson, C.J. (2014) Mechanistic aspects of evolution in the amino acid-binding protein family: reconstruction of promiscuous ancestral proteins. Poster presented at the 39th Lorne Conference on Protein Structure and Function, Lorne, VIC, Australia.

Abbreviations

AABP	Amino acid-binding protein
AAT	Aromatic aminotransferase
ABC	ATP-binding cassette
"*Ca.* P. ubique."	"*Candidatus* Pelagibacter ubique"
CDT	Cyclohexadienyl dehydratase
CM	Chorismate mutase
CsArgBP	*Caldanaerobacter subterraneus* arginine-binding protein
CyiBP	Cystine-binding protein
CysBP	Cysteine-binding protein
DEBP	Aspartate/glutamate-binding protein
DHFR	Dihydrofolate reductase
DSF	Differential scanning fluorimetry
EC	Enzyme Commission
E. coli	*Escherichia coli*
ecDEBP	*Escherichia coli* aspartate/glutamate-binding protein
ecGlnBP	*Escherichia coli* glutamine-binding protein
FRET	Förster resonance energy transfer
GlnBP	Glutamine-binding protein
GO	Gene ontology
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HisBP	Histidine-binding protein
ITC	Isothermal titration calorimetry
ISOR(-R1)	Incorporation of synthetic oligonucleotides *via* gene reassembly (round 1)
LAOBP	Lysine/arginine/ornithine-binding protein
LCA	Last common ancestor
MBP	Maltose-binding protein
MD	Molecular dynamics
ML	Maximum-likelihood
NBD	Nucleotide-binding domain
NCS	Non-crystallographic symmetry
NTA	Nitrilotriacetic acid
P. aeruginosa	*Pseudomonas aeruginosa*

PaCDT	*Pseudomonas aeruginosa* cyclohexadienyl dehydratase
PC	Principal component
PCA	Principal component analysis
PDB	Protein Data Bank
PDT	Prephenate dehydratase
PEG	Polyethylene glycol
PP	Posterior probability
RMSD	Root-mean-square-deviation
SAM	*S*-adenosyl-L-methionine
SBP	Solute-binding protein
SDS PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
S. enterica	*Salmonella enterica*
seLAOBP	*Salmonella enterica* lysine/arginine/ornithine-binding protein
smFRET	Single molecule Förster resonance energy transfer
SpArgBP	*Streptococcus pneumoniae* arginine-binding protein
StEP	Staggered extension process
TMD	Transmembrane domain
W. succinogenes	*Wolinella succinogenes*

Table of Contents

1 Introduction 1
1.1 Protein evolution 2
1.1.1 Gene duplication and the evolution of new proteins 2
1.1.2 Promiscuity and its role in protein evolution 4
1.1.3 Molecular basis for protein promiscuity 8
1.1.4 Constraints on functional evolution in proteins 11
1.2 Solute-binding proteins and amino acid-binding proteins 15
1.2.1 ABC transporters: structure, mechanism and role of SBPs 16
1.2.2 Structure of SBPs 19
1.2.3 Amino acid-binding proteins 20
1.3 Methodology 23
1.3.1 Foreword 23
1.3.2 Molecular phylogenetics and ancestral protein reconstruction 23
1.3.3 Isothermal titration calorimetry 29
1.3.4 X-ray crystallography 36
1.3.5 Molecular dynamics 41
2 Evolution of binding specificity in solute-binding proteins. 45
2.1 Foreword 46
2.2 Abstract 46
2.3 Introduction 47
2.4 Results 50
2.4.1 Reconstruction of ancestral AABPs 50
2.4.2 Characterisation of ancestral proteins 54
2.4.3 Structural basis for promiscuous binding in AncQR. 61
2.4.4 Binding energetics of AABPs. 66
2.5 Discussion 70
2.6 Significance 74
2.7 Experimental Procedures 75
2.7.1 Phylogenetic analysis and ancestral reconstruction. 75
2.7.2 Cloning 76
2.7.3 Mutagenesis 76
2.7.4 Protein expression 76

2.7.5 Protein purification 77
2.7.6 Isothermal titration calorimetry 78
2.7.7 Screening experiments 78
2.7.8 Determination of binding affinities and thermodynamic parameters 79
2.7.9 Competitive displacement experiments 79
2.7.10 Structure determination 80
2.7.11 Structure analysis 81
3 Evolution of an enzyme from a solute-binding protein. Part I: Function. 82
3.1 Introduction 83
3.1.1 The origins of enzymes: design, engineering and evolution 83
3.1.2 Cyclohexadienyl dehydratase 93
3.1.3 Objectives 99
3.2 Results 100
3.2.1 Reconstruction of ancestral sequences 100
3.2.2 Functional characterisation of ancestral proteins 104
3.2.3 Functional characterisation of Ws0279 107
3.2.4 Structure and function of Pu1068 110
3.3 Discussion 120
3.4 Materials and methods 123
3.4.1 Materials 123
3.4.2 Phylogenetics and ancestral protein reconstruction 123
3.4.3 Protein expression and purification 124
3.4.4 Differential scanning fluorimetry 124
3.4.5 Crystallisation and structure determination of Pu1068. 125
3.4.6 Intrinsic tryptophan fluorescence spectroscopy 127
4 Evolution of an enzyme from a solute-binding protein. Part II: Genetics 128
4.1 Summary 129
4.2 Results 130
4.2.1 Mutational basis for CDT evolution: site-directed mutagenesis 130
4.2.2 Mutational basis for CDT evolution: directed evolution 141
4.3 Discussion 154
4.4 Materials and methods 160
4.4.1 Materials 160
4.4.2 Molecular dynamics simulations 160

4.4.3 Mutagenesis 161
4.4.4 Genetic complementation assays 161
4.4.5 Staggered extension process (StEP) 162
4.4.6 Incorporation of synthetic oligonucleotides via gene reassembly 163
4.4.7 Library creation and selection 165
5 Evolution of an enzyme from a solute-binding protein. Part III: Dynamics. 167
5.1 Introduction 168
5.1.1 Enzyme dynamics and evolution 168
5.1.2 Conformational dynamics of amino acid-binding proteins 169
5.1.3 Objectives 175
5.2 Results 176
5.2.1 Molecular dynamics simulations of PaCDT (Part I) 176
5.2.2 Crystal structures of apo PaCDT 186
5.2.3 Molecular dynamics simulations of PaCDT (Part II) 191
5.3 Discussion 195
5.4 Methods 201
5.4.1 Materials 201
5.4.2 Molecular dynamics simulations 201
5.4.3 Structure analysis 201
5.4.4 Crystallisation and structure determination of PaCDT 202
6 Evolution of an enzyme from a solute-binding protein. Part IV: Structure.. 204
6.1 Summary 205
6.2 Results 206
6.2.1 Crystal structures of AncCDT-1 and AncCDT-3(P188L) 206
6.2.2 Structural basis for catalytic activity in PaCDT 214
6.2.3 Structural basis for evolution of CDT activity 218
6.3 Discussion 226
6.4 Methods 233
6.4.1 Crystallisation and structure determination of AncCDT-1 233
6.4.2 Crystallisation and structure determination of AncCDT-3(P188L) 234
7 Conclusions 235
8 References 242
Appendix I 264

Figures

Figure 1.1. Architecture of the ABC transporter $MalFGK_2$. 17
Figure 1.2. Representative structures of solute-binding proteins. 20
Figure 1.3. Effect of the parameter c on the shape of binding isotherms from ITC. 32
Figure 2.1. Reconstruction of ancestral amino acid-binding proteins. 52
Figure 2.2. Phylogenies of the AABP family 54
Figure 2.3. Binding specificity of ancestral and extant AABPs. 55
Figure 2.4. Promiscuous binding of amino acids by extant AABPs. 60
Figure 2.5. Crystal structures of AncQR. 63
Figure 2.6. Structural similarity between the AncQR-Arg complex and extant l-arginine-binding proteins. 65
Figure 2.7. Contrasting enthalpic and entropic modes of binding in the AABP family. 68
Figure 2.8. Binding enthalpies for the AncQR-Arg, AncQR-Gln and GlnBP-Gln interactions are independent of buffer ionisation enthalpy. 69
Figure 3.1. Architectures of two SBP-like enzymes. 92
Figure 3.2. Structural evidence for the evolutionary relationship between CDT and SBPs. 97
Figure 3.3. Evolution of CDT from SBPs. 101
Figure 3.4. Posterior probability distributions of ancestral CDT variants. 103
Figure 3.5. Genetic complementation of phenylalanine auxotrophs by ancestral and extant CDT variants. 105
Figure 3.6. Trimeric structure of Ws0279. 107
Figure 3.7. Amino acid binding profiles of Ws0279 and AncCDT-1. 109
Figure 3.8. Genomic context of Pu1068 in “Ca. P. ubique” strain HTCC1062. 110
Figure 3.9. Pu1068 is not an amino acid-binding protein. 111
Figure 3.10. Crystal structure of Pu1068. 114
Figure 3.11. DSF screening of Pu1068 against small molecule libraries. 116
Figure 3.12. Characterisation of the interaction between Pu1068 and NDSB-221 by fluorescence spectroscopy. 117
Figure 3.13. Crystal structure of the Pu1068/NDSB-221 complex. 119
Figure 4.1. Crystal structure of the PaCDT-HEPES complex. 131
Figure 4.2. Docking of l-arogenate into the active site of PaCDT. 131
Figure 4.3. Persistence of interactions between PaCDT and l-arogenate during MD simulations. 136
Figure 4.4. Trp60 and surrounding residues reshape the active site of PaCDT. 136
Figure 4.5. Sequence conservation in CDT homologues. 137
Figure 4.6. Positions of amino acid substitutions between AncCDT-1 and AncCDT-3. 139
Figure 4.7. AncCDT-3(P188L) supports rapid growth of phenylalanine auxotrophs in minimal media. 144
Figure 4.8. Expression of CDT variants obtained by directed evolution. 151
Figure 4.9. Posterior probability distributions of ancestral CDT variants at mutated positions. 153
Figure 5.1. Open and closed conformations of an amino acid-binding protein. 169

Figure 5.2. Ligand-induced conformational change by the induced fit and conformational selection mechanisms. 171
Figure 5.3. Oligomeric structure of PaCDT. 177
Figure 5.4. PaCDT samples an open conformation during MD simulations. 178
Figure 5.5. Principal component analysis of PaCDT simulations. 179
Figure 5.6. Structural interpretation of the major principal components of PaCDT trajectories. 180
Figure 5.7. Projections of PaCDT trajectories onto the PC1 axis. 182
Figure 5.8. Differences in hinge structure between open and closed conformations of PaCDT. 183
Figure 5.9. Principal component analysis of PaCDT-arogenate simulations 184
Figure 5.10. Crystal structure of apo CDT. 189
Figure 5.11. Active site cavity of apo PaCDT. 190
Figure 5.12. Packing of PaCDT crystals in space groups H3 and $P4_322$. 190
Figure 5.13. Extended molecular dynamics simulations of PaCDT. 194
Figure 6.1. Crystal structure of AncCDT-1. 209
Figure 6.2. Crystal structure of AncCDT-3(P188L). 211
Figure 6.3. Size-exclusion chromatogram of AncCDT-3(P188L). 212
Figure 6.4. Predicted substrate binding modes in apo-PaCDT and implications for catalysis. 215
Figure 6.5. CDT inherited the amino acid-binding structural motif from AABPs. 219
Figure 6.6. Role of Trp60 and surrounding residues in the evolution of CDT. 220
Figure 6.7. Functionalisation of the AncCDT-1 binding site for CDT activity. 222
Figure 6.8. Indirect mutational effects in the evolution of CDT. 225

Tables

Table 2.1. Binding affinities and thermodynamic parameters for amino acid binding to ancestral and extant AABPs. 56
Table 2.2. Verification of the binding specificity of ancestral and extant AABPs. 58
Table 2.3. Data collection and refinement statistics for AncQR structures. 64
Table 3.1. Enzymes with the type II SBP fold. 90
Table 3.2. Mean posterior probabilities (PPs) of ancestral CDT variants. 102
Table 3.3. Data collection and refinement statistics for Pu1068 113
Table 4.1. Variants of AncCDT-2 tested for CDT activity by genetic complementation. 140
Table 4.2. Composition of StEP libraries. 143
Table 4.3. Sequences of CDT variants isolated from the AncCDT-2D2/3 StEP library. 145
Table 4.4. Composition of ISOR libraries. 147
Table 4.5. Sequences of CDT variants isolated from ISOR libraries. 148
Table 4.6. Site-directed mutagenesis of AncCDT-2D2 and AncCDT-3. 150
Table 4.7. Sequences of primers used for directed evolution. 163
Table 4.8. Mutagenic oligonucleotides used for ISOR. 164
Table 5.1. Data collection and refinement statistics for PaCDT. 187

Table 5.2. Overlap in conformational space sampled during different simulations of PaCDT. 192
Table 6.1. Data collection and refinement statistics for AncCDT-1 and AncCDT-3(P188L). 207

Schemes

Scheme 3.1. Mechanism of the Kemp elimination, a model reaction for enzyme design. 85
Scheme 3.2. Multiple pathways of l-phenylalanine biosynthesis in P. aeruginosa. 94
Scheme 3.3. Possible mechanisms for the elimination of CO_2 and H_2O from prephenate. 98
Scheme 6.1. Proposed basis for transition state stabilisation in PaCDT..... 216

Chapter One

Introduction

1.1 Protein evolution

1.1.1 Gene duplication and the evolution of new proteins

The classic model describing how new proteins arise by gene duplication events is the mutation during non-functionality model, also known as the neofunctionalisation model, originally proposed by Ohno (Bergthorsson et al., 2007; Conant and Wolfe, 2008; Innan and Kondrashov, 2010; Ohno, 1970). According to the mutation during non-functionality model, gene duplication is neutral, that is, without fitness cost or benefit. Because one copy of the duplicated gene is redundant, it is freed from selective pressure and can accumulate mutations. By chance, the gene may acquire a mutation that confers a new function with a corresponding fitness benefit, and as a result, it will be maintained in a population by purifying selection and improved by positive selection. Although some examples of mutation during non-functionality have been uncovered (Conant and Wolfe, 2008), the model is undermined by the observation that gene duplication is not neutral, but has a fitness cost associated with DNA synthesis, protein expression, and deleterious gene dosage effects (Bergthorsson et al., 2007; Soskine and Tawfik, 2010). More importantly, deleterious mutations are far more common than gain-of-function mutations, and the cumulative fitness costs of deleterious mutations in a gene are exponential or steeper than exponential (Bergthorsson et al., 2007; Soskine and Tawfik, 2010). Thus, when a gene is removed from selection for the original function, loss of function is far more likely than neofunctionalisation. On the other hand, it is assumed that if the gene duplicate is not removed from selection, functional divergence cannot occur. This contradiction has been described as Ohno's dilemma (Bergthorsson et al., 2007).

In recognition of the limitations of the mutation during non-functionality model, alternative mechanisms for protein functional evolution by gene duplication have been proposed (Conant and Wolfe, 2008; Innan and Kondrashov, 2010). Examples include the

duplication-degeneration-complementation or “subfunctionalisation” model, the escape from adaptive conflict model, and the innovation-amplification-divergence model (Barkman and Zhang, 2009; Innan and Kondrashov, 2010; Des Marais and Rausher, 2008; Voordeckers et al., 2012). Although there is currently no consensus on which mechanism is predominant, they can together be distinguished from the mutation during non-functionality model in two respects: (i) functional divergence is assumed to occur before gene duplication, such that functional evolution proceeds via multifunctional genes as intermediates, and (ii) gene duplicates are assumed to be subject to continuous purifying selection.

For example, the escape from adaptive conflict model posits that a new function arises in a single-copy gene (Hughes, 1994; Des Marais and Rausher, 2008). The resulting bifunctional gene is ultimately subject to an adaptive conflict, whereby one function cannot be improved without a detrimental effect on the other function. A gene duplication event, which is assumed to be neutral, allows this adaptive conflict to be resolved, as each gene duplicate can become specialised for a single function under positive selection. On the other hand, the innovation-amplification-divergence model begins with a protein that has a secondary function (for example, an enzymatic activity) present at a low level in addition to its primary function (Bergthorsson et al., 2007; Näsvall et al., 2012). When this secondary function becomes valuable, gene duplicates can be fixed by positive selection, since the secondary function can be amplified by increased dosage of the gene. After mutations that improve the secondary function are fixed by positive selection, selection for the remaining duplicates of the original gene is relaxed and superfluous duplicates are eliminated from the population. The particular strength of the innovation-amplification-divergence model is that each step is governed by positive selection, providing an impetus for gene duplication and reducing the risk of loss of function during genetic drift.

1.1.2 Promiscuity and its role in protein evolution

As discussed in the previous section, current models of protein functional evolution by gene duplication can resolve Ohno's dilemma but depend on the existence of proteins with secondary functions. Consistent with this requirement, proteins with secondary functions are common. Many proteins are multifunctional; for example, a recent survey of the *Escherichia coli* proteome found that 37% of metabolic enzymes are multifunctional and together account for 65% of metabolic reactions (Nam et al., 2012). Furthermore, proteins are often promiscuous – that is, they have functions that are present at a low level, but have no physiological significance and are not maintained by selection (Khersonsky and Tawfik, 2010). In the parlance of evolutionary biochemistry, promiscuity is usually distinguished from multi-specificity or multi-functionality, which refer to the presence of multiple functions that are maintained by selection (Copley, 2015).

The most compelling evidence for the prevalence of promiscuity in proteins comes from genome-wide genetic complementation experiments (Desai and Miller, 2010; Patrick et al., 2007; Soo et al., 2011) and systematic profiling of promiscuous activities within enzyme superfamilies (Baier and Tokuriki, 2014; Colin et al., 2015; Huang et al., 2015). For example, it has been shown that 20% of auxotrophic phenotypes originating from single-gene knockouts in *E. coli* can be suppressed by overexpression of a different *E. coli* gene, in many cases due to substrate promiscuity or catalytic promiscuity of an unrelated suppressor gene (Patrick et al., 2007). As another example, substrate profiling of the haloalkanoate dehalogenase superfamily showed that substrate promiscuity was very common; of 217 enzymes screened for phosphatase activity against a library of 167 substrates, 70% acted on more than 5 substrates, while 23% acted on more than 40 substrates (Huang et al., 2015).

The prevalence of promiscuity probably results from both biophysical and evolutionary constraints on protein specificity (Tawfik, 2010). Perfect specificity in enzymes, for example, may be unachievable because of the inherent reactivity of nucleophiles, general acids, general bases or cofactors in enzyme active sites and the impossibility of excluding all molecules except the native substrate from the active site (Copley, 2015). Obligatory trade-offs between activity and specificity have also been suggested as an explanation for imperfect specificity in enzymes; the enzyme D-ribulose-1,5-bisphosphate carboxylase/oxygenase, responsible for carbon fixation, is an important example (Tcherkez et al., 2006). The idea of biophysical constraints on enzyme specificity is also supported by the existence of proofreading mechanisms in enzymes such as DNA polymerases, in which high fidelity is essential and is achieved by correction of incorrectly incorporated nucleotides after the fact, rather than prevention of incorrect incorporation by exceptional specificity (Copley, 2015). Evolutionary constraints on protein specificity arise because promiscuous functions are neutral, by definition, so elimination of a promiscuous function confers no fitness benefit. This principle has been invoked to explain patterns of promiscuous binding in a family of steroid receptors; it was shown that steroid receptors evolved the ability to discriminate non-cognate hormones from their specific regulatory hormone only if the non-cognate hormone was endogenously present in the cellular environment (Eick et al., 2012).

The promiscuous functions of proteins constitute a latent source of functional novelty that can be exploited for the evolution of proteins with new specialities given the appropriate selective pressure (Copley, 2015; Khersonsky and Tawfik, 2010; Schulenburg and Miller, 2014). Directed evolution experiments, in which the properties of proteins are improved by mimicking the evolutionary process of iterative random mutagenesis and selection in a controlled laboratory setting, and other protein engineering experiments have shown that promiscuous functions can often be increased to native-like

levels by sequential point mutations (Bloom and Arnold, 2009). Reconstruction and characterisation of ancestral proteins (discussed in Section 1.3.2), has also provided some evidence for improvement of promiscuous functions as a mechanism for adaptive evolution (Boucher et al., 2014; Bridgham et al., 2006). For example, it was shown that the aldosterone-responsive mineralocorticoid receptor evolved by recruitment of an ancestral corticoid receptor that bound aldosterone promiscuously as a by-product of its affinity for structurally related ligands (Bridgham et al., 2006). Notably, because the ancestral corticoid receptor predated the evolution of aldosterone biosynthesis, it could be shown conclusively that the interaction between aldosterone and the ancestral receptor was promiscuous rather than adaptive. Adaptive improvement of a weak secondary activity in an enzyme has been demonstrated in the experimental evolution of bacterial populations *in vitro* (Näsvall et al., 2012). Finally, the evolution of adaptive traits from initially non-adaptive traits is observed analogously in higher biological systems such as metabolic networks and organismal morphology (Barve and Wagner, 2013; Gould and Vrba, 1982).

Differences in the plasticity of promiscuous and native functions of proteins have important consequences for functional evolution. Promiscuous functions often have high plasticity and evolvability, in that large improvements in a promiscuous function can be achieved with few mutations (Aharoni et al., 2005; Khanal et al., 2015). Promiscuous functions often vary between different members of a protein family and during neutral drift; neutral drift can therefore expose new promiscuous functions and open up new opportunities for functional expansion of protein families (Amitai et al., 2007; Bloom et al., 2007). On the other hand, it is commonly observed that the native functions of proteins have greater mutational robustness and lower plasticity, which implies that improvements in a promiscuous function can be achieved without loss of the native function (Khersonsky and Tawfik, 2010). Weak trade-offs between the promiscuous and native

function of a protein result in bifunctional evolutionary intermediates (Khersonsky et al., 2010).

The promiscuous functions of proteins can provide information about their evolutionary histories, since the native function of one member of a protein superfamily is often identified as a promiscuous function in other members of the superfamily (Khersonsky and Tawfik, 2010). For example, the identification of promiscuous lactonase activity in a phosphotriesterase, an enzyme that evolved recently in response to anthropogenic insecticides, led to the discovery of lactonases with promiscuous phosphotriesterase activity, which likely resemble the ancestor of the phosphotriesterase family (Afriat et al., 2006). A credible mutational pathway between the two enzymes could then be deciphered (Afriat-Jurnou et al., 2012).

Multifunctional or promiscuous proteins have long been considered to have an important historical role in the evolution of specialised proteins in modern organisms. Jensen argued that primordial proteins were versatile generalists that possessed broad substrate specificity (Jensen, 1976), an idea that is readily extended to incorporate catalytic promiscuity (O'Brien and Herschlag, 1999). Although this hypothesis is an attractive explanation for how diversity in metabolic functions could be achieved in a primordial cell with relatively few genes, the evidence for this hypothesis is mostly indirect. For example, the large number of extant multifunctional enzymes has been interpreted as a vestige of the evolution of specialised modern enzymes from generalist ancestral enzymes; for the enzymes that remained multifunctional, selection pressure for higher metabolic flux, which could be achieved by subfunctionalisation of the ancestral enzymes, may have been insufficient to offset the costs of gene duplication and maintenance (Nam et al., 2012). On the other hand, some attempts have been made to assess the promiscuity of ancient proteins directly using ancestral protein reconstruction

(Devamani et al., 2016; Risso et al., 2013). Reconstruction and experimental characterisation of β-lactamases from the Precambrian era provided evidence that the modern penicillin-hydrolysing β-lactamase TEM-1 evolved from a generalist ancestor that could also hydrolyse other antibiotics efficiently (Risso et al., 2013). However, given the lack of evidence for pronounced promiscuity-to-specificity trends in other protein families (Wheeler et al., 2016), further work is necessary to determine whether these results can be generalised.

1.1.3 Molecular basis for protein promiscuity

Given the importance of promiscuity as a source of phenotypic novelty, as well the biomedical significance of promiscuous binding as a source of adverse side effects when drugs bind to unintended targets, an improved understanding of the biophysical basis for binding promiscuity is desirable (Babtie et al., 2010; Gatti-Lafranconi and Hollfelder, 2013; Nobeli et al., 2009). Binding promiscuity – the fortuitous ability of a protein to bind a non-native ligand or substrate – is often a prerequisite for functional promiscuity (Nobeli et al., 2009). The classical hydrophobic effect – displacement of water molecules from hydrophobic surfaces for entropic benefit – appears to be particularly important for mediating interactions between proteins and non-native ligands or substrates. In contrast to specific electrostatic interactions and directional hydrogen bonds, hydrophobic interactions do not require precise and complementary orientation of protein and ligand functional groups, and are therefore intrinsically permissive (Gatti-Lafranconi and Hollfelder, 2013). In several enzymes, such as the non-ribosomal peptide synthetase TcyA, a correlation between substrate hydrophobicity and catalytic efficiency is observed for promiscuous substrates (Khersonsky and Tawfik, 2005; Villiers and Hollfelder, 2009). In other cases, however, adventitious hydrogen bonding interactions are more important for promiscuous binding than non-specific hydrophobic effects (Eick et al., 2012; James and Tawfik, 2003a). For example, promiscuous binding of certain steroids

in ancestral steroid receptors was shown to originate from the excess hydrogen bonding capacity of unsatisfied polar residues in a permissively spacious binding pocket (Eick et al., 2012). Although these alternative interaction modes would be expected to differ in their thermodynamic signature (see Section 1.3.3), there is a paucity of experimental thermodynamic data for promiscuous protein-ligand interactions, and the thermodynamic basis for binding promiscuity is therefore poorly understood (Nobeli et al., 2009).

The properties of water molecules, such as their capacity to accept or donate hydrogen bonds, and their lack of steric constraints, give them particular adaptability for mediating promiscuous interactions between proteins and ligands (Ladbury, 1996). The oligopeptide-binding protein OppA, a multi-specific solute-binding protein that binds KXK peptides with any central residue, provides an illustrative example. Exhaustive crystallographic investigations showed that the multi-specificity of OppA originates from a hydrated binding cavity that can accommodate any peptide side-chain; the water molecules offer electrostatic shielding for charged side chains, satisfy the hydrogen bonding potential of polar side chains, and can be displaced for entropic benefit by non-polar side chains (Sleigh et al., 1999; Tame et al., 1996).

The conformational plasticity of proteins is also considered an important factor in promiscuity (Gatti-Lafranconi and Hollfelder, 2013; James and Tawfik, 2003b; Nobeli et al., 2009; Tokuriki and Tawfik, 2009a). The role of conformational plasticity in functional promiscuity is usually framed in terms of the conformational selection model of molecular recognition (Tokuriki and Tawfik, 2009a), discussed further in Section 5.1.2, although it should be emphasised that analogous arguments are compatible with the alternative induced fit model of molecular recognition. Briefly, according to the conformational selection model, proteins are conformationally hetereogeneous and sample an ensemble of conformational sub-states that are energetically accessible at a

given temperature (Boehr et al., 2009; Ma et al., 2002). A particular ligand will preferentially stabilise a specific conformation from this ensemble, resulting in a population shift towards that conformation in the protein-ligand complex. A protein with a high degree of conformational plasticity can sample more conformations in the unbound state and therefore provides a potential promiscuous ligand with more opportunities for fortuitous interactions (James and Tawfik, 2003b; Tokuriki and Tawfik, 2009a). One example of promiscuous binding mediated by a conformational selection mechanism comes from a multi-specific antibody, which was crystallised in different conformations with different ligands and displayed multiphasic binding kinetics indicative of a conformational selection mechanism, with a fast phase corresponding to binding of the ligand to a minor conformational isomer, followed by a slow phase when interconversion between the major and minor conformational isomers becomes rate-limiting for binding (James et al., 2003).

However, the connection between conformational plasticity, promiscuity and the evolution of new protein functions has been established explicitly only in a few cases (Hudson et al., 2015; Zou et al., 2014). Hudson *et al*. showed that an ancestral steroid hormone receptor consisted of a structural ensemble that could bind to different DNA response elements responsible for activating and repressing transcription. This depended on the ability of the receptor to sample multiple DNA-bound conformations (Hudson et al., 2015). Subdivision of the conformational space accessible to the ancestral receptor *via* mutations that selectively restricted backbone conformational dynamics was responsible for functional divergence between different lineages; in the glucocorticoid receptors, improved allosteric communication between subunits led to an enhancement of transcriptional repression, whereas in the mineralocorticoid receptors, the ability to bind the negative response element and the ability to repress transcription were lost.

1.1.4 Constraints on functional evolution in proteins

The studies discussed in Section 1.1.2 demonstrate that promiscuity is a key determinant of protein evolvability. However, other structural and biophysical properties of proteins also determine whether mutational pathways that lead to the adaptive improvement of promiscuous functions are viable. The fitness landscape is a useful metaphor to explore these determinants of evolvability (Kaltenbach and Tokuriki, 2014; de Visser and Krug, 2014). Each protein sequence is represented by a node in sequence space, connected by edges to all sequences that are related by single amino acid substitutions, and each node is mapped to a fitness value. Compression of the resulting high-dimensional landscape into a representation of fitness as a one- or two-dimensional function of protein sequence yields the familiar topographical representation of a protein fitness landscape. Ultimately, the fitness of a protein is interpreted as its contribution to organismal survival and reproduction, which determines the spread of the corresponding gene through a population, but fitness landscapes based on surrogate phenotypic properties such as catalytic activity can also be considered (Kaltenbach and Tokuriki, 2014; Romero and Arnold, 2009). Mutation and selection drive the evolution of proteins towards peaks in the fitness landscape, that is, until the protein cannot be improved further by single amino acid substitutions.

Since non-functional proteins are eliminated rapidly by purifying selection, evolutionary trajectories between fitness peaks must follow a network of functional proteins (Maynard Smith, 1970). Therefore, the improvement of a promiscuous function by adaptive evolution can be visualised as a trajectory along the fitness landscape from one peak, corresponding to the native function, towards a different, overlapping peak corresponding to the promiscuous function (Baier and Tokuriki, 2014; Kaltenbach and Tokuriki, 2014). In the simplest case, this trajectory would be smooth, with each mutation contributing to an increase in fitness, as seen frequently in directed evolution experiments

(Romero and Arnold, 2009; Tracewell and Arnold, 2009). In nature, however, evolutionary trajectories through sequence space are often constrained by epistasis – the context dependence of mutations. For example, a mutation required for a new function may be tolerated in one genetic background, in which case the fitness peak would be accessible, but deleterious in another genetic background, in which case the fitness peak would be inaccessible. Thus, epistasis makes the topology of a fitness landscape "rugged" and is a major constraint on protein evolution.

There is extensive evidence for the prevalence of epistatic interactions between amino acid substitutions in proteins (reviewed in Starr and Thornton, 2016), including bioinformatic analysis of orthologous proteins showing that substitutions are far less frequent than expected from site-specific amino acid tolerances and the rate of neutral evolution (Breen et al., 2012), demonstration of the differential effects of mutations at different points in natural, experimental and *in silico* evolutionary trajectories (Gong et al., 2013; Kaltenbach et al., 2015; Miton and Tokuriki, 2016; Shah et al., 2015), and the observation that mutations that interconvert residues in orthologous proteins are frequently deleterious (Lunzer et al., 2010). On the other hand, epistasis is not a universal phenomenon; the conservation of site-specific amino acid preferences (Doud et al., 2015; Risso et al., 2014) and the parallel evolution of biochemical traits through identical substitutions in different lineages (Harms and Thornton, 2013) provide examples of mutational effects that are independent of genetic background.

The non-linear correlation between protein thermostability and fitness is common source of epistasis (Starr and Thornton, 2016). Natural selection imposes a threshold requirement on thermostability; a functional protein typically needs to be stable enough to fold under physiological conditions, but otherwise there is little selection for thermostability (Bershtein et al., 2006). Stability-mediated epistasis arises because a

destabilising mutation in a marginally stable protein results in an unfolded and non-functional protein, imposing a severe fitness penalty, while an equally destabilising mutation in a protein with excess stability is neutral. The "permissive" effects of mutational robustness in proteins have been documented in natural and experimental evolution (Bloom et al., 2006; Gong et al., 2013; Tokuriki and Tawfik, 2009b; Wang et al., 2002). For example, reconstruction of intermediates in the evolution of influenza nucleoprotein showed that the protein acquired three destabilising substitutions that were neutral at the time they were fixed, but deleterious to viral replication in the ancestral background (Gong et al., 2013). The fitness effects of the three destabilising substitutions were counteracted by compensatory stabilising substitutions that were acquired during evolution. Notably, one of the compensatory mutations could rescue any of the destabilised variants, supporting the hypothesis that epistasis originated from global thermostability.

Although stability-mediated epistasis is non-specific, in the sense that the effect of a destabilising mutation at one site can be countered by a stabilising mutation at any other site, direct interactions between residues, which cause a mutation at one site to modulate the effect of a mutation at a neighbouring site, result in specific epistasis (Starr and Thornton, 2016). Specific epistasis is a source of historical contingency in evolution, since the tolerance of one mutation might depend on a rare, neutral and permissive mutation at a neighbouring site (Harms and Thornton, 2014). For example, the evolution of the cortisol-responsive glucocorticoid receptor from an ancestral mineralocorticoid-responsive steroid receptor required five mutations that switched the specificity of the ancestral receptor from mineralocorticoids to cortisol, as well as two permissive mutations (Ortlund et al., 2007). These permissive mutations were neutral in the ancestral background, but were necessary for the function-switching mutations to be tolerated, because they stabilised specific structural elements that were destabilised by the function-

switching mutations. In subsequent work, no alternative permissive mutations could be recovered from a library of ~12500 variants, showing that mutations that could stabilise the specific structural changes associated with the new function, while also being tolerated in the ancestral background, were extremely rare (Harms and Thornton, 2014).

Other consequences of epistasis for protein evolution, in addition to making evolution contingent on stochastically fixed neutral mutations, include limitations on the reversibility of genotypic evolution (Bridgham et al., 2009; Kaltenbach et al., 2015), constraints on the viability of different evolutionary trajectories to a fitness peak (Noor et al., 2012; Weinreich et al., 2006), and limitations on the accessibility of the global fitness optimum, as opposed to local fitness optima (Dickinson et al., 2013; Salverda et al., 2011).

Finally, the potential for functional evolution in proteins is dependent on protein fold. Some folds, such as the triosephosphate isomerase barrel fold, support a large number of enzymatic activities, while others, such as the dihydrofolate reductase fold, support only a few enzymatic activities; the diversity of enzymatic activities in superfamilies diminishes according to a power law distribution (Nobeli et al., 2009; Tóth-Petróczy and Tawfik, 2014). Folds with natural functional diversity are also more likely to be suitable scaffolds for protein design (Röthlisberger et al., 2008) and promising targets for protein engineering (O'Loughlin et al., 2006). Thermostability, which confers mutational robustness, and conformational plasticity, which enables sampling of conformations that may mediate alternative functions, are two determinants of protein evolvability that are likely to be fold-dependent. Although thermostability and conformational plasticity may seem to be mutually exclusive, they can be achieved simultaneously in folds with high "polarity", that is, folds in which a robust scaffold containing critical catalytic residues is juxtaposed with flexible loops that can modulate

new substrate specificities and reactivities (Dellus-Gur et al., 2013; Tóth-Petróczy and Tawfik, 2014). For example, the evolvable triosephosphate isomerase barrel fold consists of a core α/β barrel structure with mobile and adaptable loops on its periphery. Unfortunately, systematic explorations of fold evolvability have not yet been extended to non-catalytic proteins.

1.2 Solute-binding proteins and amino acid-binding proteins

Solute-binding proteins (SBPs; also known as periplasmic binding proteins or substrate-binding proteins) are soluble extracytoplasmic receptors for small molecules. SBPs are predominantly involved in solute transport and signal transduction; transport is achieved by association of SBPs with the integral membrane components of ATP-binding cassette (ABC) importers or tripartite ATP-independent periplasmic transporters, whereas signal transduction across the cell membrane is achieved by association of SBPs with G-protein coupled receptors, ligand-gated ion channels, chemotactic receptors or two-component regulatory systems (Berntsson et al., 2010). Some proteins related to SBPs have intracellular functions as transcriptional regulators (for example, the *lac* repressor) or enzymes (see Section 3.1.1). SBPs bind a large range of solutes, such as amino acids, carbohydrates, vitamins, metals, and osmolytes, and bind their physiological ligands with high affinity (K_d ~10 nM to 10 μM) (Davidson et al., 2008). This section gives a general overview of the structure, function and evolution of SBPs, particularly those associated with ABC importers, before focussing on a specific family of ABC importer-associated SBPs: the polar amino acid-binding protein (AABP) family, which is the subject of this thesis.

1.2.1 ABC transporters: structure, mechanism and role of SBPs

ABC transporters are integral membrane proteins that function as primary active transporters, transporting solutes against a concentration gradient by coupling transport to the hydrolysis of ATP. Both ABC importers and exporters are found in prokaryotes, while eukaryotes possess ABC exporters only (Davidson et al., 2008). The basic architecture of ABC transporters (reviewed in Rees et al., 2009) consists of two transmembrane domains (TMDs), embedded in the cell membrane, and two cytoplasmic nucleotide-binding domains (NBDs), located in the cytoplasm (Figure 1.1). Whereas the TMDs are highly diverse in sequence and structure, reflecting the wide variety of solutes transported, the NBDs, which form the catalytic site for ATP hydrolysis, are highly conserved and contain a number of characteristic sequence motifs. NBDs consist of a RecA-like subdomain, containing the Walker A motif (GXXGXGK(S/T)), the Walker B motif (φφφφD, where φ is a hydrophobic residue), the Q-loop and the H-loop (named for conserved Gln and His residues), and a helical subdomain containing the ABC signature motif (LSGGQ). The NBDs form a head-to-tail dimer, in which a single catalytic site is formed between the Walker A motif of one domain and the ABC signature motif of another domain. Each of the sequence motifs has a role in coordination of ATP or Mg^{2+}, or polarisation of the hydrolytic water molecule (Oldham and Chen, 2011a).

In the case of ABC importers, solute transport usually depends on an extracytoplasmic SBP in addition to the TMDs and NBDs; SBP-independent ABC importers of the energy coupling factor family are an exception (Zhang, 2013). SBPs generally diffuse freely in the periplasm in Gram-negative bacteria, whereas in Gram-positive bacteria and archaea, they are anchored to the cell membrane by lipids or fused to the TMDs of the transporter (Davidson et al., 2008). Although both SBPs and TMDs have ligand binding sites, the SBP binds with higher affinity and is the primary determinant of the specificity of an ABC importer (Davidson et al., 2008).

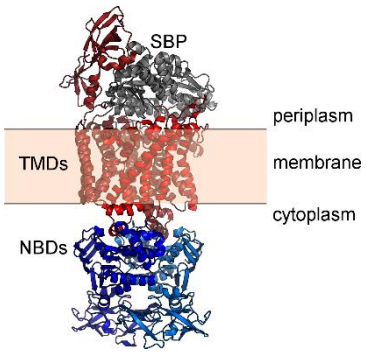


Figure 1.1. Architecture of the ABC transporter MalFGK$_2$. The crystal structure of the maltose transporter MalFGK$_2$ (PDB: 3RLF) shows the three components of an ABC importer: an extracytoplasmic solute-binding protein (SBP, grey), a dimer of transmembrane domains (TMDs, red), and a dimer of nucleotide-binding domains (NBDs, blue).

The mechanism by which ATP importers couple ATP hydrolysis to translocation of a solute across the cell membrane can be explained by the alternating access model. In the case of the maltose transporter from *E. coli* (MalFGK$_2$), the alternating access model is supported by multiple crystal structures of the transporter at different stages of the translocation cycle (Chen, 2013; Khare et al., 2009; Oldham and Chen, 2011b; Oldham et al., 2007), which can be reconciled with a wealth of biochemical data (reviewed in Bordignon et al., 2010). According to the alternating access model, ABC importers alternate between an "inward-facing" conformation, in which the solute binding site of the TMD is accessible from the cytoplasm, and an "outward-facing" conformation, in which the solute binding site of the TMD is externally accessible; these conformational changes are triggered by liganded SBP and ATP. MalFGK$_2$ rests in an inward-facing conformation (Oldham et al., 2007). Binding of liganded maltose-binding protein (MBP) to the TMDs and ATP to the NBDs is coupled to a conformational change to the outward-facing conformation; concomitantly, ATP is poised for hydrolysis by dimerisation of the NBDs, and maltose is released from MBP into the maltose binding site of the TMDs.

Hydrolysis of ATP triggers a conformational change to the inward-facing conformation, maltose dissociates from the TMDs and diffuses into the cytoplasm, and ADP dissociates from the NBDs, completing the translocation cycle. Although this alternating access model is consistent with available crystallographic and biochemical data for the model ABC importer $MalFGK_2$, it should be noted that this mechanism is not universal among ABC importers (Bordignon et al., 2010) and alternate models have also been proposed (Jones and George, 2014).

Binding of a liganded SBP to the TMDs is a necessary step in the translocation cycle of an ABC importer; in the absence of liganded SBPs, ABC importers have very low ATPase activity and cannot import solutes (Davidson et al., 2008). SBP-independent mutants of some ABC importers have been isolated; these mutants appear to destabilise the resting state of the importer, lowering the energy barrier for interconversion between the inward-facing and outward-facing conformations in the absence of SBP (Oldham and Chen, 2011b). This observation suggests that in wild-type ABC importers, SBPs are essential for lowering the energy barrier for the conformational change associated with solute translocation. Two plausible explanations have been proposed to rationalise the evolution of SBP-dependence in ABC importers specialised for high-affinity transport of scarce nutrients (Bosdriesz et al., 2015). Firstly, SBPs can be expressed at higher levels than the TMDs due to the limited availability of membrane space, effectively increasing the number of solute binding sites per transporter; for example, in *E. coli* there is a >30-fold molar excess of MBP to $MalFGK_2$. Secondly, SBPs concentrate their ligands close to the cell membrane, increasing the encounter rate between the transporter and its substrate, which is the liganded SBP for SBP-dependent systems.

1.2.2 Structure of SBPs

SBPs contain two α/β domains connected by a flexible hinge (Figure 1.2). Each domain typically consists of a core of five β sheets surrounded by α helices, with the ligand binding site located at the interface of the two domains. In the absence of ligand, SBPs typically adopt an open conformation with a large cavity between the two domains, whereas in the presence of ligand, they adopt a closed conformation with the ligand enclosed at the domain interface. The two conformations are related by a pronounced rigid body rotation about a hinge region, which has been described as a "Venus flytrap" motion (Felder et al., 1999). The conformational dynamics of SBPs are discussed further in Section 5.1.2. SBPs were initially classified into two types based on their β sheet topology (Fukami-Kobayashi et al., 1999): type I SBPs contain five parallel β sheets with the strand order $\beta_2\beta_1\beta_3\beta_4\beta_5$, whereas in type II SBPs, the strand order is $\beta_2\beta_1\beta_3\beta_n\beta_4$, where β_n is the first β strand after the crossover between the two domains and is antiparallel to the remaining strands. A more recent structural classification reorganised the SBP superfamily into six clusters to reflect the diversity of recently acquired structures (Berntsson et al., 2010). In addition to different β sheet topologies, these classes of SBPs have key differences in hinge structure, the presence of additional small α/β domains, and the number of crossovers between the two domains. For example, cluster A SBPs, typified by the vitamin B_{12}-binding protein BtuF, are characterised by four or five parallel β sheets in each domain, a single domain crossover, and a rigid helical hinge, which permits only a small domain rotation (~8°) upon ligand binding (Hvorup et al., 2007).

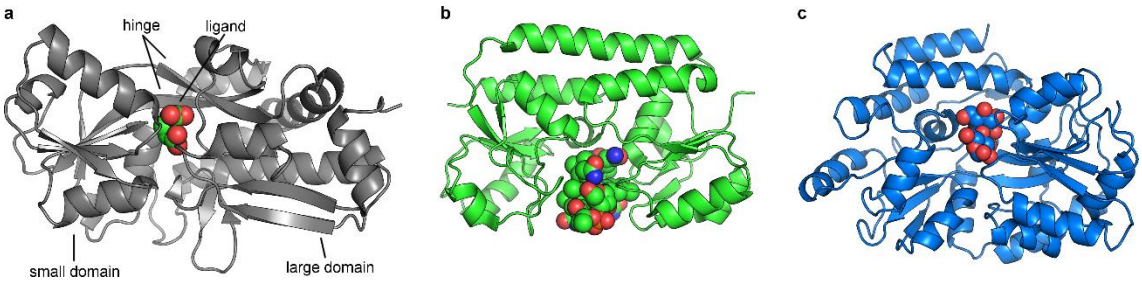


Figure 1.2. Representative structures of solute-binding proteins. (**a**) Aspartate/glutamate-binding protein from *E. coli* (PDB: 2VHA). (**b**) BtuF from *E. coli* (PDB: 1N2Z). (**c**) MBP from *E. coli* (PDB: 1ANF). According to the classification scheme of Fukami-Kobayashi et al., these structures belong to types II, III and I, respectively. According to the classification scheme of Berntsson et al., they belong to clusters F, A, and B, respectively. Ligands are shown as spheres.

1.2.3 Amino acid-binding proteins

The polar amino acid-binding protein (AABP) family is one of several phylogenetically distinct families in the SBP superfamily that have gained the ability to bind amino acids through convergent evolution (Berntsson et al., 2010). Functional diversification of the AABP family has produced proteins specific for most polar and charged proteinogenic amino acids, as well as some non-proteinogenic amino acids such as D-alanine, L-ornithine and L-cystine (Tam and Saier, 1993a). AABP-dependent transporters therefore enable bacteria to scavenge for a variety of physiologically relevant amino acids in nutrient-poor environments for use as carbon/nitrogen sources and building blocks in protein synthesis, reducing the need for energetically expensive amino acid biosynthesis. In eukaryotes, AABP homologues have been co-opted into roles in synaptic signalling and other signal transduction pathways; for example, the extracellular ligand-binding domains of ionotropic glutamate receptors are homologous to bacterial AABPs (Felder et al., 1999).

In bacteria, AABPs and their attendant ABC transporters have a range of specialised physiological roles beyond scavenging and detection of nutrients. For

example, in *E. coli*, the L-cystine-binding protein FliY has a role in the oxidative stress response; L-cysteine is exported to the periplasm as a reducing agent, and the product, L-cystine, is re-imported by a FliY-dependent ABC importer (Ohtsu et al., 2010). In *Rhizobium leguminosarum*, AABPs have an important role in the *Rhizobium*-legume symbiosis through uptake of amino acids produced by the plant in exchange for nitrogen fixation by the bacterium (Lodwig et al., 2006). AABPs can also have an important role in pathogenesis (Härtel et al., 2011; Lewis et al., 2012; Osborne et al., 2012), either indirectly, by satisfying niche in-host nutritional requirements (Müller et al., 2005) or directly, for example, by moonlighting as surface adhesins (Leon-Kempis et al., 2006).

In addition to their physiological importance, AABPs are useful from a protein engineering perspective because they can be used to engineer fluorescent sensors for amino acids (Dwyer and Hellinga, 2004). These sensors rely on the open-closed conformational change that AABPs undergo upon ligand binding, which can be coupled to an observable fluorescence output using an environmentally sensitive dye or a pair of fluorescent proteins that undergo distance-dependent Förster resonance energy transfer (FRET). The importance of amino acids such as L-glutamate, L-glutamine and L-arginine in metabolism and neurological function has driven demand for AABP-based fluorescent sensors, which can be used to monitor fluxes of these amino acids *in vivo* with high spatial and temporal resolution (Gruenwald et al., 2012; Marvin et al., 2013; Okumoto et al., 2005; Whitfield et al., 2015).

AABPs adopt the type II SBP fold described in Section 1.2.2, with two α/β domains connected by two flexible β strands (Figure 1.2a). The binding site can be conceptually divided into a structural motif that binds the amino acid moiety, which is highly conserved throughout the family, and a more variable binding pocket for the side chain of the amino acid ligand. The structural basis for binding specificity and multi-

specificity in AABPs has been addressed in previous studies (Fulyani et al., 2013; Hu et al., 2008; Oh et al., 1994; Stamp et al., 2011). However, the structural basis for promiscuous binding and the evolutionary origin of functional diversity in the AABP family have not yet been addressed explicitly. Furthermore, detailed studies of binding thermodynamics across the AABP family have been attempted only recently (Pulido et al., 2015). These topics are addressed in Chapter 2. Additionally, previous studies of AABPs have focussed on the structural and mutational basis for relatively minor changes in binding specificity; it is therefore unclear how AABPs have been co-opted into cellular processes other than solute transport, including metabolism, signal transduction, and transcriptional regulation. This topic is addressed in the remainder of this thesis, which focusses on the evolution of an enzyme, cyclohexadienyl dehydratase, from an ancestral AABP.

1.3 Methodology

1.3.1 Foreword

Part of Section 1.3.2 is derived from the following publication:

> Clifton, B.E., Whitfield, J.H., Sanchez-Romero, I., Herde, M.K., Henneberger, C., Janovjak, H., Jackson, C.J. (2016) Ancestral protein reconstruction and circular permutation for improving the stability and dynamic range of FRET sensors. *Methods Mol. Biol. 1596*, 71-87.

1.3.2 Molecular phylogenetics and ancestral protein reconstruction

The idea that extinct proteins could be studied through statistical reconstruction of their sequences originated from Pauling and Zuckerkandl (Pauling and Zuckerkandl, 1963). Following developments in statistical phylogenetic analysis, including the development of maximum-parsimony, maximum-likelihood and Bayesian phylogenetic analyses, advances in gene synthesis technology, and the wide availability of sequence data from genome sequencing, this idea has evolved into a practical methodology called ancestral protein reconstruction. Ancestral protein reconstruction is used for three main purposes: (i) elucidation of molecular mechanisms behind protein evolution (see examples discussed in Section 1.1: Eick et al., 2012; Harms and Thornton, 2014; Hudson et al., 2015); (ii) "paleobiochemistry", that is, to investigate the biochemistry, physiology and evolution of extinct organisms (Gaucher et al., 2008; Kratzer et al., 2014; Perez-Jimenez et al., 2011; Shi and Yokoyama, 2003); and (iii) protein engineering, since ancestral proteins or ancestor-like proteins may have useful properties such as improved thermostability or improved catalytic activity on alternative substrates (Alcolombri et al., 2011; Chen et al., 2010; Watanabe et al., 2006; Whitfield et al., 2015).

Resurrection of an ancestral protein is achieved by (i) collection and multiple sequence alignment of a sequence dataset representative of a particular protein family; (ii) inference of a phylogeny describing the evolution of the protein family; (iii)

probabilistic reconstruction of the sequences of ancestral nodes in the phylogeny; (iv) synthesis and cloning of genes encoding the ancestral proteins; and (v) expression, purification and characterisation of the ancestral proteins. The remainder of this section gives an overview of the methodology and theoretical background of molecular phylogenetic analysis and ancestral protein reconstruction, with a particular focus on the maximum-likelihood (ML) method. Broader discussions are given in textbooks (Wiley and Lieberman, 2011; Yang, 2014) and reviews (Merkl and Sterner, 2016; Thornton, 2004; Yang and Rannala, 2012).

Reconstruction of ancestral protein sequences requires a phylogenetic tree – a model that describes the evolutionary relationships between the gene products of interest. In a phylogenetic tree inferred from protein sequences (the focus of the following discussion), each internal node represents the ancestor of a new lineage, created by gene duplication, speciation or horizontal gene transfer. Each external node or "tip" represents an extant protein, and each branch represents the accumulation of amino acid substitutions over time.

Phylogenetic trees can be inferred from a multiple sequence alignment using the ML method. The goal of phylogenetic inference using ML is to identify the tree topology and parameters (for example, branch lengths) that maximise the likelihood function. In this context, likelihood is defined as the probability of observing the data - that is, the probability that the present-day sequences would have evolved - given the tree topology and parameters. This calculation requires an explicit probabilistic model of protein evolution, usually the general time-reversible model, which assumes that (i) sequence evolution is a stochastic Markov process; (ii) individual sites in a protein sequence evolve independently and under identical conditions; and (iii) sequence evolution is time reversible, that is, the rate of substitution (q_{ij}) from amino acid *i* to *j* satisfies the condition

$\pi_i q_{ij} = \pi_j q_{ji}$, where π_i is the equilibrium frequency of amino acid *i*. Parameters in the general time-reversible model are specified by a symmetric substitution matrix that encodes the relative rate of each possible amino acid substitution and the equilibrium frequency of each amino acid. These 208 parameters are empirically determined from analysis of protein sequences.

The limitations of the general time-reversible model become apparent when proteins are considered as functional molecules with three-dimensional structures rather than character strings; however, the general time-reversible model can be modified to address these limitations in some cases. For example, the basic general time-reversible model assumes that each site evolves at the same rate. This is an unrealistic assumption, since the strength of selection and thus the substitution rate at a site varies depending on the functional and structural role of that site in the protein. Among-site rate heterogeneity can be modelled using the discrete gamma model (denoted by +Γ), in which sites are divided equally into a small number of categories evolving at different rates; these rates are drawn from the gamma distribution. This model is specified by a single parameter, α, which controls the shape of the gamma distribution; heterogeneity in evolutionary rates between sites decreases as α tends to infinity. Another common modification to the general time-reversible model is to estimate equilibrium frequencies for each amino acid from the sequence data (denoted by +F), rather than using the equilibrium frequencies specified in the model. Even with these modifications, however, the statistical models used for phylogenetic analysis are simplistic representations of protein evolution.

The likelihood associated with a phylogenetic tree can be calculated given a multiple sequence alignment, a tree topology, a model of sequence evolution and a set of parameters including branch lengths. In practice, the parameters (such as branch lengths) are not known *a priori* and the likelihood is treated as a function of these parameters,

which are optimised to maximise the likelihood associated with a tree topology. The ML tree can be found, theoretically, by maximising the likelihood for each possible tree topology, but this computation is impossible due to the astronomical number of possible topologies. Instead, an initial tree is generated either randomly or using a simple method, such as the algorithmic distance-based *neighbour joining* method. Next, the initial tree is modified using an algorithm such as the *nearest neighbour interchange* algorithm, which swaps two pairs of subtrees separated by a branch, or the *subtree pruning and regrafting* algorithm, which removes a subtree and reattaches it elsewhere on the tree. The alternative trees generated by these algorithms are evaluated by calculating their likelihoods, and the tree with the maximum likelihood is improved through further iterations until convergence upon the ML tree is achieved.

Phylogenetic trees inferred using ML can be validated in several ways. Firstly, if the tree describes the evolution of orthologous proteins in different species, it should be consistent with established species trees, notwithstanding complicating factors such as horizontal gene transfer or incomplete lineage sorting. Secondly, the tree topology should be robust to variations in the statistical evolutionary models used in the calculation; alternative substitution matrices should give the same topology. Thirdly, the level of support for individual clades in a tree can be assessed using the bootstrap method. In this method, columns from the original multiple sequence alignment are randomly resampled to create pseudoreplicate data sets, which are used to repeat the ML analysis; usually, 100 pseudoreplicates are analysed. The bootstrap value associated with a branch in the ML tree is the percentage of pseudoreplicate trees that contain the same grouping of sequences demarcated by that branch; a high bootstrap value indicates that a branch is robust to variations in the sequence dataset and is strongly supported.

Bayesian statistics provide an alternative model-based method for inference of phylogenetic trees. The goal of Bayesian phylogenetic analysis is to identify the tree with the highest posterior probability, which is calculated using Bayes' theorem (Eq. 1.1):

$$\mathrm{P}(T, \boldsymbol{\theta}|D) = \frac{\mathrm{P}(T, \boldsymbol{\theta}) \times \mathrm{P}(D|T, \boldsymbol{\theta})}{\mathrm{P}(D)} \quad \textbf{(1.1)}$$

where P(T,$\boldsymbol{\theta}$|D) is the posterior probability of the tree and associated parameters, P(T,$\boldsymbol{\theta}$) is the prior probability of the tree and associated parameters, P(D|T,$\boldsymbol{\theta}$) is the likelihood of the data, and P(D) is a normalising constant. The posterior probability cannot be calculated directly and is instead determined using Markov chain Monte Carlo simulations, in which the number of times a particular tree topology is sampled is proportional to its posterior probability. Although ML and Bayesian methods use the same statistical models of protein evolution, the underlying statistical frameworks are very different. The two methods have different strengths and weaknesses, and coexist as state-of-the-art methods for phylogenetic inference (Merkl and Sterner, 2016).

The most common method for reconstruction of ancestral protein sequences is the empirical Bayes method (Yang 1995). The likelihood of the set of ancestral sequences associated with a phylogenetic tree is essentially the probability that the observed extant sequences would have evolved from those ancestral sequences. More formally, the empirical Bayes method uses Bayes' theorem to calculate, site-by-site, the conditional probability of each possible set of character states associated with each ancestral node in a phylogeny, given the extant sequence data, a fixed phylogenetic tree, and ML estimates of relevant parameters. The ML reconstruction at a particular site and particular ancestral node is the amino acid that makes the greatest contribution to this conditional probability, and the ML ancestral sequence is reconstructed using the ML reconstruction at every site.

The degree of confidence in an ancestral amino acid reconstruction at a site, given the sequence data and the assumptions inherent in the phylogenetic tree and evolutionary model, can be measured by calculating its posterior probability, which is defined as the fractional contribution of the reconstructed state to the total likelihood over all possible ancestral states. Analysis of the posterior probability distribution of an ancestral sequence invariably reveals the existence of ambiguously reconstructed sites, for which the posterior probability of the ML state is less than 1 and alternative plausible states have non-zero posterior probabilities. Sequence variations at these ambiguously reconstructed sites could affect the phenotype of the protein, including its functional properties; it is therefore essential to determine whether the experimentally observed phenotype of an ancestral protein is robust to statistical uncertainty in its reconstructed sequence.

Several methods have been employed to assess the phenotypic robustness of ancestral proteins to statistical uncertainty in their sequences (see Eick et al., 2016 for a thorough discussion). Firstly, individual variants of the ML ancestral protein with the plausible alternative state at each ambiguously reconstructed site can be generated and characterised, where an arbitrary posterior probability cut-off (usually 0.2) is used to define a "plausible" alternative state. Although this method does not account for the possibility of epistasis between ambiguously reconstructed sites and is impractical if there are many ambiguously reconstructed sites, characterisation of individual variants can be useful if a particular alternative state has obvious structural or functional significance, for example, for ambiguously reconstructed sites within the active site of an enzyme.

Secondly, a single variant of the ML ancestral protein with the second-most likely state at every ambiguously reconstructed site (the "AltAll" protein), representing a worst plausible case scenario, can be generated and characterised. The true ancestral sequence most likely lies between the ML sequence and the AltAll sequence in sequence space,

much closer to the ML sequence than the AltAll sequence. If the ML and AltAll variants have similar phenotypes, it may be assumed that variants in the intervening sequence space, which likely contains the true ancestral sequence, also have similar phenotypes to the ML variant. The AltAll method addresses the limitations associated with characterising the effects of individual sequence variations at ambiguously reconstructed sites (the neglect of epistatic effects and difficulty of characterising many variants) and is a highly conservative indicator of phenotypic robustness, since the most plausible ancestral sequences are much more similar to the ML sequence than the AltAll sequence. However, the conservativeness of the AltAll method could also be a weakness, because a single unlikely and erroneous state in the AltAll sequence could render the corresponding protein artefactually non-functional.

Finally, Bayesian sampling can be used to assess the phenotypic robustness of ancestral proteins to sequence uncertainty. In this case, a large set of alternative ancestral sequences is generated by sampling from the posterior probability distribution at each site, and several of these sequences are chosen randomly and characterised. However, recent work has shown that this strategy, although intuitively appealing, has a major disadvantage: sequences generated by Bayesian sampling generally contain very low-probability ancestral states (posterior probability <0.1) and are therefore frequently and artefactually non-functional, particularly when statistical uncertainty in the sequence reconstruction is high (Eick et al., 2016).

1.3.3 Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) is a method for determination of thermodynamic parameters associated with protein-ligand interactions that relies on the direct measurement of heat. From a single ITC experiment it is possible to determine the enthalpic (ΔH) and entropic (ΔS) components of the Gibbs free energy of binding (ΔG),

which together constitute the thermodynamic signature of a protein-ligand interaction. The association constant (K_a) and stoichiometry (n) of the interaction can also be determined. This section gives a methodological overview of ITC (Chaires, 2008; Freyer and Lewis, 2008; Velazquez-Campoy et al., 2004) and a brief discussion on the interpretation of the thermodynamic signature of a protein-ligand interaction in terms of molecular structure.

Modern ITC instruments operate on the principle of power compensation. Power compensation calorimeters contain a sample cell and a reference cell, which are maintained at a constant temperature through application of constant power to each cell. A protein-ligand titration is performed by incremental injection of the ligand into the sample cell containing the protein. Heat associated with formation of the protein-ligand complex is compensated by a change in the power applied to the sample cell, so that the temperatures of the sample cell and reference cell are equalised. For example, the heat produced by an exothermic protein-ligand interaction is compensated by a decrease in power applied to the sample cell. The power applied to the sample cell over the course of the titration is recorded and integrated to calculate the total heat evolved or absorbed following each injection. The resulting data, called the binding isotherm, can be fitted to a model to extract thermodynamic parameters for the interaction. In the simplest case, where the protein has a single binding site, or multiple identical and non-cooperative binding sites, the data can be fitted to the independent binding sites model to determine K_a, n and ΔH, which can then be used to calculate ΔG and ΔS.

The shape of the binding isotherm resulting from a protein-ligand titration is governed by the value of a parameter c:

$$c = [P_{tot}] \times K_a \times n \qquad \textbf{(1.2)}$$

c is a critical parameter for experimental design, and its effect on the shape of the binding isotherm is shown in Figure 1.3. ITC experiments are most informative when *c* is between ~1 and 1000 (Velazquez-Campoy et al., 2004), producing a sigmoidal curve from which K_a, *n* and ΔH can be determined accurately (Figure 1.3b). As *c* becomes very large (*c* >1000), the binding isotherm becomes insensitive to changes in K_a, and only ΔH and *n* can be determined (Figure 1.3a). This limit on the value of *c*, together with a limit on the minimum protein concentration required to produce an adequate heat signal during the experiment, implies an upper limit on the K_a values (~10^8 – 10^9 M^{-1}) that can be determined by ITC (Freyer and Lewis, 2008).

Understanding the thermodynamic basis for binding promiscuity requires measurement of thermodynamic parameters for low affinity interactions. A persistent misconception is that these parameters cannot be obtained by ITC, since titration of a protein with a small excess of ligand under low *c* conditions (*c* <1) produces a flat and featureless binding isotherm (Figure 1.3c) (Tellinghuisen, 2012). In fact, informative binding isotherms can be obtained provided that the protein is titrated with sufficient ligand to ensure complete formation of the protein-ligand complex, which results in a hyperbolic binding isotherm from which K_a can be determined precisely (Figure 1.3d) (Tellinghuisen, 2008, 2012; Turnbull and Daranas, 2003). The main limitation of this protocol is that ΔH and *n* become correlated at very low *c*. Although ΔH can be estimated by fixing *n* at the known stoichiometry, deviations in *n* from the ideal value due to errors in protein concentration, ligand concentration, effective cell volume, titrant volume, and impure or inactive protein will produce proportional errors in ΔH – errors that are absorbed into *n* when this parameter can be estimated from the data. Thus, high protein purity and homogeneity, ideally confirmed through titration of the sample with a high affinity ligand, proper instrument calibration, and cautious interpretation of the data are

required for accurate determination of thermodynamic parameters for low affinity interactions.

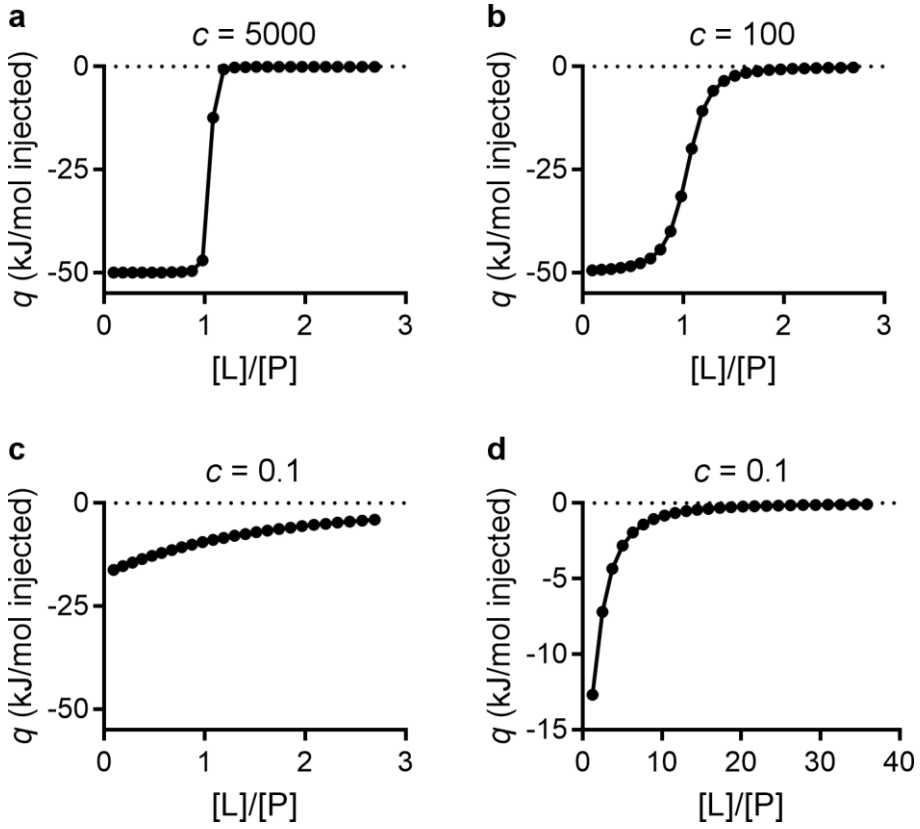


Figure 1.3. Effect of the parameter *c* on the shape of binding isotherms from ITC. The binding isotherms (plots of heat, *q*, as a function of the molar ratio of ligand to protein, [L]/[P], in the cell) represent simulated ITC experiments in which 25 × 2 µL injections of 0.75 mM ligand were made into 0.1 mM protein in a 164 µL cell. To show the effect of *c* on the shape of the binding isotherm, the values of ΔH and *n* for the protein-ligand interaction were kept constant while K_a was varied. (**a**) $K_a = 5 \times 10^7$ M^{-1} (c = 5000). (**b**) $K_a = 1 \times 10^6$ M^{-1} (c = 100). (**c–d**) $K_a = 5 \times 10^3$ M^{-1} (c = 0.1). In (**d**), the concentration of the injected ligand was increased to 10 mM, showing that useful data can be obtained with low *c* values provided that complete formation of the protein-ligand complex is achieved. ITC data were simulated using NanoAnalyze software (TA Instruments).

An alternative method for determination of thermodynamic parameters for low affinity interactions by ITC is the competitive displacement titration (Velazquez-Campoy and Freire, 2006; Zhang and Zhang, 1998). In this method, a low affinity protein-ligand

interaction is coupled to a high affinity protein-ligand interaction by displacement of the low affinity ligand from the protein by the high affinity ligand. The apparent affinity of the protein for the strongly binding ligand ($K_{a,app}$) is reduced depending on the affinity for the weakly binding ligand ($K_{a,w}$; Eq. 1.3)), and the apparent binding enthalpy of the strongly binding ligand (ΔH_{app}) changes depending on the binding enthalpy of the weakly binding ligand (ΔH_w; Eq. 1.4)).

$$K_{a,app} = \frac{K_{a,s}}{1 + K_{a,w}[L]} \quad \textbf{(1.3)}$$

$$\Delta H_{app} = \Delta H_s - \frac{1}{1 + \frac{1}{K_{a,w}[L]}} \Delta H_w \quad \textbf{(1.4)}$$

where $K_{a,s}$ and ΔH_s are the binding affinity and enthalpy for the strongly binding ligand, respectively, and [L] is the concentration of the weakly binding ligand. The main limitation of competitive displacement titrations is the requirement for an appropriate competing ligand, with K_a sufficiently high that a sigmoidal isotherm is obtained, and ΔH sufficiently different from the low affinity ligand that the heat of binding of the high affinity ligand is not completely compensated by the heat of unbinding of the low affinity ligand. Additionally, the thermodynamic parameters for the low affinity ligand are strongly dependent on the thermodynamic parameters measured separately for the high affinity ligand, which can lead to problematic propagation of errors.

The heat observed in an ITC experiment originates from heat associated with the protein-ligand interaction itself in addition to background heat associated with heat of dilution, heat of mixing, and other non-specific physical phenomena such as frictional effects. Background heat can be minimised through careful matching of the titrant and titrate buffers and corrected for using a control titration in which the ligand is injected into buffer. Equilibria linked to the protein-ligand interaction, most importantly

protonation of the ligand, protein or buffer, also contribute to the observed heat. If proton transfer occurs upon binding, the observed heat (ΔH_{obs}) depends on the heat of ionisation of the buffer (ΔH_{ion}) and the number of protons transferred to (or from) the buffer (n_p):

$$\Delta H_{obs} = \Delta H_{int} + n_p \Delta H_{ion} \quad \textbf{(1.5)}$$

Thus comparisons of the thermodynamic signatures of different protein-ligand complexes can be confounded by differences in the number of protons transferred between protein, ligand and buffer. The extrinsic heat associated with buffer protonation can be minimised by using a buffer with low ionisation enthalpy (such as phosphate or acetate), and corrections for buffer protonation can be accomplished by repeating the titration in buffers with different ionisation enthalpies.

One of the main advantages of ITC is that since it relies on the universal phenomenon of heat, it is applicable to a very wide range of protein-ligand complexes without the need for labelling. An important exception is that binding will not be observed if ΔH is very close to zero. However, this problem can often be circumvented by changing the experimental conditions; protein-ligand interactions are usually associated with a large negative change in heat capacity (ΔC_p), such that ΔH becomes more negative as temperature increases. If a protein-ligand interaction is linked with buffer protonation, ΔH can be changed by changing the buffer, according to Equation 1.5. Finally, protein-ligand interactions with ΔH close to zero can be detected through competitive displacement experiments, since a competing ligand with ΔH close to zero will still reduce the apparent affinity for another ligand according to Equation 1.3.

The interpretation of the thermodynamic signature of an interaction in terms of molecular structure has been reviewed thoroughly elsewhere (Baron and McCammon, 2013; Holdgate, 2001; Martin and Clements, 2013; Mobley and Dill, 2009; Olsson et al.,

2008). Briefly, the enthalpic term of the thermodynamic signature reflects the strength of interactions between protein, ligand and solvent; ΔH is made more favourable by a net increase in the strength of hydrogen bonds, electrostatic interactions and other intermolecular interactions in the system in the ligand-bound state compared to the unbound state. The entropic term of the thermodynamic signature reflects the number of configurations available to protein, ligand and solvent; ΔS is made more favourable by an increase in the number of possible configurations of the system upon binding. This can be achieved, for example, by the classical hydrophobic effect – displacement of ordered water molecules from hydrophobic surfaces of the protein or ligand into the disordered bulk solvent, which results in an increase in solvation entropy. The loss of translational and rotational entropy of a ligand associated with conformational restriction in a protein binding pocket is another example of an entropic factor governing protein-ligand interactions.

1.3.4 X-ray crystallography

The goal of protein X-ray crystallography is to generate a model of the three-dimensional structure of a crystalline protein from its X-ray diffraction pattern. Radiation in the X-ray range of the electromagnetic spectrum has a wavelength on the order of Ångstroms (Å = 10^{-10} m) and can therefore be used to resolve the positions of individual atoms in a protein, which are bonded at an average distance of 1.5 Å. Scattering of X-rays by a protein crystal produces a characteristic diffraction pattern. The scattering angles at which diffracted X-rays are observed are indicative of the geometry of the crystal lattice, whereas the intensities of diffracted X-rays contain information about the spatial distribution of electron density within the crystal. The intensities and phases of diffracted X-rays (the latter of which are inferred separately) can be used to calculate the electron density map, which can be combined with prior knowledge of protein composition, geometry, and stereochemistry to produce a complete atomic model of the structure of a protein. The theoretical background associated with protein X-ray crystallography is discussed thoroughly in recent textbooks (Rhodes, 2006; Rupp, 2010).

A protein crystal is an ordered and periodic three-dimensional array of protein molecules, which can be produced by carefully controlled precipitation of a protein from a supersaturated solution using a precipitant such as the polymer polyethylene glycol. The geometry of a crystal can be conceptualised as a crystal lattice, a grid of three infinite sets of equidistant and parallel planes that divide three-dimensional space into unit lattices. A unit lattice and its molecular contents together constitute the unit cell, which is the smallest unit of the crystal that can be used to generate the entire crystal by translation operations. The unit cell may also have internal symmetry, in which case the crystal can be defined by an asymmetric unit and a set of symmetry operators that can be used to generate the entire unit cell. The two types of crystallographic symmetry operators that are compatible with periodic assembly of chiral three-dimensional objects like proteins

are plain rotation axes and screw axes, which are rotation axes along which an object is also translated.

Crystals are divided into seven crystal systems (triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal and hexagonal) based on the relative lengths of the edges of the unit cell (a, b, c) and the angles between the edges of the unit cell (α, β, γ). Different crystal systems have different requirements for the minimum internal symmetry of the unit cell. For example, the unit cell of a tetragonal crystal has two edges equal in length ($a = b$), the angle between each edge is 90° ($\alpha = \beta = \gamma = 90°$), and the minimum internal symmetry is a 4-fold rotation axis parallel to the unit cell vector **c**. Each crystal system is associated with a primitive lattice, except the trigonal crystal system, which shares the hexagonal lattice. In addition to these six primitive lattices, there are eight translationally centred lattices, which have lattice points within the unit cell or on the faces of the unit cell, as well as on its vertices. The primitive and translationally centered lattices together comprise the 14 Bravais lattices in three-dimensional space. Combination of the Bravais lattices with compatible symmetry elements (rotations and screw axes for crystals of asymmetric, chiral objects) gives a total of 65 chiral space groups – that is, there are 65 unique ways to construct translationally periodic assemblies of chiral molecules in three-dimensional space.

X-ray scattering occurs when an X-ray photon (wave packet) induces oscillations of electrons through its oscillating electric field vector. The X-ray photon simultaneously excites all electrons within its coherence length. As a result, the electrons emit coherent partial waves that interfere constructively and destructively to create a resultant scattered wave; constructive interference is maximised by superposition of waves in phase, whereas destructive interference is maximised by superposition of waves out-of-phase

with a phase difference of 180°. The X-ray photon reappears in a certain direction with a probability proportional to the amplitude of the scattered wave in that direction.

Each reflection in an X-ray diffraction pattern originates from X-ray scattering from a discrete set of parallel, equidistant planes in the crystal. These planes can be categorised by their Miller indices *hkl*, which correspond to the number of intersections per unit cell between the planes and the **a**, **b** and **c** unit cell vectors, respectively. For example, the (210) planes intersect the **a** edge of the unit cell twice, intersect the **b** edge of the unit cell once, and are parallel to the **c** edge of the unit cell. Each set of planes has a constant interplanar spacing d_{hkl}; planes with higher Miller indices subdivide the crystal more finely and have lower interplanar spacing. As a result of constructive and destructive interference, diffraction from the (*hkl*) planes is observed if and only if the angle, θ, at which X-rays strike and are reflected from the planes satisfy Bragg's law (Eq. 1.6):

$$n\lambda = 2d_{hkl} \sin\theta \qquad \textbf{(1.6)}$$

where λ is the X-ray wavelength and *n* is an integer.

The relationship between Bragg's law and the positions of reflections in the diffraction pattern becomes evident in reciprocal space, a space spanned by three vectors **a***, **b*** and **c*** with the following mathematical relationship to the real space unit cell vectors **a**, **b** and **c**:

$$\begin{pmatrix} \mathbf{a}\cdot\mathbf{a}^* & \mathbf{b}\cdot\mathbf{a}^* & \mathbf{c}\cdot\mathbf{a}^* \\ \mathbf{a}\cdot\mathbf{b}^* & \mathbf{b}\cdot\mathbf{b}^* & \mathbf{c}\cdot\mathbf{b}^* \\ \mathbf{a}\cdot\mathbf{c}^* & \mathbf{b}\cdot\mathbf{c}^* & \mathbf{c}\cdot\mathbf{c}^* \end{pmatrix} = \mathbf{I} \qquad \textbf{(1.7)}$$

Each set of planes (*hkl*) corresponds to a single vector in reciprocal space, $\mathbf{d}^*_{hkl}$, with magnitude $1/d_{hkl}$, between the reciprocal lattice origin and a reciprocal lattice point *hkl*. The Bragg condition for diffraction from the (*hkl*) planes can be interpreted geometrically as the intersection of reciprocal lattice point *hkl* with the so-called Ewald sphere, a sphere

of radius 1/λ centred on the crystal such that the reciprocal lattice origin is positioned on the Ewald sphere and collinear with the incident X-ray beam and the crystal. When this condition is fulfilled, a diffracted X-ray emerges from the crystal along the vector between the centre of the Ewald sphere and the reciprocal lattice point, producing a discrete reflection. Rotation of the crystal also rotates the reciprocal lattice and causes different reciprocal lattice points to intersect the Ewald sphere, allowing different reflections to be observed.

Bragg's law gives the necessary condition for a reflection to be observed, but provides no information about the relationship between the intensity of the diffracted X-ray and the electron density along the corresponding set of lattice planes. This relationship is provided by the structure factor equation:

$$\mathbf{F}_{hkl} = \int_V \rho(x,y,z) \exp[2\pi \mathrm{i}(hx + ky + lz)]\, dV \qquad \textbf{(1.8)}$$

The structure factor $\mathbf{F}_{hkl}$, a vector that describes the amplitude and phase of the *hkl* reflection, can thus be obtained by integrating the electron density, $\rho(x, y, z)$, found in infinitesimal volume elements over the volume of the unit cell, V. The structure factor equation can also be expressed in terms of the scattering contribution of individual atoms in the unit cell rather than volume elements:

$$\mathbf{F}_{hkl} = \sum_i f_i \exp[2\pi \mathrm{i}(hx + ky + lz)] \qquad \textbf{(1.9)}$$

where f_i represents the partial wave resulting from X-ray diffraction by atom i. Equation 1.8 has the form of a Fourier transform; the sum of $\mathbf{F}_{hkl}$ over all reflections is the Fourier transform of the electron density function. Equivalently, the electron density $\rho(x, y, z)$ can be obtained from the structure factors $\mathbf{F}_{hkl}$ through the following inverse Fourier transformation:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\mathbf{F}_{hkl}\exp[-2\pi\,\mathrm{i}(hx+ky+lz)] \quad \textbf{(1.10)}$$

The magnitude of $\mathbf{F}_{hkl}$ is proportional to the intensity of reflection *hkl*, which is measured in the X-ray diffraction experiment. However, the phase angles cannot be measured directly. Solving this so-called phase problem is one of the major challenges in X-ray crystallography.

If the structure of a homologous protein with high structural similarity (C_α root-mean-square-deviation (RMSD) < 2 Å) is available, phases can be calculated from the known structure and used as initial estimates for the phases associated with a new crystal structure; this method is called molecular replacement. The proper position of the phasing model in the new crystal structure must be identified before its phases can be calculated. The orientation and translational position of the phasing model are optimised successively by trial and error, to maximise agreement between the observed structure factor amplitudes ($|\mathbf{F}_{\mathrm{obs}}|$) and the structure factor amplitudes calculated from the phasing model ($|\mathbf{F}_{\mathrm{calc}}|$). Ideally, the estimated phases from the correctly positioned model can then be used, together with the observed structure factor amplitudes, to generate an interpretable electron density map. The electron density map is used to build an atomic model of the target protein. An improved model can be used to calculate new phases and improve the electron density map. The goal of this iterative process of structure refinement is to maximise agreement between the experimental diffraction data ($|\mathbf{F}_{\mathrm{obs}}|$) and the data expected given the protein model ($|\mathbf{F}_{\mathrm{calc}}|$), that is, to minimise the R-factor defined in Equation 1.11.

$$R = \frac{\sum ||\mathbf{F}_{\mathrm{obs}}| - |\mathbf{F}_{calc}||}{\sum |\mathbf{F}_{\mathrm{obs}}|} \quad \textbf{(1.11)}$$

1.3.5 Molecular dynamics

In molecular dynamics (MD) simulations, the conformational dynamics of proteins are modelled using classical physics (Adcock and McCammon, 2006; Dror et al., 2012; Durrant and McCammon, 2011). Starting with an initial protein structure obtained by crystallography or otherwise, the trajectory of each atom in the protein over time is calculated using Newton's laws of motion and simple, classical approximations to the potential energy of the solvated protein. An MD simulation can be viewed as a crude approximation to the solution of the time-dependent Schrödinger wave equation, which provides a complete description of the quantum mechanical motion of atoms but is computationally unworkable.

The total potential energy, U(**r**), can be considered a function of the position of every atom in the system, **r**. A classical approximation of the potential energy function is called a force field. The force field contains terms that describe bonded forces, which control bond lengths, bond angles and dihedral angles, and non-bonded forces, which describe interactions between atoms, using simple mathematical functions. Bonded atoms are modelled simply as spheres connected by springs, such that deviations from ideal bond lengths are penalised according to Hooke's law. Deviations from ideal bond angles are penalised similarly, and dihedral angles are controlled by sinusoidal functions that favour staggered conformations over eclipsed conformations. Van der Waals interactions are treated by a Lennard-Jones potential, with an attractive term proportional to r^{-6} and a repulsive term proportional to r^{-12}, while electrostatic interactions are treated using Coulomb's law. Each atom is assigned a partial charge, so hydrogen bonds and other dipole interactions can also be modelled using Coulomb's law. Force fields are parameterised to maximise agreement between the force field and experimental data or high-level quantum mechanical calculations; for example, the GROMOS 53A6 force field was optimised by re-parameterisation of earlier force fields to reproduce the free

energies of solvation of amino acid analogues (Oostenbrink et al., 2004). Examples of force field parameters that need to be optimised include partial atomic charges, van der Waals radii, and force constants that determine the stiffness of different bonds.

A force field can be used to calculate the force acting on each atom, **F**(**r**), which is given by the negative derivative of the potential energy function (Eq. 1.12):

$$\boldsymbol{F}(\boldsymbol{r}) = -\frac{dU(\boldsymbol{r})}{d\boldsymbol{r}} \quad \textbf{(1.12)}$$

If the force acting on an atom is known, its motion can be calculated by numerical integration of Newton's second law (Eq. 1.13):

$$\boldsymbol{F}(\boldsymbol{r}) = m\frac{d^2\boldsymbol{r}(t)}{dt} \quad \textbf{(1.13)}$$

Equation 1.13 is numerically integrated by periodically (once every 1–2 fs of simulation time) calculating the force acting on each atom and using the calculated force to update the position and velocity of each atom.

The atomic positions, atomic velocities and single-point energy obtained at each time point during an MD simulation specify a microstate of the system. In principle, macroscopic and observable thermodynamic properties of the system, such as enthalpy or pressure, can be derived from MD simulations using the framework of statistical mechanics, which connects the probability distribution of microstates in an ensemble with the thermodynamic properties of the corresponding macrostate. MD simulations must be designed such that the microstates sampled during the trajectory are consistent with a physiological macrostate. The NPT or NVT ensembles, corresponding to macrostates with a fixed number of particles (N), temperature (T) and volume (V) or pressure (P), are most commonly simulated; this is achieved by modifying atomic positions or velocities

periodically according to some algorithm to maintain constant temperature and/or pressure.

The main advantage of MD simulations for studying the conformational dynamics of proteins is that the position of every atom in a protein can be monitored simultaneously, femtosecond by femtosecond. However, two major trade-offs are necessary to achieve this spatial and temporal resolution, unmatched by any experimental method. One problem is that the accuracy of force fields is limited by their simplicity. For example, modelling of polar interactions using Coulomb's law with constant atomic partial charges requires that the quantum mechanical effect of polarisation is ignored. The second problem is that limitations on simulation time may prevent observation of conformational changes on the μs–ms scale that are slow or infrequent, or prevent sufficient microscopic sampling to determine the macroscopic properties of a protein. However, the computational demands of MD simulations are becoming dramatically less restrictive due to rapid improvements in software and hardware.

The output of an MD simulation, the position of each protein atom as a function of time, has high dimensionality ($3N$, where N is the number of atoms), but many of these coordinates are correlated; for example, the position of an atom will be correlated with the position of nearby atoms, or atoms in the same secondary structure element or domain. Large-scale correlated motions often correspond to functionally important conformational changes, and can be extracted and visualised using principal component analysis (PCA), a technique for reducing the dimensionality of datasets (Grant et al., 2006). The protein structure at each time point is superimposed on a reference structure, and a $3N \times 3N$ covariance matrix is constructed from the covariance in position for each pair of atoms in the structural ensemble. The $3N$ orthogonal eigenvectors of the diagonalised covariance matrix are called principal component axes, and their

eigenvalues reflect the amount of variance in the structural dataset described by the eigenvector. In MD simulations, a principal component axis has a physical interpretation as a specific correlated motion of atoms, that is, a conformational fluctuation. The eigenvalue associated with the principal component axis reflects the magnitude of the conformational fluctuation. Projection of the structural dataset onto several of the largest principal component axes gives a low-dimensional representation of the trajectory that shows the largest conformational fluctuations only. In the case of SBPs, the dominant principal component axes typically represent bending or twisting about the hinge connecting the two α/β domains, corresponding to the open-closed conformational transition (Bucher et al., 2011a; Pang et al., 2005; Silva et al., 2011a).

Chapter Two

Evolution of binding specificity in solute-binding proteins.

2.1 Foreword

Chapter 2 is derived from the following publication:

> Clifton, B.E. and Jackson, C.J. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem. Biol. 23*, 236–245.

2.2 Abstract

The promiscuous functions of proteins are an important reservoir of functional novelty in protein evolution, but the molecular basis for binding promiscuity remains elusive. We used ancestral protein reconstruction to experimentally characterise evolutionary intermediates in the functional expansion of the polar amino acid-binding protein family, which has evolved to bind a variety of amino acids with high affinity and specificity. High-resolution crystal structures of an ancestral arginine-binding protein in complex with L-arginine and L-glutamine show that the promiscuous binding of L-glutamine is enabled by multi-scale conformational plasticity, water-mediated interactions and selection of an alternate conformational sub-state productive for L-glutamine binding. Evolution of specialised glutamine-binding proteins from this ancestral protein was achieved by displacement of water molecules from the protein-ligand interface, reducing the entropic penalty associated with the promiscuous interaction. These results provide a structural and thermodynamic basis for the co-option of a promiscuous interaction in the evolution of binding specificity.

2.3 Introduction

A central goal in the study of molecular evolution is to identify the genetic and structural mechanisms behind the diversification of protein families (Dean and Thornton, 2007; Harms and Thornton, 2013). Proteins often have promiscuous functions that are present at low levels and are not necessarily physiologically relevant or maintained by purifying selection (Khersonsky and Tawfik, 2010). Current models of protein evolution suggest an important role for promiscuity in functional diversification (Conant and Wolfe, 2008; Khersonsky and Tawfik, 2010; Näsvall et al., 2012), since the promiscuous functions of proteins represent a latent source of evolutionary novelty that can be improved readily by adaptive evolution (Aharoni et al., 2005). Recognition of the importance of promiscuity in protein evolution has led to resurgence of a hypothesis proposed by Jensen that primordial proteins, in contrast to their specialised modern counterparts, had broad specificity and were multi-functional (Copley, 2012; Jensen, 1976; O'Brien and Herschlag, 1999). However, evidence for this hypothesis, for example, the observation that promiscuity is widespread in proteins (Patrick et al., 2007), is mostly circumstantial. In addition, there are few examples where the structural basis of a promiscuous activity that has been co-opted during protein evolution has been precisely determined (Eick et al., 2012; Ortlund et al., 2007).

The polar amino acid-binding protein (AABP) family is a valuable model system for investigating the functional diversification of protein families. The AABP family is one of several lineages in the ubiquitous solute-binding protein (SBP) superfamily that are involved in amino acid transport and chemotaxis in bacteria (Berntsson et al., 2010; Tam and Saier, 1993a). AABPs bind extracellular amino acids with high affinity and effect transport or chemotaxis by interacting with ATP-binding cassette (ABC) importers or chemotactic receptors in the cell membrane (Tam and Saier, 1993a). AABPs also have

important roles in pathogenicity (Leon-Kempis et al., 2006; Osborne et al., 2012) and are useful from a protein engineering perspective because they can be used to construct fluorescent sensors for amino acids (Dwyer and Hellinga, 2004). The presence of AABP homologues in eukaryotes, where they are involved in synaptic signalling and other signal transduction pathways, indicates an ancient origin for this protein family (Felder et al., 1999). Gene duplication and divergence over hundreds of millions of years has expanded the binding repertoire of the AABP family to include most polar and charged proteinogenic amino acids as well as some non-proteinogenic amino acids (Tam and Saier, 1993a).

Although extensive structural characterisation of AABPs has improved our understanding of the mechanisms by which they achieve specific binding of amino acids (Berntsson et al., 2010; Fulyani et al., 2013), little is known about the evolutionary origin of their functional diversity. One possibility that is consistent with the current view of protein evolution is that ancestral AABPs were generalists that bound a wide range of amino acids, in contrast to modern, specialised AABPs that generally display specificity towards a small number of amino acids. This would have permitted transport of a variety of amino acids in a primordial organism with fewer genes. The functional diversity of modern AABPs could then be explained by partitioning of the binding activities of a multi-functional ancestor into multiple proteins by subfunctionalisation. An alternate hypothesis is that ancestral AABPs were also specialised for binding particular amino acids and new specificities evolved successively. These hypotheses cannot be tested simply through comparisons of extant AABPs, but can be addressed using ancestral protein reconstruction, a technique that allows the structures and functions of extinct proteins to be characterised experimentally (Thornton, 2004).

We have shown previously that an ancestral AABP resurrected by ancestral protein reconstruction can be used to produce a robust genetically-encoded fluorescent sensor for L-arginine (Whitfield et al., 2015). Here, we expand on our previous phylogenetic and functional analysis to provide evidence that ancestral AABPs were similar in specificity to extant AABPs, although they did exhibit some promiscuous binding. We show that promiscuous binding of L-glutamine by an ancestral arginine-binding protein is enabled by conformational plasticity, selection of an alternate low-energy conformational sub-state, and water-mediated hydrogen bonding networks. We argue that evolution of specialised glutamine-binding proteins was enabled by entropic improvements to this promiscuous binding mode through replacement of water-mediated interactions with polar protein-ligand interactions.

2.4 Results

2.4.1 Reconstruction of ancestral AABPs

To reconstruct ancestors of the AABP family that predated the divergence of modern AABP subfamilies, we collected protein sequences representing a range of AABP subfamilies with specificity for different amino acids. As in our previous phylogenetic analysis (Whitfield et al., 2015), these sequences included bacterial homologues of six widespread AABPs: aspartate-/glutamate-binding protein (DEBP), lysine-/arginine-/ornithine-binding protein (LAOBP), histidine-binding protein (HisBP), glutamine-binding protein (GlnBP), cysteine-binding protein (CysBP) and cystine-binding protein (CyiBP). However, to improve the accuracy of the reconstructed ancestral sequences, we expanded our previous analysis to include a total of 340 sequences from phylogenetically diverse bacteria. The improvements resulting from more comprehensive taxon sampling included increased bootstrap values on major branches of the AABP phylogeny and increased mean posterior probabilities (PPs) of the ancestral sequences.

We used maximum likelihood methods to reconstruct the phylogeny of the AABP family (Figure 2.1; Figure 2.2). The evolutionary relationships between the different subfamilies were generally well-resolved. The main uncertainty was the position of the clade containing homologues of LAOBP; the use of different substitution models and heuristic search algorithms suggested alternate plausible topologies where the LAOBP clade is positioned closer to the DEBP and CysBP clades (Figure 2.2), accounting for several low bootstrap values on the maximum likelihood tree. However, these alternate tree topologies gave rise to very similar ancestral sequences at the nodes of interest (sequence id. >90%).

We focussed on four ancestral nodes, which we called AncQR, AncCE, AncQ and AncE, representing nodes at which various AABP subfamilies diverged (Figure 2.1B),

and reconstructed their sequences by maximum likelihood. Accurately rooting the phylogenetic tree was unfeasible because the closest known relatives of the AABP family – a family of osmolyte-binding proteins (Berntsson et al., 2010) – share very low sequence identity with AABPs (4-15% based on structure guided alignments), making their use as an outgroup impracticable. Hence it is not known *a priori* which of the four ancestral nodes are the most ancient; nonetheless, the positions of the ancestral nodes deep within the AABP phylogeny guarantee that they predated functional divergence of different AABP subfamilies.

The mean PPs for the ancestral proteins characterised in this work ranged from 0.82 to 0.87 (Figure 2.1C). The PPs at positions that are known to be functionally important from crystal structures of extant AABPs are high (Figure 2.1D); most of these positions are reconstructed unambiguously (PP >0.9) and the ambiguities generally represent conservative substitutions. Thus the reconstructed ancestral sequences are likely to provide plausible approximations of the phenotypes of the extinct ancestral AABPs.

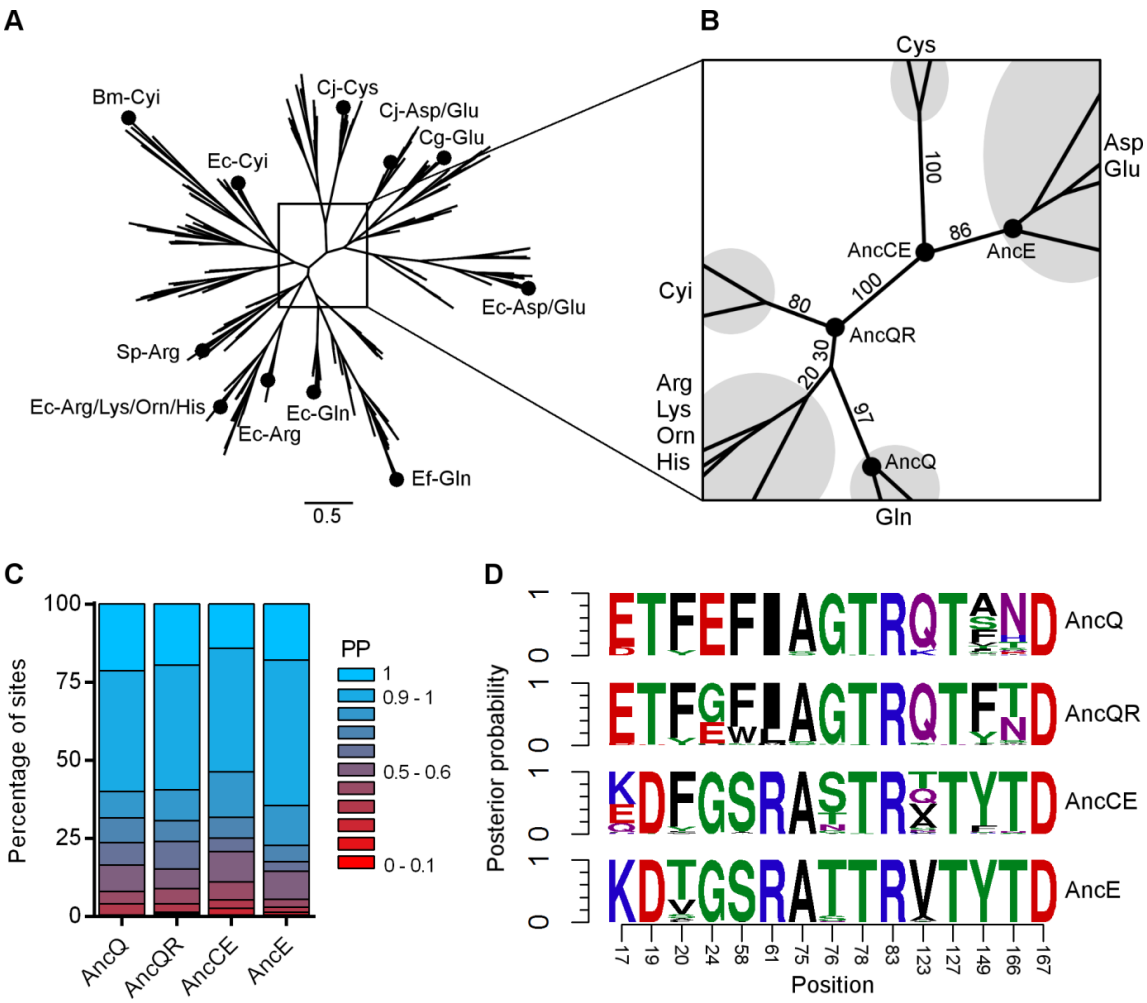


Figure 2.1. Reconstruction of ancestral amino acid-binding proteins. (**A**) Maximum likelihood phylogeny of the AABP family used for ancestral reconstruction. Tips corresponding to representative AABPs that have been characterised previously are annotated with the source organism and the amino acids bound by the protein. The scale bar represents the mean number of substitutions per site. Abbreviations: bm, *Bacillus megaterium*; cg, *Corynebacterium glutamicum*; cj, *Campylobacter jejuni*; ec, *Escherichia coli*; ef, *Enterococcus faecalis*; sp, *Streptococcus pneumoniae*; Cyi, L-cystine; Orn, L-ornithine. (**B**) Expanded view of (A) showing the four ancestral nodes characterised in this work and bootstrap values from 100 replicates on major branches. (**C**) Posterior probability (PP) distributions of the reconstructed ancestral sequences. (**D**) Posterior probability distributions for individual positions near the amino acid binding sites of ancestral AABPs. Residues are numbered according to the equivalent position in AncQR.

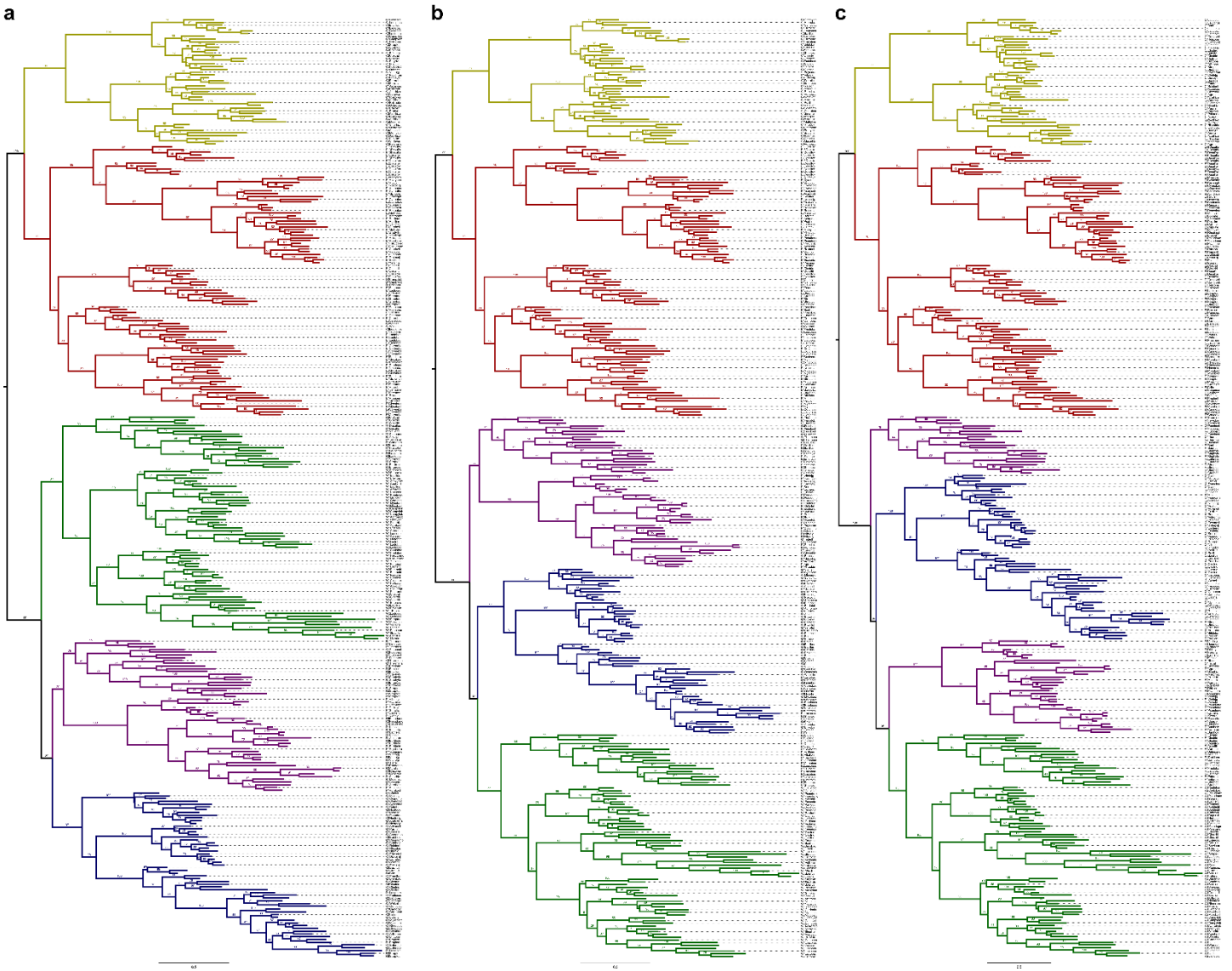

a
b
c

Figure 2.2. Phylogenies of the AABP family (previous page). Sequences are grouped by colour into five clades as in Figure 2.1b: CysBP (yellow), DEBP (red), CyiBP (green), LAOBP (purple), GlnBP (blue). Branches are labelled with bootstrap values from 100 replicates. (**a**) Maximum likelihood tree used for ancestral reconstruction generated using the LG substitution matrix (ln likelihood = -95431.5). (**b**) Alternate tree generated using the WAG substitution matrix (ln likelihood = -95591.5). (**c**) Alternate tree generated using the LG substitution matrix (ln likelihood = -95432.2). A high-resolution version of this figure is available online (http://dx.doi.org/10.1016/j.chembiol.2015.12.010; Figure S1).

2.4.2 Characterisation of ancestral proteins

We cloned synthetic genes encoding the four ancestral AABPs into expression vectors and expressed the corresponding proteins in *Escherichia coli*. The binding specificities of the ancestral proteins were assessed by isothermal titration calorimetry (ITC). Initially, interactions between the ancestral proteins and amino acids were identified *via* qualitative single-injection screening experiments. Binding affinities and thermodynamic parameters for each protein-ligand interaction were then measured in quantitative ITC experiments (Figure 2.3; Table 2.1). Finally, we confirmed the specificity of the ancestral proteins using competitive displacement experiments, in which each ancestral protein was titrated with a binding amino acid in the presence of a cocktail of non-binding amino acids (Table 2.2). This was necessary to ensure that interactions with ΔH close to zero, which would not produce significant heat exchanges in direct ITC experiments, were not overlooked.

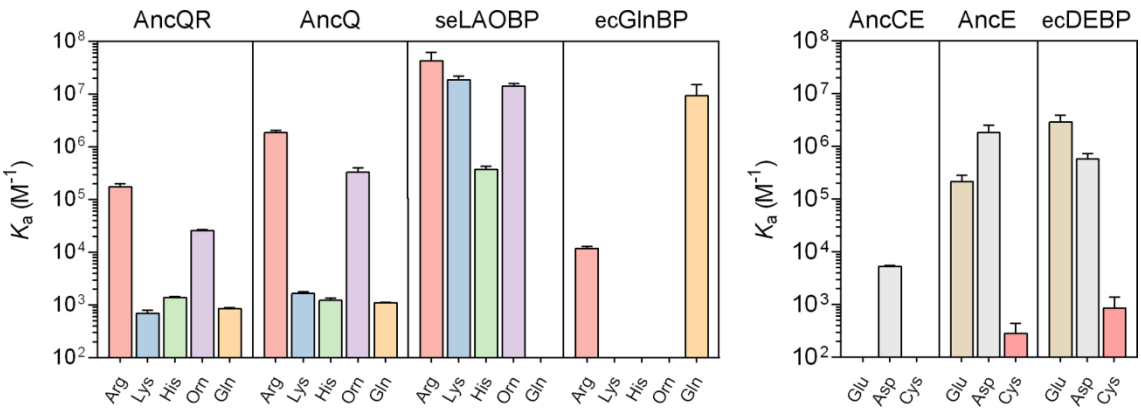


Figure 2.3. Binding specificity of ancestral and extant AABPs. Association constants (K_a) for AABP-amino acid interactions determined by ITC at 25 °C. Data represent the mean ± s.d. from at least three titrations.

Table 2.1. Binding affinities and thermodynamic parameters for amino acid binding to ancestral and extant AABPs. K_a and ΔH_{obs} values represent mean ± s.d. from at least three titrations. K_d, $T\Delta S$ and ΔG values were calculated from the average K_a and ΔH_{obs} values. Literature values for K_d and ΔH are given where applicable. Parameters for the AncE-Cys and DEBP-Cys interactions were determined by competitive displacement experiments as described in Section 2.7.9.

Protein	**Ligand**	**K_a (M^{-1})**	**K_d (μM)**	**ΔH_{obs} (kJ/mol)**	**$T\Delta S$ (kJ/mol)**	**ΔG (kJ/mol)**
AncQR	Arg	$(1.75 \pm 0.28) \times 10^5$	5.73 ± 0.91	-46.8 ± 2.3	-16.8 ± 2.4	-29.9 ± 0.4
	Lys	$(7.00 \pm 0.88) \times 10^2$	1430 ± 180	-20.5 ± 2.3	-4.2 ± 2.3	-16.2 ± 0.3
	His	$(1.39 \pm 0.04) \times 10^3$	717 ± 22	-53.9 ± 1.0	-36.0 ± 1.0	-17.9 ± 0.1
	Orn	$(2.57 \pm 0.14) \times 10^4$	38.9 ± 2.2	-40.3 ± 0.9	-15.1 ± 0.9	-25.2 ± 0.1
	Gln	$(8.57 \pm 0.43) \times 10^2$	1170 ± 60	-68.7 ± 0.3	-51.9 ± 0.4	-16.7 ± 0.1
AncQ	Arg	$(1.86 \pm 0.22) \times 10^6$	0.537 ± 0.062	-41.4 ± 1.4	-5.7 ± 1.4	-35.8 ± 0.3
	Lys	$(1.66 \pm 0.13) \times 10^3$	604 ± 46	-19.5 ± 1.6	-1.2 ± 1.7	-18.4 ± 0.2
	His	$(1.23 \pm 0.11) \times 10^3$	814 ± 75	-53.8 ± 4.1	-36.2 ± 4.1	-17.6 ± 0.2
	Orn	$(3.33 \pm 0.71) \times 10^5$	3.01 ± 0.64	-26.1 ± 0.9	+5.4 ± 1.0	-31.5 ± 0.5
	Gln	$(1.10 \pm 0.03) \times 10^3$	910 ± 26	-43.1 ± 2.6	-25.8 ± 2.6	-17.4 ± 0.1
AncCE	Asp	$(5.33 \pm 0.25) \times 10^3$	188 ± 9	+23.5 ± 0.8	+44.8 ± 0.8	-21.3 ± 0.1
AncE	Glu	$(2.17 \pm 0.69) \times 10^5$	4.61 ± 1.46	+9.2 ± 0.8	+39.7 ± 1.1	-30.5 ± 0.8
	Asp	$(1.82 \pm 0.69) \times 10^6$	0.549 ± 0.207	+7.6 ± 0.3	+43.4 ± 1.0	-35.7 ± 0.9
	Cys	$(2.87 \pm 1.55) \times 10^2$	3480 ± 1880	+34.5 ± 4.9	+48.5 ± 5.0	-14.0 ± 1.3

Protein	Ligand	K_a (M^{-1})	K_d (μM)	ΔH_{obs} (kJ/mol)	$T\Delta S$ (kJ/mol)	ΔG (kJ/mol)
LAOBP	Arg	$(4.31 \pm 1.97) \times 10^7$	0.0232 ± 0.0106 *lit.* 0.014 [1] *lit.* 0.00098 [2]	-45.6 ± 0.8 *lit.* -47.3 [2]	-2.0 ± 1.4	-43.6 ± 1.1
	Lys	$(1.87 \pm 0.33) \times 10^7$	0.0536 ± 0.0096 *lit.* 0.015 [1]	-54.0 ± 1.1	-12.5 ± 1.2	-41.5 ± 0.4
	His	$(3.78 \pm 0.51) \times 10^5$	2.65 ± 0.35 *lit.* 0.5 [1] *lit.* 2.7 [2]	-57.2 ± 1.5 *lit.* -34.3 [2†]	-25.4 ± 1.6	-31.8 ± 0.3
	Orn	$(1.42 \pm 0.19) \times 10^7$	0.0705 ± 0.0094 *lit.* 0.029 [1]	-50.7 ± 0.4	-9.9 ± 0.5	-40.8 ± 0.3
GlnBP	Gln	$(9.44 \pm 5.77) \times 10^6$	0.106 ± 0.065 *lit.* 0.3 [3]	-60.2 ± 3.9	-20.4 ± 4.1	-39.8 ± 1.5
	Arg	$(1.19 \pm 0.11) \times 10^4$	83.9 ± 7.4	-26.7 ± 5.3*	-3.2 ± 5.3*	-23.3 ± 0.2
DEBP	Glu	$(2.91 \pm 1.02) \times 10^6$	0.344 ± 0.120 *lit.* 0.8 [4]	+12.8 ± 0.8	+49.7 ± 1.2	-36.9 ± 0.9
	Asp	$(5.77 \pm 1.62) \times 10^5$	1.73 ± 0.49 *lit.* 1.2 [4]	+22.1 ± 0.4	+55.0 ± 0.8	-32.9 ± 0.7
	Cys	$(8.69 \pm 5.13) \times 10^2$	1150 ± 680	+49.8 ± 6.2	+66.6 ± 6.4	-16.8 ± 1.5

[1]Nikaido and Ames, 1992

[2]Pulido et al., 2015

[3]Weiner and Heppel, 1971

[4]Willis and Furlong, 1976

†corrected for protonation enthalpy

*****estimated assuming $n = 0.74 \pm 0.13$ (mean ± s.d. from two GlnBP + Gln titrations using the same batch of protein). The reported errors in ΔH_{obs} and $T\Delta S$ account for the uncertainty in n.

Table 2.2. Verification of the binding specificity of ancestral and extant AABPs. Apparent association constants ($K_{a,app}$) and apparent binding enthalpies (ΔH_{app}) for AABP-ligand interactions measured by ITC in the presence of potential competing ligands. Each competing ligand was included at a concentration of 1 mM unless otherwise indicated. Titrations highlighted in grey showed a reduction in $K_{a,app}$ and/or change in ΔH_{app} indicative of an interaction between the protein and the competing ligand.

Protein	Titrant	Competing ligands	$K_{a,app}$ (M^{-1})	ΔH_{app} (kJ/mol)
AncQR	Arg	None	$(1.75 \pm 0.28) \times 10^5$	-46.8 ± 2.3
		Ala, Gly, Ser	1.49×10^5	-48.0
		Asp, Glu, Asn	1.67×10^5	-47.0
		D-Ala, D-Ser, Thr	1.83×10^5	-46.7
		Phe, Leu	1.05×10^5	-47.3
		Cys (5 mM)	1.72×10^5	-45.2
AncQ	Arg	None	$(1.86 \pm 0.22) \times 10^6$	-41.4 ± 1.4
		Ala, Gly, Ser	1.76×10^6	-44.1
		Asp, Glu, Asn	1.75×10^6	-45.2
		D-Ala, D-Ser, Thr	1.72×10^6	-44.9
		Phe, Leu	1.64×10^6	-43.7
		Cys (5 mM)	1.60×10^6	-37.0
AncCE	Asp	None	$(5.33 \pm 0.25) \times 10^3$	+23.5 ± 0.8
		Ala, Gly, Ser	5.28×10^3	+22.9
		D-Ala, D-Ser, Thr	5.31×10^3	+22.1
		Asp, Glu, Gln	4.74×10^3	+21.6
		Arg (10 mM)	4.46×10^3	+24.0
		His, Lys, Orn	6.10×10^3	+21.3
		Phe, Leu	5.24×10^3	+22.3
		Cys (5 mM)	4.23×10^3	+25.6
AncE	Glu	None	$(2.17 \pm 0.69) \times 10^5$	+9.2 ± 0.8
		Ala, Gly, Ser	2.25×10^5	+7.4
		D-Ala, D-Ser, Thr	1.94×10^5	+6.1
		Asn, Gln	1.19×10^5	+8.6
		Phe, Leu	2.95×10^5	+9.1
		Arg, Lys, Orn	2.13×10^5	+9.2
		His	2.06×10^5	+8.9
		Cys (5 mM)	1.30×10^5	-8.5
LAOBP	Orn	None	$(1.42 \pm 0.19) \times 10^7$	-50.7 ± 0.4
		Gln	1.53×10^7	-44.1
GlnBP	Gln	None	$(9.44 \pm 5.77) \times 10^6$	-60.2 ± 3.9
		Arg	2.16×10^5	-41.6
		Lys, His, Orn	5.86×10^6	-53.4
DEBP	Glu	None	$(2.91 \pm 1.02) \times 10^6$	+12.8 ± 0.8
		Cys (5 mM)	4.18×10^5	-29.3

These ITC experiments showed that the ancestral proteins are not generalists that bind a significantly expanded range of amino acids, but have specificity comparable to extant AABPs, albeit with significant promiscuous activities in most cases. AncQR and AncQ are primarily arginine-/ornithine-binding proteins that also exhibit promiscuous, low-affinity binding of L-histidine, L-lysine and L-glutamine (Figure 2.3). These ancestral proteins exhibit similar binding profiles to homologues of LAOBP, which bind L-lysine, L-arginine, L-ornithine and L-histidine with K_d in the nanomolar to micromolar range (Nikaido and Ames, 1992); the main difference between these modern and ancestral AABPs is that AncQR and AncQ have a stronger preference for L-arginine and L-ornithine, binding L-histidine, L-lysine and L-glutamine only weakly. AncQ represents the ancestor of the GlnBP subfamily (assuming that this subfamily is monophyletic); thus AABPs specific for L-glutamine likely evolved through co-option of promiscuous L-glutamine binding in an arginine-binding protein. Likewise, AABPs with high affinity for L-lysine and L-histidine appear to have evolved from arginine-binding proteins capable of promiscuous L-lysine and L-histidine binding, similar to AncQR and AncQ. AncCE is a low-affinity aspartate-binding protein, and AncE is a high-affinity aspartate/glutamate-binding protein with low affinity for L-cysteine (Figure 2.3). These proteins are comparable to extant homologues of DEBP (Willis and Furlong, 1976).

We reassessed the specificity of several extant AABPs descended from the ancestral AABPs, namely LAOBP from *Salmonella enterica* (seLAOBP), and DEBP and GlnBP from *E. coli* (ecDEBP and ecGlnBP), to confirm that they did not display the same promiscuous cross-reactivities as the ancestral AABPs (Figure 2.3; Figure 2.4). seLAOBP was not inhibited by 1 mM L-glutamine, suggesting that the promiscuous binding of L-glutamine is unique to the ancestral proteins AncQR and AncQ. In contrast, ecGlnBP displays weak binding of L-arginine (K_d = 84 μM) that competes with L-glutamine binding, which has not been reported previously. Directed evolution experiments have

provided evidence that a protein's ancestral functions are often retained as vestigial promiscuous activities following the evolution of a new function (Tokuriki et al., 2012). Thus the promiscuous binding of L-arginine by ecGlnBP could be interpreted as a vestige of its evolutionary history, retained as the arginine-binding ancestral protein AncQ evolved specificity towards L-glutamine, providing further evidence that the GlnBP subfamily originates from an arginine-binding protein. ecDEBP also exhibits weak binding of L-cysteine (K_d = 1.2 mM) that competes with L-glutamate and L-aspartate binding, similar to AncE. Altogether, our re-evaluation of the specificity of ecGlnBP, seLAOBP and ecDEBP provides further evidence that promiscuous binding of amino acids is not limited to ancestral AABPs.

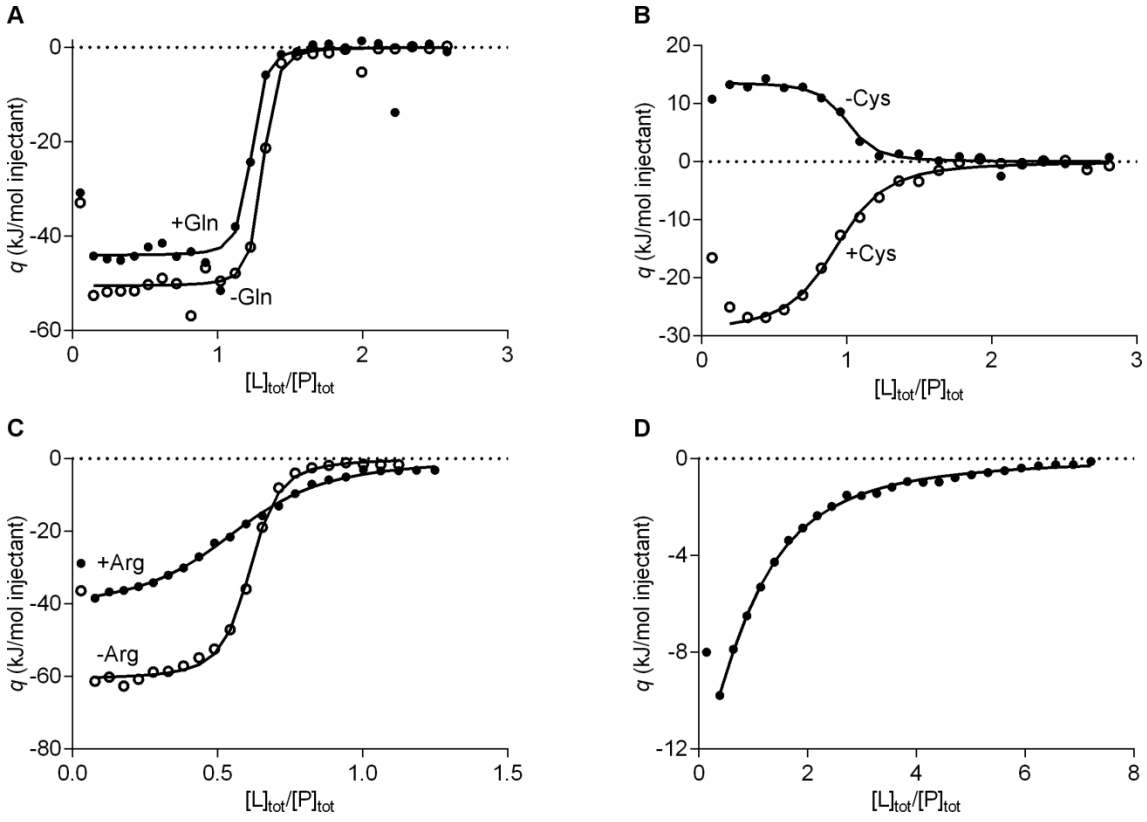


Figure 2.4. Promiscuous binding of amino acids by extant AABPs. (**A**-**C**) ITC data for titrations of extant AABPs with high-affinity ligands in the presence and absence of competing ligands. Competing ligands were included in the cell and syringe solutions at the concentration specified. The heat associated with each injection is plotted as a function of the molar ratio of ligand to protein, and the data were fitted to the independent binding sites model as described in Section 2.7.8. (**A**) seLAOBP titrated with L-ornithine

(± 1 mM L-glutamine). (**B**) ecDEBP titrated with L-glutamate (± 5 mM L-cysteine). (**C**) ecGlnBP titrated with L-glutamine (± 1 mM L-arginine). (**D**) ITC data from direct titration of ecGlnBP with L-arginine.

The affinities of the ancestral AABPs for amino acids are similar to those of extant AABPs (Davidson et al., 2008), except in the case of AncCE. The low affinity of this protein for amino acids could reflect errors in the reconstructed ancestral sequence. Two residues in the binding site of AncCE have plausible alternate reconstructions with PP >0.2 (Figure 2.1D). Accordingly, we also characterised alternate versions of AncCE with the second-best reconstruction at these sites. However, amino acid binding to the resulting variants of AncCE ($AncCE^{K23E}$, $AncCE^{T128Q}$ and $AncCE^{K23E\ T128Q}$) was not detected. The low affinity of AncCE for L-aspartate possibly reflects the fact that this protein is an early intermediate in the evolution of high-affinity aspartate-/glutamate-binding proteins, but the possibility of errors in the reconstructed sequence outside the immediate vicinity of the binding site cannot be excluded.

2.4.3 Structural basis for promiscuous binding in AncQR.

We obtained a high-resolution (1.52 Å) crystal structure of AncQR in complex with L-arginine (K_d 5.7 μM) to identify the structural basis for L-arginine specificity (see Table 2.3 for data collection and refinement statistics). AncQR has the type II periplasmic binding protein fold typical of the AABP family, which consists of two globular α/β domains connected by a flexible two-stranded hinge, with the ligand located at the interface of the two domains. The key residues that bind the guanidinium group of L-arginine are Glu17, Ala75 and Gln123, which form hydrogen bonds or ion-dipole interactions, and Phe20 and Phe58, which form π-stacking interactions (Figure 2.5A). Similar binding geometries are found in other arginine-binding proteins (e.g. Protein Data Bank (PDB) codes 1LAF, 2Q2A, 2Y7I; Figure 2.6), supporting the hypothesis that AncQR is specialised for L-arginine binding.

We then solved the crystal structure of the promiscuous AncQR-glutamine complex (K_d 1.2 mM) at high resolution (1.43 Å; see Table 2.3 for data collection and refinement statistics). The orientation of the ligand is dictated by the stereochemistry of the amino acid binding site and is identical in the AncQR-Arg and AncQR-Gln structures. However, significant conformational changes are required to accommodate binding of the two ligands, given their different sizes and chemical functionalities (Figure 2.5B). The amide nitrogen of the L-glutamine ligand interacts with the carbonyl group of Ala75 and the carboxyl group of Glu17. This requires movement of a binding site loop, relative to the AncQR-Arg structure, to bring Glu17 closer to the smaller ligand. The carbonyl group of the ligand does not interact with the protein directly; instead, two ordered water molecules satisfy its hydrogen bonding potential. Unexpectedly, Gln123 does not form a direct hydrogen bond with L-glutamine, instead rotating away from the ligand and, together with Glu17, participating in a water-mediated hydrogen bonding network.

One unusual feature of the AncQR-Arg structure is the conformational heterogeneity observed in the vicinity of the binding site in Gln123 and the loop extending from Lys145 to Glu151 (Figure 2.5C). Residual electron density in the AncQR-Arg structure matches the alternate conformation of these residues observed in the AncQR-Gln structure that enables promiscuous L-glutamine binding, showing that this conformational sub-state is also sampled in the presence of L-arginine. L-Glutamine preferentially stabilises one of the two conformational sub-states; there is no evidence for conformational heterogeneity in the AncQR-Gln structure. This observation provides direct evidence for selection of an alternate low-energy conformational sub-state as a mechanism for promiscuous binding.

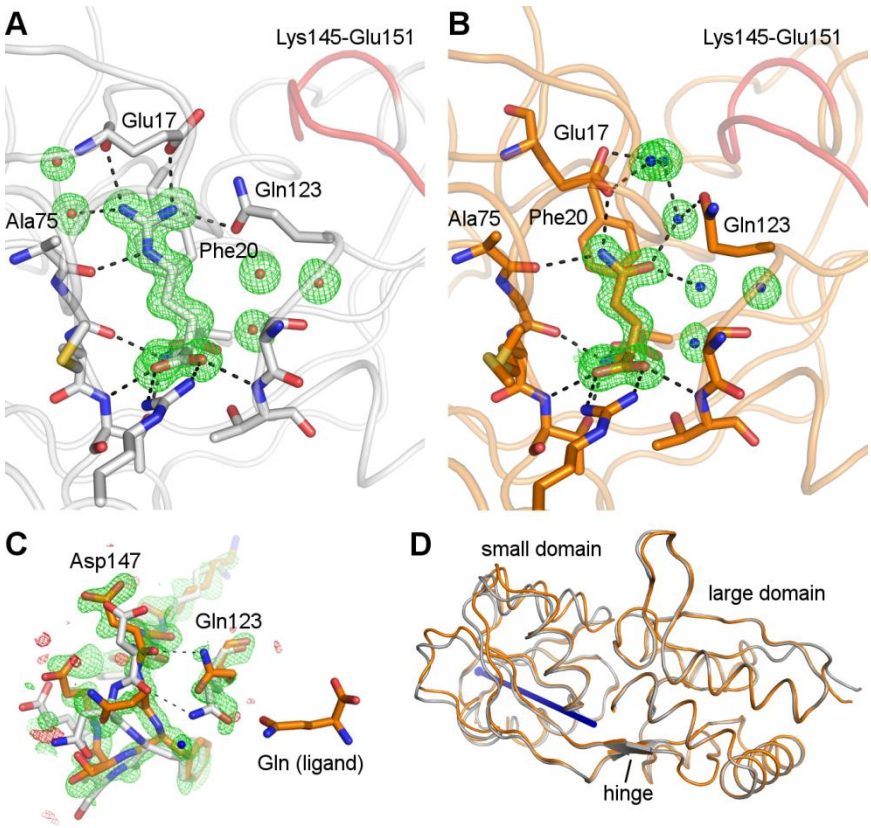


Figure 2.5. Crystal structures of AncQR. (**A-B**) Binding sites of the AncQR-Arg (A) and AncQR-Gln (B) complexes. Electron density for the ligands and ordered water molecules is shown by omit F_o-F_c maps contoured at +3σ. Phe58 is positioned on top of the ligands and is omitted for clarity. The flexible loop Lys145-Glu151 is shown in red. (**C**) Existence of multiple low-energy conformational sub-states in the binding site of the AncQR-Arg complex. Residues 123 and 145-151 were modelled in the conformation unique to the AncQR-Arg structure (grey) with occupancy of 0.5. The F_o-F_c electron density map resulting from refinement of this model is contoured at ±3σ and matches the alternate conformation observed in the AncQR-Gln structure (orange). (**D**) Global conformational differences between the AncQR-Arg (grey) and AncQR-Gln (orange) structures. Backbone atoms of the large domain of each protein (residues 7-95 and 193-232) were superimposed, revealing a rigid body displacement of the small domain, which corresponds to a 5.3° rotation of the small domain about the axis shown by the blue arrow. The two hinge strands (residues 94-97 and 190-195) connecting the two domains are shown in cartoon representation. See Section 2.7.11 for further details.

Table 2.3. Data collection and refinement statistics for AncQR structures.

Structure	**AncQR/L-arginine**	**AncQR/L-glutamine**
PDB code	4ZV1	4ZV2
Data collection		
Wavelength (Å)	0.9537	0.9655
Space group	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions		
a, *b*, *c* (Å)	41.46, 60.41, 103.99	36.20, 61.30, 104.57
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Resolution range (Å)	28.56 – 1.52 (1.55 – 1.52)*	36.20 – 1.43 (1.45 – 1.43)
R_{merge} (%)	9.3 (61.2)	17.8 (187.1)
$CC_{1/2}$ (%)†	99.8 (67.4)	99.7 (52.8)
$I / \sigma(I)$	12.0 (2.4)	9.2 (1.5)
Completeness (%)	99.6 (94.4)	99.7 (96.7)
Multiplicity	6.9 (4.7)	12.7 (9.7)
Refinement		
Resolution range (Å)	28.56 – 1.52	34.21 – 1.43
Number of reflections	38763	41372
R_{work}/R_{free} (%)	17.00/20.05	18.17/20.49
No. of atoms		
Protein	1853	1764
Ligand	12	10
Water	308	266
Average B factors (Å^2)		
Protein	13.74	14.69
Ligand	6.31	8.91
Water	25.26	22.85
R.m.s. deviations		
Bond lengths (Å)	0.0239	0.0248
Bond angles (°)	2.318	2.230

*Values in parentheses refer to highest resolution shell.

†Karplus and Diederichs, 2012

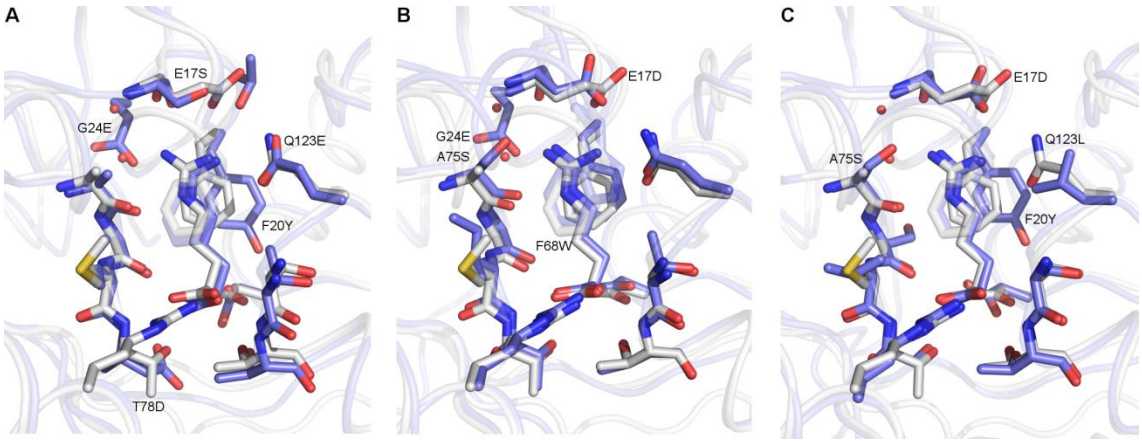


Figure 2.6. Structural similarity between the AncQR-Arg complex and extant L-arginine-binding proteins. (**A**) STM4351, an L-arginine-specific binding protein from *Salmonella enterica* (PDB: 2Y7I). (**B**) ArtJ from *Geobacillus stearothermophilus* (PDB: 2Q2A). (**C**) LAOBP from *S. enterica* (PDB: 1LAF). AncQR is shown in grey and extant proteins are shown in purple. Significant sequence differences between AncQR and extant AABPs are indicated.

A subtle difference in tertiary structure is also observed between the AncQR-Arg and AncQR-Gln complexes (backbone RMSD 0.65 Å overall). This structural difference reflects a 5.3° rigid body rotation of one domain of AncQR about an axis ~45° from the hinge axis (Figure 2.5D), as shown by the decrease in RMSD when the two structures are aligned by superimposition of each domain separately (backbone RMSD 0.41 Å for the small domain and 0.49 Å for the large domain). Notably, such differences are not observed for other AABPs crystallised with different ligands (e.g. LAOBP (Silva et al., 2011a)), but are observed for homologous ligand-binding domains of eukaryotic ionotropic glutamate receptors (Jin et al., 2003). Studies of these AABP homologues have shown that the energetic landscape governing the open-closed conformational transition is strongly ligand-dependent (Lau and Roux, 2007, 2011); flexibility about the hinge connecting the two domains evidently allows AABPs to close to different degrees to optimise interactions with different ligands.

2.4.4 Binding energetics of AABPs.

ITC can provide insight into the mechanisms of molecular recognition because it can be used to deconvolute the free energy of binding (ΔG) into enthalpic (ΔH) and entropic (ΔS) components. The ancestral proteins examined here have variable thermodynamic signatures: binding of cationic and neutral amino acids to AncQR and AncQ is exothermic and enthalpy-driven, whereas binding of anionic amino acids to AncCE and AncE is endothermic and entropy-driven (Figure 2.7A). These contrasting thermodynamic signatures of AABP-amino acid interactions are largely shared by extant AABPs; ecGlnBP and seLAOBP exhibit enthalpy-driven binding to each of their ligands, whereas ecDEBP exhibits entropy-driven binding to each of its ligands. Since ecDEBP binds L-glutamate in a comparatively compact, solvent-occluded and highly charged pocket (Hu et al., 2008), the distinct thermodynamic signatures observed for the anionic AABPs may reflect favourable entropy changes associated with displacement of rigid water networks balanced against unfavourable enthalpies of desolvation. Overall, the remarkable range of ΔH and TΔS for the interactions studied here (>100 kJ/mol, compared with a range of ~30 kJ/mol for ΔG) reflects that different thermodynamic strategies are available for interactions between AABPs and amino acids, as noted previously (Pulido et al., 2015), despite the fact that the key structural motif for binding the amino acid moiety is conserved throughout the family.

The promiscuous binding of L-glutamine by AncQR is enabled by conformational flexibility on three scales: alternate side chain conformations, alternate backbone conformations in loops near the binding site, and rigid body domain movements. One consequence of the adventitious binding mode of L-glutamine through adaption of these alternate protein conformations is that the binding site is not pre-organised for L-glutamine binding. Yet in the AncQR-Gln structure there is no evidence of conformational heterogeneity, suggesting that just one conformational sub-state of the protein is

productive for L-glutamine binding. Accordingly, since several rotors (in Gln123 and Glu17) are restricted in a binding site that is not pre-organised, and since a network of ordered water molecules is restricted at the binding interface, it might be expected that the interaction between AncQR and L-glutamine is not entropically favourable.

Indeed, the calorimetric data shows that binding of L-glutamine to AncQR is associated with a favourable enthalpy term and a highly unfavourable entropy term (Figure 2.7B). The higher affinity of AncQR for L-arginine compared to L-glutamine can be attributed to a smaller entropic penalty associated with L-arginine binding. Likewise, ecGlnBP retains an enthalpic mode of binding towards L-glutamine, but the entropic penalty is far lower than for the interaction between AncQR and L-glutamine, which results in $>10^4$-fold higher binding affinity. A likely explanation is that evolutionary optimisation of the interaction between AncQR and L-glutamine has occurred primarily through optimisation of binding entropy, as ordered water molecules have been displaced by new protein side chains that interact directly with the ligand, *via* the mutations Gln123Lys and Thr166His (Figure 2.7C). ITC experiments in various buffers with different ionisation enthalpies confirmed that the apparent differences in thermodynamic signature for the AncQR-Arg, AncQR-Gln and ecGlnBP-Gln interactions were not obscured by protonation events (Figure 2.8).

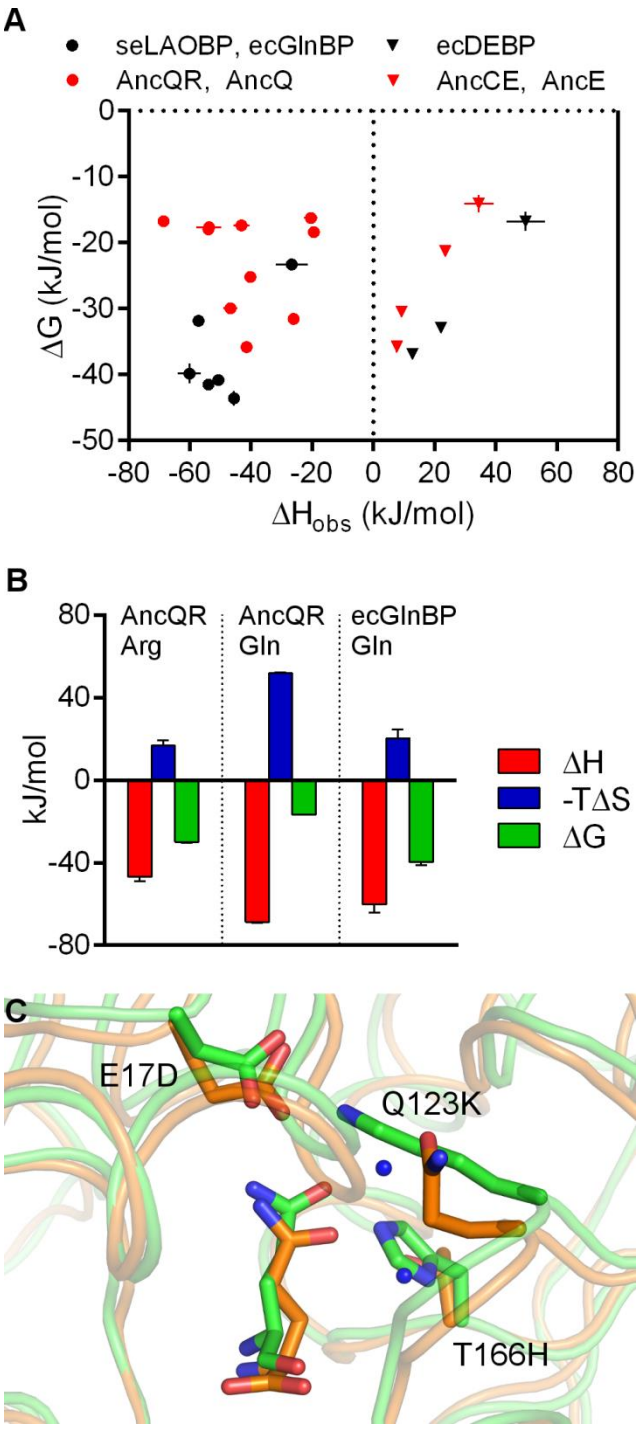


Figure 2.7. Contrasting enthalpic and entropic modes of binding in the AABP family. Thermodynamic parameters for AABP-amino acid interactions were determined by ITC at 25 °C. ΔH_{obs} values represent the mean ± s.d. from at least three titrations. $T\Delta S$ and ΔG values were calculated from mean ΔH_{obs} and K_a values, and errors in these quantities were propagated. (**A**) Distribution of ΔH_{obs} and ΔG values for AABP-amino acid interactions showing enthalpic binding for ligands of AncQR, AncQ, seLAOBP and ecGlnBP, and entropic binding for ligands of AncCE, AncE and ecDEBP. (**B**) Thermodynamic signatures for interactions between AncQR and L-arginine, AncQR and

L-glutamine, and ecGlnBP and L-glutamine. (**C**) Comparison of AncQR-Gln (orange) and ecGlnBP-Gln (green; PDB: 1WDN) complexes shows that water molecules (blue) in the AncQR-Gln complex are displaced through mutations to binding site residues.

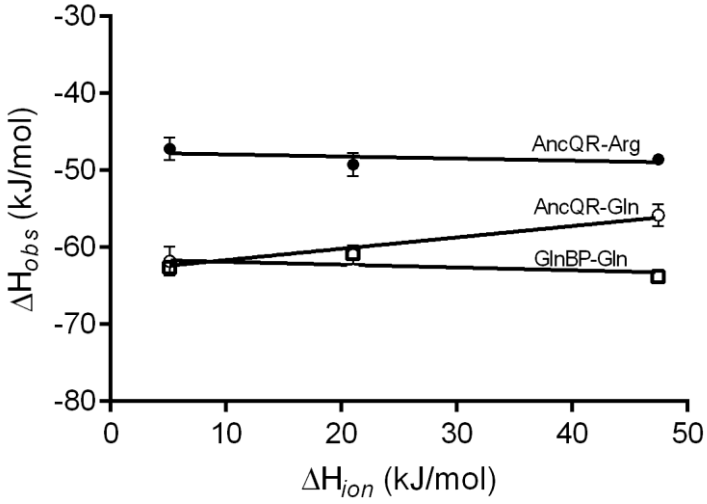


Figure 2.8. Binding enthalpies for the AncQR-Arg, AncQR-Gln and GlnBP-Gln interactions are independent of buffer ionisation enthalpy. Binding enthalpies (ΔH_{obs}) were determined by ITC experiments in three buffers: K_2HPO_4 (ΔH_{ion} = 5.12 kJ/mol; Fukuda and Takahashi, 1998), HEPES (ΔH_{ion} = 21.01 kJ/mol; Fukuda and Takahashi, 1998) and Tris (ΔH_{ion} = 47.45 kJ/mol; Goldberg et al., 2002). Error bars represent 95% confidence intervals estimated from individual titrations in NanoAnalyze software (TA Instruments).

2.5 Discussion

Since the promiscuous functions of proteins are a key source of phenotypic novelty, understanding the structural and thermodynamic basis for promiscuous protein-ligand interactions is vital. In this work we have used ancestral protein reconstruction to evaluate the role of promiscuity in the functional expansion of the AABP family. We identified two ancestral arginine-binding proteins that exhibit promiscuous binding of L-glutamine (AncQR and AncQ), which could account for the evolution of glutamine-binding proteins from arginine-binding proteins. Structural characterisation of AncQR in complex with L-arginine and L-glutamine revealed that the promiscuous interaction with L-glutamine was mediated by conformational plasticity, selection of an alternate low-energy conformational sub-state, and water-mediated hydrogen bonding networks, features that are enthalpically favourable but incur a severe entropic penalty.

The four ancestral AABPs examined in this work exhibit specificities for amino acids that are comparable to modern AABPs; we did not find evidence for subfunctionalisation of a proficient generalist ancestral protein in the functional expansion of the AABP family. The limited functional diversity of the ancestral proteins could be a result of obligatory trade-offs between multi-specificity and high affinity, but this is unlikely since SBPs (including AABPs) that have high affinity for multiple chemically diverse amino acids (Walshaw and Poole, 1996) and oligopeptides (Guyer et al., 1986) are known. Although we cannot rule out the possibility that the ancestor of the entire AABP family was a generalist, since it is not possible to reconstruct this protein using currently available sequence data, our results suggest that specialisation of the AABP family occurred early in its evolutionary history. Specificity in ancestral SBP-dependent amino acid transport systems may have been favoured by differences in the metabolic cost or availability of different amino acids. In general, the evidence that the

ancient progenitors of modern proteins had a larger range of physiologically relevant functions compared to their descendants remains limited (Perez-Jimenez et al., 2011; Risso et al., 2013).

The crystal structure of AncQR complexed with L-glutamine provides further evidence for the importance of multi-scale conformational flexibility and water-mediated interaction networks in mediating promiscuous protein-ligand interactions. The importance of conformational flexibility can be understood in light of the conformational selection model of molecular recognition, whereby proteins exist in ensembles of energetically accessible conformational states that can be preferentially bound and stabilised by different ligands (Boehr et al., 2009; Ma et al., 2002). The implication of conformational plasticity for promiscuity is that a protein with multiple accessible conformational states has more opportunities for interactions with different ligands (James and Tawfik, 2003b; Tokuriki and Tawfik, 2009a). This can be seen directly in AncQR, where L-glutamine preferentially stabilises one of two conformational sub-states accessible in the presence of L-arginine. The alternate rotameric states of the binding site residues Gln123 and Glu17, movement of binding site loops, and rigid body domain motions observed in the arginine- and glutamine-bound structures of AncQR illustrate how various degrees of freedom in conformational space can be exploited. The flexibility of water in satisfying the hydrogen bonding requirements of promiscuously bound ligands is well-known (Ladbury, 1996). Water-mediated interaction networks are also a key feature of high-affinity multi-specific SBPs such as LAOBP (Oh et al., 1994) and the oligopeptide-binding protein OppA (Tame et al., 1996).

The thermodynamic factors that distinguish adaptive and promiscuous protein-ligand interactions are not well understood (Nobeli et al., 2009). Binding of promiscuous substrates in enzymes is thought to be driven frequently by non-specific hydrophobic

interactions (Babtie et al., 2010; Khersonsky and Tawfik, 2010), presumably favoured by solvation entropy, and an example of a promiscuous protein-protein interaction governed by favourable configurational entropy has also been described (Chang et al., 2008). In contrast, cross-reactive antibodies have been shown to bind chemically diverse antigens using different combinations of hydrogen bonds with fortuitously placed protein residues (James and Tawfik, 2003a) – this type of interaction is presumably enthalpic. In this work, we have interpreted the thermodynamic signature of the enthalpy-driven promiscuous interaction between AncQR and L-glutamine as a consequence of the adventitious satisfaction of hydrogen bonding potential of the ligand by restricted water molecules and protein side chains. However, even in AncQR itself, alternate modes of promiscuous binding may be possible, as suggested by the large difference in ΔH for the interactions with L-glutamine and L-lysine, which bind with approximately equal affinity ($\Delta\Delta H$ ~40 kJ/mol and $\Delta\Delta G$ <1 kJ/mol; Table 2.1). Calorimetric measurement of ΔH for low-affinity interactions is often avoided because errors in protein concentration and deviations from 1:1 stoichiometry result in proportional errors in ΔH (Tellinghuisen, 2008); however, in the case of AncQR, we could confirm the accuracy of the active protein concentration *via* titrations with L-arginine.

With some exceptions (Thorpe and Brooks, 2007), there are still few examples where the thermodynamics of protein-ligand interactions have been linked explicitly to the increases in binding affinity gained through evolution. A future challenge will be to correlate thermodynamic and structural changes in a protein-ligand complex along an evolutionary trajectory defined by directed evolution or ancestral protein reconstruction, which could shed light on how nature overcomes enthalpy-entropy compensation in the evolution of high-affinity binding proteins. Furthermore, functional evolution of the AABP family has produced enzymes in addition to new binding specificities (Tam and Saier, 1993b). Reconstruction of the ancestral AABPs and recapitulation of evolutionary

trajectories towards catalytic activity could provide insight into how the various constraints associated with the production of enzymes from non-catalytic scaffolds might be overcome by natural selection or protein design.

2.6 Significance

Although the current literature provides many examples for the evolution of new protein functions from pre-existing promiscuous functions, the structural and biophysical bases for these promiscuous functions and their improvement by adaptive evolution are often unclear. In this work, we addressed this issue by characterising intermediates in the evolution of a family of solute-binding proteins using ancestral protein reconstruction. High-resolution crystal structures provide evidence for an attractive but unproven hypothesis regarding the role of conformational diversity in protein evolution – that promiscuity can arise from selection of pre-existing conformational sub-states by alternate substrates or ligands – and show how conformational flexibility on multiple scales can influence promiscuous binding. Our calorimetric data provides insight into the relationship between the thermodynamic driving forces for promiscuous and adaptive protein-ligand interactions, showing that an adventitious binding mode enabled by recruitment of alternate protein conformations, together with the fixation of ordered water molecules at the binding interface, results in a promiscuous interaction that is enthalpically favourable but entropically unfavourable. Altogether, our structural and thermodynamic data provide a detailed view of a promiscuous protein-ligand interaction that preceded the evolution of a binding protein with high affinity and specificity.

2.7 Experimental Procedures

2.7.1 Phylogenetic analysis and ancestral reconstruction.

340 phylogenetically diverse bacterial homologues of GlnBP (UniProt: P0AEQ3), DEBP (UniProt: P37902), HisBP (UniProt: P0AEU0), LAOBP (UniProt: P09551), and CyiBP (UniProt: P0AEM9) from *E. coli*, and CysBP from *Campylobacter jejuni* (UniProt: Q0P9S0) were selected from the NCBI database of reference proteins using the BLAST server. Alignments were constructed by dividing sequences into subfamilies, aligning sequences within their subfamilies using MUSCLE (Edgar, 2004), editing the alignments to remove large insertions and N-terminal signal sequences, and combining the edited alignments by profile-profile alignment in MUSCLE. Unrooted phylogenetic trees were computed using the maximum likelihood method implemented in PhyML (Guindon et al., 2010) and bootstrapped with 100 replicates. Substitution models and other parameters for phylogenetic analysis were chosen using ProtTest (Abascal et al., 2005; Guindon et al., 2010); model selection was done on the basis of the Akaike information criterion. The tree used for ancestral reconstruction was computed using the LG model of sequence evolution, and rate heterogeneity was modelled using the discrete-gamma model with four rate categories. The fraction of invariant sites and equilibrium amino acid frequencies were estimated from the data. Heuristic tree searches conducted by optimizing a BIONJ distance tree with nearest neighbour interchange and subtree pruning and regrafting moves became trapped in a local minimum (Figure 2.2C). Randomisation of the starting tree resulted in convergence upon a tree with a different topology and improved likelihood (Figure 2.2A). The robustness of this maximum likelihood tree was assessed by reconstructing trees with alternate substitution models (WAG and JTT), starting from BIONJ and randomised trees. This resulted in identification of a third plausible tree topology (Figure 2.2B). Ancestral protein sequences were reconstructed using the

empirical Bayes method implemented in PAML (Yang, 2007). The posterior probability distribution at each site for each ancestral node was also calculated using PAML.

2.7.2 Cloning

The ancestral protein sequences were back-translated and codon-optimised for expression in *E. coli*. The protein sequences of DEBP from *E. coli* (UniProt: P37902, residues 28-302), GlnBP from *E. coli* (UniProt: P0AEQ3, residues 23-248), and LAOBP from *S. enterica* (UniProt: P02911, residues 23-260), each with signal peptide removed, were likewise back-translated and codon-optimised. The resulting genes were synthesised (GeneArt) and cloned into the *Nde*I/*Eco*RI site of the pETMCSIII plasmid (Neylon et al., 2000) for expression of the proteins with N-terminal hexahistidine tags. Cloning was done using standard restriction-ligation methods. Correct ligation of the genes into the vector was confirmed by sequencing, which was done at the Biomolecular Resource Facility at the Australian National University.

2.7.3 Mutagenesis

Site-directed mutagenesis of AncCE was done using Gibson assembly (Gibson et al., 2009). Briefly, the desired mutations were encoded in complementary primers ~30 bp in length, with the mutation in the middle of the primer. Using these primers, overlapping fragments of the AncCE gene containing the desired mutations were amplified by PCR. These fragments were assembled together with the linearised pETMCSIII vector by Gibson assembly. Plasmid DNA was isolated from *E. coli* Top10 cells transformed with the assembly reaction mixture, and successful mutagenesis was confirmed by sequencing.

2.7.4 Protein expression

AncQR, AncE and extant AABPs were expressed in *E. coli* BL21(DE3) cells, grown in auto-induction media (per L: 20 g tryptone, 5 g yeast extract, 5 g NaCl, 6 g Na_2HPO_4, 3 g KH_2PO_4, 6 mL glycerol, 2 g lactose, 0.5 g glucose, 100 mg ampicillin) at 37 °C for 24

h. AncCE and AncQ were expressed in BL21-AI cells, grown in Terrific Broth (TB) supplemented with 100 mg/L ampicillin at 37 °C to OD_{600} 0.8, then induced with 1 mM IPTG and 0.2% (w/v) arabinose for 3 h. Cells were harvested and stored at -80 °C prior to protein purification.

2.7.5 Protein purification

For ITC experiments, proteins were purified by Ni-NTA affinity chromatography with on-column refolding, followed by size exclusion chromatography (SEC). This ensured complete removal of endogenously bound ligands and high protein purity. Cells were thawed, resuspended in binding buffer (20 mM NaH_2PO_4, 500 mM NaCl, 20 mM imidazole, pH 7.4) and lysed by sonication. The cell lysate was fractionated by ultracentrifugation (24200*g*, 1 h, 4 °C), and the soluble fraction was filtered and loaded onto a 5 mL HisTrap HP column (GE Healthcare) equilibrated with binding buffer. The column was washed with 25 mL binding buffer followed by 25 mL unfolding buffer (8 M urea, 20 mM NaH_2PO_4, 500 mM NaCl, pH 7.4). Proteins were refolded on-column by application of a gradient from unfolding buffer to binding buffer over 75 min at 2.5 mL/min using an ÄKTA Purifier (GE Healthcare), then eluted in elution buffer (20 mM NaH_2PO_4, 500 mM NaCl, 500 mM imidazole, pH 7.4). In order to remove misfolded aggregates, refolded proteins were purified by SEC on a HiLoad 26/600 Superdex 200 column (GE Healthcare), eluting in phosphate buffer (20 mM Na_2HPO_4, 100 mM NaCl, pH 7.40). Protein purity was verified by SDS-PAGE. For crystallisation, AncQR was purified by Ni-NTA affinity chromatography without refolding, concentrated, saturated with excess L-arginine or L-glutamine, and purified by SEC, eluting in Tris buffer (20 mM Tris, 100 mM NaCl, pH 8.40) additionally containing 0.1 mM L-arginine or 5 mM L-glutamine. The purified protein was concentrated to 10 mg/mL prior to crystallisation.

2.7.6 Isothermal titration calorimetry

ITC experiments were generally performed using a Nano ITC low volume calorimeter (TA Instruments); some of the initial screening experiments were done using a VP-ITC microcalorimeter (GE Healthcare). The calorimetric constant of the Nano ITC instrument was determined by titrating excess Tris with 20 × 2.5 µL injections of 1.00 mM HCl, with both solutions made up in water thoroughly degassed by boiling (calibration factor = -0.978 ± 0.007, mean ± s.d. from three titrations). The syringe volume was calibrated by mass of water ejected (49.8 mg per 50 µL, calibration factor = 1.00), and the active cell volume was calibrated by the manufacturer (164 µL). ITC experiments were done at 25 °C with stirring at 150-250 rpm. Samples were generally prepared in phosphate buffer (20 mM Na_2HPO_4, 100 mM NaCl, pH 7.4), except where high concentrations (≥5 mM) of acidic or basic ligands were used in the titration, in which case the concentration of Na_2HPO_4 was increased to 50 mM. Buffers used to make up amino acid solutions were matched precisely to the protein buffer by dialysis or desalting. Amino acid stock solutions were prepared in 50 mL volumetric flasks from commercial samples (Sigma-Aldrich, Alfa Aesar) with stated purity ≥98%. Samples were degassed by vacuum before use.

2.7.7 Screening experiments

AncQR, AncCE, AncQ and AncE were screened for binding to L-Ala, L-Asp, L-Glu, L-His, L-Lys, L-Leu, L-Asn, L-Gln, L-Arg, L-Ser, D-Ser, L-Thr, L-ornithine and L-cystine by ITC. Variants of AncCE were screened for binding to L-Asp, L-Glu, L-His, L-Lys, L-Asn, L-Gln, L-Arg, and L-ornithine only. Typically 30 µL of 2.5 mM amino acid solution was injected continuously into 100 µM protein solution over 300 s. The signal was recorded for a further 300 s after the injection. Control experiments, where each amino acid solution was injected into phosphate buffer, were also performed.

2.7.8 Determination of binding affinities and thermodynamic parameters

The experimental parameters for quantitative ITC experiments were varied according to the affinity of the interaction being studied. In general, 15-25 injections were made into 100 μM protein solution, with the ligand concentration and injection volume chosen to ensure complete formation of the protein-ligand complex. Where appropriate, the background heat was estimated as the average heat of the last few injections. Otherwise, the background heat was estimated as the average heat associated with an injection in a titration of ligand into buffer, using an identical protocol to the corresponding protein-ligand titration. All titrations used for determination of thermodynamic parameters were performed in triplicate at least. Data analysis was done in NanoAnalyze software (TA Instruments). The raw power signals were integrated, background heat was subtracted from each data point, and thermodynamic parameters were determined by fitting the resulting binding isotherm to the independent binding sites model. For titrations with c >1, where c is the product of the association constant (K_a) for the interaction and the protein concentration in the cell, the stoichiometry (n), enthalpy (ΔH) and K_a for the interaction were determined. For titrations with c <1, n was fixed at 1 while ΔH and K_a were determined.

2.7.9 Competitive displacement experiments

Competitive displacement ITC experiments were performed to confirm the binding specificity of the ancestral and extant proteins. Titrations of AABPs with high-affinity ligands were repeated with a cocktail of up to three amino acids, each at a concentration of 1 mM, included in both the cell and syringe solutions. A similar procedure was used to test for binding of each ancestral protein to L-cysteine, which could not be injected directly into the cell in ITC experiments, due to heat associated with oxidation of this amino acid. Instead, titrations of AABPs with high-affinity ligands were repeated with 5

mM L-cysteine included in the cell and syringe solutions, to test for a reduction in apparent binding affinity or a change in apparent binding enthalpy indicative of L-cysteine binding. Thermodynamic parameters for the AncE-cysteine and ecDEBP-cysteine interactions were estimated as follows: 5 mM L-cysteine, included in the cell and syringe solutions, was displaced from the protein by titration with L-aspartate (AncE) or L-glutamate (ecDEBP). Apparent values for K_a and ΔH (K_{app} and ΔH_{app}) were obtained by fitting the data to the independent binding sites model; these values were used to calculate K_a and ΔH for the protein-cysteine interaction ($K_{a,Cys}$ and ΔH_{Cys}) , using the equations $K_{a,Cys} = ((K_{a,Glu} / K_{app}) - 1) / [Cys]$ and $\Delta H_{Cys} = (\Delta H_{Glu} - \Delta H_{app}) \times (1 + (1 / (K_{a,Cys} \times [Cys])))$ (Zhang and Zhang, 1998). Errors in $K_{a,Cys}$ and ΔH_{Cys} were propagated through these equations.

2.7.10 Structure determination

Crystallisation was done using the vapour diffusion method at 4 °C. AncQR-Arg crystals used for structure determination grew from a hanging drop containing 2 µL protein and 2 µL 0.2 M Li_2SO_4, 0.1 M HEPES pH 7.50, 27.5% (w/v) PEG 3350 as the precipitant. AncQR-Gln crystals used for structure determination were obtained by serial microseeding from crystals grown in a hanging drop containing 2 µL protein and 2 µL 0.2 M $MgCl_2$, 0.1 M HEPES pH 7.50, 24% (w/v) PEG 3350 as the precipitant. Crystals were cryoprotected in 30% (w/v) PEG 3350 and flash-frozen in a nitrogen stream at 100 K. X-ray diffraction data were collected on the MX1 beamline (AncQR-Arg) or MX2 beamline (AncQR-Gln) at the Australian Synchrotron (Melbourne, Australia). Data were indexed and integrated in iMOSFLM (Battye et al., 2011), and scaled in Aimless in the CCP4 suite (Winn et al., 2011). The structure of the AncQR-Arg complex was solved by molecular replacement in PHASER (McCoy et al., 2007), using the structure of the ecGlnBP-Gln complex (PDB: 1WDN) as a search model. The structure of the AncQR-Gln complex was solved by molecular replacement in MOLREP (Vagin and Teplyakov,

1997) using the AncQR-Arg structure as a search model. Models were built manually in Coot (Emsley et al., 2010) and refined by iterative reciprocal- and real-space refinement in REFMAC5 (Murshudov et al., 1997) and Coot. Data collection and refinement statistics are given in Table 2.3. The coordinates and structure factors for the crystal structures of AncQR complexed with L-arginine and L-glutamine have been deposited in the PDB under accession codes 4ZV1 and 4ZV2, respectively.

2.7.11 Structure analysis

RMSD values were calculated in Bio3D (Grant et al., 2006). Identification of the rotation axis and calculation of the degree of rotation for the rigid body domain displacement between the AncQR-Arg and AncQR-Gln structures was done using DynDom (Hayward and Berendsen, 1998). The angle between the rotation axis and the hinge axis (44.8°) was calculated taking the hinge axis as the vector between the centres of mass of the following groups of backbone atoms in the AncQR-Arg structure: (1) residues 94 and 195; and (2) residues 96 and 193.

Chapter Three

Evolution of an enzyme from a solute-binding protein. Part I: Function.

3.1 Introduction

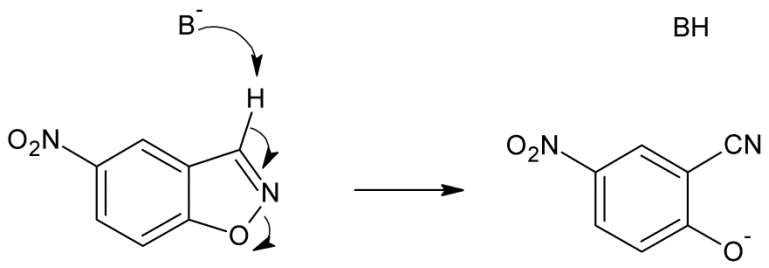
3.1.1 The origins of enzymes: design, engineering and evolution

As shown in Section 1.1.2, "molecular tinkering" (Bridgham et al., 2010; Jacob, 1977) accounts for much of the functional diversity found in modern enzyme superfamilies; specialised enzymes frequently evolve from enzymes with pre-existing functional diversity *via* improvement of promiscuous functions, and previous studies of enzyme evolution have mostly focussed on this kind of evolutionary process. More dramatic functional changes are also observed in enzyme evolution; for example, enzymes can evolve from non-catalytic proteins (Todd et al., 2002). However, the genetic, biophysical or structural basis for the evolution of enzymes from non-catalytic proteins has not been studied previously, despite the fact that introducing catalytic activity into protein scaffolds is currently a major goal of protein engineering and design.

This section gives an overview of the strategies that have been developed to engineer catalytic proteins from non-catalytic scaffolds: *de novo* computational enzyme design, selection of combinatorial libraries, co-option of the mammalian immune system to produce catalytic antibodies, and "minimalist" enzyme design. The current limitations of these strategies and possible implications for the evolution of enzymes from non-catalytic proteins are discussed. Finally, I discuss examples of enzymes that have evolved from non-catalytic SBPs and argue that an improved understanding of these evolutionary processes could inform new strategies for enzyme design.

De novo *computational enzyme design.* The potential impact of *de novo* computational enzyme design is that the advantages of enzyme catalysis, including regioselectivity, stereoselectivity, rapid turnover under mild conditions, and non-toxicity, can be extended to synthetically useful reactions outside the scope of naturally occurring enzymes. Recent advances in computational enzyme design have resulted in enzyme catalysts for the Kemp

elimination (Privett et al., 2012; Röthlisberger et al., 2008), the retro-aldol reaction (Jiang et al., 2008), the Diels-Alder reaction (Siegel et al., 2010), and ester hydrolysis (Rajagopalan et al., 2014). The reactivity of designed enzymes can be extended using cofactors such as metal ions (Khare et al., 2012). The so-called inside-out approach to enzyme design begins with design of an idealised active site called a theozyme (Hilvert, 2013; Kiss et al., 2013). The theozyme specifies an idealised transition state for the enzyme-catalysed reaction, and includes protein functional groups designed to stabilise the transition state. For example, theozymes for the Kemp elimination (Scheme 3.1) typically include a desolvated carboxylate group as a general base, a hydrogen bond donor to stabilise negative charge developing on the phenolic oxygen, and a π stacking residue to assist substrate binding and stabilise positive charge delocalised across the aromatic ring in the transition state (Privett et al., 2012). The precise three-dimensional geometry of the theozyme is optimised using quantum mechanical calculations. Next, the theozyme is grafted into an existing protein scaffold, which is chosen such that the geometry of functional groups of the protein in the theozyme is compatible with the positions of backbone atoms in the scaffold (Zanghellini et al., 2006). Finally, the catalytic residues are restrained while the remainder of the active site is redesigned, ideally stabilising the catalytic residues in the necessary conformation and optimising shape complementarity with the transition state. The anticipated result is a stable protein that selectively binds the transition state for the desired chemical transformation, resulting in an efficient catalyst.



Scheme 3.1. Mechanism of the Kemp elimination, a model reaction for enzyme design.

The low catalytic efficiencies of enzymes designed to date expose deficiencies in our understanding of enzyme catalysis (Korendovych and DeGrado, 2014). Many initial designs are inactive or exhibit modest rate accelerations: k_{cat}/K_M ~1–100 M^{-1} s^{-1} (Lassila et al., 2010), compared with a median k_{cat}/K_M of 1.25×10^5 M^{-1} s^{-1} for natural enzymes (Bar-Even et al., 2011). Measures of catalytic proficiency that account for the rate of the uncatalysed reaction are even less encouraging; the median K_{tx}^{-1} for natural enzymes is 5.2×10^{18} M^{-1} compared with 8.3×10^6 M^{-1} for computationally designed enzymes (Mak and Siegel, 2014)[1]. In the case of the Kemp elimination reaction, even the most proficient designed enzymes are outpaced by an appropriate small molecule control; the acetate-catalysed Kemp elimination in acetonitrile, mimicking a desolvated general base in an enzyme active site, has a pseudo-first order rate constant of ~5600 s^{-1}, compared with a k_{cat} of 700 s^{-1} for the most efficient designed Kemp eliminase following extensive optimisation by directed evolution (Korendovych and DeGrado, 2014).

Baker argues that computational enzyme design can fail at three stages: firstly, the idealised active site geometry might be inadequate to stabilise the transition state; secondly, the idealised active site geometry might not be realised in the designed protein; and thirdly, the active site geometry might be realised in the designed protein but

[1] $K_{tx}^{-1} = (k_{cat}/K_M)/k_{uncat}$ (Radzicka and Wolfenden, 1995).

incompatible with other features of the protein such as long-range electrostatics and dynamics (Baker, 2010). Directed evolution of designed enzymes has been used to identify important features that were missing in the original designs (Blomberg et al., 2013; Giger et al., 2013; Preiswerk et al., 2014). For example, the catalytic efficiency of a designed Kemp eliminase was increased >500-fold by directed evolution: a new hydrogen bond donor was introduced to stabilise charge developing in the transition state, and the shape complementarity of the active site was improved, discouraging the substrate from binding in non-productive conformations and improving the alignment of the catalytic base with the substrate (Blomberg et al., 2013). In this case, the design for the active site was apparently adequate, but directed evolution was required to implement the design in the protein scaffold with the precision required for efficient catalysis.

However, even after extensive optimisation by directed evolution and other protein engineering strategies, designed enzymes have not yet reached the catalytic proficiency of the most efficient natural enzymes, which can achieve rate accelerations (k_{cat}/k_{uncat}) up to ~10^{26} and catalytic efficiency (k_{cat}/K_M) up to the diffusion limit (10^8 – 10^9 M^{-1} s^{-1}) (Mak and Siegel, 2014). In that case, what features of natural enzymes are not currently reproduced in designed enzymes? Factors such as protein dynamics, hydrogen tunnelling, substrate binding, product release or long-range electrostatics may need to be taken into account in the design process (Baker, 2010; Blomberg et al., 2013; Boehr et al., 2006; Nagel and Klinman, 2009). For example, a recent attempt to transplant the active site of an enantioselective haloalkane dehalogenase into a homologous non-selective dehalogenase failed despite precise conservation of the transplanted active site geometry in the engineered protein; it was suggested that the functional transformation was unsuccessful because structural dynamics and hydration at the active site entrance of the target dehalogenase were not reproduced in the engineered protein (Sykora et al., 2014).

Selection from combinatorial libraries. Selection of *de novo* enzymes from combinatorial libraries is not only useful for producing enzymes with practical applications, but informative for understanding the emergence of natural enzymes (Smith and Hecht, 2011; Urvoas et al., 2012). For example, what fraction of sequence space is populated by proteins that bind small molecules or have catalytic activity? What fraction of proteins with a well-defined tertiary structure have catalytic activity? Does selection for catalytic activity alone produce well-structured proteins? By addressing these questions using enzymes derived from combinatorial libraries, the role of chance in the emergence of natural enzymes can be better understood. For example, Keefe and Szostak showed that ATP-binding proteins could be recovered with a frequency of ~10^{-11} from a library of random polypeptide sequences; four specific ATP-binding proteins were selected by mRNA display from a library of 6×10^{12} sequences 80 residues long (Keefe and Szostak, 2001). Surprisingly, one of the ATP-binding proteins was shown to catalyse the hydrolysis of ATP *in crystallo*, showing that selection for binding alone can result in promiscuous catalytic activity (Simmons et al., 2009).

Enzymes can be recovered at much higher frequencies from combinatorial libraries specifically designed to encode folded proteins. For example, Hecht and co-workers designed a superfamily of *de novo* four-helix bundles using binary patterning, in which protein sequences are composed of random polar and nonpolar residues in particular patterns that result in the formation of amphipathic helices (Patel et al., 2009). A surprisingly high proportion of proteins from these naïve binary patterned libraries exhibit catalytic activity: in one library, 50% of proteins exhibited heme-dependent peroxidase activity, 30% exhibited esterase activity, and 20% exhibited lipase activity (Patel et al., 2009). In addition, 18 artificial four-helix bundles selected from a binary patterned library rescued four different auxotrophic *E. coli* strains missing conditionally essential genes, which showed that artificial proteins from a combinatorial library can

replace metabolic enzymes *in vivo*, albeit *via* mechanisms currently unknown (Fisher et al., 2011).

Combinatorial selection of enzymes can also make use of features in pre-existing protein scaffolds. For example, enzymes with a unique RNA ligase activity were isolated from a combinatorial library of 4×10^{12} non-catalytic zinc finger scaffolds with two completely randomised loops (Seelig and Szostak, 2007). One of these RNA ligases was structurally characterised and exhibited a new fold; the zinc-binding sites were remodelled and the two helices in the original scaffold were replaced with an unstructured loop (Chao et al., 2013).

Catalytic antibodies. The development of catalytic antibodies has clear parallels with the evolution of enzymes from non-catalytic proteins: the primary role of antibodies is antigen binding rather than catalysis, and catalytic activity is gained by affinity maturation, an algorithm of iterative mutagenesis and selection similar to evolution. Immunisation against hapten molecules that mimic the conformational, stereochemical and electronic properties of the transition state for a particular reaction produces antibodies that selectively bind the transition state and are therefore catalytic (Hilvert, 2000; Schultz et al., 2002). Immunisation against mechanism-based inhibitors instead of transition state analogues can be used to introduce catalytic residues into the antibody scaffold – a method called reactive immunisation. Although antibody catalysts have been produced for over 100 different reactions, and some of these catalysts have desirable properties such as chemo-, regio-, and stereospecificity, their catalytic efficiencies have never approached those of natural enzymes (Hilvert, 2000). Potential explanations include (i) insufficient selection pressure in the immune system for hapten binding tight enough to generate enzyme-like catalytic efficiency (K_d $<10^{-10}$ M); (ii) selection is for binding rather than catalysis, preventing, for example, the evolution of pathways for

product dissociation; (iii) haptens cannot perfectly replicate the properties of transition states; and (iv) limitations of the immunoglobulin fold itself – for example, intrinsic flexibility in the complementarity determining regions that limit pre-organisation of the active site (Hilvert, 2000; Hollfelder et al., 2000; Padlan, 1994).

Minimalist enzyme design. As discussed above, designing an enzyme by attempting to engineer complementarity between a protein scaffold and the transition state for a particular reaction is a challenging and computationally demanding task. DeGrado and co-workers have instead advocated a "minimalist" strategy for designing enzymes from non-catalytic proteins, in which a single reactive residue is inserted into a hydrophobic cavity where it can interact productively with a given substrate (Korendovych and DeGrado, 2014). This strategy is reminiscent of the efficient catalysis of model reactions by serum albumins, which are not normally considered enzymes; for example, a reactive lysine residue (as a general base) in a hydrophobic pocket is sufficient for catalysis of the Kemp elimination reaction in serum albumin (Hollfelder et al., 1996, 2000). Using this minimalist strategy, calmodulin derivatives that catalyse the Kemp elimination, the retro-aldol reaction and ester hydrolysis have been designed (Korendovych et al., 2011; Moroz et al., 2015; Raymond et al., 2014). Notably, these enzymes could evolve naturally given the appropriate selective pressure, since they are produced by a single point mutation in a functional non-catalytic protein, and could be optimised further by evolution, as shown by the 220-fold increase in activity in a designed Kemp eliminase achieved by directed evolution (Moroz et al., 2013). It remains to be seen whether the accessibility of enzymatic activities *via* single point mutations is a peculiarity of calmodulin – a protein with a flexible hydrophobic pocket that has evolved to interact with multiple partners – or a general feature of non-catalytic proteins (Moroz et al., 2013).

Evolution of enzymes from non-catalytic proteins. As shown above, rudimentary enzymes can be designed rationally by insertion of the appropriate catalytic machinery into non-catalytic scaffolds and arise frequently by chance in folded proteins; therefore, we should also expect to see examples of enzymes that have evolved from non-catalytic proteins in nature. Indeed, in the SBP superfamily (for example), six enzymes with the type II SBP fold have been discovered (Table 3.1). Although extensive divergent evolution has obscured sequence similarity between some of these enzymes (e.g. ATP phosphoribosyltransferase) and non-catalytic SBPs, the conserved topology consisting of two α/β domains connected by two flexible β strands is a clear indicator of their evolutionary history.

Table 3.1. Enzymes with the type II SBP fold. These enzymes were identified by searching the PDB for proteins belonging to CATH superfamily 3.40.190.10 (periplasmic binding protein-like II) (Sillitoe et al., 2015). The PDB code of a representative structure, the Enzyme Commission (EC) number, and the gene ontology (GO) annotation are given for each enzyme.

Name	**PDB**	**EC**	**GO**
ATP phosphoribosyltransferase	1H3D	2.4.2.17	Histidine biosynthesis
Thiaminase I	3THI	2.5.1.2	Thiamine catabolism
Porphobilinogen deaminase	1PDA	2.5.1.61	Porphyrin biosynthesis
1,4-dihydroxy-6-naphthoate synthase	3A3U	4.1.-.-	Menaquinone biosynthesis
Prephenate dehydratase	3MWB	4.2.1.51	Phenylalanine biosynthesis
Cyclohexadienyl dehydratase	3KBR	4.2.1.51 4.2.1.91	Phenylalanine biosynthesis

The structures of SBP-like enzymes have revealed the various ways in which the SBP fold has been adapted to highly specialised and chemically complex roles in primary metabolism, including recruitment of cofactors and additional domains and the evolution of new oligomeric structures (Figure 3.1). For example, porphobilinogen deaminase catalyses the condensation of four porphobilinogen molecules to form the porphyrin precursor 1-hydroxymethylbilane. This is achieved by sequential nucleophilic addition of deaminated porphobilinogen monomers to a dipyrromethane cofactor, which occupies the capacious cavity between the two SBP-like domains and is covalently linked *via* a cysteine sulfhydryl group to a third α/β domain (Figure 3.1A) (Louie et al., 1992). The active site is covered by a flexible loop. MD simulations have shown that the flexibility of the active site loop and flexibility of the SBP-like domains about the adjoining hinge are both needed to accommodate the growing polypyrrole chain (Bung et al., 2014). As another example, ATP phosphoribosyltransferase catalyses the transfer of a phosphoribosyl group from 5′-phosphoribosyl 1′-pyrophosphate onto the N1 atom of the adenosine ring of ATP – a condensation reaction dependent on a Mg^{2+} cofactor. The active site of this enzyme is formed between two SBP-like domains in an open conformation (Figure 3.1B). A third histidine-binding regulatory domain controls catalytic activity by altering the oligomeric state of the enzyme, which is active only in the dimeric form (Cho et al., 2003; Lohkamp et al., 2004).

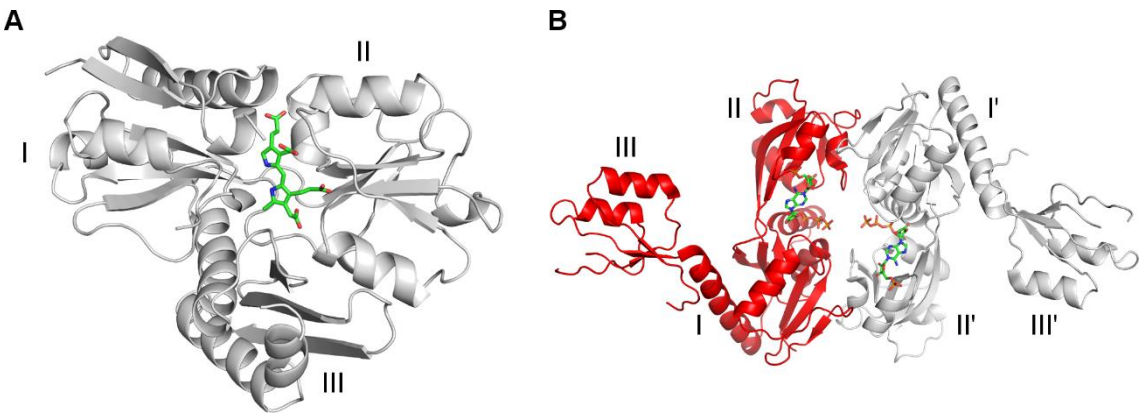


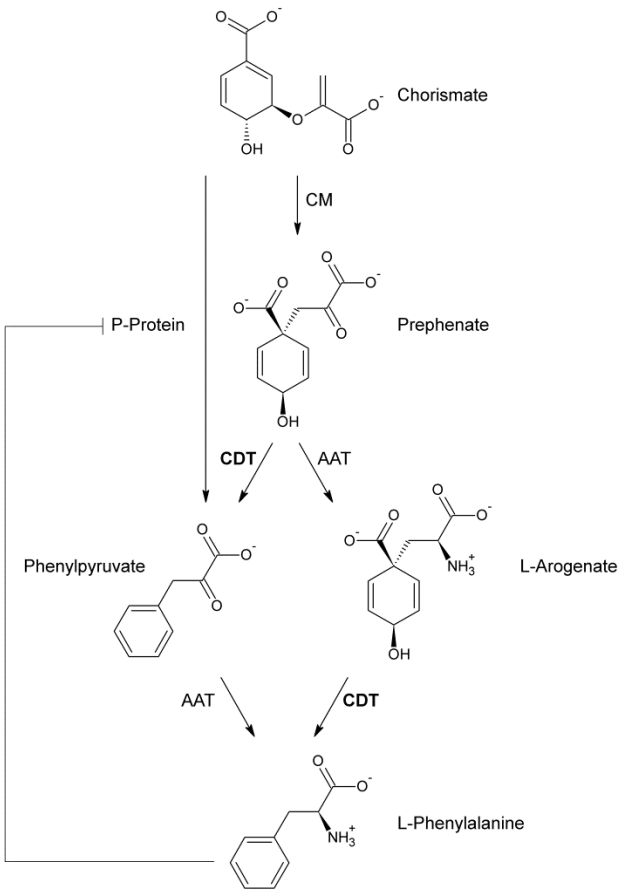
Figure 3.1. Architectures of two SBP-like enzymes. (**A**) Porphobilinogen deaminase bound to dipyrromethane cofactor (PDB: 1PDA). (**B**) ATP phosphoribosyltransferase dimer bound to 1-(5-phospho-D-ribosyl)-ATP (PDB: 1Q1K). In both structures, domains I and II have the type II SBP fold.

The precise molecular mechanisms for the evolution of enzymes from non-catalytic SBPs – or non-catalytic proteins in general – have not been investigated, leaving important questions unresolved. For example, the SBP superfamily has undergone functional diversification to interact specifically with an extensive range of solutes; the SBP scaffold is clearly evolvable. Moreover, these proteins represent a potential reservoir of substrate-binding proteins that could become catalysts upon introduction of the appropriate catalytic machinery. Why, then, are only a few examples of highly specialised SBP-derived enzymes known? How are the properties of SBPs that are necessary for transport functions – for example, conformational cycling between open and closed structures – reconciled with the apparently conflicting demand of a pre-organised active site for efficient catalysis? How were the positioning and reactivity of catalytic groups optimised, and complementarity between the enzyme and transition state of the reaction improved? As discussed above, previous attempts at protein design have not yet achieved catalytic efficiencies similar to natural enzymes; identifying the mechanisms by which nature has improved *de novo* enzymes could help protein engineers to do the same. To

this end, we investigated the evolution of the model enzyme cyclohexadienyl dehydratase from a non-catalytic SBP precursor.

3.1.2 Cyclohexadienyl dehydratase

Cyclohexadienyl dehydratase (CDT; EC 4.2.1.51, 4.2.1.91) is encoded by the gene *pheC* and catalyses the conversion of prephenate to phenylpyruvate and the conversion of L-arogenate to L-phenylalanine. These reactions are involved in the biosynthesis of L-phenylalanine. In bacteria, L-phenylalanine and other aromatic amino acids are synthesised through the shikimate pathway (Scheme 3.2) (Dosselaere and Vanderleyden, 2001). Chorismate, a major branch point metabolite in the shikimate pathway, is converted to prephenate, the common precursor of L-phenylalanine and L-tyrosine, by a Claisen rearrangement reaction. L-Phenylalanine is then obtained by transamination and Grob-like fragmentation of prephenate, *via* the intermediate phenylpyruvate or L-arogenate. In many Gram-negative bacteria, including *E. coli* and *Pseudomonas aeruginosa*, the primary biosynthetic pathway from chorismate to L-phenylalanine is mediated by a bifunctional chorismate mutase-prephenate dehydratase (P-protein), which converts chorismate to phenylpyruvate *via* prephenate, and an aromatic aminotransferase which converts phenylpyruvate to L-phenylalanine. This pathway is localised in the cytoplasm and is regulated by feedback inhibition of P-protein by L-phenylalanine.



Scheme 3.2. Multiple pathways of L-phenylalanine biosynthesis in *P. aeruginosa*. In the core pathway, chorismate is channelled through to phenylpyruvate by P-protein, a bifunctional chorismate mutase-prephenate dehydratase (Calhoun et al., 1973). An aromatic aminotransferase (AAT) converting phenylpyruvate to L-phenylalanine completes the pathway. This pathway is controlled by feedback inhibition of P-protein by L-phenylalanine (Calhoun et al., 1973). In the periplasmic overflow pathway, chorismate is converted to prephenate by a monofunctional chorismate mutase (CM) (Calhoun et al., 2001). Subsequent reactions catalysed by cyclohexadienyl dehydratase (CDT) and various aromatic aminotransferases *via* phenylpyruvate or L-arogenate complete the pathway to L-phenylalanine (Patel et al., 1977).

CDT is involved in an alternative "overflow" pathway to L-phenylalanine found in addition to the cytoplasmic pathway in some Gram-negative bacteria such as *P. aeruginosa* (Scheme 3.2). In this pathway, chorismate is converted to prephenate by a monofunctional chorismate mutase, and prephenate is converted to L-phenylalanine by CDT and various aromatic aminotransferases, *via* the intermediate L-arogenate or

phenylpyruvate. The overflow pathway to L-phenylalanine is localised in the periplasm (Calhoun et al., 2001; Zhao et al., 1993) and is not feedback regulated (Fiske et al., 1983). Flux through the overflow pathway was first demonstrated in mutants of *P. aeruginosa* that lack feedback inhibition of the L-tyrosine-regulated isotype of 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase; these mutants accumulate L-phenylalanine through the overflow pathway because they lack early-pathway regulation of the shikimate pathway (Fiske et al., 1983). However, under normal laboratory growth conditions, flux through the overflow pathway is disfavoured by the relatively low affinities of CDT and aromatic aminotransferases for prephenate and L-arogenate and limited substrate availability, especially due to substrate channelling (Zhao et al., 1992). The spatial separation of cytoplasmic chorismate biosynthesis from the periplasmic overflow pathway may also disfavour flux through the latter pathway.

In addition to its potential role in L-phenylalanine biosynthesis, CDT has a role in L-arogenate catabolism in *P. aeruginosa*, which can use L-arogenate as a sole carbon or nitrogen source (Fischer et al., 1997). One pathway for L-arogenate catabolism begins with conversion of L-arogenate to L-phenyalanine by CDT, followed by hydroxylation of L-phenylalanine to give L-tyrosine, which is catabolised through the homogentisate pathway (Ramos, 2004). The involvement of CDT in this pathway accounts for the three-fold increase in doubling time of *pheC* knockouts of *P. aeruginosa* (relative to wild type) cultured in media containing L-arogenate as the sole nitrogen source (Fischer et al., 1997). In *P. aeruginosa*, L-arogenate is a chemoattractant with a receptor distinct from the chemoreceptors for aromatic amino acids (Fischer et al., 1997), suggesting that catabolism of L-arogenate by CDT is important under certain environmental conditions.

The overflow pathway to L-phenylalanine has a potential role in interactions between phytopathogens and their hosts. The *pheC* gene has a limited phylogenetic

distribution and it has been noted that several of the species possessing the gene are opportunistic or obligate pathogens of plants, which primarily use the L-arogenate pathway rather than the phenylpyruvate pathway for L-phenylalanine biosynthesis (Zhao et al., 1993). Furthermore, the periplasmic localisation of the pathway is suggestive of environmental responsiveness (Zhao et al., 1993). Genome-wide studies of pathogenicity determinants have identified components of the phenylalanine overflow pathway as being essential for phytopathogenicity; for example, disruption of the *pheC* gene encoding a homolog of CDT in *Dickeya chrysanthemi* markedly reduces maceration of African violet leaves by this pathogen (Okinaka et al., 2006). Together, these observations suggest an important role for CDT in plant-pathogen interactions. This view is also supported by recent work on the role of chorismate mutase (the enzyme immediately upstream of CDT) in plant infection: a chorismate mutase secreted by the causative agent of maize smut, *Ustilago maydis*, enhances virulence by redirecting flux away from the salicylic acid pathway to the phenylpropanoid pathway in the host plant (Djamei et al., 2011); salicylic acid is a phytohormone synthesized from chorismate that induces pathogenesis defence genes (Wildermuth et al., 2001). The mechanism by which CDT enhances pathogenicity is currently unknown but could similarly involve rerouting metabolic flux through the shikimate pathway towards the phenylalanine overflow pathway, in order to undermine biosynthesis of salicylic acid and other defensive compounds by the host.

CDT was recognized as a member of the AABP family on the basis of sequence homology soon after the nucleotide sequence of the gene was made available (Tam and Saier, 1993b; Zhao et al., 1992). More recently, the crystal structure of CDT from *P. aeruginosa* (PDB: 3KBR; Midwest Center for Structural Genomics, unpublished) shows that the SBP fold has been conserved in CDT and provides structural evidence for a close evolutionary relationship between CDT and AABPs (Figure 3.2).

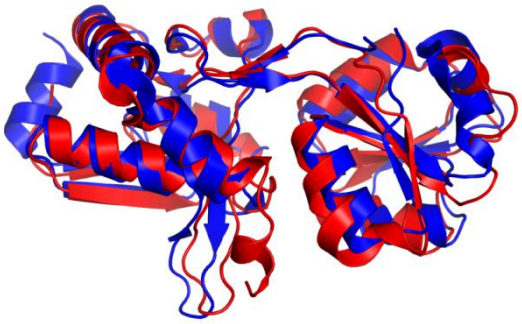
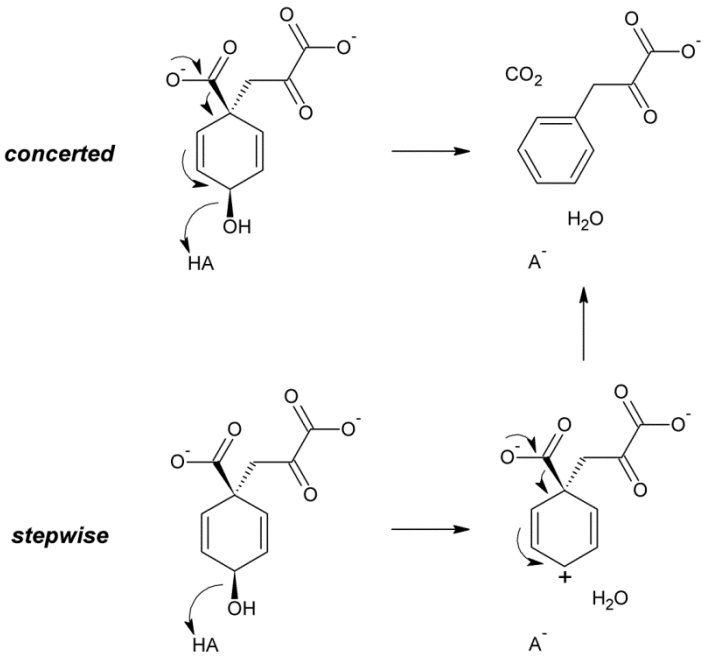


Figure 3.2. Structural evidence for the evolutionary relationship between CDT and SBPs. The structure of CDT from *P. aeruginosa* (blue) superimposed on the structure of the amino acid-binding protein Ws0279 from *Wolinella succinogenes* (red).

A major advantage of CDT as a model system for studying the emergence of catalytic activity is that enzyme activity can be assessed using high-throughput genetic complementation assays. Since P-protein, encoded by the gene *pheA*, is the sole source of prephenate dehydratase activity in *E. coli*, *pheA* knockouts of *E. coli* are phenylalanine auxotrophs (Joyce et al., 2006). Since the chorismate mutase activity of P-protein is redundant, expression of CDT rescues *pheA* knockouts of *E. coli* grown on minimal media by restoring prephenate dehydratase activity (Zhao et al., 1992). Prephenate dehydratase variants with different catalytic efficiencies can be differentiated using this complementation assay (Kleeb et al., 2007). More sophisticated genetic selection strategies for prephenate dehydratase activity have also been developed; the intracellular concentration of prephenate (and thus the strength of selection) can be controlled by inducible expression of a cyclohexadienyl dehydrogenase, which converts prephenate to 4-hydroxyphenylpyruvate, channelling prephenate into L-tyrosine biosynthesis. This method enables selection of prephenate dehydratase variants with a broader range of catalytic efficiencies (Kleeb et al., 2007).

The enzyme mechanism of CDT is currently unknown. Two possibilities can be envisioned: a general acid-mediated concerted elimination of CO_2 and H_2O from the cyclohexadiene substrate, or alternatively, a stepwise mechanism whereby elimination of

H_2O produces a stabilised divinyl carbocation, which is then quenched by elimination of CO_2 (Scheme 3.3) (Hermes et al., 1984). ^{18}O-labelling experiments showed that the reaction occurs *via* the stepwise mechanism in acidic solution (Hermes et al., 1984). On the other hand, ^{13}C kinetic isotope effect experiments implicated the concerted mechanism in the prephenate dehydratase from *Methanocaldococcus jannaschii* (which is not homologous to CDT) (Van Vleet et al., 2010).



Scheme 3.3. Possible mechanisms for the elimination of CO_2 and H_2O from prephenate.

3.1.3 Objectives

The aim of the work described in this chapter was to characterise the evolutionary trajectory from a non-catalytic SBP to a catalytically active CDT using ancestral protein reconstruction. Using phylogenetic analysis and ancestral protein reconstruction, the sequences of extinct intermediates in the evolution of CDT from AABPs were reconstructed, allowing these ancestral proteins to be expressed and characterised, and enabling the genetic and structural mechanisms underpinning the emergence of catalytic activity to be studied directly. Functional characterisation of the ancestral proteins, in addition to several extant homologues of CDT, showed that CDT evolved from a cationic amino acid-binding protein *via* an intermediate of unknown function. Subsequent chapters address the role of conformational dynamics in the evolution of CDT and the mutational and structural bases for the emergence of catalytic activity in the SBP fold.

3.2 Results

3.2.1 Reconstruction of ancestral sequences

The SBP with the highest sequence identity with CDT that has been structurally or functionally characterised is Ws0279 from *W. succinogenes*, which has 26% sequence identity with CDT from *P. aeruginosa* (PaCDT), excluding the signal peptides. The sequence identity between Ws0279 and PaCDT is therefore comparable with pairwise sequence identities between AABPs with different amino acid specificities (for example, 20% – 35% for the AABPs of *E. coli*). The structure of Ws0279 has been solved in complex with L-lysine (PDB: 3K4U; New York SGX Research Center for Structural Genomics, unpublished), indicating that Ws0279 is a SBP rather than an enzyme, although the binding specificity of the protein has not been reported.

To reconstruct the evolutionary history of PaCDT, we obtained the sequences of 131 homologues of Ws0279 and PaCDT from phylogenetically diverse organisms and reconstructed the phylogeny of these sequences using the maximum-likelihood (ML) method (Figure 3.3). The outgroup for this phylogeny consisted of 271 representative sequences from the previous AABP phylogenetic analysis (Section 2.4.1). The topology of the phylogeny was robust to the use of alternative evolutionary models, and convergence of the heuristic tree search to the ML tree was confirmed by repeating the phylogenetic analysis using randomised initial trees. Major branches on the ML phylogeny were supported by high bootstrap values (Figure 3.3).

Five ancestral nodes, designated AncCDT-1 to AncCDT-5, were selected for experimental characterisation (Figure 3.3). These nodes were chosen because patterns of sequence conservation in the extant sequences (discussed in Section 4.2.1) suggested that the evolution of CDT activity occurred between AncCDT-1 and AncCDT-3; the descendants of AncCDT-3 contain conserved residues that are putatively important for

CDT activity. AncCDT-1 represents the last common ancestor (LCA) of Ws0279 and PaCDT, whereas the other ancestral nodes represent intermediates in the evolution of PaCDT from AncCDT-1.

Alternative versions of the ancestral sequences, designated AncCDT-1W to AncCDT-5W, were reconstructed using an alternative substitution model (WAG+I+Γ+F, compared with LG+I+Γ+F for the ML ancestral sequences) and a phylogeny inferred using this alternative substitution model. These alternative sequences were reconstructed to assess the robustness of the phenotypes of the reconstructed ancestral proteins to variation in the substitution model, and had sequence identities of >90% with the corresponding ML ancestral sequences.

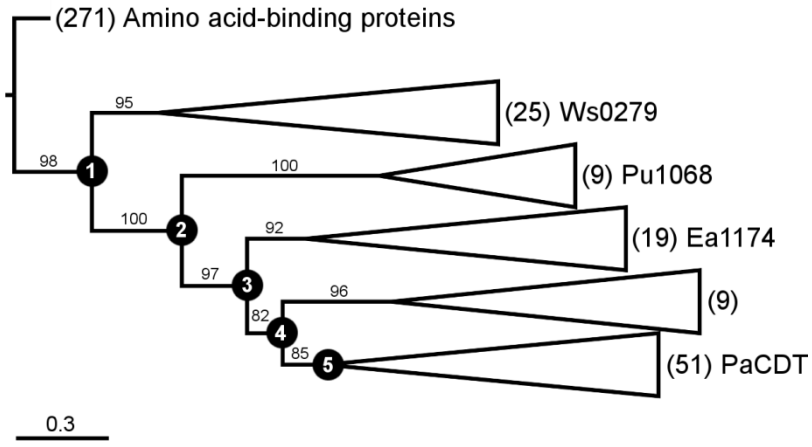


Figure 3.3. Evolution of CDT from SBPs. Condensed ML phylogeny of CDT homologues and AABPs inferred using the LG+I+Γ+F evolutionary model. Branches are labelled with bootstrap values from 100 replicates. The scale bar represents the mean number of substitutions per site. The five compressed clades are labelled with the corresponding number of sequences and the representative protein characterised in this study, if applicable. The five ancestral nodes that were experimentally characterised (AncCDT-1 to AncCDT-5) are labelled.

The mean posterior probabilities (PPs) of the ancestral sequences ranged from 0.78 (AncCDT-2) to 0.89 (AncCDT-5; Table 3.2). The PP distributions at positions close to the binding site, which are putatively important for amino acid binding or CDT activity, are summarised in Figure 3.4. For AncCDT-3, 4 and 5, the ancestral states at these positions were reconstructed unambiguously. For AncCDT-1 and 2, some positions have plausible alternative reconstructions that correspond to conservative substitutions. However, these uncertainties are unlikely to affect the inferred phenotypes of the ancestral proteins, as discussed further in the context of the structure of AncCDT-1 in Section 6.2.1.

Table 3.2. Mean posterior probabilities (PPs) of ancestral CDT variants.

Protein	Mean PP
AncCDT-1	0.80
AncCDT-2	0.79
AncCDT-3	0.85
AncCDT-4	0.85
AncCDT-5	0.89

Figure 3.4. Posterior probability distributions of ancestral CDT variants. The posterior probability (PP) distribution is shown at positions in the binding site putatively important for amino acid binding or CDT activity. The PP distributions of AncCDT-4 and AncCDT-5 are identical to that of AncCDT-3, with the exception of position 80 in AncCDT-5 (*P*(Ser) = 0.70, *P*(Thr) = 0.30), and are therefore not shown. The sequences of Ws0279, Pu1068, Ea1174 and PaCDT at the same positions are also given for comparison.

3.2.2 Functional characterisation of ancestral proteins

This section gives a summary of the functional characterisation of the ancestral proteins completed by Joe Kaczmarski under my supervision. Each ancestral protein was tested for CDT activity using the genetic complementation assay described in Section 3.1.2, in which expression of CDT rescues the growth of *E. coli* phenylalanine auxotrophs that lack prephenate dehydratase activity encoded by the gene *pheA*. Expression of AncCDT-3, 4 and 5 rescued the growth of Δ*pheA* cells in minimal media, with growth rates increasing along the evolutionary trajectory, whereas expression of AncCDT-1 and 2 did not rescue growth (Figure 3.5a). Likewise, expression of AncCDT-3W, 4W and 5W, but not 1W or 2W, rescued growth of Δ*pheA* cells in minimal media, although there was some variation in growth rates compared to AncCDT-3, 4 and 5 (Figure 3.5b). In particular, AncCDT-3W transformants grew considerably faster than AncCDT-3 transformants, reaching maximal OD_{600} in ~2 days rather than ~6 days.

AncCDT-1 to AncCDT-5 were also tested for amino acid binding by ITC. Amino acid binding was detected only for AncCDT-1, which bound cationic amino acids with high affinity: L-arginine (K_d = 0.32 μM), L-ornithine (1.2 μM), L-histidine (2.3 μM) and L-lysine (6.7 μM).

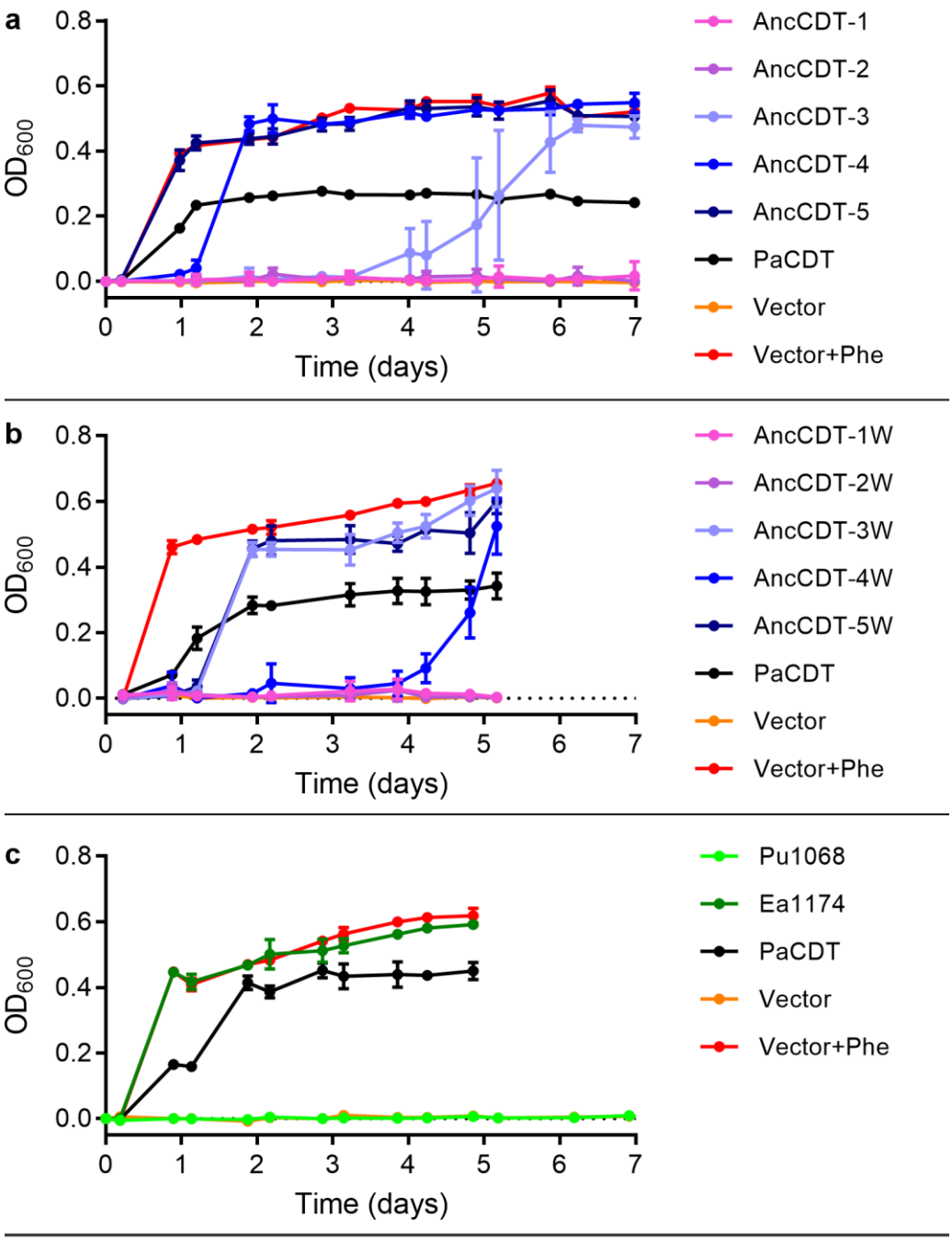


Figure 3.5. Genetic complementation of phenylalanine auxotrophs by ancestral and extant CDT variants. Growth of *E. coli* Δ*pheA* cells transformed with (**a**) ancestral proteins reconstructed using the LG substitution matrix, (**b**) ancestral proteins reconstructed using the WAG substitution matrix, and (**c**) extant homologues of PaCDT in M9–F media at 37 °C. Data represent mean ± s.e.m. (n = 3 – 5). Growth of empty vector transformants in selective M9–F media and unselective M9+F media are also shown. The experiments represented in this figure were performed by Joe Kaczmarski.

These results confirmed the hypothesis that CDT activity evolved on the branch connecting AncCDT-1 and AncCDT-3. However, the function of the intermediate AncCDT-2 remained unknown. To test whether this intermediate ancestral protein had a function distinct from AncCDT-1 (an amino acid-binding protein) and AncCDT-3 (a CDT), or was inactive because of errors in its reconstructed sequence, extant proteins descended from AncCDT-2 and AncCDT-3 were also tested for CDT activity by genetic complementation. The representative extant proteins that were characterised are Pu1068 from "*Candidatus* Pelagibacter ubique" and Ea1174 from *Exiguobacterium antarcticum*; AncCDT-2 is the LCA of Pu1068 and PaCDT, while AncCDT-3 is the LCA of Ea1174 and PaCDT (Figure 3.3). Pu1068 did not exhibit CDT activity in the complementation assay, whereas Ea1174 exhibited CDT activity comparable with PaCDT, providing further evidence that CDT activity evolved between AncCDT-2 and AncCDT-3 (Figure 3.5c). Additionally, AncCDT-2 was purified and tested for promiscuous prephenate dehydratase activity using an *in vitro* spectrophotometric assay for phenylpyruvate formation, but no activity was detected.

The functions of the intermediate ancestral protein AncCDT-2 and a clade of proteins descended from this ancestral protein therefore remained unresolved. In addition, the function of the putative lysine-binding protein Ws0279 had not been confirmed. To strengthen the argument that CDT evolved from an amino acid-binding protein *via* an intermediate with a different function, functional characterisation of Ws0279 and Pu1068 was attempted.

3.2.3 Functional characterisation of Ws0279

Ws0279 has been annotated as a putative lysine-binding protein on the basis of electron density for L-lysine in the crystal structure (PDB: 3K4U). However, the electron density in the binding site is somewhat ambiguous due to the low resolution of the structure (2.62 Å), and because the electron density possibly reflects a mixture of structurally related ligands retained during the purification and crystallisation processes. The protein was therefore expressed in *E. coli* BL21(DE3) cells and purified by nickel affinity chromatography and SEC. Ws0279 eluted from the size-exclusion column as a trimer (*v* ~ 194 mL, calculated MW ~ 75 kDa, theoretical MW for dimer/trimer = 56 kDa/84 kDa), which was surprising because oligomerisation is rare in SBPs (Berntsson et al., 2010; Ruggiero et al., 2014). However, the asymmetric unit of the crystal structure of Ws0279 contains two trimeric assemblies of the protein, supporting the conclusion that the protein is a trimer in solution (Figure 3.6).

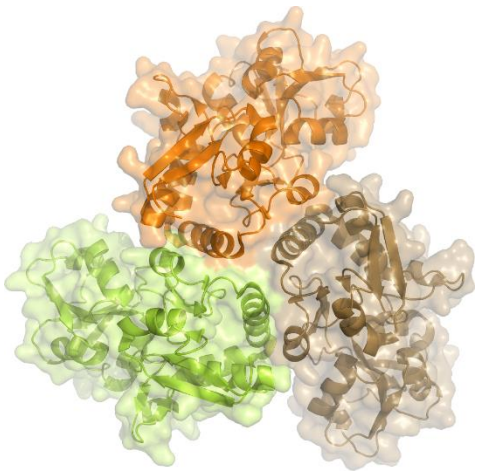


Figure 3.6. Trimeric structure of Ws0279. The Ws0279 trimer (PDB: 3K4U) is viewed down the three-fold non-crystallographic symmetry (NCS) axis. The amount of surface area buried per subunit, calculated using PISA (Krissinel and Henrick, 2007), is 3740 Å^2.

The binding specificity of Ws0279 was confirmed by differential scanning fluorimetry (DSF). This method relies on the fact that binding of a high-affinity ligand, in most cases, increases the melting temperature (T_m) of a protein due to thermodynamic coupling of unbinding and unfolding, which increases the free energy of unfolding (ΔG_u) (Niesen et al., 2007). The thermal denaturation of a protein is monitored using a hydrophobic fluorescent dye, which binds to the hydrophobic regions of a protein that are exposed in the unfolded state. As the dye is transferred from an aqueous environment to a hydrophobic environment, its fluorescence is unquenched. Since fluorescence can be monitored conveniently in 96-well or 384-well plates, DSF is a useful high-throughput method to test the effect of ligands on protein thermostability as a proxy for binding.

Using DSF, Ws0279 was tested for binding of all proteinogenic amino acids except L-cysteine, as well as four non-proteinogenic amino acids that are also bound by some AABPs (D-alanine, D-serine, L-ornithine and L-cystine). This experiment confirmed that Ws0279 is a specific L-lysine binding protein; Ws0279 was significantly stabilised by 10 mM L-lysine (ΔT_m = 7.2 °C) and 10 mM L-arginine, to a lesser extent (2.4 °C) (Figure 3.7a). The T_m increase observed for Ws0279 in the presence of 1 mM L-lysine (6.1 °C) is comparable with the T_m increases reported for other AABPs in the presence of their physiological ligands under similar conditions (≥5 °C) (Giuliani et al., 2008).

We also tested binding of amino acids to AncCDT-1 by DSF to compare the binding specificities of AncCDT-1 and Ws0279 directly. As expected, AncCDT-1 was significantly stabilised by L-arginine (ΔT_m = 6.8 °C), L-ornithine (5.8 °C), L-lysine (4.5 °C) and L-histidine (2.7 °C), each at a concentration of 10 mM (Figure 3.7b). Comparison of the ΔT_m profiles of AncCDT-1 and Ws0279 shows that AncCDT-1 has broader ligand specificity than Ws0279. If our reconstruction of AncCDT-1 faithfully reproduces the

phenotype of the LCA of Ws0279 and PaCDT, specificity towards L-lysine was not the ancestral phenotype and must have evolved in the Ws0279 lineage.

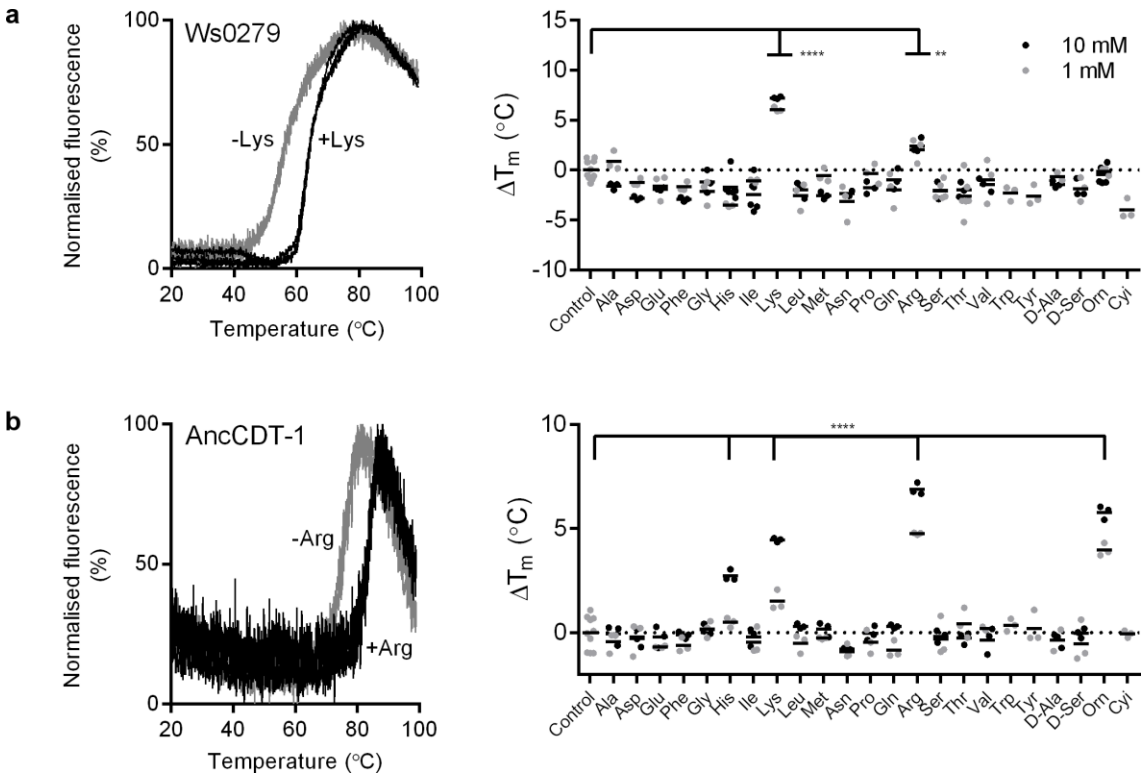


Figure 3.7. Amino acid binding profiles of Ws0279 and AncCDT-1. (**a**) Ws0279; (**b**) AncCDT-1. Left panel: Examples of thermal denaturation data from DSF: normalised fluorescence intensity as a function of temperature. L-lysine and L-arginine give the largest increases in melting temperature (T_m) for Ws0279 and AncCDT-1, respectively. Three replicates are shown for each sample (protein only and protein + 10 mM ligand). Right panel: T_m values for each protein in the presence of 1 mM or 10 mM amino acids, derived from a Boltzmann fit to the temperature-dependent fluorescence data, compared to a protein-only control. The T_m of Ws0279 was 57.6 ± 1.0 °C and the T_m of AncCDT-1 was 77.2 ± 0.8 °C (mean ± s.d., n = 10). Asterisks indicate a significant increase in T_m with 10 mM ligand compared to the control by one-way ANOVA with Dunnett's test for multiple comparisons (**P < 0.01, ****P < 0.0001).

3.2.4 Structure and function of Pu1068

Pu1068 originates from the oligotrophic oceanic bacterium "*Candidatus* Pelagibacter ubique" of the SAR11 clade of α-proteobacteria. "*Ca.* P. ubique" has the smallest genome (1.31×10^6 bp) and the fewest genes (1354 open reading frames) of any known free-living organism, but has retained the majority of core metabolic functions (Giovannoni et al., 2005). Its streamlined genome also contains a relatively high proportion of transport proteins, especially high affinity SBP-dependent ABC transporters, which are important for scavenging in a nutrient-poor marine environment (Giovannoni et al., 2005). Genomic context suggests that the gene *Pu1068* encodes a solute-binding protein involved in an ABC transport system rather than an enzyme; *Pu1068* is immediately upstream of a gene encoding the transmembrane component of an ABC transporter (*glnP*) (Figure 3.8). *Pu1068* also appears to be co-transcribed with four genes encoding an amine demethylase of the heterotetrameric sarcosine oxidase family (*soxBDAG*) (Figure 3.8).

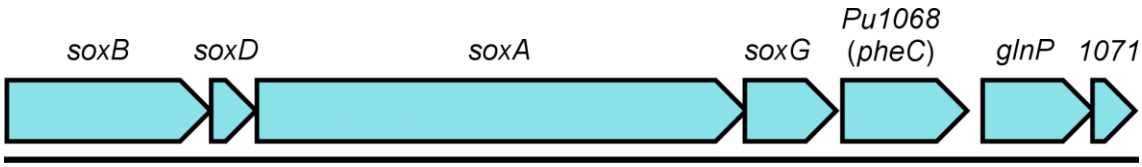


Figure 3.8. Genomic context of *Pu1068* in "*Ca.* P. ubique" strain HTCC1062. The length of the genomic region shown is 6945 nt.

Pu1068 was expressed in *E. coli* BL21(DE3) cells and purified by nickel affinity chromatography and SEC. The protein eluted as a monomer (v ~ 234 mL, calculated MW ~ 19 kDa; theoretical MW for monomer = 28 kDa). DSF experiments confirmed that Pu1068 is not an amino acid-binding protein; no significant increase in T_m (>2 °C) indicative of a potential interaction was observed in the presence of any amino acid (Figure 3.9).

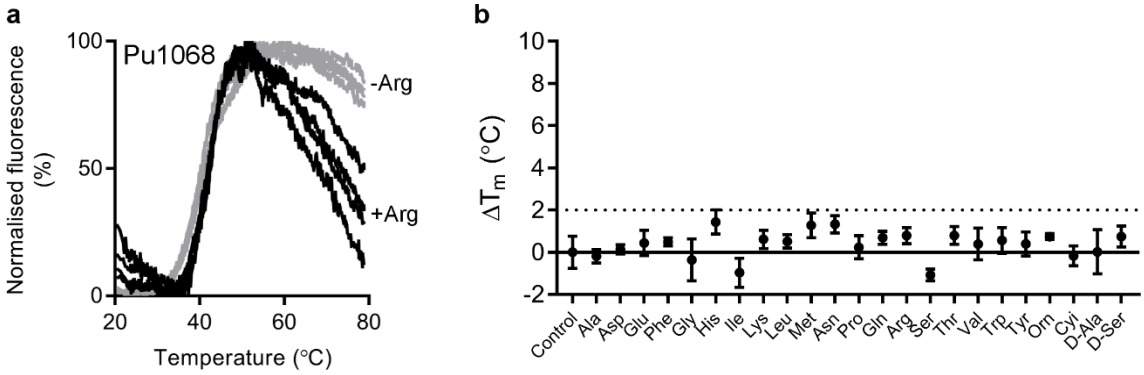


Figure 3.9. Pu1068 is not an amino acid-binding protein. (**a**) Example of thermal denaturation data for Pu1068 from DSF. Normalised fluorescence intensity as a function of temperature in the absence (grey) and presence (black) of 10 mM L-arginine. Four replicates are shown for each condition. (**b**) T_m of Pu1068 in the presence of amino acids, derived from a Boltzmann fit to the temperature-dependent fluorescence data, compared to a protein-only control. T_m values represent mean ± s.d. (n = 2 – 4). Each amino acid was tested at a concentration of 10 mM, except Ser, Trp, Tyr and Cyi, which were tested at a concentration of 1 mM. The T_m of Pu1068 in the absence of amino acids was 41.3 ± 0.8 °C (mean ± s.d., n = 14).

We next attempted to identify the function of Pu1068 by solving its X-ray crystal structure. Since SBPs bind their cognate ligands with high affinity, they often co-purify and co-crystallise with their ligands if they are available in the expression system. Electron density in crystal structures of sufficiently high resolution (>2 Å) can guide proposals for the structure of the ligand, which can be confirmed using mass spectrometry and functional assays such as DSF or ITC. A major advantage of this approach is that it requires no prior knowledge of the physiological ligand of the SBP. There are numerous examples of the use of X-ray crystallography to identify the ligands of AABPs (Deka et al., 2004; Müller et al., 2005); another notable example of this strategy is a recent functional genomics project that resulted in identification of ten novel ligands for SBPs of the tripartite ATP-independent periplasmic transporter class from X-ray crystal structures (Vetting et al., 2015).

Crystallisation conditions for Pu1068 were identified using sparse-matrix screens. A crystal grew in a drop containing 18% (w/v) PEG 8000, 10% (v/v) PEG 200, 0.1 M Bis-Tris propane pH 9.0 as the precipitant. An optimised crystal grew in a drop containing 24% (w/v) PEG 8000, 5% (v/v) PEG 400, 0.1 M Bis-Tris propane pH 9.0 as the precipitant. Using X-ray diffraction data collected from this crystal, the structure of Pu1068 was solved by molecular replacement at 1.6 Å resolution. Data collection and refinement statistics are given in Table 3.3.

Pu1068 adopted an open conformation in the crystal structure, which was indicative of an absence of bound ligand (Figure 3.10a). There are several possible explanations for the absence of bound ligand in the crystal structure: Pu1068 may bind a metabolite that is not produced by *E. coli* in sufficient amounts under the given culture conditions, the ligand may have unbound during the purification process, or Pu1068 may be an enzyme rather than a binding protein.

Table 3.3. Data collection and refinement statistics for Pu1068.

Structure	Pu1068 (*apo*)	Pu1068/NDSB-221
PDB code	5HMT	5KKW
Data collection		
Wavelength (Å)	1.0332	0.9537
Space group	$P2_12_12_1$	$P4_3$
Cell dimensions		
a, *b*, *c* (Å)	38.73, 65.87, 90.97	77.69, 77.69, 44.80
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution range (Å)	37.43 – 1.57 (1.59 – 1.57)	38.84 – 1.88 (1.92 – 1.88)
R_{merge} (%)	12.9 (167.8)	7.8 (58.1)
$CC_{1/2}$ (%)	99.8 (56.6)	99.8 (67.7)
$I / \sigma(I)$	12.7 (1.9)	17.2 (3.0)
Completeness (%)	97.3 (96.6)	95.7 (67.7)
Multiplicity	11.5 (10.8)	7.4 (6.2)
Refinement		
Resolution range (Å)	37.43 – 1.57	38.86 – 1.88
Number of reflections	30761	19884
R_{work}/R_{free} (%)	18.25/21.30	20.50/23.76
No. of atoms		
Protein	1885	1858
Water	207	95
Ligand	–	14
Sulfate	–	5
Average B factors (Å^2)		
Protein	14.96	28.84
Water	23.50	29.92
Ligand	–	43.85
Sulfate	–	73.61
R.m.s. deviations		
Bond lengths (Å)	0.0249	0.0196
Bond angles (°)	2.20	1.92
Ramachandran† (%)		
Favoured	97.9	95.8
Allowed	2.1	4.2
Disallowed	0	0

*Values in parentheses refer to highest resolution shell.

†From PDB validation report.

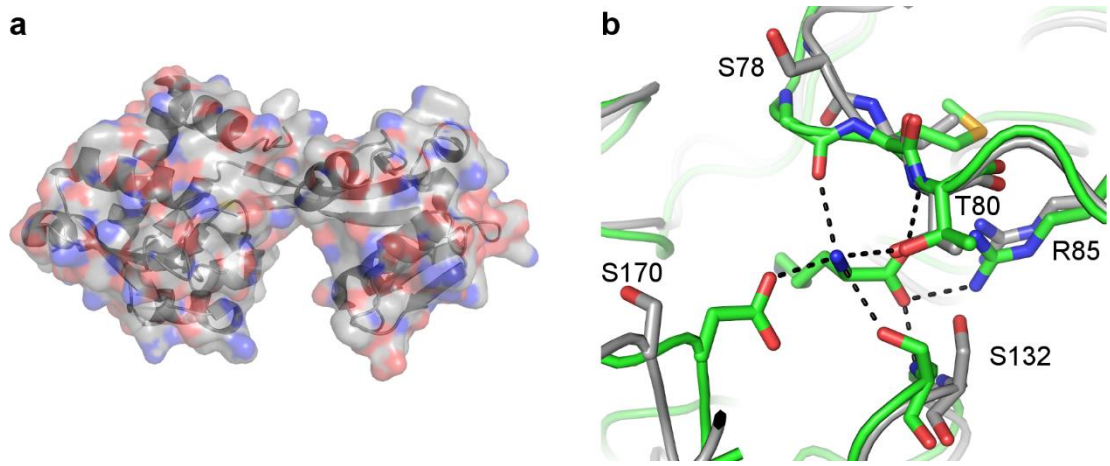


Figure 3.10. Crystal structure of Pu1068. (**a**) Crystal structure of Pu1068. The protein adopts an open conformation with a large cavity between the two domains. (**b**) Comparison of the binding sites of Pu1068 (grey) and Ws0279, bound to L-lysine (green). The closed conformation of Pu1068 was modelled by superimposition of the two domains separately onto the structure of Ws0279. Two major structural differences explain the inability of Pu1068 to bind amino acids: the substitution D170S and the deletion I76Δ, which reorients the carbonyl group of Ser78, both remove conserved interactions with the amino group of the bound amino acid.

Comparison of the binding sites of Ws0279 and Pu1068 suggests that there are two major structural differences that account for the inability of Pu1068 to bind amino acids (Figure 3.10b): the substitution D170S and the deletion I76Δ, which repositions the carbonyl group of Ser78 in a binding site loop, both remove conserved interactions with the amino group of the bound amino acid[2]. On the other hand, interactions between Pu1068 and the carboxylic acid group, *via* Thr80, Arg85, and Ser132, appear to be conserved, suggesting that Pu1068 has retained the ability to bind carboxylic acids. The structure of Pu1068 also has a number of implications for the evolution of CDT activity, which are explored in Section 6.2.3.

[2] Throughout this thesis, residues in extant CDT homologues (Ws0279, Pu1068, Ea1174 and PaCDT) are numbered according to the equivalent position in the ancestral proteins.

Continuing under the assumption that Pu1068 is an SBP, the protein was screened for binding against two libraries of small molecules by DSF; these experiments were done by the CSIRO Collaborative Crystallisation Centre. Firstly, Pu1068 was screened against the Silver Bullets Bio Screen, a commercial screen containing 260 biologically relevant molecules, including amino acids, cofactors, nucleotides, sugars, enzyme inhibitors, and metabolic intermediates in a redundant 96-well plate format (i.e., each molecule appears in multiple wells). No compounds in the Silver Bullets Bio Screen significantly increased the thermostability of Pu1068 under the given conditions (Figure 3.11a), although no melting transitions could be observed for 17 conditions, including the six conditions containing fluorescent compounds; thus, use of an alternative screen or further optimisation of the protocol for use of the Silver Bullets Bio screen in a DSF format would be beneficial.

Next, the protein was screened for increased stability against a subset of the Hampton Solubility & Stability Screen, which contains compounds intended to increase the stability of proteins, including some non-biological compounds (Figure 3.11b). Large increases in the T_m of Pu1068 were observed in the presence of the sulfobetaines NDSB-221 (3-(1-methylpiperidinium-1-yl)propane-1-sulfonate) and NDSB-256 (3-(benzyldimethylammonio)propane-1-sulfonate), which gave ΔT_m values of 7.5 °C and 4.4 °C, respectively, at the highest concentration. Smaller increases in T_m were also observed in the presence of dicarboxylates such as sodium succinate (ΔT_m = 4.1 °C at the highest concentration), sodium malonate (3.6 °C), and DL-malic acid (3.6 °C), supporting the hypothesis that Pu1068 retained the ability to bind carboxylates. However, since the screen contains very high concentrations of these molecules (10–400 mM), the modest T_m increases observed with dicarboxylates are not indicative of high-affinity binding.

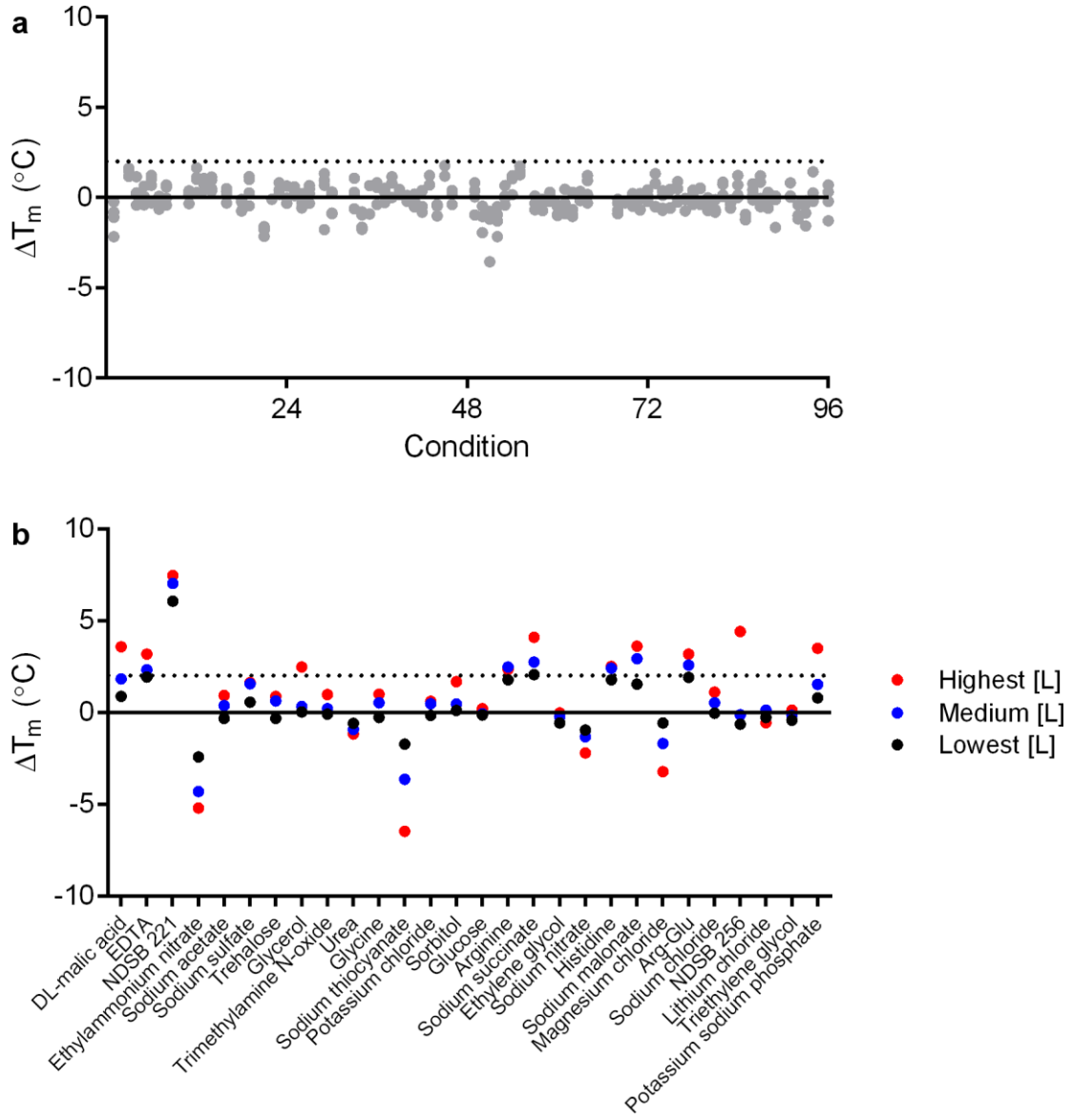


Figure 3.11. DSF screening of Pu1068 against small molecule libraries. (**a**) Silver Bullets Bio Screen. ΔT_m for each condition compared to the median T_m for all conditions (41.6 °C) is shown. (**b**) Hampton Solubility & Stability Screen. ΔT_m for each condition compared to the T_m of a protein-only control (39.7 ± 0.2 °C, mean ± s.d.) is shown. The concentration of each ligand used in the Hampton Solubility & Stability screen is given in Appendix I.

We suspected that the non-biological sulfobetaines that increased the thermostability of Pu1068 in the DSF assay might mimic the physiological ligand of the protein. To first provide evidence that NDSB-221 binds specifically to Pu1068, rather than non-specifically stabilising the protein, the interaction was characterised by intrinsic

tryptophan fluorescence spectroscopy. NDSB-221 elicited an increase in the fluorescence of Pu1068 of up to 20%, and titration of the protein with the ligand gave a K_d of 0.53 mM for the interaction (Figure 3.12). The large increase in the tryptophan fluorescence of Pu1068 upon addition of NDSB-221, indicative of a substantial change in the microenvironment of tryptophan residues in the protein, supports the hypothesis that NDSB-221 is a specific ligand of Pu1068; the binding site contains two tryptophan residues (Trp22 and Trp60) that are solvent-exposed in the unliganded state and are the most likely source of the increase in tryptophan fluorescence upon binding of the sulfobetaine.

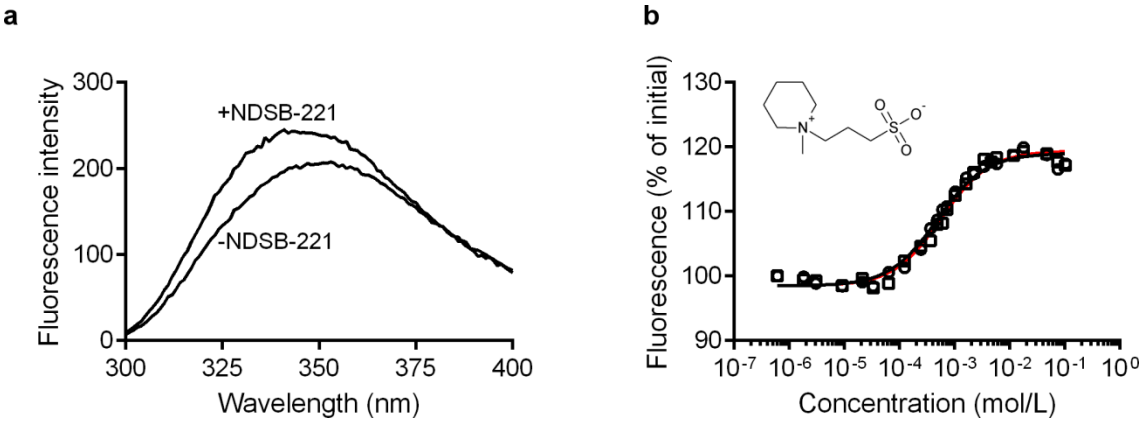


Figure 3.12. Characterisation of the interaction between Pu1068 and NDSB-221 by fluorescence spectroscopy. (**a**) Fluorescence spectrum of Pu1068 in the presence and absence of 10 mM NDSB-221, with an excitation wavelength of 280 nm. (**b**) Fluorescence titration of Pu1068 with NDSB-221: peak fluorescence is plotted against the concentration of ligand. Two separate titrations are shown. Fitting the data to a Boltzmann function gives a K_d of 0.53 mM and a maximum fluorescence change of 20%. The structure of NDSB-221 is inset.

Computational metabolite docking has emerged as a useful strategy for the functional annotation of enzymes (Jacobson et al., 2014; Zhao et al., 2013). This method can also be applied to identify the physiological ligand of an SBP, provided that an accurate model of the closed conformation adopted in the presence of the ligand is available. Therefore, we next solved the crystal structure of the Pu1068/NDSB-221 complex to obtain a model of the ligand-bound conformation of the protein that might be suitable for metabolite docking. New conditions for the co-crystallisation of Pu1068 and NDSB-221 were identified using sparse-matrix screens. Crystals were obtained using 0.1 M Tris pH 8.5, 0.2 M lithium sulfate, 30% PEG 4000 as the precipitant. Optimisation of this condition yielded the crystal used for structure determination, which grew in a hanging drop containing 0.1 M MES pH 6.5, 0.1 M lithium sulfate, 27% PEG 3350 as the precipitant and diffracted to ~1.9 Å. This crystal belonged to space group $P4_3$ and was not isomorphous with the crystal of unliganded Pu1068 (space group $P2_12_12_1$). The structure of the Pu1068/NDSB-221 complex was solved by molecular replacement using the two domains of the unliganded Pu1068 structure as separate search models. Data collection and refinement statistics are given in Table 3.3 (p. 113).

The crystal structure of the Pu1068/NDSB-221 complex shows that the protein remains in an open conformation when bound to NDSB-221 (Figure 3.13a). NDSB-221 binds at the expected binding site of Pu1068, but interacts only with the large domain and remains solvent-exposed, probably because this low-affinity ligand is unable to induce closure of the protein (Figure 3.13b). The sulfonate group of the ligand forms interactions with Thr80 and Arg85; analogous interactions with the carboxylate group of α-amino acids are seen in AABPs. As expected from the fluorescence spectroscopy data, the ligand also forms hydrophobic and cation-π interactions with Trp22 and Trp60. Although the electron density surrounding the piperazine ring of the ligand is ambiguous, possibly

reflecting partial occupancy or multiple binding modes, the identity of the ligand is established by strong electron density for the heavy sulfur atom (Figure 3.13b).

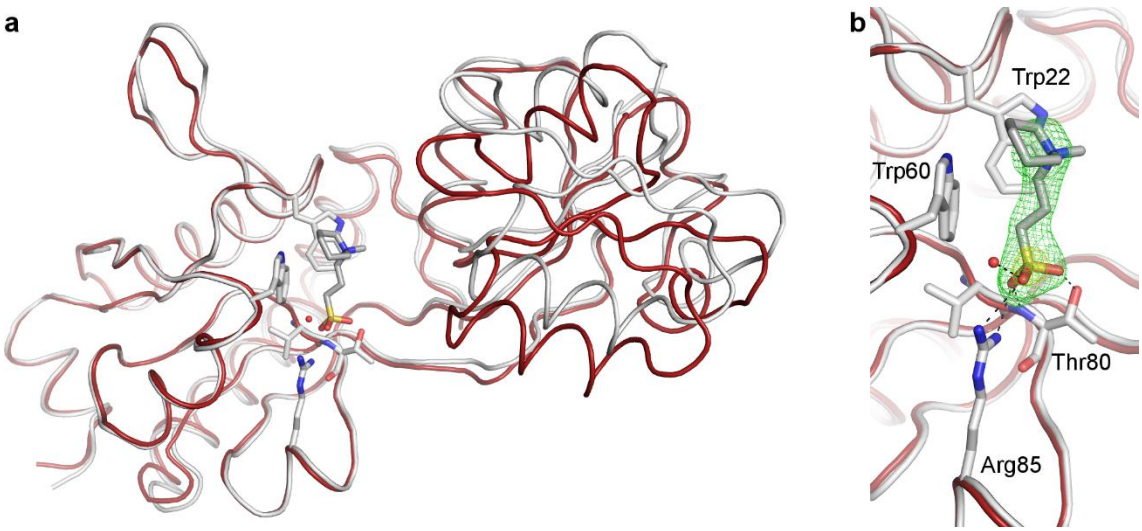


Figure 3.13. Crystal structure of the Pu1068/NDSB-221 complex. (**a**) Comparison of the structures of unliganded Pu1068 (red) and the Pu1068/NDSB-221 complex (grey). (**b**) Binding mode of NDSB-221. Electron density for the ligand is shown by an $mF_o - dF_c$ omit map contoured at +3σ (green) and +10σ (yellow).

Having shown *via* fluorescence spectroscopy and X-ray crystallography that NDSB-221 is a specific ligand of Pu1068, and considering the possibility that this sulfobetaine mimics the physiological ligand of Pu1068, we evaluated binding of a small number of biologically relevant osmolytes with betaine-like and carboxylic acid or sulfonic acid functionalities (glycine betaine, DL-carnitine, γ-butyrobetaine, taurine and dimethylsulfoniopropionate) using a combination of DSF and fluorescence spectroscopy. However, no evidence for binding of these molecules to Pu1068, even with low affinity, was obtained.

3.3 Discussion

Overall, reconstruction and experimental characterisation of intermediates in the evolution of CDT, reinforced by functional characterisation of extant homologues of the enzyme, supported the conclusion that CDT evolved from a cationic amino acid-binding protein, most likely *via* an intermediate SBP of unknown function. AncCDT-1 and Ws0279 are both cationic amino-acid binding proteins, albeit with different binding specificities: Ws0279 exhibits specificity towards L-lysine, whereas AncCDT-1 additionally has high affinity for L -arginine, L -histidine and L -ornithine. Loss of affinity towards amino acids appears to have occurred on the branch of the CDT phylogeny joining AncCDT-1 and AncCDT-2; AncCDT-2 and its extant descendant Pu1068 do not display affinity towards amino acids. The observation that AncCDT-3, Ea1174 and PaCDT, but not AncCDT-2 and Pu1068, have CDT activity is consistent with the evolution of CDT activity having occurred on the branch of the CDT phylogeny joining AncCDT-2 and AncCDT-3. The functions of the intermediate proteins, AncCDT-2 and Pu1068, remain unresolved.

Although the physiological ligand of Pu1068 has not yet been identified, some general conclusions are supported by the experimental data. Firstly, the structural motif that interacts with the carboxylate group of amino acid ligands in AABPs is also conserved in Pu1068, and the protein is stabilised by high concentrations of carboxylates and sulfonates (which are analogous), together suggesting that the physiological ligand of the protein is a carboxylate. Thus, focussed metabolite libraries containing high proportions of carboxylates, which have been developed for the functional annotation of SBPs of the tripartite ATP-independent periplasmic transporter family (Vetting et al., 2015), may also be useful for the functional annotation of Pu1068. Secondly, Pu1068 binds NDSB-221 with low affinity, and this molecule could have some structural

similarity with the physiological ligand of the protein. Although the sulfobetaine functionality of NDSB-221 was strongly suggestive of osmolytes that are abundant in the marine environment, such as glycine betaine and dimethylsulfoniopropionate (Reisch et al., 2011; Sleator and Hill, 2002), binding of these molecules to Pu1068 was not detected; therefore, other classes of compounds should be considered in future experiments. Thirdly, the fact that Pu1068 could not be co-purified with its ligand suggests that the ligand is not likely to be present in the *E. coli* metabolome, although dissociation of the Pu1068-ligand complex during purification is an alternative possibility. Given that the secondary metabolic capabilities of "*Ca.* P. ubique" are extremely limited due to its streamlined genome (Giovannoni et al., 2005), absence of the ligand of Pu1068 from the *E. coli* metabolome would place a significant constraint on the identity of this ligand.

Consideration of the biological context of Pu1068 could also assist with the functional characterisation of this protein. Metaproteomic analysis showed that Pu1068 has relatively low abundance compared with other transport-associated proteins in SAR11 bacteria (Sowell et al., 2009). However, *Pu1068* transcripts were significantly enriched in the SAR11 metatranscriptome following addition of high-molecular weight dissolved organic matter, whereas transporters in general were underrepresented under the same conditions (Sharma et al., 2014). Other changes in gene expression, particularly the enrichment of the formate-tetrahydrofolate ligase gene, suggested that nitrogen acquisition and energy production stimulated by the addition of high-molecular weight dissolved organic matter resulted from flux of methylated nitrogen compounds through one-carbon oxidation pathways. Given the differential expression of *Pu1068* under these conditions, and the co-transcription of *Pu1068* with a homologue of sarcosine oxidase, *soxBDAG*, which would also be involved in one-carbon oxidation of *N*-methylamines (Sun et al., 2011), it was suggested that Pu1068 might be involved in the uptake of methylated nitrogen compounds (Sharma et al., 2014). *N*-methyl-L-glutamate, which

contains both *N*-methylamine and carboxylate functionalities and has been proposed as a potential substrate of SoxBDAG (Sun et al., 2011), is one conspicuous possibility for the physiological ligand of Pu1068. Continuing research into the metabolism, physiology and nutritional requirements of SAR11 bacteria (Carini et al., 2013, 2014; Sun et al., 2011; Tripp, 2013) will provide further information relevant to the biological role of Pu1068.

Although the function of AncCDT-2 has not been confirmed, this ancestral protein, corresponding to the LCA of Pu1068 and PaCDT, most likely had a similar function to Pu1068, given that the reconstructed ancestral protein did not exhibit CDT activity or affinity towards amino acids. Although this phenotype could have resulted from errors in the reconstruction of the sequence of AncCDT-2, the hypothesis that AncCDT-2 had the same function as Pu1068 is nevertheless more plausible given the phylogeny of CDT. The proposition that AncCDT-2 had CDT activity implies the evolution of an SBP (Pu1068) from a CDT, which is unlikely because the association between CDT and the transmembrane component of the ancestral SBP-dependent ABC transporter would have degenerated after the change in function from solute transport to catalysis. Furthermore, directed evolution experiments (Section 4.2.2) showed that a considerable number of substitutions are required to recapitulate CDT activity in AncCDT-2, including substitutions that are not represented in the posterior probability distribution of AncCDT-2 (i.e. not including substitutions towards alternative plausible reconstructions of AncCDT-2). On the other hand, the proposition that AncCDT-2 was an amino acid-binding protein, similar to AncCDT-1, would imply convergent evolution of a number of structural features shared by Pu1068 and PaCDT (discussed in Section 6.2.3), and is therefore unparsimonious. Therefore, the most likely evolutionary scenario given the existing experimental data is that CDT evolved from a cationic amino acid-binding protein *via* an intermediate with the same function as Pu1068, making this ancestral protein a functional intermediate between AABPs and CDT.

3.4 Materials and methods

3.4.1 Materials

pDOTS7 is a derivative of pQE-82L (QIAGEN) modified to enable Golden Gate cloning (Engler et al., 2008), and was created by removal of the *Sap*I site from pQE-82L and introduction of two reciprocal *Sap*I sites following the His_6 tag, with the *Sap*I sites separated by a 28 bp stuffer fragment. This vector was obtained from Prof. Harald Janovjak (IST Austria). Codon-optimised synthetic genes encoding Ws0279 (UniProt: Q7MAG0; residues 24–258) and Pu1068 (UniProt: Q4FLR5; residues 19–255), cloned into the *Sap*I site of the pDOTS7 vector using the Golden Gate method, were obtained from Joe Kaczmarski.

3.4.2 Phylogenetics and ancestral protein reconstruction

The protein sequences of 113 homologues of Ws0279 and PaCDT were collected from the NCBI reference sequence database using the BLAST server. The sequences were aligned in MUSCLE (Edgar, 2004), and the alignment was edited to remove N-terminal signal peptides and large insertions. The resulting alignment was combined with a subset of a previous alignment of representative AABP sequences (Section 2.4.1) by profile-profile alignment in MUSCLE. The final alignment contained sequences of 113 CDT homologues and 271 outgroup AABPs. Phylogenetic analysis was done using the ML method implemented in PhyML (Guindon et al., 2010). Evaluation of BIONJ trees reconstructed using different amino acid substitution models, using the Akaike information criterion as implemented in ProtTest (Abascal et al., 2005), supported the use of the WAG substitution matrix with gamma-distributed rate heterogeneity, a fixed proportion of invariant sites, and equilibrium amino acid frequencies estimated from the data (WAG+I+Γ+F model). Phylogenies were reconstructed in PhyML by optimisation of an initial BIONJ tree by the nearest-neighbour interchange and subtree pruning and

regrafting algorithms. Robustness of the resulting tree topology to the substitution model was assessed by repeating the analysis using the LG and JTT substitution matrices (LG/JTT+I+Γ+F models), and convergence to the ML tree was checked by repeating the analyses with ten randomised initial trees. Although the resulting trees had essentially identical topologies, the tree inferred using the LG+I+Γ+F model had the highest likelihood and was therefore taken as the ML tree. Reconstruction of ancestral protein sequences was performed using the empirical Bayes method implemented in PAML (Yang, 2007).

3.4.3 Protein expression and purification

Ws0279 and AncCDT-1 were expressed in *E. coli* BL21(DE3) cells grown in Luria-Bertani (LB) medium (Ws0279) or Terrific Broth (TB) medium (AncCDT-1) supplemented with 100 mg/L ampicillin to OD_{600} ~0.7 at 37 °C, induced with 1 mM IPTG, and incubated for a further 20 h at 37 °C. Pu1068 was expressed in *E. coli* BL21(DE3) cells grown in auto-induction media (Section 2.7.4) at 37 °C for 24 h. The proteins were purified by Ni-NTA affinity chromatography under native conditions, followed by SEC, as described in Section 2.7.5.

3.4.4 Differential scanning fluorimetry

Differential scanning fluorimetry (DSF) experiments to test Ws0279, AncCDT-1 and Pu1068 for amino acid binding were done using a ViiA 7 real-time PCR instrument (Thermo Scientific). 100 mM aqueous amino acid solutions were stored at –20 °C. Reaction mixtures contained 5 μM protein in DSF buffer, 5× SYPRO orange dye (Sigma-Aldrich) and 1 mM or 10 mM ligand in a total volume of 20 μL, and were dispensed onto a 384-well PCR plate, at least in triplicate. At least ten replicates of ligand-free control were also included on each plate. Fluorescence intensities were monitored continuously as the samples were heated from 20 °C to 99 °C at a rate of 0.05 °C/s, with excitation at

580 nm and emission at 623 nm. Melting temperatures were determined by fitting the data to a Boltzmann function (Eq. 3.1), with the parameters *A* and *C* (accounting for the slopes of the pre- and post-transition baselines) fixed at zero if possible.

$$y = Ax + B + \frac{Cx + D}{1 + \exp(\frac{T_m - x}{E})} \quad \textbf{(3.1)}$$

Pu1068 was screened for thermostability against the Silver Bullets Bio Screen (Hampton Research) and a subset of the Solubility and Stability Screen (Hampton Research) by the CSIRO Collaborative Crystallisation Centre (www.csiro.au/C3), Melbourne, Australia. For the Silver Bullets Bio Screen, reaction mixtures containing 0.1 μg Pu1068, 1.875× SYPRO orange, and 2 μL ligand mixture in a total volume of 10 μL were dispensed onto a 384-well PCR plate in quadruplicate. For the Solubility and Stability Screen, the reaction mixtures contained 0.3 μg Pu1068, 3.75× SYPRO orange and 5 μL ligand in a total volume of 20 μL, in a 96-well plate format; each ligand was tested at three concentrations and three replicates of a ligand-free control were also included. Fluorescence intensities were measured on a BioRad CFX384 real-time PCR instrument with excitation at 490 nm and emission at 570 nm. The temperature was ramped from 20 °C to 100 °C at a rate of 0.05 °C/s, and the fluorescence intensity was measured at 0.5 °C intervals. Melting temperatures were taken as the temperature at the minimum of the first derivative of the melt curve, which was determined by fitting the data to a quadratic function in the vicinity of the melting temperature using GraphPad Prism 7 or Meltdown software (Rosa et al., 2015).

3.4.5 Crystallisation and structure determination of Pu1068.

Initial crystallisation conditions for *apo* Pu1068 were identified from the Crystal Screen, Index, PEGRx and PEG/Ion high-throughput crystal screens (Hampton Research); sitting drops containing 0.2 μL protein (18 mg/mL in 10 mM Tris pH 8.0, 100 mM NaCl) and

0.2 μL precipitant were prepared in 96 well plates using a Cartesian Honeybee instrument. One crystal grew in a drop containing 18% (w/v) PEG 8000, 10% (v/v) PEG 200, 0.1 M Bis-Tris propane pH 9.0 as the precipitant. The crystal used for data collection grew from a hanging drop at 18 °C containing 1.5 μL protein (18 mg/mL Pu1068 in 10 mM Tris pH 8.0, 100 mM NaCl, 10% glycerol) and 1.5 μL 24% (w/v) PEG 8000, 5% (v/v) PEG 400, 0.1 M Bis-Tris propane pH 9.0 as the precipitant. The crystal was flash frozen in a nitrogen stream at 100 K without cryoprotectant. Diffraction data were collected at 100 K on the MX2 beamline at the Australian Synchrotron at a wavelength of 1.0332 Å. The diffraction data were indexed and integrated in iMOSFLM (Battye et al., 2011), and scaled in Aimless in the CCP4 package (Winn et al., 2011). The structure was solved by molecular replacement in Phaser (McCoy et al., 2007), using the two domains of PaCDT (PDB: 3KBR) as separate search models. Residues 27–122 and 221–258 of PaCDT were taken as the large domain, and residues 123–220 were taken as the small domain. The model of Pu1068 was built manually in Coot (Emsley et al., 2010) and refined by iterative reciprocal space-real space refinement in REFMAC5 (Murshudov et al., 1997) and Coot.

Crystallisation conditions for the Pu1068/NDSB-221 complex were identified using the SG1 sparse-matrix screen (Molecular Dimensions) by preparing sitting drops containing 1 μL protein (24 mg/mL Pu1068 in 10 mM Tris pH 8.0, 50 mM NaCl, 10 mM NDSB-221) and 1 μL precipitant. Crystals were obtained using 0.1 M Tris pH 8.5, 0.2 M Li_2SO_4, 30% PEG 4000 as the precipitant. Optimisation of this crystal condition yielded the crystal used for structure determination, which grew from a hanging drop at 18 °C containing 1 μL protein and 1 μL 0.1 M MES pH 6.5, 0.1 M Li_2SO_4, 27% (w/v) PEG 3350 as the precipitant. The crystal was flash frozen without cryoprotection in a nitrogen stream at 100 K. X-ray diffraction data was collected at 100 K on the MX1 beamline of the Australian Synchrotron at a wavelength of 0.9537 Å. The data were indexed and integrated in XDS (Kabsch, 2010), and scaled in Aimless (Winn et al., 2011). The

structure of the Pu1068/NDSB-221 complex was solved by molecular replacement in Phaser (McCoy et al., 2007), using the two domains of the *apo* Pu1068 structure as separate search models. The structure was refined by iterative reciprocal space-real space refinement in REFMAC5 (Murshudov et al., 1997) and Coot (Emsley et al., 2010). Geometric restraints for NDSB-221 were generated using eLBOW in the Phenix package (Adams et al., 2010). Translation-libration-screw parameters for one group of atoms were included in the final round of reciprocal space refinement in REFMAC5. The coordinates and structure factors for the crystal structures of Pu1068 have been deposited in the PDB under accession codes 5HMT (*apo*) and 5KKW (NDSB-221 complex).

3.4.6 Intrinsic tryptophan fluorescence spectroscopy

Intrinsic tryptophan fluorescence spectra were recorded using a Cary Eclipse fluorescence spectrophotometer. Protein samples were prepared at a concentration of 5 μM in DSF buffer. The excitation wavelength was 280 nm, and emission was measured between 300 nm and 400 nm. Following addition of each ligand aliquot, the sample was incubated at ambient temperature for 1 min before the fluorescence spectrum was recorded.

Chapter Four

Evolution of an enzyme from a solute-binding protein. Part II: Genetics.

4.1 Summary

The aim of the work described in this chapter was to identify the amino acid substitutions required for the gain of CDT activity in ancestral SBPs. Initial attempts to identify the substitutions required to introduce catalytic activity into AncCDT-2 by site-directed mutagenesis, based on molecular docking and MD simulations of the PaCDT-arogenate complex and analysis of sequence conservation in CDT homologues, were unsuccessful. However, directed evolution of the mutagenised AncCDT-2 gene using incorporation of synthetic oligonucleotides *via* gene reassembly (ISOR) did yield variants with CDT activity. Two further rounds of ISOR were used to remove unnecessary substitutions from these AncCDT-2 variants, culminating in the identification of a CDT variant with only six substitutions relative to AncCDT-2. Furthermore, recombination of AncCDT-2D2 and AncCDT-3 using the staggered extension process (StEP) yielded variants of AncCDT-3 with higher catalytic activity.

4.2 Results

4.2.1 Mutational basis for CDT evolution: site-directed mutagenesis

Having shown the plausibility of an evolutionary transition from a cationic amino acid-binding protein to a CDT, our next goal was to identify the specific amino acid substitutions that were necessary for this functional transition. We considered this problem in two parts, attempting to identify the substitutions necessary to introduce catalytic activity into AncCDT-2 before identifying the remaining substitutions necessary to introduce catalytic activity into AncCDT-1. AncCDT-1 and AncCDT-3 are separated by 77 substitutions, while AncCDT-2 and AncCDT-3 are separated by 36 substitutions. To assess the potential role of each of these substitutions for catalytic activity, and to choose substitutions to introduce by site-directed mutagenesis, the structure of PaCDT and the sequences of CDT homologues were analysed.

The structure of the PaCDT-arogenate complex was modelled by molecular docking in AutoDock Vina (Trott and Olson, 2010), using the crystal structure of PaCDT bound to 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), which shares some fortuitous similarities with the cyclohexadiene substrates of the enzyme and binds at the interface of the two α/β domains (Figure 4.1). The amino acid binding site, which is strictly conserved in all amino acid-binding proteins, is also conserved in PaCDT; therefore, only poses that placed the amino acid moiety of L-arogenate near this site were considered plausible. Two poses with energies of –31.4 kJ/mol and –25.9 kJ/mol fulfilled this criterion (Figure 4.2).

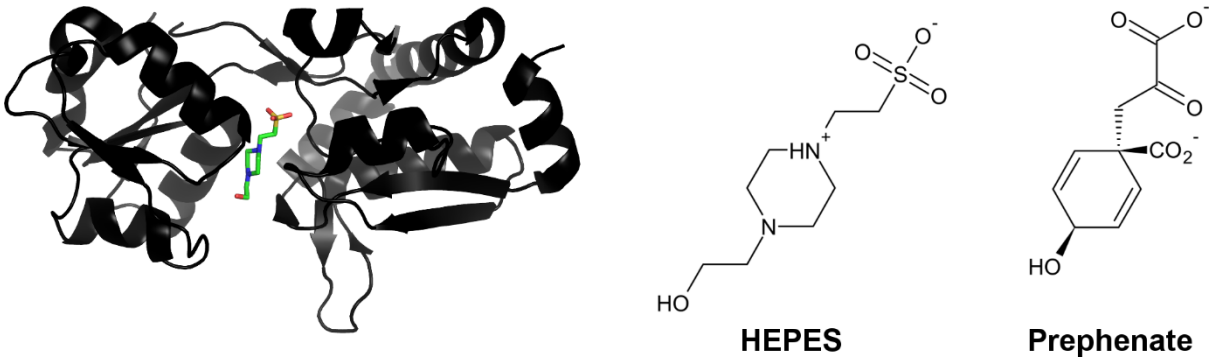


Figure 4.1. Crystal structure of the PaCDT-HEPES complex. HEPES binds at the interface of the two domains of PaCDT (PDB: 3KBR). The structures of HEPES and prephenate are shown for comparison.

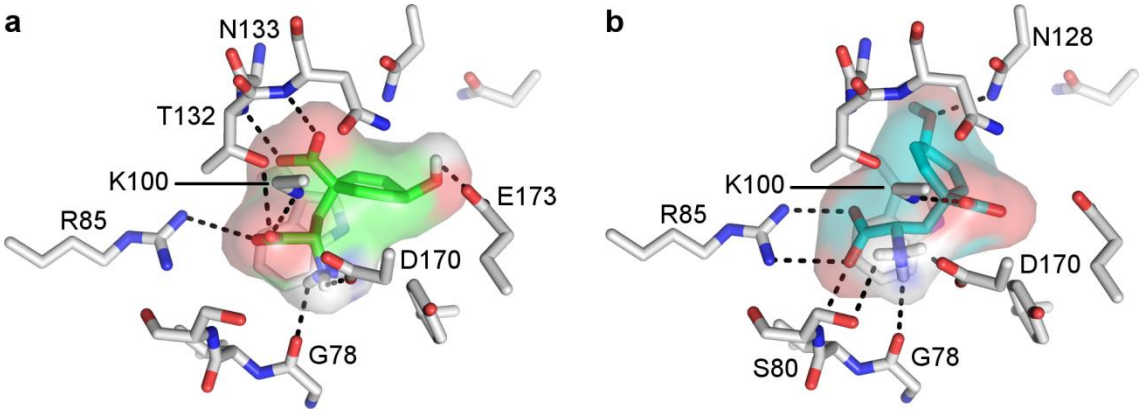


Figure 4.2. Docking of L-arogenate into the active site of PaCDT. Two poses are shown: (**a**) Pose 1: *E* = –31.4 kJ/mol. (**b**) Pose 2: *E* = –25.9 kJ/mol. Residues forming direct polar interactions with the substrate are labelled.

In the lowest energy pose of the PaCDT-arogenate complex (pose 1), Gly78, Arg85, Thr132 and Asp170 form canonical interactions with the amino acid moiety of L-arogenate, and Lys100 forms an additional electrostatic interaction with the carboxylate group. The backbone amide groups of Thr132 and Asn133 interact with the departing carboxyl group. Glu173 is positioned near the departing hydroxyl group; this residue is therefore a candidate for the general acid necessary for catalysis of the elimination reaction. Indeed, the pK_a of Glu173 estimated using PROPKA (Olsson et al., 2011) is 6.57, consistent with this residue being protonated close to physiological pH.

In the other pose of the PaCDT-arogenate complex (pose 2), Gly78, Ser80, Arg85 and Asp170 form canonical interactions with the amino acid group of the substrate, while Lys100 binds the departing carboxyl group and Asn128 binds the hydroxyl group. Overall, pose 2 reproduces the expected interactions between PaCDT and the amino acid moiety of L-arogenate more closely, whereas pose 1 appears to enable stronger interactions between PaCDT and the departing carboxylate group of the substrate, and suggests a plausible catalytic residue.

To test the robustness of the docking results, MD simulations of the PaCDT-arogenate complex were performed. L-arogenate was used for the simulations rather than prephenate because the amino group is already parameterised in the GROMOS 53a6 force field, whereas the ketone group is not. The main advantage of MD simulations over molecular docking in this context is that the assumption that the enzyme is rigid or partially rigid, with flexibility limited to sampling of different side chain rotamers, can be relaxed; in an MD simulation, conformational space around the docked pose can be sampled to optimise interactions with the substrate. This was considered particularly important in the case of CDT, since its active site is located at the interface of two domains connected by a flexible hinge. Indeed, as shown in Sections 5.2.1 and 5.2.2, the

conformation of PaCDT observed in the HEPES-bound structure is substantially different from the conformation of the *apo*-enzyme. In addition, the quality of the force field parameters for the substrate, L-arogenate, could be assured as most of the functional groups in L-arogenate (all except the alkene group) are found in proteinogenic amino acids and could be extracted from existing, validated force fields.

Two 50 ns simulations were initialised from each of the docking poses shown in Figure 4.2. Although PaCDT is trimeric, the enzyme-substrate complex was simulated as an isolated monomer to reduce computational cost. Because no evidence of cooperativity between subunits has been found (Zhao et al., 1992) and the ancestral CDT variants are monomeric (Joe Kaczmarski, unpublished results), indicating that oligomerisation is not required for catalytic activity, simulation of the monomer was considered sufficient for the purposes of identifying interactions between the substrate and enzyme. The quaternary structure of PaCDT and its possible consequences for catalysis are discussed in further detail in Chapter 5.

The substrate remained bound to the enzyme for the duration of the four simulations. The poses of the enzyme-substrate complex that are most frequently observed are similar to pose 1, even for the simulations initialised from pose 2 (Figure 4.3). The amino acid moiety of L-arogenate interacts frequently with Tyr22, Gly78, Ser80 and Arg85, whereas direct interactions between the amino group and Asp170 (a highly conserved residue that binds the amino group in AABPs) are rare. The departing carboxylate group of L-arogenate is persistently bound by a set of backbone amide groups in an active site loop (Gly131, Thr132, Asn133) and the side-chain of Asn133. Asn128 and Lys100 also interact occasionally with the departing carboxylate group. Fewer interactions between PaCDT and the departing hydroxyl group of L-arogenate are observed; Asn128 and Asn152 appear to interact with the departing hydroxyl group

primarily *via* hydrogen bonding between the amide side-chain and the oxygen atom of the hydroxyl group. Direct interactions between Glu173 and L-arogenate were infrequent, which likely reflects the inability of basic MD simulations to capture protonation dynamics of residues with pK_a close to the pH of solution[3] and sample geometries close to the transition state of a reaction.

Modelling of the PaCDT-arogenate complex using molecular docking and MD simulations suggested a number of residues in PaCDT that are likely to interact directly with the substrate, enabling substrate binding. In addition, comparison of the models of the PaCDT-arogenate complex with the crystal structure of Ws0279 suggested a second group of residues that indirectly affect the active site of the enzyme by stabilising an alternative conformation of Trp60 (Figure 4.4). In Ws0279, Trp60 forms hydrophobic interactions with the aliphatic chain of the ligand, L-lysine. Although Trp60 is also conserved in PaCDT, this residue adopts a different conformation in which packing against the cyclohexadiene ring of the substrate is possible. Thus, the conformational change of Trp60 reshapes the active site of PaCDT and enables binding of the larger cyclohexadiene substrates. Surrounding residues, including Thr19 and Asp21, generate a hydrogen bonding network that appears to stabilise the new conformation of Trp60.

[3] Simulations of the PaCDT-arogenate complex with Glu173 in the protonated state were also attempted, but these simulations showed immediate dissociation of the substrate from the enzyme.

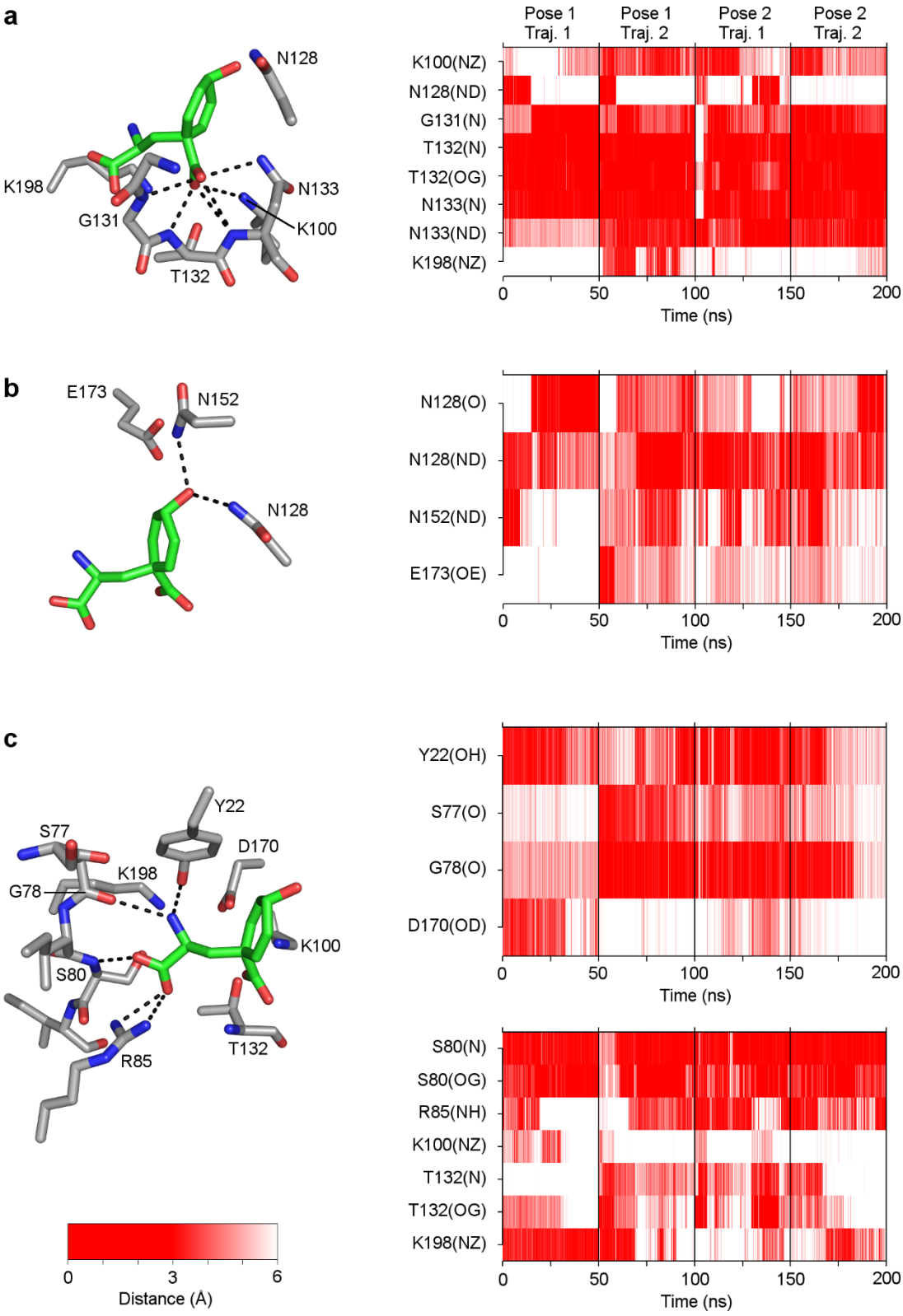

a
N128
K198
N133
G131
K100
T132
Pose 1 Traj. 1
Pose 1 Traj. 2
Pose 2 Traj. 1
Pose 2 Traj. 2
K100(NZ)
N128(ND)
G131(N)
T132(N)
T132(OG)
N133(N)
N133(ND)
K198(NZ)
0
50
100
150
200
Time (ns)
b
E173
N152
N128
N128(O)
N128(ND)
N152(ND)
E173(OE)
0
50
100
150
200
Time (ns)
c
Y22
S77
D170
K198
G78
K100
S80
T132
R85
0
3
6
Distance (Å)
Y22(OH)
S77(O)
G78(O)
D170(OD)
0
50
100
150
200
Time (ns)
S80(N)
S80(OG)
R85(NH)
K100(NZ)
T132(N)
T132(OG)
K198(NZ)
0
50
100
150
200
Time (ns)

Figure 4.3. Persistence of interactions between PaCDT and L-arogenate during MD simulations. (Previous page) Left: typical interactions between PaCDT and the (**a**) departing carboxylate group, (**b**) departing hydroxyl group, and (**c**) amino acid moiety of L-arogenate during MD simulations. The same snapshot of the PaCDT-arogenate complex is shown in each panel. Right: heat maps illustrating the distance between the indicated protein atoms and the (**a**) departing carboxylate group, (**b**) departing hydroxyl group, and (**c**) amino group and carboxylate group of the amino acid moiety of L-arogenate over a total 200 ns of simulation time. Distances less than 3 Å are shown in red.

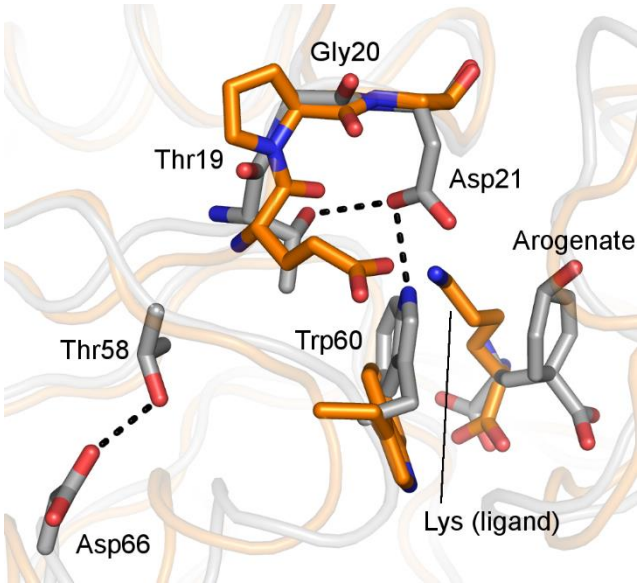


Figure 4.4. Trp60 and surrounding residues reshape the active site of PaCDT. Grey: Model of the PaCDT-arogenate complex (as in Figure 4.3). Orange: Crystal structure of Ws0279 bound to L-lysine (PDB: 3K4U). Residues in PaCDT are labelled.

The sequence conservation of CDT homologues was analysed to identify residues that became highly conserved after the divergence of AncCDT-3 (Figure 4.5). Unsurprisingly, residues in the active site of CDT, such as Lys100, Asn128, Asn133 and Glu173, which make frequent interactions with the substrate *via* their side-chains, are highly conserved. Other highly conserved residues include Gly20, Asp66 and Gly131. Gly131 interacts with the departing carboxylate group of L-arogenate *via* its backbone amide group. Alternative residues at this position would likely clash with the substrate, and the conformational flexibility of glycine may also be necessary for the flexibility of this active site loop. Gly20 and Asp66 have possible structural roles in supporting the new conformation of Trp60 (Figure 4.4). Gly20 adopts a conformation less favoured for non-glycine residues ($\varphi = 67.7°$, $\psi = 17.2°$), which is necessary to accommodate Thr19 in a conformation that enables a hydrogen-bonding network between Thr19 and Trp60. Asp66, which interacts with Thr58 in the crystal structure of PaCDT *via* its side chain, may stabilise the active site loop between these two residues.

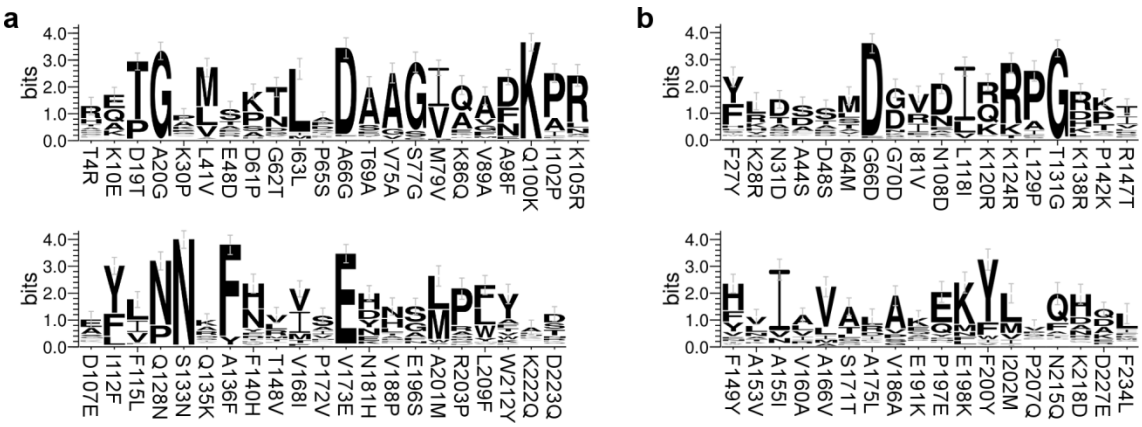


Figure 4.5. Sequence conservation in CDT homologues. The logos show sequence conservation in proteins that evolved from AncCDT-3, based on the sequence alignment used for phylogenetic analysis. (**a**) Sequence conservation (measured as information content in bits) at positions substituted between AncCDT-1 and AncCDT-2. (**b**) Sequence conservation at positions substituted between AncCDT-2 and AncCDT-3.

Most of the substitutions that appeared to have reshaped and functionalised the active site of PaCDT occurred between AncCDT-1 and 2; few substitutions between AncCDT-2 and 3 are located in the first or second shells of the active site (Figure 4.6). Nonetheless, seven substitutions between AncCDT-2 and 3 were hypothesised to affect CDT activity:

T131G and L129P: Gly131 and Pro129 are located in the active site loop that makes backbone interactions with the departing carboxylate group of the substrate. Gly131 is completely conserved in CDT homologues, and Thr at this position is likely to clash with the departing carboxylate group of the substrate.

G66D: Asp66 appears to have a structural role, as described above. This residue is highly conserved in CDT homologues.

A155I and A166V: Ile155 and Val166 are moderately conserved. These residues are in the second shell of the active site and have a possible structural role; in particular, Ile155 is adjacent to Asn128 and Asn152, and may have a role positioning these residues that are putatively important for catalysis.

P197E and L198K: Glu197 and Lys198 are located in the hinge region of CDT and may affect conformational sampling of the enzyme. In particular, Lys198 projects towards the active site and makes an additional interdomain contact with Asp170, which may increase catalytic activity by stabilising the closed, catalytically competent conformation (see Chapter 5). Additionally, Lys198 may contribute to electrostatic stabilisation of the dianionic substrate prephenate in the active site.

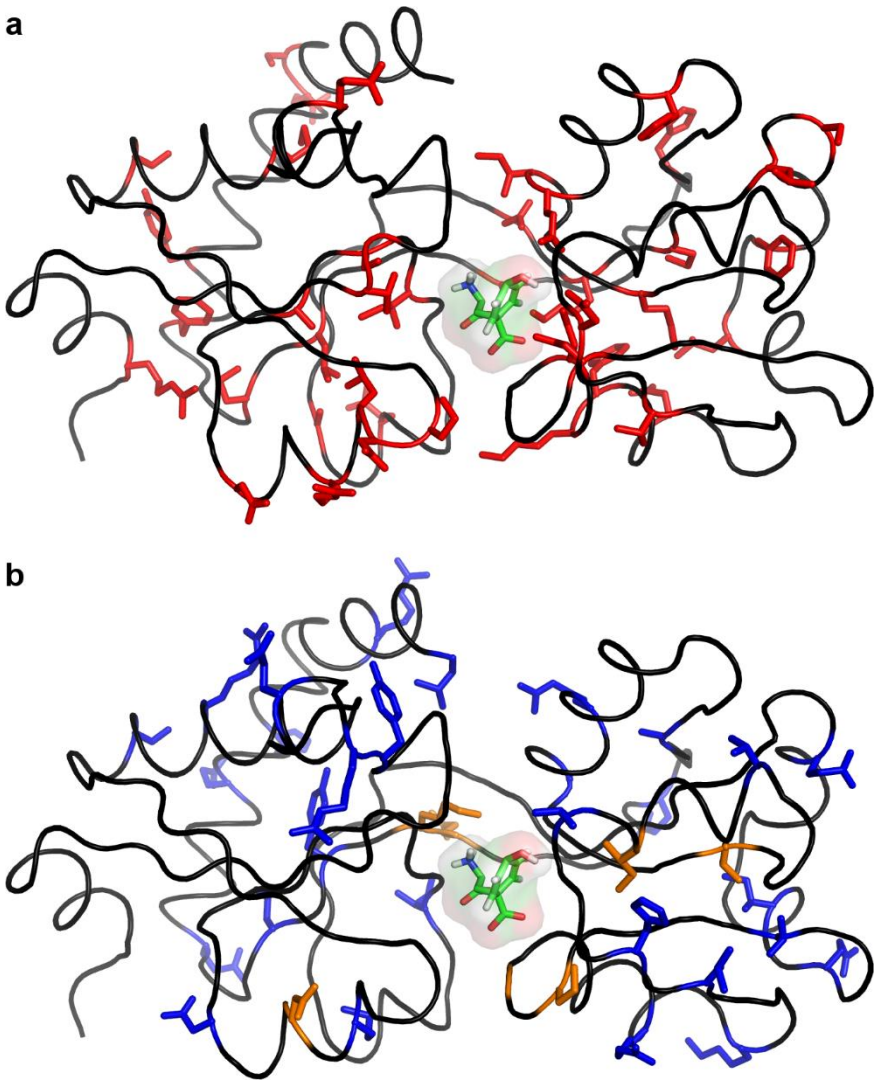


Figure 4.6. Positions of amino acid substitutions between AncCDT-1 and AncCDT-3. Residues that were substituted between AncCDT-1 and AncCDT-3 are shown by sticks at the equivalent position in the PaCDT structure docked with L-arogenate. (**a**) Substitutions between AncCDT-1 and AncCDT-2. (**b**) Substitutions between AncCDT-2 and AncCDT-3. The seven substitutions whose effect was tested by site-directed mutagenesis are shown in orange, while the remaining substitutions are shown in blue.

Using site-directed mutagenesis, these seven substitutions were combined in various permutations in the AncCDT-2 background (Table 4.1). The final variant, AncCDT-2D2, contained all seven substitutions. The CDT activity of each of the variants was assessed by genetic complementation; however, none of the variants possessed detectable CDT activity (visible growth within 8–10 days of incubation in liquid M9–F minimal media at 37 °C). As such, we began to use directed evolution strategies to identify the remaining substitutions necessary to recapitulate CDT activity in AncCDT-2.

Table 4.1. Variants of AncCDT-2 tested for CDT activity by genetic complementation.

Variant	Substitutions						
2A1	T131G						
2A2	T131G	G66D					
2B1	T131G		L129P				
2B2	T131G	G66D	L129P				
2C1	T131G	G66D	L129P			P197E	
2C2	T131G	G66D	L129P				L198K
2C3	T131G	G66D	L129P			P197E	L198K
2C4	T131G	G66D	L129P	A155I			
2C5	T131G	G66D	L129P		A166V		
2C6	T131G	G66D	L129P	A155I	A166V		
2D1	T131G	G66D	L129P	A155I		P197E	L198K
2D2	T131G	G66D	L129P	A155I	A166V	P197E	L198K

4.2.2 Mutational basis for CDT evolution: directed evolution

Two directed evolution strategies, the staggered extension process (StEP) (Zhao et al., 1998) and incorporation of synthetic oligonucleotides *via* gene reassembly (ISOR) (Herman and Tawfik, 2007), were used to identify mutations required for the gain of catalytic activity in AncCDT-2. StEP is a PCR-based method for the recombination of homologous genes. Recombination of two or more template genes is accomplished using a PCR reaction with very short (<5 s) extension time so that only a fragment of the gene is synthesised in each cycle. Between extension steps, the gene fragment can anneal to a different template, which results in recombination. ISOR is achieved by fragmenting a gene using DNAse I, then reassembling the fragments together with mutagenic oligonucleotides, in order to generate a library containing different combinations of targeted, oligonucleotide-encoded mutations. StEP has the advantage of experimental simplicity, whereas ISOR offers the possibility of greater control over library creation. As discussed in Section 3.1.2, both of these strategies are amenable to the directed evolution of CDT variants because the ability of CDT to rescue *E. coli* phenylalanine auxotrophs lacking the *pheA* gene provides a convenient genetic selection method for assessing catalytic activity.

Concurrently with our attempts to identify gain-of-function mutations in AncCDT-2, we used StEP to identify the mutations responsible for the difference in catalytic activity between the two versions of AncCDT-3 (based on the LG and WAG substitution models). As discussed previously (Section 3.2.2), AncCDT-3 and AncCDT-3W differ at 10 residues, and AncCDT-3W enables substantially faster growth of *E. coli* Δ*pheA* cells. A library was created by recombination of AncCDT-3 and AncCDT-3W using StEP, and *E. coli* Δ*pheA* cells transformed with the StEP library were plated on M9 minimal media to identify fast-growing clones. Colonies appeared after incubation for 2 days at 37 °C, as expected based on the growth of AncCDT-3W, and

eight of these clones were sequenced (Table 4.2). Of the ten substitutions between AncCDT-3 and AncCDT-3W, only two were recovered in every sequenced clone: H181D and P188L. Because the P188L substitution was also recovered in fast-growing clones from the AncCDT-2D2/3 StEP library (see below), we presumed that the P188L substitution was responsible for the difference in growth rate between AncCDT-3 and AncCDT-3W, whereas the H181D substitution had hitchhiked due to its proximity to the P188L substitution. Indeed, AncCDT-3(P188L) exhibited a growth rate in the complementation assay similar to AncCDT-3W, with comparable protein expression (Figure 4.7). Subsequently, AncCDT-3(P188L) was shown to possess prephenate dehydratase activity *in vitro*, with a k_{cat} of 4.53×10^{-2} s^{-1} and K_M of 280 μM, compared with k_{cat} of 1.84×10^{1} s^{-1} and K_M of 19 μM for PaCDT (Joe Kaczmarski, unpublished results).

Table 4.2. Composition of StEP libraries. The sequences of clones from the AncCDT-3/3W and AncCDT-2D2/3 StEP libraries, isolated from unselective (LBA) and selective (M9–F) media, are summarised. For each property, the mean value is given, followed by the range in parentheses. n.d. not determined.

Library	**AncCDT-3/3W**	**AncCDT-2D2/3**
Before selection		
Unique sequences obtained	0	8
Template-switching frequency	n.d.	2.3 (0 – 5)
Random mutations	n.d.	0.4 (0 – 2)
Percentage of ancestral states[1]	n.d.	44% (0 – 100%)
After selection		
Library size	10^3	5×10^3
Unique sequences obtained[2]	8	5
Template-switching frequency	1.9 (0 – 3)	1.6 (0 – 4)
Random mutations	0.3 (0 – 1)	1.4 (1 – 3)
Percentage of ancestral states[1]	38% (0 – 70%)	5% (0 – 10%)

[1] Percentage of positions at which the ancestral and derived states differed between the two template genes and the ancestral state was observed. AncCDT-3 was taken as the ancestral sequence in the AncCDT-3/3W library.

[2] In the case of the AncCDT-2D2/3 library, only clones with confirmed growth in liquid M9–F media were considered.

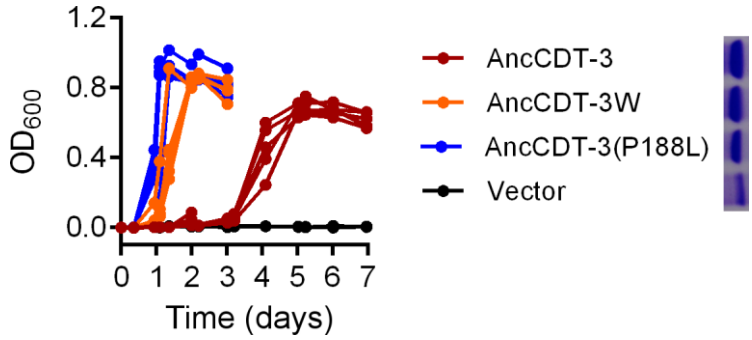


Figure 4.7. AncCDT-3(P188L) supports rapid growth of phenylalanine auxotrophs in minimal media. Five replicate growth curves are shown for *E. coli* Δ*pheA* transformants of AncCDT-3, AncCDT-3W and AncCDT-3(P188L), grown in M9–F media at 37 °C. The P188L substitution in AncCDT-3 is sufficient to recapitulate the faster growth of AncCDT-3W transformants. The SDS PAGE gel fragment shows expression of each protein in the soluble fractions of crude cell lysates of M9+F cultures incubated at 37 °C overnight.

To identify substitutions in AncCDT-2 required for the gain of CDT activity, AncCDT-2D2 (Table 4.1, p. 140) and AncCDT-3 were recombined using StEP. The composition of the resulting library is summarised in Table 4.2 (p. 143). Transformation of *E. coli* Δ*pheA* cells with the StEP library and incubation on selective agar plates yielded colonies within two days. Thus, recombination of AncCDT-2D2 and AncCDT-3 by StEP yielded variants with higher catalytic activity than AncCDT-3 itself, which yields colonies of similar size after four days under the same conditions[4]. Sequencing of individual colonies gave five unique sequences that closely resembled the sequence of AncCDT-3 and were therefore uninformative for identifying the substitutions required for the gain of catalytic activity in AncCDT-2D2 (Table 4.2 and Table 4.3). This bias towards the sequence of AncCDT-3 was not present in the library itself, as shown by sequencing of clones isolated from unselective media (Table 4.2). The clones isolated from the AncCDT-2D2/3 StEP library contained several point mutations (P188L, P188S,

[4] Δ*pheA* cells transformed with the pDOTS7 vector form colonies on M9–F plates after incubation at 37 °C for six days, possibly due to non-enzymatic conversion of prephenate to phenylpyruvate. However, these transformants exhibit no visible growth in M9–F liquid media after incubation at 37 °C for ten days.

V187D, V187A and F98S) that were apparently responsible for the increase in CDT activity relative to AncCDT-3. In particular, the importance of the P188S and V187A substitutions is evident, because these substitutions were isolated in single point mutants of AncCDT-3, and the importance of the P188L substitution was established above.

Table 4.3. Sequences of CDT variants isolated from the AncCDT-2D2/3 StEP library.

Clone ID	Reversions [1]	Point mutations	Growth time (days) [2]	Distance from AncCDT-2 [3]
S1	-	P188S	2	37
S3	S44A, S48D	S3G, P188L, K232R	2	37
S21	-	V187A	3	37
S23	S44A, S48D, K191E	F98S	3	34
S27	M64I, D70G	V187D	3 – 4	35

[1] Sequences are given as variants of AncCDT-3 with reversions to the ancestral state (AncCDT-2).
[2] Time required for the clone to reach OD_{600} of 0.2 in M9–F liquid media at 37 °C, given as a range for two biological replicates.
[3] Number of amino acid substitutions separating the sequence from AncCDT-2.

Because the sequences of CDT variants obtained from the AncCDT-2D2/3 StEP library did not differ substantially from the sequence of AncCDT-3, ISOR was used as an alternative strategy to identify the minimal set of mutations required for the gain of catalytic activity in AncCDT-2. In the first round of ISOR (ISOR-R1), AncCDT-2D2 was used as the template gene, and the remaining 29 substitutions between AncCDT-2 and AncCDT-3 were encoded in mutagenic oligonucleotides. The composition of the library is summarised in Table 4.4, and the sequences of the seven unique clones isolated from selective media are given in Table 4.5.

Surprisingly, none of the oligonucleotide-encoded mutations included in ISOR-R1 were observed in every clone isolated from selective media, indicating that none of the remaining mutations on the AncCDT-2/3 branch were indispensable for catalytic activity in AncCDT-2D2 (Table 4.5). The most common encoded substitutions were I64M (5/7 clones) and N108D (6/7 clones). As in the StEP experiments, ISOR yielded CDT variants with higher catalytic activity than AncCDT-3 itself, apparently due to a number of random point mutations. A substitution at Pro102 (P102L or P102S) was observed in each clone, together with nearby substitutions in some cases (G99S or T101A). Notably, these substitutions are clustered near the substitutions observed in the AncCDT-2D2/3 StEP library (P188L, P188S, V187D, V187A and F98S), at the boundary between the hinge region and small domain of the protein. The clone exhibiting fastest growth in minimal media, CDT-J3, which contained a fortuitous reversion of a mutation introduced by site-directed mutagenesis (A166V), had the fewest substitutions (11) relative to AncCDT-2.

Table 4.4. Composition of ISOR libraries. The sequences of clones from the three ISOR libraries, isolated from unselective (LBA) and selective (M9–F) media, are summarised. For each property, the mean value is given, followed by the range in parentheses. Out of the clones isolated from selective media, only those with confirmed growth in liquid M9–F media were considered. n.d. not determined.

Round	**R1**	**R2**	**R3**
Before selection			
Unique sequences	7	8	0
Encoded mutations	2.9 (0 – 5)	3.4 (1 – 6)	n.d.
Random mutations	2.0 (0 – 3)	3.1 (1 – 7)	n.d.
Total mutations	4.9 (3 – 10)	6.5 (2 – 12)	n.d.
After selection			
Library size	10^5	10^3	10^3
Unique sequences	7	6	2
Encoded mutations	4.0 (3 – 6)	9.8 (8 – 11)	6 (6 – 6)
Random mutations	2.7 (2 – 4)	1.3 (0 – 2)	1.5 (0 – 3)
Total mutations	6.7 (6 – 9)	11.2 (8 – 12)	7.5 (6 – 9)

Table 4.5. Sequences of CDT variants isolated from ISOR libraries.

Clone ID	Encoded mutations [1]	Random mutations	Growth time (days) [2]	Distance from AncCDT-2 [3]
Round 1	Background: AncCDT-2D2			
J3	i64M, p142K, f149Y	G99*S*, P102*L*, V166a	2	11
J4	f27Y, i64M, n108D, p142K, s171T, i202M	K23*R*, P102*S*, F209*S*	3	16
J5	i64M, n108D, i202M	F25*L*, P102*L*	3	12
J8	f27Y, a44S, i64M, n108D, v186A, e191K	T101*A*, P102*L*	3	15
J15	g70D, n108D, d227E	G12*S*, F25*V*, P102*L*, Q221*R*	3 – 4	14
J23	g70D, n108D, d227E	F25*V*, P102*L*, Q221*R*	4	13
J27	f27Y, a44S, i64M, n108D, S161*A*	P102*L*	7 – 9	13
Round 2	Background: AncCDT-2			
L1	F25*L*, i64M, g66D, T101*A*, P102*L*, l129P, t131G, a155I, a166V, p197E, l198K	D110*V*	3	12
L4	F25*L*, i64M, g66D, T101*A*, P102*L*, l129P, t131G, a155I, p197E, l198K	G12*D*, T97*A*	3	12
L5	F25*L*, i64M, g66D, P102*L*, l129P, t131G, a155I, a166V, p197E, l198K	D6*G*, n215*S*	4	12
L6	F25*L*, i64M, g66D, G99*S*, P102*L*, t131G, a155I, a166V, p197E, l198K	A69*S*, E116*G*	3	12
L9	F25*L*, i64M, g66D, P102*L*, t131G, a155I, p197E, l198K	-	4 – 5	8
L13	F25*L*, G99*S*, T101*A*, P102*L*, l129P, t131G, a155I, P188*L*, p197E, l198K	M76*V*	3 – 4	11
Round 3	Background: AncCDT-2			
M1	F25*L*, G99*S*, P102*L*, t131G, a155I, l198K	n31D, f149*L*, R163*H*	7 – 8	9
M5	F25*L*, G99*S*, P102*L*, t131G, a155I, l198K	-	7 – 8	6

[1] Ancestral states (from AncCDT-2) are denoted by lower-case letters, derived states (from AncCDT-3) by upper-case letters, and alternative states by italicised letters.

[2] Time required for the clone to reach OD_{600} of 0.2 in M9–F media at 37 °C, given as a range for at least three biological replicates.

[3] Number of amino acid substitutions separating the sequence from AncCDT-2.

To determine the importance of individual point mutations observed in ISOR-R1 for catalytic activity, several of these mutations were incorporated into AncCDT-2D2 by site-directed mutagenesis (Table 4.6). Although the P102L mutant of AncCDT-2D2 did not exhibit detectable CDT activity, addition of further mutations (F25L, I64M, G99S or T101A) to this mutant yielded active CDT variants. Δ*pheA* cells transformed with the AncCDT-2D2(F25L/P102L) construct grew in M9–F media within 4–6 days, whereas Δ*pheA* cells transformed with the other double mutants grew less rapidly (≥ 6 days) and less consistently. Thus, the F25L substitution appears to have a greater effect on catalytic activity than the I64M, G99S or T101A substitutions. Given the proximity of Pro102 and Pro188 and the positive effect of the P188L substitution on CDT activity in AncCDT-3, we also tested the effect of the P188L substitution on AncCDT-2D2 and the effect of the P102L substitution on AncCDT-3 (Table 4.6). The P188L substitution did not elicit detectable CDT activity in AncCDT-2D2, but the P102L substitution substantially increased the activity of AncCDT-3, decreasing the growth time of Δ*pheA* transformants from four days to two days.

Table 4.6. Site-directed mutagenesis of AncCDT-2D2 and AncCDT-3.

Mutations	Growth time (days) [1]	Distance from AncCDT-2 [2]
Background: AncCDT-2D2		
P102L	>9 (n.d.) [3]	8
P188L	>8 (n.d.)	8
F25L, P102L	4 – 6	9
I64M, P102L	≥6 [4]	9
G99S, P102L	≥6 [4]	9
T101A, P102L	≥6 [4]	9
Background: AncCDT-3		
P102L	2	37
P188L	1	37

[1] Time required for the clone to reach OD_{600} of 0.2 in M9–F liquid media at 37 °C, given as a range for at least three biological replicates.
[2] Number of amino acid substitutions separating the sequence from AncCDT-2.
[3] n.d. growth not detected.
[4] Growth within 8–9 days was not observed for all replicates.

Finally, two further rounds of directed evolution using ISOR were performed to purge any of the seven rational substitutions that were unessential for CDT activity in AncCDT-2. The composition of these two libraries is summarised in Table 4.4, and the sequences of clones isolated from selective media are given in Table 4.5. In ISOR-R2, the fragmented AncCDT-2 gene was reassembled with oligonucleotides encoding the seven rational substitutions and five substitutions from previous experiments (F25L, G99S, T101A, P102L and P188L). Six substitutions were conserved in every sequenced clone: four rational substitutions (T131G, A155I, P197E and L198K), and two substitutions from ISOR-R1 (F25L and P102L). In most cases, the growth rates of these clones were greater than AncCDT-3, and the clone with the fewest substitutions relative to AncCDT-2 contained eight substitutions. In ISOR-R3, the number of mutagenic oligonucleotides was reduced further: the six substitutions conserved in ISOR-R2, in addition to the G99S substitution, were included. The two clones with unique sequences that were isolated in this round both contained three rational substitutions (T131G, A155I

and L198K) and three substitutions from ISOR-R1 (F25L, P102L and G99S). In addition, one clone (CDT-M5) that contained only these six substitutions was isolated, although growth of this clone was slow compared with AncCDT-3. Major differences in protein expression between the original ancestral proteins and the CDT variants obtained from directed evolution were not observed, and cannot account for the observed differences in growth rates (Figure 4.8).

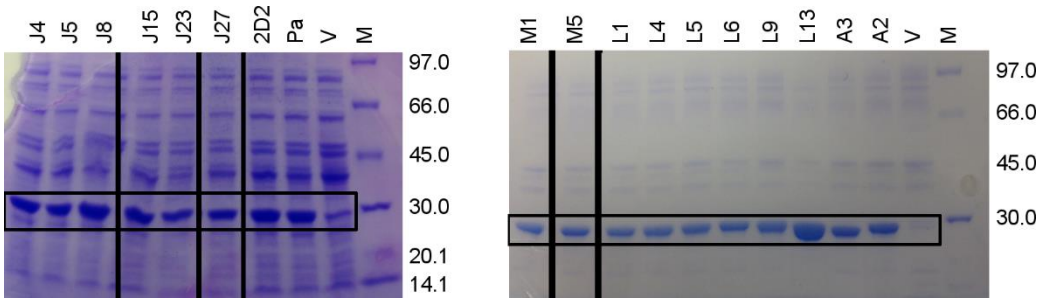


Figure 4.8. Expression of CDT variants obtained by directed evolution. SDS PAGE gels showing protein expression in the soluble crude lysates of Δ*pheA* cells transformed with CDT variants and incubated in M9+F media at 37 °C overnight. The removal of irrelevant lanes from the gel images is indicated by the vertical black lines. The boxes show the bands corresponding to CDT (~28 kDa). Abbreviations: 2D2, AncCDT-2D2; Pa, PaCDT; V, vector; M, marker (with the molecular weight of each band listed in kDa); A3, AncCDT-3; A2, AncCDT-2.

Preservation of the T131G, A155I and L198K substitutions throughout two rounds of ISOR suggests that these substitutions were historically important for the emergence of CDT activity in AncCDT-2. However, AncCDT-2 represents only the most probabilistic ancestral sequence out of an ensemble of plausible ancestral sequences, and the substitutions observed in the directed evolution experiments could simply represent sampling of alternative ancestral states. Thus, to evaluate whether these substitutions actually occurred during the evolution of CDT, the PP distributions of the ancestral CDT variants at these positions were analysed (Figure 4.9). Given the assumptions underlying the phylogenetic analysis, the substitutions T131G and A155I (or, less probably, A155V) are virtually certain to have occurred between AncCDT-1 and AncCDT-3. However, the historical importance of the L198K substitution is less certain; Leu198 is reconstructed with low confidence in AncCDT-1 and AncCDT-2, and is a plausible alternative to the most likely state (Lys) in AncCDT-3. Indeed, Leu198 is found in several extant homologues of Ea1174; thus, the L198K substitution must be dispensable if other compensatory substitutions have occurred. The PP distributions of the ancestral proteins were also analysed at positions where other noteworthy substitutions (F25L/V, I64M, G99S, T101A, P102S/L, P188S/L) occurred in the directed evolution experiments (Figure 4.9). The residues that accrued random substitutions (Phe25, Gly99, Thr101, Pro102, Pro188), with the notable exception of Pro188, were reconstructed unambiguously, indicating that the substitutions do not represent sampling of alternative ancestral states. Furthermore, the historical substitution I64M, which was observed in ISOR-R1 and ISOR-R2, was predicted with high statistical confidence to have occurred between AncCDT-1 and AncCDT-3.

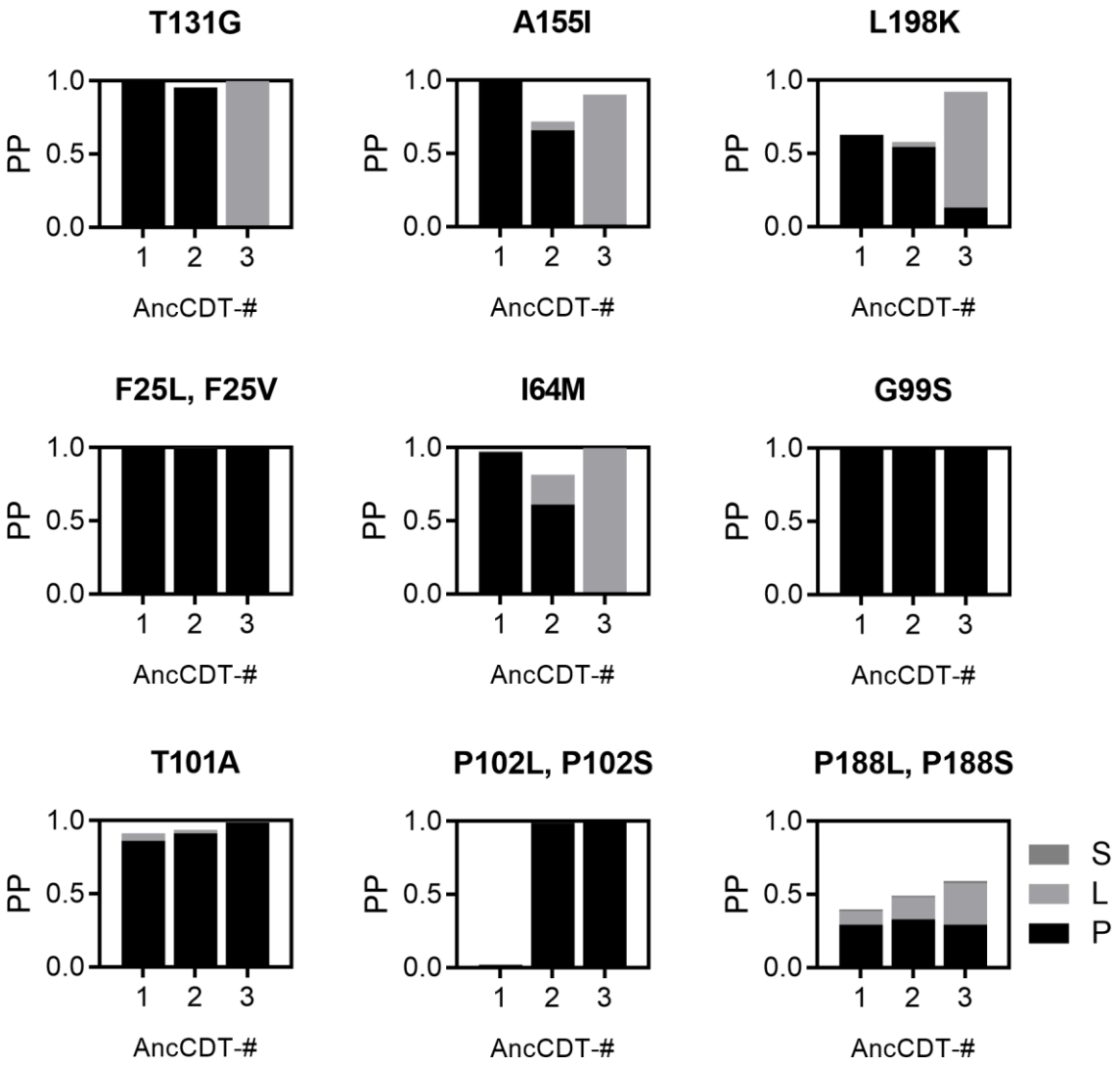


Figure 4.9. Posterior probability distributions of ancestral CDT variants at mutated positions. For each substitution, the posterior probabilities (PPs) of the ancestral state (black) and the derived state(s) (grey) are shown for AncCDT-1, AncCDT-2 and AncCDT-3.

4.3 Discussion

Using a combination of site-directed mutagenesis and directed evolution, we have shown that no more than six substitutions are required to recapitulate the evolution of CDT activity in AncCDT-2. Three of these substitutions (T131G, A155I, L198K) occurred on the branch of the CDT phylogeny joining AncCDT-2 and AncCDT-3 and therefore had plausible historical importance in the evolution of CDT, whereas the three remaining substitutions (F25L, G99S, P102L) constitute an alternative evolutionary trajectory towards higher CDT activity, which is different from the historical trajectory. Incorporation of additional mutations in the AncCDT-2 background was associated with further increases in apparent catalytic activity. A number of substitutions that improve the apparent catalytic activity of AncCDT-3 were also identified, including the substitution P188L, which is responsible for the phenotypic difference between AncCDT-3 and AncCDT-3W. The structural effects of the substitutions observed in these experiments, and the possible implications for catalysis in CDT, are considered in Section 6.2.3.

Altogether, three substitutions (T131G, A155I, L198K) that occurred on the AncCDT-2/3 branch of the CDT phylogeny were conserved in all derivatives of AncCDT-2 confirmed to have sufficient CDT activity to support the growth of *E. coli* phenylalanine auxotrophs in minimal media. The historical importance of the T131G and A155I substitutions is evident from the phylogenetic analysis; the reconstruction of these ancestral states in AncCDT-1 and AncCDT-3 was unambiguous, with the exception that A155V was a plausible alternative to A155I. The historical importance of the L198K substitution, however, was not unequivocal; the reconstruction of the corresponding ancestral states was ambiguous. These predictions from the phylogenetic analysis can be reconciled with patterns of sequence conservation in extant CDT homologues (Figure 4.5,

p. 137); Gly131 is conserved without exception and Ala155 is very rare, usually replaced by Ile, Val, or Asn, whereas at position 198, Leu and other hydrophobic residues are frequently observed in place of Lys. Notably, none of the three substitutions are accessible *via* single nucleotide mutations, implying that alternative residues must have been sampled at these positions between AncCDT-1 and AncCDT-3. Site-saturation mutagenesis could be used to determine whether any of these intermediate states, accessible from the ancestral state *via* single nucleotide mutations, were compatible with CDT activity.

Directed evolution and mutagenesis of AncCDT-2 and AncCDT-3 uncovered an unexpected mutational hotspot, at which a disproportionate number of substitutions that increased the apparent catalytic activity of the ancestral proteins occurred (F98S, G99S, T101A, P102S, P102L, V187A, V187D, P188S and P188L). These substitutions are located on the two hinge strands of the enzyme, close to the small domain. Given the considerable distance between this region of the enzyme and its active site, as well as the structural flexibility of the outer hinge strand, the mechanism by which these substitutions increase CDT activity is not obvious. Interestingly, several of these substitutions represent reversions to residues similar to the original ancestral state in AncCDT-1: F98S (Ala in AncCDT-1), P102L (Ile in AncCDT-1) and P188L (Ile in AncCDT-1). Thus, the substitutions that occurred at these positions were probably not required for the evolution of CDT activity.

The availability of multiple mutational pathways to higher catalytic activity in the ancestral CDT variants suggests that these proteins are highly evolvable. The availability of multiple mutational pathways is evidenced by the number of substitutions that increased the catalytic activity of AncCDT-2 variants and AncCDT-3 (Tables 4.3, 4.5, 4.6), as well as the fact that the substitutions acquired by AncCDT-2 during directed

evolution were not historical substitutions, that is, the historical evolutionary trajectory between AncCDT-2 and AncCDT-3 was different from the trajectory observed in the directed evolution experiments. Additionally, many of the substitutions encountered in the directed evolution of AncCDT-2 appeared to have an additive effect on CDT activity. For example, whereas six substitutions in AncCDT-2 provide sufficient CDT activity to enable slow growth of phenylalanine auxotrophs, earlier variants from ISOR-R1 and R2 with additional substitutions exhibited faster growth (Table 4.5). Likewise, site-directed mutagenesis showed that various substitutions in the AncCDT-2D2 background, in combination with P102L, yielded slow growth of phenylalanine auxotrophs, but the variants from ISOR-R1 with different combinations of these substitutions exhibited faster growth (Table 4.5, Table 4.6). Altogether, these observations give the impression that the fitness landscape of CDT is relatively smooth, with multiple mutational pathways to higher catalytic activity and different mutations contributing additively to catalytic efficiency. Thus, although specific substitutions would have been needed to introduce the catalytic machinery required for CDT activity (e.g., Lys100, Asn128, Asn133, Asn152 and Glu173) in the ancestral SBP, once these catalytic residues are present, as in AncCDT-2 and AncCDT-3, there appears to be multiple solutions to the refinement of catalytic activity by remote substitutions.

The main limitation of the genetic selection for CDT activity used in these directed evolution experiments is its dynamic range; the power to distinguish CDT variants with different catalytic efficiencies is quite limited. Variants with promiscuous CDT activity cannot be distinguished by genetic selection because of the leakiness of phenylalanine auxotrophy in the Δ*pheA* strain, a result of slow non-enzymatic conversion of prephenate to phenylpyruvate, which leads to slow growth in the absence of CDT activity (Kleeb et al., 2007). Thus, without recourse to *in vitro* assays, we cannot exclude the possibility that some of the variants constructed by site-directed mutagenesis, which appeared to be

inactive, possess promiscuous CDT activity. On the other hand, highly efficient CDT variants cannot be differentiated from moderately efficient CDT variants because phenylalanine production is not rate-limiting for growth given sufficient CDT activity. Thus, the k_{cat}/K_M of PaCDT is 6000-fold greater than that of AncCDT-3(P188L), but the two enzymes support similar growth rates in Δ*pheA* cells. Previous work has shown that the range of k_{cat}/K_M that could be differentiated by genetic selection for prephenate dehydratase activity was approximately 24 to 3000 M^{-1} s^{-1} (Kleeb et al., 2007). Although our results are not directly comparable due to differences in the bacterial strain, plasmid and growth temperature, it is reasonable to assume that the conditions for genetic selection used in this work yield a similar dynamic range (~100-fold difference in k_{cat}/K_M). The stringency of selection could be increased by introducing an inducible cyclohexadienyl dehydrogenase into the system to reduce the intracellular substrate concentration, as described previously (Kleeb et al., 2007), or by using a different plasmid offering tighter control over expression of the CDT variants. A second limitation of the genetic selection strategy is the prephenate dehydratase and arogenate dehydratase activities of CDT cannot be distinguished; presumably, prephenate dehydratase activity is responsible for growth, although conversion of prephenate to L-arogenate by promiscuous aminotransferases upon accumulation of prephenate at a non-physiological concentration may also enable L-phenylalanine production from arogenate dehydratase activity. Thus, although genetic selection is a useful high-throughput method for initial assessment of CDT activity, *in vitro* enzyme assays will be needed for a complete and quantitative investigation of the evolution of CDT.

Site-directed mutagenesis and recombination of AncCDT-3 and AncCDT-3W using StEP revealed that the P188L substitution is responsible for the phenotypic difference between the two alternative reconstructions of this ancestral protein. In the maximum-likelihood reconstruction that yielded AncCDT-3, Pro188 and Leu188 are

roughly equal in posterior probability (*P*(Pro) = 0.293, *P*(Leu) = 0.280; Figure 4.9); use of the alternative WAG substitution matrix, yielding AncCDT-3W, slightly perturbed these posterior probabilities to favour Leu at this position (*P*(Pro) = 0.328, *P*(Leu) = 0.410). Errors in the multiple sequence alignment likely contributed to the low posterior probabilities at this position; Pro188 is located in the second hinge strand of the protein, which is flexible and variable in length among CDT homologues, and corresponds to the only portion of the alignment with significant uncertainty. Thus, in statistical terms, AncCDT-3(P188L) is effectively an equally probable reconstruction of the LCA of CDTs, compared with AncCDT-3. For this reason, and because of the higher catalytic activity of AncCDT-3(P188L), the AncCDT-3(P188L) variant was characterised in subsequent work (Chapter 6) rather than AncCDT-3 itself.

The directed evolution experiments described in this chapter provide a basis for further experiments needed to propose an evolutionary trajectory between non-catalytic SBPs and catalytic CDTs that is both complete (starting from a high-affinity SBP and concluding with a highly active CDT) and biologically plausible (accessible *via* single point mutations and excluding non-functional intermediates). Completion of the evolutionary trajectory would require the function of AncCDT-2 to be identified, most likely following functional characterisation of Pu1068, or identification of the substitutions required for the gain of CDT activity in the high-affinity AABP AncCDT-1. To exclude non-functional intermediates from the evolutionary trajectory, substitutions that are necessary for the gain of CDT activity but neutral with respect to the ancestral solute-binding function would need to be identified.

Evolutionary optimisation of AncCDT-3(P188L) could also be productive for understanding the catalytic proficiency of PaCDT, which has a k_{cat}/K_M ~6000-fold higher than the ancestral protein. Further improvements to AncCDT-3(P188L) could be

achieved using random mutagenesis together with a more stringent genetic selection, as outlined above. These experiments would also have implications for the reproducibility of protein evolution; it would be instructive to determine whether alternative mutational pathways to higher catalytic activity are possible in AncCDT-3(P188L), as in AncCDT-2 and AncCDT-3, or whether certain historical substitutions towards extant CDT homologues must be recapitulated for further gains in catalytic activity.

4.4 Materials and methods

4.4.1 Materials

Codon-optimised synthetic genes encoding the ancestral proteins, cloned into the *Sap*I site of the pDOTS7 vector using the Golden Gate method, were obtained from Joe Kaczmarski. *E. coli* strain JW2580-1 was obtained from the Coli Genetic Stock Center at Yale University, New Haven, CT. This strain is from the Keio collection of *E. coli* knockouts (Baba et al., 2006) and exhibits phenylalanine auxotrophy as a result of a knockout in *pheA*.

4.4.2 Molecular dynamics simulations

The structure of PaCDT (PDB: 3KBR) was prepared by modelling a missing residue (Gln190) and missing side chains in MODELLER (Sali and Blundell, 1993). An acetyl cap was added to the N-terminal residue and an amide cap was added to the C-terminal residue using MODELLER and Coot (Emsley et al., 2010). Initial poses for the PaCDT/L-arogenate complex were obtained by docking L-arogenate into the rigid PaCDT structure using AutoDock Vina (Trott and Olson, 2010). The topology file for L-arogenate was generated using Automated Topology Builder (Koziara et al., 2014), and charge groups and partial atomic charges were edited (Lemkul et al., 2010) to ensure consistency with equivalent functional groups parameterised in the GROMOS 53a6 force field.

MD simulations of the PaCDT-arogenate complex were performed using GROMACS version 4.5.5 (Pronk et al., 2013) with the GROMOS 53a6 force field (Oostenbrink et al., 2004). The enzyme-substrate complex was solvated in a rhombic dodecahedron with SPC water molecules, such that the minimal distance of the protein to the periodic boundary was 15 Å, and six Na^+ ions were added to neutralize the system. Energy minimisation was done using the steepest descent algorithm. A 100 ps isothermal (NVT) MD simulation with position restraints on the protein and substrate was used to

equilibrate the system at 300 K. For production MD simulations of the NPT ensemble, the temperature was maintained at 300 K using Berendsen's thermostat (τ_T = 0.1 ps), and the pressure was maintained at 1 bar using Berendsen's barostat (τ_p = 0.5 ps, compressibility = 4.5×10^{-5} bar^{-1}). All protein bonds were constrained with the LINCS algorithm; water molecules were constrained using the SETTLE algorithm; the time step for numerical integration was 2 fs; the cut-offs for short-range electrostatics and van der Waals forces were 9 Å and 14 Å, respectively; the Particle-Mesh Ewald method was used to evaluate long-range electrostatics; neighbour lists were updated every 10 steps. Following a 1 ns equilibration phase, which was not considered in the analysis, the four production runs were continued for 50 ns.

4.4.3 Mutagenesis

Mutagenesis was achieved using Gibson assembly as described in Section 2.7.3, except that the mutagenised gene fragments were assembled together with the pDOTS7 vector rather than the pETMCSIII vector.

4.4.4 Genetic complementation assays

E. coli strain JW2580-1 (Δ*phe*A) cells were transformed with the appropriate plasmid by electroporation, plated on LB agar supplemented with 100 mg/L ampicillin (LBA), and incubated at 37 °C overnight. Single colonies were used to inoculate 20 mL M9 minimal media supplemented with L-tyrosine, ampicillin and IPTG (M9–F; per L: 6 g Na_2HPO_4, 3 g KH_2PO_4, 0.5 g NaCl, 1 g NH_4Cl, 20 mL 20% (w/v) glucose, 2 mL 1 M $MgCl_2$, 0.1 mL 1 M $CaCl_2$, 2 mL 2.5 mg/mL L-tyrosine, 1 mL 100 mg/mL ampicillin, 0.2 mL 1 M IPTG). The cultures were incubated at 37 °C with shaking at 180 rpm for 8 – 10 days and OD_{600} was measured periodically. For assessment of protein expression, single colonies were used to inoculate 20 mL M9–F media supplemented with 20 µg/mL L-phenylalanine (M9+F), which was incubated at 37 °C overnight. 1 mL of the resulting culture was

pelleted and resuspended in Ni-equilibration buffer. The cells were lysed using BugBuster Protein Extraction Reagent (Merck-Millipore) and fractionated by centrifugation at 14000 rpm for 10 min, and the soluble fraction of the cell lysate was analysed by SDS-PAGE using an ExpressPlus 4–20% polyacrylamide gel (GenScript) stained with Coomassie Blue.

4.4.5 Staggered extension process (StEP)

Recombination using the staggered extension process (StEP) was performed following a literature protocol (Zhao and Zha, 2006). The StEP reaction mixture contained 5 μL 10× *Taq* buffer, 1.5 mM $MgCl_2$, 0.2 mM each dNTP, 75 fmol each template plasmid, 30 pmol each primer, and 2.5 U *Taq* polymerase (New England Biolabs) in a total volume of 50 μL. The primers used in the reaction were the 5′ flanking primer P7XF and the 3′ flanking primer P7XR (Table 4.7), which amplify ~100 bp on either side of the *Sap*I site of the pDOTS7 vector. The thermocycling program consisted of 80 cycles of (i) a denaturation step for 30 s at 95 °C; and (ii) an annealing/extension step for 5 s at 52 °C. 2 μL of the resulting PCR product was incubated with 10 U *Dpn*I (Thermo Scientific) in a reaction volume of 10 μL at 37 °C for 1 hr to digest the parental plasmid DNA. 5 μL of the *Dpn*I-digested StEP product was then amplified in a nested PCR reaction using *Taq* polymerase, in a total volume of 100 μL. The primers used for the nested PCR reaction, P7NF and P7NR (Table 4.7), target the *Eco*RI site on the 5′ strand and the *Hind*III site on the 3′ strand of the pDOTS7 vector, respectively. The nested PCR product was run on a 1% agarose gel. The target band was excised from the gel and purified using the QiaQUICK Gel Extraction Kit.

Table 4.7. Sequences of primers used for directed evolution.

P7XF	CGTCTTCACCTCGAGAAATC
P7XR	CAACCGAGCGTTCTGAAC
P7NF	CAATTTCACACAGAATTCATTAAAG
P7NR	GCTCAGCTAATTAAGCTTTTATTAG
P7SF	GGCCCTTTCGTCTTCAC
P7SR	AGCTTGGATTCTCACCAAT

4.4.6 Incorporation of synthetic oligonucleotides via gene reassembly

Incorporation of synthetic oligonucleotides *via* gene reassembly (ISOR) was done following literature protocols (Herman and Tawfik, 2007; Rockah-Shmuel et al., 2014). The template gene was amplified by PCR using Phusion Hot Start II Polymerase (Thermo Scientific) using the primers P7XF and P7XR (Table 4.7). The purified PCR product was digested with DNAse I (New England Biolabs) in a reaction mixture containing 100 mM Tris pH 7.5, 10 mM $MnCl_2$, 4 μg PCR product and 0.3 U DNAse I in a total volume of 40 μL. The reaction mixture was incubated at 37 °C for 1 – 2 min and quenched by the addition of 20 μL 0.1 M EDTA pH 8.0 pre-incubated at 80 °C, followed by heat inactivation at 80 °C for 15 min. The digested PCR product was run on a 2% agarose gel, and fragments 50 – 250 bp in size were excised from the gel and purified using the Wizard SV Gel and PCR Clean-Up System (Promega). The fragments were reassembled using *Taq* polymerase: each reaction contained 40 ng gene fragments, 2 μL 10× buffer, 0.2 mM dNTPs, 1.25 U *Taq* polymerase and varied concentrations of equimolar mutagenic oligonucleotides (5 – 800 nM total concentration) in a volume of 20 μL (see Table 2.1 for a list of oligonucleotides included in each round). The thermocycling protocol consisted of (i) an initial denaturation step at 95 °C for 2 min; (ii) 40 cycles of a denaturation step at 95 °C for 30 s, then 13 hybridization steps from 65 °C to 41 °C in 2 °C steps, each for 90 s (total 13.5 min), then an extension step at 72 °C for 1 min; and

(iii) a final extension step at 72 °C for 7 min. 0.5 μL of the unpurified assembly reaction mixture was amplified in a 50 μL nested PCR reaction using *Taq* polymerase and the primers P7NF and P7NR (Table 4.7). The nested PCR product was run on a 1% agarose gel and purified by gel extraction.

Table 4.8. Mutagenic oligonucleotides used for ISOR.

ISOR-R1 Substitution	**Oligonucleotide**
F27Y	aaaccgtttagctataaagatccgaacggtca
K28R	accgtttagctttcgtgatccgaacggtc
N31D	gctttaaagatccggatggtcagtatacc
A44S	gatgttgcaaaaagcctggcaaaagatctggg
D48S	agcactggcaaaaagcctgggtgttaaag
I64M	tggccgaccctgatgagcgatctgcag
G70D	gatctgcaggcagataaatttgatatcgcaatgg
I81V	ggtggtgttaccgtgacaccggaacgt
N108D	gctggttcgtaaagaagatgccgataaattcaa
L118I	aagcctggaagatattaataaaccggatgt
K120R	cctggaagatctgaatcgtccggatgttaaag
K124R	aaaccggatgttcgtgtggcagttaatccg
K138R	accaatgaaaaatttgcccgtgaacatctgccg
P142K	gccaaagaacatctgaagaaagccaaaattc
R147T	gccgaaagccaaaattaccgtgtttgaaaataatgcc
F149Y	agccaaaattcgtgtgtatgaaaataatgccg
A153V	cgtgtgtttgaaaataatgtggaaatttttcaagaggttg
V160A	tttcaagaggttgcgagcggtcgtgccgatgt
S161A	caagaggttgttgcgggtcgtgccgatgt
S171T	gatgtgatgattaccgataccgttgaagcagcat
A175L	ccgatagcgttgaagcactgtattacgcaaaa
V186A	ccgggtctggcagcggttccggttgat
E191K	gttccggttgataaaccgtttacccatagt
F200Y	acccatagtgaaaaaggttatatgattccgaaagg
I202M	aaaggtttcatgatgccgaaaggtgatccg
P207Q	tccgaaaggtgatcaggaatttctgaactatgtg
N215Q	cggaatttctgaactatgtgaaccagtggctgaaacaa
K218D	aacaattggctggatcaaatgaaacagcaggg
Q223N	ggctgaaacaaatgaaacagaacggcacctatgataaa
D227E	acagcagggcacctatgaaaaactgtatgaaa
F234L	gataaactgtatgaaaaatggctgaaataataaaagc
ISOR-R2 Substitution	**Oligonucleotide**
F25L	ccggtgattataaaccgcttagctttaaagatcc
I64M/G66D	gccgaccctgatgagcgatctgcag
G66D	ccctgattagcgatctgcaggcagg
G99S/P102L	ccgtatatgacctttagtaaaacactgctg
T101A/P102L	gacctttggtaaagcactgctggttc
P102L	ctttggtaaaacactgctggttcgtaaag
L129P/T131G	gatgttaaagtggcagttaatccgggcggcaccaatgaaaaatttg
T131G	ggcagttaatctgggcggcaccaatgaaaaatttg
A155I	gtttgaaaataatgccgaaatttttcaagaggttgttagcgg
A166V	gcggtcgtgccgatgtgatgattaccgatagcg
P188L	ctggcagcagttctggttgataaacc

P197E	gaaccgtttacccatagtgaactgggtttcatgattccg
L198K	ccgtttacccatagtccgaaaggtttcatgattccgaaag
P197E/L198K	ggttgatgaaccgtttacccatagtgaaaaaggtttcatgattccgaaaggtgatc

ISOR-R3 Substitution	**Oligonucleotide**
F25L	ccggtgattataaaccgcttagctttaaagatcc
G99S/P102L	ccgtatatgacctttagtaaaacactgctg
P102L	ctttggtaaaacactgctggttcgtaaag
T131G	ggcagttaatctgggcggcaccaatgaaaaatttg
A155I	gtttgaaaataatgccgaaatttttcaagaggttgttagcgg
P197E	gaaccgtttacccatagtgaactgggtttcatgattccg
L198K	ccgtttacccatagtccgaaaggtttcatgattccgaaag
P197E/L198K	ggttgatgaaccgtttacccatagtgaaaaaggtttcatgattccgaaaggtgatc

4.4.7 Library creation and selection

Purified PCR products (0.5 μg) from StEP or ISOR reactions were digested with 2.5 μL each of *Hind*III FD and *EcoR*I FD (Thermo Scientific) in a 50 μL reaction at 37 °C for 30 min. The reaction mixture was purified immediately using a PCR purification kit. The pDOTS7 vector containing the AncCDT-2 insert (2.5 μg) was digested using 2.5 μL each of *Hind*III FD, *EcoR*I FD, and *Pst*I FD (which cuts within the AncCDT-2 insert) in a 50 μL reaction at 37 °C for 30 min. The digested vector was purified immediately using a PCR purification kit, then run on a 1% agarose gel and purified by gel extraction. Ligation reaction mixtures contained 100 ng pDOTS7 vector, a 3-fold molar excess of insert, 2 μL 10× T4 DNA ligase buffer, and 5 U T4 DNA ligase (Thermo Scientific) in a volume of 20 μL, and were incubated at room temperature for 1 hr. Following purification of the ligation reaction mixture using a PCR purification kit, electrocompetent *E. coli* strain JW2580-1 (Δ*pheA*) cells were transformed with 1 μL ligation product by electroporation and plated on LBA. Following overnight incubation of the plates at 37 °C, colonies were scraped into LB media, then resuspended in 20 mL fresh LBA media. 100 μL of the resulting cell suspension was used to inoculate 20 mL fresh LBA media, which was then incubated at 37 °C until the OD_{600} reached ~0.5. A 1 mL aliquot of the culture was washed twice with 1 mL M9 salts (6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L NH_4Cl, 0.5 g/L NaCl),

and resuspended in 1 mL M9 salts. Serial dilutions of the cell suspension were made in M9 salts, plated on M9–F agar, and incubated at 37 °C. The resulting colonies were streaked onto LBA agar, and their plasmid DNA was amplified by PCR using the sequencing primers P7SF and P7SR (Table 4.7). The resulting PCR products were sequenced by GENEWIZ (South Plainfield, N.J., U.S.A.) or the Biomolecular Resource Facility at ANU. Single colonies from the streaked LBA plates were used to confirm growth of the clone in liquid M9–F media, as described in Section 4.4.4, and to inoculate LBA cultures, from which plasmid DNA was extracted using a FavorPrep Plasmid DNA Extraction Mini Kit (Favorgen).

Chapter Five

Evolution of an enzyme from a solute-binding protein. Part III: Dynamics.

5.1 Introduction

5.1.1 Enzyme dynamics and evolution

Enzymes are intrinsically dynamic molecules and a complete description of enzyme catalysis requires consideration of their conformational landscapes (Henzler-Wildman and Kern, 2007; Ma and Nussinov, 2010). This is because rate enhancements by enzymes depend not only on their ability to reduce the free energy barrier for a reaction through, for example, pre-organisation of the active site for transition state stabilisation, but on their ability to cycle through the conformational states that are required for catalysis. For example, conformational changes are often required to facilitate substrate binding or product release, and these steps can be rate-limiting (Jackson et al., 2009). Alternatively, conformational changes may be required for an enzyme to access a high-energy catalytically competent state (Henzler-Wildman et al., 2007).

Given the importance of enzyme dynamics for catalysis, the dynamic properties of enzymes might be expected to be under evolutionary selection. However, since the dynamics of a protein are ultimately a product of its structure, it has been challenging to distinguish whether conservation of protein dynamics reflects selection for dynamic properties or is simply a by-product of structural conservation (Marsh and Teichmann, 2014). Nonetheless, there are some compelling examples of the adaptation of enzyme dynamics to different functions or cellular environments. For example, *E. coli* dihydrofolate reductase (DHFR) undergoes conformational motions that enable the enzyme to release and replenish its cofactor rapidly, whereas human DHFR, which is adapted to lower intracellular concentrations of cofactor, has evolved altered conformational sampling that enables the enzyme to release its cofactor more slowly (Bhabha et al., 2013). Enzyme dynamics are also important from an evolutionary perspective because unproductive conformational sampling can limit the catalytic

efficiency of recently evolved enzymes, prior to the stabilisation of the catalytically competent conformations of newly introduced active site residues by more remote substitutions (Bar-Even et al., 2015; Mabbitt et al., 2016).

5.1.2 Conformational dynamics of amino acid-binding proteins

As mentioned in Section 1.2.2, SBPs undergo conformational cycling between an open conformation in the unliganded state and a closed conformation in the liganded state (Figure 5.1). This characteristic hinge-bending motion appears to be an intrinsic consequence of the SBP architecture (Keskin et al., 2000), although the magnitude of the conformational change appears to vary between SBP classes (Berntsson et al., 2010). The open conformation of SBPs is generally more flexible than the closed conformation, as illustrated by free energy calculations of the conformational landscape, showing broad free energy minima for the open conformation (Lau and Roux, 2007; Yao et al., 2013), and by the observation of different open conformations in different crystal structures of the same SBP (Björkman and Mowbray, 1998).

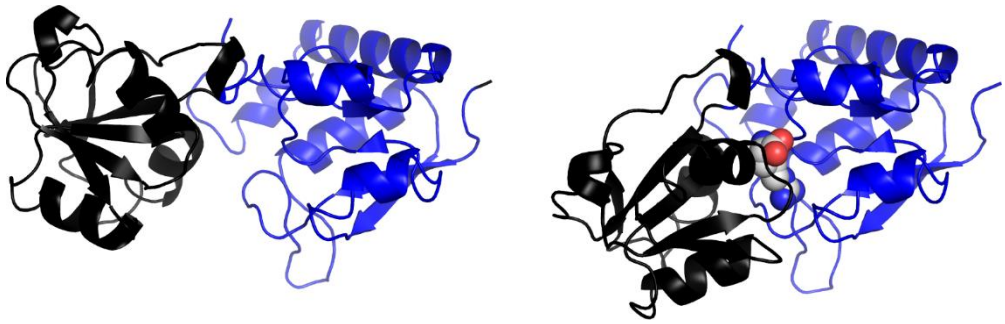


Figure 5.1. Open and closed conformations of an amino acid-binding protein. Crystal structures of lysine-/arginine-/ornithine-binding protein in the open conformation (left; PDB: 2LAO) and closed conformation in complex with L-arginine (right; PDB: 1LAF). The two domains of the protein are shown in different colours.

Although the open conformation is the ground state for unliganded SBPs and the closed conformation is the ground state for liganded SBPs (Davidson et al., 2008),

multiple lines of evidence have established that SBPs can sample closed or partially closed conformations even in the absence of ligands. Several SBPs have been crystallised in the unliganded closed state (Flocco and Mowbray, 1994; Oswald et al., 2008). Conformation-specific antibodies for liganded HisBP also react with unliganded HisBP (Wolf et al., 1994). Paramagnetic relaxation enhancement NMR experiments (Tang et al., 2007) and accelerated MD simulations (Bucher et al., 2011a) have shown that unliganded MBP samples a minor partially closed conformation with an occupancy of ~5%. Finally, single-molecule FRET has been used to observe open-closed transitions in unliganded SBPs of the AABP-dependent glutamine/asparagine transport system GlnPQ from *Lactococcus lactis* (Gouridis et al., 2014) and variants of MBP (Kim et al., 2013a). However, the closed conformation may not be energetically accessible for all unliganded SBPs; paramagnetic relaxation enhancement measurements on unliganded GlnBP, unlike MBP, could be explained without accounting for minor populations of closed conformers (Bermejo et al., 2010).

The ability of SBPs to sample closed or partially closed conformations in the absence of ligand raises the possibility that ligand binding occurs by a conformational selection mechanism (Bucher et al., 2011b). Conformational selection is one mechanism that has been proposed to account for the formation of protein-ligand complexes where ligand binding and protein conformational changes are coupled, the other principal mechanism being induced fit (Figure 5.2). According to the induced fit model, a ligand binds to an SBP in the open conformation and triggers a transition to the closed conformation. In contrast, the conformational selection model posits that a small population of the SBP exists in the closed conformation prior to ligand binding, and that preferential binding of the ligand to the closed conformation results in a population shift towards the closed liganded SBP. Pre-equilibrium between the open and closed conformations in the unliganded state is a necessary but insufficient condition for

conformational selection; an additional requirement is that the ligand must bind preferentially to the closed conformation. Many SBPs engulf their ligands completely, such that the binding site is not accessible in the closed conformation and a conformational selection mechanism is impossible. Thus, induced fit has been regarded historically as the predominant mechanism of ligand-induced conformational change in SBPs.

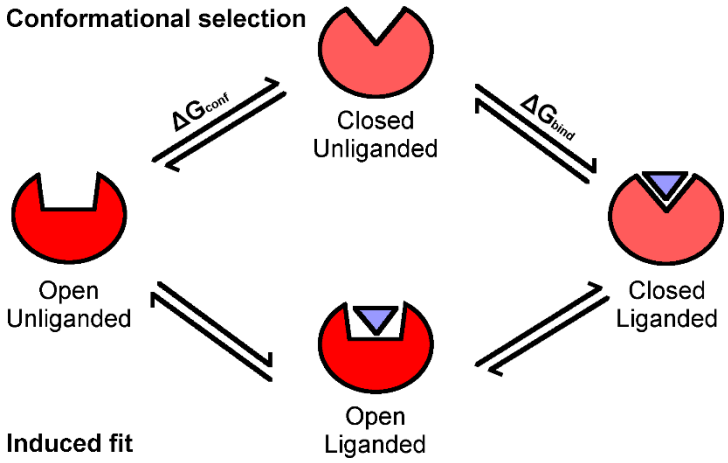


Figure 5.2. Ligand-induced conformational change by the induced fit and conformational selection mechanisms. According to the conformational selection model (upper pathway), the conformational change of the protein precedes ligand binding, whereas according to the induced fit model (lower pathway), ligand binding precedes the conformational change. This simplified scheme can be extended to include additional conformational states and hybrid induced fit-conformational selection models. The total free energy of ligand binding (ΔG) can be expressed as the sum of the intrinsic free energy difference between the open and closed states of the protein (ΔG_{conf}) and the free energy change associated with ligand binding to the closed state of the protein (ΔG_{bind}).

Recent experimental and theoretical work has confirmed that ligand binding to SBPs occurs either by the induced fit mechanism or by hybrid induced fit-conformational selection mechanisms involving intermediate semi-closed states (Bucher et al., 2011b; Gouridis et al., 2014; Kim et al., 2013a; Silva et al., 2011b). Unlike bulk measurements

of equilibrium relaxation kinetics (Vogt and Di Cera, 2012, 2013), single-molecule FRET (smFRET) experiments allow conformational changes to be monitored independently of ligand binding, and can thus distinguish the induced fit and conformational selection mechanisms. smFRET studies of MBP and GlnPQ have shown the open-closed transition follows first-order kinetics, as predicted by the induced fit model (Gouridis et al., 2014; Kim et al., 2013a). Moreover, three-colour smFRET experiments, in which binding events and conformational changes can be detected independently but simultaneously, have shown that fluorescently labelled maltose binds more frequently to the open conformation of MBP than the semi-closed conformation (Kim et al., 2013a). Theoretical work has supported these conclusions; Bucher *et al*. showed using free energy calculations based on the thermodynamic integration method that the semi-closed state of MBP has a higher affinity for maltotriose than the open state, but concluded that the low occupancy of the semi-closed state and the slow rate of the open to semi-closed transition disfavours the conformational selection pathway; in contrast, calculation of the conformational free energy landscape of maltose-bound MBP using the adaptive biasing force method showed that the open-closed transition in maltotriose-bound MBP is barrierless, favouring the induced fit pathway (Bucher et al., 2011b).

The open-closed conformational equilibrium of an SBP is inextricably linked to its binding affinity. The total free energy of ligand binding can be decomposed using thermodynamic linkage relationships into the intrinsic free energy change associated with the conformational change and the free energy change associated with ligand binding to the closed state (Figure 5.2). It follows that the affinity of SBPs can be altered by allosteric mutations that manipulate the intrinsic equilibrium between the open and closed states, without affecting the protein-ligand interface. Many such allosteric mutations that increase or decrease affinity have been identified in MBP (Marvin and Hellinga, 2001; Seo et al., 2014; Telmer and Shilton, 2003). Likewise, antibodies that target the closed

conformation of MBP increase ligand affinity through similar allosteric effects (Rizk et al., 2011). Evolutionary selection for increased affinity in SBPs could manifest in stabilisation of the closed conformation (Telmer and Shilton, 2003). The periplasmic domains of the sensor-kinase BvgS, which are homologous to SBPs, provide an extreme example of selection for the position of the open-closed equilibrium; these proteins are constitutively closed and therefore active except in the presence of negative modulators (Dupré et al., 2015; Herrou et al., 2010).

Solute translocation by SBP-dependent ATP transporters requires ligand release into the transmembrane domains, enabled by the SBP transitioning from the closed conformation to the open conformation; thus, the conformational landscapes of SBPs, specifically, the relative free energies of the open and closed conformations and the rate of exchange between the two conformations, are relevant to the process of solute translocation. Recent work on GlnPQ showed that solute transport can be inhibited by closed unliganded SBPs, suggesting that sampling of the closed conformation in the absence of ligand can regulate the rate of solute transport, and showed that the intrinsic lifetime of the closed conformation of the SBP affects the rate of transport, with longer lifetimes giving slower transport (Gouridis et al., 2014). Hinge mutants of MBP with higher maltose affinity due to stabilisation of the closed conformation also have longer lifetimes in the closed conformation (Seo et al., 2014), suggesting that there is an association between the position of the open-closed equilibrium and the transport rate, and that there is a possible trade-off between binding affinity and transport rate. Finally, it has been proposed that interactions between closed unliganded SBPs and the TMDs of ABC transporters could be responsible for futile ATP hydrolysis observed in the absence of ligand, although in $MalFGK_2$ at least, futile ATP hydrolysis can be stimulated by open MBP (Gould et al., 2009).

In summary, the conformational dynamics of SBPs have important consequences for the mechanism of binding (induced fit versus conformational selection), binding affinity, and the rate of solute transport, which suggests that the conformational landscape of SBPs is under evolutionary selection. Given that the conformational dynamics of SBPs appear to have been optimised for their function in solute transport, it is likely that the evolution of catalytic activity in the SBP fold required adaptation of conformational dynamics for catalysis.

We therefore hypothesised that different functional constraints on the conformational landscape for binding proteins and enzymes would have necessitated adaptation of protein dynamics during the evolution of CDT from an SBP. We expected that CDT would retain the ligand-dependent (substrate-dependent) conformational change of SBPs, since large-scale motions of proteins are largely intrinsic to protein architecture and are therefore conserved during evolution (Hollup et al., 2011; Keskin et al., 2000; Marsh and Teichmann, 2014). We reasoned that the closed conformation of CDT is likely the catalytically competent conformation (Section 4.2.1), and that the open-closed conformational change, having been decoupled from its role in solute translocation, would be necessary for catalytic activity only insofar as necessary to allow substrate access to the active site. Therefore, the catalytic activity of CDT could have been improved during evolution by restricting the magnitude of the open-closed conformational change, to restrict unproductive sampling of the open conformation, or by increasing the rate of exchange between the two conformations to enable rapid substrate capture and product release.

5.1.3 Objectives

In this chapter, the dynamics of CDT were characterised in the context of their potential role in the evolutionary adaptation of the SBP scaffold for catalysis. MD simulations of PaCDT were performed to characterise the conformational landscape of the protein and to determine whether the conformational dynamics of SBPs are conserved in PaCDT. New crystal structures of PaCDT in the absence of substrate analogues were also solved in order to experimentally validate the results of the MD simulations.

5.2 Results

5.2.1 Molecular dynamics simulations of PaCDT (Part I)

In the deposited crystal structure of PaCDT (PDB: 3KBR), the protein adopts a closed conformation in complex with 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES). To confirm that the open-closed conformational transition observed in SBPs was conserved during the evolution of CDT activity and to characterise the conformational landscape sampled by PaCDT, including any open conformations, we performed MD simulations of the *apo*-protein.

The ambiguity of the oligomeric state of PaCDT posed an initial problem for the MD simulations. The protein has been reported to be dimeric, based on SEC data (Zhao et al., 1992), although SEC data recently obtained in the Jackson lab showed better agreement with a trimeric structure (observed MW ~ 85 kDa, theoretical MW for dimer/trimer = 59 kDa/88 kDa) (Joe Kaczmarski, unpublished results). In the crystal structure, PaCDT adopts a hexameric structure that would be compatible with either a dimeric or trimeric structure in solution, although the trimer interface has the larger surface area (721 $Å^2$ vs 506 $Å^2$ per interface, Figure 5.3). The trimer interface in PaCDT is formed by interactions between the large domains only, which would enable unrestricted motion of the small domain with respect to the large domain. On the other hand, the dimer interface in PaCDT is formed by interactions between the large domain of one subunit and the small domain of the other subunit, which would restrict large-scale motion of the two domains. Thus, the two oligomeric structures were expected to differ substantially in their dynamic properties, and simulation of the oligomeric structure dominant in solution was considered important. Given the SEC data, PaCDT was simulated as a trimer; further crystallographic evidence that PaCDT is trimeric was obtained subsequently, as shown in Section 5.2.2.

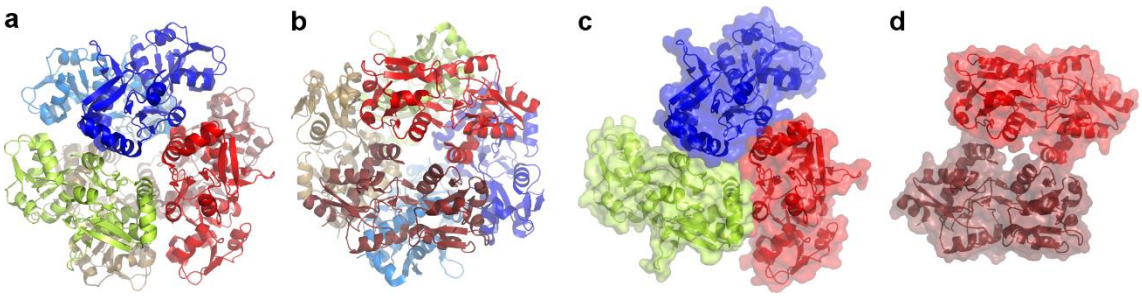


Figure 5.3. Oligomeric structure of PaCDT. (**a-b**) Hexameric assembly of PaCDT (PDB: 3KBR), viewed (**a**) down the 3-fold symmetry axis and (**b**) down the 2-fold symmetry axis. (**c**) Trimeric assembly of PaCDT. (**d**) Dimeric assembly of PaCDT.

Four replicate simulations of 100 ns duration were initialised from the crystallographic trimer assembly of PaCDT with the bound HEPES molecules removed, effectively yielding 12 trajectories of PaCDT monomers with a total simulation time of 1.2 µs. Visual inspection of the trajectories showed that the open conformation of PaCDT was sampled frequently during the MD simulations. This observation was corroborated by analysis of the radius of gyration and interdomain angle (the angle between the centres of mass of the large domain, the hinge region, and the small domain) (Figure 5.4). The radius of gyration of individual PaCDT subunits had a broad distribution from 1.84 nm to 2.08 nm, indicative of substantial conformational changes (Figure 5.4a). The peak in the radius of gyration distribution at 1.89 nm corresponds to the closed conformation (*c.f.* 1.87 nm for the initial, energy-minimized structure), whereas the peak at 2.03 nm corresponds to the open conformation. Likewise, the interdomain angle had a broad distribution, with a peak at ~110° corresponding to the closed conformation and a broad peak at ~150° corresponding to the open conformation (Figure 5.4b). For comparison, the differences in interdomain angles between the crystallographic open and closed conformations of LAOBP and GlnBP are 33° and 24° respectively.

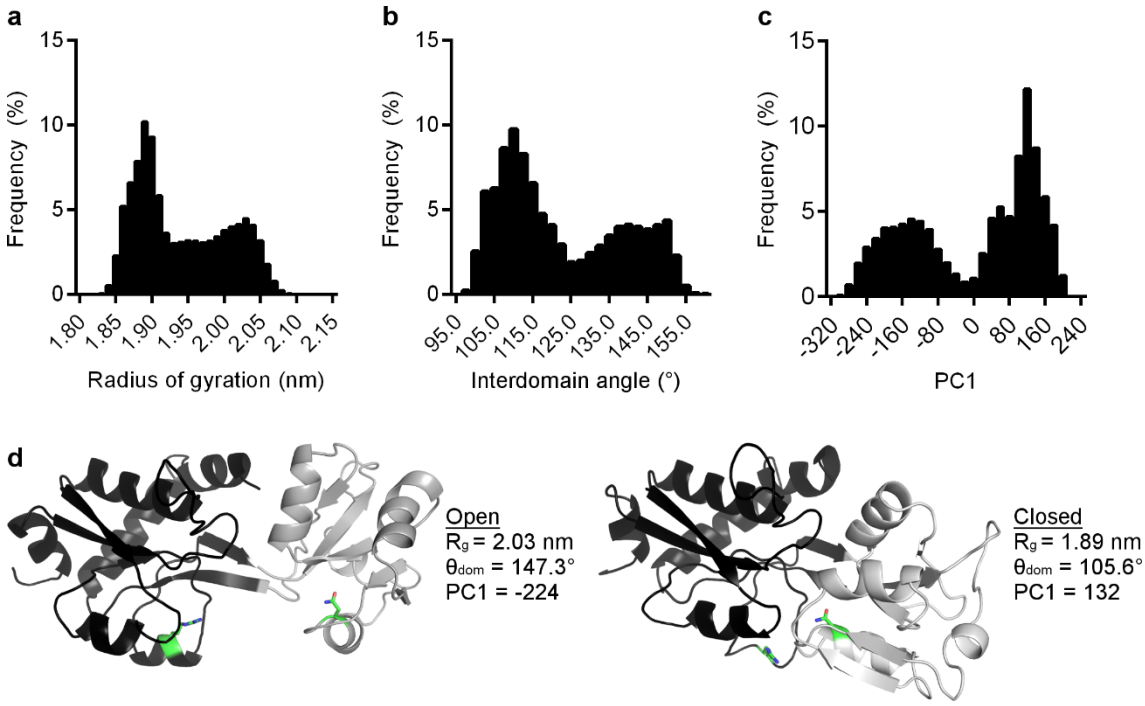


Figure 5.4. PaCDT samples an open conformation during MD simulations. The frequency histograms show (**a**) the radius of gyration (R_g), (**b**) the interdomain angle (θ_{dom}; see Section 0 for definition), and (**c**) the projection onto the first principal component (PC1), for individual PaCDT subunits across the four 100 ns simulations. (**d**) Representative open and closed PaCDT structures. The active site residues Arg85 and Asn133, which are located in different domains, are shown in green.

Principal component analysis (PCA) was used to provide further insight into the large-scale domain motions observed in the MD simulations (Figure 5.5). More than 90% of the variance in atomic positions observed during the simulations could be described using three principal components (PCs) that describe three rigid-body domain motions (Figure 5.5a). PC1 and PC2, accounting for 74% and 12% of conformational variance respectively, correspond to two orthogonal hinge-bending motions (Figure 5.6). PC3, accounting for 5% of conformational variance, corresponds to a hinge-twisting motion, that is, a relative rotation of one domain about the axis between the centres of mass of the two domains (Figure 5.6).

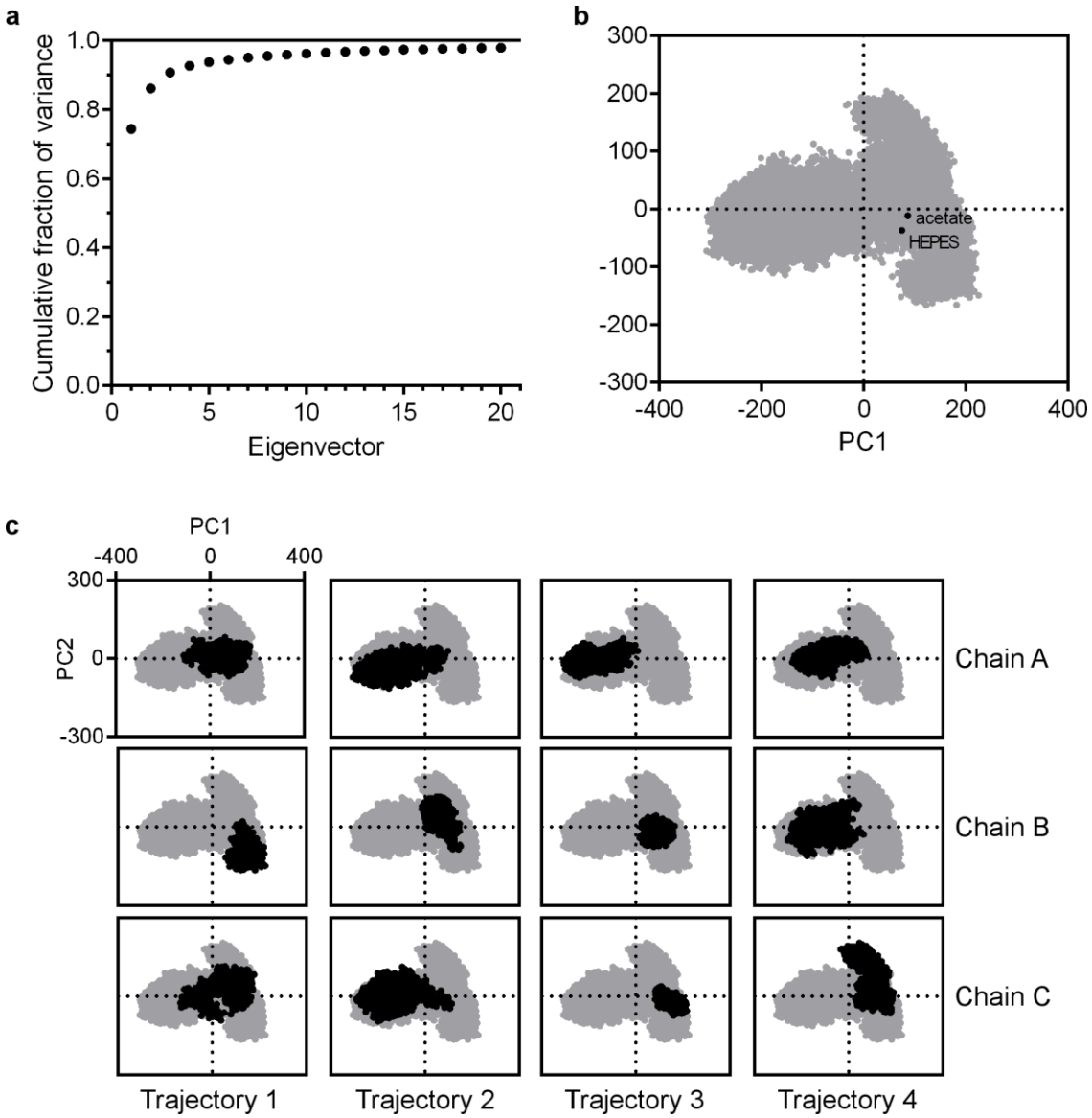


Figure 5.5. Principal component analysis of PaCDT simulations. (**a**) Cumulative fraction of variance in atomic positions during the MD simulations accounted for by the first 20 eigenvectors. One eigenvector alone captures 74% of the variance; three eigenvectors capture >90% of variance. (**b**) Projection of the PaCDT trajectories onto the first two principal components. Projections of the two crystal structures (HEPES-bound and acetate-bound, see Section 5.2.2) onto the first two principal components are also shown. (**c**) Conformational space sampled during individual trajectories. Projections of individual subunits from individual trajectories onto the first two principal components (black), compared with projection of the full simulation dataset onto the first two principal components (grey).

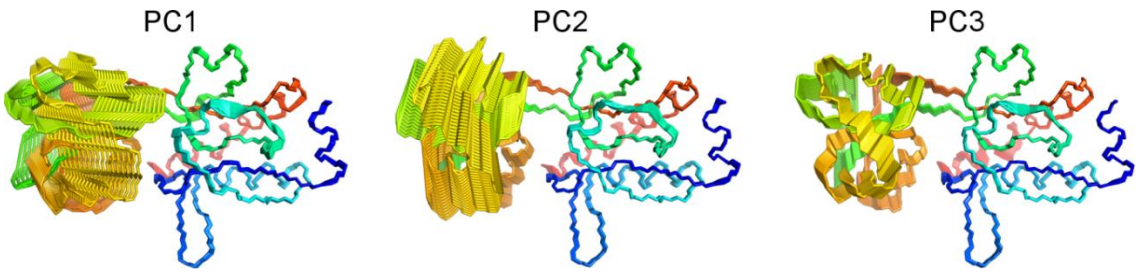


Figure 5.6. Structural interpretation of the major principal components of PaCDT trajectories. The structures illustrating the physical interpretation of the first three principal components were generated by interpolating between structures at the extremities of each principal component axis. PC1 and PC2 reflect orthogonal hinge-flexing motions responsible for the open-closed conformational transition, whereas PC3 reflects a hinge-twisting motion. These three principal components account for >90% of variance in atomic positions during the MD simulations. Note that only one domain makes a significant contribution to each principal component, as expected for a rigid-body domain motion.

PC1 is a useful descriptor of the open-closed conformational change of PaCDT and is negatively correlated with the interdomain angle (Pearson correlation: $r = –0.937$, $P < 10^{-4}$) and the radius of gyration (Pearson correlation: $r = –0.915$, $P < 10^{-4}$). The distribution of PC1 values mirrors the distributions of the interdomain angle and the radius of gyration (Figure 5.4c). The conformational space sampled by PaCDT during the simulations can be visualised by projection of the trajectories onto the PC1 and PC2 axes, as shown in Figure 5.5b.

The time dependence of PC1 for each PaCDT subunit in each trajectory, shown in Figure 5.7, illustrates several points. Firstly, transitions between the open and closed conformations are infrequent in these simulations. In five of the twelve subunit trajectories, a transition from the closed conformation to the open conformation occurred before a stable closed conformation was adopted, suggesting that the conformational change was dependent on forces acting on the protein prior to equilibration. In the remaining seven subunit trajectories, the closed conformation was stable. Secondly,

several abortive transitions between the open and closed conformation are observed, where a closed subunit transitions briefly to a semi-open conformation before reverting to the closed conformation, or *vice versa* for an open subunit. These partial transitions are indicative of a significant energy barrier for both the open-closed and closed-open transitions, and the transient opening motions may also be important for access of the substrate to the PaCDT active site. Thirdly, the open conformation of PaCDT is more dynamic (in terms of large-amplitude domain motion) than the closed conformation, as shown by the larger fluctuations in PC1 for open subunits compared to closed subunits. A similar phenomenon is observed in SBPs, leading to the proposal that SBPs have a natural tendency to oscillate in a way that leads to a productive open-closed transition (Bucher et al., 2011a; Loeffler and Kitao, 2009). Finally, there is no obvious positive or negative cooperativity between subunits of the PaCDT trimer, which is stable throughout the simulations; multiple subunits can exist in the open conformation at the same time. Notably, the quaternary structure of PaCDT appears to have evolved such that it is compatible with the conserved conformational dynamics of the SBP family.

The open-closed conformational change of PaCDT is dependent on structural changes in the two antiparallel β-sheets (hinge strands) connecting the two domains (Figure 5.8). In the crystal structure, the two hinge strands interact via four hydrogen bonds between the following atoms: $Y95^{NH}$ and $K198^{O}$; $L96^{O}$ and $K198^{NH}$; $D98^{NH}$ and $A196^{O}$; $D98^{O}$ and $A196^{NH}$. Similar hydrogen bonding networks are observed for the closed conformation throughout the simulations. However, in the open conformation, the hydrogen bonding network between the two strands is extended, with an additional interaction between $K100^{NH}$ and $D194^{O}$. The extended β-sheets are responsible for holding the two domains further apart, resulting in an open conformation.

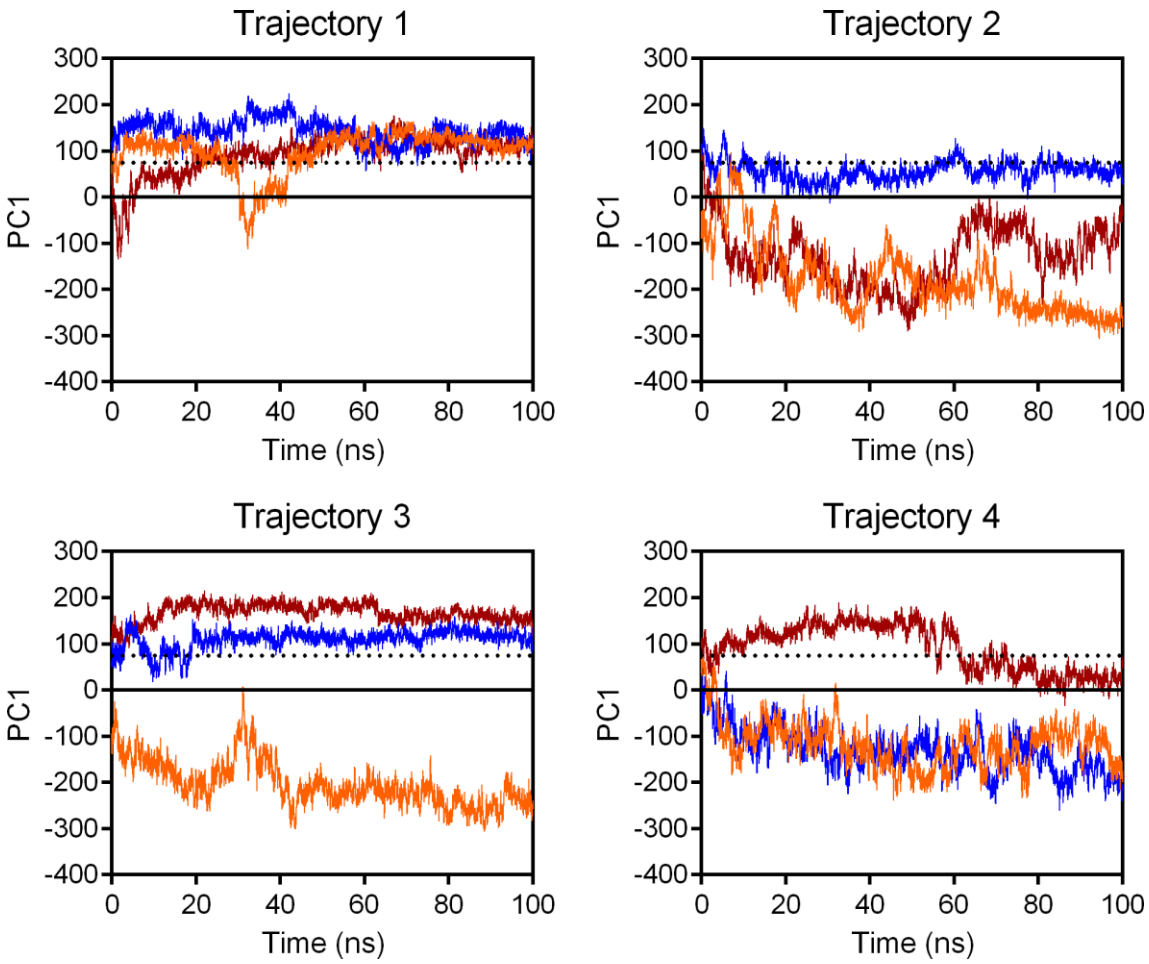


Figure 5.7. Projections of PaCDT trajectories onto the PC1 axis. The three colours represent the three subunits of PaCDT. The dotted line (PC1 = 74.6) represents the crystal structure (PDB: 3KBR).

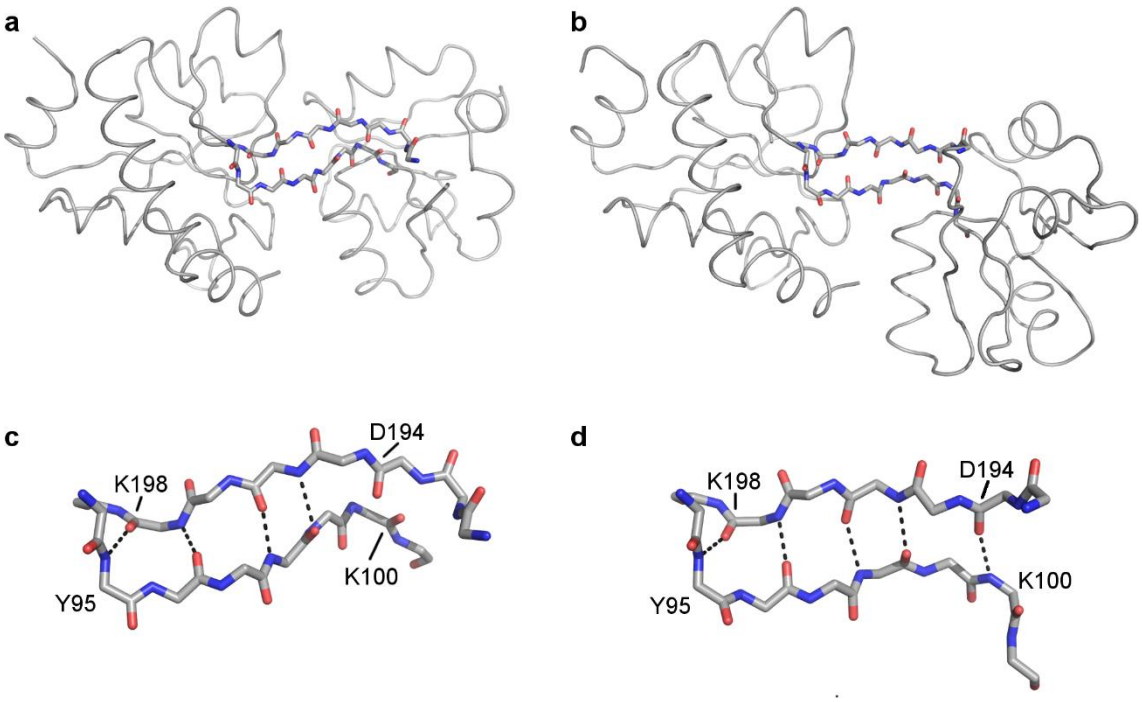


Figure 5.8. Differences in hinge structure between open and closed conformations of PaCDT. (**a**-**b**) The crystallographic closed conformation (**a**) and a representative open conformation (**b**) of PaCDT, with the hinge region shown in stick representation. (**c**-**d**) Interactions between the two hinge strands in the closed conformation (**c**) and open conformation (**d**) of PaCDT. The additional backbone interaction between Lys100 and Asp194 extends the β-sheet structure of the hinge, reorienting the two domains to yield the open conformation, as shown in (**a**-**b**).

The four trajectories of the PaCDT/arogenate complex (Section 4.2.1) were projected onto the PC1 and PC2 axes from the *apo*-PaCDT simulations to compare the large-scale dynamics of each species (Figure 5.9). As expected, PaCDT remained in the closed conformation while bound to substrate. However, the crystallographic closed conformation of the HEPES-bound protein used for docking was rarely sampled, with the enzyme-substrate complex exhibiting a substantial difference in the PC2 coordinate compared to the crystal structure. This result suggests that HEPES stabilises a conformation of PaCDT that is sampled by the *apo*-enzyme but is substantially different from the substrate-bound conformation, and justifies the use of MD simulations to optimise the model of the enzyme-substrate complex obtained by molecular docking.

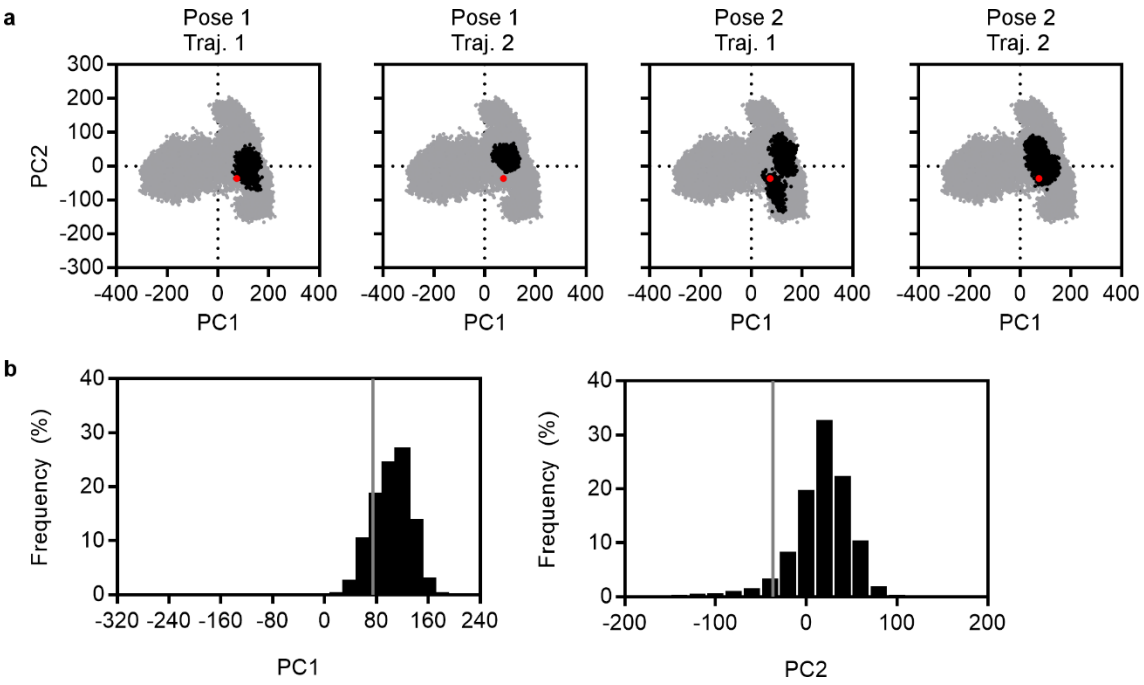


Figure 5.9. Principal component analysis of PaCDT-arogenate simulations. (**a**) Projection of trajectories of the PaCDT-arogenate complex onto the principal components obtained from the *apo*-PaCDT simulations. The red dot represents the crystal structure (PDB: 3KBR). (**b**) Frequency histograms showing conformational sampling along the PC1 and PC2 axes across the complete set of PaCDT-arogenate simulations. The grey lines represent the crystal structure (PDB: 3KBR). The PC2 value associated with the crystal structure is rarely observed during the simulations.

The dynamical behaviour of PaCDT in MD simulations is similar to that previously reported for AABPs, in terms of the large-amplitude hinge-bending and hinge-twisting motions, the magnitude of the open-closed conformational change, and the metastability of the *apo*-closed state (Chu et al., 2014; Loeffler and Kitao, 2009; Pang et al., 2005; Silva et al., 2011a). For example, in the case of HisBP, the *apo*-closed structure was stable over three 70 ns simulations, showing only transient sampling of a semi-open conformation with an interdomain angle ~15° greater than the closed conformation (Chu et al., 2014). The two motions with the largest amplitude, obtained from principal component analysis of simulations of the *apo*-open structure, were hinge-bending and hinge-twisting, and the difference in interdomain angle between the open and closed conformations was ~30°. LAOBP exhibited somewhat different behaviour in MD simulations, with the *apo*-closed state transitioning to the open conformation in all 23 replicates of a 20 ns simulation (Silva et al., 2011a); however, this difference can be accounted for by the higher temperature (318 K) used for the LAOBP simulations compared to the HisBP and PaCDT simulations (300 K). For LAOBP, the principal components accounting for the most conformational variance corresponded to hinge-twisting and hinge-bending motions, respectively.

In summary, PaCDT exhibits similar dynamical properties to SBPs in MD simulations. The open conformation of the enzyme was sampled during the simulations, as shown by radius of gyration and interdomain angle distributions and principal component analysis. However, the closed state of the *apo*-enzyme was stable on the timescale of tens of nanoseconds, and transitions to the open state were observed only in the equilibration phase of each simulation. The motions with the largest amplitude were rigid-body hinge-bending and hinge-twisting motions. To experimentally validate the results of the MD simulations, and to obtain an experimental benchmark for future

computational work (such as adaptive biasing force calculations), we next attempted to solve the crystal structure of PaCDT in the open conformation.

5.2.2 Crystal structures of *apo* PaCDT

Crystallisation of PaCDT in an open conformation required the identification of new crystallisation conditions without HEPES or other substrate mimics using sparse matrix screens. Crystals were obtained from a drop containing 0.2 M ammonium acetate, 0.1 M Tris pH 8.5, 25% (w/v) PEG 3350 as the precipitant. Preliminary diffraction data was collected on optimised crystals grown using 0.2 M ammonium acetate, 0.1 M Tris pH 8.0, 18% (w/v) PEG 3350 as the precipitant; these crystals belonged to the tetragonal space group $P4_322$ and diffracted to 3.1–3.5 Å. An almost identical crystallisation condition (0.2 M ammonium acetate, 0.1 M Tris pH 8.2, 18% (w/v) PEG 3350) produced a crystal belonging to the hexagonal space group *H*3, which diffracted to 2.1 Å. Crystal structures in both space groups were solved by molecular replacement. For the high-resolution structure, the two domains of the previous PaCDT structure (PDB: 3KBR) were used as separate search models to account for the expected rigid body displacement of the two domains corresponding to the open-closed conformational change. The low-resolution structure was then solved by molecular replacement using the refined high-resolution structure as a search model. Data collection and refinement statistics are given in Table 5.1.

Table 5.1. Data collection and refinement statistics for PaCDT.

Structure	**PaCDT** (*P*4$_3$22)	**PaCDT** (*H*3)
PDB code	5JOT	5HPQ
Data collection		
Wavelength (Å)	0.9537	0.9537
Space group	*P*4$_3$22	*H*3
Cell dimensions		
a, *b*, *c* (Å)	95.36, 95.36, 187.87	124.89, 124.89, 40.63
α, *β*, *γ* (°)	90.00, 90.00, 90.00	90.00, 90.00, 120.00
Resolution range (Å)	47.68 – 3.11 (3.32 – 3.11)	38.03 – 2.05 (2.11 – 2.05)
R_{merge} (%)	64.5 (165.3)	11.0 (57.9)
R_{pim} (%)	18.1 (48.2)	5.5 (46.2)
$CC_{1/2}$ (%)	95.7 (63.1)	99.5 (71.4)
$I / \sigma(I)$	5.9 (1.8)	8.3 (1.6)
Completeness (%)	100.0 (100.0)	97.9 (82.3)
Multiplicity	13.3 (12.3)	4.8 (2.5)
Refinement		
Resolution range (Å)	46.72 – 3.11	36.05 – 2.05
Number of reflections	15344	13739
R_{work}/R_{free} (%)	24.25/29.52	17.59/24.10
No. of atoms (*chain A / B / C*)		
Protein	1899 / 1895 / 1883	1870
Acetate	4 / 4 / 4	4
Water	12	55
Average B factors (Å^2) (*chain A / B / C*)		
Protein	32.02 / 31.29 / 39.19	30.88
Acetate	17.70 / 17.21 / 26.17	24.62
Water	10.77	26.86
R.m.s. deviations		
Bond lengths (Å)	0.011	0.016
Bond angles (°)	1.40	1.78
Ramachandran† (%)		
Favoured	94.1	94.9
Allowed	5.5	4.2
Disallowed	0.4	0.8

*Values in parentheses refer to highest resolution shell.

†From PDB validation report.

PaCDT adopted a closed conformation in the absence of HEPES, rather than an open conformation (Figure 5.10), which was unexpected because crystallisation of unliganded SBPs in the closed conformation is rare (but not unprecedented; Flocco and Mowbray, 1994; Oswald et al., 2008). The overall backbone RMSD between the *apo* and HEPES-bound structures is 1.45 Å. The RMSD between the two structures is lower when the two domains are considered separately (1.18 Å for the small domain, 0.62 Å for the large domain), indicating that the structural difference between the two structures is partially explained by a rigid-body displacement of the two domains. PaCDT is more closed in the *apo* structure than in the HEPES-bound structure; specifically, analysis in DynDom (Hayward and Berendsen, 1998) showed that the conformational difference corresponds to an 11° rotation of one domain about an axis perpendicular to the vector between the centres of mass of the two domains (Figure 5.10a). The difference between the two PaCDT structures can also be visualised by projecting the structures onto the principal component axes obtained from the MD simulations (Figure 5.5b); the higher PC1 value for the *apo* structure indicates that the structure is more closed.

The structural changes in PaCDT responsible for the difference in conformation between the *apo* and HEPES-bound structures are minor; interactions between the hinge strands are identical in both structures. A small 25° change in the φ angle of Gly99 in the first hinge strand appears to be responsible for the difference in conformation. Compared with the HEPES-bound structure, an additional interdomain interaction between Asn152 and Asp21 is observed in the *apo* structure, which may contribute to the stability of the closed conformation in the absence of substrate (Figure 5.10b).

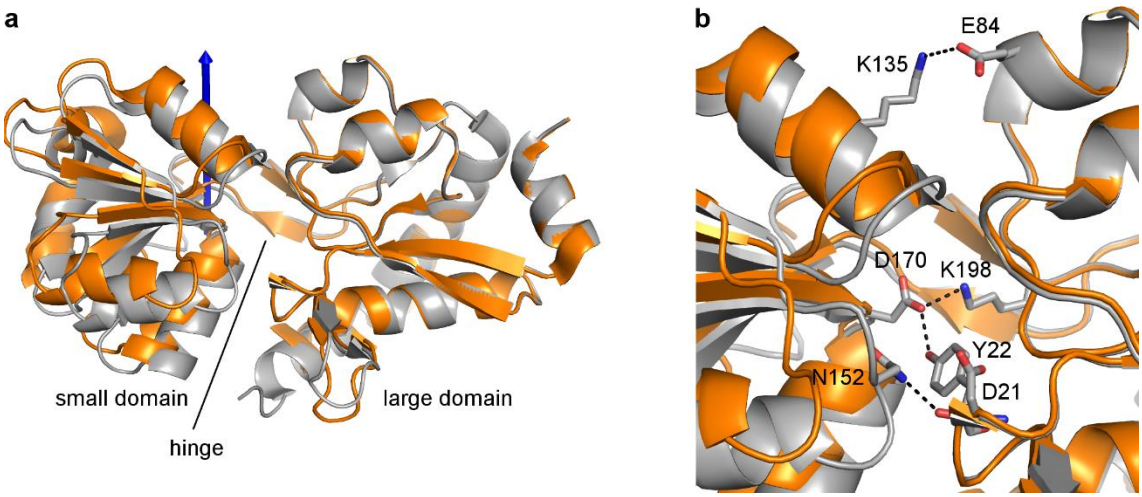


Figure 5.10. Crystal structure of *apo* CDT. (**a**) Comparison of the *apo* (grey) and HEPES-bound (orange) structures of PaCDT. The structures were superimposed using the large domain of each structure. An 11° rotation of the small domain about the axis indicated by the blue arrow is required to superimpose the small domains of the two structures. (**b**) Polar interdomain interactions in *apo*-PaCDT (grey). The HEPES-bound structure (orange) has similar interdomain contacts, except for the interaction between Asn152 and Asp21, which is only possible in the more closed *apo* structure.

Since the active site of PaCDT is located at the interface of the two domains, the rigid-body displacement of the two domains between the *apo* and HEPES-bound structures alters the shape of the active site. Compared to the large and solvent-accessible cavity in the HEPES-bound structure, the active site cavity of the *apo*-PaCDT structure is small and occluded (Figure 5.11). Clearly, a more open conformation is necessary for HEPES, which is bulkier than the cyclohexadiene substrates of the enzyme, to be accommodated in the active site. In the *apo* structure, the highly polar active site cavity contains an acetate molecule from the crystallisation buffer, which binds at the canonical amino acid binding site, and four ordered water molecules (Figure 5.11). The adventitiously bound acetate molecule makes several interactions with the large domain of PaCDT, but just one interaction with the small domain, and is therefore unlikely to make a significant contribution to stabilisation of the closed conformation of the enzyme.

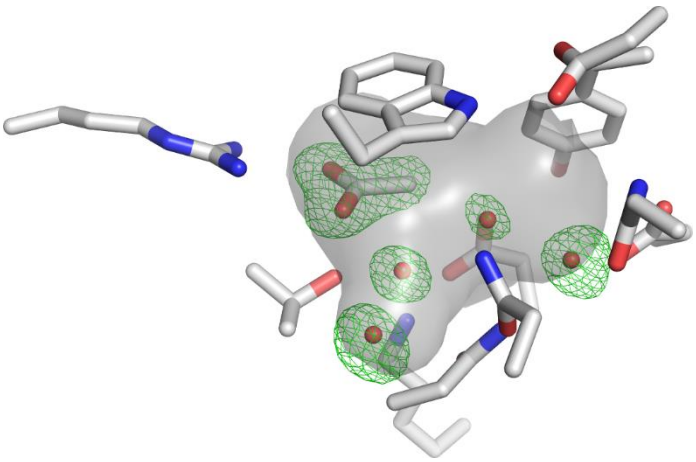


Figure 5.11. Active site cavity of *apo* PaCDT. The surface of the occluded active site cavity is shown in grey. Electron density for the acetate and water molecules is shown by an $mF_o - DF_c$ omit map contoured at +3σ.s

The *apo* structures of PaCDT provide further evidence that the protein has a trimeric structure rather than a dimeric structure. In the HEPES-bound structure and the *apo* structure in space group $P4_322$, PaCDT is assembled into hexamers, which would be compatible with either a dimeric structure or trimeric structure in solution. On the other hand, in the crystal structure in space group *H*3, only the trimeric structure is observed; the trimers are packed head-to-head, rather than packing head-to-tail to form hexamers as in the other crystal structures (Figure 5.12).

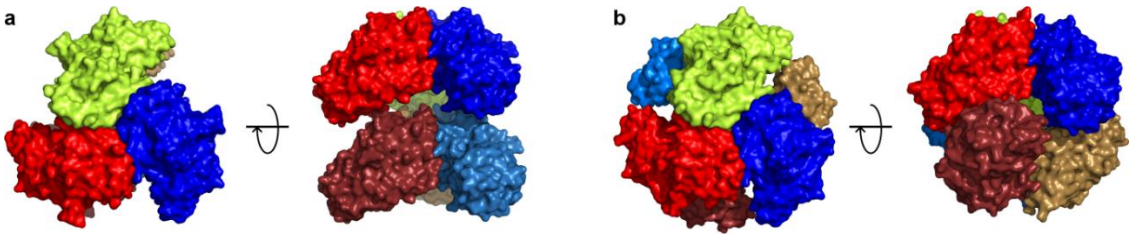


Figure 5.12. Packing of PaCDT crystals in space groups *H*3 and $P4_322$. (**a**) *H*3; (**b**) $P4_322$. For each structure, two PaCDT trimers are shown in two orientations; the first view is down the three-fold symmetry axis and the second view is obtained by a 90° rotation about the axis shown.

5.2.3 Molecular dynamics simulations of PaCDT (Part II)

The crystal structure of *apo*-PaCDT enables an important limitation of the MD simulations reported in Section 5.2.1 to be addressed. The five transitions from the closed conformation to the open conformation that were observed in these simulations each occurred at the beginning of the simulation, before a stable closed conformation had been adopted. These transitions could have been caused by artefactual forces acting on the enzyme prior to equilibration, resulting from the assignment of initial velocities, the perturbative removal of the buffer molecule from the active site, or structural distortions in the initial energy-minimised structure, for example. The crystal structure of *apo*-PaCDT provides a more realistic initial conformation of the enzyme, mitigating some of these problems, and allows the robustness of the observed conformational dynamics to the choice of initial structure to be assessed; the MD simulations were therefore repeated using the *apo*-PaCDT structure.

Another possible critique of the previously reported MD simulations is that only one force field (GROMOS 53A6) was tested; it is therefore unknown whether the observed conformational dynamics are robust to the choice of force field. The GROMOS 53A6 force field uses a united-atom approach, in which a single "united" atom is used to represent an aliphatic carbon atom and any hydrogen atoms bonded to it. This approximation reduces the number of atoms in the system and accordingly reduces computational demand, allowing longer simulations for a given amount of computational time. Use of this approximation is justified by the ability of the GROMOS 53A6 force field to reproduce experimental free energies of solvation (Oostenbrink et al., 2004); nonetheless, to ensure that the simulated conformational dynamics of PaCDT could be reproduced using an all-atom force field, one simulation was performed using the OPLS3 force field.

Four 170 ns simulations were initialised from the crystal structure of the unliganded PaCDT homotrimer, using the GROMOS 53A6 force field (5HPQ-GROMOS simulations). An additional 150 ns simulation was performed using the same structure and the OPLS3 force field (5HPQ-OPLS simulation). Combined with the previous simulations initialised from the HEPES-bound structure (3KBR simulations), these simulations yielded a data set of 27 trajectories of PaCDT monomers with a total simulation time of 4.69 μs.

To compare the conformational space sampled in each set of trajectories, each set of trajectories was analysed separately using principal component analysis, and the resulting principal components were compared quantitatively by computing their pairwise squared inner products and root-mean-square inner products (Table 5.2). These indicators showed that there was considerable overlap in the conformational space sampled in each set of trajectories (root-mean-square inner products for first 10 principal components >0.7). Most importantly, the first and second principal components, corresponding to the open-closed conformational change, described similar motions in each set of trajectories (pairwise squared inner products >0.7). These results demonstrate convergence of the conformational space described by each set of simulations.

Table 5.2. Overlap in conformational space sampled during different simulations of PaCDT. The principal components from the 5HPQ-GROMOS simulations and the 5HPQ-OPLS simulations are compared with the original 3KBR simulations.

Simulation set	5HPQ-GROMOS	5HPQ-OPLS
Pairwise squared inner products		
PC1	0.956	0.893
PC2	0.753	0.888
PC3	0.698	0.533
Root-mean-square inner product (first 10 eigenvectors)	0.731	0.726

The open-closed conformational dynamics of the 3KBR and 5HPQ-GROMOS trajectories were analysed by principal component analysis and interdomain angle and radius of gyration calculations to determine whether the conclusions based on the 3KBR simulations were robust to the choice of initial structure (Figure 5.13). In the 5HPQ-GROMOS simulations, nine subunits remained closed throughout the 170 ns simulation and three subunits arrived at an open conformation, confirming that the closed conformation of PaCDT is stable on the timescale of hundreds of ns, although the open conformation is also energetically accessible (Figure 5.13d). Likewise, projection of the 5HPQ-OPLS trajectories onto the principal components of the 3KBR and 5HPQ-GROMOS trajectories showed one subunit remaining closed for 150 ns and two subunits adopting an open conformation (Figure 5.13d). In contrast to the 3KBR simulations, the five transitions to the open conformation in the 5HPQ-GROMOS and 5HPQ-OPLS simulations occurred after equilibration of the closed structure and are therefore unlikely to be artefacts caused by forces acting on the protein prior to equilibration. The 5HPQ-GROMOS and 5HPQ-OPLS simulations therefore corroborate the main conclusions from the 3KBR simulations, that the closed conformation of PaCDT is stable on the 100 ns timescale, and that sampling of the open conformation is also possible.

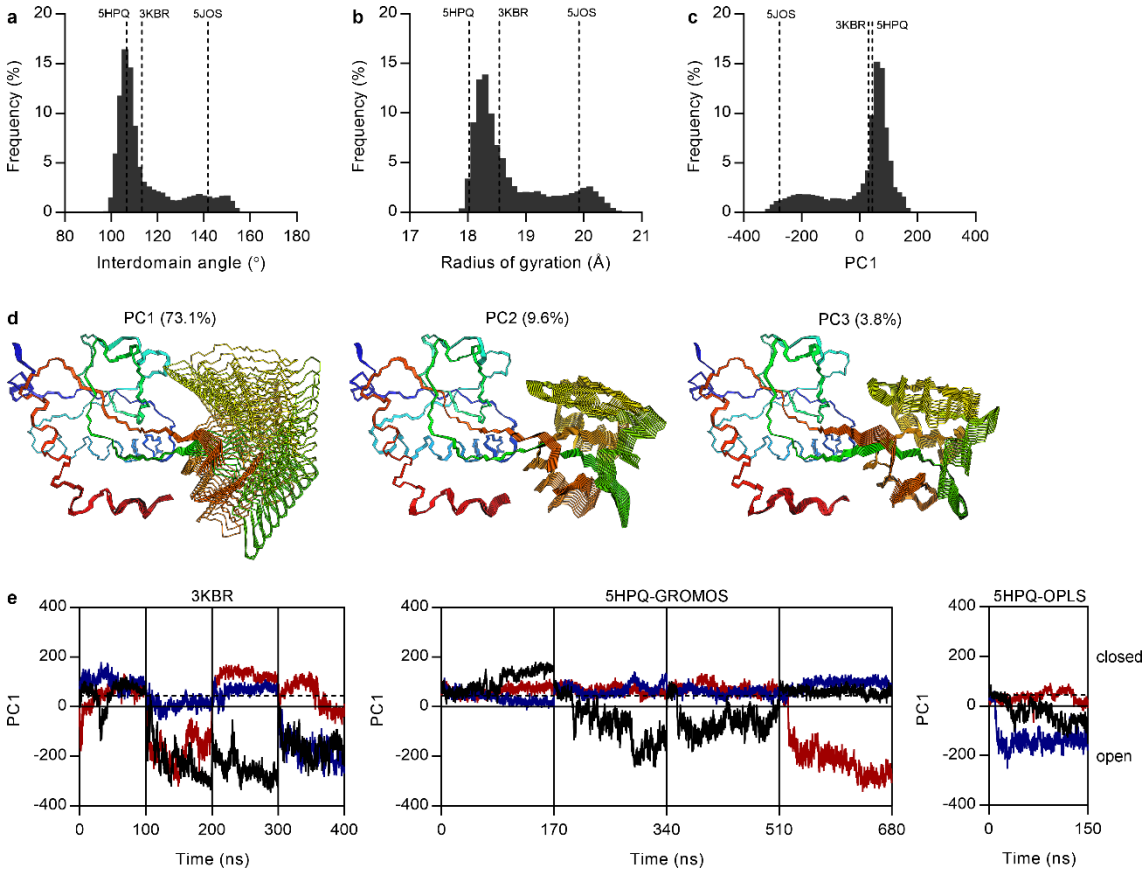


Figure 5.13. Extended molecular dynamics simulations of PaCDT. (a-c) Frequency histograms of the interdomain angle (**a**), the radius of gyration (**b**), and the projection onto the first principal component (PC1) (**c**) for individual PaCDT subunits during the 3KBR and 5HPQ-GROMOS simulations. The corresponding values for the crystal structures of PaCDT (PDB: 5HPQ, 3KBR) and AncCDT-3(P188L) (PDB: 5JOS) are also shown. (**d)** Physical interpretation of the first three principal components derived from the 3KBR and 5HPQ-GROMOS trajectories. The structures were generated by interpolating between structures at the extremities of each principal component axis. The variance in atomic position associated with each principal component is listed in brackets. (**e**) Projection of individual PaCDT trajectories onto the PC1 axis derived from the 3KBR and 5HPQ-GROMOS simulations (4 × 100 ns for 3KBR, 4 × 170 ns for 5HPQ-GROMOS, 1 × 150 ns for 5HPQ-OPLS). Each color represents a subunit of the PaCDT homotrimer. The dotted line represents the crystallographic conformation (5HPQ).

5.3 Discussion

Efficiency in enzyme catalysis depends on pre-organisation of active site residues in a rigid, catalytically competent conformation, and the low efficiency of rationally designed, recently evolved, and promiscuous enzymes is partially attributable to conformational disorganisation or "floppiness" (Bar-Even et al., 2015; Khersonsky et al., 2012; Mabbitt et al., 2016). The binding site of an SBP is not pre-organised; formation of the SBP-ligand complex depends on a significant conformational change from an open conformation to a closed conformation, and the equilibrium between these two conformations controls binding affinity and the rate of solute transport. We therefore questioned how the apparent conflict between the conserved and functionally important conformational dynamics of the SBP fold and the requirement of a pre-organised active site for efficient catalysis could have been resolved in the evolution of CDT from an SBP. To this end, we investigated the conformational dynamics of PaCDT using MD simulations and X-ray crystallography. The MD simulations showed that *apo*-PaCDT undergoes large-scale conformational fluctuations similar to SBPs and can adopt an open conformation, although the closed conformation was also stable on the timescale of tens of nanoseconds. The crystal structure of PaCDT showed that the enzyme can adopt a closed conformation even in the absence of substrate or substrate analogues. Thus, although PaCDT has retained the open-closed conformational equilibrium characteristic of the SBP fold, this equilibrium may have been shifted towards the closed conformation. These results support the view that stabilisation of the closed, catalytically competent conformation of PaCDT may have been an adaptation towards higher catalytic activity in the evolution of this enzyme from a non-catalytic precursor.

Systematic comparison of structural dynamics across protein superfamilies has shown that low-frequency motions are generally a consequence of the global architecture

of a protein rather than specific structural elements and interactions, and that these low-frequency motions are the motions most conserved within protein superfamilies (Keskin et al., 2000; Maguid et al., 2008; Marsh and Teichmann, 2014). Consistent with this view, principal component analysis of the MD simulations showed that the low-frequency motions associated with the SBP fold, hinge-bending and hinge-twisting, are conserved in PaCDT. This result was foreshadowed by an early comparison of the dynamics of three distantly related proteins with the SBP fold using coarse-grained Gaussian network models; the binding protein LAOBP, the transcriptional regulator CysB, and the enzyme porphobilinogen deaminase displayed similar normal modes, based on these simplified models of protein flexibility, despite their significant functional divergence (Keskin et al., 2000). Nonetheless, these conserved collective motions have different consequences in the context of different cellular functions; for example, in porphobilinogen deaminase, which catalyses the oligomerisation of porphobilinogen, SBP-like hinge bending is essential for accommodation of the growing polypyrrole chain in the active site of the enzyme (Bung et al., 2014).

Although PaCDT retained the characteristic conformational dynamics of the SBP fold, the unusual stability of the closed, catalytically competent conformation, evidenced by the crystal structure of the *apo*-enzyme, suggests that the conformational equilibrium was modified as an adaptation towards higher catalytic efficiency. Precedent for evolutionary selection for the position of the open-closed equilibrium of an SBP is evident in the homologous periplasmic domains of the sensor-kinase BvgS, which are constitutively closed (Herrou et al., 2010). In the early stages of the evolution of CDT, prior to optimisation of the conformational equilibrium for catalysis, slow binding of the substrate in a productive conformation may have limited the efficiency of the enzyme. Indeed, the intrinsic lack of a pre-organised binding site in SBPs may have been a major barrier to the evolution of SBP-derived enzymes that catalyse more challenging reactions.

The MD simulations reported in this chapter provide a useful description of the conformational space sampled by PaCDT. The dynamical properties observed in these MD simulations could be reproduced using different initial structures and force fields, and comparison of the principal components obtained from different sets of simulations demonstrated convergence of the conformational space sampled in each set of simulations, that is, sampling of both the open conformation and the closed conformation was sufficient to achieve an adequate description of the conformational space spanned by PaCDT. However, a significant limitation of these MD simulations is that transitions between the open and closed conformations were infrequent, implying that the occupancies of the two conformations are not indicative of their relative energies, that is, the free energy difference between the two conformations did not converge during the simulations. Since experimentally determined rates of the open-closed conformational transition in SBPs range can extend to the millisecond-second range (Gouridis et al., 2014), it is unsurprising that convergence was not achieved within 100–170 ns simulations. Determination of thermodynamic and kinetic parameters for motions on the millisecond-second timescale is well beyond the scope of conventional atomistic MD simulations; alternative computational methods, such as accelerated MD (Hamelberg et al., 2004) or the adaptive biasing force method (Darve et al., 2008), or experimental methods would be required to determine the relative free energies of the open and closed conformations, or kinetic information about the rate of exchange between the open and closed conformations.

The possibility that the closed conformation of PaCDT observed in the crystal structure is a crystal artefact should also be considered, since it is possible that the closed conformation could be stabilised by favourable crystal contacts. However, the observation of the same conformation in multiple, differently packed crystals provides evidence that the stability of the closed conformation cannot be explained solely by

crystal packing. Conversely, given the present lack of experimental evidence that PaCDT samples an open conformation, the validity of the open conformation observed in the MD simulations might be questioned; however, the crystal structure of AncCDT-3(P188L) in an open conformation (Section 6.2.1) provides experimental evidence that the open conformation is accessible in some CDT variants.

Alongside the SEC data, the crystal structure of *apo*-PaCDT establishes that the enzyme is trimeric, contrary to previous reports (Zhao et al., 1992). Since the two trimer interfaces of PaCDT are located on the large domain of the protein, the trimeric assembly was not anticipated to impose additional restrictions on the motion of the small domain with respect to the large domain. Indeed, the MD simulations confirmed that the open-closed transition is fully compatible with the oligomeric assembly of the protein. Thus, the quaternary structure of PaCDT has evolved in such a way that the conserved hinge-bending and hinge-twisting motions of the SBP superfamily can be retained, as seen in the few examples of oligomeric AABPs (Ruggiero et al., 2014). Similar observations of the conservation of large-scale dynamics despite differences in function and quaternary structure have been reported for other protein superfamilies (Luebbering et al., 2012). Consistent with the absence of allosteric regulation in PaCDT, the hinge-bending and hinge-twisting motions of PaCDT subunits are not cooperative; the conformational state of one subunit is independent of the conformational states of the other subunits. Given that CDT is not allosterically regulated, the role of quaternary structure in PaCDT is unclear. The ancestral CDT variants are monomeric, indicating that oligomerisation is not required for catalysis; thus, the quaternary structure of PaCDT may have evolved neutrally. However, it is suggestive that directed evolution of AncCDT-2 for higher prephenate dehydratase activity *in vivo* yielded oligomeric CDT variants; CDT-J3 (Table 4.5, p. 148) exists as a mixture of monomer, dimer and trimer in solution, with the dimer

being the major species (Joe Kaczmarski, unpublished results). Further work is therefore needed to assess the role of quaternary structure in the evolution of CDT.

Experimental characterisation of the solution-state dynamics of PaCDT would be invaluable for confirming the importance of the conformational landscape in the evolution of CDT activity. Fusion of fluorescent or paramagnetic tags to the enzyme would enable "spectroscopic ruler" measurements based on FRET, electron paramagnetic resonance or paramagnetic relaxation enhancement NMR, which could be used to determine whether the open conformation or closed conformation of PaCDT is favoured in solution, identify any minor semi-open or semi-closed conformations with potential functional significance, and determine whether the enzyme adopts a different conformation during substrate turnover. The intrinsic opening and closing rates of the enzyme could be determined by single molecule FRET spectroscopy. Although the quaternary structure of PaCDT would pose problems for these experimental methods, these problems could be dispelled by disruption of the trimer interface by targeted mutagenesis. These experimental techniques could be complemented by theoretical investigation of the conformational landscape of the enzyme, using umbrella sampling simulations to calculate the free energy surface along the coordinate associated with the open-closed conformational change, or using alternative MD methods such as accelerated MD to increase sampling of the conformational landscape.

Additionally, the impact of the open-closed conformational equilibrium on the catalytic activity of CDT has not yet been proven. This could be addressed, for example, by designing allosteric mutations to stabilise and destabilise different conformations of PaCDT and testing the effects of these mutations on catalytic activity. The viscosity dependence of the kinetic parameters of PaCDT could be measured to determine whether the substrate binding or product release steps are rate-limiting, as might be expected if

these steps are dependent on significant conformational changes. Finally, comparison of the conformational dynamics of the less active ancestral CDT variants with the highly active PaCDT could be used to determine whether changes in conformational dynamics are associated with changes in catalytic efficiency and to identify individual mutations responsible for these differences.

5.4 Methods

5.4.1 Materials

A codon-optimised gene encoding PaCDT (UniProt: Q01269; residues 26–268), cloned into the pDOTS7 vector (see Section 4.4.1), was obtained from Joe Kaczmarski.

5.4.2 Molecular dynamics simulations

MD simulations using the GROMOS 53a6 force field (Oostenbrink et al., 2004) were initialised from the HEPES-bound and unliganded PaCDT structures (PDB: 3KBR, 5HPQ). The structure of PaCDT trimer was generated from the monomer structure by application of the crystallographic three-fold rotation operation. The HEPES molecules were removed, and missing residues, missing side-chains, N-terminal acetyl caps, and C-terminal amide caps were modelled as described in Section 4.4.2. MD simulations were performed using GROMACS version 4.5.5 (Pronk et al., 2013) for the HEPES-bound structure and GROMACS version 4.6.5 for the unliganded structure, as described in Section 4.4.2. Following a 1 ns equilibration phase, which was not considered in the analysis, the four simulations of the HEPES-bound structure were continued for 100 ns, and the four simulations of the unliganded structure were continued for 170 ns. The 5HPQ-OPLS simulations were performed using the OPLS3 force field (Harder et al., 2016) in Desmond software (Bowers et al., 2006) by Joe Kaczmarski and Elaaf Mohamed.

5.4.3 Structure analysis

The amount of surface area buried in the hexameric PaCDT assembly was calculated using the PISA server (Krissinel and Henrick, 2007). Principal component analysis, interdomain angle calculations, and radius of gyration calculations were done in Bio3D (Skævern et al., 2014). Analysis was restricted to protein backbone atoms (N, C, Cα) at 0.1 ns intervals for the 3KBR and 5HPQ-GROMOS simulations, and 0.15 ns intervals for

the 5HPQ-OPLS simulations. Similarity between the conformational space sampled in each set of simulations was assessed by performing principal component analysis separately for each set of simulations and computing the root-mean-square inner product of the first 10 eigenvectors. The interdomain angle, θ_{dom}, was calculated as the angle between the centres of mass of three groups of backbone atoms: the large domain (residues 3–97 and 196–233), the hinge region (residues 96–98 and 196–198) and the small domain (residues 98–195).

5.4.4 Crystallisation and structure determination of PaCDT

PaCDT was expressed in *E. coli* BL21(DE3) cells grown in LB media containing 100 mg/L ampicillin to OD_{600} 0.6 at 37 °C, induced with 0.8 mM IPTG, and incubated for a further 20 h at 37 °C. The protein was purified by nickel affinity chromatography under native conditions and size-exclusion chromatography, eluting in Tris buffer (20 mM Tris pH 8.0, 100 mM NaCl, 0.5 mM DTT, 10% glycerol). The crystal belonging to space group $P4_322$ grew in a hanging drop containing 2 μL 10 mg/mL PaCDT and 2 μL 18% (w/v) PEG 3350, 0.2 M ammonium acetate, 0.1 M Tris pH 8.0 at 18 °C. This crystal was cryoprotected in 30% (w/v) PEG 3350, 0.2 M ammonium acetate, 0.1 M Tris pH 8.0. The crystal belonging to space group *H*3 grew in a hanging drop containing 1 μL 10 mg/mL PaCDT and 1 μL 18% (w/v) PEG 3350, 0.2 M ammonium acetate, 0.1 M Tris pH 8.2 at 18 °C. This crystal was cryoprotected in a mixture of 30% (v/v) PEG 400 and 70% (v/v) mother liquor. Both crystals were flash frozen in a nitrogen stream at 100 K and stored in liquid nitrogen. Diffraction data were collected at 100 K on the MX1 beamline at the Australian Synchrotron. The data were indexed and integrated in iMOSFLM (*H*3 structure) (Battye et al., 2011) or XDS ($P4_322$ structure) (Kabsch, 2010), and scaled in Aimless in the CCP4 suite (Winn et al., 2011). The structures were solved by molecular replacement in Phaser (McCoy et al., 2007). For the *H*3 structure, the two domains of the published PaCDT structure (PDB: 3KBR) were used as separate search models. For the

$P4_322$ structure, the refined $H3$ structure was used as a search model; three protein molecules were located in the asymmetric unit. The $H3$ structure was rebuilt manually from a polyalanine model and refined by real space-reciprocal space refinement in Coot (Emsley et al., 2010) and REFMAC5 (Murshudov et al., 1997). Minor adjustments to the P_4322 structure were made in Coot, and reciprocal space refinement with NCS restraints was done in REFMAC5. Data collection and refinement statistics are given in Table 5.1. Structure factors and coordinates for the crystal structures of PaCDT have been deposited in the PDB under accession codes 5HPQ (space group $H3$) and 5JOT (space group $P4_322$).

Chapter Six

Evolution of an enzyme from a solute-binding protein. Part IV: Structure.

6.1 Summary

The aim of the work described in this chapter was to determine the structural basis for the emergence of CDT activity in ancestral SBPs. Firstly, the crystal structures of AncCDT-1 and AncCDT-3(P188L) were solved. These ancestral proteins represent two key intermediates in the evolution of CDT: AncCDT-1 represents the last common ancestor of AABPs and CDTs, and AncCDT-3(P188L) represents the last common ancestor of modern CDTs. We anticipated that comparison of these structures would highlight the historical structural changes that resulted in the evolution of CDT activity, unobscured by the divergent structural changes that altered binding specificity in the Ws0279 lineage and the (largely) functionally neutral changes between AncCDT-3 and PaCDT. Secondly, the crystal structure of *apo*-PaCDT was used to model the CDT-substrate complex and propose a catalytic mechanism for the enzyme. Finally, based on the structural and mutational data acquired in this project, the roles of individual mutations in effecting the gain of CDT activity in AncCDT-1 are discussed.

6.2 Results

6.2.1 Crystal structures of AncCDT-1 and AncCDT-3(P188L)

Crystals of the AncCDT-1/arginine complex grew readily in a wide range of PEG-based conditions, which were identified using sparse matrix crystallisation screens, but these crystals were typically irregular, intergrown and unsuitable for X-ray diffraction. However, single crystals could be obtained by serial microseeding from crystals grown in a hanging drop using 0.2 M lithium sulfate, 0.1 M Tris pH 8.2, 22% (w/v) PEG 3350 as the precipitant, and X-ray diffraction data was collected to a resolution of ~2.6 Å using synchrotron radiation. The crystal structure of the AncCDT-1/arginine complex was solved in space group $P2_12_12_1$ by molecular replacement, using the AncQR/glutamine complex (60% sequence identity) as a search model, with four protein molecules in the asymmetric unit. The crystal used for structure determination possessed an orthorhombic unit cell with one especially long axis (a = 47.0 Å, b = 68.9 Å, c = 318.6 Å), causing the reflections to partially overlap and limiting the accuracy of the integrated data; nonetheless, the structure of the AncCDT-1/arginine complex could be refined to an R_{free} value of 28.7%, and the resulting model was mostly complete, except for a surface loop in the small domain (residues 191–194) and a flexible part of the outer hinge strand (residues 201–204) in some NCS-related subunits. Although the data exhibited an atypical intensity distribution suggestive of twinning (twin fraction estimated from cumulative distribution function for $|L|$ = 17.8%), no merohedral or pseudo-merohedral twin laws are possible in space group $P2_12_12_1$; therefore, the atypical intensity distribution was likely a product of the quality of the data. Complete data collection and refinement statistics are given in Table 6.1.

Table 6.1. Data collection and refinement statistics for AncCDT-1 and AncCDT-3(P188L).

Structure	AncCDT-1	AncCDT-3(P188L)
PDB code	5T0W	5JOS
Data collection		
Wavelength (Å)	0.9501	0.9537
Space group	$P2_12_12_1$	$P4_32_12$
Cell dimensions		
a, *b*, *c* (Å)	47.0, 68.9, 318.6	90.4, 90.4, 101.8
α, *β*, *γ* (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution range (Å)	33.66–2.59 (2.71–2.59)	45.19–2.10 (2.16–2.10)
R_{merge} (%)	15.8 (77.5)	7.3 (76.0)
$CC_{1/2}$ (%)	99.3 (75.5)	99.9 (61.9)
$I / \sigma(I)$	8.2 (2.4)	19.7 (2.0)
Completeness (%)	99.4 (99.3)	99.8 (98.2)
Multiplicity	6.8 (6.9)	11.8 (5.0)
Refinement		
Resolution range (Å)	33.19 – 2.59	44.35 – 2.10
Number of reflections	31387	23894
R_{work}/R_{free} (%)	25.35 / 28.73	17.56 / 20.68
No. of atoms (*chain A* / *B* / *C* / *D*)		
Protein	1769 / 1734 / 1718 / 1721	1932
Arginine	12 / 12 / 12 / 12	-
Citrate	-	13
Benzoate	-	9
Na^+	-	1
Water	21	70
Average B factors (Å^2) (*chain A* / *B* / *C* / *D*)		
Protein	32.99/45.93/58.04/43.58	48.82
Arginine	8.33/24.29/44.08/17.99	-
Water	18.10	44.64
Other	-	38.70
R.m.s. deviations		
Bond lengths (Å)	0.0108	0.0213
Bond angles (°)	1.37	2.05
Ramachandran† (%)		
Favoured	94.9	95.4
Allowed	4.6	3.3
Disallowed	0.4	1.3

*Values in parentheses refer to highest resolution shell.

†From PDB validation report.

In the AncCDT-1/arginine complex, the key residues that interact with the side chain of the amino acid ligand, and thereby determine the binding specificity of the ancestral protein, are Asp19, Ser77 and Gln128, which form hydrogen bonds with the guanidinum group of L-arginine, and Tyr22 and Trp60, which form cation-π interactions (Figure 6.1a). The geometry of the binding site is typical of extant arginine-binding proteins, particularly homologues of LAOBP (PDB: 2Q2A, 1LAF), as expected given the high affinity of AncCDT-1 for multiple cationic amino acids. Several differences between the binding sites of AncCDT-1 and Ws0279 could account for the reduction in affinity for L-arginine from the ancestral protein to the extant protein: the substitution Q128K removes a hydrogen bond with L-arginine, while the substitutions T169F and T131V alter the positions of Tyr22 and Trp60, disrupting the cation-π interactions with the ligand (Figure 6.1b).

Although some residues in the binding site of AncCDT-1 were reconstructed ambiguously (Figure 3.4, p. 103), comparison of the structure of AncCDT-1 with the structures of homologous AABPs suggests that replacement of these residues with alternative predictions of the ancestral state is unlikely to abolish binding of L-arginine. The conservative substitution D19E is found in a number of other cationic AABPs, such as AncQR (PDB: 4ZV1), and maintains an interaction with the guanidinium group of L-arginine. The substitution S77G would likely have the effect of increasing specificity towards L-arginine; in LAOBPs (PDB: 1LAF, 2Q2A), Ser77 is important for maintaining a water-mediated hydrogen bonding network required for binding of L-lysine and L-histidine (Oh et al., 1994). In contrast, in AABPs with higher specificity for L-arginine (PDB: 4ZV1, 2Y7I, 4YMX), this residue is replaced with Ala; the loss of an interaction between Ser and L-arginine can be compensated by recruitment of a water molecule. Finally, the substitution Q128K would abolish a hydrogen bond with L-arginine and likely reduce the affinity of AncCDT-1 for this ligand; nonetheless, this interaction is not

essential for binding of cationic amino acids, as shown by replacement of Gln128 by Leu (PDB: 1LAF) and Lys (PDB: 3K4U) in other AABPs with similar specificity. Thus, the conclusion that AncCDT-1 is an AABP specific for cationic amino acids is apparently robust to uncertainty in the reconstruction of the ancestral sequence, although variations in the precise specificity of the ancestral protein cannot be excluded.

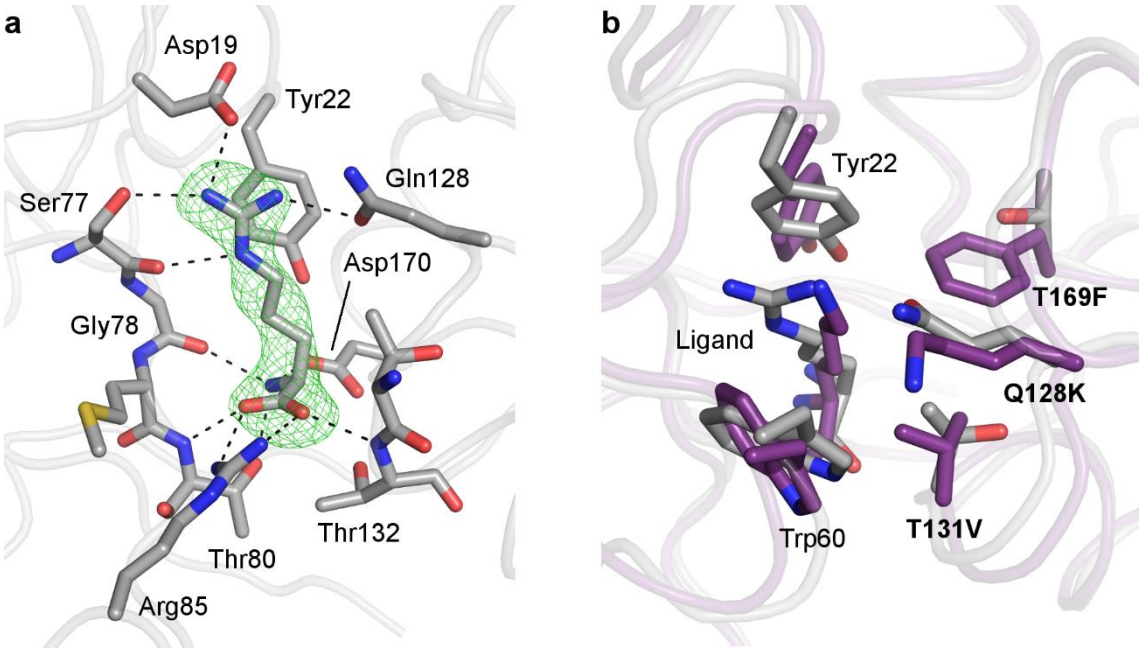


Figure 6.1. Crystal structure of AncCDT-1. (**a**) Binding site of AncCDT-1 (chain D). Electron density for the ligand, L-arginine, is shown by an $mF_o - DF_c$ omit map contoured at +3σ. Trp60 is situated on top of the ligand and is omitted for clarity. Minor NCS differences are not shown. (**b**) Comparison of the AncCDT-1/arginine (grey) and Ws0279/lysine (purple) complexes. Substitutions between AncCDT-1 and Ws0279 with potential relevance for the difference in binding specificity are shown.

Next, the crystal structure of AncCDT-3(P188L) was solved. Crystals of AncCDT-3(P188L) were obtained from crystallisation conditions containing high concentrations (>1 M) of sodium citrate or ammonium citrate. The crystal used for structure determination grew from a sitting drop containing 0.1 M Tris pH 8.0, 1.2 M sodium citrate as the precipitant, and diffracted to ~2.1 Å using synchrotron radiation. The structure of AncCDT-3(P188L) was solved by molecular replacement, using the two domains of PaCDT as separate search models, and refined to an R_{free} value of 20.7%. Data collection and refinement statistics are given in Table 6.1 (p. 207).

The crystal structure of AncCDT-3(P188L) shows the protein in an open conformation (Figure 6.2a), providing experimental confirmation that the open-closed conformational change associated with SBPs is compatible with CDT activity. The protein forms a crystallographic dimer, with the small domain of one subunit situated in the cavity between the two domains of a second subunit in a neighbouring asymmetric unit. This oligomeric structure is most likely a crystal packing artefact; size-exclusion chromatography indicated that the AncCDT-3(P188L) monomer predominates in solution (Figure 6.3). Residual electron density in the crystal structure indicated that the crystallographic dimer is bridged by small molecules (Figure 6.2). One molecule, tentatively modelled as the benzoate anion based on the electron density, chemical environment and its plausibility as an impurity in the sodium citrate solution used for crystallisation, binds in a hydrophobic pocket between the two domains of one subunit. The two small domains of the crystallographic dimer are bridged by a second molecule located on the two-fold symmetry axis, which was modelled as the citrate anion in two possible orientations. The putative binding site for the departing carboxylate group of the substrate of CDT is occupied by this citrate molecule, whose terminal carboxylate groups interact with Asn128, Thr132 and Asn133 in AncCDT-3(P188L) (Figure 6.2d).

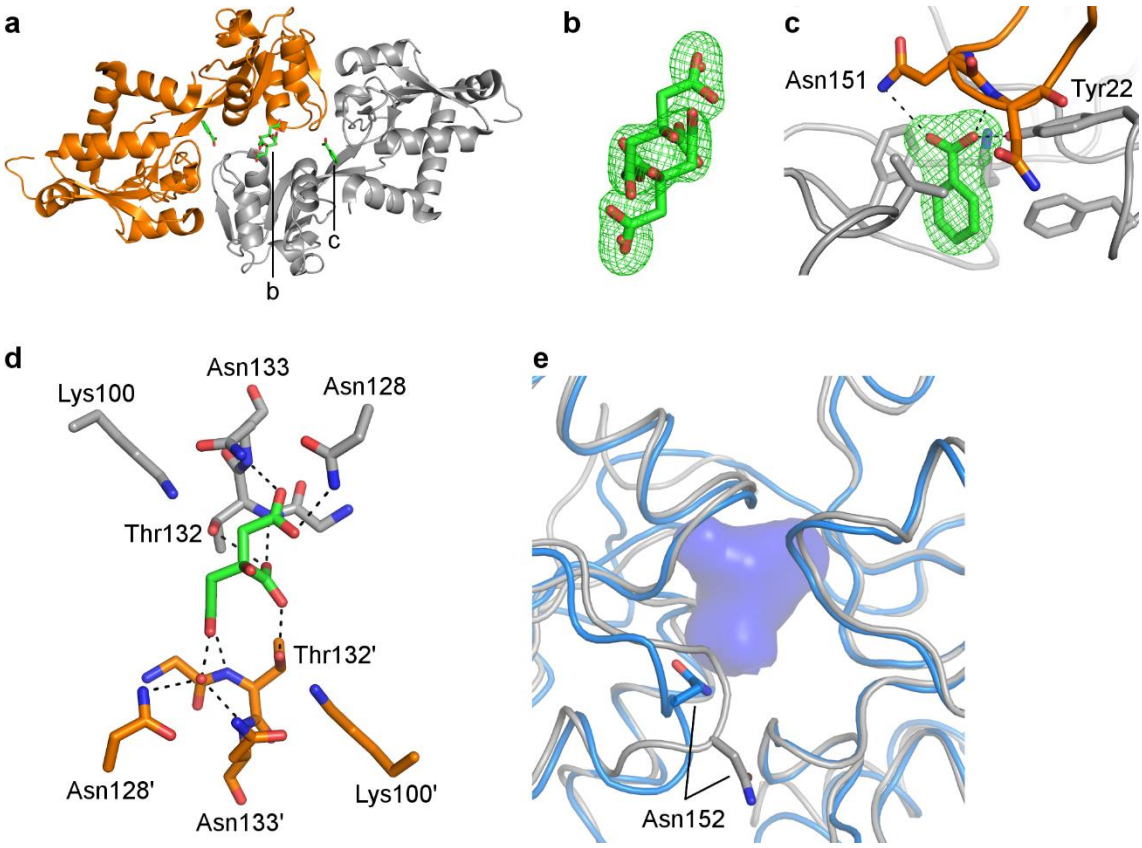


Figure 6.2. Crystal structure of AncCDT-3(P188L). (**a**) Crystallographic dimer of AncCDT-3(P188L) in an open conformation. (**b**–**c**) Electron density for citrate (**b**) and benzoate (**c**) in the AncCDT-3(P188L) structure, illustrated by $mF_o - DF_c$ omit maps contoured at +3σ. The citrate molecule is situated on the crystallographic two-fold symmetry axis in two possible orientations, as shown. (**d**) Binding mode of citrate to AncCDT-3(P188L). (**e**) Alignment of the AncCDT-3(P188L) (grey) and *apo*-PaCDT (blue) structures. The two domains of AncCDT-3(P188L) were superimposed separately onto the PaCDT structure. The surface of the active site cavity of PaCDT is shown.

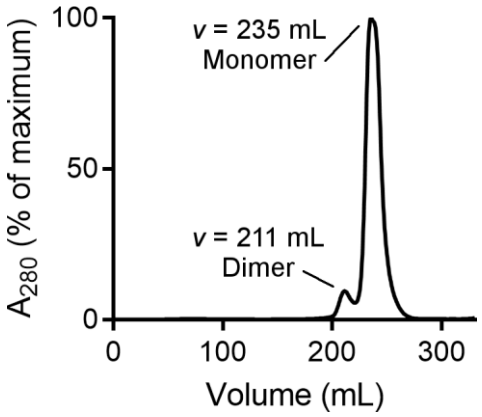


Figure 6.3. Size-exclusion chromatogram of AncCDT-3(P188L). The peak at 235 mL corresponds to the monomer (calculated MW ~ 18 kDa, theoretical MW = 28 kDa) and the peak at 211 mL corresponds to the dimer (calculated MW ~ 42 kDa, theoretical MW = 56 kDa). Although the SEC data is not quantitative, the monomer peak at 235 mL can be assigned confidently by analogy with other monomeric AABPs that eluted at a similar volume. Calibration data for the SEC column was provided by Nicholas Fraser.

The structures of AncCDT-3(P188L) and PaCDT can be compared using the metrics introduced in Section 5.2.1. The interdomain angle of the ancestral protein is 141°, compared with an interdomain angle of 106° for the crystal structure of *apo*-PaCDT. Projection of the AncCDT-3(P188L) structure onto the principal component axes obtained from MD simulations of PaCDT gives a PC1 value of –236 and a PC2 value of –62. Thus, the interdomain angle and PC values for the AncCDT-3(P188L) structure fall within the ranges expected from MD simulations of the open conformation of PaCDT (Figure 5.4), which provides evidence that the global conformation of AncCDT-3(P188L) is not grossly distorted by the unusual crystal packing arrangement. Separate superimposition of the two domains of AncCDT-3(P188L) onto the structure of *apo*-PaCDT gives a backbone RMSD of 1.07 Å for the large domain and 1.81 Å for the small domain. Most of the residues surrounding the active site adopt a similar conformation in AncCDT-3(P188L) and PaCDT. A notable exception is seen in the loop containing Asn152, which is oriented away from the active site in AncCDT-3(P188L) (Figure 6.2e). However, since this loop is located at the dimer interface of the protein and interacts with the bridging benzoate molecule, this alternative conformation is most likely stabilised by crystal contacts and irrelevant for function.

6.2.2 Structural basis for catalytic activity in PaCDT

The crystal structure of *apo*-PaCDT showed some important differences with the HEPES-bound structure initially used for docking (Section 4.2.1): in the *apo* structure, PaCDT adopted a more closed conformation and exhibited an occluded active site cavity, whereas in the more open HEPES-bound structure, the active site cavity was larger and solvent-accessible (Section 5.2.2). We therefore hypothesised that the *apo* structure of PaCDT would provide a conformation of the enzyme more relevant for substrate binding and catalysis, and modelled the structure of the L-arogenate and prephenate complexes using molecular docking based on the *apo* structure. In their respective lowest energy poses, L-arogenate and prephenate adopted the expected orientation (Figure 6.4); the α-amino acid or α-keto acid moieties bind at the conserved structural motif that recognises the same functional groups in AABPs, and the departing carboxylate group binds at the N-terminus of a helix around Asn128-Asn133, as expected based on previous MD simulations of the PaCDT/arogenate complex and binding of a carboxylate at this site in the AncCDT-3(P188L) structure. The shape complementarity between the active site of *apo*-PaCDT and its substrates was high, suggesting pre-organisation of the active site for catalysis. In this respect, CDT can be distinguished from SBPs, which adopt an open conformation in the unliganded state, such that only partial pre-organisation of the binding site is possible (by separate pre-organisation of the binding interfaces of each domain).

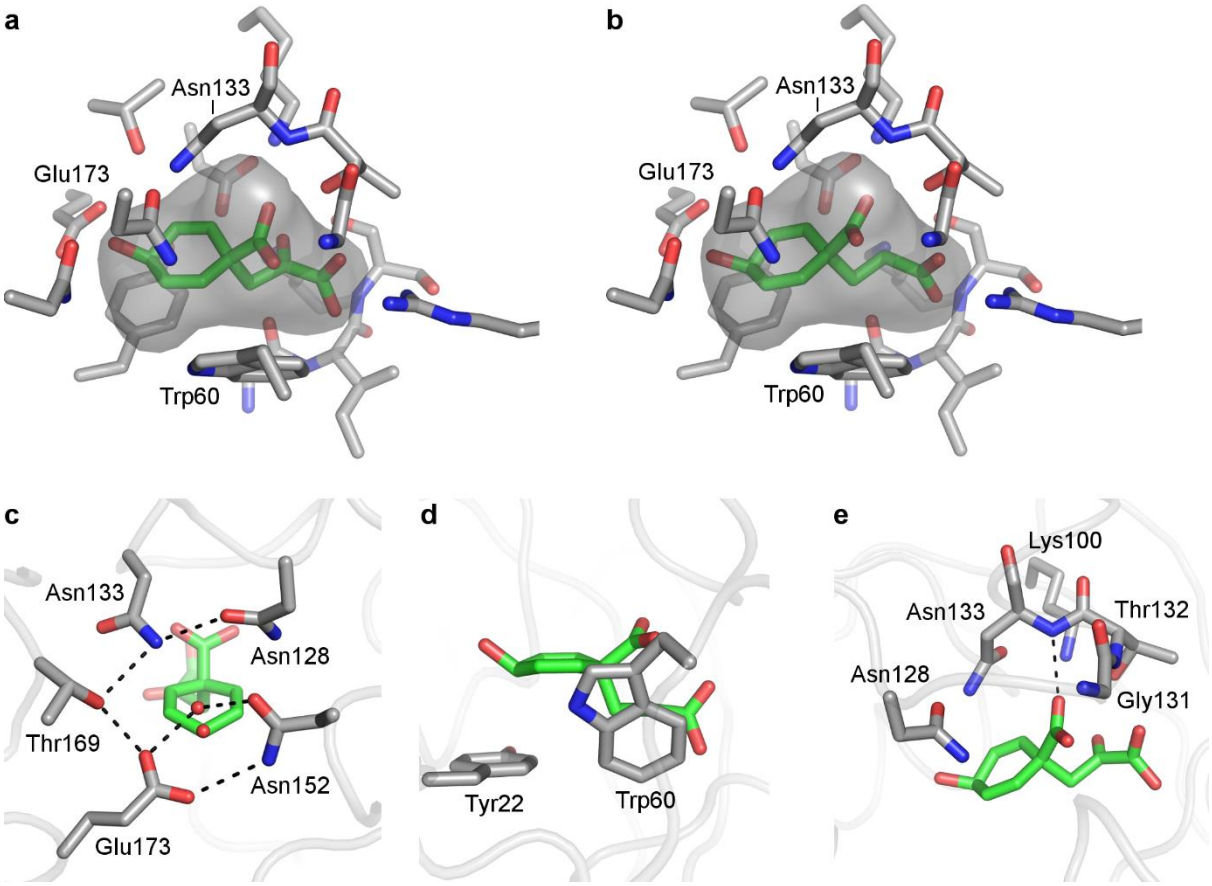
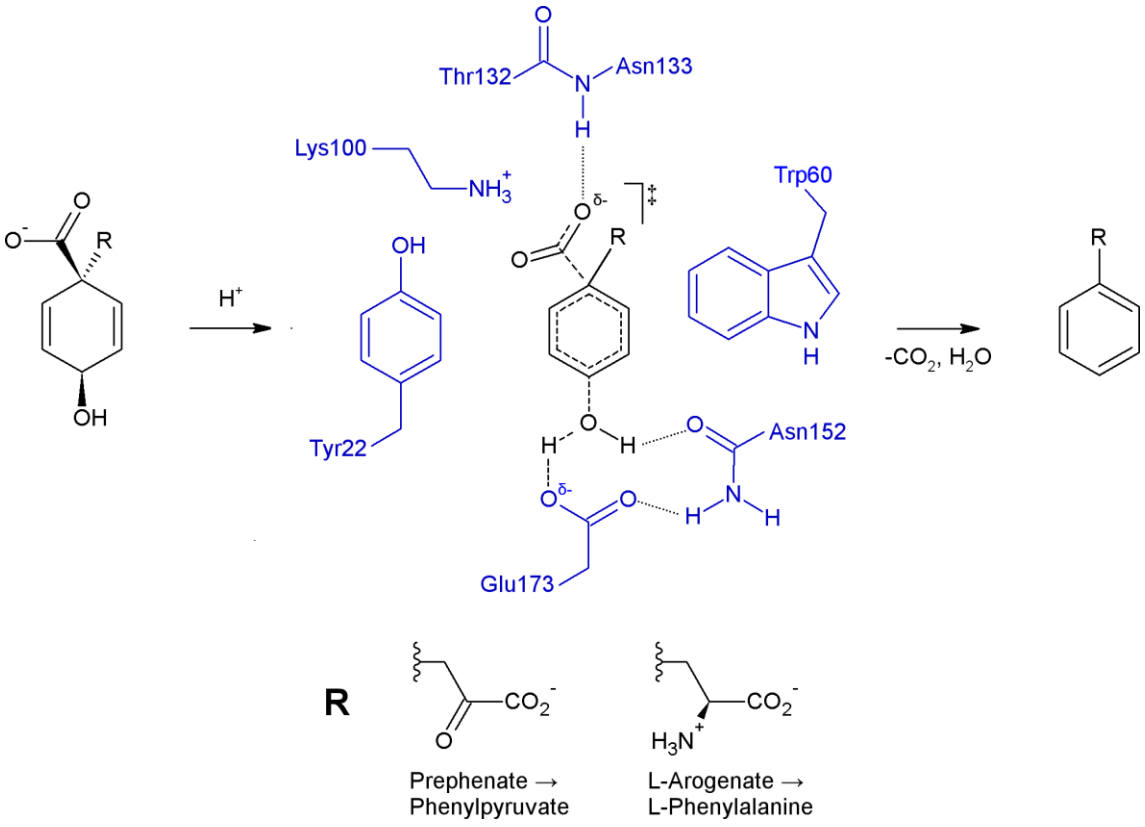


Figure 6.4. Predicted substrate binding modes in *apo*-PaCDT and implications for catalysis. (**a-b**) Lowest energy poses of (**a**) prephenate (E = –25.9 kJ/mol) and (**b**) L-arogenate (E = –19.7 kJ/mol) in the *apo*-PaCDT structure. The surface of the active site cavity is shown in grey. (**c**) Glu173 is poised for proton donation to the departing hydroxyl group of prephenate by hydrogen bonding interactions with neighbouring residues. An ordered water molecule in the crystal structure is shown in red. (**d**) π stacking interactions with Tyr22 and Trp60 could stabilise the developing π system in the transition state. (**e**) Lys100, Asn128, Thr132 and Asn133 could stabilise the departing carboxylate group of prephenate.

In the lowest energy pose for each of the PaCDT-substrate complexes, the hydroxyl group of the substrate is in close proximity with Glu173. This residue is proposed to be the general acid required for protonation and elimination of the hydroxyl group of the substrate (Scheme 6.1). Prediction of the pK_a of Glu173 in PROPKA (Olsson et al., 2011) using the *apo*-PaCDT structure gives a pK_a of 7.75, compared with a pK_a of 6.57 using the HEPES-bound structure; thus, Glu173 is expected to be protonated at neutral pH, consistent with its proposed role as a general acid. Desolvation is the primary cause of the elevated pK_a of Glu173 in PaCDT; this residue protrudes from a hydrophobic pocket and is surrounded by Tyr22, Phe156 and Met167. The deprotonated state of Glu173 is also destabilised by electrostatic interactions with Asp21 and Asp170, which make a minor contribution to the elevated pK_a of this residue.



Scheme 6.1. Proposed basis for transition state stabilisation in PaCDT. The transition state for the concerted mechanism is shown, but similar interactions could also contribute to transition state stabilisation in the stepwise mechanism.

The structure of *apo*-PaCDT reveals an intricate hydrogen bonding network that appears to stabilise Glu173 and poise the enzyme for protonation and elimination of the hydroxyl group of prephenate or L-arogenate (Figure 6.4c). Glu173 interacts directly with Asn153 and Thr169, which is connected *via* this hydrogen bonding network to Asn133 and Asn128. In the HEPES-bound structure, the hydrogen bonding network is disrupted by a clash between Asn152 and the hydroxyethyl group of HEPES, which forces Asn152 from the active site of the enzyme. Given the configuration of active site residues in the *apo*-PaCDT structure, the hydroxyl group of the substrate could plausibly interact with Asn133, Asn152 and Glu173 *via* hydrogen bonding interactions. The feasibility of this binding mode of the hydroxyl group of the substrate is demonstrated by the fact that this position is occupied by a water molecule in the *apo*-PaCDT structure (Figure 6.4c). The position of this water molecule possibly reflects the optimal position of the departing hydroxyl group in the transition state; compared with the position of this group in the model of the enzyme-substrate complex, the water molecule is displaced by ~1 Å away from the approximate plane of the cyclohexadiene ring.

In addition to protonation and elimination of the hydroxyl group of prephenate and L-arogenate, PaCDT must also promote decarboxylation and aromatisation of these substrates. As discussed in Section 3.1.2, it is not known whether these steps occur by a concerted mechanism, in which CO_2 and H_2O are eliminated from the substrate simultaneously, or by a stepwise mechanism, in which elimination of H_2O, yielding the stabilised divinyl carbocation as an intermediate, precedes decarboxylation (Scheme 3.3, p. 98). Given the low basicity of the alcohol group ($pK_a \sim -2$), the concerted mechanism is probably more likely *a priori*. In either case, however, Trp60 and Tyr22 have a potential role in transition state stabilisation; these residues are positioned such that they could stabilise the developing π system of the aromatic ring in the transition state *via* π-stacking interactions (in the concerted mechanism), or could stabilise the carbocation intermediate

via cation-π interactions (in the stepwise mechanism) (Figure 6.4d). The predicted poses of L-arogenate and prephenate in *apo*-PaCDT also show that Lys100, Asn128, Thr132 and Asn133 could form electrostatic interactions and hydrogen bonds that stabilise the departing carboxylate group of the substrate and promote decarboxylation (Figure 6.4e).

6.2.3 Structural basis for evolution of CDT activity

Given the substrate binding mode and catalytic mechanism for CDT proposed in the previous section, the emergence of CDT activity in the SBP fold can be rationalised by comparing the structures of CDT variants (PaCDT and AncCDT-3(P188L)) with the structures of the ancestral AABP (AncCDT-1) and extant SBPs that diverged from this ancestral protein (Ws0279 and Pu1068).

The structural motif that binds the α-amino acid functionality of L-arogenate in CDT was inherited from its ancestral AABP (Figure 6.5). Comparison of the model of the PaCDT/arogenate complex with the structure of AncCDT-1 suggests that the binding mode of the α-amino acid functionality is identical in each protein; the carboxylate group is bound by Ser/Thr80, Arg85 and Thr132, and the amino group is bound by Gly78 and Asp170. Unlike AABPs, however, CDT has dual specificity for α-amino acids and α-keto acids (L-arogenate and prephenate, respectively). The substitutions Q100K and L198K appear to contribute to the change in specificity; electrostatic shielding of Asp170 by these Lys residues would promote binding of the ketone group. Thr80 could also act as a hydrogen bond donor to the ketone group of prephenate.

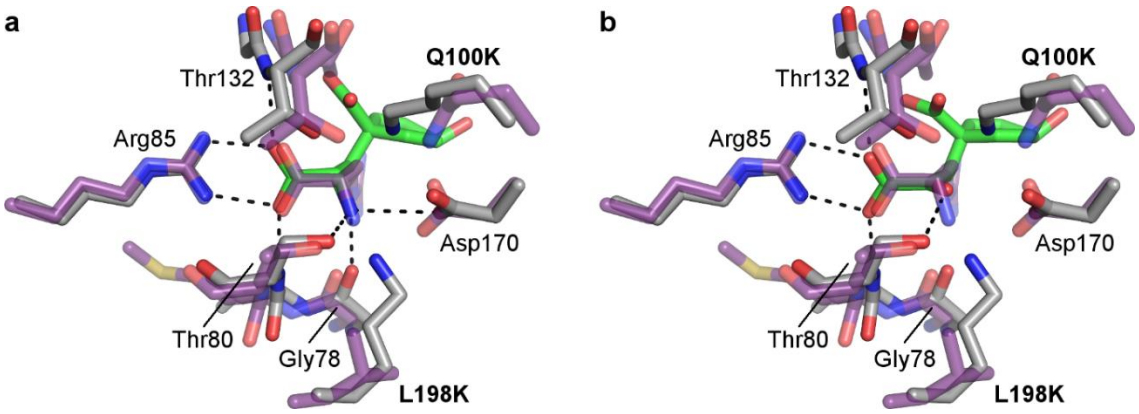


Figure 6.5. CDT inherited the amino acid-binding structural motif from AABPs. Comparison between AncCDT-1 and *apo*-PaCDT, docked with (**a**) L-arogenate and (**b**) prephenate. Positions are labelled with the corresponding residue in AncCDT-1 and AncCDT-3, if conserved in both proteins, or with the corresponding substitution between AncCDT-1 and AncCDT-3.

The evolution of CDT from AABPs required a significant conformational change in Trp60, which reshaped the binding site of the ancestral SBP and obstructed the binding site for the amino acid side chain (Figure 6.6a). Two substitutions appear to be responsible for this conformational change in AncCDT-1: D19T and A20G. In CDT variants, Trp60 is stabilised by a hydrogen bond from Asp21. However, Asp21 is solvent-exposed in AncCDT-1; the substitution D19T is needed to reorient this residue towards the binding site by creating a hydrogen bond between Thr19 and Asp21. The substitution A20G enables rotation of Gly20 into a backbone conformation disfavoured for non-glycine residues (φ = 72.3°, ψ = 19.0° in *apo*-PaCDT), which is needed to accommodate Thr19 in the conformation required for the hydrogen bonding network with Asp21 and Trp60. The effect of this substitution can be seen by comparing PaCDT with ArgBPs from *Caldanaerobacter subterraneus* (CsArgBP; PDB: 4YMX) and *Streptococcus pneumoniae* (SpArgBP; PDB: 4H5F), which have a similar constellation of residues in this loop (Ser19, Ala/Pro20, Asp21); in these ArgBPs, the side chain of Ser19 is prevented

from adopting the conformation observed in PaCDT due to steric exclusion by the carbonyl group of Ser19, which is repositioned by the A20G substitution (Figure 6.6b).

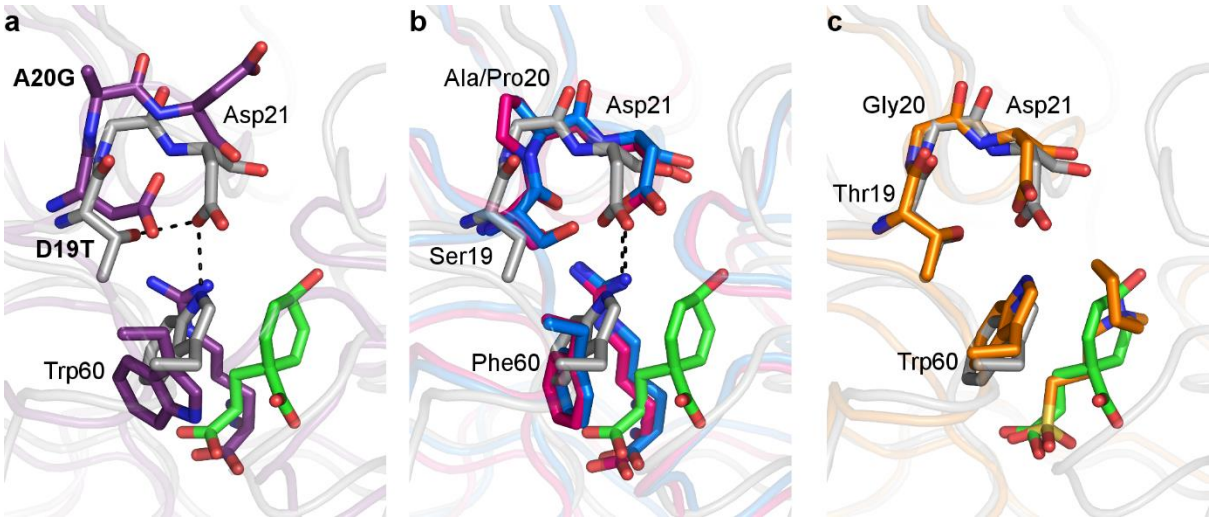


Figure 6.6. Role of Trp60 and surrounding residues in the evolution of CDT. *Apo*-PaCDT (grey, with docked prephenate shown in green), AncCDT-1 (purple), SpArgBP (pink; PDB: 4H5F), CsArgBP (blue; PDB: 4YMX), Pu1068 (orange, bound to NDSB-221) are shown for comparison. (**a**) Comparison of PaCDT with AncCDT-1 shows that the conformational change of Trp60 reshapes the ancestral binding site. (**b**) Comparison of PaCDT with SpArgBP and CsArgBP shows that the A20G substitution enabled rotation of the carbonyl group of Thr19, permitting an alternative side-chain conformation. Residue labels refer to SpArgBP and CsArgBP. (**c**) Trp60 adopts the same conformation in PaCDT and Pu1068.

Comparison of CsArgBP and SpArgBP with AncCDT-1 suggests a plausible sequence for the D19T and A20G substitutions in AncCDT-1 to effect the conformational change of Trp60. With Asp21 and Trp60 already present in AncCDT-1, the D19T substitution could have occurred first, removing the interaction between L-arginine and Asp19, but enabling rotation of Asp21 towards the binding site and producing a compensatory interaction between L-arginine and Asp21, as seen in CsArgBP and SpArgBP. Thus, the substitution D19T would likely allow retention of the ancestral arginine-binding function. The substitution A20G would finally enable the conformational change of Trp60, supported by Thr19 and Asp21 in the appropriate geometry.

In addition to CDT, the Thr19-Gly20-Asp21-Trp60 tetrad is also found in Pu1068, and reconstructed with high statistical confidence in AncCDT-2 (Figure 3.4). As expected on the basis of conservation of this sequence motif, Trp60 adopts the same conformation in Pu1068 and PaCDT (Figure 6.6c). This result suggests that the conformational change of Trp60 occurred as an evolutionary adaptation towards a different function prior to the evolution of CDT activity.

As discussed in the previous section, catalysis of the decarboxylative aromatisation of cyclohexadienols by CDT is seemingly dependent on a conserved general acid, Glu173, as well as a network of hydrogen bonding interactions that poise Glu173 for proton donation and stabilise the departing hydroxyl and carboxylate groups of the substrate. The substitutions required to recapitulate these critical interactions and functionalise the active site of AncCDT-1 are V173E, Q100K, Q128N and S133N; Asn152 and Thr169 are already present in the ancestral protein (Figure 6.7a). An additional substitution, T131G, resolves a steric clash between Thr131 and the departing carboxylate group of the substrate; the importance of this substitution was evidenced by the directed evolution experiments described in Chapter 4. The reconstruction of Gln128,

Ser133, Asn152 and Val173 was ambiguous in AncCDT-1 (Figure 3.4); thus, the possibility that alternative substitutions were responsible for the historical gain of CDT activity cannot be excluded. Of these residues, only Gln128 is involved in protein-ligand interactions in AncCDT-1. The ambiguous reconstruction of Ser133 and Val173, in particular, was a consequence of the fact that these residues are located at the periphery of the binding site in AABPs and are not important for ligand binding; these positions are therefore quite variable in extant homologues of Ws0279. In AncCDT-1, Asn133 is a plausible alternative reconstruction to Ser133 (*P*(Ser) = 0.383, *P*(Asn) = 0.175), raising the possibility that the emergence of this catalytic residue, like Asn152 and Thr169, could have preceded any functional divergence.

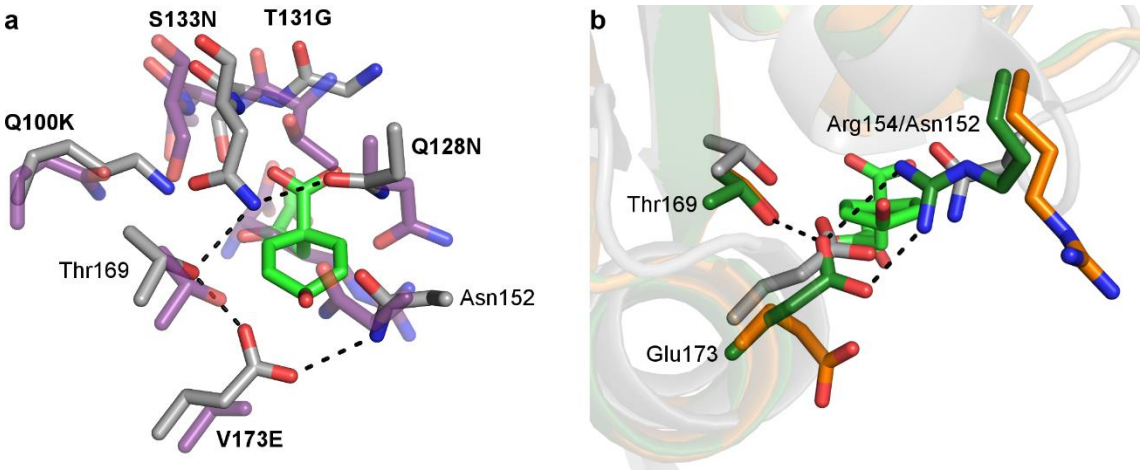


Figure 6.7. Functionalisation of the AncCDT-1 binding site for CDT activity. (**a**) Comparison between AncCDT-1 (purple) and *apo*-PaCDT (grey, with docked prephenate in green), showing substitutions in AncCDT-1 required for binding and stabilisation of the departing carboxyl group of prephenate and for recapitulation of the hydrogen bonding network around Glu173. (**b**) The configuration of the Thr169-Glu173 dyad in *apo*-PaCDT (grey) shows similarity with *apo*-Pu1068 (dark green) but not the Pu1068/NDSB-221 complex (orange).

Comparison of Pu1068 and PaCDT reveals another important commonality in their active (or binding) sites: in *apo*-Pu1068, Glu173 is located in a partially desolvated environment and interacts with Thr169 (Figure 6.7b). However, the interaction between Glu173 and Asn152 in CDT is replaced by interactions with Arg154 in Pu1068. Arg154 would be expected to decrease the pK_a of Glu173, due to stabilisation of the Glu173 anion *via* hydrogen bonding and electrostatic effects. Indeed, Glu173 is predicted by PROPKA (Olsson et al., 2011) to have a pK_a of 3.24 in *apo*-Pu1068, compared with a pK_a of 5.26 in AncCDT-3(P188L), which is an appropriate comparison because both proteins adopt an open conformation in which Glu173 is more solvent-accessible. The difference in the pK_a of Glu173 between *apo*-Pu1068 and AncCDT-3(P188L) suggests that Glu173 is less likely to have a catalytic role as a general acid in Pu1068, consistent with the hypothesis that this protein is an SBP. However, it should be noted that Glu173 and Arg154 do not interact in the Pu1068/NDSB-221 structure, which increases the predicted pK_a of Glu173 (5.29); it is not clear which, if either, geometry of Glu173 and Arg154 is functionally relevant. In any case, the presence of Glu173 in Pu1068, with a similar interaction geometry and non-polar environment in the *apo*-Pu1068 structure as in the PaCDT structure, suggests that this structural motif, apparently vital for catalytic activity in CDT, was initially an adaptation for a different function in Pu1068. Thus, in several respects, the structure of Pu1068 is intermediate between the structures of AABPs and CDTs.

Substitutions in the second shell of the active site of CDT appear to have contributed to catalytic efficiency by refining the positions of active site residues (Figure 6.8). For example, the functionally important A155I substitution (Section 4.2.2), which occurred between AncCDT-2 and AncCDT-3, could have improved the positioning of Asn128 and Asn152; Ile155 packs closely against these residues and disrupts the helical secondary structure near Asn152 (Figure 6.8b). Conservation of Phe136 (or Tyr136) in CDT homologues (Figure 4.5) suggests an important role for the substitution A136F;

Phe136 interacts with the hydrophobic portion of Lys100, orientating Lys100 towards the active site (Figure 6.8c). Finally, the substitutions F25L and F25V, which were observed to increase CDT activity in some AncCDT-2 derivatives (Section 4.2.2), occur adjacent to the active site residue Tyr22. The possible consequences of these substitutions include adjustment of the position of Tyr22 to optimise π-stacking interactions in the transition state, or modify the hydrophobic surface around Glu173 to optimise the chemical environment of this general acid (Figure 6.8d).

Finally, substitutions extending further from the active site have also been observed to affect CDT activity, with a prominent mutational hotspot at the boundary between the small domain and hinge region (Figure 6.8a). The P102L substitution, although apparently not involved in the historical evolution of CDT, was particularly important for the gain of CDT activity in AncCDT-2. Pro102 is located in a β sheet near the inner hinge strand of CDT; the P102L substitution enables an additional backbone hydrogen bond with Val187, which extends the adjacent β sheet near the outer hinge strand, as shown by the structure of AncCDT-1, which has Ile at position 102 (Figure 6.8e). Given the remoteness of this substitution from the active site of CDT and the role of the hinge region in mediating the open-closed conformational change of the enzyme, it is tempting to speculate that modification of hydrogen bonding networks in the hinge region affects catalysis by altering the relative stabilities of the open and closed conformations. L198K, which also occurs in the hinge region of the enzyme, possibly has a similar effect: in the *apo*-PaCDT structure, Lys198 bridges the two domains of the protein *via* interactions with Gly78, Ser/Thr80 and Asp170, which could contribute to the stability of the *apo*-closed state of the enzyme (Figure 6.8f).

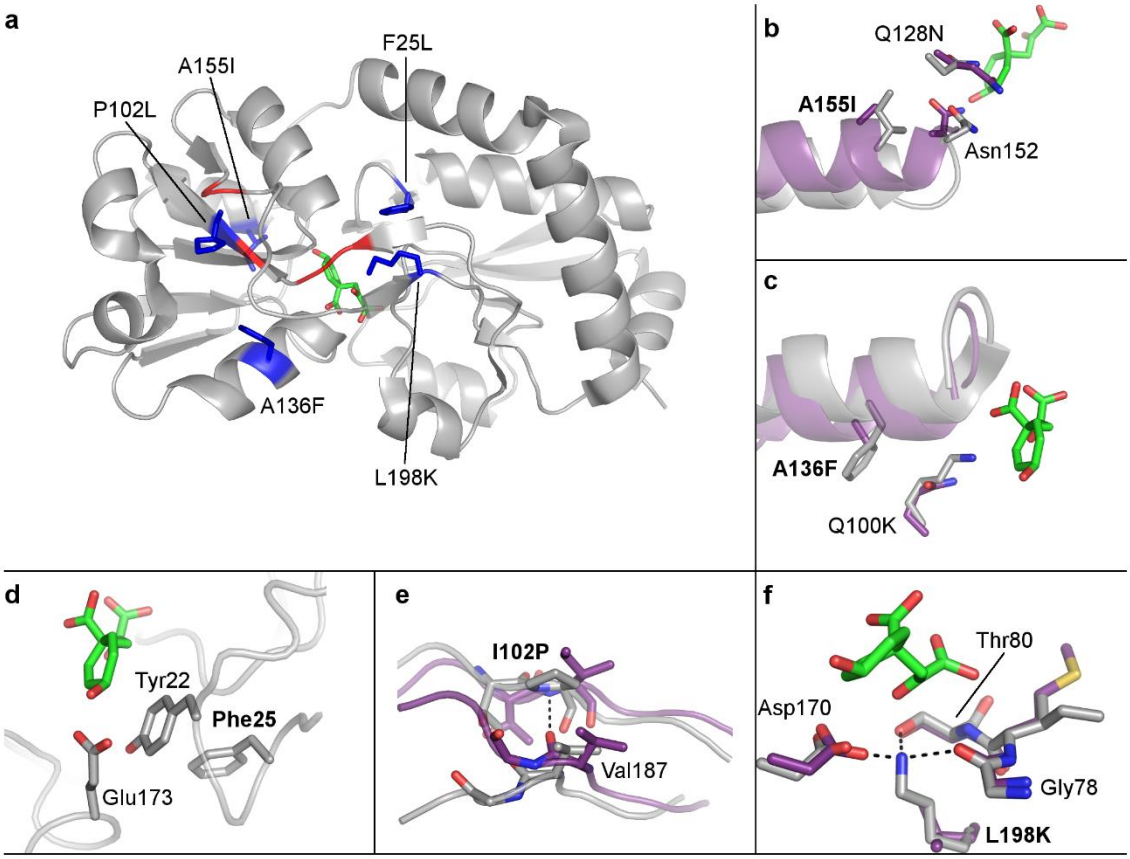


Figure 6.8. Indirect mutational effects in the evolution of CDT. (**a**) The positions of substitutions with putative indirect effects on catalysis in *apo*-PaCDT. The substitutions detailed in the other panels are shown in blue; the positions of the other substitutions that occurred in the mutational hotspot near the hinge domain are shown in red. The effects of the (**b**) A155I, (**c**) A136F, (**d**) F25L/F25V, (**e**) P102L and (**f**) L198K substitutions are shown by comparison of the AncCDT-1 (purple) and *apo*-PaCDT (grey) structures. Positions are labelled with the corresponding residue in AncCDT-1 and AncCDT-3, if conserved in both proteins, or with the corresponding substitution between AncCDT-1 and AncCDT-3.

6.3 Discussion

Altogether, based on the structures of extant and ancestral SBPs and CDTs, we propose that the following structural features and adaptations enabled the emergence of CDT activity in amino acid-binding proteins:

- The ancestral α-amino acid binding motif was retained in CDT with minor modifications: electrostatic shielding of Asp170 by Lys100 and Lys198 contributed to the dual specificity for α-amino acid and α-keto acid substrates.
- The ancestral binding site was remodelled by a conformational change in Trp60, driven by mutations in a neighbouring loop, which facilitated complementarity between the active site of CDT and its substrates.
- Insertion of Glu173 into an appropriate non-polar environment yielded a reactive general acid with an elevated pK_a, and a hydrogen bonding network extending from this residue enabled pre-organisation of the active site.
- Trp60 and Tyr22 could contribute to transition state stabilisation *via* π-stacking or cation-π interactions with the cyclohexadiene ring, while Lys100, Asn128 and Asn133 could contribute to transition state stabilisation *via* electrostatic and hydrogen bonding interactions with the departing carboxyl group.
- A number of substitutions radiating from the active site refined the positions and chemical environment of active site residues. Some remote substitutions potentially affected the conformational equilibrium of the enzyme, favouring the closed conformation over the open conformation.

Two separate issues must be considered to resolve the catalytic mechanism of CDT: the identity of the proton donor to the departing hydroxyl group must be established, and the possibilities of stepwise and concerted mechanisms for elimination of CO_2 and H_2O must be differentiated. We propose that CDT utilises Glu173 as a general acid to protonate the departing hydroxyl group of its substrates, based on the proximity of this residue to the substrates in their predicted binding modes, its conservation in CDT homologues, and its chemical environment: desolvation and repulsive electrostatic interactions increase the pK_a of Glu173 substantially, making this residue a potentially effective general acid at neutral pH. The use of Glu as a proton donor in CDT would contrast with the use of much weaker acids in other enzymes that catalyse the aromatisation of prephenate with elimination of water, including prephenate dehydratase (PDT) and carboxy-*S*-adenosyl-L-methionine (carboxy-SAM) synthase. In PDT, the putative general acid is Thr172 (pK_a ~16), as indicated by extensive mutagenesis studies showing that Thr172 is required for catalysis, whereas mutation of conserved acidic residues in the active site has no significant effect (Hsu et al., 2004; Kleeb et al., 2007; Van Vleet et al., 2010; Zhang et al., 2000). Carboxy-SAM synthase, which transfers a carboxyl group from prephenate to the *S*-methyl group of SAM, exhibits a substrate-assisted mechanism in which the hydroxyl group of prephenate abstracts a proton from the *S*-methyl group of SAM (pK_a ~19), giving a nucleophilic ylide intermediate; strong evidence for this mechanism was obtained from hydrogen-deuterium exchange experiments (Kim et al., 2013b).

The weak acidity of the proposed catalytic acids in PDT and carboxy-SAM synthase suggests an alternative driving force for the decarboxylative aromatisation of prephenate in these enzymes. It has been speculated previously that geometric distortion of the substrate, for example, forcing the keto-acid side chain towards the plane of the cyclohexadiene ring and the departing carboxylate group away from it, could promote

decarboxylation; the favourable energetics associated with aromatisation could also weaken the C–O bond to the hydroxyl group, increasing its pK_a towards that of hydroxide (Kim et al., 2013b; Van Vleet et al., 2010). Similar considerations could also apply to CDT, although deprotonation of the stronger general acid would make a much greater energetic contribution to catalysis in CDT than in PDT and carboxy-SAM synthase.

In PDT and carboxy-SAM synthase, the large discrepancy between the pK_a of the hydroxyl group of prephenate and the pK_a of the general acid implies that protonation of the hydroxyl group is energetically unfavourable and unlikely to occur prior to decarboxylation; thus, their reaction mechanisms are thought to be concerted (Kim et al., 2013b; Van Vleet et al., 2010). In the case of PDT, ^{13}C kinetic isotope effect experiments also supported the concerted mechanism (Van Vleet et al., 2010). Consideration of Hammond's postulate together with the favourable kinetics and thermodynamics of the reaction suggests that the transition state geometry for the concerted mechanism would resemble the substrate geometry, but with elongation of the C–C bond to the departing carboxylate group and the C–O bond to the hydroxyl group, and with the keto-acid side chain approaching the plane of the developing aromatic ring. Although a similar concerted mechanism is likely in CDT, two important differences with PDT also motivate consideration of the stepwise mechanism: the stronger general acid and the presence of aromatic residues (Tyr22, Trp60) that could stabilise a carbocation intermediate through cation-π interactions in the active site. In contrast, in PDT, the substrate is stabilised by a conserved Phe residue, which is less preferred in cation-π interactions, and the remaining aromatic residues in the active site have been shown to be highly mutable (Kleeb et al., 2007). Given the convergent evolution of CDT and PDT, which is another rare enzyme with the type II SBP fold (Table 3.1, p. 90), these possible mechanistic differences between the two enzymes deserve further investigation; the independent emergence of

two enzymes with the same activity in the same fold (generally associated with non-catalytic proteins), but with different mechanisms, would be quite remarkable.

A variety of experiments could be used to critically evaluate the proposed catalytic mechanism of CDT. The pH dependence of CDT activity could be measured to determine the pK_a of the general acid and test the hypothesis that Glu173, with a predicted pK_a of ~7, is the general acid. Isotope labelling and kinetic isotope effect experiments previously used to determine the mechanisms of PDT and the non-enzymatic conversion of prephenate to phenylpyruvate could also be applied to CDT; for example, incorporation of ^{18}O water into partially reacted substrate would provide strong evidence for a long-lived carbocation intermediate (Hermes et al., 1984). Site-directed mutagenesis of key residues in PaCDT, particularly Glu173, and continuation of the directed evolution experiments described in this thesis could be used to assess the essentiality of key active site residues for CDT activity. Structures of CDT variants complexed with transition state analogues and non-reactive substrate analogues could be used to validate the predicted substrate binding mode and provide further information relevant to the catalytic mechanism. Inhibitors of prephenate dehydratase have been reported and could be suitable for this purpose (Bushweller and Bartlett, 1989); these inhibitors are prephenate analogues in which the cyclohexadiene ring is saturated or partially saturated and the hydroxyl group (HC–OH) is replaced with a trigonal pyramidal sulfoxide group (S^+–O^-). Co-crystallisation of CDT variants with substrate or transition-state analogues would also be useful for stabilising the closed, catalytically competent conformation of the enzyme, enabling direct comparison between (closed) PaCDT and (open) AncCDT-3(P188L), as well as any other variants that would crystallise in the open conformation in the *apo* state.

The complexity of the active site of CDT – manifested in the reticulated hydrogen bonding networks extending from the catalytic residues, extensive opportunities for

enzyme-substrate interactions, and dependence of the active site structure on second shell residues – presents a challenge in explaining how the functional evolution of CDT could have occurred by a gradual process of mutation and selection. The structural data presented in this work, particularly for AncCDT-1 and Pu1068, shows how various pre-existing structural features were co-opted during the evolution of CDT and partly explains how CDT activity could have emerged *via* evolutionary tinkering with an ancestral amino acid binding site. Most obviously, the specificity of the ancestral amino acid binding motif in AncCDT-1 was exploited for substrate binding in CDT. Additionally, some residues that are ostensibly important for CDT activity, but not the ancestral amino acid binding activity, are observed in AncCDT-1 (Asp21, Asn152, Thr169). The presence of these residues in AncCDT-1 (whether or not they accurately represent the ancestral state) shows that they were compatible with the ancestral function and could have evolved neutrally. Consideration of the sequence reconstruction of AncCDT-2 together with the structure of Pu1068 suggests that other major structural adaptations important for CDT activity, including the conformational change of Trp60 and the insertion of Glu173 into a hydrophobic pocket, were actually adaptations for binding a different, as-yet unidentified solute (the ligand of Pu1068). This finding reinforces the importance of AncCDT-2 as a functional intermediate between AABPs and CDTs. Altogether, these results suggest that the significant change in function between AABPs and CDT depended on the co-option of various pre-existing structural features, both adaptive and neutral with respect to the ancestral functions.

Although PaCDT is a specialised and rapid enzyme ($k_{cat}/K_M \sim 10^6$ M^{-1} s^{-1}), it catalyses a mechanistically simple and energetically unchallenging transformation of an intrinsically reactive substrate. Thus, rate acceleration may have been achieved by very simple means in early intermediates in the evolution of CDT (preceding AncCDT-3). The introduction of a reactive, desolvated general acid into the binding site of an ancestral

SBP could have been sufficient for initial, promiscuous CDT activity. Fortuitous enzyme-substrate interactions, direct or water-mediated, may also have contributed to stabilisation of the departing carboxylate group of the substrate. The catalytic efficiency of this promiscuous enzyme could then have been improved by subsequent optimisation of complementarity between the enzyme and the transition state, including the introduction of hydrogen bonding networks to position the catalytic residue precisely and stabilise the departing carboxylate group. This hypothetical scenario for the emergence of CDT activity is reminiscent of recent attempts to engineer enzymes from non-catalytic proteins *via* point mutations that introduce a nucleophilic or basic residue into a hydrophobic pocket; proteins that catalyse Kemp elimination, the retro-aldol reaction and ester hydrolysis have been designed in this way (Moroz et al., 2015). Our results suggest that the evolution of catalytic activity in non-catalytic proteins could occur by a similar mechanism, by introduction of a single reactive functional group in an appropriate environment. In particular, similar processes may have been important in the evolution of secondary metabolic pathways, in which enzymes have evolved to catalyse reactions with relatively low energy barriers, and in which product flux is less tightly coupled with organismal fitness (Bar-Even and Tawfik, 2013; Bar-Even et al., 2011).

The open conformation of AncCDT-3(P188L) observed in the crystal structure suggests that the conformational dynamics of AncCDT-3(P188L) are typical of SBPs, with the enzyme adopting an open conformation in the *apo* state and (presumably) a closed conformation in the enzyme-substrate complex, in contrast to PaCDT, which adopted a closed conformation in the *apo* state. Differences in the conformational dynamics of AncCDT-3(P188L) and PaCDT may partially account for the ~6000-fold difference in catalytic efficiency (k_{cat}/K_M 1.62×10^2 M^{-1} s^{-1} and 9.68×10^5 M^{-1} s^{-1}, respectively). Given that conformational cycling in AncCDT-3(P188L) is apparently similar to conformational cycling in SBPs, optimisation of conformational dynamics for

enzymatic activity, if required for the high catalytic efficiency observed in PaCDT, must have occurred later in the evolutionary trajectory. This is a heartening conclusion from the perspective of enzyme engineering and *de novo* design: even when the collective dynamics associated with a protein fold are not optimised for catalysis, they are probably not the limiting factor in the early stages of the *de novo* evolution of catalytic activity, and it may be possible to overcome unproductive dynamics by directed evolution.

6.4 Methods

6.4.1 Crystallisation and structure determination of AncCDT-1

AncCDT-1 was expressed in *E. coli* BL21(DE3) cells grown in TB medium containing 100 mg/L ampicillin to OD_{600} 0.6 at 37 °C, induced with 0.5 mM IPTG, and incubated for a further 20 h at 37 °C. The protein was purified by nickel affinity chromatography under native conditions and size-exclusion chromatography, eluting in 3× crystallisation buffer (60 mM HEPES pH 7.5, 150 mM NaCl). The protein was diluted in water, and L-arginine was added to give a final buffer of 20 mM HEPES pH 7.5, 50 mM NaCl, 1 mM L-arginine. Crystallisation was achieved using the vapour diffusion method at 18 °C. The crystal used for structure determination was obtained by serial microseeding: irregularly shaped crystals obtained from a hanging drop containing 2 µL AncCDT-1 (18 mg/mL in 20 mM HEPES pH 7.5, 50 mM NaCl, 1 mM L-arginine) and 2 µL 0.2 M Li_2SO_4, 0.1 M Tris pH 8.2, 22% (w/v) PEG 3350 as the precipitant were crushed and serially diluted in the precipitant, and a hanging drop was prepared by mixing 2 µL of the resulting microseed suspensions with 2 µL protein. Three iterations of microseeding using the resulting crystals yielded the final crystal used for structure determination. This crystal was cryoprotected in 0.2 M Li_2SO_4, 0.1 M Tris pH 8.0, 30% (w/v) PEG 3350 and flash frozen in a nitrogen stream at 100 K. Diffraction data were collected on the MX1 beamline of the Australian Synchrotron. The data were indexed and integrated in iMOSFLM (Battye et al., 2011) and scaled in Aimless (Winn et al., 2011). The structure was solved by molecular replacement in Phaser (McCoy et al., 2007) using the AncQR-Gln structure (PDB: 4ZV2), processed using CHAINSAW (Winn et al., 2011) to remove non-conserved side-chains, as a search model. Four protein molecules were located in the asymmetric unit (V_M = 2.34 $Å^3$/Da, 47% solvent content). The AncCDT-1 model was built manually in Coot (Emsley et al., 2010) and refined with NCS restraints by simulated

annealing and restrained refinement in PHENIX (Adams et al., 2010) and REFMAC5 (Murshudov et al., 1997). In the final round of refinement in REFMAC5, translation-libration-screw parameters were refined for four groups of atoms per subunit, chosen using TLSMD (Painter and Merritt, 2006). Data collection and refinement statistics are given in Table 6.1, and coordinates and structure factors have been deposited in the PDB under accession code 5T0W.

6.4.2 Crystallisation and structure determination of AncCDT-3(P188L)

AncCDT-3(P188L) was expressed and purified in the same way as AncCDT-1 (Section 6.4.1) except that L-arginine was not added to the crystallisation buffer; the final buffer was 20 mM HEPES pH 7.5, 50 mM NaCl. Crystallisation was achieved using the vapour diffusion method at 18 °C. The crystal used for structure determination grew from a sitting drop containing 1 μL 18 mg/mL AncCDT-3(P188L) and 1 μL 1.2 M sodium citrate, 0.1 M Tris pH 8.0. The crystal was cryoprotected in 1.8 M sodium citrate, 0.05 M Tris pH 8.0 and flash frozen in a nitrogen stream at 100 K. Diffraction data were collected at 100 K on the MX1 beamline of the Australian Synchrotron. The data were indexed and integrated in iMOSFLM (Battye et al., 2011) and scaled in Aimless in the CCP4 suite (Winn et al., 2011). The structure was solved by molecular replacement in Phaser (McCoy et al., 2007), using the two domains of PaCDT (PDB: 3KBR) as separate search models. The AncCDT-3(P188L) model was built manually in Coot (Emsley et al., 2010) and refined by real space and reciprocal space refinement in Coot and REFMAC5 (Murshudov et al., 1997). In the final round of refinement in REFMAC5, translation-libration-screw parameters were refined for three groups of residues (7–109, 110–205, and 206–247), chosen using TLSMD (Painter and Merritt, 2006). Data collection and refinement statistics are given in Table 6.1, and coordinates and structure factors have been deposited in the PDB under accession code 5JOS.

Chapter Seven

Conclusions

Functional evolution of the SBP superfamily over hundreds of millions of years has contributed to the metabolic and environmental adaptability of bacteria by enabling transport of a wide variety of solutes with high affinity and specificity. The utility of SBPs has been extended further by their recruitment into other cellular processes, including chemotaxis, signal transduction, transcriptional regulation and metabolism. This work presented two case studies of functional evolution of SBPs – the evolution of binding specificity and the emergence of catalytic activity in amino acid-binding proteins – that advance our understanding of the evolutionary origin of this functional diversity.

Evolution of binding specificity in amino acid-binding proteins. Reconstruction and functional characterisation of ancestral AABPs showed that ancestral AABPs were similar in specificity compared with modern AABPs, suggesting that the evolution of new binding specificities in the AABP family occurred successively, not by subfunctionalisation of a generalist ancestor. However, the ancestral AABPs also displayed promiscuous binding activities with potential evolutionary significance. Specifically, we showed that the promiscuous binding of glutamine in an ancestral arginine-binding protein was co-opted in the evolution of specialised glutamine-binding proteins; promiscuous binding of glutamine depended on water-mediated interactions, conformational plasticity, and stabilisation of an alternative low energy conformational sub-state by the promiscuous ligand. The favourable binding enthalpy for this promiscuous interaction was offset by an unfavourable binding entropy, and the evolution of high-affinity glutamine-binding proteins occurred by reduction of this entropic penalty to binding. Altogether, structural and thermodynamic characterisation of this promiscuous binding mode provided a detailed view of the starting point for the evolution of a protein-ligand interaction with high affinity and specificity.

***Evolution of an enzyme from an amino acid-binding protein*.** Phylogenetic analysis and functional characterisation of AncCDT-1 and Ws0279 showed that CDT ultimately evolved from a cationic amino acid-binding protein, most likely with broad specificity for L-arginine, L-lysine, L-histidine and L-ornithine. We also identified a functional intermediate between AABPs and CDT, typified by the extant protein Pu1068, which did not exhibit CDT activity or affinity towards amino acids. Although the physiological function of Pu1068 has not yet been resolved, determination of the crystal structure and identification of a weak non-physiological ligand of the protein represent tangible progress towards this goal.

We solved the crystal structure of PaCDT in the *apo* state, which displayed an occluded and pre-organised active site, complementary to the substrates of the enzyme. Based on modelling of the PaCDT-substrate complex, we propose that the decarboxylative aromatisation of cyclohexadienols by CDT depends on general acid catalysis by a desolvated glutamate residue. Electrostatic interactions and hydrogen bonding networks extending from the general acid appear to contribute to the stabilisation of the departing carboxylate and hydroxyl groups in the transition state.

Comparison of the crystal structures of PaCDT, AncCDT-1 and AncCDT-3(P188L) revealed the contribution of individual substitutions to the evolution of CDT by reshaping, functionalising and refining the active site of enzyme. The observation that the structure of Pu1068 is intermediate between AncCDT-1 and AncCDT-3(P188L) shows that several structural elements important for catalysis in CDT were likely adaptations towards a different function, and contributes to an understanding of how the major functional transition between AABPs and CDT, requiring extensive structural changes, could have occurred gradually.

Site-directed mutagenesis and directed evolution experiments were used to identify the substitutions required to introduce CDT activity into the intermediate ancestral protein AncCDT-2; altogether, no more than six substitutions were necessary for sufficient CDT activity to facilitate complementation of phenylalanine auxotrophs. A range of other substitutions that increased the catalytic efficiency of AncCDT-2 and AncCDT-3 were also discovered, suggesting the existence of many evolutionary trajectories towards higher catalytic activity *via* remote substitutions. These experiments also demonstrated the importance of the historical substitutions T131G, A155I and L198K for the evolution of CDT activity.

Finally, MD simulations of PaCDT showed that the open-closed conformational dynamics associated with the SBP superfamily were conserved during the evolution of CDT. The crystal structure of AncCDT-3(P188L) provided experimental confirmation that sampling of the open conformation is possible in CDT variants. However, the crystal structure of PaCDT in the *apo*-closed state and MD simulations, showing stability of the closed conformation of the enzyme on the 100 ns timescale, suggest that stabilisation of the closed, catalytically competent conformation may have been an important adaptation for higher catalytic efficiency in CDT.

Altogether, these results provide insight into the historical sequence-structure-function relationships underlying the evolution of CDT, showing how evolutionary tinkering with the non-catalytic SBP scaffold yielded a specialised and efficient enzyme. Because the mechanisms underlying the evolution of enzymes from non-catalytic proteins have not been described previously, it is not clear whether the evolutionary trajectory of CDT is generally representative of the evolutionary trajectories of *de novo* enzymes. However, based on this work, we speculate that several conclusions about the evolution of enzymes from non-catalytic proteins are valid.

Firstly, the emergence of enzyme activity in non-catalytic proteins can occur gradually and in distinct stages, similar to the evolution of new activities in existing enzymes. The most important adaptation for the evolution of CDT activity appears to have been the incorporation of a desolvated general acid into the binding pocket of an ancestral SBP, although this was initially an adaptation towards a different function. Following introduction of the reactive general acid, optimisation of enzyme-substrate complementarity and the introduction of hydrogen-bonding networks to position the catalytic residue precisely and stabilise the departing carboxylate group of the substrate appear to have occurred. Further improvements in catalytic efficiency were likely gained by second- and third-shell substitutions that refined the structure of the active site and optimised conformational sampling to favour catalytically relevant conformations. This type of evolutionary process, in which major changes in enzyme chemistry, mediated by active site substitutions, are gradually refined by substitutions radiating from the active site, which subtly optimise active site structure and dynamics, has been documented in laboratory evolution experiments (Campbell et al., 2016; Tokuriki et al., 2012). Thus, similar evolutionary processes may be common in the evolution of enzymes from both catalytic and non-catalytic proteins.

Secondly, the evolution of enzymes from non-catalytic proteins might be constrained by chemistry. CDT has a simple one-step or two-step reaction mechanism that depends on a single catalytic residue, and the substrate of CDT is a high-energy metabolic intermediate that is predisposed to the reaction catalysed by the enzyme. Simple reaction mechanisms and high-energy intermediates are also features of other enzymes that have evolved from non-catalytic proteins (Ngaki et al., 2012), suggesting that the emergence of enzymes in non-catalytic proteins may be constrained by the feasibility of simple reaction mechanisms and availability of intrinsically reactive substrates. The simplicity of the enzyme mechanisms in these cases could be an important

factor in the availability of multiple mutational pathways to higher catalytic activity, as observed in the case of CDT.

Thirdly, our results suggest that protein dynamics might not be the most important factor restricting the emergence of enzyme activity in non-catalytic proteins. In the case of CDT, even though the conformational dynamics of the SBP fold are not optimised for enzyme activity, the large-scale open-closed conformational change intrinsic to this fold is not incompatible with enzyme activity, as shown by the substantial catalytic activity of AncCDT-3(P188L) (k_{cat} = 1.04 × 10^{-2} s^{-1}). However, our results provide evidence that adaptation of protein dynamics may have been required later in the evolutionary trajectory, between AncCDT-3(P188L) and PaCDT, for the optimisation of catalytic efficiency. More generally, our results suggest that enzyme-like conformational dynamics may be required for enzyme-like catalytic efficiency, but may not be required for the emergence of catalytic activity.

Finally, the evolutionary trajectory of CDT has striking similarities with the optimisation of rationally designed enzymes by directed evolution; catalytic activity can be initialised by computationally guided grafting of a reactive catalytic motif (e.g. a desolvated carboxylate) into a protein scaffold that can accommodate the transition state for a given reaction, and directed evolution can be used to introduce additional stabilising interactions, optimise positioning of catalytic groups, improve enzyme-transition state complementarity, and optimise conformational sampling, frequently *via* remote substitutions (Blomberg et al., 2013; Khersonsky et al., 2012). Thus, the strategies that have been used to improve catalytic activity in computational design and directed evolution experiments appear to mirror those that drove the emergence of an enzyme from a non-catalytic protein by natural selection, suggesting that through continued iteration

of these processes, engineered enzymes should be able to match the complexity and catalytic efficiency of natural enzymes.

Proposals for future work have been discussed throughout this thesis. Briefly, further progress in understanding the genetic and structural mechanisms underpinning the evolution of CDT could be achieved by: (1) characterisation of AncCDT-2 and Pu1068, which would elucidate the functional transition between AABPs and CDT more completely; (2) continuation of directed evolution experiments to characterise the fitness landscape of CDT in further detail and determine how connectivity between the fitness landscapes of CDT and SBPs enabled the evolution of CDT without non-functional intermediates; (3) experimental characterisation of the conformational landscape of CDT to determine the importance of the open/closed conformational equilibrium for rapid catalysis; and (4) confirmation of the catalytic mechanism of CDT using techniques such as site-directed mutagenesis, kinetic assays, and co-crystallisation of the enzyme with substrate or transition state analogues. More broadly, studying different examples of functional innovation in the SBP superfamily, such as the recruitment of SBPs into signalling complexes, would provide insight into the evolution of complex cellular processes and demonstrate how the utility of SBPs, which from a practical perspective represent a large family of tractable and readily available modules for ligand recognition, might be extended in protein engineering and synthetic biology applications. Further examples of functionally novel SBPs, including enzymes, also await discovery by bioinformatic analysis and functional annotation of genomic data.

References

Citation convention: *Cell*

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104–2105.

Adams, P.D., Afonine, P. V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: A comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. Sect. D Biol. Crystallogr. *66*, 213–221.

Adcock, S.A., and McCammon, J.A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. Chem. Rev. *106*, 1589–1615.

Afriat, L., Roodveldt, C., Manco, G., and Tawfik, D.S. (2006). The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. Biochemistry *45*, 13677–13686.

Afriat-Jurnou, L., Jackson, C.J., and Tawfik, D.S. (2012). Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. Biochemistry *51*, 6047–6055.

Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S.M., Roodveldt, C., and Tawfik, D.S. (2005). The "evolvability" of promiscuous protein functions. Nat. Genet. *37*, 73–76.

Alcolombri, U., Elias, M., and Tawfik, D.S. (2011). Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. J. Mol. Biol. *411*, 837–853.

Amitai, G., Gupta, R.D., and Tawfik, D.S. (2007). Latent evolutionary potentials under the neutral mutational drift of an enzyme. HFSP J. *1*, 67–78.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Babtie, A., Tokuriki, N., and Hollfelder, F. (2010). What makes an enzyme promiscuous? Curr. Opin. Chem. Biol. *14*, 200–207.

Baier, F., and Tokuriki, N. (2014). Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily. J. Mol. Biol. *426*, 2442–2456.

Baker, D. (2010). An exciting but challenging road ahead for computational enzyme design. Protein Sci. *19*, 1817–1819.

Bar-Even, A., and Tawfik, D.S. (2013). Engineering specialized metabolic pathways--is there a room for enzyme improvements? Curr. Opin. Biotechnol. *24*, 310–319.

Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D.S., and Milo, R. (2011). The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry *50*, 4402–4410.

Bar-Even, A., Milo, R., Noor, E., and Tawfik, D.S. (2015). The moderately efficient enzyme: futile encounters and enzyme floppiness. Biochemistry *54*, 4969–4977.

Barkman, T., and Zhang, J. (2009). Evidence for escape from adaptive conflict? Nature *462*, E1–E2.

Baron, R., and McCammon, J.A. (2013). Molecular recognition and ligand association. Annu. Rev. Phys. Chem. *64*, 151–175.

Barve, A., and Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature *500*, 203–206.

Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R., and Leslie, A.G.W. (2011). iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr. D Biol. Crystallogr. *67*, 271–281.

Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. *104*, 17004–17009.

Bermejo, G.A., Strub, M.-P., Ho, C., and Tjandra, N. (2010). Ligand-free open-closed transitions of periplasmic binding proteins: the case of glutamine-binding protein. Biochemistry *49*, 1893–1902.

Berntsson, R.P.-A., Smits, S.H.J., Schmitt, L., Slotboom, D.-J., and Poolman, B. (2010). A structural classification of substrate-binding proteins. FEBS Lett. *584*, 2606–2617.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D.S. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature *444*, 929–932.

Bhabha, G., Ekiert, D.C., Jennewein, M., Zmasek, C.M., Tuttle, L.M., Kroon, G., Dyson, H.J., Godzik, A., Wilson, I.A., and Wright, P.E. (2013). Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat. Struct. Mol. Biol. *20*, 1243–1249.

Björkman, J., and Mowbray, S.L. (1998). Multiple open forms of ribose-binding protein trace the path of its conformational change. J. Mol. Biol. *279*, 651–664.

Blomberg, R., Kries, H., Pinkas, D.M., Mittl, P.R.E., Grütter, M.G., Privett, H.K., Mayo, S.L., and Hilvert, D. (2013). Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature *503*, 418–421.

Bloom, J.D., and Arnold, F.H. (2009). In the light of directed evolution: pathways of adaptive protein evolution. Proc. Natl. Acad. Sci. *106*, 9995–10000.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. *103*, 5869–5874.

Bloom, J.D., Romero, P.A., Lu, Z., and Arnold, F.H. (2007). Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biol. Direct *2*, 17.

Boehr, D.D., Dyson, H.J., and Wright, P.E. (2006). An NMR perspective on enzyme dynamics. Chem. Rev. *106*, 3055–3079.

Boehr, D.D., Nussinov, R., and Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. Nat. Chem. Biol. *5*, 789–796.

Bordignon, E., Grote, M., and Schneider, E. (2010). The maltose ATP-binding cassette

transporter in the 21st century - towards a structural dynamic perspective on its mode of action. Mol. Microbiol. *77*, 1354–1366.

Bosdriesz, E., Magnúsdóttir, S., Bruggeman, F.J., Teusink, B., and Molenaar, D. (2015). Binding proteins enhance specific uptake rate by increasing the substrate-transporter encounter rate. FEBS J. *282*, 2394–2407.

Boucher, J.I., Jacobowitz, J.R., Beckett, B.C., Classen, S., and Theobald, D.L. (2014). An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. eLife *3*, e02304.

Bowers, K., Chow, E., Xu, H., Dror, R., Eastwood, M., Gregersen, B., Klepeis, J., Kolossvary, I., Moraes, M., Sacerdoti, F., et al. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. Proc. ACM/IEEE SC Conf. Supercomput. November 11–17.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535–538.

Bridgham, J.T., Carroll, S.M., and Thornton, J.W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. Science *312*, 97–101.

Bridgham, J.T., Ortlund, E.A., and Thornton, J.W. (2009). An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature *461*, 515–519.

Bridgham, J.T., Eick, G.N., Larroux, C., Deshpande, K., Harms, M.J., Gauthier, M.E.A., Ortlund, E.A., Degnan, B.M., and Thornton, J.W. (2010). Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. PLoS Biol. *8*, e1000497.

Bucher, D., Grant, B.J., Markwick, P.R., and McCammon, J.A. (2011a). Accessing a hidden conformation of the maltose binding protein using accelerated molecular dynamics. PLoS Comput. Biol. *7*, e1002034.

Bucher, D., Grant, B.J., and McCammon, J.A. (2011b). Induced fit or conformational selection? The role of the semi-closed state in the maltose binding protein. Biochemistry *50*, 10530–10539.

Bung, N., Pradhan, M., Srinivasan, H., and Bulusu, G. (2014). Structural insights into *E. coli* porphobilinogen deaminase during synthesis and exit of 1-hydroxymethylbilane. PLoS Comput. Biol. *10*, e1003484.

Bushweller, J.H., and Bartlett, P.A. (1989). Sulfoxide analogues of dihydro- and tetrahydroprephenate as inhibitors of prephenate dehydratase. J. Org. Chem. *54*, 2404–2409.

Calhoun, D.H., Pierson, D.L., and Jensen, R.A. (1973). Channel-shuttle mechanism for the regulation of phenylalanine and tyrosine synthesis at a metabolic branch point in *Pseudomonas aeruginosa*. J. Bacteriol. *113*, 241–251.

Calhoun, D.H., Bonner, C.A., Gu, W., Xie, G., and Jensen, R.A. (2001). The emerging periplasm-localized subclass of AroQ chorismate mutases, exemplified by those from *Salmonella typhimurium* and *Pseudomonas aeruginosa*. Genome Biol. *2*, 0030.1.

Campbell, E., Kaltenbach, M., Correy, G., Carr, P., Porebski, B.T., Livingstone, E., Jurnou, L., Hyvönen, M., Buckle, A.M., Weik, M., et al. (2016). The role of protein

dynamics in the evolution of new enzyme function. Nat. Chem. Biol. *12*, 944–950.

Carini, P., Steindler, L., Beszteri, S., and Giovannoni, S.J. (2013). Nutrient requirements for growth of the extreme oligotroph "Candidatus Pelagibacter ubique" HTCC1062 on a defined medium. ISME J. *7*, 592–602.

Carini, P., Campbell, E.O., Morré, J., Sañudo-Wilhelmy, S.A., Thrash, J.C., Bennett, S.E., Temperton, B., Begley, T., and Giovannoni, S.J. (2014). Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. ISME J. *8*, 1727–1738.

Chaires, J.B. (2008). Calorimetry and thermodynamics in drug design. Annu. Rev. Biophys. *37*, 135–151.

Chang, C.-E.A., McLaughlin, W.A., Baron, R., Wang, W., and McCammon, J.A. (2008). Entropic contributions and the influence of the hydrophobic environment in promiscuous protein-protein association. Proc. Natl. Acad. Sci. *105*, 7456–7461.

Chao, F.-A., Morelli, A., Haugner III, J.C., Churchfield, L., Hagmann, L.N., Shi, L., Masterson, L.R., Sarangi, R., Veglia, G., and Seelig, B. (2013). Structure and dynamics of a primordial catalytic fold generated by in vitro evolution. Nat. Chem. Biol. *9*, 81–83.

Chen, J. (2013). Molecular mechanism of the *Escherichia coli* maltose transporter. Curr. Opin. Struct. Biol. *23*, 492–498.

Chen, F., Gaucher, E.A., Leal, N.A., Hutter, D., Havemann, S.A., Govindarajan, S., Ortlund, E.A., and Benner, S.A. (2010). Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. Proc. Natl. Acad. Sci. *107*, 1948–1953.

Cho, Y., Sharma, V., and Sacchettini, J.C. (2003). Crystal structure of ATP phosphoribosyltransferase from *Mycobacterium tuberculosis*. J. Biol. Chem. *278*, 8333–8339.

Chu, B.C.H., Chan, D.I., DeWolf, T., Periole, X., and Vogel, H.J. (2014). Molecular dynamics simulations reveal that apo-HisJ can sample a closed conformation. Proteins *82*, 386–398.

Colin, P.-Y., Kintses, B., Gielen, F., Miton, C.M., Fischer, G., Mohamed, M.F., Hyvönen, M., Morgavi, D.P., Janssen, D.B., and Hollfelder, F. (2015). Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. Nat. Commun. *6*, 10008.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. *9*, 938–950.

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

Copley, S.D. (2015). An evolutionary biochemist's perspective on promiscuity. Trends Biochem. Sci. *40*, 72–78.

Darve, E., Rodríguez-Gómez, D., and Pohorille, A. (2008). Adaptive biasing force method for scalar and vector free energy calculations. J. Chem. Phys. *128*, 144120.

Davidson, A.L., Dassa, E., Orelle, C., and Chen, J. (2008). Structure, function, and

evolution of bacterial ATP-binding cassette systems. Microbiol. Mol. Biol. Rev. *72*, 317–364.

Dean, A.M., and Thornton, J.W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. Nat. Rev. Genet. *8*, 675–688.

Deka, R.K., Neil, L., Hagman, K.E., Machius, M., Tomchick, D.R., Brautigam, C.A., and Norgard, M. V (2004). Structural evidence that the 32-kilodalton lipoprotein (Tp32) of *Treponema pallidum* is an L-methionine-binding protein. J. Biol. Chem. *279*, 55644–55650.

Dellus-Gur, E., Toth-Petroczy, A., Elias, M., and Tawfik, D.S. (2013). What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. J. Mol. Biol. *425*, 2609–2621.

Desai, K.K., and Miller, B.G. (2010). Recruitment of genes and enzymes conferring resistance to the nonnatural toxin bromoacetate. Proc. Natl. Acad. Sci. U. S. A. *107*, 17968–17973.

Devamani, T., Rauwerdink, A.M., Lunzer, M., Jones, B.J., Mooney, J.L., Tan, M.A.O., Zhang, Z.-J., Xu, J.-H., Dean, A.M., and Kazlauskas, R.J. (2016). Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. J. Am. Chem. Soc. *138*, 1046–1056.

Dickinson, B.C., Leconte, A.M., Allen, B., Esvelt, K.M., and Liu, D.R. (2013). Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. Proc. Natl. Acad. Sci. *110*, 9007–9012.

Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., Osorio, S., Tohge, T., Fernie, A.R., Feussner, I., et al. (2011). Metabolic priming by a secreted fungal effector. Nature *478*, 395–398.

Dosselaere, F., and Vanderleyden, J. (2001). A metabolic node in action: chorismate-utilizing enzymes in microorganisms. Crit. Rev. Microbiol. *27*, 75–131.

Doud, M.B., Ashenberg, O., and Bloom, J.D. (2015). Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. Mol. Biol. Evol. *32*, 2944–2960.

Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H., and Shaw, D.E. (2012). Biomolecular simulation: a computational microscope for molecular biology. Annu. Rev. Biophys. *41*, 429–452.

Dupré, E., Herrou, J., Lensink, M.F., Wintjens, R., Vagin, A., Lebedev, A., Crosson, S., Villeret, V., Locht, C., Antoine, R., et al. (2015). Virulence regulation with Venus flytrap domains: structure and function of the periplasmic moiety of the sensor-kinase BvgS. PLoS Pathog. *11*, e1004700.

Durrant, J.D., and McCammon, J.A. (2011). Molecular dynamics simulations and drug discovery. BMC Biol. *9*, 71.

Dwyer, M.A., and Hellinga, H.W. (2004). Periplasmic binding proteins: a versatile superfamily for protein engineering. Curr. Opin. Struct. Biol. *14*, 495–504.

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

Eick, G.N., Colucci, J.K., Harms, M.J., Ortlund, E.A., and Thornton, J.W. (2012). Evolution of minimal specificity and promiscuity in steroid hormone receptors. PLoS Genet. *8*, e1003072.

Eick, G.N., Bridgham, J.T., Anderson, D.P., Harms, M.J., and Thornton, J.W. (2016). Robustness of reconstructed ancestral protein functions to statistical uncertainty. Mol. Biol. Evol. *34*, 247–261.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. PLoS One *3*, e3647.

Felder, C.B., Graul, R.C., Lee, A.Y., Merkle, H., and Sadee, W. (1999). The Venus flytrap of periplasmic binding proteins: an ancient protein module present in multiple drug receptors. AAPS PharmSci *1*, E2.

Fischer, R.S., Song, J., Gu, W., and Jensen, R.A. (1997). L-Arogenate is a chemoattractant which can be utilized as the sole source of carbon and nitrogen by *Pseudomonas aeruginosa*. Appl. Environ. Microbiol. *63*, 567–573.

Fisher, M.A., McKinley, K.L., Bradley, L.H., Viola, S.R., and Hecht, M.H. (2011). De novo designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. PLoS One *6*, e15364.

Fiske, M.J., Whitaker, R.J., and Jensen, R.A. (1983). Hidden overflow pathway to L-phenylalanine in *Pseudomonas aeruginosa*. J. Bacteriol. *154*, 623–631.

Flocco, M.M., and Mowbray, S.L. (1994). The 1.9 Å X-ray structure of a closed unliganded form of the periplasmic glucose/galactose receptor from *Salmonella typhimurium*. Biochemistry *269*, 8931–8936.

Freyer, M.W., and Lewis, E.A. (2008). Isothermal titration calorimetry: experimental design, data analysis, and probing macromolecule/ligand binding and kinetic interactions. Methods Cell Biol. *84*, 79–113.

Fukada, H., and Takahashi, K. (1998). Enthalpy and heat capacity changes for the proton dissociation of various buffer components in 0.1 M potassium chloride. Proteins *33*, 159–166.

Fukami-Kobayashi, K., Tateno, Y., and Nishikawa, K. (1999). Domain dislocation: a change in core structure in periplasmic binding proteins in their evolutionary history. J. Mol. Biol. *286*, 279–290.

Fulyani, F., Schuurman-Wolters, G.K., Zagar, A.V., Guskov, A., Slotboom, D.-J., and Poolman, B. (2013). Functional diversity of tandem substrate-binding domains in ABC transporters from pathogenic bacteria. Structure *21*, 1879–1888.

Gatti-Lafranconi, P., and Hollfelder, F. (2013). Flexibility and reactivity in promiscuous enzymes. Chembiochem *14*, 285–292.

Gaucher, E.A., Govindarajan, S., and Ganesh, O.K. (2008). Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature *451*, 704–707.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchinson III, C.A., and Smith,

H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 12–16.

Giger, L., Caner, S., Obexer, R., Kast, P., Baker, D., Ban, N., and Hilvert, D. (2013). Evolution of a designed retro-aldolase leads to complete active site remodeling. Nat. Chem. Biol. *9*, 494–498.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242–1245.

Giuliani, S.E., Frank, A.M., and Collart, F.R. (2008). Functional assignment of solute-binding proteins of ABC transporters using a fluorescence-based thermal shift assay. Biochemistry *47*, 13974–13984.

Goldberg, R.N., Kishore, N., and Lennen, R.M. (2002). Thermodynamic quantities for the ionization reactions of buffer. J Phys Chem Ref Data *31*, 231–370.

Gong, L.I., Suchard, M.A., and Bloom, J.D. (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. eLife *2*, e00631.

Gould, S.J., and Vrba, E.S. (1982). Exaptation - a missing term in the science of form. Paleobiology *8*, 4–15.

Gould, A.D., Telmer, P.G., and Shilton, B.H. (2009). Stimulation of the maltose transporter ATPase by unliganded maltose binding protein. Biochemistry *48*, 8051–8061.

Gouridis, G., Schuurman-Wolters, G.K., Ploetz, E., Husada, F., Vietrov, R., de Boer, M., Cordes, T., and Poolman, B. (2014). Conformational dynamics in substrate-binding domains influences transport in the ABC importer GlnPQ. Nat. Struct. Mol. Biol. *22*, 57–64.

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., and Caves, L.S.D. (2006). Bio3D: an R package for the comparative analysis of protein structures. Bioinformatics *22*, 2695–2696.

Gruenwald, K., Holland, J.T., Stromberg, V., Ahmad, A., Watcharakichkorn, D., and Okumoto, S. (2012). Visualization of glutamine transporter activities in living cells using genetically encoded glutamine sensors. PLoS One *7*, e38591.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Guyer, C.A., Morgan, D.G., and Staros, J. V (1986). Binding specificity of the periplasmic oligopeptide-binding protein from *Escherichia coli*. J. Bacteriol. *168*, 775–779.

Hamelberg, D., Mongan, J., and McCammon, J.A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J. Chem. Phys. *120*, 11919–11929.

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J.Y., Wang, L., Lupyan, D., Dahlgren, M.K., Knight, J.L., et al. (2016). OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J. Chem. Theory Comput. *12*, 281–296.

Harms, M.J., and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat. Rev. Genet. *14*, 559–571.

Harms, M.J., and Thornton, J.W. (2014). Historical contingency and its biophysical basis in glucocorticoid receptor evolution. Nature *512*, 203–207.

Härtel, T., Klein, M., Koedel, U., Rohde, M., Petruschka, L., and Hammerschmidt, S. (2011). Impact of glutamine transporters on pneumococcal fitness under infection-related conditions. Infect. Immun. *79*, 44–58.

Hayward, S., and Berendsen, H.J. (1998). Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins *30*, 144–154.

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. Nature *450*, 964–972.

Henzler-Wildman, K.A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M.A., Petsko, G.A., Karplus, M., et al. (2007). Intrinsic motions along an enzymatic reaction trajectory. Nature *450*, 838–844.

Herman, A., and Tawfik, D.S. (2007). Incorporating Synthetic Oligonucleotides via Gene Reassembly (ISOR): a versatile tool for generating targeted libraries. Protein Eng. Des. Sel. *20*, 219–226.

Hermes, J.D., Tipton, P.A., Fisher, M.A., O'Leary, M.H., Morrison, J.F., and Cleland, W.W. (1984). Mechanisms of enzymatic and acid-catalyzed decarboxylations of prephenate. Biochemistry *23*, 6263–6275.

Herrou, J., Bompard, C., Wintjens, R., Dupré, E., Willery, E., Villeret, V., Locht, C., Antoine, R., and Jacob-Dubuisson, F. (2010). Periplasmic domain of the sensor-kinase BvgS reveals a new paradigm for the Venus flytrap mechanism. Proc. Natl. Acad. Sci. *107*, 17351–17355.

Hilvert, D. (2000). Critical analysis of antibody catalysis. Annu. Rev. Biochem. *69*, 751–793.

Hilvert, D. (2013). Design of protein catalysts. Annu. Rev. Biochem. *82*, 447–470.

Holdgate, G.A. (2001). Making cool drugs hot : the use of isothermal titration calorimetry as a tool to study binding energetics. Biotechniques *31*, 164–184.

Hollfelder, F., Kirby, A.J., and Tawfik, D.S. (1996). Off-the-shelf proteins that rival tailor-made antibodies as catalysts. Nature *383*, 60–63.

Hollfelder, F., Kirby, A.J., Tawfik, D.S., Kikuchi, K., and Hilvert, D. (2000). Characterization of proton-transfer catalysis by serum albumins. J. Am. Chem. Soc. *122*, 1022–1029.

Hollup, S.M., Fuglebakk, E., Taylor, W.R., and Reuter, N. (2011). Exploring the factors determining the dynamics of different protein folds. Protein Sci. *20*, 197–209.

Hsu, S.-K., Lin, L.-L., Lo, H.-H., and Hsu, W.-H. (2004). Mutational analysis of feedback inhibition and catalytic sites of prephenate dehydratase from *Corynebacterium glutamicum*. Arch. Microbiol. *181*, 237–244.

Hu, Y., Fan, C.-P., Fu, G., Zhu, D., Jin, Q., and Wang, D.-C. (2008). Crystal structure of a glutamate/aspartate binding protein complexed with a glutamate molecule: structural basis of ligand specificity at atomic resolution. J. Mol. Biol. *382*, 99–111.

Huang, H., Pandya, C., Liu, C., Al-Obaidi, N.F., Wang, M., Zheng, L., Toews Keating, S., Aono, M., Love, J.D., Evans, B., et al. (2015). Panoramic view of a superfamily of phosphatases through substrate profiling. Proc. Natl. Acad. Sci. *112*, E1974–E1983.

Hudson, W.H., Kossmann, B.R., Mitchelle, I., Vera, S. De, Chuo, S., Weikum, E.R., Eick, G.N., Thornton, J.W., Ivanov, I.N., Kojetin, D.J., et al. (2015). Distal substitutions drive divergent DNA specificity among paralogous transcription factors through subdivision of conformational space. Proc. Natl. Acad. Sci. *113*, 326–331.

Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. *256*, 119–124.

Hvorup, R.N., Goetz, B.A., Niederer, M., Hollenstein, K., Perozo, E., and Locher, K.P. (2007). Asymmetry in the structure of the ABC transporter-binding protein complex BtuCD-BtuF. Science *317*, 1387–1390.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Jackson, C.J., Foo, J., Tokuriki, N., Afriat, L., Carr, P.D., Kim, H., Schenk, G., Tawfik, D.S., and Ollis, D.L. (2009). Conformational sampling, catalysis, and evolution of the bacterial phosphotriesterase. Proc. Natl. Acad. Sci. *106*, 21631–21636.

Jacob, F. (1977). Evolution and tinkering. Science *196*, 1161–1166.

Jacobson, M.P., Kalyanaraman, C., Zhao, S., and Tian, B. (2014). Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. Trends Biochem. Sci. *39*, 363–371.

James, L.C., and Tawfik, D.S. (2003a). The specificity of cross-reactivity : Promiscuous antibody binding involves specific hydrogen bonds rather than nonspecific hydrophobic stickiness. Protein Sci. *12*, 2183–2193.

James, L.C., and Tawfik, D.S. (2003b). Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. Trends Biochem. Sci. *28*, 361–368.

James, L.C., Roversi, P., and Tawfik, D.S. (2003). Antibody multispecificity mediated by conformational diversity. Science *299*, 1362–1367.

Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. *30*, 409–425.

Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas, C.F., et al. (2008). De novo computational design of retro-aldol enzymes. Science *319*, 1387–1391.

Jin, R., Banke, T.G., Mayer, M.L., Traynelis, S.F., and Gouaux, E. (2003). Structural basis for partial agonist action at ionotropic glutamate receptors. Nat. Neurosci. *6*, 803–810.

Jones, P.M., and George, A.M. (2014). A reciprocating twin-channel model for ABC transporters. Q. Rev. Biophys. *3*, 1–32.

Joyce, A.R., Reed, J.L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S.A., Palsson, B., and Agarwalla, S. (2006). Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. J. Bacteriol. *188*, 8259–8271.

Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125–132.

Kaltenbach, M., and Tokuriki, N. (2014). Dynamics and constraints of enzyme evolution. J. Exp. Zool. B. Mol. Dev. Evol. *322*, 468–487.

Kaltenbach, M., Jackson, C.J., Campbell, E.C., Hollfelder, F., and Tokuriki, N. (2015). Reverse evolution leads to genotypic incompatibility despite functional and active site convergence. eLife *4*, e06492.

Karplus, P.A., and Diederichs, K. (2012). Linking crystallographic model and data quality. Science *336*, 1030–1033.

Keefe, A.D., and Szostak, J.W. (2001). Functional proteins from a random-sequence library. Nature *410*, 715–718.

Keskin, O., Jernigan, R.L., and Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys. J. *78*, 2093–2106.

Khanal, A., Yu McLoughlin, S., Kershner, J.P., and Copley, S.D. (2015). Differential effects of a mutation on the normal and promiscuous activities of orthologs: implications for natural and directed evolution. Mol. Biol. Evol. *32*, 100–108.

Khare, D., Oldham, M.L., Orelle, C., Davidson, A.L., and Chen, J. (2009). Alternating access in maltose transporter mediated by rigid-body rotations. Mol. Cell *33*, 528–536.

Khare, S.D., Kipnis, Y., Greisen, P., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y., Gallaher, J.L., Silman, I., Leader, H., et al. (2012). Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat. Chem. Biol. *8*, 294–300.

Khersonsky, O., and Tawfik, D.S. (2005). Structure-reactivity studies of serum paraoxonase PON1 suggest that its native activity is lactonase. Biochemistry *44*, 6371–6382.

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

Khersonsky, O., Röthlisberger, D., Dym, O., Albeck, S., Jackson, C.J., Baker, D., and Tawfik, D.S. (2010). Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. J. Mol. Biol. *396*, 1025–1042.

Khersonsky, O., Kiss, G., Rothlisberger, D., Dym, O., Albeck, S., Houk, K.N., Baker, D., and Tawfik, D.S. (2012). Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. Proc. Natl. Acad. Sci. *109*, 10358–10363.

Kim, E., Lee, S., Jeon, A., Choi, J.M., Lee, H.-S., Hohng, S., and Kim, H.-S. (2013a). A single-molecule dissection of ligand binding to a protein with intrinsic dynamics. Nat. Chem. Biol. *9*, 313–318.

Kim, J., Xiao, H., Bonanno, J.B., Kalyanaraman, C., Brown, S., Tang, X., Al-Obaidi, N.F., Patskovsky, Y., Babbitt, P.C., Jacobson, M.P., et al. (2013b). Structure-guided

discovery of the metabolite carboxy-SAM that modulates tRNA function. Nature *498*, 123–126.

Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D., and Houk, K.N. (2013). Computational enzyme design. Angew. Chem. Int. Ed. Engl. *52*, 5700–5725.

Kleeb, A.C., Edalat, M.H., Gamper, M., Haugstetter, J., Giger, L., Neuenschwander, M., Kast, P., and Hilvert, D. (2007). Metabolic engineering of a genetic selection system with tunable stringency. Proc. Natl. Acad. Sci. U. S. A. *104*, 13907–13912.

Korendovych, I. V, and DeGrado, W.F. (2014). Catalytic efficiency of designed catalytic proteins. Curr. Opin. Struct. Biol. *27*, 113–121.

Korendovych, I. V., Kulp, D.W., Wu, Y., Cheng, H., Roder, H., and DeGrado, W.F. (2011). Design of a switchable eliminase. Proc. Natl. Acad. Sci. *108*, 6823–6827.

Koziara, K.B., Stroet, M., Malde, A.K., and Mark, A.E. (2014). Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies. J. Comput. Aided. Mol. Des. *28*, 221–233.

Kratzer, J.T., Lanaspa, M.A., Murphy, M.N., Cicerchi, C., Graves, C.L., Tipton, P.A., Ortlund, E.A., Johnson, R.J., and Gaucher, E.A. (2014). Evolutionary history and metabolic insights of ancient mammalian uricases. Proc. Natl. Acad. Sci. *111*, 3763–3768.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774–797.

Ladbury, J.E. (1996). Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. Chem. Biol. *3*, 973–980.

Lassila, J.K., Baker, D., and Herschlag, D. (2010). Origins of catalysis by computationally designed retroaldolase enzymes. Proc. Natl. Acad. Sci. *107*, 4937–4942.

Lau, A.Y., and Roux, B. (2007). The free energy landscapes governing conformational changes in a glutamate receptor ligand-binding domain. Structure *15*, 1203–1214.

Lau, A.Y., and Roux, B. (2011). The hidden energetics of ligand binding and activation in a glutamate receptor. Nat. Struct. Mol. Biol. *18*, 283–287.

Lemkul, J.A., Allen, W.J., and Bevan, D.R. (2010). Practical considerations for building GROMOS-compatible small-molecule topologies. J. Chem. Inf. Model. *50*, 2221–2235.

Leon-Kempis, M.D.R., Guccione, E., Mulholland, F., Williamson, M.P., and Kelly, D.J. (2006). The *Campylobacter jejuni* PEB1a adhesin is an aspartate/glutamate-binding protein of an ABC transporter essential for microaerobic growth on dicarboxylic amino acids. Mol. Microbiol. *60*, 1262–1275.

Lewis, V.G., Ween, M.P., and McDevitt, C.A. (2012). The role of ATP-binding cassette transporters in bacterial pathogenicity. Protoplasma *249*, 919–942.

Lodwig, E.M., Hosie, A.H.F., Bourdès, A., Findlay, K., Allaway, D., Karunakaran, R., Downie, J.A., and Poole, P.S. (2006). Amino-acid cycling drives nitrogen fixation in the legume–*Rhizobium* symbiosis. Nature *422*, 722–726.

Loeffler, H.H., and Kitao, A. (2009). Collective dynamics of periplasmic glutamine

binding protein upon domain closure. Biophys. J. *97*, 2541–2549.

Lohkamp, B., McDermott, G., Campbell, S.A., Coggins, J.R., and Lapthorn, A.J. (2004). The structure of *Escherichia coli* ATP-phosphoribosyltransferase: identification of substrate binding sites and mode of AMP inhibition. J. Mol. Biol. *336*, 131–144.

Louie, G. V, Brownlie, P.D., Lambert, R., Cooper, J.B., Blundell, T.L., Wood, S.P., Warren, M.J., Woodcock, S.C., and Jordan, P.M. (1992). Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site. Nature *356*, 133–135.

Luebbering, E.K., Mick, J., Singh, R.K., Tanner, J.J., Mehra-Chaudhary, R., and Beamer, L.J. (2012). Conservation of functionally important global motions in an enzyme superfamily across varying quaternary structures. J. Mol. Biol. *423*, 831–846.

Lunzer, M., Golding, G.B., and Dean, A.M. (2010). Pervasive cryptic epistasis in molecular evolution. PLoS Genet. *6*, e1001162.

Ma, B., and Nussinov, R. (2010). Enzyme dynamics point to stepwise conformational selection in catalysis. Curr. Opin. Chem. Biol. *14*, 652–659.

Ma, B., Shatsky, M., Wolfson, H.J., and Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci. *11*, 184–197.

Mabbitt, P.D., Correy, G.J., Meirelles, T., Fraser, N.J., Coote, M.L., and Jackson, C.J. (2016). Conformational disorganization within the active site of a recently evolved organophosphate hydrolase limits its catalytic efficiency. Biochemistry *55*, 1408–1417.

Maguid, S., Fernandez-Alberti, S., and Echave, J. (2008). Evolutionary conservation of protein vibrational dynamics. Gene *422*, 7–13.

Mak, W.S., and Siegel, J.B. (2014). Computational enzyme design: Transitioning from catalytic proteins to enzymes. Curr. Opin. Struct. Biol. *27C*, 87–94.

Des Marais, D.L., and Rausher, M.D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature *454*, 762–765.

Marsh, J.A., and Teichmann, S.A. (2014). Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. BioEssays *36*, 209–218.

Martin, S.F., and Clements, J.H. (2013). Correlating structure and energetics in protein-ligand interactions: paradigms and paradoxes. Annu. Rev. Biochem. *82*, 267–293.

Marvin, J.S., and Hellinga, H.W. (2001). Manipulation of ligand binding affinity by exploitation of conformational coupling. Nat. Struct. Mol. Biol. *8*, 795–798.

Marvin, J.S., Borghuis, B.G., Tian, L., Cichon, J., Harnett, M.T., Akerboom, J., Gordus, A., Renninger, S.L., Chen, T.-W., Bargmann, C.I., et al. (2013). An optimized fluorescent probe for visualizing glutamate neurotransmission. Nat. Methods *10*, 162–170.

Maynard Smith, J. (1970). Natural selection and the concept of a protein space. Nature *225*, 563–564.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and

Read, R.J. (2007). Phaser crystallographic software. J. Appl. Cryst. *40*, 658–674.

Merkl, R., and Sterner, R. (2016). Ancestral protein reconstruction: techniques and applications. Biol. Chem. *397*, 1–21.

Miton, C.M., and Tokuriki, N. (2016). How mutational epistasis impairs predictability in protein evolution and design. Protein Sci. *25*, 1260–1272.

Mobley, D.L., and Dill, K.A. (2009). Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". Structure *17*, 489–498.

Moroz, O. V., Moroz, Y.S., Wu, Y., Olsen, A.B., Cheng, H., Mack, K.L., McLaughlin, J.M., Raymond, E.A., Zhezherya, K., Roder, H., et al. (2013). A single mutation in a regulatory protein produces evolvable allosterically regulated catalyst of nonnatural reaction. Angew. Chemie Int. Ed. *52*, 6246–6249.

Moroz, Y.S., Dunston, T.T., Makhlynets, O. V., Moroz, O. V., Wu, Y., Yoon, J.H., Olsen, A.B., McLaughlin, J.M., Mack, K.L., Gosavi, P.M., et al. (2015). New tricks for old proteins: single mutations in a non-enzymatic protein give rise to various enzymatic activities. J. Am. Chem. Soc. *137*, 14905–14911.

Müller, A., Thomas, G.H., Horler, R., Brannigan, J.A., Blagova, E., Levdikov, V.M., Fogg, M.J., Wilson, K.S., and Wilkinson, A.J. (2005). An ATP-binding cassette-type cysteine transporter in *Campylobacter jejuni* inferred from the structure of an extracytoplasmic solute receptor protein. Mol. Microbiol. *57*, 143–155.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. *53*, 240–255.

Nagel, Z.D., and Klinman, J.P. (2009). A 21st century revisionist's view at a turning point in enzymology. Nat. Chem. Biol. *5*, 543–550.

Nam, H., Lewis, N.E., Lerman, J.A., Lee, D., Chang, R.L., Kim, D., and Palsson, B.O. (2012). Network context and selection in the evolution to enzyme specificity. Science *6185*, 1101–1104.

Näsvall, J., Sun, L., Roth, J.R., and Andersson, D.I. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. Science *338*, 384–387.

Neylon, C., Brown, S.E., Kralicek, A. V, Miles, C.S., Love, C.A., and Dixon, N.E. (2000). Interaction of the *Escherichia coli* replication terminator protein (Tus) with DNA: a model derived from DNA-binding studies of mutant proteins by surface plasmon resonance. Biochemistry *39*, 11989–11999.

Ngaki, M.N., Louie, G. V., Philippe, R.N., Manning, G., Pojer, F., Bowman, M.E., Li, L., Larsen, E., Wurtele, E.S., and Noel, J.P. (2012). Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis. Nature *485*, 530–533.

Niesen, F.H., Berglund, H., and Vedadi, M. (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat. Protoc. *2*, 2212–2221.

Nikaido, K., and Ames, G.F. (1992). Purification and characterization of the periplasmic lysine-, arginine-, ornithine-binding protein (LAO) from *Salmonella typhimurium*. J. Biol. Chem. *267*, 20706–20712.

Nobeli, I., Favia, A.D., and Thornton, J.M. (2009). Protein promiscuity and its implications for biotechnology. Nat. Biotechnol. *27*, 157–167.

Noor, S., Taylor, M.C., Russell, R.J., Jermiin, L.S., Jackson, C.J., Oakeshott, J.G., and Scott, C. (2012). Intramolecular epistasis and the evolution of a new enzymatic function. PLoS One *7*, e39822.

O'Brien, P.J., and Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. Chem. Biol. *6*, R91–R105.

O'Loughlin, T.L., Patrick, W.M., and Matsumura, I. (2006). Natural history as a predictor of protein evolvability. Protein Eng., Des. Sel. *19*, 439–442.

Oh, B.-H., Ames, G.F.-L., and Kim, S.-H. (1994). Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. J. Biol. Chem. *269*, 26323–26330.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer).

Ohtsu, I., Wiriyathanawudhiwong, N., Morigasaki, S., Nakatani, T., Kadokura, H., and Takagi, H. (2010). The L-cysteine/L-cystine shuttle system provides reducing equivalents to the periplasm in *Escherichia coli*. J. Biol. Chem. *285*, 17479–17487.

Okinaka, Y., Perna, N.T., Yang, S., Keen, N.T., and Yang, C.-H. (2006). Identification of potential virulence genes in *Erwinia chrysanthemi* 3937: transposon insertion into plant-upregulated genes. J. Gen. Plant Pathol. *72*, 360–368.

Okumoto, S., Looger, L.L., Micheva, K.D., Reimer, R.J., Smith, S.J., and Frommer, W.B. (2005). Detection of glutamate release from neurons by genetically encoded surface-displayed FRET nanosensors. Proc. Natl. Acad. Sci. *102*, 8740–8745.

Oldham, M.L., and Chen, J. (2011a). Snapshots of the maltose transporter during ATP hydrolysis. Proc. Natl. Acad. Sci. *108*, 15152–15156.

Oldham, M.L., and Chen, J. (2011b). Crystal structure of the maltose transporter in a pretranslocation intermediate state. Science *332*, 1202–1205.

Oldham, M.L., Khare, D., Quiocho, F.A., Davidson, A.L., and Chen, J. (2007). Crystal structure of a catalytic intermediate of the maltose transporter. Nature *450*, 515–521.

Olsson, M.H.M., Søndergaard, C.R., Rostkowski, M., and Jensen, J.H. (2011). PROPKA3: consistent treatment of internal and surface residues in empirical pKa calculations. J. Chem. Theory Comput. *7*, 525–537.

Olsson, T.S.G., Williams, M.A., Pitt, W.R., and Ladbury, J.E. (2008). The thermodynamics of protein-ligand interaction and solvation: insights for ligand design. J. Mol. Biol. *384*, 1002–1017.

Oostenbrink, C., Villa, A., Mark, A.E., and van Gunsteren, W.F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J. Comput. Chem. *25*, 1656–1676.

Ortlund, E.A., Bridgham, J.T., Redinbo, M.R., and Thornton, J.W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. Science *317*, 1544–1548.

Osborne, S.E., Tuinema, B.R., Mok, M.C.Y., Lau, P.S., Bui, N.K., Tomljenovic-Berube, A.M., Vollmer, W., Zhang, K., Junop, M., and Coombes, B.K. (2012). Characterization of DalS, an ATP-binding cassette transporter for D-alanine, and its role in pathogenesis in *Salmonella enterica*. J. Biol. Chem. *287*, 15242–15250.

Oswald, C., Smits, S.H.J., Höing, M., Sohn-Bösser, L., Dupont, L., Le Rudulier, D., Schmitt, L., and Bremer, E. (2008). Crystal structures of the choline/acetylcholine substrate-binding protein ChoX from *Sinorhizobium meliloti* in the liganded and unliganded-closed states. J. Biol. Chem. *283*, 32848–32859.

Padlan, E.A. (1994). Anatomy of the antibody molecule. Mol. Immunol. *31*, 169–217.

Painter, J., and Merritt, E.A. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr. Sect. D Biol. Crystallogr. *62*, 439–450.

Pang, A., Arinaminpathy, Y., Sansom, M.S.P., and Biggin, P.C. (2005). Comparative molecular dynamics - similar folds and similar motions? Proteins *61*, 809–822.

Patel, N., Pierson, D.L., and Jensen, R.A. (1977). Dual enzymatic routes to L-tyrosine and L-phenylalanine via pretyrosine in *Pseudomonas aeruginosa*. J. Biol. Chem. *252*, 5839–5846.

Patel, S.C., Bradley, L.H., Jinadasa, S.P., and Hecht, M.H. (2009). Cofactor binding and enzymatic activity in an unevolved superfamily of *de novo* designed 4-helix bundle proteins. Protein Sci. *18*, 1388–1400.

Patrick, W.M., Quandt, E.M., Swartzlander, D.B., and Matsumura, I. (2007). Multicopy suppression underpins metabolic evolvability. Mol. Biol. Evol. *24*, 2716–2722.

Pauling, L., and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular restoration studies of extinct forms of life. Acta Chem. Scand. *17*, 9–16.

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., et al. (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat. Struct. Mol. Biol. *18*, 592–596.

Preiswerk, N., Beck, T., Schulz, J.D., Milovník, P., Mayer, C., Siegel, J.B., Baker, D., and Hilvert, D. (2014). Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. Proc. Natl. Acad. Sci. U. S. A. *111*, 8013–8018.

Privett, H.K., Kiss, G., Lee, T.M., Blomberg, R., Chica, R.A., Thomas, L.M., Hilvert, D., Houk, K.N., and Mayo, S.L. (2012). Iterative approach to computational enzyme design. Proc. Natl. Acad. Sci. *109*, 3790–3795.

Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D., et al. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics *29*, 845–854.

Pulido, N.O., Silva, D.-A., Tellez, L.A., Pérez-Hernández, G., García-Hernández, E., Sosa-Peinado, A., and Fernández-Velasco, D.A. (2015). On the molecular basis of the high affinity binding of basic amino acids to LAOBP, a periplasmic binding protein from *Salmonella typhimurium*. J. Mol. Recognit. *28*, 108–116.

Radzicka, A., and Wolfenden, R. (1995). A proficient enzyme. Science *267*, 90–93.

Rajagopalan, S., Wang, C., Yu, K., Kuzin, A.P., Richter, F., Lew, S., Miklos, A.E., Matthews, M.L., Seetharaman, J., Su, M., et al. (2014). Design of activated serine-containing catalytic triads with atomic-level accuracy. Nat. Chem. Biol. *10*, 386–391.

Ramos, J.-L. (2004). *Pseudomonas*: Biosynthesis of Macromolecules and Molecular Metabolism, Volume 3 (New York: Springer US).

Raymond, E.A., Mack, K.L., Yoon, J.H., Moroz, O. V, Moroz, Y.S., and Korendovych, I. V (2014). Design of an allosterically regulated retroaldolase. Protein Sci *24*, 561–570.

Rees, D.C., Johnson, E., and Lewinson, O. (2009). ABC transporters: the power to change. Nat. Rev. Mol. Cell Biol. *10*, 218–227.

Reisch, C.R., Moran, M.A., and Whitman, W.B. (2011). Bacterial catabolism of dimethylsulfoniopropionate (DMSP). Front. Microbiol. *2*, 1–12.

Rhodes, G. (2006). Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models (San Diego: Elsevier/Academic Press).

Risso, V.A., Gavira, J.A., Mejia-Carmona, D.F., Gaucher, E.A., and Sanchez-Ruiz, J.M. (2013). Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. J. Am. Chem. Soc. *135*, 2899–2902.

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., et al. (2014). Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. Mol. Biol. Evol. *32*, 440–455.

Rizk, S.S., Paduch, M., Heithaus, J.H., Duguid, E.M., Sandstrom, A., and Kossiakoff, A.A. (2011). Allosteric control of ligand-binding affinity using engineered conformation-specific effector proteins. Nat. Struct. Mol. Biol. *18*, 437–442.

Rockah-Shmuel, L., Tawfik, D.S., and Goldsmith, M. (2014). Generating Targeted Libraries by the Combinatorial Incorporation of Synthetic Oligonucleotides During Gene Shuffling. In Directed Evolution Library Creation: Methods and Protocols, E.M.J. Gillam, J.N. Copp, and D.F. Ackerley, eds. (New York: Springer-Verlag), pp. 129–137.

Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. *10*, 866–876.

Rosa, N., Ristic, M., Seabrook, S.A., Lovell, D., Lucent, D., and Newman, J. (2015). Meltdown: a tool to help in the interpretation of thermal melt curves acquired by differential scanning fluorimetry. J. Biomol. Screen. *20*, 898–905.

Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., et al. (2008). Kemp elimination catalysts by computational enzyme design. Nature *453*, 190–195.

Ruggiero, A., Dattelbaum, J.D., Staiano, M., Berisio, R., D'Auria, S., and Vitagliano, L. (2014). A loose domain swapping organization confers a remarkable stability to the dimeric structure of the arginine binding protein from *Thermotoga maritima*. PLoS One *9*, e96560.

Rupp, B. (2010). Biomolecular Crystallography: Principles, Practice and Application to Structural Biology (New York: Garland Science).

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Salverda, M.L.M., Dellus, E., Gorter, F.A., Debets, A.J.M., van der Oost, J., Hoekstra, R.F., Tawfik, D.S., and de Visser, J.A.G.M. (2011). Initial mutations direct alternative pathways of protein evolution. PLoS Genet. *7*, e1001321.

Schulenburg, C., and Miller, B.G. (2014). Enzyme recruitment and its role in metabolic expansion. Biochemistry *53*, 836–845.

Schultz, P.G., Yin, J., and Lerner, R.A. (2002). The chemistry of the antibody molecule. Angew. Chem. Int. Ed. *41*, 4427–4437.

Seelig, B., and Szostak, J.W. (2007). Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. Nature *448*, 828–831.

Seo, M.-H., Park, J., Kim, E., Hohng, S., and Kim, H.-S. (2014). Protein conformational dynamics dictate the binding affinity for a ligand. Nat. Commun. *5*, 3724.

Shah, P., McCandlish, D.M., and Plotkin, J.B. (2015). Contingency and entrenchment in protein evolution under purifying selection. Proc. Natl. Acad. Sci. *112*, E3226–E3235.

Sharma, A.K., Becker, J.W., Ottesen, E.A., Bryant, J.A., Duhamel, S., Karl, D.M., Cordero, O.X., Repeta, D.J., and Delong, E.F. (2014). Distinct dissolved organic matter sources induce rapid transcriptional responses in coexisting populations of *Prochlorococcus*, *Pelagibacter* and the OM60 clade. Environ. Microbiol. *16*, 2815–2830.

Shi, Y., and Yokoyama, S. (2003). Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. Proc. Natl. Acad. Sci. *100*, 8308–8313.

Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St Clair, L.J., Gallaher, J.L., Hilvert, D., Gelb, M.H., Stoddard, B.L., et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science *329*, 309–313.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. *43*, D376–D381.

Silva, D.-A., Domínguez-Ramírez, L., Rojo-Domínguez, A., and Sosa-Peinado, A. (2011a). Conformational dynamics of L-lysine, L-arginine, L-ornithine binding protein reveals ligand-dependent plasticity. Proteins *79*, 2097–2108.

Silva, D.-A., Bowman, G.R., Sosa-Peinado, A., and Huang, X. (2011b). A role for both conformational selection and induced fit in ligand binding by the LAO protein. PLoS Comput. Biol. *7*, e1002054.

Simmons, C.R., Stomel, J.M., McConnell, M.D., Smith, D.A., Watkins, J.L., Allen, J.P., and Chaput, J.C. (2009). A synthetic protein selected for ligand binding affinity mediates ATP hydrolysis. ACS Chem. Biol. *4*, 649–658.

Skævern, L., Yao, X.-Q., Scarabelli, G., and Grant, B.J. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics *15*, 399.

Sleator, R.D., and Hill, C. (2002). Bacterial osmoadaptation: The role of osmolytes in bacterial stress and virulence. FEMS Microbiol. Rev. *26*, 49–71.

Sleigh, S.H., Seavers, P.R., Wilkinson, A.J., Ladbury, J.E., and Tame, J.R. (1999). Crystallographic and calorimetric analysis of peptide binding to OppA protein. J. Mol. Biol. *291*, 393–415.

Smith, B.A., and Hecht, M.H. (2011). Novel proteins: from fold to function. Curr. Opin. Chem. Biol. *15*, 421–426.

Soo, V.W.C., Hanson-Manful, P., and Patrick, W.M. (2011). Artificial gene amplification reveals an abundance of promiscuous resistance determinants in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. *108*, 1484–1489.

Soskine, M., and Tawfik, D.S. (2010). Mutational effects and the evolution of new protein functions. Nat. Rev. Genet. *11*, 572–582.

Sowell, S.M., Wilhelm, L.J., Norbeck, A.D., Lipton, M.S., Nicora, C.D., Barofsky, D.F., Carlson, C.A., Smith, R.D., and Giovanonni, S.J. (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J. *3*, 93–105.

Stamp, A.L., Owen, P., El Omari, K., Lockyer, M., Lamb, H.K., Charles, I.G., Hawkins, A.R., and Stammers, D.K. (2011). Crystallographic and microcalorimetric analyses reveal the structural basis for high arginine specificity in the *Salmonella enterica* serovar Typhimurium periplasmic binding protein STM4351. Proteins *79*, 2352–2357.

Starr, T.N., and Thornton, J.W. (2016). Epistasis in protein evolution. Protein Sci. *25*, 1204–1218.

Sun, J., Steindler, L., Thrash, J.C., Halsey, K.H., Smith, D.P., Carter, A.E., Landry, Z.C., and Giovannoni, S.J. (2011). One carbon metabolism in SAR11 pelagic marine bacteria. PLoS One *6*, e23973.

Sykora, J., Brezovsky, J., Koudelakova, T., Lahoda, M., Fortova, A., Chernovets, T., Chaloupkova, R., Stepankova, V., Prokop, Z., Smatanova, I.K., et al. (2014). Dynamics and hydration explain failed functional transformation in dehalogenase design. Nat. Chem. Biol. *10*, 428–430.

Tam, R., and Saier, M.H. (1993a). Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. Microbiol. Rev. *57*, 320–346.

Tam, R., and Saier, M.H. (1993b). A bacterial periplasmic receptor homologue with catalytic activity: cyclohexadienyl dehydratase of *Pseudomonas aeruginosa* is homologous to receptors specific for polar amino acids. Res. Microbiol. *144*, 165–169.

Tame, J.R., Sleigh, S.H., Wilkinson, A.J., and Ladbury, J.E. (1996). The role of water in sequence-independent ligand binding by an oligopeptide transporter protein. Nat. Struct. Mol. Biol. *3*, 998–1001.

Tang, C., Schwieters, C.D., and Clore, G.M. (2007). Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. Nature *449*, 1078–1082.

Tawfik, D.S. (2010). Messy biology and the origins of evolutionary innovations. Nat. Chem. Biol. *6*, 692–696.

Tcherkez, G.G.B., Farquhar, G.D., and Andrews, T.J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. Proc. Natl. Acad. Sci. *103*, 7246–7251.

Tellinghuisen, J. (2008). Isothermal titration calorimetry at very low c. Anal. Biochem. *373*, 395–397.

Tellinghuisen, J. (2012). Designing isothermal titration calorimetry experiments for the study of 1:1 binding: problems with the "standard protocol". Anal. Biochem. *424*, 211–220.

Telmer, P.G., and Shilton, B.H. (2003). Insights into the conformational equilibria of maltose-binding protein by analysis of high affinity mutants. J. Biol. Chem. *278*, 34555–34567.

Thornton, J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. Nat. Rev. Genet. *5*, 366–375.

Thorpe, I.F., and Brooks, C.L. (2007). Molecular evolution of affinity and flexibility in the immune system. Proc. Natl. Acad. Sci. *104*, 8821–8826.

Todd, A.E., Orengo, C.A., and Thornton, J.M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. Structure *10*, 1435–1451.

Tokuriki, N., and Tawfik, D.S. (2009a). Protein dynamism and evolvability. Science *324*, 203–207.

Tokuriki, N., and Tawfik, D.S. (2009b). Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature *459*, 668–673.

Tokuriki, N., Jackson, C.J., Afriat-Jurnou, L., Wyganowski, K.T., Tang, R., and Tawfik, D.S. (2012). Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. Nat. Comm. *3*, 1257.

Tóth-Petróczy, Á., and Tawfik, D.S. (2014). The robustness and innovability of protein folds. Curr. Opin. Struct. Biol. *26*, 131–138.

Tracewell, C.A., and Arnold, F.H. (2009). Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr. Opin. Chem. Biol. *13*, 3–9.

Tripp, H.J. (2013). The unique metabolism of SAR11 aquatic bacteria. J. Microbiol. *51*, 147–153.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J. Comput. Chem. *31*, 455–461.

Turnbull, W.B., and Daranas, A.H. (2003). On the value of c: can low affinity systems be studied by isothermal titration calorimetry? J. Am. Chem. Soc. *125*, 14859–14866.

Urvoas, A., Valerio-Lepiniec, M., and Minard, P. (2012). Artificial proteins from combinatorial approaches. Trends Biotechnol. *30*, 512–520.

Vagin, A., and Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. J. Appl. Cryst. *30*, 1022–1025.

Velazquez-Campoy, A., and Freire, E. (2006). Isothermal titration calorimetry to

determine association constants for high-affinity ligands. Nat. Methods *1*, 186–191.

Velazquez-Campoy, A., Ohtaka, H., Nezami, A., Muzammil, S., and Freire, E. (2004). Isothermal Titration Calorimetry. In Current Protocols in Cell Biology, (John Wiley & Sons, Inc.), pp. 17.8.1–17.8.24.

Vetting, M.W., Al-Obaidi, N., Zhao, S., San Francisco, B., Kim, J., Wichelecki, D.J., Bouvier, J.T., Solbiati, J.O., Vu, H., Zhang, X., et al. (2015). Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. Biochemistry *54*, 909–931.

Villiers, B.R.M., and Hollfelder, F. (2009). Mapping the limits of substrate specificity of the adenylation domain of TycA. ChemBioChem *10*, 671–682.

de Visser, J.A.G.M., and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. Nat. Rev. Genet. *15*, 480–490.

Van Vleet, J., Kleeb, A., Kast, P., Hilvert, D., and Cleland, W.W. (2010). 13C isotope effect on the reaction catalyzed by prephenate dehydratase. Biochim. Biophys. Acta *1804*, 752–754.

Vogt, A.D., and Di Cera, E. (2012). Conformational selection or induced fit? A criticial appraisal of the kinetic mechanism. Biochemistry *51*, 5894–5902.

Vogt, A.D., and Di Cera, E. (2013). Conformational selection is a dominant mechanism of ligand binding. Biochemistry *52*, 5723–5729.

Voordeckers, K., Brown, C.A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., and Verstrepen, K.J. (2012). Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol. *10*, e1001446.

Walshaw, D.L., and Poole, P.S. (1996). The general L-amino acid permease of *Rhizobium leguminosarum* is an ABC uptake system that also influences efflux of solutes. Mol. Microbiol. *21*, 1239–1252.

Wang, X., Minasov, G., and Shoichet, B.K. (2002). Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. J. Mol. Biol. *320*, 85–95.

Watanabe, K., Ohkuri, T., Yokobori, S., and Yamagishi, A. (2006). Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree. J. Mol. Biol. *355*, 664–674.

Weiner, J.H., and Heppel, L.A. (1971). A binding protein for glutamine and its relation to active transport in *Escherichia coli*. J. Biol. Chem. *246*, 6933–6941.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science *312*, 111–114.

Wheeler, L.C., Lim, S.A., Marqusee, S., and Harms, M.J. (2016). The thermostability and specificity of ancient proteins. Curr. Opin. Struct. Biol. *38*, 37–43.

Whitfield, J.H., Zhang, W., Herde, M.K., Clifton, B.E., Radziejewski, J., Janovjak, H., Henneberger, C., and Jackson, C.J. (2015). Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. Protein Sci. *24*, 1412–1422.

Wildermuth, M.C., Dewdney, J., Wu, G., and Ausubel, F.M. (2001). Isochorismate synthase is required to synthesize salicylic acid for plant defence. Nature *414*, 562–571.

Wiley, E.O., and Lieberman, B.S. (2011). Phylogenetics: Theory and Practice of Phylogenetic Systematics (New Jersey: Wiley-Blackwell).

Willis, R.C., and Furlong, C.E. (1976). Purification and properties of a periplasmic glutamate-aspartate binding protein from *Escherichia coli* K12 Strain W3092*. J. Biol. Chem. *250*, 2574–2580.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. *67*, 235–242.

Wolf, A., Shaw, E.W., Nikaido, K., and Ames, G.F.-L. (1994). The histidine-binding protein undergoes conformational changes in the absence of ligand as analyzed with conformation-specific monoclonal antibodies. J. Biol. Chem. *269*, 23051–23058.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, Z. (2014). Molecular Evolution: A Statistical Approach (Oxford: Oxford University Press).

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314.

Yao, Y., Belcher, J., Berger, A.J., Mayer, M.L., and Lau, A.Y. (2013). Conformational analysis of NMDA receptor GluN1, GluN2, and GluN3 ligand-binding domains reveals subtype-specific characteristics. Structure *21*, 1788–1799.

Zanghellini, A., Jiang, L.I.N., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., and Ro, D. (2006). New algorithms and an in silico benchmark for computational enzyme design. Protein Sci. *15*, 2785–2794.

Zhang, P. (2013). Structure and mechanism of energy-coupling factor transporters. Trends Microbiol. *21*, 652–659.

Zhang, Y.L., and Zhang, Z.Y. (1998). Low-affinity binding determined by titration calorimetry using a high-affinity coupling ligand: a thermodynamic study of ligand binding to protein tyrosine phosphatase 1B. Anal. Biochem. *261*, 139–148.

Zhang, S., Wilson, D.B., and Ganem, B. (2000). Probing the catalytic mechanism of prephenate dehydratase by site-directed mutagenesis of the *Escherichia coli* P-protein dehydratase domain. Biochemistry *39*, 4722–4728.

Zhao, H., and Zha, W. (2006). In vitro “sexual” evolution through the PCR-based staggered extension process (StEP). Nat. Protoc. *1*, 1865–1871.

Zhao, G., Xia, T., Fischer, R.S., and Jensen, R.A. (1992). Cyclohexadienyl dehydratase from *Pseudomonas aeruginosa*: molecular cloning of the gene and characterization of the gene product. J. Biol. Chem. *267*, 2487–2493.

Zhao, G., Xia, T., Aldrich, H., and Jensen, R.A. (1993). Cyclohexadienyl dehydratase from *Pseudomonas aeruginosa* is a periplasmic protein. J. Gen. Microbiol. *139*, 807–813.

Zhao, H., Giver, L., Shao, Z., Affholter, J.A., and Arnold, F.H. (1998). Molecular evolution by staggered extension (StEP) in vitro recombination. Nat. Biotechnol. *16*, 258.

Zhao, S., Kumar, R., Sakai, A., Vetting, M.W., Wood, B.M., Brown, S., Bonanno, J.B., Hillerich, B.S., Seidel, R.D., Babbitt, P.C., et al. (2013). Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature *502*, 698–702.

Zou, T., Risso, V.A., Gavira, J.A., Sanchez-Ruiz, J.M., and Ozkan, S.B. (2014). Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. Mol. Biol. Evol. *32*, 132–143.

Appendix I

Composition of the Hampton Solubility & Stability Screen subset used for DSF analysis of Pu1068. Each compound was tested at 1×, 0.5× and 0.25× the maximum concentration.

Compound	Maximum concentration
DL-malic acid, pH 7.0	250 mM
Ethylenediaminetetracetic acid	50 mM
NDSB-221	250 mM
Ethylammonium nitrate	12.5% (v/v)
Sodium acetate	125 mM
Sodium sulfate	300 mM
Trehalose	250 mM
Glycerol	12.5% (v/v)
Trimethylamine-*N*-oxide	250 mM
Urea	125 mM
Glycine	125 mM
Sodium thiocyanate	500 mM
Potassium chloride	500 mM
Sorbitol	500 mM
Glucose	250 mM
L-Arginine	75 mM
Sodium succinate, pH 7.0	375 mM
Ethylene glycol	5% (v/v)
Sodium nitrate	500 mM
L-Histidine	50 mM
Sodium malonate, pH 7.0	375 mM
Magnesium chloride	500 mM
L-Arginine + L-glutamic acid	75 mM each
Sodium chloride	500 mM
NDSB-256	250 mM
Lithium chloride	500 mM
Triethylene glycol	5% (v/v)
Potassium sodium phosphate	500 mM