Catalytic Multitasking in MetC: One Enzyme, Multiple Reactions

A dissertation presented in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Biochemistry
at the University of Otago, Dunedin, New Zealand.

Matteo Paolo Ferla

2014
Abstract

Unlike most modern enzymes, which perform a single reaction, primordial enzymes are believed to have each performed a broad range of reactions. Consequently, most modern enzymes are poor comparative models for inferring the properties of primordial enzymes. The aim of this thesis was to identify and characterise modern enzymes that have evolved to catalyse multiple reactions on multiple substrates, in order to better understand the properties of primordial enzymes.

Using bioinformatics and phylogenetics, I have discovered that, in three distinct clades, one enzyme (MetC) has taken over the role of the absent alanine racemase (Alr). Two of the three MetC enzymes have also taken over the role of the absent glutamate racemase (MurI). Even though MetC, MurI and Alr are not homologous. In other organisms, such as Escherichia coli, MetC catalyses the β-elimination of cystathionine in methionine biosynthesis. The E. coli MetC has a small promiscuous alanine racemisation activity, but no detectable glutamate racemisation activity.

Two of the clades—the genus Pelagibacter and the family Anaplasmataceae, which includes Wolbachia—were in the same class, the Alphaproteobacteria, but the precise location was under debate in the literature. Therefore, I used phylogenetic methods to determine that they did belong to sister orders and that the AT-richness of their genomes is ancestral, rather than a source of bias.

The three enzymes investigated are located in different groups along the MetC tree. Using in vitro activity assays, I found that the enzyme most similar to E. coli MetC, Pelagibacter ubique MetC, could racemise alanine, but could not racemise glutamate. Further away on the tree, Wolbachia MetC has glutamate and alanine racemisation activities that are both stronger than its cystathionine β-elimination activity in terms of turnover numbers (kcat). While, in a basal group, Thermotoga maritima MetC possess the strongest glutamate racemising activity of the enzymes studied. These three enzymes were also able to promiscuously cleave cysteine.

The three enzymes had Michaelis constants (Km) for the various substrates that are comparable to those of each dedicated enzyme (MetC, Alr and MurI) in other
organisms, while having turnover numbers ($k_{cat}$) that are much lower than those of each dedicated enzyme. It was conjectured that this balance of kinetic parameters in these enzymes is due to the physiological necessity of operating with low concentrations of each substrate, yet producing a limited amount of product due to the low cellular demand for each metabolite.

The crystal structure of $T. maritima$ MetC was solved, showing that it may possess glutamate racemisation activity because an active site tryptophan forms a hydrogen bond with the terminal carboxyl group of the bound glutamate. The structure also revealed a latch-like loop close to the active site entrance.

I used directed evolution in an attempt to improve the cystathionine elimination activity of $T. maritima$ MetC. A mutant (S86T/S305C) was identified that possessed a more permissive active site with decreased Michaelis constants for all activities, including those that had not been under selection. This may indicate that wild-type $T. maritima$ MetC has a tight balance of kinetic parameters that needs to be relaxed before it can be evolved into an enzyme with a single activity.

Overall, the properties of these enzymes reveal how multiple activities are balanced in a single active site and offer new insights into the likely nature of primordial enzymes.
ACKNOWLEDGEMENT INFOGRAPHICS

Overview of acknowledgements

Breakdown of research acknowledgements

Wayne Patrick (Supervisor)
Guidance and advice
Patience shown
Constructive disapproval shown
Paying for lab events
Much more

Kurt Krause and Craig Marshall
(Thesis committee, University of Otago)
Austen Ganley
(Second supervisor, Massey University)
# Table of Contents

1 Primordiality, Promiscuity and Multitasking ........................................... 1

1.1 In the beginning ....................................................................................... 2
1.2 Many substrates, many mechanisms ...................................................... 3
1.3 Promiscuity in enzymes ......................................................................... 4
1.4 Origin of promiscuous activities ............................................................ 6
1.5 Gene duplications .................................................................................... 7
1.6 Multitasking enzymes ........................................................................... 8
1.7 Evolvability and robustness ..................................................................... 11
1.8 Other factors affecting enzymes ............................................................... 13
1.9 Aims of this thesis .................................................................................. 14

2 The Genetics of metC in Wolbachia, P. ubique and T. maritima ............. 17

2.1 Introduction .............................................................................................. 18

2.1.1 Reciprocal promiscuity of alr and metC ........................................... 18
2.1.2 MetC and methionine biosynthesis ................................................... 18
2.1.3 Alr and peptidoglycan ..................................................................... 20
2.1.4 Non-homologous gene displacement of alr ........................................ 21

2.2 Results ...................................................................................................... 23

2.2.1 Search for species without alr (in silico) ........................................... 23
2.2.2 Cloning of candidate metC genes .................................................... 26
2.2.3 Complementation (in vivo) ............................................................... 27
2.2.4 Investigation of alr alternatives in T. maritima (in silico) ................. 28
2.2.5 Investigation of alr alternatives in P. ubique (in silico) ....................... 32
2.2.6 Investigation of alr alternatives in Wolbachia (in silico) .................... 33
2.2.7 Operon order ...................................................................................... 33
2.2.8 MetC as a Cystathionine γ-synthase ........................................... 34
2.2.9 Evolutionary history of metC (in silico) ........................................ 34
2.2.10 Evolutionary history of the Alphaproteobacteria in terms of metC, murI and alr (in silico) ............................................................ 36
2.2.11 Filling pathway holes in methionine biosynthesis .......................... 37

2.3 Discussion ......................................................................................... 39
  2.3.1 Gene displacement ........................................................................ 39
  2.3.2 MetC as a cystathionine β-lyase .................................................. 39
  2.3.3 Peptidoglycan and D-alanine requirement .................................... 42
  2.3.4 MetC as sole alanine racemase .................................................... 43
  2.3.5 Evolution and streamlining .......................................................... 44
  2.3.6 Conclusions ................................................................................. 46

2.4 Methods ............................................................................................ 47
  2.4.1 Identification of genomes of interest ........................................... 47
  2.4.2 Verification and gene searches ..................................................... 48
  2.4.3 Cloning ........................................................................................ 49
  2.4.4 Complementation tests ................................................................. 50
  2.4.5 Phylogeny of MetC and Alr .......................................................... 51
  2.4.6 Covariance of metY and metC across genomes ............................ 53

3 The Phylogeny of the Alphaproteobacteria .......................................... 55

3.0 Preface .............................................................................................. 56
  3.0.1 Reasoning .................................................................................... 56
  3.0.2 Author contribution ..................................................................... 56
  3.0.3 Copyright notice .......................................................................... 57

3.1 Introduction ......................................................................................... 58
  3.1.1 Merits of rRNA-based phylogeny ............................................... 58
3.1.2 The diverse members of the *Alphaproteobacteria* ........................................... 58
3.1.3 Debate over the phylogeny of the *Alphaproteobacteria* .............................. 59
3.1.4 Chapter aims .................................................................................................. 60

3.2 Results .............................................................................................................. 62
3.2.1 GC content of rRNA genes is comparatively unbiased................................. 62
3.2.2 Tree building strategy..................................................................................... 63
3.2.3 The 16S and 23S rRNA gene trees ............................................................... 64
3.2.4 The complete concatenated trees ................................................................. 70
3.2.5 The complete RY-, MK- and RYMK-recoded trees....................................... 75
3.2.6 The trimmed concatenated trees ................................................................. 77
3.2.7 Trees with single clades removed .................................................................. 83
3.2.8 Mitochondrial placement .............................................................................. 83
3.2.9 16S-23S dataset with several 16S-only sequences ......................................... 86
3.2.10 Approximately Unbiased tests ................................................................. 86

3.3 Discussion ......................................................................................................... 90
3.3.1 rRNA focus ................................................................................................. 90
3.3.2 GC bias ........................................................................................................ 90
3.3.3 Effects of alignment method, rate model and taxon sampling...................... 91
3.3.4 Taxonomic observations ............................................................................. 92
3.3.5 Proposal of new subclasses ......................................................................... 94

3.4 Materials and Methods .................................................................................... 96
3.4.1 Taxon selection............................................................................................ 96
3.4.2 GC content .................................................................................................. 96
3.4.3 Sequence QC ............................................................................................... 96
3.4.4 Alignment/gap-removal ............................................................................. 97
3.4.5 RY-, MK- and RYMK-recoded datasets ...................................................... 97
3.4.6 Maximum likelihood trees ................................................................. 97
3.4.7 Topology evaluation ........................................................................ 97
3.4.8 Approximately Unbiased tests ......................................................... 98

4 The biochemistry of MetC from Wolbachia, P. ubique and T. maritima ..... 99

4.1 Introduction ............................................................................................ 100
  4.1.1 From genes to enzymes .................................................................. 100
  4.1.2 PLP ................................................................................................. 102
  4.1.3 Reactions ........................................................................................ 103
  4.1.4 Promiscuity vs. multifunctionality .................................................. 104

4.2 Results ..................................................................................................... 104
  4.2.1 Assay overview ................................................................................ 104
  4.2.2 E. coli MetC kinetics ........................................................................ 105
  4.2.3 Wolbachia MetC kinetics ................................................................ 107
  4.2.4 P. ubique MetC kinetics .................................................................. 108
  4.2.5 T. maritima MetC kinetics ............................................................... 110
  4.2.6 Background control ........................................................................ 112
  4.2.7 Analysis of cystathionine elimination product by mass spectrometry .... 113
  4.2.8 T. maritima MetC structure .............................................................. 116
  4.2.9 Structural comparison (T. maritima MetC – E. coli MetC) ............... 118

4.3 Discussion ............................................................................................... 126
  4.3.1 Summary of results .......................................................................... 126
  4.3.2 Glutamate racemisation ................................................................. 126
  4.3.3 Alanine racemisation ...................................................................... 128
  4.3.4 Cysteine β-elimination .................................................................. 129
  4.3.5 Cystathionine elimination specificity ............................................ 129
  4.3.6 Temperature dependence of enzyme kinetics ................................ 130
4.3.7 Balance of activities in *P. ubique* MetC ........................................ 130
4.3.8 Balance of activities in *Wolbachia* MetC ....................................... 131
4.3.9 Balance of activities in *T. maritima* MetC ..................................... 132
4.3.10 Physiological Michaelis constants .................................................. 132
4.3.11 Interplay of activities ........................................................................ 134

4.4 Methods .................................................................................................. 135
4.4.1 Expression of the MetC enzymes ......................................................... 135
4.4.2 Purification of the MetC enzymes ........................................................ 135
4.4.3 Spectrophotometric assays .................................................................. 136
4.4.4 MS of cystathionine elimination product ............................................ 137
4.4.5 Thermal shift assay .............................................................................. 137
4.4.6 Determination of *T. maritima* MetC structure ................................... 137

5 The Directed Evolution of *T. maritima* MetC ........................................ 139

5.1 Introduction ............................................................................................. 140
5.1.1 Evolvability of *T. maritima* MetC ...................................................... 140
5.1.2 Broad substrate ambiguity and primordiality ....................................... 140
5.1.4 Methodologies and requirements for error-prone PCR ....................... 142
5.1.5 Effect of directed evolution on *T. maritima* MetC .............................. 143

5.2 Results ..................................................................................................... 144
5.2.1 Library construction and analysis of diversity ...................................... 144
5.2.2 Attempts at racemisation selection ...................................................... 146
5.2.3 Cystathionine elimination selection .................................................... 147
5.2.4 Kinetics of the S86T/S305C mutant ..................................................... 148
5.2.5 Structural predictions for the S86T/S305C mutant ............................ 151
5.2.6 Structural predictions based on the kinetics of the mutant .................. 152

5.3 Discussion ............................................................................................... 154
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 1.1</td>
<td>The rise of early metabolism</td>
<td>3</td>
</tr>
<tr>
<td>Fig. 1.2</td>
<td>Cartoon of a fitness landscape</td>
<td>12</td>
</tr>
<tr>
<td>Fig. 2.1</td>
<td>Structures of the monomers of <em>E. coli</em> MetC and Alr</td>
<td>19</td>
</tr>
<tr>
<td>Fig. 2.2</td>
<td>Methionine and cysteine metabolism</td>
<td>20</td>
</tr>
<tr>
<td>Fig. 2.3</td>
<td>D-amino acid in proteobacterial peptidoglycan</td>
<td>21</td>
</tr>
<tr>
<td>Fig. 2.4</td>
<td>Rarity of taxa lacking <em>alr</em></td>
<td>24</td>
</tr>
<tr>
<td>Fig. 2.5</td>
<td>pBAD/metC&lt;sub&gt;wMel&lt;/sub&gt;(Kpn&lt;sup&gt;+&lt;/sup&gt;)</td>
<td>27</td>
</tr>
<tr>
<td>Fig. 2.6</td>
<td>Suppression of methionine and D-alanine auxotrophy</td>
<td>28</td>
</tr>
<tr>
<td>Fig. 2.7</td>
<td>Inferred history of <em>metC</em>, <em>alr</em> and <em>murI</em> in the phylum <em>Thermotogae</em></td>
<td>30</td>
</tr>
<tr>
<td>Fig. 2.8</td>
<td>Inability of the <em>metC</em> orthologues to rescue an <em>E. coli</em> Δ<em>metB</em> strain</td>
<td>34</td>
</tr>
<tr>
<td>Fig. 2.9</td>
<td>Condensed phylogenetic relation between between <em>metC</em> gene</td>
<td>35</td>
</tr>
<tr>
<td>Fig. 2.10</td>
<td>Inferred history of <em>metC</em>, <em>alr</em> and <em>murI</em></td>
<td>36</td>
</tr>
<tr>
<td>Fig. 2.11</td>
<td>Copresence of <em>metB</em>, <em>metC</em>, <em>metZ</em> and <em>metY</em></td>
<td>38</td>
</tr>
<tr>
<td>Fig. 2.12</td>
<td>Proposed routes to homocysteine biosynthesis in <em>T. maritima</em></td>
<td>41</td>
</tr>
<tr>
<td>Fig. 3.1</td>
<td>Relationship of rRNA gene <em>vs.</em> genomic GC content</td>
<td>63</td>
</tr>
<tr>
<td>Fig. 3.2</td>
<td>Summary of bootstrap supports of 16S and 23S only trees</td>
<td>66–67</td>
</tr>
<tr>
<td>Fig. 3.3</td>
<td>Tree inferred with ARB-SINA aligned complete dataset under GTRΓ</td>
<td>68–69</td>
</tr>
<tr>
<td>Fig. 3.4</td>
<td>Summary of bootstrap supports of complete dataset</td>
<td>71–74</td>
</tr>
<tr>
<td>Fig. 3.5</td>
<td>Box plot of the distributions of GC contents</td>
<td>76</td>
</tr>
<tr>
<td>Fig. 3.6</td>
<td>Summary of bootstrap supports of trimmed datasets</td>
<td>78–82</td>
</tr>
<tr>
<td>Fig. 3.7</td>
<td>Summary of bootstrap supports of mtDel dataset</td>
<td>85</td>
</tr>
<tr>
<td>Fig. 3.8</td>
<td>Tree inferred with ARB-SINA aligned combo dataset under GTRΓ</td>
<td>88–89</td>
</tr>
<tr>
<td>Fig. 3.9</td>
<td>Proposed subclasses of the <em>Alphaproteobacteria</em></td>
<td>95</td>
</tr>
<tr>
<td>Fig. 4.1</td>
<td>PLP-dependant racemisation and β-elimination</td>
<td>101</td>
</tr>
</tbody>
</table>
Fig. 4.2 Representative aliquots of the various enzymes used in the study..............104
Fig. 4.3 Michaelis–Menten plots of D→L alanine racemisation in E. coli MetC.............106
Fig. 4.4 Michaelis–Menten plots of various activities in Wolbachia MetC.................108
Fig. 4.5 Michaelis–Menten plots of various activities in P. ubique MetC...............109
Fig. 4.6 Michaelis–Menten plots of various activities in T. maritima MetC..........110
Fig. 4.7 The effect of temperature on the parameters of T. maritima MetC.........112
Fig. 4.8 Mass spectra of cystathionine elimination products ......................114–6
Fig. 4.9 T. maritima MetC crystal and resolved structure..............................117
Fig. 4.10 Domains of the T. maritima MetC monomer................................118
Fig. 4.11 Difference in the entranceway..................................................119
Fig. 4.12 Difference in active site entrance ............................................120
Fig. 4.13 The occluding loop .................................................................121
Fig. 4.14 Conservation of residues across the protein.................................122
Fig. 4.15 Subtle differences between the active site ................................123
Fig. 4.16 R59–E235 in E. coli MetC and G45–K235 in T. maritima MetC.........124
Fig. 4.17 Potential glutamate position in the active site ...............................125
Fig. 4.18 Residues at positions equivalent to 108 across the MetC tree ..............125
Fig. 5.1 Distribution of number of mutations........................................145
Fig. 5.2 Michaelis–Menten plots of vv. activities in MetC S86T/S305C.............149
Fig. 5.3 Location of S305 and S86 residues.............................................151
LIST OF TABLES

Table 2.1 Cell wall composition of taxa without *alr*.................................................. 25
Table 3.1 Summary statistics for the datasets used.......................................................... 65
Table 4.1 *E. coli* MetC activities measured in other studies........................................ 104
Table 4.2 *E. coli* MetC activities measured in this study.............................................. 104
Table 4.3 *Wolbachia* MetC activities ............................................................................. 106
Table 4.4 *P. ubique* MetC activities ................................................................................. 110
Table 5.1 Mutational frequencies in the sampled population ............................................. 143
Table 5.2 Kinetics of *T. maritima* MetC S86T/S305C...................................................... 148
Table 5.3 Differences between the mutant and wild-type ................................................ 148
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH</td>
<td>Alanine dehydrogenase</td>
</tr>
<tr>
<td>SINA</td>
<td>SILVA incremental aligner</td>
</tr>
<tr>
<td>ASKA</td>
<td>A complete set of <em>E. coli</em> K-12 ORF archive</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>CHES</td>
<td>N-cyclohexyl-2-aminoethanesulfonic acid</td>
</tr>
<tr>
<td>COG</td>
<td>Cluster of orthologous genes</td>
</tr>
<tr>
<td>Cth</td>
<td>Cystathionine</td>
</tr>
<tr>
<td>DAAO</td>
<td>D-amino acid oxidase</td>
</tr>
<tr>
<td>DTNB</td>
<td>5,5'-dithiobis-2-nitrobenzoic acid (Ellman's reagent)</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>GDH</td>
<td>Glutamate dehydrogenase</td>
</tr>
<tr>
<td>GTRCAT</td>
<td>Generalised time reversible model with CAT (discrete category based) distribution</td>
</tr>
<tr>
<td>GTRΓ</td>
<td>Generalised time reversible model with Γ distribution</td>
</tr>
<tr>
<td>His₆</td>
<td>Hexahistidine tag</td>
</tr>
<tr>
<td>IAD</td>
<td>Innovation, amplification and divergence model</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>iTOL</td>
<td>Interactive tree of life</td>
</tr>
<tr>
<td>JGI IMG</td>
<td>Joint Genome Institute, Integrated Microbial Genomes server</td>
</tr>
<tr>
<td>$k_{cat}$</td>
<td>Turnover number</td>
</tr>
<tr>
<td>$K_M$</td>
<td>Michaelis constant</td>
</tr>
<tr>
<td>LB</td>
<td>Lysogeny broth</td>
</tr>
<tr>
<td>LC/MS</td>
<td>Liquid chromatography in tandem with mass spectroscopy</td>
</tr>
<tr>
<td>LDH</td>
<td>Lactate dehydrogenase</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LPSN</td>
<td>List of prokaryotes with standing in nomenclature</td>
</tr>
<tr>
<td>LSU</td>
<td>Large subunit of ribosomal RNA</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>Multiple sequence comparison by log-expectation</td>
</tr>
<tr>
<td>NADH/NAD⁺</td>
<td>Nicotinamide adenine dinucleotide (hydride)</td>
</tr>
<tr>
<td>NADPH/NADP⁺</td>
<td>Nicotinamide adenine dinucleotide phosphate (hydride)</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>OD₆₀₀</td>
<td>Optical density at λ=600 nm</td>
</tr>
<tr>
<td>epPCR</td>
<td>Error prone polymerase chain reaction</td>
</tr>
<tr>
<td>PEDel-AA</td>
<td>Programme for estimating diversity in error-prone PCR libraries</td>
</tr>
<tr>
<td>PLP</td>
<td>Pyrodoxa-5'-phosphate</td>
</tr>
<tr>
<td>Pub</td>
<td><em>Pelagibacter ubique</em></td>
</tr>
<tr>
<td>RAxML</td>
<td>Randomized accelerated [sic.] maximum likelihood</td>
</tr>
<tr>
<td>SOB</td>
<td>Super optimal broth</td>
</tr>
<tr>
<td>SOC</td>
<td>Super optimal broth with catabolite repression</td>
</tr>
<tr>
<td>SSU</td>
<td>Small subunit of ribosomal RNA</td>
</tr>
<tr>
<td>TEMED</td>
<td>Tetramethylthlenediamine</td>
</tr>
<tr>
<td>TEV</td>
<td>Tobacco etched virus</td>
</tr>
<tr>
<td>Tma</td>
<td><em>Thermotoga maritima</em></td>
</tr>
<tr>
<td>Tris</td>
<td>Tris(hydroxymethyl)aminomethane</td>
</tr>
<tr>
<td>WAG</td>
<td>S. Whelan and N. Goldman substitution model</td>
</tr>
<tr>
<td>wMel</td>
<td><em>Wolbachia sp.</em> endosymbiont of <em>Drosophila melanogaster</em></td>
</tr>
</tbody>
</table>
Chapter 1

Primordiality, Promiscuity and Multitasking
1.1 In the beginning

DNA is often celebrated as the molecule that makes life possible, but it is the encoded proteins that actually carry out the tasks that allow life to exist. Their tasks range from organising large structures to making complex organic molecules out of simple inorganic ones. The latter task is done by enzymes: these proteins bind specific substrates and lower the energy barriers of certain reactions, thus enabling these reactions to occur, even when they would be effectively impossible spontaneously. For the system to work efficiently enzymes have generally evolved to be highly specific towards certain compounds in order to not to waste other cellular metabolites and, in many cases, to maximise the rate at which these substrates are turned over.

Whereas modern enzymes are often highly specific, primordial enzymes are believed to have possessed very broad specificities. In 1976 Roy Jensen conjectured that primordial enzymes may have recognised a broad range of substrates and, as a result of their combined broad range of substrates and products, a patchwork metabolic network might have arisen that was able to produce complex compounds (Fig. 1.1). This network included not only those that were depleted from abiotic stocks, but also novel compounds, some of which were synthesised from unstable intermediates (Jensen, 1976). Several pieces of evidence support and expand on this hypothesis. Specifically, chromosomes are believed to be a late evolutionary invention and instead protocells may have possessed unlinked RNA genes that encoded enzymes. This would be problematic if the number of different genes required to be inherited were large and consequently genes encoding enzymes capable of a broad range of activities were favoured (Szilágyi et al., 2012). Further evidence in favour of the patchwork model is that modern central metabolism is achieved thanks to a very limited number of protein families: 88% of the protein domains in small molecule biosynthesis in Escherichia coli fall into only 213 fold superfamilies, which could indicate that the primordial ancestors of these folds could catalyse the broad range of activities seen subdivided among their modern descendants (Teichmann et al., 2001). As metabolism assembled gradually, possibly only nine different surviving folds might have been present right when primordial protein-based metabolism first started, after which other proteins with novel folds appeared (Caetano-Anollés et al., 2007, Caetano-Anollés et al., 2009). After billions of years of evolution the multitude of descendants seen today still bear a
structural resemblance, but have profoundly different properties to their primordial ancestors.

**Figure 1.1.** The rise of early metabolism. It has been hypothesised that in order for metabolism to assemble, primordial enzymes (Pacman-like cartoons in left hand panel) might have been able to catalyse a large number of reactions resulting in a network of many metabolites (small circles), including some useful metabolites (small stars). Enzymes in this network duplicated and specialised to the enzymes found today that perform a single step in a pathway (right hand panel). For details of image creation of all figures, see § I.II.XIV.

### 1.2 Many substrates, many mechanisms

Modern enzymes are most often specialists that physiologically perform a single reaction on a single substrate. This occurs despite the fact that a single substrate can be subjected to many possible reactions and, conversely, a particular mechanism can be adopted on many substrates. As a consequence of this, most modern enzymes are poor models to infer the properties of primordial enzymes. In this thesis, modern enzymes that have evolved to be able to perform multiple reactions on multiple substrates are sought and studied in order to better understand the properties of primordial enzymes.

The current comparative models for primordial enzymes are enzymes that possess a property called promiscuity, and certain enzymes that have lost their specificity. The promiscuity, discussed in section 1.3, is a biochemical property that is inherent to enzymes and is important in the evolution of new functions. Enzymes that have lost their specificity, discussed in section 1.7, are intermediates in directed evolution studies aimed at changing
the activity of an enzyme. These enzymes lose their ancestral specificity in order to gain a new specificity and, therefore, progress through a generalist state, able to catalyse several reactions equally well. Consequently, enzyme promiscuity and generalist intermediates will be introduced and, throughout this thesis, compared to the properties of the enzymes identified.

1.3 Promiscuity in enzymes

Even though modern enzymes are highly specialised, many, if not all, possess small side activities that arise fortuitously as a consequence of the chemistry performed by the active site and of the accommodative layout of the active site. These activities, called promiscuous activities, are maintained neutrally, i.e. not subjected to an evolutionary pressure (Khersonsky and Tawfik, 2010b). Despite arising without evolutionary pressures, enzyme promiscuity is important in the evolution of new functions. If a new evolutionary pressure occurs, under which the promiscuous activity is beneficial, the promiscuous activity can act as a starting point for the evolution of the new function (Khersonsky et al., 2006).

There are various forms and degrees of promiscuity that enzymes can accommodate. Catalysis on different substrates via the same mechanism is called substrate ambiguity, while catalysis via a different mechanism on the same or a different substrate is called catalytic promiscuity (Khersonsky et al., 2006). Regarding the latter case, different mechanisms entail the cleavage or formation of different bonds, or the presence of a different transition state (Hult and Berglund, 2007). Examples of both can be found in chymotrypsin, a serine protease whose physiological role is to hydrolyse the amide bonds of peptides, but that is also able to hydrolyse several other types of bonds on non-physiological substrates (O’Brien and Herschlag, 1999). Chymotrypsin can, for example, hydrolyse esters via an identical reaction to its native amide hydrolytic activity (i.e. substrate ambiguity), but it can also hydrolyse phosphonate esters (i.e. catalytic promiscuity) via a reaction that does not proceed via a tetrahedral transition state like in the native reaction, but via a triangular bi-pyramidal intermediate (O’Brien and Herschlag, 1999).

Catalytic promiscuity can also occur when an enzyme catalyses more than one transformation on a single substrate. An apparent example is the ability of E. coli MenF to promiscuously aminate chorismate in addition to catalysing its native activity, the
isomerisation of chorismate. Both reactions share an initial elimination reaction, but differ in the position and identity of the attacking nucleophile, which is a hydroxide in the physiological reaction and an amine in the promiscuous one (Patrick et al., 2007). This is also a good example of the variable nature of promiscuity: while MenF can catalyse this promiscuous activity, its isozyme, EntC, probably lacks this promiscuous activity (Patrick et al., 2007).

In addition to these two major subdivisions (substrate ambiguity and catalytic promiscuity), there are several special cases. A case similar to substrate ambiguity is the ability to accept a non-physiological cofactor. For example, several NADH-dependent enzymes can use the photolabile 2-azido derivative of NAD⁺ in place of NAD⁺ (Kim and Haley, 1990). In some instances the different cofactor can induce altered properties, including the ability to catalyse other reactions. For example, when carbonic anhydrase was bound to manganese as opposed to zinc, it gained peroxidation and epoxidation activities (Fernandez-Gacio et al., 2006; Okrasa and Kazlauskas, 2006), while carbonic anhydrase bound to rhodium gained olefin reduction activity (Jing et al., 2009). This phenomenon has been defined as cofactor promiscuity (Khersonsky and Tawfik, 2010a) or cofactor-induced catalytic promiscuity (Hult and Berglund, 2007).

Enzyme promiscuity, both in the form of substrate ambiguity and of catalytic promiscuity, is widespread in nature. This is demonstrated by several studies that used a genome-wide approach to uncover this diversity. This was done by probing a library of pooled plasmids bearing each E. coli gene, for genes whose encoding enzyme possessed a promiscuous activity that is beneficial under a particular environment. This diversity of promiscuous activities has been dubbed the “promiscuome” (Andersson, 2011). One study found that 17 out of the 104 auxotrophic E. coli single-gene knock-outs could be rescued by a different over-expressed E. coli gene (Patrick et al., 2007). Another study found that 15 over-expressed E. coli genes could impart an improved growth in specific toxic environments, out of the 237 environments tested (Soo et al., 2011). A third found that several genes could assemble into three different pathways to circumvent an auxotrophy caused by the deletion of a pyridoxine biosynthetic gene, pdxB (Kim et al., 2010).

Consequently, promiscuity is not a peculiarity of some enzymes, but a widespread phenomenon, which plays an important part in the evolution of new enzyme activities as
will be discussed in the following sections. The evolution of new enzyme activities, in turn, can be used to extrapolate how primordial enzymes evolved.

### 1.4 Origin of promiscuous activities

**Evolution** can only select pre-existing traits and cannot cause mutants with certain advantageous traits to arise. This is true for enzymes too: in order for the positive selection of a new enzymatic activity to occur this activity needs to be present. It needs to be sufficient to provide a fitness advantage for selection. Enzyme promiscuity therefore acts as a starting point in the evolution of new activities. Because this also applies to primordial enzymes, early primordial enzymes must have possessed a large repertoire of promiscuous activities in order for primordial metabolism to start.

By definition, enzyme promiscuity is not under selection. The neutral theory of molecular evolution describes how mutations fix in the population by random drift as their influence is below the threshold for being either deleterious or advantageous (Kimura, 1968). These neutral mutations do not affect the main activity, but they affect the presence, absence and kinetic parameters of the various promiscuous activities that an enzyme may have. In general, the kinetic parameters of similar promiscuous activities tend to vary together (Bloom *et al.*, 2007). This has been shown, for example, with the changes in specificities in mutants of an engineered P450 peroxygenase under negative selection. Mutants of the enzyme were selected solely on the basis of the retention of the “native” ability to hydroxylate 12-<i>p</i>-nitrophenyloxydecanoic acid. While the native activity was unchanged, the parameters of the promiscuous activities shifted depending on the chemistry, for example, in several mutants the preference for various bicyclic substrates increased together and concomitantly the activity towards phenolic ether substrates decreased together (Cirino and Arnold, 2003, Bloom *et al.*, 2007). In a similar way, neutral mutations may result in the appearance of new promiscuous activities due to the fact that residues that are critical in coordinating substrates for promiscuous activities might not play a role in the main activity.

However, not all extant promiscuous activities arise neutrally. Some were ancestral activities that are no longer under selection and have since drifted into becoming vestigial promiscuous activities. One such example is found in members of the phosphatase superfamily; these enzymes physiologically hydrolyse phosphate monoesters, phosphate
diesters, phosphate triesters or sulfate esters, yet still possess some (vestigial) promiscuous activity towards other kinds of esters (Mohamed and Hollfelder, 2013).

1.5 Gene duplications

Despite being the result of neutral drift, promiscuous activities play a large part in the evolution of enzyme activities. When an activity stops being a promiscuous activity by conferring a selectable fitness advantage, the main activity of the ancestor must be still maintained. The end result of these two opposing pressures is (often) gene duplication and divergence, which allows one copy to encode an enzyme with the ancestral activity and another to encode an enzyme with the newly evolved activity (O’Brien and Herschlag, 1999). The number of mutations to improve the new activity that occurs before the gene is duplicated varies. The gene could be caught between two different selective pressures before it duplicates (Escape from Adaptive Conflict (EAC) model, Des Marais and Rausher, 2008) or the amplification step itself could allow the promiscuous activity to confer a selectable advantage, which would otherwise be too small (Innovation, Amplification and Divergence (IAD) model, Bergthorsson et al., 2007).

In one laboratory study designed to test the feasibility of the IAD model, *Salmonella enterica* hisA was made to evolve to compensate for the loss of a gene encoding an enzyme with a similar function (*trpF*), by starting with a hisA mutant encoding an enzyme that had a small activity towards the substrate of the latter. As a result of selective pressure on this promiscuous activity during serial passaging, the encoding gene was found to evolve by amplifying and then diverging in many, but not all lineages. After 3,000 generations, the result in most lineages was the maintenance of a duplication due to the specialisation of the genes, either as two specialists (16/30 lineages) or as a specialist and a generalist (8/30) (Näsvall et al., 2012).

Several studies have characterised the properties of the last common ancestors of groups of genes that have since diverged apart enabling the encoded enzymes to possess different main activities. This can be investigated by ancestral sequence reconstruction, which involves the prediction and synthesis of the sequence of an ancestor (Williams et al., 2006). Whereas the results of these studies generally confirm that new enzyme activities arise due to selection of a side activity in the ancestral enzyme, the degree of this ancestral activity varies between models studied. In some, the ancestral gene encodes a generalist,
such as an enzyme capable of catalysing the activities of the descendants moderately well (e.g. the broad specificity of the last common ancestor of a superfamily of serine proteases seen in Wouters et al. (2003)). In others, the ancestor encodes a specialist with a small promiscuous activity, such as the ancestral metazoan steroid receptor (Thornton et al., 2003). In one study, a family of genes encoding enzymes (fungal α-glucosidases) that specialised in different lineages towards maltose or isomaltose hydrolysis was found to descend from a gene encoding a generalist in one lineage, but from a gene encoding a maltose specialist with a small side activity in another (Voordeckers et al., 2012).

1.6 Multitasking enzymes

Even though most genes evolved by duplication and divergence, this is not always the case as not all enzymes have evolved to catalyse a single reaction on a specific substrate. Similarly to enzyme promiscuity, the enzymes that present the exception can be subdivided into enzymes with dual specificity or broad specificity, such as chymotrypsin (§1.2), and into enzymes that physiologically catalyse two or more different reactions. The former group is analogous to substrate ambiguity, while the latter is analogous to catalytic promiscuity. The major difference between these enzymes and promiscuous enzymes is that their activities have been under selective pressures to perform multiple physiological roles.

While enzymes that catalyse the same physiologically relevant reaction on different substrates are commonly referred to as having broad specificity, the case in which a single enzyme catalyses different physiological reactions does not have a clear name, due to the rarity of such cases. The term “multifunctionality” is a general term that often refers to proteins that perform more than one physiological activity as a result of fusion of two different enzymes. The term “moonlighting” is also similar, but most often encompasses cases when a peptide adopts a non-catalytic role, such as a structural or signalling one (Copley, 2003). Consequently, herein the term “multitasking” is adopted to describe both dual/broad-specificity and the case of enzymes that catalyse two or more different reactions; this latter case will be referred to as “catalytic multitasking” in analogy with catalytic promiscuity.

Multitasking with respect to physiological substrates (“substrate multitasking”) can be further subdivided depending on the degree of specificity that is displayed. This can range from dual specificity, i.e. specificity towards two substrates, but not, presumably, other
similar compounds to broad specificity, *i.e.* indiscriminate specificity towards a certain group of substrates.

In the case of broad specificity, several enzymes, such as xenobiotic-biodegrading enzymes (Summers *et al.*, 2011) and aminopeptidases that target exogenous peptides (Jankiewicz and Bielawski, 2003), are multitasking by necessity as they target an unknown but variable set of substrates.

In the case of dual specificity, an example is *Mycobacterium tuberculosis* PriA, which catalyses two reactions that in other organisms are catalysed by separate enzymes (TrpF and HisA, mentioned in § 1.4). This enzyme can isomerise two different phosphoribose derivatives, but not others (Due *et al.*, 2011). Several other examples are known. In *Bacillus subtilis*, HemY catalyses two consecutive steps in heme biosynthesis (Hansson and Hederstedt, 1994). In some species a folate-dependent formimino-/formyl-transferase catalyses the same reaction on two different substrates, one of which is processed differently in other species (Jeanguenin *et al.*, 2010). In *Haloferax mediterranei*, D-2-hydroxyacid dehydrogenase accepts both NADH and NADPH, therefore displaying dual coenzyme specificity, analogously to the substrate promiscuity in terms of cofactor (Domenech and Ferrer, 2006).

Enzymes generally display substrate ambiguity towards substrates that are similar to the physiological substrate, therefore the recognition of multiple substrates for the main activities might also result in a broader substrate ambiguity in these enzymes compared to enzymes that catalyse a single activity. For example, whereas there are four dedicated deoxynucleotid kinase in most animals, in *Drosophila melanogaster*, a single gene encoding a multispecific deoxynucleotid kinase is present. This kinase also has a reduced ability to discriminate against non-natural nucleotide analogues (Munch-Petersen *et al.*, 1998). Similarly, the dihydroxy-acid dehydratase from *Sulfolobus solfataricus* can dehydrate the physiological metabolites dihydroxyisovalerate and D-glucanate, but can also, to a lesser degree, dehydrate non-physiological substrates that are structural intermediates between the two substrates (Kim and Lee, 2006).

While many cases of substrate multitasking are known, very few cases of catalytic multitasking are known. One example is that, in *Thermoproteus neutrophilus*, two consecutive gluconeogenesis steps are catalysed by a bifunctional fructose-1,6-bisphosphate aldolase/phosphatase, rather than two specialized enzymes (*e.g.* FbaA and Fbp in *E. coli*).
These two different reactions are possible because the active site remodels itself (Du et al., 2011). Another example is the dihydroneopterin aldolase (FolB) in *M. tuberculosis*, which can act on its substrate not only as an aldolase, but also as an epimerase and oxygenase in order to produce its own competitive inhibitors. This potentially allows better regulation of the production of the main product, a folate precursor (hydoxymethylidihydropterin) (Czekster and Blanchard, 2012).

It should be noted, however, that if an enzyme performs a single reaction catalysed in other organisms by a series of enzymes, it does not necessarily mean it is multitasking. For example, the flavin-dependent thymidylate synthase (*thyX*-encoded) converts deoxuryridine monophosphate to thymidine monophosphate and is found in place of both thymidylate synthase (*thyA*-encoded) and folate reductase (*folA*-encoded) (Leduc et al., 2007). However, it actually does not catalyse both reactions, but instead catalyses a single more direct reaction (Koehne et al., 2009). A similar situation is seen in pyridoxal-5’-phosphate (PLP) biosynthesis in *B. subtilis* where PLP synthase (*pdxS*-encoded), aided by a glutamate deaminase (*pdxT*-encoded), produces PLP directly from three common metabolites, as opposed to the eight enzymatic steps in *E. coli* (Raschle et al., 2007).

A further issue is that it is not always clear-cut whether an activity is under evolutionary selection. The boundary between promiscuity and multitasking is especially blurred in secondary metabolism. Enzymes involved in plant secondary metabolism have broadened specificities, which allow plants to produce a large variety of compounds, including some that confer a small fitness advantage, such as by deterring herbivores (Fischbach and Clardy, 2007). This lack of a clear-cut boundary between promiscuity and multitasking applies for both substrate ambiguity and for catalytic promiscuity. An example of the latter is that some sequiterpene synthases cyclise a single compound, farnesyl pyrophosphate, into a variety of different products via subtly different mechanisms, resulting from the ring closure at different positions after the elimination of the pyrophosphate (Fischbach and Clardy, 2007).

Consequently, even though multitasking enzymes represent a minority, the group is diverse, ranging in terms of both range of activities and selection pressure for these activities. On one end of this spectrum lie catalytic multitasking enzymes, because they are able to balance the requirements for the catalysis of different types of reactions, potentially making them good comparative models for primordial enzymes.
1.7 Evolvability and robustness

The property to generate mutations that could be beneficial in certain environments is called evolvability. This property is not directly selectable, but, as an increased evolvability can occasionally allow beneficial mutations to arise, it can be carried along (unselected) with the beneficial mutations (selected) (Masel and Trotter, 2010). A closely linked concept is robustness, which is the ability of a system to tolerate mutations, without which any mutation would be deleterious and evolvability would be impossible (Masel and Trotter, 2010). One extreme example of the importance of evolvability is the non-specific cleavage activity (star activity) in restriction endonucleases. Despite the fact this activity is highly deleterious and kept to a minimum, bacterial mutants with more promiscuous restriction endonucleases are much better at adapting against mutant phages (Vasu et al., 2012).

When a promiscuous activity is put under selection, in many cases the native activity is robust and it suffers little from the increase in activity of the new function. This weak trade-off was seen, for example, in the evolution of esterase activity in serum paraoxonase, in bacterial phosphotriesterase and in carbonic anhydrase II (Aharoni et al., 2005). These enzymes additionally possessed other promiscuous activities, which were not placed under selection. These were found to be less robust than the native activity and instead fluctuated greatly (Aharoni et al., 2005). Nevertheless, the trade-off in activity is not always weak between the nascent activity and the native activity. In some instances there can be a strong negative trade-off, where the native ability is greatly diminished. This is the case for a mutation of *Rhodobacter sphaeroides* tyrosine ammonia-lyase (H89F), which made the enzyme 20,000 times less catalytically efficient at deaminating tyrosine (to coumaric acid, the native activity), but only 200 times more active at deaminating phenylalanine (to cinnamic acid) (Watts et al., 2006).

Directed evolution is a method whereby a mutant pool of a gene is made and the variants that encode the most active enzymes for a new activity are selected *in vitro* or *in vivo*. In many of these studies, the evolution progresses with a weak trade-off in activities and a generalist enzyme is initially encountered. This enzyme is able not only to catalyse both reactions fairly well, but has broadened substrate ambiguity. This is because the ancestral specificity needs to be lost before a new specificity can be gained and as a result the intermediate is non-specific. After the generalist stage, further mutations towards an
improved activity favour specialisation. One example is the intermediate encountered during the directed evolution of *E. coli* β-glucuronidase into a β-galactosidase, which was able to catalyse the hydrolysis of other sugars, including galactoside-6-phosphate, that neither the ancestor nor the final specialised mutant could hydrolyse (Matsumura and Ellington, 2001). The ability of these evolutionary intermediates to catalyse a broad range of reactions is a characteristic that primordial enzymes are believed to have possessed (§ 1.1).

The fitness landscape (Fig. 1.2) is an illustrative plot of fitness for activities across sequence space (Romero and Arnold, 2009). Many paths from one sequence to another are possible, but in the case of directed evolution studies the paths taken are those of the most successful mutants out of a large population, even though other variants are possible. Multitasking enzymes are subjected to different evolutionary pressures than generalist intermediates and the path that leads to and from a multitasking enzyme might differ from one that progresses through a generalist intermediate. In the case of tyrosine and phenylalanine ammonia lyases, the two reactions are not always mutually exclusive, as demonstrated by the maize tyrosine/phenylalanine ammonia lyase that can catalyse both reactions with similar efficiency, albeit a hundred fold less efficiently than the phenylalanine specialist in parsley or *Arabidopsis thaliana* (Watts *et al.*, 2006).

![Figure 1.2](image-url)

**Figure 1.2.** Cartoon of a fitness landscape. The surface is sequence space and the fitness conferred by an activity (either red or blue) is depicted as a contour map. The landscape is not smooth, but possesses residues that work synergistically or antagonistically when in combination with others, occasionally producing local maxima. Evolution from one activity to the other might progress with a strong trade-off (*i.e.* a gain for the new activity comes at the cost of the ancestral) or a weak trade-off (*i.e.* a gain for the new activity come at a minor cost to the ancestral activity).
As mentioned, evolvability is enabled by robustness. There have been some studies to quantify the robustness of an activity in an unbiased way. One of these studies used HIV-1 protease and found that 54% of the single mutations were non-functional, 32% were functional but deleterious, 12% neutral and 2% increased the catalytic efficiency of the enzyme. When the starting point harboured a clinical mutation (D30N) that decreased activity 40-fold, but conferred resistance to a competitive inhibitor (nelfinavir), the robustness increased (52% were lethal, 13% deleterious, 26% neutral and 9% beneficial). This was due to several mutations behaving differently when in combination with other mutations resulting in a rugged fitness landscape (Parera et al., 2007).

1.8 Other factors affecting enzymes

In addition to the optimisation of the main activity, there are other factors at work influencing the fitness endowed by a gene, such as repression of deleterious side activities and stability of the encoded enzyme.

An example of repressing deleterious side activities is seen in Bacillus brevis tyrocidine synthetase (tycA-encoded), an adenylyltransferase involved in producing a non-ribosomal peptide (tyrocidine). The enzyme has evolved to discriminate between its substrate, phenylalanine, and other natural amino acids, such as the closely related tyrosine, which is accepted a thousand fold less. This is because the incorporation of incorrect amino acids is deleterious as it would result in a non-functional product. However, it has not evolved to discriminate against unnatural amino acids, such as D-phenylalanine, β-cyclohexyl-L-alanine, 4-amino-L-phenylalanine, L-norleucine, which are accepted as substrates more readily than tyrosine due to the lack of a selective pressure against this substrate ambiguity (Villiers and Hollfelder, 2009). This repression of deleterious activities may come at the cost of the main activity: in one study a five-residue mutant of the HaeIII methyltransferase evolved to recognise a different sequence, displayed not only a relaxed specificity similar to generalist intermediates, but an unselected nine-fold higher catalytic efficiency for the ancestral target sequence as opposed to a trade-off (Cohen et al., 2004).

Regarding protein stability, broad specificity is often a result of a more accessible or disordered active site (Perona and Craik, 1995), which may in turn have an effect on stability. The transition from the hydrolysis of penicillins to cephalosporins in several clinically isolated TEM-1 β-lactamases came at the cost not only of a decrease in the native
activity, but also of stability, a trade-off that was rectified in other clinical isolates thanks to a stabilizing mutation (M182T) (Wang et al., 2002). Another example is that clinical isolates of AmpC mutants capable of hydrolysing cefotaxime showed a weak trade-off with the ancestral activity, the hydrolysis of cephalothin, but possessed a 3–7°C decrease in thermostability (Thomas et al., 2010). Conversely, the reconstructed Precambrian β-lactamase ancestor for Firmicutes, Actinobacteria and Proteobacteria was hyperstable and had a broad specificity to the point that it could cleave even man-made cephalosporin antibiotics, such as ceftazidime and ceftriaxone. This was in contrast to the modern descendant, TEM-1 β-lactamase, which is a penicillin specialist (Risso et al., 2013). A trade-off for the hyperstability and broad specificity of the ancestor is a hundred fold lower turnover for penicillin compared to the specialist TEM-1.

In summary, reaction catalysis, substrate specificity, evolvability, robustness and thermostability form a complex balance of factors that shape the evolution of enzymes.

### 1.9 Aims of this thesis

The overall aim of this thesis was to explore some of the factors in enzyme evolution. In particular I sought to investigate the biochemical differences between catalytic multitasking and catalytic promiscuity, thus ascertaining the best comparative model for primordial enzymes. This study built on previous work that had identified the E. coli cystathionine β-lyase (MetC) as having promiscuous alanine racemase (Alr) activity (Soo, 2012). My overall approach was to use bioinformatics and phylogenetics to identify candidates for multitasking (rather than promiscuous) MetC homologues encoded in bacterial genomes, and then to use biochemistry, directed evolution and structural biology (via a collaboration) to study the properties of the enzymes. The specific objectives of each chapter are described below.

This project was initially conceived to identify bacterial species in nature in which a gene (metC) had displaced another gene (alr). This was done to determine whether the promiscuous activity of MetC for the activity of Alr (alanine racemisation) previously determined in the lab had any evolutionary utility in nature. This was done and three candidate genes found, but the questions addressed changed when it was discovered that two of the enzymes catalysed both reactions physiologically as opposed to losing the
ancestral activity. Consequently, the focus was to determine the properties of a multitasking enzyme as little is known about them and as they could provide good models for primordial enzymes, which are believed to be multitasking.

The aim for chapter 2 was to identify multitasking MetC enzymes. Achieving this aim required addressing a series of specific objectives. It was necessary to determine whether any bacterial species has metC, but not alr using bioinformatics. Once candidate genes were identified, the second objective was to clone them and test whether they could rescue E. coli strains that lacked metC and alr. The third objective was to assess whether the metC genes were physiologically relevant for methionine biosynthesis in their host organisms.

The aim for chapter 3 was to determine the phylogenetic relationship of two of the organisms, P. ubique and Wolbachia. The phylogeny of the Alphaproteobacteria was inferred using a concatenated 16S and 23S dataset. To allay fears of GC-bias driven artifacts several hypotheses were tested. The first objective was to determine whether rRNA has the same GC-bias as genomic GC-bias in the Alphaproteobacteria. The second was establishing whether the same result was obtained when certain orders are removed. The third objective was to establish whether the same result is obtained when the dataset is recoded to remove GC-content.

The aim of chapter 4 was to determine the kinetic parameters of the enzymes and the structural features of T. maritima MetC. This was done in order to determine the differences between catalytic multitasking and catalytic promiscuity. This was accomplished by assaying each enzyme for cystathionine elimination, alanine racemisation, glutamate racemisation and cysteine elimination. Additionally, the crystal structure of T. maritima MetC was solved by our collaborators and analysed in order to determine which residues are key to the improved racemisation.

The aim for chapter 5 was to investigate the change in activities when T. maritima MetC is mutated This was done in order to determine the interplay of the constraints imposed by the various activities. This required the creation of an error-prone PCR library of T. maritima metC, followed by the identification of mutants that rescued a knockout faster than wild-type T. maritima metC and their biochemical characterisation. This data allowed a comparison of the mutants and wild-type.
The genetics of \textit{metC} in \textit{Wolbachia, P. ubique} and \textit{T. maritima}

\textbf{Acknowledgements:}

- Dr Valerie Soo (formerly at Institute of Natural Sciences, Albany, Massey University) constructed pCA24N/\textit{metC\textsubscript{Eco}}
- Natasha le Roux (formerly at Institute of Natural Sciences, Albany, Massey University) constructed pMAL/\textit{metC\textsubscript{Pub–malE}}
2.1 Introduction

2.1.1 Reciprocal promiscuity of alr and metC

As mentioned in the previous chapter (§ 1.5), some studies have partially uncovered the abundance of promiscuity in the E. coli translated genome, often with unpredicted results (Patrick et al., 2007, Kim et al., 2010, Soo et al., 2011). One unexpected case was the ability of overexpressed alr to suppress the methionine auxotrophy of a ΔmetC strain, despite the fact that the two genes encoded enzymes with different structures and with physiological activities involving different mechanisms and substrates (Patrick et al., 2007). Despite the difference, the reciprocal was also true, namely that metC could rescue a knockout strain that lacked alr and its close homologue (Soo, 2012).

2.1.2 MetC and methionine biosynthesis

The enzymes encoded by the metC and the alr genes (MetC and Alr respectively) both rely on the cofactor pyridoxal-5’-phosphate (PLP) for activity (Clausen et al., 1996, Strych et al., 2000), but are non-homologous. In fact, PLP-dependent enzymes belong to seven non-homologous families, which have some overlapping activities because the PLP cofactor enables catalysis thanks to its ability to act as an electron sink in order to stabilize any carbanionic intermediates (Eliot and Kirsch, 2004, revised in Percudani and Peracchi, 2009). MetC belongs to the PLP-dependent fold type I family, while Alr belongs to the type III family, which in turn belongs to the (β/α)_{8} barrel superfamily (Eliot and Kirsch, 2004) (Fig. 2.1). Moreover, the two enzymes catalyse substantially different reactions: MetC catalyses an elimination on a C–S bond (EC 4.4.1.1), while Alr catalyses a racemisation (EC 5.1.1.1) (Webb, 1992, Keseler et al., 2009).

MetC is a cystathionine β-lyase, an enzyme that synthesises homocysteine in the penultimate step of methionine biosynthesis (Fig. 2.2). Its substrate, cystathionine, is a thioether of two α-amino acids conjoined via their one- and two-carbon long straight side chains and consequently the bridging sulfur bonds a β-carbon on one side and a γ-carbon on the other. When the thioether is eliminated at the shorter side, i.e. a β-elimination, the products are pyruvate, ammonia and homocysteine, which is the unmethylated precursor of methionine (Clausen et al., 1996).
Figure 2.1. Structures of the monomers of \textit{E. coli} MetC (A) and Alr (B). Both protein bind PLP (grey), but have different structures. A) MetC (PDB: 1cl1, resi. 62–375 shown from one chain, 29–62 of another shown in darker shade) is a PLP-dependent enzyme of the AAT-like family (fold type I), and is found as a homotetramer and the active site of \textit{E. coli} MetC is formed by two adjoining monomers. B) Alr (PDB: 2rjg, resi. 10–216 shown) is a homodimer, where the N-terminus (shown) is a (\(\beta/\alpha\))\(_8\) barrel.

MetC is in one of several overlapping biosynthetic routes of methionine biosynthesis that are utilised in different bacteria (Fig. 2.2; Belfaiza et al., 1986; Born and Blanchard, 1999; Alaminos and Ramos, 2001; Auger et al., 2002; Picardeau et al., 2003; Goudarzi and Born, 2006; Seiflein and Lawrence, 2006; Hwang et al., 2007; Zubieta et al., 2008; Tran et al., 2011). Methionine is synthesised from the branch point metabolite homoserine, which is first converted into an O-succinyl or an O-acetyl ester via one of two non-homologous enzymes (encoded by \textit{metA} and \textit{metX}; Bourhy et al., 1997; Zubieta et al., 2008). Next, the acyl group is replaced by a sulfur compound, either directly with free sulfide by acyl-homoserine sulfhydrase (encoded by several alternative genes, including \textit{metI} (Auger et al., 2002), \textit{metY} (Hwang et al., 2002) and \textit{metZ} (Alaminos and Ramos, 2001)) or indirectly (trans-sulfurylation) with cysteine to form a thioether (cystathionine) by cystathionine \(\gamma\)-synthase (encoded by \textit{metB} (Aitken et al., 2003) or by the bifunctional \textit{metI} (Auger et al., 2002)). The product, cystathionine, is the substrate for MetC. Each of the two routes produces homocysteine, which is methylated by a methyltransferase (\textit{metH} or \textit{metE} encoded) to form methionine (Alaminos and Ramos, 2001).
Figure 2.2. Methionine and cysteine biosynthesis. MetC (shaded in blue) is part of the forward trans-sulfurylation route for methionine biosynthesis (green arrows), which utilises cysteine as the sulfur source. Other route for methionine and cysteine biosynthesis are the reverse trans-sulfurylation (homocysteine to cysteine, blue arrows), the direct-sulfurylation for homocysteine (yellow arrows) and the direct-sulfurylation for cysteine (purple arrows). Cysteine and methionine biosynthesis and catabolism are heavily intertwined and all operations involving C-S bonds utilise either a homologue of MetC (♣ in figure, type I PLP-dependent enzyme, PF01053, CL0061) or a type II PLP-dependent enzyme (♣, type II PLP-dependent enzyme (PF00291). The acyl transferases belong to the class I amidotransferase fold (♣, PF04204, CL0014) or the α/β hydrolase fold (♣, PF00561, CL0028).

2.1.3 Alr and peptidoglycan

In *E. coli* Alr and its isoyme encoded by *dadX* are alanine racemases, namely enzymes that catalyse the interconversion between L-alanine and D-alanine (Wild et al., 1985). Whereas DadX is involved in alanine catabolism, Alr is involved in peptidoglycan biosynthesis (Strych et al., 2000). Peptidoglycan is a cross-linked polymer of sugars and...
unusual amino acids (Schleifer and Kandler, 1972, Vollmer et al., 2008). Specifically, the amino acids in peptidoglycan are alternating D and L amino acids with some difference present in different bacterial phyla (Schleifer and Kandler, 1972, Vollmer et al., 2008). In most Proteobacteria these are L-alanine, D-glutamate, meso-diaminopimelic acid and D-alanine (Fig. 2.3) (Schleifer and Kandler, 1972, Vollmer et al., 2008).

![Figure 2.3. D-amino acids in proteobacterial peptidoglycan.](image)

2.1.4 Non-homologous gene displacement of alr

Racemases are not only involved in peptidoglycan biosynthesis, but also in a variety of roles. Furthermore, several different racemases are known (appendix II); these are not all homologues of Alr (PLP-dependent fold III family), but fall into four different families (viz. references in Appendix II).

As alanine racemase activity is expected to be present in all bacterial species with peptidoglycan, if an organism were to lack alr, a non-orthologous gene may have taken over the role. This phenomenon is called non-orthologous gene displacement (Koonin et al., 1996, Osterman and Overbeek, 2003, Green and Karp, 2004).

Consequently, the aim of this chapter was to identify metC genes that have displaced alr. Three steps were required to do so, namely (i) an in silico search for candidates, (ii) in vivo testing of the activities of the enzymes encoded in these genes and (iii) an in silico analysis of whether the activities are likely to be physiologically relevant.
A displacement of \textit{alr} by \textit{metC} is of interest because it would mean that promiscuous activity seen in \textit{E. coli} MetC has become a main activity in another organism, therefore showing the importance of the evolutionary potential of promiscuous activities encoded in a genome. Furthermore, if the enzyme encoded by this \textit{metC} catalyses not only the alanine racemisation activity, but also maintains its ancestral cystathionine elimination activity, it would be a case of a catalytic multitasking enzyme discussed in chapter 1 (§ 1.6), which may hold unusual biochemical properties.
2.2 Results

2.2.1 Search for species without \textit{alr (in silico)}

A search was conducted to identify bacterial genomes without \textit{alr}, but with \textit{metC}. 1023 fully sequenced and annotated prokaryotic genomes (as of 6\textsuperscript{th} August 2010) were used for the analysis (methods § 2.4.1). Only fully sequenced genomes were used because the absence of a gene cannot be confirmed for draft assemblies, in which the gene sought may lie in a contig gap.

The presence or absence of \textit{alr} and \textit{metC} was determined via COG (cluster of orthologous genes) annotation. A COG is a cluster of genes that share sequence homology (as determined using BLASTp) with each other (Tatusov \textit{et al.}, 2003). Whereas homologues of alanine racemase (\textit{alr}-encoded) cluster together by themselves, there is no distinction between cystathionine $\beta$-lyase (\textit{metC}-encoded), cystathionine $\gamma$-synthase (\textit{metB}-encoded) and acetyl-homoserine (thiol)-lyase (\textit{metZ}-encoded); therefore these form a single cluster. Despite this limitation, COG annotations were preferred because gene symbols and gene names are subjected to a large amount of ambiguity. For example, the locus RoseRS_2559 (Genbank: ABQ90935) from \textit{Roseiflexus sp. RS-1 (phylum Chloroflexi)} is annotated as “Alanine racemase”, but it is actually a homologue of \textit{lacI}, a repressor; this error seems to have been propagated to 76 genomes. Ideally, PFAM annotations would have been more precise as they are motif based; however, few genomes have all of their encoded proteins annotated by PFAM and matching each protein to a motif would have been computationally unfeasible.

The dataset was visualised on a tree, based on NCBI taxonomy to sort the results into taxa of similar composition (Fig. 2.4). From the resulting tree it is clear that \textit{metC/B/Z} was not very conserved while \textit{alr} was highly conserved and absent in only ten taxa (listed in Table 2.1).
Figure 2.4. Rarity of taxa lacking alr. The tree is a taxonomic tree based on NCBI annotation and the clades are coloured based on the presence or absence of metC and alr. Specifically: in green are taxa with both metC/B/Z and alr; in blue are taxa without metC/B/Z, but with alr; in gold are taxa without alr, but with metC/B/Z; and in red are taxa without either gene.

This list of ten taxa was further reduced with the exclusion of unlikely candidates. Six of the taxa are known to lack or possess unusual peptidoglycan (Table 2.1), meaning that they have no requirement for an alanine racemase. Members of the Chlamydiaceae are known to possess peptidoglycan, but appear to lack both alr and metC; consequently they were also excluded from further analysis. Furthermore, a BLASTP search was done to verify the absence of alanine racemase using the E. coli protein as a query (methods § 2.4.2). This revealed that one taxon on the list, Thermomicrobium roseum, was a false positive due to the fact that the analysis did not take into account plasmids. T. roseum possess a 0.9 Mb mega-
plasmid (Wu et al., 2009), where the alr homologue was located (GenBank: YP_002523438).

Table 2.1. Cell wall compositions of taxa without alr. PG: peptidoglycan presence/absence.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>PG?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>No</td>
<td>Members of this domain do not use D-alanine in their pseudopeptidoglycan (Kandler and Konig, 1998).</td>
</tr>
<tr>
<td>Mollicutes</td>
<td>No</td>
<td>This firmicute class, which includes the genus <em>Mycoplasma</em>, is characterised by the use of mycolic acids instead of peptidoglycan (Garrity, 2005).</td>
</tr>
<tr>
<td>Rhodopirellula baltica</td>
<td>No</td>
<td>This is the sole sequenced species of the <em>Planctomycetes</em>, a bacterial phylum with a characteristic membrane enclosed nucleoid (pirellulosome). It is known to lack peptidoglycan and to possess instead a disulfide cross-linked proteinaceous cell wall (S-layer), like archaea (Fuerst and Sagulenko, 2011).</td>
</tr>
<tr>
<td>Dehalococcoides</td>
<td>No</td>
<td>This genus belongs to the class <em>Dehalococcoidetes</em> of the <em>Chloroflexi</em>, a deep-branching phylum that unusually stain Gram-negative due to a very thin or absent peptidoglycan, but that possess instead a thick S-layer (protein) (Surcliffe, 2011). The class <em>Dehalococcoidetes</em> is known to lack peptidoglycan (Fazi et al., 2008).</td>
</tr>
<tr>
<td>Endosymbiotic gammaproteobacteria</td>
<td>No</td>
<td>These are endosymbionts of plant sap-sucking insects with minimal genomes, probably forming a monophyletic clade sister to the family <em>Enterobacteriaceae</em> (Williams et al., 2010). These species are known to lack cell walls (Nikoh and Nakabachi, 2009), consequently they lack alanine racemase.</td>
</tr>
<tr>
<td>Sulcia and Amoebophilus spp.</td>
<td>No</td>
<td>These two genera of the phylum <em>Bacteroidetes</em>, are obligate intracellular symbionts of insects and amoebas, respectively, and have heavily reduced genomes (McCutcheon et al., 2009, Schmitz-Esser et al., 2010). In <em>Amoebophilus asiaticus</em>, alr is present as a pseudogene (Aasi_0149).</td>
</tr>
<tr>
<td>Thermotoga spp.</td>
<td>Yes</td>
<td>This genus belongs to the deep-branching phylum, <em>Thermotogae</em>, members of which feature an unusual sheath-like outer membrane called the toga (Garrity, 2005). These species also possess an unusual peptidoglycan cross-link (1/L-lysine), but still require D-alanine and are sensitive to β-lactam antibiotics, which target cell wall biosynthesis (Boniface et al., 2009). Species of this genus do not have alanine racemase, but the related <em>Thermosipho</em> and <em>Petrotoga</em> spp. do. Two <em>Thermotoga</em> species have a homologue of the regulator lacI (e.g. YP_001471200) misannotated as alanine racemase.</td>
</tr>
<tr>
<td>Chlamydiaceae</td>
<td>Yes</td>
<td>The members of this family (<em>Chlamydia</em> and <em>Chlamydaphila</em>) are obligate intracellular parasites sensitive to β-lactam antibiotics and possess most cell wall genes, but no cell wall has been detected to date. This has been termed the “chlamydial anomaly” (McCoy and Maurelli, 2006).</td>
</tr>
<tr>
<td>Thermomonas roseum</td>
<td>Yes</td>
<td>This bacterium, of the <em>Chloroflexi</em>, possesses a peptidoglycan, layer albeit thin with several unusual features such as acetyl galacosamine substituted for acetylglicosamine (Merkel et al., 1980, Wu et al., 2009).</td>
</tr>
<tr>
<td>Rickettsiales</td>
<td>Yes</td>
<td>Peptidoglycan is absent in some members, but present in others (discussed in § 2.3.3).</td>
</tr>
</tbody>
</table>
Members of the other two groups do not possess an alr gene, but possess metC, making the latter genes candidates for the genes that displaced alr. These two groups included members of the Thermotoga genus and an incomplete subset of species annotated at the time on NCBI as members of the order Rickettsiales. As will be further elaborated in Chapter 3, in reality the latter group consists of Pelagibacter ubique, a member of the Pelagibacterales (which is a sister order to the Rickettsiales), and a subset of the Anaplasmataceae, which is one of the three families in the Rickettsiales.

Having found these three groups, a representative metC gene was chosen from each and used for genetic complementation tests to verify the activity, while the evolutionary history of these metC homologues was investigated and the genomes were checked for other genes that might be acting as alanine racemase enzymes in vivo.

### 2.2.2 Cloning of candidate metC genes

The metC from Wolbachia “endosymbiont of Drosophila melanogaster” was chosen as a representative of one of the three taxa; the common abbreviation for this organism is wMel (Wu et al., 2004). The organism was first characterised in Culex pipiens (mosquito) and named Wolbachia pipientis, but taxonomically it was not established whether Wolbachia symbionts in other arthropods were strains of this species and consequently the usage of the species epithet is discouraged (Lo et al., 2007).

The Wolbachia metC gene (locus: WD0925) was codon optimised for expression in E. coli and commercially synthesised with the addition of an N-terminal His₆-tag followed by a TEV protease site (methods § 2.4.3). The Wolbachia metC gene was cloned (via NcoI and HindIII) into pBAD, which possesses an arabinose inducible promoter that allows titratable expression (Guzman et al., 1995).

Next, whole circle PCR (methods § 2.4.3) was used to introduce a KpnI restriction site into pBAD/metC<sub>Wol</sub>(<strike>kpnI</strike><sup>-</sup>) by introducing a codon for glycine after the codon encoding threonine-15 in the TEV protease site (Fig. 2.5) in order to use the N-terminal tag with T. maritima metC and other genes.
Figure 2.5. pBAD/metC<sub>ωMel</sub>(KpnI<sup>+</sup>) plasmid with and without the KpnI site. The three restriction enzymes used on the plasmid are marked in red on the plasmid map. The sequence validation is the chromatogram trace from Sanger sequencing of the plasmid (Genetics analysis service, University of Otago). Image generated using MacVector and Illustrator CS4.

The metC gene from T. maritima MSB8 was amplified from genomic DNA by PCR. The primers contained restriction sites for KpnI and XbaI, allowing the amplicon to be cloned into pBAD/metC<sub>ωMel</sub>(KpnI<sup>+</sup>) after the Wolbachia metC gene had been removed by digestion with the same enzymes. The resulting plasmid, pBAD/metC<sub>Tma</sub>, was confirmed by DNA sequencing and was used for complementation tests, as well as protein purification in chapter 4.

A codon optimised form of the metC gene from P. ubique HTCC1062 was cloned by Natasha le Roux into pMAL-c5X, creating a fusion protein with N-terminal maltose-binding protein (le Roux, 2012). Technically, P. ubique is a candidate species (Rappe et al., 2002) and should be officially written as “candidatus Pelagibacter ubique” (Stackebrandt et al., 2002); however, for consistency and clarity it will be written as Pelagibacter ubique herein.

2.2.3 Complementation (in vivo)

The complementation tests in E. coli were done with a air dadX double-knockout strain (MB2795), which is auxotrophic for D-alanine (Strych et al., 2001), and a metC
knockout strain (JW2975 {BW25113 metC::kan}), which is auxotrophic for methionine (Baba et al., 2006) (§ 2.4.4).

All three metC genes were able to rescue both the D-alanine auxotroph, E. coli MB2795, and the methionine auxotroph, E. coli JW2975 (Fig. 2.6), albeit at different rates. In the case of the suppression of the D-alanine auxotrophy of MB2795, all rescuers grew within one day nearly as well as the D-alanine supplemented control. In contrast, for the suppression of the methionine auxotrophy of JW2975 different rescuing rates were seen (E. coli metC 2 days, methionine supplementation 2 days, P. ubique metC 5 days, T. maritima metC about 10 days and Wolbachia metC two weeks). For P. ubique metC, growth was seen without IPTG (leaky expression) and with 100 µM IPTG, but not at 500 µM IPTG (toxic over-expression) in both strains.

**Figure 2.6.** Suppression of methionine and D-alanine auxotrophy (on M9 and LB medium, respectively) by Wolbachia, P. ubique and T. maritima metC genes. The complementation tests were performed as described in § 2.4.4. See main text (§ 2.2.2) for growth times. The following plasmids were used: pCA24N/metC<sub>Eco</sub>-gfp (labelled as E. coli metC above the corresponding panel), pCA24N/alr<sub>Eco</sub>-gfp (E. coli alr), pMAL/malE-metC<sub>pub</sub> (P. ubique metC), pBAD/metC<sub>woM</sub> (Wolbachia metC), pBAD/metC<sub>Tma</sub> (T. maritima metC) and pBAD/myc-his(B) (noIns, empty vector control).

**2.2.4 Investigation of alr alternatives in T. maritima (in silico)**

The complementation test (Fig. 2.6) demonstrated that the metC genes from P. ubique, T. maritima and Wolbachia encode enzymes with alanine racemase activities. However, it was possible that the respective genomes might encode other
proteins with alanine racemase activity; that is, the alanine racemase activity of these MetC proteins may not be physiologically relevant. To investigate this, a series of BLASTp searches were done to find potential racemases encoded in the genome (§ 2.4.2).

A caveat to this approach is that BLASTp searches will reveal the presence of annotated genes with sequence homology to the query and therefore misannotated genes, very distant homologues and non-homologous genes with the same function (analogues) may be missed. However, genes annotated as split may have been covered in the search, as in fact, in the searches described below, one gene was found present as two fragments, which revealed that it was misannotated as two separate reading frames due a stop codon potentially introduced by sequencing error (appendix Table § VI.1). Another shortcoming of homology searches is provided by cases of non-homologous gene displacement, such as the displacement of alr by metC, which this chapter is dedicated to determining. The best way to avoid this issue is doing genome-wide in vivo tests as discussed in § 2.3.4. With these caveats in mind, BLASTp was nevertheless deemed to be the best available tool to begin assessing gene presence in P. ubique, T. maritima and Wolbachia.

Firstly, expected racemases were searched for their presence. For example, the peptidoglycan of T. maritima is known to contain D-alanine, D-glutamate and D-lysine (Boniface et al., 2009). Therefore, glutamate and lysine racemase genes were expected to be present in the genome.

Glutamate racemase is the product of the murI gene (Doublet et al., 1993). In some organisms, a gene is present that racemises aspartate (racX-encoded; Yohda et al., 1991). The E. coli glutamate racemase shares 23% sequence identity with Bacillus subtilis aspartate racemase, making them close homologues. BLASTp searches revealed that, unexpectedly, murI genes are only present in two species of Thermotoga and absent in the rest. These two species are T. lettingae and T. thermarum, which will be called for simplicity herein the T. thermarum subgroup, while the remainder will be referred to as the T. maritima subgroup. The racX gene was also absent from the T. maritima subgroup, ruling out both MurI and RacX as sources of alanine racemase activity. The two subgroups are monophyletic (appendix Fig. III.1). The murI gene in the T. thermarum subgroup is most likely ancestral, as 7 out of the top 10 best scoring BLASTp hits to the murI gene of T. lettingae (YP_001470155) are from members of the Thermotogae phylum. Consequently, the most likely explanation for the observed distribution of genes is that murI was lost in
the ancestor of the *T. maritima* subgroup (Fig. 2.7). With the exception of *murI* and *alr*, all other peptidoglycan genes are present in *T. maritima* (appendix Table VI.11).

*Thermotogae* require D-lysine in their peptidoglycan (Boniface et al., 2009). Ornithine is structurally similar to lysine, but has a side chain that is one carbon shorter; therefore ornithine racemase was used for BLASTP searches to identify potential lysine racemases. All *Thermotogae* members were found to possess this racemase gene (*lyr*), despite it being uncommon in other taxa. This result is highly suggestive that the gene (locus tag TM1597 in *T. maritima*) does encode a lysine racemase. In light of the fact that its copy number does not change between *Thermotogae* species with and without *alr* and that the *lyr* gene is absent in *P. ubique* and *Wolbachia*, it is unlikely that *lyr* took over the roles of *alr* and *murI*.

![Figure 2.7. Inferred history of metC, alr and murI in the phylum Thermotogae.](image)

As noted in section 2.1.3, there are also other bacterial enzymes known to possess amino acid racemase activities. Several of these proteins used as queries were homologues of other proteins with different functions, some of which are present in *E. coli*. Previous work in the lab where a rescue experiment in which the *E. coli* D-alanine auxotroph, MB2795, was transformed with a pooled library of all *E. coli* genes (the ASKA collection; Kitagawa et al. 2005) found only *alr*, *dadX* and *metC* as rescuers, whereas other genes, including homologues of genes encoding racemases, were not found (Soo, 2012). For example *malY*, homologue of *metC*, rescues the *metC* knockout, *E. coli* JW2975, but not
the \textit{alr} double knockout (Patrick \textit{et al.}, 2007, Soo, 2012). The phylogenetic conservation of promiscuous activities has not been investigated directly, so it cannot be unambiguously concluded that paralogues in other species behave similarly. A homologue of an \textit{E. coli} gene with the same function in different organisms is likely to share more features, such as the promiscuous activities of the encoded enzyme, than compared to more distant homologues with different functions. This applies to several genes present in \textit{T. maritima}.

Diaminopimelate epimerase (\textit{dapF}-encoded) converts \textit{LL}-diaminopimelate to \textit{meso}-diaminopimelate, an intermediate in the most common bacterial lysine biosynthetic pathway (Usha \textit{et al.}, 2009). In \textit{E. coli} \textit{dapF} was not found to be able to rescue the \textit{D}-alanine auxotrophic strain. Most \textit{Thermotoga} have \textit{dapF} and in \textit{T. maritima} (locus tag: TM1522) it is an operon with other lysine biosynthetic genes. Moreover, the \textit{dapF} gene appears to be vertically inherited in the \textit{Thermotogae} (as well as the \textit{Pelagibacterales} and the \textit{Rickettsiales}) and is present in a single copy throughout the close relatives of these taxa, regardless of the presence or absence of \textit{alr}. Consequently, despite being an amino acid epimerase it is unlikely to racemise alanine in \textit{T. maritima}.

A gene of unknown function, \textit{yhfX}, is annotated as encoding a “predicted amino acid racemase” in \textit{E. coli} (Keseler \textit{et al.}, 2009), but was not found as a rescuer of \textit{E. coli} MB2795. A BLASTP search for \textit{yhfX} revealed the presence of a homologue in all \textit{Thermotogae} species (TM1731 in \textit{T. maritima}) that is more closely related to the \textit{E. coli} gene \textit{yggS}, which is not a racemase.

Human serine racemase is a serine dehydratase that has gained racemase activity and is a homologue of threonine dehydratase (Smith \textit{et al.}, 2010), an enzyme involved in isoleucine biosynthesis (Gallagher \textit{et al.}, 1998). Whereas threonine dehydratase (\textit{ilvA}-encoded) is present in the \textit{Thermotoga} genus (TM0356), serine dehydratase/racemase is not. Furthermore, unlike \textit{E. coli} with \textit{ilvA} and its homologue \textit{tdcB} (Guillouet \textit{et al.}, 1999), \textit{T. maritima} has only one gene encoding threonine dehydratase and a full pathway for isoleucine biosynthesis to which the enzyme belongs.

\textit{D}-amino acid transaminase (\textit{dat}-encoded) acts upon \textit{D}-alanine and \textit{D}-glutamate in some \textit{Firmicutes} (Pucci \textit{et al.}, 1995) and is a homologue of the branched-chain amino acid transaminase (\textit{ilvE}-encoded). The latter (TM0831) is present in \textit{T. maritima}, but not the former.
As an ancillary search to find other genes that could act as alanine racemases, the Phylogenetic Profiler on the JGI IMG server (Markowitz et al., 2012) was used to find cases of newly acquired genes. 31 genes are present in the genomes of *Thermotoga* species that are absent in the other members of the *Thermotoga* phylum, and therefore could encode candidates for newly-acquired alanine racemase functionality under the assumption that the gene that displaced *alr* is horizontally acquired. The majority of the encoded proteins are hypothetical proteins and only five are enzymes, namely threonine dehydratase (*ilvA*-encoded) and four esterases (encoded by TM0077, TM0195, TM0356, TM1062 and TM1281). The *ilvA* gene is flagged here, because the other members of the *Thermotoga* do not possess a known pathway for isoleucine biosynthesis.

Therefore, this bioinformatics analysis suggests that *metC* is the best candidate to encode a physiologically relevant alanine racemases in *Thermotoga maritima*, while the *lyr*, *dapF*, *ilvA*, *ilvE*, and *yggS* genes are less likely candidates. On the other hand, the candidate that was considered most logical, *a priori*, was *murI* (encoding glutamate racemase) and this was found to be absent in *T. maritima*, like *alr* itself.

### 2.2.5 Investigation of *alr* alternatives in *P. ubique* (in silico)

The free-living bacterium *P. ubique* is a member of the order *Pelagibacterales* in the *Alphaproteobacteria*. Similar to *T. maritima* (§ 2.2.3), *P. ubique* lacks both *alr* and *murI*. Instead it possesses *metC* and the aspartate racemase gene, *racX* (SAR11_1277). For methionine biosynthesis, it possesses *metX* (SAR11_0217) and *metY* (SAR11_1030; Fig. 2.2), in addition to *metC* (SAR11_0829). Members of the order *Pelagibacterales* that do not belong to the genus *Pelagibacter* (also known as subgroup IA, Grote et al., 2012a) possess instead *alr* and *metZ*, a homologue of *metC*.

Similar to *T. maritima*, *P. ubique* possess *dapF* (SAR11_0257), *yggS* (SAR11_0361), *ilvE* (SAR11_0086, but not *dat*). *P. ubique* also lacked *lyr* and *ilvA*.

In the case of newly acquired genes, 147 genes are present in genomes of *Pelagibacter* species and absent in the other members of the *Pelagibacterales*. Five of these encode PLP-dependent enzymes, including *racX* and *metC*. Of these, only *metC* is present in the other two species investigated.
2.2.6 Investigation of \textit{alr} alternatives in \textit{Wolbachia (in silico)}

The obligate endosymbiont \textit{Wolbachia} is a member of the family \textit{Anaplasmataceae} in the order \textit{Rickettsiales} in the \textit{Alphaproteobacteria}. \textit{Wolbachia} lacks \textit{alr} and \textit{murI}, but possess all other genes associated with peptidoglycan biosynthesis, including genes involved in \textit{meso}-diaminopimelate biosynthesis, such as \textit{dapF} (WD1208). The absence of the \textit{murI} gene is shared by all members of the \textit{Rickettsiales} and \textit{Pelagibacterales} (§ 2.2.10) and another gene must have taken over its role. The \textit{dapF} gene is, however, not present in all members of the \textit{Rickettsiales} with peptidoglycan (the exception being \textit{Midichloria mitochondrii}), which further weakens the possibility of \textit{dapF} as an \textit{alr} alternative. Apart from \textit{metC} (WD0925), \textit{Wolbachia} has no methionine biosynthetic genes (data not shown). Furthermore, 85 genes have been acquired at the time of the loss of \textit{alr}, one of these is \textit{metC}, while the other genes that encode enzymes are involved in nucleotide biosynthesis. Therefore, \textit{metC} is the most likely candidate for encoding the physiological alanine racemase.

2.2.7 Operon order

Operons often are composed of functionally related genes (Tamames \textit{et al.}, 1997), consequently the operons of the \textit{metC} genes were investigated. In \textit{T. maritima} and \textit{P. ubique} the other genes in the operon with \textit{metC} do not appear to have a functional relationship, while in \textit{Wolbachia} the gene may be in an operon with other cell division genes. Operons can be predicted \textit{in silico} or they can be determined experimentally via their transcripts. Here, \textit{in silico} annotations from the BioCyc server were used for convenience for \textit{Wolbachia} and \textit{P. ubique} (Caspi \textit{et al.}, 2010); while data from transcriptome studies were used for \textit{T. maritima} (Latif \textit{et al.}, 2013) and \textit{E. coli} (Keseler \textit{et al.}, 2009). In \textit{Wolbachia}, \textit{metC} is predicted to be in an operon with \textit{murE} (WD0924) and \textit{bolA} (WD0926). The former encodes a peptidoglycan ligase, while the second encodes a factor that determines cell shape, both therefore involved with peptidoglycan biosynthesis. This provides evidence that \textit{metC} may be involved in peptidoglycan biosynthesis in \textit{Wolbachia}. 
2.2.8 MetC as a Cystathionine $\gamma$-synthase

Earlier results (Fig. 2.7) confirmed that plasmid-encoded copies of the *Wolbachia*, *T. maritima* and *P. ubique* metC genes were able to rescue an *E. coli* ΔmetC strain, as expected. However, the metC gene in *T. maritima* is actually annotated as metB (encoding cystathionine $\gamma$-synthase) in NCBI, accession number NP_229075. This is due to lack of a distinct clustering between metB genes and metC genes in terms of homology (vide infra, §2.2.9).

Consequently, the *Wolbachia*, *T. maritima*, *P. ubique* and *E. coli* metC genes were tested for their ability to rescue an *E. coli* ΔmetB strain (JW3973) (methods §2.4.4). Whereas a plasmid-encoded copy of *E. coli* metB was able to rescue this strain, the three metC genes were not; that is, no colony growth was observed within two weeks of incubation on M9 medium with 0.4% glucose (Fig. 2.8).

![Figure 2.8. Inability of the metC orthologues to rescue an E. coli ΔmetB strain on M9 medium.](image)

2.2.9 Evolutionary history of metC (in silico)

The phylogeny of metC was investigated due to the problematic discrimination of metC, metB and metZ (§2.2.8) and in order to glean the relationship between the metC genes from *E. coli*, *P. ubique*, *Wolbachia* and *T. maritima* (§2.4.5).

A protein tree was inferred with every sequence homologue of metC, including metY as an outgroup, in a manually selected group of taxonomically diverse organisms (appendix Fig. III.11). This revealed that the sequences annotated as metB are interspersed with other sequences, while the group of sequences annotated as metY, those annotated as metZ, and a
subset of sequences annotated as \textit{metC} form clades with high support. However, not all sequences annotated as \textit{metC} cluster together. The group with high support includes \textit{E. coli} \textit{metC} and \textit{S. cerevisiae IRC7}, while a smaller \textit{metC} cluster was also present and includes \textit{S. cerevisiae STR3}.

To determine the relationship between various \textit{metC} homologues, a tree was inferred from a dataset obtained from a shrunken set of \textsc{blastp} search results (Fig. 2.9). The first division on the \textit{metC} tree is between a group of \textit{metC} sequences mainly from \textit{Firmicutes}, \textit{Thermotoga}, \textit{Rhizobiales} and a family of \textit{Bacteroidetes}, and another group with several diverse sequences, including the enterobacterial \textit{metC} genes. At the base of the latter group are the sequences from \textit{Wolbachia} and \textit{Anaplasma}, while within the main group are sequences from various organisms, including \textit{P. ubique} and members of the \textit{Enterobacteriales}. Consequently, out of the three \textit{metC} genes investigated in this thesis, the \textit{metC} from \textit{P. ubique} is evolutionarily closest to the \textit{E. coli} homologue, while the \textit{T. maritima} is the furthest.

\textbf{Figure 2.9.} Condensed phylogenetic relation between \textit{metC} genes. The groups of \textit{metC} genes originating predominantly from related species are collapsed and annotated as the taxa, despite the presence of some \textit{metC} genes from distant species. The expanded version of this tree is present in appendix III (§ III.11).
2.2.10 Evolutionary history of the Alphaproteobacteria in terms of metC, murI and alr (in silico)

A consequence of the previous section is that the Wolbachia and P. ubique metC genes are not from the same clade. In contrast, the alr gene tree (Appendix III) shows that the alr genes from the Alphaproteobacteria, the class which Wolbachia and P. ubique belong to, are vertically inherited. Therefore, the most parsimonious solution is that two different metC genes have displaced alr, with one displacement event occurring in the common ancestor of the Anaplasmataceae (the family in which Wolbachia is classified) and the other event occurring in the ancestor of Pelagibacter species (Fig. 2.10).

Figure 2.10. Inferred history of metC, alr and murI in the Alphaproteobacteria. The murI gene was lost in a common ancestor of the Rickettsiidae and Caulobacteridae (yellow cross). New (distantly-related) copies were acquired in the Caulobacteridae and in the Pelagibacter lineage (yellow circles). In this lineage metC was also acquired (green circle), while the alr gene was lost (blue cross). This resulted in a lineage with metC and racX, but no alr (cyan branch). The alr gene was also displaced by a metC gene in the common ancestor of the Anaplasmataceae (blue cross and green circle). This metC gene was lost in some lineages, along with other peptidoglycan genes (green crosses). Consequently, in the Rickettsiales, there are some lineages with no peptidoglycan (red branches), some with peptidoglycan and alr (blue branches), and others with peptidoglycan and metC (green branches). This cartoon is based on the cladograms in Fig. 3.3.
As mentioned (§ 2.2.6), the murI gene is absent in the Rickettsiales and in the Pelagibacterales, but is present in the other two clades of the Alphaproteobacteria. Furthermore, the murI gene (encoding glutamate racemase) does not appear to be conserved vertically throughout the Alphaproteobacteria as the murI genes present in the most basal clade of Alphaproteobacteria (Magnetococcus marinus) and in other Proteobacteria are not closely related to the murI genes from the other clade of Alphaproteobacteria (Caulobacteridae). This suggests that murI was absent in the last common ancestor of the Rickettsidae and Caulobacteridae.

2.2.11 Filling pathway holes in methionine biosynthesis

In E. coli the substrate for MetC is produced by its homologue MetB (§ 2.1.2). However, in T. maritima and P. ubique, metC is found without a metB. To determine whether this is a common phenomenon the co-presence of the various methionine biosynthetic genes was investigated.

To discern between metC, metB and metZ, the NCBI COG annotation used for the search (§ 2.2.1) could not be used. Instead the InterPro annotation was used (§ 2.4.6), because the former are based on clusters of sequenced homology, but the latter is based on protein motifs. The analysis was conducted on 1666 annotated species obtained from the InterPro server on the 12th October 2012.

The first step in methionine biosynthesis is the formation of a homoserine ester via either metX or metA. A priori these are fully mutually redundant, a conclusion with agrees with the empirical data where they negatively correlate (50% of species with metX are without metA, 37% with metA are without metX, 4% have with both metX and metA). Antithetically, the two alternative genes which encode the enzymes for the direct sulfurylation step, metZ and metY, do not negatively correlate (60% with metY are without metZ, 6% with metZ are without metY and 16% have both metY and metZ; Fig. 2.11) and instead metZ is found predominantly with metY, which indicates that the current picture is incomplete, for example metY may play an additional role.

In terms of whole pathways, 52% of strains have simply metA or metX with metY, while only 4% have the E. coli-like pathway. However, whereas metB is not found without
metC, metC can be found without metB in 13% of the species, although in half of these cases metC is with metY (9% of species). This indicates that metB is found only with metC, but metC can be found with either metB or metY.

Figure 2.11. Co-presence of metB, metC, metZ and metY. The percentages represent fractions of the bacterial species annotated, which despite having biases towards clades rich in pathogenic bacteria, e.g. Enterobacteriaceae, was used as a proxy for general diversity. This graph reveals that metC is more often found without metB (10%) than with (9%) even though both are required for methionine biosynthesis in the current model. Instead, metC is frequently found with metY and without metB (10%). This lends support to the idea that some MetY enzymes may be able to synthesise cystathionine.
2.3 Discussion

2.3.1 Gene displacement

Previous studies have revealed some of the abundance of promiscuous activities present in the *E. coli* translated genome. These side activities represent a pool that may be advantageous under certain selective pressures. In this chapter, the question of whether this potential has become actuality for a given gene in any of the sequenced bacterial species was addressed. The model chosen was the displacement of *alr* by *metC*. Analysis of sequenced genomes revealed that *P. ubique*, *Wolbachia* and *T. maritima* possess peptidoglycan genes, but lack the *alr* gene. The *metC* genes of these species were able to rescue *E. coli* strains that lacked either *metC* or the two alanine racemase genes, *alr* and *dadX* (Fig. 2.6). These results showed that the three genes have both activities, but do not indicate whether they are main activities or promiscuous activities.

Ideally, it would have been preferable to knock out the *metC* genes in *P. ubique*, *Wolbachia* and *T. maritima*, so that their auxotrophies for methionine and/or D-alanine could have been verified experimentally. Unfortunately, there are no tools available for genetic manipulation of these species. A selectable plasmid vector is available for *T. maritima* (Han et al., 2012); however, no gene knockouts have been reported. Consequently, a series of in silico investigations were done to determine:

- The physiological requirement for cystathionine β-elimination in methionine biosynthesis (§ 2.3.2).
- The requirement for alanine racemisation (§ 2.3.3).
- The absence of other genes encoding alanine racemases (§ 2.3.4–5).

2.3.2 MetC as a cystathionine β-lyase

The three *metC* genes tested could rescue *E. coli ΔmetC*, albeit slowly (Fig. 2.6). The speed of colony formation does not necessarily reflect enzyme activity, since it can also depend on how well the foreign gene is expressed in *E. coli* and on the physiological requirements of the host cell. The requirements for D-alanine and methionine differ by roughly three orders of magnitude (240,000–700,000 residues of D-alanine per cell given that there are that many monomers of peptidoglycan/cell and each has one D-alanine
residue (Sundararaj et al., 2004), vs. $4 \cdot 10^7$ residues of methione per cells given that there are $\sim 3,600,000$ protein/cell (Sundararaj et al., 2004), each composed of 360 amino acids on average (Sundararaj et al., 2004) and 2.8% of which are methionine (Lobry and Gautier, 1994)). Nevertheless, despite the slower rescue time, these experiments showed that the three MetC enzymes could catalyse the $\beta$-elimination of cystathionine in vivo. The in vitro kinetic parameters of the three enzymes will be investigated further in Chapter 4.

The canonical role of MetC was established in E. coli. In this bacterium, the substrate of MetC (cystathionine), is produced by the metB-encoded enzyme, cystathionine $\gamma$-synthase (Aitken et al., 2011). However, the analysis in section 2.2.11 demonstrated that this is not the case in all bacteria, with about one in six sequenced species possessing a metC gene without a metB.

The metB gene is not present in the three lineages of interest in this work. Indeed, Wolbachia does not possess any other methionine biosynthetic genes and therefore the cystathionine $\beta$-lyase activity of Wmel MetC does not appear to be physiologically relevant. On the other hand, P. ubique and T. maritima possess complete methionine biosynthetic pathways, due to the presence of metY. This gene encodes acetylhomoserine sulphydrylase, which catalyses the direct formation of homocysteine from acetyl-homoserine, bypassing the need for the enzymes encoded by metB and metC (Fig. 2.2).

At first glance, the lack of metB implies that cystathionine is not produced in P. ubique and T. maritima, which in turn would imply that the cystathionine $\beta$-lyase activities of Pub MetC and Tma MetC are not physiologically relevant. However, two other enzymes are known to be able to catalyse the formation of cystathionine; these are cystathionine $\gamma$-lyase from yeast (CYS3p) (Ono et al., 1993) and acetyl-homoserine thiolyase from B. subtilis (metI-encoded) (Auger et al., 2002). The case of the B. subtilis MetI is particularly informative. It can catalyse the synthesis of homocysteine and cystathionine, by accepting either sulfide or cysteine as the nucleophile that can attack the other substrate, acetyl-homoserine (Auger et al., 2002). This ambiguity towards various thiolated substrates is not uncommon. The acetyl-homoserine thiolyase (metY-encoded) of some other organisms, including Rhodospirillum rubrum, is able to utilise not only sulfide, but also menthanethiol to produce methionine directly (Erb et al., 2012). The presence of metC with metY, as opposed to with metB, is also seen in many different sequenced genomes, and is the most frequent pairing after metC with metB.
The *T. maritima* and *P. ubique* genomes lack *metB*, but both contain *metY* genes. Consequently, there is a high possibility that the MetY enzyme in these species is similar to the *B. subtilis* MetI and able to catalyse the formation of cystathionine, thus providing a physiological route to the substrate for MetC (Fig. 2.12). Testing this hypothesis through auxotroph rescue experiments would not be straightforward, because *E. coli* uses succinyl-homoserine as a key intermediate, while *T. maritima* and *P. ubique* use acetyl-homoserine (Fig. 2.12).

**Figure 2.12.** The two proposed routes to homocysteine biosynthesis in *T. maritima*. In addition to the direct sulfurylation route catalysed by MetY (blue line), it is hypothesised here that a second route may be present (green line), where the acetyl group of acetyl-homoserine is replaced not by hydrogen sulfide, but by an organothiol (cysteine). A similar situation is probably seen in *P. ubique*, except for the presence of *metX* and *metW* instead of *metA*, and *bhmT* (encoding betaine:homocysteine methyltransferase) instead of *metE*.

Conversely, the possibility that MetC is acting as a bifunctional MetB/MetC in these organisms seems highly unlikely. On one hand, phylogenetically there seem to be two clusters of *metC* genes within the cluster of *metB* genes, which would suggest that the activities might overlap. This hypothesis is, however, not supported empirically: the inability of *E. coli*, *T. maritima*, *Wolbachia* and *P. ubique* metC to rescue an *E. coli ΔmetB* strain (Fig. 2.8) indicates that these genes do not encode an enzyme that can synthesise cystathionine from succinyl-homoserine and cysteine. The possibility that the *T. maritima* and *P. ubique* MetC enzymes can synthesise cystathionine from acetyl-homoserine (not succinyl-homoserine) and cysteine is not ruled out, but it is made improbable. It is likely that the requisite γ-elimination and β-elimination are mutually exclusive in these enzymes, especially in light of a recent study that did not find both activities in a library of MetC-MetB chimeras (Manders *et al.*, 2013).
The probable bifunctionality of MetY, in conjunction with MetC, indicates that 
\textit{T. maritima} and \textit{P. ubique} possess two routes to produce methionine, one via free thiols 
and the other via cysteine (Fig. 2.12). If correct, this is a case of distributed robustness, a 
common biochemical strategy where alternative routes confer a distinct advantage under 
certain environments. For example, the pathway to degrade histidine in several 
\textit{Pseudomonas} species splits into a route that is advantageous in low nitrogen environments 
(because it yields free ammonium ions at several steps), and another route for high nitrogen 
environments that requires fewer steps (Gerth \textit{et al.}, 2012).

\subsection*{2.3.3 Peptidoglycan and D-alanine requirement}

The physiological requirement for an alanine racemase is more clear cut: this activity 
is required for the synthesis of peptidoglycan. Whereas archaeal pseudomurein lacks 
D-amino acids, no bacterial peptidoglycan is known to be without them (Schleifer and 
Kandler, 1972, Vollmer \textit{et al.}, 2008).

All three species studied here require peptidoglycan, although some differences 
compared to \textit{E. coli} are present in terms of their peptidoglycan.

\textit{T. maritima} possesses an unusual morphology, where its outer membrane forms a 
sheath, known as a toga (Huber \textit{et al.}, 1986) and its peptidoglycan has been extensively 
studied due to several differences with the majority of bacteria, but it does contain D-
alanine (Boniface \textit{et al.}, 2009).

The peptidoglycan of the comma-shaped \textit{P. ubique} has not been studied directly, but it 
is an alphaproteobacterium (Thrash \textit{et al.}, 2011) and most proteobacteria have a similar 
peptidoglycan structure composed of L-alanine, D-glutamate, meso-diaminopimelate and D-
alanine (Schleifer and Kandler, 1972). The peptidoglycan of a related species, \textit{Rickettsia 
prowazekii}, is known to be composed in that way (Pang and Winkler, 1994). Nevertheless, 
\textit{P. ubique} has a doubling time of 10 h (Rappe \textit{et al.}, 2002, Carini \textit{et al.}, 2012), which would 
imply a slower demand for D-alanine than actively-dividing \textit{E. coli}.

In the \textit{Rickettsiales} the peptidoglycan plays a smaller role than it does in free-living 
organisms. Only \textit{Rickettsia spp.} have a confirmed peptidoglycan cell wall, while \textit{Orientia 
tsutsugamushi} has a confirmed absence of peptidoglycan ((Amano \textit{et al.}, 1987)). The 
members of the \textit{Anaplasmataceae} have not been tested directly, but no cell wall is visible in 
electromicrographs (Garrity, 2005). This is likely to indicate that peptidoglycan is necessary
solely for septum formation during cell division, in a way that is akin to L-form spheroplasts of *E. coli* (Casadesus, 2007). No peptidoglycan has yet been detected by staining in *Wolbachia* and no homologues of peptidoglycan crosslinking enzymes are present in its genome. However, there is evidence that the peptidoglycan precursor, lipid II, is produced at low levels and that it lipid II necessary for the coesion of the inner and outer member thanks to its association with a membrane spanning complex as demonstrated by the disruption caused by the lipid-II–targeting antibiotic fosfomycin (Vollmer *et al.*, 2013). Recently, it was found that *Wolbachia* is sensitive to fosfomycin, an antibiotic that targets peptidoglycan biosynthesis (Vollmer *et al.*, 2013). Therefore, peptidoglycan (and thus, alanine racemase) is required in *Wolbachia*, albeit at lower levels than in *E. coli*. In this regard, the three organisms of interest (*Wolbachia, T. maritima* and *P. ubique*) are all alike.

This commonality has several consequences, which will be discussed in chapter 4, where their biochemistry is explored.

### 2.3.4 MetC as sole alanine racemase

The ability of MetC to racemise alanine does not preclude other encoded enzymes from doing so and nor does it indicate that the activity is physiological, even though D-alanine is required (§ 2.3.3). This is problematic due to fact that the gene knockouts cannot be done with the species being studied here (§ 2.3.1). One way to prove the physiological role is by demonstrating that there are no other genes capable of racemising alanine. Creating ORF collections similar to the ASKa collection (Kitagawa *et al.*, 2005) or even a pooled equivalent amplified using a primer pool generated with modern photolithographic DNA synthesis methods (*e.g.* GeneArt) is prohibitively expensive. Consequently, homologues of known amino acid racemases were searched.

As described in § 2.2.3, each of the three species possess a few genes whose encoded enzymes could potentially have alanine racemase activity, but this is unlikely for a variety of reasons.

Firstly, three of these (*yggS, ilvA* and *ilvE*) are more closely related to genes present in *E. coli*, which cannot racemise alanine, than to known amino acid racemases (§ 2.2.3). Similarly, the gene *dapF*, which encodes a racemase required for lysine biosynthesis, can be
potentially ruled out as an alanine racemase as the *E. coli* homologue cannot rescue the D-alanine auxotrophic strain MB2795.

Secondly, only one of these genes (*dapF*) is present in all three species.

Thirdly, these genes are essential for the species possessing them and are present as conserved single copies throughout the orders to which the three investigated species belong. This would change if a duplication and neofunctionalisation (*i.e.* as alanine racemases) event had occurred.

The enzymes encoded by *T. maritima* *dapF*, *yggS*, *ilvA*, *ilvE* and *lyr* could be tested to rescue the *E. coli* D-alanine auxotrophic strain, but the evidence against their role as alanine racemases does not warrant it.

Therefore, it is likely that *metC* encodes the only enzyme with alanine racemase activity in *T. maritima, Wolbachia* and *P. ubique*.

### 2.3.5 Evolution and streamlining

I have established that the *metC* gene displaced *alr* in *Wolbachia, T. maritima* and *P. ubique*. Furthermore, it has retained its physiological role in methionine biosynthesis in the latter two species: I have therefore found a gene that encodes a multitasking enzyme. Evolutionary pressures that might give rise to this scenario shall now be considered.

In the innovation, amplification and divergence (IAD) model of enzyme evolution described in chapter 1 (§ 1.5), the benefit conferred by a promiscuous activity is increased by the amplification of its gene. This increases the expression levels of the enzyme while also allowing one copy to be maintained and the other to accumulate mutations such that it encodes a new specialist (Bergthorssen *et al.*, 2007).

As mentioned, both the cystathionine elimination and the alanine racemisation activities are most likely to be required at a lower level in the three species compared to *E. coli* (§ 2.3.2–3), a phenomenon that will be further discussed in chapter 4. Consequently, as a multitasking enzymes evolves to catalyse reactions that are not required to be highly productive, a small promiscuous activity in the ancestor with a minor change in expression levels is probably sufficient enough to confer a selective advantage. The change in expression levels in the IAD model arises from gene duplication, whereas here a deregulation of expression may be sufficient. In this case, however, there is a single gene
that is constrained to compromise between the two activities as opposed to two or more genes that diverge unconstrained. Therefore, multitasking enzymes might present a special case in the IAD model, where the second step, amplification, differs by instead being a change in expression levels and the last step, divergence, is absent.

The three species with bifunctional MetC enzymes have unusually small genomes. It is plausible that this observation is related to the selective pressure that resulted in the possession of bifunctional MetC as opposed to dedicated Alr and MetC enzymes.

*P. ubique* possesses the smallest known genome (1,308,759 bp) of any free-living organism (Giovannoni et al., 2005). This is due to its extreme oligotrophic lifestyle in the open ocean, where phosphate, the backbone for DNA, is one of the limiting nutrients (Rappe et al., 2002). Therefore, there is a clear selection pressure to streamline the *P. ubique* genome (Rappe et al., 2002) and the presence of a gene that encodes for two functions is more advantageous than the presence of two separate genes.

*Wolbachia* is an intracellular parasite that infects a variety of arthropods (Lo et al., 2007), and it has lost a large number of genes compared to related species (wMel strain: 1,267,782 bp) (Foster et al., 2005).

*T. maritima* is unusual because over 24% of its genes appear to be of archaeal origin (Nelson et al., 1999). Its genome contains 1,860,725 bp, which is smaller than most bacteria (Nelson et al., 1999). This characteristic is shared with archaea, which seem to be under strong streamlining pressures since descending from an ancestor that was predicted to have had a large genome. It has been suggested that *T. maritima* may have undergone streamlining (Latif et al., 2013). There is some correlation between streamlining and thermophily and it has been hypothesised to be due to a reduced habitat variability (Sabath et al., 2013).

The shared genome-streamlining trend in the three species with multitasking MetC enzymes raises the hypothesis that under such environments, the possession of certain multitasking genes is probably more advantageous that the possession of two separate genes, despite the conflict between the activities and their regulation.
2.3.6 Conclusions

In this chapter, the question of whether a promiscuous activity encoded by an *E. coli* gene has become a main activity in a different species led to the identification of three enzymes (*P. ubiquae*, *Wolbachia* and *T. maritima* MetC) in which this occurred. These enzymes not only catalyse that activity physiologically (alanine racemisation), but also retained the activity that is the sole main activity in *E. coli* MetC (cystathionine elimination), making them multitasking enzymes.

The *alr* gene is present in the overwhelming majority of bacterial taxa. Consequently, the absence of *alr* as seen in *P. ubiquae*, *Wolbachia* and *T. maritima* is a very rare event. The alanine racemase activity of MetC is likely to be physiological as no other genes stood out as potentially capable of racemising alanine. The cystathionine elimination activity was hypothesised to be physiological in *T. maritima* and *P. ubiquae* because several pieces of evidence indicate that the other enzyme involved in the sulfurylation of homoserine (MetY) may be capable of catalysing the preceding step to MetC.

MetC and Alr are not homologous, but encode enzymes that require the same cofactor, PLP, a cofactor with a large catalytic repertoire (Eliot and Kirsch, 2004). Furthermore, the two activities, alanine racemisation and cystathionine elimination, are different types of reactions. In chapter 4 the biochemistry of the reactions of the three enzymes is investigated. One shared feature of *P. ubiquae*, *Wolbachia* and *T. maritima* is that they possess streamlined genomes, which could be the reason why a multitasking MetC is advantageous. Two of these species, *P. ubiquae* and *Wolbachia*, belong to the *Alphaproteobacteria*. In chapter 3 the relationship between the two is explored.
2.4 Methods

In appendix I are listed the general techniques employed (§ I.11), the sources of chemicals and the composition of various solutions (§ I.1.1), the *E. coli* strains (§ I.1.11) and plasmids (§ I.1.111) used, and the sequences of primers (§ I.1.1v).

2.4.1 Identification of genomes of interest

To discover whether there were any organisms without an *alr* homologue but with a *metC* homologue, fully sequenced, annotated prokaryotic genomes from NCBI were parsed with a Perl 5 script with the aid of LWP module (*investigator.pl*, Appendix IV.II).

Specifically, NCBI “protein tables” were used, which summarise information about each protein encoded in a genome, including automated COG annotation. The Perl script downloaded each protein table using a URL extracted from the HTML source code of the NCBI prokaryotic genome page (*http://www.ncbi.nlm.nih.gov/ genomes/lproks.cgi*), as protein tables are not available directly in NCBI FTP. In the case of prokaryotes with two or more chromosomes, e.g. *Agrobacterium tumefaciens* or *Brucella melitensis*, the protein tables were concatenated.

Subsequently, the Perl script parsed each table for the presence or absence of genes annotated as members of the COG0787 (alanine racemase) and COG0626 (cystathionine β-lyase/γ-synthase) clusters and combined this information with the general organismal information from the main prokaryotic genome page in NCBI.

The table that was generated included several instances of multiple strains of a single species; therefore, the data was organised by annotating a “taxonomy tree” (i.e. a polytomic tree based on taxonomic classification) that was generated by the iTOL server (*itol.embl.de*, Letunic and Bork, 2007). The NCBI taxonomy ID number of each species was used as input. The branches of the tree were then colour-coded according to their Alr and MetC compositions using the iTOL labelling feature and a mapping file prepared in Excel 2008.

2.4.2 Verification and gene searches

Once a shortlist of taxa lacking the *alr* gene was generated, this absence was verified by a BLASTP search in the NCBI database (*blast.ncbi.nlm.nih.gov*, Sayers *et al.*, 2009). This approach was used also to ascertain the presence or absence of certain other
genes in the species of interest (Table 2.2). The presence of some homologues and selected operon structures were obtained from the BioCyc server (biocyc.org, Karp et al., 2002). The current taxonomic distribution and nomenclature of certain taxa was ascertained with the LPSN server (www.bacterio.net, Ezeebzy, 1997). The JGI IMG server (img.jgi.doe.gov, Markowitz et al., 2012) was used to find cases of genes that positively or negatively correlated.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Genbank ID</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetC</td>
<td>Cth $\beta$-lyase</td>
<td>NP_417481</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>Alr</td>
<td>Ala racemase</td>
<td>NP_418477</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>MurI</td>
<td>Glu racemase</td>
<td>NP_418402</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>YhfX</td>
<td>Unknown</td>
<td>NP_417426</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>DapF</td>
<td>Dap epimerase</td>
<td>NP_418254</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>PrdF</td>
<td>Pro racemase</td>
<td>CAB71312</td>
<td>Clostridium sticklandii</td>
</tr>
<tr>
<td>Orr</td>
<td>Orn racemase</td>
<td>CBH21408</td>
<td>Clostridium sticklandii</td>
</tr>
<tr>
<td>RacX</td>
<td>Asp racemase</td>
<td>NP_391323</td>
<td>Bacillus subtilis</td>
</tr>
<tr>
<td>SRR</td>
<td>Ser racemase</td>
<td>NP_068766</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MetA</td>
<td>Hser Su-transferase</td>
<td>NP_418437</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>MetX</td>
<td>Hser Ac-transferase</td>
<td>NP_249081</td>
<td>Pseudomonas aeruginosa</td>
</tr>
<tr>
<td>MetB</td>
<td>Cth $\gamma$-synthase</td>
<td>NP_418374</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>MtcB</td>
<td>Cth $\beta$-synthase</td>
<td>ABG54359</td>
<td>Klebsiella pneumoniae</td>
</tr>
<tr>
<td>MtcC</td>
<td>Cth $\gamma$-lyase</td>
<td>ABG54358</td>
<td>Klebsiella pneumoniae</td>
</tr>
<tr>
<td>MetY</td>
<td>Hser (thiol)-lyase</td>
<td>NP_253712</td>
<td>Pseudomonas aeruginosa</td>
</tr>
<tr>
<td>MetZ</td>
<td>Hser (thiol)-lyase</td>
<td>NP_251797</td>
<td>Pseudomonas aeruginosa</td>
</tr>
<tr>
<td>MetI</td>
<td>Cth/hcys synthase</td>
<td>NP_389069</td>
<td>Bacillus subtilis</td>
</tr>
</tbody>
</table>

Table 2.2. GenBank ID of the major genes used as BLASTP queries. The symbols are derived from the gene symbol, even though alternative names may be in use (e.g. CBL for MetC). In the protein name the following abbreviations appear in addition to the standard three-letter abbreviations for amino acids: Ac (acetyl group), cth (cystathionine), dap (diaminopimelate), hcys (homocysteine), hser (homoserine), orn (ornithine) and Su (succinyl group).

2.4.3 Cloning

A codon-optimised wMel metC gene (peptide: NP_966663) with N-terminal tag was designed with the GeneOptimizer online software (GeneArt, a division of Life Sciences, Carlsbad, CA) and ordered via GeneArt (order: 11AAHZEP). The sequence of the gene with tags (green and blue AAs, see chapter 4 for biochemical information) is:
The plasmid (§ I.1.11) was amplified in DH5α (§ I.11.1–I.11.11) and then the gene was cloned into pBAD/myc-his(B), using the standard protocol provided in § I.11.111. The gene was excised from GeneArt’s pMK-RQ vector with NcoI and HindIII. The reaction also contained BspHI to better differentiate wMel metC from the backbone fragments resulting from a NcoI site in the kan gene (1275 bp (metC), 894 bp, 450 bp, 332 bp vs. 1275 bp, 1226 bp and 1063 bp). Simultaneously, pBAD was linearised with NcoI and HindIII. The bands corresponding to the vector backbone (5319 bp) and the metC gene were excised and purified, and subsequently ligated. The sequence verified plasmid was designated pBAD/metCwMel.

A KpnI site was introduced into the sequence encoding the linker between the His6 tag and wMel metC gene in pBAD/metCwMel, by whole circle PCR with Phusion polymerase (New England Biolabs, Ipswich, MA) using tev_kpnI_F and tev_kpnI_R (sequences in appendix § I.1.11; 60°C annealing temperature). The circular amplicon was purified and
electroporated according to the protocols in appendix I. The sequence was verified with pBAD_F and pBAD_R (§ I.1.III).

The \textit{metC} gene from \textit{Thermotoga maritima} MSB8 was amplified by PCR with Phusion polymerase from genomic DNA kindly donated by Prof. Emily Parker (University of Canterbury). The primers were Tma\_metC\_kpnI\_F and Tma\_metC\_xbaI\_R (§ I.I.IV; 62°C annealing temperature), which contained 5’ restriction sites for KpnI and XbaI. As noted in section 2.2.3, the codon-optimised \textit{Pub metC} gene was cloned into pMAL-c5X by Natasha le Roux (2012).

### 2.4.4 Complementation tests

To perform complementation tests \textit{E. coli} JW2975 (\textit{metC::kan}), \textit{E. coli} JW2973 (\textit{metB::kan}), \textit{E. coli} MB2795 (\textit{alr::frt dadX::frt}) and \textit{E. coli} WM335 (\textit{murI’ gltR’}) were used.

Cells were grown overnight in 1 m\ell\ LB, washed in milliQ H\textsubscript{2}O with 1× M9 salts (§ I.I.I) quantified by absorbance at 600 nm (1 AU = 2–3 \cdot 10^8 cells/m\ell; 1 m\ell saturated culture yields approx. 10^9 cells). After which, the cells were diluted to 10^6 cells/m\ell and 100 µ\ell (10^5 cells) were plated on 0.4% glucose M9 plates supplemented with relevant antibiotics (100 µg/m\ell ampicillin or carbenicillin, 34 µg/m\ell chloramphenicol or 30 µg/m\ell kanamycin for \textit{bla}, \textit{cat} or \textit{kan} markers, respectively) and inducer (50 µM IPTG, or 0.02% or 0.002% w/v arabinose). In the case of strain MB2795, LB (rich media) was used instead of M9 media as LB lacks D-alanine.

Cells were incubated at 28°C or at 37°C. Images of the plates were taken either with a flatbed scanner (Canon CanoScan\textsuperscript{®} 3800) or with a bucket of light (Parkinson, 2007) and a point-and-shoot digital camera (Nikon Coolpix\textsuperscript{®} P300).

As a negative control, a strain harbouring either the pBAD plasmid without insert or the pCA24N/gfp plasmid was used, and in certain instances the absence of inducer was used as a further negative control (10–50 colonies escape repression, data not shown). For a positive control, either a strain harbouring a plasmid with the missing \textit{E. coli} gene from the \textit{ASKA} collection (Kitagawa \textit{et al.}, 2005) \textit{(i.e. self-complementation)} was used or supplementation with the appropriate metabolite was used. Specifically, 50 µg/m\ell
methionine was added for JW2973 or JW2975, 50 µg/ml D-alanine for MB2795, 10 µg/ml folate for JW1082 and its derivative and 150 µg/ml D-glutamate for WM335.

2.4.5 Phylogeny of MetC and Alr

Several trees were inferred from datasets with sequences that were assembled with different criteria. Once a set of sequences was assembled, these sequences were aligned with MUSCLE under default settings (Edgar, 2004), trimmed with Gblocks set to eliminate positions that were over half filled with gaps (Talavera and Castresana, 2007) and used to infer a tree with RAxML 7.2.8 (multithread) with 500 bootstrap replicates under a WAG model (a simple model named after its authors) with empiric frequencies (Stamatakis, 2006). The resulting trees were visualised either with the iTOL server or with FigTree v1.1 (tree.bio.ed.ac.uk/software/figtree).

To determine the relationship between the various proteins related to MetC (Fig. 2.3), the 2003 curated COG annotation group present on the NCBI website (www.ncbi.nlm.nih.gov/COG/grace/wiew.cgi?COG0626) was used as a reference by picking each gene annotated as a member of COG0626 in a subset of species that was chosen to maximise phylogenetic diversity. These species were:

- *Escherichia coli*, an enterobacterium (gammaproteobacterium)
- *Salmonella enterica*, member of the sister genus to *Escherichia*
- *Yersinia pestis antiqua*, distant enterobacterium
- *Vibrio cholerae*, member of a sister clade to the *Enterobacteriales*
- *Pseudomonas aeruginosa*, a gammaproteobacterium
- *Legionella pneumophila*, a basal gammaproteobacterium
- *Burkholderia pseudomallei*, a betaproteobacterium
- *Bordetella pertussis*, a betaproteobacterium
- *Pelagibacter ubique*, member of Pelagibacteriales
- alpha proteobacterium HIMB114, member of Pelagibacteriales
- *Wolbachia*, member of the Rickettsiales
- *Anaplasma marginale*, member of the Rickettsiales
- *Rhodobacter spheroides*, close to the Rickettsiales
- *Magnetospirillum magneticum*, member of the Alphaproteobacteria
- *Agrobacterium tumefaciens*, member of the Alphaproteobacteria
- *Xanthobacter autotrophicus*, member of the Alphaproteobacteria
• *Sphingopyxis alaskensis*, member of the *Alphaproteobacteria*
• *Lautropia mirabilis*, an alphaproteobacterium
• *Magnetococcus marinus*, basal alphaproteobacterium
• *Campylobacter jejuni*, a deep-branching proteobacterium
• *Clostridium acetobutylicum*, a firmicute (non-proteobacterium)
• *Bacillus subtilis*, a firmicute
• *Corynebacterium glutamicum*, actinobacterium
• *Fusobacterium nucleatum*, a deep branching bacterium
• *Thermotoga maritima*, a deep branching bacterium
• *Deinococcus radiodurans*, a deep branching bacterium
• *Methanosarcina acetivorans*, an euryarchaeote
• *Thermoplasma acidophilum*, an euryarchaeote
• *Sulfolobus solfataricus*, a crenarchaeote
• *Saccharomyces cerevisiae*, a yeast
• *Arabidopsis thaliana*, a plant

To determine the structure of the MetC protein tree, a series of BLASTp searches were conducted with the following queries:

• *E. coli* MetC: NP_417481, 250 hits (i.e. maximum number of results found)
• *Lautropia mirabilis* MetC: WP_005674411, 50 hits (default setting)
• *P. ubique* MetC: YP_266249, 50 hits
• *T. maritima* MetC: NP_229075, 50 hits
• *Wolbachia* (*Drosophila melanogaster*) MetC: NP_966663, 50 hits

A file was made containing first the MetC sequences from the previous tree and then the results of the searches. This large dataset was aligned and trimmed as described above. As this dataset was far too large for tree inferences and was highly enriched in certain species, the dataset size was reduced in size by removing highly similar sequences. This was done thanks to the -L argument in RAxML, which removes sequences that are closer than a set cut-off, while keeping a single representative of the cluster based on its location in file (hence the spike-in of the initial dataset). The ideal cut-off was empirically found to be 95% identity as it left 206 sequences and kept *Wolbachia* MetC.
To determine the Alr protein tree (used for the analysis in § 2.2.10), two methods of collecting the sequences were used. Firstly, a preliminary BLASTP search was done with *Rickettsia prowazekii* Alr (NP_220488) as the query to see whether the most similar Alr sequences were alphaproteobacterial as would be the case for a highly conserved sequence and to identify any sequences that were not from a member of the *Alphaproteobacteria*. Secondly, the Alr sequences from several species across the *Alphaproteobacteria* were manually chosen. Together these sets of sequence were used to create the dataset for the tree inference, which was done as above, namely the sequences were aligned in Muscle aligned and trimmed in Gblocks before being inferred with RAxML.

To determine the Glr (*murI* encoded) protein tree and specifically if the *murI* from the *Caulobacteridae* and the *murI* from *Magnetococcus marinus* were closely related, the collated BLASTP searches for the glutamate racemase from *Rhodospirillum centenum* (YP_002299498) and *Magnetococcus marinus* (YP_867127) were aligned, trimmed and reduced in size (90% identity cut-off), before being used to infer the phylogeny of the protein using the same methods that were described above.

### 2.4.6 Covariance of *metY* and *metC* across genomes

To analyse the co-occurrence of the various methionine biosynthetic genes, the annotations from the InterPro server were used (*www.ebi.ac.uk/interpro*; Hunter *et al.*, 2012) as the COG annotations did not have sufficient discriminatory power. This was done by downloading the “protein matches” for each family of interest, which were the following:

- *metA*. Homoserine succinyl/acetyltransferase (IPR005697)
- *metX*. Homoserine acetyltransferase (IPR008220)
- The families within the top-level family IPR000277:
  - *metC*. Cystathionine β-lyase, bacterial (IPR006233)
  - *STR3*. Cystathionine β-lyase, eukaryotic (IPR006238)
  - *mdeA*. Methionine γ-lyase (IPR006237)
  - *metY*. O-acetylhomoserine sulphydrylase (IPR006235)
  - *metB*. Cystathionine γ-synthase (IPR011821)
  - *metZ*. O-succinylhomoserine sulphydrylase (IPR006234)
A Perl script cross-tabulated the resulting data to give the number of genes of each family for each organism. This table was further cross-tabulated in Excel 2008 (Microsoft, Redmond, WA) via the pivot table operation.
Chapter 3

The Phylogeny of the

*Alphaproteobacteria*

**Acknowledgements:**

- Dr. J. Cameron Thrash, Dr. Wayne Patrick and Prof. Stephen Giovanonni contributed substantially to this chapter: their contribution is elaborated upon in section 3.0.
3.0 Preface

3.0.1 Reasoning

In this chapter, the ribosomal rRNA tree of *Alphaproteobacteria* was investigated. Two of the three *metC* genes discovered in Chapter 2 were from alphaproteobacterial species (*Wolbachia* and *P. ubique*). Given my interest in understanding the evolution of *metC*, *alr* and *murI* genes, and particularly gene loss during genome streamlining, it became important to understand the phylogenetic relationships within the *Alphaproteobacteria*.

In mid 2011, the phylogenetic positions of several species with sequenced genomes were not fully known (e.g. some members of the *Pelagibacterales*, *Magnetococcus marinus* and *Midichloria mitochondrii*). In light of the fact that 16S ribosomal RNA (rRNA) is the gold standard for phylogeny, I began by constructing a preliminary tree using this marker. Despite the low support, it clustered the *Pelagibacterales* with the *Rickettsiales*, similar to previous studies (Williams *et al.*, 2007, Thrash *et al.*, 2011). This allowed me to map the presence and absence of *metC*, *alr* and *murI* in the sequenced genomes (Fig. 2.10).

However, in late 2011 and early 2012, three publications disagreed with previous studies and cited the strong AT-bias of the *Pelagibacterales* and the *Rickettsiales* as a source of systematic error in previous phylogenies (Brindefalk *et al.*, 2011, Viklund *et al.*, 2011, Rodriguez-Ezpeleta and Embley, 2012). I hypothesised that AT-bias would be weaker in sequences encoding rRNA than it was in entire genomes, because rRNA plays a constrained structural role in the ribosome. To test this hypothesis, we initiated a collaboration with Prof. Stephen J. Giovannoni and his postdoctoral fellow, Dr J. Cameron Thrash (Oregon State University). The result was the work presented in this chapter, which has also been published in *PLoS One*, with the title “New rRNA-based phylogenies of the *Alphaproteobacteria* provide perspective on major groups, mitochondrial ancestry and phylogenetic instability” (Ferla *et al.*, 2013).

3.0.2 Author contribution

During the dataset assembly, Dr Cameron Thrash provided all available 16S and 23S sequences from the IMG (v350) database; I selected and checked them, with his input. I aligned, trimmed and recoded (when appropriate) the datasets, using a series of Perl 5.0
scripts that I wrote. Due to the lack of available computational power in the Institute of Natural Sciences, Massey University, Dr Thrash ran the majority of the inferences. After moving to the University of Otago in September 2012, I ran some of the inferences on the biocperformance machine operated by Darren Hart, the computer resources manager of the Biochemistry Department. To analyse the multitude of trees obtained, I wrote some scripts and analysed the resulting data. Of these analyses, Dr Thrash suggested the comparison of the datasets via an approximately unbiased test. I wrote the paper with substantial input from the other authors.

3.0.3 Copyright notice

All works published in PLoS are published under a Creative Commons Attribution (CC BY) license; therefore, no permission was required from the publisher for reprint of this article. Under the CC BY license, authors retain ownership of the copyright for their article.
3.1 Introduction

3.1.1 Merits of rRNA-based phylogeny

The 16S rRNA gene has traditionally been the most heavily used molecular taxonomy marker because of its universal presence, its vertical inheritance, and its constant and slow evolution. However, due to drawbacks such as a limited number of informative characters, new markers have also been sought (Pace, 2009). Thanks to the quantity of sequenced genomes and higher computational power, many recent studies have used concatenations of large numbers of genes to infer phylogenetic history. However, due to differences in phylogenetic strategies, actual gene histories, and systematic error, gene concatenation studies occasionally disagree with each other and with rRNA gene-based studies (e.g. McInerney et al., 2008).

3.1.2 The diverse members of the Alphaproteobacteria

An example of such incongruence is the phylogeny of the Alphaproteobacteria (Williams et al., 2007, Brindefalk et al., 2011, Georgiades et al., 2011, Thrash et al., 2011, Grote et al., 2012, Rodríguez-Ezpeleta and Embley, 2012, Viklund et al., 2012), which has received considerable attention because it contains many important taxa, including the ancestor of the mitochondria. The Alphaproteobacteria contains members that are pathogens of humans, such as Rickettsia, and livestock, such as Ehrlichia, as well as agriculturally valuable species, such as Rhizobium radiobacter (formerly Agrobacterium tumefaciens), and several highly abundant marine groups such as Roseobacter, SAR116, and SAR11. The commonly accepted alphaproteobacterial orders are the Rhizobiales, the Rhodobacterales, the Caulobacterales, the Parvularculales, the Sphingomonadales, the Rhodospirillales, the Rickettsiales (Garrity, 2005, Lee et al., 2005) and the recently validated Magnetococcales (Bazylinski et al., 2013). Several orders that are represented by a single deep-branching species (namely Kilonieliales (Wiese et al., 2009), Kopriimonadales (Quinn et al., 2012), Kordiimonadales (Kwon et al., 2005), Sneathiellales (Kurahashi et al., 2008) and Rhodothalassiales) have been proposed but the relationships of these orders have not been addressed. The most controversial order is the Rickettsiales, which is composed of the families Rickettsiaceae, Anaplasmataceae, Midichloriaceae (Montagna et al., 2013) and
Holosporaceae, with the membership of the SAR11 clade (Pelagibacterales, Grote et al., 2012) currently under debate (vide infra).

### 3.1.3 Debate over the phylogeny of the Alphaproteobacteria

Several major differences exist between the various phylogenetic studies of the Alphaproteobacteria, especially in taxon and marker selection. In the case of taxa, it is hard to shortlist a subset of taxa that is small enough to be computationally feasible, but large enough to cover the diversity of the group. In the case of marker selection, many studies choose highly conserved housekeeping genes (e.g. Williams et al., 2007), but in some cases subsets of these genes with particular properties are chosen (e.g. Viklund et al., 2012), and the criteria for inclusion vary between studies. These methodological differences sometimes result in the poor choice of markers, such as horizontally transferred genes or those with adaptive properties, genes not universally conserved, or genes inadequately screened against contaminated draft assemblies. Moreover, it has been demonstrated that the most important factor in the correct resolution of a phylogeny is the selection of only genes with a strong phylogenetic signal and without significant incongruence, whereas an increase in the number of genes used does not result in a better resolution (Salichos and Rokas, 2013).

In 2005, Lee et al. inferred the most comprehensive phylogeny of the Alphaproteobacteria at that time, by using the 16S rRNA gene and all existing type strains (Lee et al., 2005). This study became the basis of current classifications. However, it excluded candidate species, such as members of the Pelagibacterales and many members of the Holosporaceae, and many more species have been discovered and sequenced since. As a result, further studies have been conducted. In 2007, Williams et al. used a thoroughly selected set of 104 protein-encoding genes and found that Candidatus Pelagibacter ubique (P. ubique) was basal in the Rickettsiales and that mitochondria were sister to the Rickettsiaceae and Anaplasmataceae (Williams et al., 2007). Unfortunately, the study predated the sequencing of Magnetococcus marinus (formerly Magnetococcus sp. MC-1, Schübbe et al., 2009, Bazylnski et al., 2013) and Odyssella thessalonicensis (Georgiades et al., 2011), so the clades of these two species were absent from the tree.

In 2011 and 2012, several studies on the phylogenetic placement of SAR11 and mitochondria in the Alphaproteobacteria were published near-simultaneously, with conflicting conclusions. Thrash et al. used a variable number of conserved genes and
included several newly sequenced members of the “Pelagibacterales” in their analysis (Thrash et al., 2011). They found the same topology for the Alphaproteobacteria as that obtained previously (Williams et al., 2007), but with evidence for Pelagibacterales and mitochondria as sister groups. Viklund et al. raised concerns of AT-driven artefacts by finding that the trees inferred from concatenations of proteins with high GC bias favoured the Pelagibacterales as a sister clade to the Rickettsiales, whereas those from less biased proteins favoured the Pelagibacterales as a sister clade to the group of Rhizobiales, Caulobacterales and Rhodobacterales (Viklund et al., 2012). However, when the species sampling was increased or when maximum likelihood was used instead of Bayesian inference, the resulting trees supported the membership of Pelagibacterales within the clade Rickettsiales. In that study, mitochondria, Odyssella thessalonicensis and Magnetococcus marinus were omitted. On the other hand, Georgiades et al. found that the Pelagibacterales and the mitochondria formed a sister clade to the Rickettsiales, whereas Odyssella thessalonicensis was found to be basal to the clade composed of the remaining alphaproteobacterial orders (Georgiades et al., 2011). Rodríguez-Ezpeleta and Embley used a variety of different approaches and found the Pelagibacterales–Rickettsiales topology with several methods, including RY-recoding in an attempt to account for GC bias, but a different topology was concluded to be correct (Rodríguez-Ezpeleta and Embley, 2012). This study also omitted Odyssella thessalonicensis and Magnetococcus marinus, but included mitochondria, which clustered in different locations depending on the methodology.

### 3.1.4 Chapter aims

The key nodes in the alphaproteobacterial tree, such as the branch leading to modern mitochondria, are very ancient (dating to >2 billion years ago; Brocks et al., 1999). We chose to revisit the debate over alphaproteobacterial phylogeny using rRNA genes. Being universally conserved and under strong structural and functional constraints, we assert that the rRNA genes are ideal for shedding light on the relationships between the major groups. The 16S rRNA gene remains the gold standard for microbial taxonomy and current ecological studies depend on classifying organisms based on this marker. Furthermore, the large quantity of 23S sequences now available allowed us to use concatenated 16S and 23S sequences, to improve on the only limitation of the 16S (i.e. limited characters) and to provide a better signal in ascertaining problematic inner nodes (Konstantinidis and Tiedje, 2005). While rRNA genes are problematic with regards to long
branch attraction artefacts in Eukaryotes (Philippe and Germot, 2000), we also show that
they drastically reduce GC content bias in Bacteria, and thus they allowed us to substantially
alleviate this potential source of systematic error. Our results do not support grouping the
Holosporaceae family with the Rickettsiales, and do support the hypothesis that the
Pelagibacterales is a sister group to the Rickettsiales, in a new subclass (Rickettsidae subcl.
 nov.) that also includes the mitochondria.
3.2 Results

3.2.1 GC content of rRNA genes is comparatively unbiased

Several alphaproteobacterial groups have extremely AT-rich genomes. This has led to speculation that the topology in which the *Pelagibacterales* cluster with the *Rickettsiales* may be an artefact, due to a spurious attraction of AT-rich taxa (Rodríguez-Ezpeleta and Embley, 2012). We hypothesized that there may be less freedom for GC content to change in rRNA genes, compared with protein-coding genes, because rRNA is under structural constraints. To test this hypothesis, we analyzed the GC content of genomes, compared to rRNA genes, for single members of each species that was available in the IMG v350 database. We found that the GC content of rRNA genes is considerably less varied than that of the corresponding genomes (Fig. 3.1A). Linear regression analyses for the alphaproteobacterial groups gave a slope of 0.15 for the 16S sequences \( (R^2 = 0.63) \) and a slope of 0.20 for the 23S sequences \( (R^2 = 0.72) \), demonstrating that there is 5- to 6-fold less variation in rRNA gene GC content than in genomic GC content. There was even less bias after the highly variable sites were removed from our alignments (Fig. 3.1B and 3.1C). These results validated our decision to use rRNA gene sequences for building trees that do not suffer from artefactual attraction of AT-rich taxa. However, the trend did not hold for mitochondrial LSU and SSU rRNA gene sequences. Linear regression yielded slopes of 0.69 \( (R^2 = 0.82) \) and 0.71 \( (R^2 = 0.79) \) for the 16S and 23S sequences in Fig. 3.1A, respectively. Therefore, the mitochondrial sequences were treated with caution and topologies were inferred both with and without them.
Figure 3.1. Relationship of rRNA gene vs. genomic GC content for Alphaproteobacteria and mitochondria. The rRNA gene GC content was calculated using a Perl script, while the genomic GC content was taken from the IMG database. The “other orders” group includes the Caulobacterales, Sphingomonadales, Rhizobiales, Rhodobacterales and Parvularculales. A) The GC content was calculated from the entirety of the rRNA sequences. B) The GC content was calculated from the rRNA sequences after alignment and removal of gap-rich regions. C) The GC content was calculated from the rRNA sequences after alignment and removal of gap-rich regions and invariant sites.

3.2.2 Tree building strategy

We examined the phylogeny of the Alphaproteobacteria using single members of each species for which both the 16S and 23S rRNA genes were available, together with seven outgroup taxa from adjacent Proteobacteria classes (‘complete’ trees). Using this complete dataset, trees were made from 16S and 23S rRNA genes alone to generate initial topologies, and subsequently with concatenations of the 16S and 23S alignments. We also tested the effect of reducing taxon selection across the tree by subsampling to an idealized membership based on the monophyletic groupings observed in the complete trees (‘trimmed’ dataset). All sets of taxa were aligned with both MUSCLE (Edgar, 2004) and ARB-SINA (Pruesse et al., 2012) to test for incongruence between common rRNA gene alignment methods, and computed with maximum-likelihood (RAxML; Stamatakis, 2006) using both GTR+ and GTRCAT rate models. Each of these datasets was evaluated for compositional bias using RY recoding (Woese et al., 1991), as well as a novel method that we name RYMK recoding, and all were evaluated with and without mitochondrial sequences. Additionally, to understand the effect of sampling bias from the lack of any one group, we employed a clade-specific jackknifing approach in which we removed each of the
monophyletic groups from the dataset and re-calculated the tree. This was done for both alignment methods, both rate categories, and plus or minus mitochondrial sequences. Finally, because many important under-examined alphaproteobacterial taxa have neither genome sequences nor 23S rRNA gene sequences available, we examined their taxonomic affiliation by aligning their 16S sequences with the concatenated 16S–23S bacteria dataset (‘combo’ dataset). This dataset was also tested under both alignment and rate model algorithms. These variations generated a total of 140 trees (summarized in Table 3.1), all of which are available in the Supporting Information of the published paper.

### 3.2.3 The 16S and 23S rRNA gene trees

We began by creating phylogenies of the 16S and 23S rRNA genes separately (the ‘16S’ and ‘23S’ datasets in Table 3.1). These topologies provided a baseline to which we could compare our concatenated rRNA gene trees. The phylogenies were completed with taxa from the complete dataset, both alignment methods, both rate models, and both with and without mitochondrial taxa, for a total of eight trees each of the 16S and 23S rRNA genes. Whereas the 16S trees resolved the *Holosporaceae* within the clade including the *Rhodospirillales* and other orders (with low bootstrap support), the 23S trees did not consistently resolve the monophyly of *Holosporaceae*, nor its location (bootstrap supports are summarized in Fig. 3.2). Conversely, the 16S trees did not consistently resolve the *Pelagibacterales* in a specific location. Neither the 16S nor 23S phylogenies could consistently resolve the *Caulobacterales, Rhizobiales* and *Rhodobacterales*. A 16S tree that was used previously to classify the *Alphaproteobacteria* (Lee et al., 2005) had many similar results, but also some differences. For example, the *Holosporaceae* did not resolve monophyletically and was basal to the *Rickettsiales* (the *Pelagibacterales* were absent), whereas here the *Holosporaceae* is close to the *Rhodospirillales*.

**Figure 3.2 (pages after next).** Summary of bootstrap supports of 16S and 23S only trees. The values in the boxes are the bootstrap support percentages for the given bipartitions from each of the eight inferences for each set (with or without mitochondria, aligned with ARB-SINA or MUSCLE, GTRΓ or GTRCAT model). In the cases where the final tree did not agree with the proposed topology on the location of a clade the bootstrap support for the proposed bipartition is represented in red, despite being absent in the final tree.
<table>
<thead>
<tr>
<th>Dataset name‡</th>
<th># Taxa</th>
<th># Chars (MUSCLE/ ARB)</th>
<th>Coding</th>
<th>Average support value ± SD (%)‡</th>
<th>Average RF metric within the set ± SD‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S‡ + mito</td>
<td>190</td>
<td>1371/1078</td>
<td>regular</td>
<td>67 ± 1.7</td>
<td>57.3 ± 7.5</td>
</tr>
<tr>
<td>23S‡ + mito</td>
<td>190</td>
<td>2621/2159</td>
<td>regular</td>
<td>80.5 ± 0.6</td>
<td>30.0 ± 7.0</td>
</tr>
<tr>
<td>complete* + mito</td>
<td>190</td>
<td>3992/3237</td>
<td>regular</td>
<td>82.7 ± 0.5</td>
<td>26.3 ± 4.8</td>
</tr>
<tr>
<td>complete* + mito</td>
<td>190</td>
<td>7984/6474</td>
<td>RYMK</td>
<td>85.7 ± 0.7</td>
<td>25.5 ± 4.1</td>
</tr>
<tr>
<td>complete* + mito</td>
<td>190</td>
<td>3992/3237</td>
<td>RY</td>
<td>72.5 ± 0.6</td>
<td>48 ± 8.9</td>
</tr>
<tr>
<td>complete* + mito</td>
<td>190</td>
<td>3992/3237</td>
<td>MK</td>
<td>79.2 ± 0.8</td>
<td>31.8 ± 4.7</td>
</tr>
<tr>
<td>16S – mito</td>
<td>166</td>
<td>1412/1204</td>
<td>regular</td>
<td>69 ± 0</td>
<td>28.5 ± 5.0</td>
</tr>
<tr>
<td>23S – mito</td>
<td>166</td>
<td>2661/2246</td>
<td>regular</td>
<td>83 ± 0</td>
<td>12.8 ± 2.1</td>
</tr>
<tr>
<td>complete* – mito</td>
<td>166</td>
<td>4073/3473</td>
<td>regular</td>
<td>83.2 ± 0.5</td>
<td>15.6 ± 2.3</td>
</tr>
<tr>
<td>complete* – mito</td>
<td>166</td>
<td>8146/6946</td>
<td>RYMK</td>
<td>83 ± 0.4</td>
<td>15.8 ± 2.3</td>
</tr>
<tr>
<td>complete* – mito</td>
<td>166</td>
<td>4073/3473</td>
<td>MK</td>
<td>80.0 ± 0.7</td>
<td>22.3 ± 3.4</td>
</tr>
<tr>
<td>complete mtDel</td>
<td>166</td>
<td>3992/3237</td>
<td>MK</td>
<td>81.2 ± 0.7</td>
<td>21.6 ± 4.5</td>
</tr>
<tr>
<td>trimmed* + mito</td>
<td>111</td>
<td>3990/3335</td>
<td>regular</td>
<td>77.2 ± 0.5</td>
<td>22.8 ± 4.9</td>
</tr>
<tr>
<td>trimmed* + mito</td>
<td>111</td>
<td>7980/6670</td>
<td>RYMK</td>
<td>79 ± 0.9</td>
<td>27.8 ± 4.8</td>
</tr>
<tr>
<td>trimmed* + mito</td>
<td>111</td>
<td>3990/3335</td>
<td>MK</td>
<td>71.5 ± 0.9</td>
<td>17.2 ± 3.9</td>
</tr>
<tr>
<td>trimmed* – mito</td>
<td>87</td>
<td>4121/3542</td>
<td>regular</td>
<td>76.7 ± 0.5</td>
<td>18.3 ± 2.9</td>
</tr>
<tr>
<td>trimmed* – mito</td>
<td>87</td>
<td>8242/7084</td>
<td>RYMK</td>
<td>74.2 ± 0.2</td>
<td>20.8 ± 2.6</td>
</tr>
<tr>
<td>trimmed* – mito</td>
<td>87</td>
<td>4121/3542</td>
<td>MK</td>
<td>67 ± 2.3</td>
<td>19.5 ± 4</td>
</tr>
<tr>
<td>trimmed* – mito</td>
<td>87</td>
<td>4121/3542</td>
<td>MK</td>
<td>72.7 ± 0.7</td>
<td>12.3 ± 2.4</td>
</tr>
<tr>
<td>exoricketless* + mito</td>
<td>171</td>
<td>4000/3166</td>
<td>regular</td>
<td>82.5 ± 1.2</td>
<td>27.5 ± 6.9</td>
</tr>
<tr>
<td>exoricketless* – mito</td>
<td>147</td>
<td>4097/3488</td>
<td>regular</td>
<td>83.5 ± 0.3</td>
<td>14.8 ± 3.3</td>
</tr>
<tr>
<td>hololess* + mito</td>
<td>188</td>
<td>4009/3281</td>
<td>regular</td>
<td>83.7 ± 0.5</td>
<td>24.1 ± 5.3</td>
</tr>
<tr>
<td>hololess* – mito</td>
<td>164</td>
<td>4069/3463</td>
<td>regular</td>
<td>83.5 ± 0.3</td>
<td>14.1 ± 1.2</td>
</tr>
<tr>
<td>magnetoless* + mito</td>
<td>189</td>
<td>4015/3275</td>
<td>regular</td>
<td>83.0 ± 0.6</td>
<td>27.2 ± 6.6</td>
</tr>
<tr>
<td>magnetoless* – mito</td>
<td>165</td>
<td>4080/3448</td>
<td>regular</td>
<td>83.5 ± 0.3</td>
<td>19.5 ± 3.9</td>
</tr>
<tr>
<td>pelagiless* + mito</td>
<td>182</td>
<td>3452/3999</td>
<td>regular</td>
<td>81.0 ± 0.9</td>
<td>33.0 ± 6.7</td>
</tr>
<tr>
<td>pelagiless* – mito</td>
<td>158</td>
<td>3275/4048</td>
<td>regular</td>
<td>82.7 ± 0.7</td>
<td>20.5 ± 3.8</td>
</tr>
<tr>
<td>rhodoless* + mito</td>
<td>174</td>
<td>3984/3281</td>
<td>regular</td>
<td>82.5 ± 0.5</td>
<td>24.8 ± 5.3</td>
</tr>
<tr>
<td>rhodoless* – mito</td>
<td>150</td>
<td>4114/3466</td>
<td>regular</td>
<td>83.2 ± 0.5</td>
<td>18.8 ± 2.9</td>
</tr>
<tr>
<td>rickettsiales* + mito</td>
<td>149</td>
<td>3282/3980</td>
<td>regular</td>
<td>82.0 ± 0.9</td>
<td>25.6 ± 5.1</td>
</tr>
<tr>
<td>rickettsiales* – mito</td>
<td>149</td>
<td>3472/4096</td>
<td>regular</td>
<td>82.2 ± 0.5</td>
<td>17.5 ± 2.6</td>
</tr>
<tr>
<td>combo* + mito</td>
<td>277</td>
<td>3992/3237</td>
<td>regular</td>
<td>49.3 ± 12.6</td>
<td>74.2 ± 0.5</td>
</tr>
<tr>
<td>combo* – mito</td>
<td>253</td>
<td>4073/3473</td>
<td>regular</td>
<td>37.8 ± 8.1</td>
<td>76.0 ± 0.4</td>
</tr>
</tbody>
</table>
Figure 3.3 (Facing pages). Tree inferred with the ARB-SINA aligned complete dataset under a GTRΓ model. Bootstrap values (n = 1000) are indicated at the nodes. Red arrows indicate how a taxon or clade differs in the other regularly coded trees, with values in square brackets indicating in how many trees this is seen. If there are one or more differences within a family, this is indicated after the name of the family. The leaves of the phylogram are collapsed into taxonomic families and into the host phyla for mitochondria. The internal topology of the Rhodospirillales order is not the same in all primary trees, therefore it has been expanded to show all leaves (inset).
3.2.4 The complete concatenated trees

To increase resolution we concatenated the 16S and 23S rRNA genes, and explored topological stability between different alignment (ARB-SINA, MUSCLE) and rate (GRTT, GTRCAT) methods, as well as the variable inclusion of mitochondrial sequences. The complete trees included a representative of each defined Alphaproteobacteria species present in IMG v350 and NCBI Genbank, for which 16S and 23S rRNA gene sequences were available. Specifically, only a single strain was picked for each species, whereas in the cases of unclassified strain, all were chosen. In this set of eight trees, the Alphaproteobacteria is divided into three primary clades (representative final tree in Fig. 3.3; bootstrap summary in Fig. 3.4A). The earliest diverging clade is the Magnetococcales, represented by Magnetococcus marinus.

One of the subsequent clades has the Pelagibacterales subtending the Anaplasmataceae, Midichloria mitochondrii, the Rickettsiaceae and the mitochondria (if present). The other clade has the Holosporaceae at the base, the Rhodospirillales as the next clade branching out, followed by the Sphingomonadales, then the remaining orders. The Holosporaceae is represented by Odysella thessalonicensis and Caedibacter caryophilus, and is monophyletic in six out of the eight trees, but in two it resolves paraphyletically at the same location.

Figure 3.4 (next pages). Summary of bootstrap supports of ‘complete’ dataset. The values in the boxes are the bootstrap support percentages for the indicated bipartitions, from each of the eight inferences for each set, as described in the legend (i.e. with or without mitochondria, aligned with ARB-SINA or MUSCLE, GTR or GTRCAT rate model). In the cases where the final tree did not agree with the proposed topology on the location of a clade, the bootstrap support for the proposed bipartition is represented in red, despite being absent in the final tree. A. Regular-coded complete datasets. B. RY-recoded complete datasets. C. MK-recoded complete datasets. D. RYMK-recoded complete datasets.
Rhizobiales
Rhodobacterales
Caulobacterales
Sphingomonadales
Rhodospirillaceae
Acetobacteraceae
Holosporaceae
Anaplasmataceae
Rickettsiaceae
Mitochondria
Pelagibacterales
Magnetococcales
outgroup
The *Holosporaceae* family is currently classified as a member of the *Rickettsiales* (Garrity, 2005, Lee et al., 2005), but this topology (with monophyletic *Holosporaceae*) was not seen in our final tree and it is only supported in a very small fraction of the bootstrap trees (2.0%, 2.0%, 0.3%, 0.4%, 0.4%, 6.8%, 4%, 1.3%). The support of the monophyly of the *Rhodobacterales*, *Caulobacterales* and *Rhizobiales* increased greatly when some species were reclassified, namely when *Maricaulis maris* was removed from the *Hyphomonadaceae* (*Caulobacterales*) and when the clade formed by *Labrenzia* and *Roseibium* species was moved from the *Rhodobacteraceae* (*Rhodobacterales*) to the *Rhizobiales*. The support for the monophyly of the *Rhodospirillales* was not high, varying from 56% to 92% in the complete trees, and the monophyly of the *Rhodospirillaceae* had even less support (42–52%).

### 3.2.5 The complete RY-, MK- and RYMK-recoded trees

Accepting the topology formed with the regular-coding complete datasets, we compared the GC content of the crown group to that of the *Rickettsiales–Pelagibacterales* and to that of the mitochondria, separately. As expected based on the data in Fig. 3.1, GC content of the SSU and LSU rRNA genes exhibited much smaller variation than that of the genomes (Fig. 3.5). Nevertheless, there were small, statistically significant differences when the GC contents of the crown group SSU and LSU sequences were compared to the *Rickettsiales–Pelagibacterales* sequences (Wilcoxon rank-sum *p*-values = 2.34 x 10^{-15} and 4.70 x 10^{-15}, for SSU and LSU sequences respectively). This analysis indicated that using the rRNA genes mitigated GC bias substantially, but did not eliminate it completely.
In light of this analysis, together with previous concerns surrounding the AT richness in the Alphaproteobacteria (Brindefalk et al., 2011, Rodríguez-Ezpeleta and Embley, 2012, Viklund et al., 2012), we tested RY-recoding of the dataset. This approach (Woese et al., 1991, Phillips and Penny, 2003) masks GC content bias by recoding adenosine (A) and guanosine (G) as purines (R), and cytidine (C) and thymidine (T) as pyrimidines (Y). For consistency with our other analyses, we analyzed the recoded dataset with the GTR model. This approach has also been taken by others (Rodríguez-Ezpeleta and Embley, 2012). Support for most of the orders and families dropped in the RY-recoded datasets, compared to those with regular coding (bootstrap summary in Fig. 3.4B; summary statistics in Table 3.1). There was a large decrease in support for the monophyly of the Rhodospirillales, Rhizobiales and Caulobacterales. However, support for the monophyly of some orders, such as the Pelagibacterales and Sphingomonadales, did not change greatly. The trees of the RY-recoded datasets that included mitochondria resolved the Pelagibacterales as a sister clade to the mitochondria–Rickettsiaceae–Midichloria mitochondrii–Acetobacteraceae clade (without the Holosporaceae). Conversely, three of the four trees without the mitochondria placed the Pelagibacterales as paraphyletic with the Rickettsiales.

As a consequence of the poor resolution obtained with the RY-recoded dataset, we utilized a novel variant: each RY-recoded dataset was concatenated to an MK-recoded dataset (‘RYMK’). In the MK recoding, adenosine (A) and cytosine (C) were recoded as amino bases (M), and guanosine (G) and thymidine (T) as keto bases (K). Thus, the GC bias information is lost in the same way as RY-recoding, however the concatenated RYMK
dataset retains the overall number of characters. This approach was designed as a tool to improve phylogenetic analyses of datasets with GC bias, rather than reflecting any biological process(es). We note that RY-recoding has been used similarly (Phillips and Penny, 2003). The site-independent nature of the GTR model ensured that it remained appropriate for analyzing the concatenated RYMK dataset.

Used on its own, MK-recoding suffered from similar issues to RY-recoding (Fig. 3.4C). In particular, the MK-recoded datasets do not resolve the *Rhodospirillales* correctly; consequently the basal *Holosporaceae* is destabilized to the point that it clusters with the *Rickettsiales*. However, average support values and Robinson-Foulds metrics (Table 3.1) showed that the MK-recoded trees were more consistent than the RY-recoded trees, with respect to alignment method and rate model. On the other hand, in the RYMK-recoded datasets the grouping of the *Pelagibacterales* with the *Rickettsiaceae*, *Midichloria mitochondrii*, *Anaplasmataceae* and mitochondria, and the grouping of *Holosporaceae* with the remaining orders, are both more supported than in either the RY-recoded datasets or the MK-recoded datasets (bootstrap summary in Fig. 3.4D). These groupings also have more consistent topologies (7 of 8 trees; compare Fig. 3.4B-C with Fig. 3.4D).

### 3.2.6 The trimmed concatenated trees

We have asserted that it is advantageous to focus on LSU and SSU sequences because it maximizes the number of taxa that can be included. In contrast, genome concatenation studies rely on a smaller number of (fully sequenced) taxa. In order to test the impact of reduced taxon selection on overall topology, we constructed trimmed datasets with fewer bacterial taxa. By reducing the number of taxa on short branches, in clades that had more than 95% bootstrap support in the trees from the complete datasets, we reduced the number of *Alphaproteobacteria* sequences from 166 to 86, while leaving the number of mitochondrial sequences at 24. Despite the high inclusion threshold, the trees differed substantially from the complete phylogenies. Bootstrap summaries are provided in Fig. 3.6. None of the trimmed trees with regular coding resolved the *Rhodospirillaceae* monophyletically, the location of the *Holosporaceae* within the clade containing *Caulobacterales* varied substantially, and the ARB-SINA-aligned trees without mitochondria placed the *Pelagibacterales* paraphyletic with the *Rickettsiales*, albeit with poor bootstrap support (< 40%). The trees inferred from the RY-, MK- and RYMK-recoded datasets had
even more diverse topologies (bootstrap summaries in Figs. 3.6). The Rhodospirillales were either paraphyletic with respect to the clade composed of Sphingomonadales, Rhizobiales, Rhodobacterales and Caulobacterales, or paraphyletic or polyphyletic in respect to the Holosporales. The Pelagibacterales was found basal to the remaining Alphaproteobacteria (except for Magnetococcus marinus) or basal to the clade containing the Caulobacterales in all but three trees. Despite the differences, the mitochondria resolved consistently in the trimmed and complete trees.

**Figure 3.6 (Next pages).** Summary of bootstrap supports of ‘trimmed’ datasets. A. Regular-coded complete datasets. B. RY-recoded complete datasets. C. MK-recoded complete datasets. D. RYMK-recoded complete datasets.
Rhizobiales
Rhodobacterales
Caulobacterales
Sphingomonadales
Rhodospirillaceae
Acetobacteraceae
Holosporaceae
Anaplasmataceae
Rickettsiaceae
Mitochondria
Pelagibacterales
Magnetococcales
outgroup

Different topologies with polyphyletic Rhodospirillaceae and Holosporaceae in all trees

ARB CAT no mt: 81
ARB GTRF no mt: 60
Muscle CAT no mt: 74
Muscle GTRF no mt: 47
ARB CAT mt: 80
ARB GTRF mt: 62
Muscle CAT mt: 73 (zero length branch)
### Chapter 3

<table>
<thead>
<tr>
<th>Rhizobiales</th>
<th>Rhodobacterales</th>
<th>Caulobacterales</th>
<th>Sphingomonadales</th>
<th>Rhodospirillaceae</th>
<th>Acetobacteraceae</th>
<th>Holosporaceae</th>
<th>Anaplasmataceae</th>
<th>Rickettsiaceae</th>
<th>Mitochondria</th>
<th>Pelagibacterales</th>
<th>Magnetococcales</th>
<th>Outgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.0</td>
<td>11.2</td>
<td>2.0</td>
<td>16.4</td>
<td>8.3</td>
<td>9.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.6</td>
<td>45.2</td>
<td>95.6</td>
<td>90.9</td>
<td>44.1</td>
<td>42.0</td>
<td>76.5</td>
<td>80.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21.7</td>
<td>18.5</td>
<td>25.5</td>
<td>21.6</td>
<td>21.4</td>
<td>12.6</td>
<td>27.5</td>
<td>23.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>2.2</td>
<td>0.5</td>
<td>1.7</td>
<td>1.0</td>
<td>1.6</td>
<td>1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### C.}

- Arb GTRI no mt: 80
- Muscle CAT no mt: 34
- Muscle GTRI no mt: 67
- Arb CAT mt: 47
- Arb GTRI mt: 29
- Muscle CAT mt: 52
- Muscle GTRI mt: 52

- Arb CAT no mt: 93, 93
- Arb GTRI no mt: 79, 92
- Muscle CAT no mt: 89, 42
- Muscle GTRI no mt: 81, 85
- Arb CAT mt: 33, 62
- Arb GTRI mt: 34, 24
- Muscle CAT mt: 92, 52
- Muscle GTRI mt: 72, 10
D.

Rhizobiales
Rhodobacterales
Caulobacterales
Sphingomonadales
Rhodospirillaceae
Acetobacteraceae
Holosporaceae
Anaplasmataceae
Rickettsiaceae
Mitochondria
Pelagibacterales
Magnetococcales
outgroup

Different topologies with polyphyletic Rhodospirillaceae and Holosporaceae in most trees
3.2.7 Trees with single clades removed

In an effort to identify the particular taxa or groups that may be critical to the observed instability of the complete phylogeny, we tested the effects of clade-specific jackknifing on topology. Several orders or families were removed, and trees were constructed with each of these single clades removed. The six orders and families removed were the Rhodospirillales (‘rhodoless’ dataset in Table 3.1), Magnetococcus marinus (‘magnetoless’ dataset in Table 3.1), the Rickettsiales including the Holosporaceae (‘exorickettsialss’ dataset), the Rickettsiales excluding the Holosporaceae (‘rickettsialss’ dataset), the Holosporaceae (‘hololess’ dataset), and the Pelagibacterales (‘pelagiless’ dataset). For each of the six datasets, eight trees were constructed (ARB-SINA and MUSCLE alignments; GRTT and GTRCAT rate models; plus and minus mitochondrial sequences). Removal of most orders had little effect on the tree. An exception was the problematic resolution of the Holosporaceae (e.g. paraphyletic in six trees out of 48, basal to the Pelagibacterales in two trees, basal to the Rhodospirillales in three trees, or nested in the Rhodospirillales in one). Removing the Rhodospirillales had the largest effects, destabilising the Holosporaceae and the Pelagibacterales (Holosporaceae basal to Pelagibacterales and Rickettsiales in two trees; Pelagibacterales basal to Holosporaceae and the clade of various orders in four trees).

3.2.8 Mitochondrial placement

Trees were made with and without mitochondrial sequences (Table 3.1). Despite the longer branches of the mitochondrial clade, the trees inferred from the various datasets with mitochondria better resolved the overall topology shown in Fig. 3.3, as opposed to the trees without mitochondria. For example, the primary clade composed of Rickettsiaceae, Anaplasmataceae, Midichloria mitochondrii and Pelagibacterales – in which mitochondrial sequences also cluster – and the clade with a basal Holosporaceae and various orders are supported much more strongly in the datasets with mitochondria, especially the ARB-SINA-aligned datasets. This is true not only for the regularly coded full datasets (Fig. 3.4A), but also for the RY-, MK- and RYMK-recoded ones (Fig. 3.4B–D), and the jackknifed datasets. However, this is not the case for the trimmed datasets, where the changes in support with and without mitochondria vary between alignment method and inference model.
In contrast to the increase in support for the primary clade containing the mitochondria in the full trees, the monophyly of the *Rhodospirillales* and the *Rhodospirillaceae* loses support when the mitochondria are added, especially in the RYMK-recoded dataset. The internal topology of the *Rhizobiales* also changes. In the primary trees without mitochondria, the *Hyphomicrobiaceae* is basal to a clade formed by the *Beijerinckiaceae*, *Methylcystaceae*, *Methylobacteriaceae* and *Xanthobacteraceae*, while in those trees with mitochondria the family is basal to the sister subclade of the *Rhizobiales*, composed of the remaining families bar *Parvibaculum lavamentivorans* (basal to the two subclades) and with the addition of *Labrenzia* spp. and *Roseibium* sp. (Fig. 3.3).

The datasets with and without mitochondria were separately aligned and trimmed. One hypothesis for the variations in topology between trees, with and without mitochondria, was that the trees with mitochondria were influenced by having fewer characters as a result of the Gblocks editing criteria. We tested this hypothesis by pruning mitochondrial sequences post-alignment, from all datasets that included mitochondria (‘complete mtDel’ dataset in Table 3.1). These new alignments were processed with both rate models, and the resulting trees were compared to those of the alignments that were initially done without mitochondria. In order to compare the trees with mitochondria and the trees without, the mitochondrial leaves were pruned from the former group. Overall, the locations of the *Pelagibacterales*, the *Holosporaceae* and other groups of interest are the same in the ‘mtDel’ trees as seen in the other trees. However, the bootstrap support of all groups is lower (Fig. 3.7), except for the support for the monophyly of the *Holosporales* in the trees from the ARB-SINA aligned dataset. Overall, the presence of the mitochondrial sequences adds support to the trees, but at the same time results in a smaller number of characters (due to Gblocks editing).
Figure 3.7. Summary of bootstrap supports of ‘mtDel’ datasets.
3.2.9 16S-23S dataset with several 16S-only sequences

There are a vast number of species represented in the databases solely by their 16S rRNA gene sequences. Within the Alphaproteobacteria, there are 975 validly-described type strains (Cole et al., 2009), against approximately 150 genotyped species. Several of the species for which we have only 16S rRNA gene sequences branch deeply within the tree, to the point that some are reported to be part of their own orders. As a consequence, a group of deep-branching taxa with only 16S sequences was aligned to the other 16S sequences, and the 23S rRNA gene positions left as missing data in the concatenated alignments. The result was the ‘combo’ dataset (Table 3.1).

In the resulting trees (Fig. 3.8), the deep-branching species that were added reduce the support for several nodes, indicating that this approach is not ideal. However, it does offer a snapshot into the diversity of the Alphaproteobacteria that is not covered by the genome databases. Several species classified as Rhodospirillales do not cluster with the Rhodospirillales, but instead cluster with the orders currently represented by a single genus, some of which appear to be synonyms. More specifically, in the trees, the order Kiloniellales contains Kiloniella laminariae, Kopriimonas byunsanensis, Rhodovibrio salinarum and Pelagibius litoralis. The order Kordiimonadales includes Kordiimonas gwangyangensis, Rhodothalassium saleigens and a diverse group of iodine-oxidising bacteria. The order Sneathiales contains only Sneathiella chinensis. However, in the Rhodospirillales, several species with only 16S sequences are found in basal positions. The Acetobacteraceae can be expanded to include basally Elioraea tepidiphila, Alysiosphaera europeae and Geminicoccus roseus. Furthermore, the genus Tistrella may be basal to all of the Rhodospirillales. Full genome sequences of these species may therefore increase the resolving power of future studies.

3.2.10 Approximately Unbiased tests

Trees from the 16S, 23S, complete with regular encoding, complete RY-recoded and complete RYMK-recoded datasets were assessed against the various alignments with the Approximately Unbiased (AU) test (Shimodaira and Hasegawa, 2001). A $p$-value of 0.05 was used as the cut-off, such that trees with $p$-values below this number could be rejected based on an alignment. Generally, trees from a given dataset were unable to be rejected by their respective alignments from either ARB-SINA or MUSCLE. On the other
hand, they could be rejected by most of the alignments that were based on other datasets. For the complete datasets without mitochondria, none of the four regularly coded primary trees (GTRCAT and GTRΓ with ARB-SINA and MUSCLE alignments) could be rejected based on the ARB-SINA or MUSCLE alignments. For the complete datasets with mitochondria, of the four regularly-coded trees, two could be rejected by one of the alignments. The four trees with mitochondria agree on the topology of the orders and families and differ by a Robinson-Foulds distance of less than 22 different bipartitions while the four trees without mitochondria disagree in some instances regarding families, but differ by less than 18 bipartitions. Whereas all of the trees from the 16S, 23S and RY-recoded datasets could be rejected based on the two regularly-coded complete alignments, the RYMK-recoded trees could not be rejected based on 11 of the 16 comparisons with the regularly coded alignments.

Figure 3.8 (next facing pages). Tree inferred with the ARB-SINA aligned ‘combo’ dataset under a GTRΓ model. This dataset included strains that were represented solely by a 16S sequence, therefore revealing the diversity that is not covered by genome sequences. Bootstrap values (n = 1,000) are indicated at the nodes. Red arrows indicate how a taxon or clade differs in the other trees. Insets show clades which are particularly rich in unsequenced genomes (black text) compared to sequenced genomes (green text). The internal topology of the Rhodospirillales differs between trees, but the six constant subdivisions are highlighted.
3.3 Discussion

3.3.1 rRNA focus

In this study, we updated the alphaproteobacterial rRNA gene tree and explored its topological stability by systematically varying a series of parameters: alignment method; rate model; character encoding; and taxon sampling. We focused on rRNA genes, rather than protein-encoding genes, in order to recover phylogenetic signals that may have been obscured due to the genomic AT-richness and ancient divergence events that are hallmarks of the Alphaproteobacteria. We focused on concatenated 16S-23S rRNA gene sequences to give better support than 16S alone, while also allowing the inclusion of many more taxa than what is currently feasible with available genomes for concatenation studies. One downside of our approach is the limited number of full-length 23S sequences, relative to 16S. However on balance, and given the controversy that surrounds the phylogeny of the Alphaproteobacteria (Brindefalk et al., 2011, Georgiades et al., 2011, Thrash et al., 2011, Rodríguez-Ezpeleta and Embley, 2012, Viklund et al., 2012), we hypothesized that concatenated 16S-23S rRNA gene sequences would offer the greatest insights into accurate placement of the major groups.

3.3.2 GC bias

The debate over alphaproteobacterial phylogeny focuses mainly on the location of the Pelagibacterales (SAR11 clade), and particularly whether it is a sister clade to the Rickettsiales or whether they are artefactually attracted due to shared low GC content (Brindefalk et al., 2011, Rodríguez-Ezpeleta and Embley, 2012, Viklund et al., 2012). Despite concerns about the effects of such compositional biases on tree topologies (Woese et al., 1991), little is known about the effect of genomic GC composition on 16S and 23S sequences. Here, we found that for the Alphaproteobacteria sequences that we analyzed, the GC content of rRNA genes was 5- to 6-fold less variable than the genomic GC content (Fig. 3.1). This validated the choice of rRNA genes for minimizing artefacts that were due to shared AT richness, although further analysis showed that the GC bias was not completely eliminated in our rRNA gene datasets (Fig. 3.5).

Consistent with the result in Fig. 3.1, several of our further tests argued against a GC content-driven artefact affecting the position of the Pelagibacterales. Using the jackknifing approach, we tested the hypothesis that the Pelagibacterales and the Rickettsiales
attract each other. Under this hypothesis, the AT-rich *Rickettsiales* should mask the true phylogenetic signal by attracting the AT-rich *Pelagibacterales*, and therefore the removal of the *Rickettsiales* from the dataset should reveal the location of the *Pelagibacterales* independently of this attraction. However, the *Pelagibacterales* were placed in a similar position, both in the complete dataset (Fig. 3.3) and when *Rickettsiales* were missing. Additional evidence against compositional bias-driven attraction comes from the recoded datasets. The RY-recoded trees were less supported on many nodes, which may be due to the presence of two instead of four character states. We introduced RYMK-recoding to help overcome the limitation caused by the reduced number of character states and we found that the resulting trees concurred with the regular dataset, regarding the clustering of the *Pelagibacterales* with the mitochondria, *Rickettsiaceae* and *Anaplasmataceae*. Approximate Unbiased tests confirmed that RYMK-encoding (which eliminates GC bias) was in greater agreement with regular encoding than either RY- or MK-encoding. Therefore we recommend RYMK-recoding as a superior alternative to RY-recoding for assessing artefactual attractions that may arise due to GC content biases.

Overall, in datasets of concatenated 16S and 23S sequences, we found no evidence that GC content may be contributing to a topology where the *Pelagibacterales* cluster with the *Rickettsiales*.

### 3.3.3 Effects of alignment method, rate model and taxon sampling

In general, the MUSCLE and ARB-SINA alignments generated trees that agreed well with each other when used in combination with Gblocks. Also, the trees generated with the two different rate models (GTRΓ and GTRCAT) did not differ substantially and had similar AU scores. Therefore it can be concluded that in the case of rRNA gene-based trees computed with RAxML, neither the choice of alignment program, nor the rate model, significantly affect the final topology.

In contrast, taxon selection greatly affected the final tree. The number of taxa present was reduced by trimming the dataset of leaves that were assumed to be contributing little to the overall tree topology. Contrary to expectations, the trees from the trimmed datasets had lower support values and were much less consistent with each other compared to the complete trees (compare Fig. 3.4 and Fig. 3.6). Extra species, represented by 16S sequences only, were also added to yield the ‘combo’ dataset. Trees constructed with this enlarged dataset also showed reduced support, consistent with the reduction in
characters that came from not having 23S sequences for these species. Our strategy of aligning concatenated 16S-23S sequences, using at least one representative of each species, yielded trees with the highest possible support (Fig. 3.3).

3.3.4 Taxonomic observations

The divergence of the clade with Rickettsia and the clade with Caulobacter has been estimated to have occurred 1,650–2,390 million years ago (Battistuzzi et al., 2004). Therefore, it is clear that several taxa may find themselves on long branches, which may spuriously attract (Kolaczkowski and Thornton, 2009). To minimise the quantity of long branches, one representative of each alphaproteobacterial species, represented by both 16S and 23S sequences, was chosen in this study. Moreover, by analyzing only the 16S and 23S markers it was also straightforward to screen manually for misannotation errors and to correct or exclude problematic taxa. For example, the contig NZ_AAAP01003712 for Magnetospirillum magnetotacticum was found to be contaminated with sequence from Methylobacterium populi, from position 3,624 to 8,812 (data not shown).

In contrast to previous analyses, this study covered all of the known diversity within Pelagibacterales (i.e. subgroups I, III, IV and V, as classified by Grote et al., 2012). The monophyly of the group was confirmed with more than 95% support in all trees inferred. Therefore, our data provided no evidence that Pelagibacterales may be polyphyletic, in agreement with one recent study (Grote et al., 2012b), but not another (Rodríguez-Ezpeleta and Embley, 2012). The Pelagibacterales fall basal to a clade composed of mitochondria and a Rickettsiales subclade without the Holosporaceae, in all of the full trees. This grouping has a moderate bootstrap support, but this is most likely due to the instability of the Holosporales (vide infra) and not due to an AT-attractional bias (vide supra).

In our four complete, regularly coded trees the mitochondria are a sister group to a clade formed by Anaplasmataceae and Rickettsiaceae, with high support. This result is consistent with several studies that used concatenated protein phylogenies (Williams et al., 2007, Georgiades et al., 2011, Thrash et al., 2011), although slightly different to Georgiades et al. (2011) and Thrash et al. (2011), which found support for the Pelagibacteraceae as the sister clade to the mitochondria. However, the placement of the Pelagibacterales near the branch point of the mitochondria and the Rickettsiales, regardless of the order, is in greater agreement with all three of the above studies than with the results
presented by Brindefalk et al. (Brindefalk et al., 2011), Viklund et al. (Viklund et al., 2012), and Rodríguez-Ezpeleta and Embley (Rodríguez-Ezpeleta and Embley, 2012), where the Pelagibacterales are placed elsewhere in the Alphaproteobacteria entirely.

*Magnetococcus marinus* is the sole genome-sequenced representative of the Magnetococcales, a clade that is basal to the remaining Alphaproteobacteria (Lefèvre et al., 2009, Bazylinski et al., 2013). At the outset of this study, its membership in the Alphaproteobacteria was unclear, although it has since been proven correct (Bazylinski et al., 2013). Having chosen our outgroup to include the most diverse members of the Betaproteobacteria, Gammaproteobacteria and Zetaproteobacteria, we can also confirm the membership of the Magnetococcales within Alphaproteobacteria. Its inclusion in our phylogenetic analyses greatly reduced the length of the branch leading to the remaining alphaproteobacterial orders. This enabled us to circumvent an earlier problem, in which the absence of *Magnetococcus marinus* meant that the choice of outgroup affected the topology of the alphaproteobacterial tree (Thrash et al., 2011).

Similarly, in our trees the Holosporaceae are basal to a large clade of alphaproteobacterial orders, yet they are represented solely by *Odyssella thessalonicensis* and *Caedibacter caryophilus* (Fig. 3.3). The former was only sequenced recently (Georgiades et al., 2011), while the latter is represented solely by 16S and 23S sequences. Previously, Holosporaceae has been classified as a family in the Rickettsiales (Garrity, 2005, Lee et al., 2005); however, in our datasets this clade does not cluster within this order. Instead, it lies basal to the clade that comprises several orders, but not the Rickettsiales and Magnetococcales. In light of our new analysis, we propose to remove the Holosporaceae from the Rickettsiales and to create a new order, the Holosporales ord. nov. Consequently, under this revised classification only the Rickettsiaceae, the Midichloriaceae and the Anaplasmataceae comprise the order Rickettsiales (sensu novo).

There are several peculiarities involving the Rhodospirillales in this work and in other studies (Lee et al., 2005, Rodríguez-Ezpeleta and Embley, 2012), albeit generally reported without comment. There seems to be an instability within the Rhodospirillales clade: whereas the Acetobacteriaceae resolves with high support, the Rhodospirillaceae rarely resolves monophyletically or with high support (this study, Lee et al., 2005, Rodriguez-Ezpeleta and Embley, 2012). In this study the deletion of the Rhodospirillales profoundly reduces the support of the location of the Holosporaceae. The Holosporaceae is located in a
clade with the *Rhodospirillales*, *Sphingomonadales*, *Rhizobiales*, *Caulobacterales* and *Rhodobacterales* with an average support of 82% in the trees from the complete dataset, but with only 42% support (on average) in the trees without the *Rhodospirillales*. It can be concluded that the *Rhodospirillales* play a large role in supporting the topology at the internal node of the tree. Poor support, or even paraphyly (Lee *et al.*, 2005, Rodríguez-Ezpeleta and Embley, 2012), of the *Rhodospirillales* could be indicative of an incorrect resolution of the internal node separating the clade composed of *Rickettsiales* and *Pelagibacterales* from the clade that includes the *Caulobacterales*, *Rhodospirillales* and several other orders.

The problematic internal organisation of the *Rhodospirillales* is even more aggravated in the datasets that include species represented only by 16S rRNA gene sequences. Some species classified as *Rhodospirillales* actually belong to the separate orders *Kiloniellales*, *Kordiimonadales* and *Sneathiellales*, while several *Rhodospirillales* species, such *Tistrella mobilis* (whose genome was published too recently to be included in this study, Xu *et al.*, 2012), form clades that are basal to the other *Rhodospirillales*. The implication is that the current family-level classification of the *Rhodospirillales* will need revisiting in the future.

### 3.3.5 Proposal of new subclasses

In light of the number of orders present in the *Alphaproteobacteria*, we propose the creation of three subclasses that are based on the tree topology found here (Fig. 2), and that aid in the description of the groups. In particular, we propose to distinguish the two clades that are apical to the *Magnetococcales*. Therefore we propose to place: (i) the *Magnetococcales* in the *Magnetococcidea subcl. nov.*; (ii) the *Rickettsiales* (*sensu novo*), the protomitochondrion (*i.e.* the bacterial ancestor of the eukaryotic organelle) and *Pelagibacterales ord. nov.* in the *Rickettsidea subcl. nov.*; and (iii) the *Holosporales ord. nov.*, *Rhodospirillales*, *Sphingomonadales*, *Rhizobiales*, *Caulobacterales* and other orders in the *Caulobacteridae subcl. nov*. Our proposal is summarized in Fig. 3.9. Overall, our updated analysis supports the position of the *Pelagibacterales* as a sister group to the composite clade containing *Rickettsiales* and the mitochondrial branch. Our analysis has also provided support for taxonomic assignment of several recently sequenced species, including *Odyssella thessalonicensis* and *Magnetococcus marinus*.
Figure 3.9. Proposed subclasses of the Alphaproteobacteria. The three proposed subdivisions are the Magnetococci, the Rickettsiales and the Caulobacteridae. Furthermore, the Holosporaceae should be removed from the Rickettsiales, however the identities of the family-level subdivisions of the Holosporales, such as the Holosporaceae (marked with an asterisk), are beyond the scope of this work. Under this scheme the Rickettsiales are comprised solely of the Rickettsiaceae, Anaplasmataceae and Midichloriaceae. The protomitochondrion (†) is an extinct organism that gave rise to the mitochondrial organelles of eukaryotes.
3.4 Materials and Methods

3.4.1 Taxon selection

The 16S and 23S rRNA genes used to construct the concatenated datasets were obtained from IMG (v350). The Perl scripts list_clustermaker.pl (§ IV.III) and fasta_acceptor.pl were used to select a sole sequence as a representative for each species present while checking whether the sequence matched that of the other paralogs. In negative cases, the most common sequence was chosen. To supplement the trees with additional Alphaproteobacteria without sequenced genomes, 16S or both 16S and 23S rRNA gene sequences were obtained from GenBank.

3.4.2 GC content

A Perl script, GC-counter.pl (§ IV.V), was used to determine the GC content of the rRNA gene sequences; genomic GC content was obtained from IMG. Scatter plots were created in Prism 4. The slopes of the linear regressions of both the SSU and the LSU rRNA gene GC content over genomic GC content for the Alphaproteobacteria and for the mitochondria were calculated in Excel with the slope function. Box plots were created in R using boxplot, and statistical significance was evaluated using wilcox.test.

3.4.3 Sequence QC

The quality of the sequences was corrected by trimming the ends according to the match from the ARB-SINA aligner (Pruesse et al., 2012). The length was checked with a Perl script, impostor checker.pl (§ IV.VI), and all sequences under 1200 (for 16S) or 2000 (for 23S) bases were removed. One outcome of imparting this size cut-off was to remove numerous mitochondrial sequences; however, including these mitochondrial sequences made no difference to the observed topologies (data not shown). The presence of multiple copies of rDNA in most organisms allowed sequencing errors and contaminations to be identified.
3.4.4 Alignment/gap-removal

The sequences were aligned either with Muscle (Edgar, 2004) under default settings or with Sina (Pruesse et al., 2012) set to remove terminal unaligned bases and not to reverse complement sequences (due to issues with some mitochondrial sequences). Poorly aligned sites were curated with Gblocks (Talavera and Castresana, 2007) using the settings from (Sassera et al., 2011): -b1=(n/2)+1 -b2=(n/2)+1 -b3=n/2 -b4=2 -b5=h. All dataset variants were created by deleting the targeted taxa before the alignment step.

3.4.5 RY-, MK- and RYMK-recoded datasets

For the RY-recoded datasets, the completed (aligned and trimmed) datasets were recoded, by converting all A and G bases to R and all T and C bases to Y. For the MK-recoded datasets, all A and C bases were converted to M and all T and G bases were converted to K. For the RYMK-recoded datasets, the complete dataset was duplicated and one copy was RY-recoded, while the second was MK-recoded; the two differentially recoded variants were then concatenated.

3.4.6 Maximum likelihood trees

Maximum likelihood inferences were performed with RAxML 7.2.9 (Stamatakis, 2006) with 1,000 bootstrap replicates under either a GTR or a GTRCAT model. The trees were displayed in FigTree3.1 (Drummond et al., 2012). Figure 2 was further annotated with Adobe Illustrator CS4. The supplementary tree summary figures were made with Illustrator CS4, with the aid of Newick utilities for data extraction.

3.4.7 Topology evaluation

The bootstrap support for various groupings of interest were obtained via the Perl script descriptor.pl (§ IV.VIII), using Newick utilities (Junier and Zdobnov, 2010). First the bootstrap trees were rooted to the outgroup (nuw_reroot), then the leaves in these were renamed by mapping the names to the groups of interest (nuw_rename), after which the clades composed of a single name were condensed (nuw_condense) and finally the trees were ordered (nuw_order). Once this was done, the script determined the support for monophyly of the groupings by simply counting the frequency of the names in each of the resulting trees, where a single instance would indicate monophyly.
3.4.8 Approximately Unbiased tests

The final trees of the full datasets were concatenated into a single file and the per site log-likelihoods were calculated with both the ARB-SINA-aligned and MUSCLE-aligned datasets (RAxML via option -fg), in order to be converted into a matrix (makermt --puzzle) that could be interpreted by CONSEL (conel and catpv). Robinson-Foulds distances were calculated between each tree in a set with HashRF (code.google.com/p/hashrf/).
Chapter 4

The Biochemistry of Wolbachia, P. ubique and T. maritima MetC

Acknowledgements:

- Dr Valerie Soo (formerly at Institute of Natural Sciences, Albany, Massey University) constructed pBAD/metC_{Eco}
- Natasha le Roux (formerly at Institute of Natural Sciences, Albany, Massey University) constructed pMAL/metC_{Pub}-malE
- Dr Chris Squire and Yuliana Yosaatmadja (School of Biological Sciences, University of Auckland) crystallised T. maritima MetC and solved its structure.
- Dr Torsten Kleffmann (Centre for Protein Research, University of Otago) obtained mass spectrometry data on the cleavage product of T. maritima MetC
- Prof. Dominique Mengin-Lecreulx (I.B.B.M.C., Université Paris Sud) provided E. coli strain WM335
4.1 Introduction

4.1.1 From genes to enzymes

In chapter 2 it was shown that in three different lineages across the metC gene tree, metC has displaced the alr gene and a representative gene from each lineage was able to rescue both Escherichia coli ΔmetC and Δalr ΔdadX knockout strains. In all three bacteria the alanine racemisation is essential, albeit at low levels, whereas the cystathionine β-elimination activity differs. In Wolbachia, there is no apparent physiological role for cystathionine β-elimination as the whole methionine biosynthesis pathway is absent, whereas in Pelagibacter ubique and Thermotoga maritima cystathionine β-elimination is likely to be physiological, but appears to form an auxiliary route to methionine biosynthesis. In this chapter the biochemistry of the three encoded enzymes is explored in order to elucidate the full balance of activities and to determine the similarities and differences between promiscuous activities and multitasking activities.

4.1.2 PLP

The ability of MetC to catalyse not only the β-elimination of cystathionine, but also the racemisation of alanine, the reaction of Alr, stems from the cofactor, pyridoxal-5′-phosphate (PLP, Fig. 4.1).

PLP is a derivatised pyridine ring with a reactive aldehyde (4-formyl group) that allows the molecule to reversibly form a Schiff base (secondary aldimine) with the terminal amine of a lysine residue of the enzyme, or with the amine of the substrate (Eliot and Kirsch, 2004). Without substrate, the phospho-pyridoxyl group is bound to a specific lysine residue, but is transferred to the amine of the substrate, when present. The conjugated system of the PLP and its ability to be in either aromatic or quinoid form allow the adjacent carbon of the substrate to break one of its bonds, before regaining or losing another bond at the same or at a different position (Eliot and Kirsch, 2004). MetC and Alr both utilise PLP, but are not homologous. Cofactor binding in the various families of PLP-dependent enzymes appears to be the result of convergent evolution, as the ability to bind PLP has few requirements: a pocket large enough to accommodate the cofactor, a lysine to interact with
it, an aromatic residue to π-stack with its heterocyclic ring and some positive residues to hold its phosphate (Eliot and Kirsch, 2004).

Despite its catalytic potential, the spatial constrains imparted by the cavity of the active site dictate the reaction specificity (Eliot and Kirsch, 2004). According to the Dunathan hypothesis, bond cleavage is dictated by the conformation of the substrate in respect to the plane of the phospho-pyridoxyl group; specifically, the bond to be broken is oriented parallel to the delocalised π-system of the ring (Eliot and Kirsch, 2004).

**Figure 4.1.** PLP-dependent racemisation and β-elimination. In alanine racemisation the arrow pushing is shown solely for the L to D direction. In the reverse direction, the tyrosine acts as a base by abstracting the proton, while the lysine donates a proton. In the elimination reaction the charge is rearranged, resulting in the departure of the thiol. The Schiff base is then transferred from the amino acrylate to the lysine, releasing the aminoacrylate.
4.1.3 Reactions

In the case of Alr and MetC, both reaction mechanisms (Fig. 4.1) start with the same step, namely the deprotonation of the $\alpha$-carbon, which results in a secondary ketimine with the phospho-pyridoxyl in quinonoid form. In the racemisation reaction, the $\alpha$-carbon is reprotonated on the opposite face. In the elimination reaction, the leaving of the $\beta$-thioether is favoured, therefore resetting the charge balance back to phospho-pyridoxyl in aromatic form with an aldimine bond to aminoacrylate, which can be substituted by the lysine (Clausen et al., 1996, Toney, 2011). Due to enamine–ketimine tautomerism, aminoacrylate is spontaneously hydrolysed into pyruvate and ammonia. Therefore, the overall reaction catalysed is: cystathionine $\rightarrow$ homocysteine + pyruvate + ammonia.

However, the racemisation reaction catalysed by MetC differs slightly from that catalysed by Alr. In Alr the PLP-interacting lysine is involved with the proton of D-alanine, while on the other face the tyrosine that $\pi$-stacks with the PLP is involved with L-alanine (Watanabe et al., 2002). Contrarily, in MetC the $\pi$-stacking tyrosine is involved with the proton of D-alanine and the lysine with that of L-alanine, which means that the acid–base pair is inverted in MetC compared to Alr (Soo, 2012). This is consistent with the fact that in MetC the PLP-interacting lysine abstracts the proton from cystathionine, an L-amino acid (Clausen et al., 1996).

While it is essential for racemisation, in *E. coli* MetC the $\pi$-stacking tyrosine does not play a pivotal catalytic role in the cystathionine $\beta$-elimination as demonstrated by an Y111F mutant (Lodha and Aitken, 2011).

A similar overlap between racemisation and $\beta$-elimination is seen with human serine racemase (Wang et al., 2012). This PLP-dependent racemase belongs to the type II family (Smith et al., 2010), unlike MetC (type I, § 4.1.2) and Alr (type III). The $\beta$-elimination of serine involves the same mechanism as the $\beta$-elimination of cystathionine with the difference that the leaving group is a hydroxyl ion, rather than a thiol, which is a better leaving group. In human serine racemase the proton of L-serine is abstracted by the PLP-interacting lysine, similarly to MetC and antithetically to Alr. In the case of D-serine the proton is abstracted by a serine residue (Wang et al., 2012). Therefore, in human serine racemase the racemisation and the $\beta$-elimination compete over the same substrate. It is clear that PLP-driven racemisation and $\beta$-elimination are not as dissimilar as the different EC numbers would suggest (EC 5.1.1.1 and EC 4.4.1.8 respectively), but the reactions are
influenced by the substrate. For example, it could be hypothesised that if the bridging sulfur of cystathionine were a carbon (2,7-diaminosuberic acid) the molecule might proceed through with the racemisation. This indicates that the two activities share several commonalities and explains why it is possible to have both activities within the same active site in MetC.

4.1.4 Promiscuity vs. multifunctionality

As mentioned in chapter 1, the difference between a main activity and a promiscuous activity is that the former is physiologically beneficial and selectively maintained, while the latter is non-physiological and neutrally drifting. At the biochemical level, this does not imply that the main activity is more efficient than the promiscuous activity: some enzymes have a higher turnover ($k_{cat}$) of unnatural compounds than towards their natural substrates, especially when the former have groups with strong polar effects, such as activated leaving groups. An example of this is a broad-specificity reductase that displayed a higher turnover for an unnatural ketone with a trifluoromethyl group on the adjacent carbon than towards natural compounds (Ma et al., 2013). Furthermore, the catalytic efficiencies (i.e. the ratio of turnover number to Michaelis constant, $k_{cat}/K_M$) vary quite substantially between cases of promiscuity (Khersonsky et al., 2006) as do the main activities of enzymes (Bar-Even et al., 2011).

Therefore, the experiment in this chapter aimed to investigate an open question from chapter 2, that is, whether there are any discernible biochemical differences between promiscuous activities and low-yield main activities, such as those possessed by multitasking enzymes. This question was answered for the specific case of MetC, by characterising the biochemistry of P. ubique MetC, Wolbachia MetC and particularly T. maritima MetC, whose structure is also analysed.
4.2 Results

4.2.1 Assay overview

Firstly, the MetC enzymes from *E. coli*, *Wolbachia*, *T. maritima* and *P. ubique* were purified (Fig. 4.2, specific details in § 4.2.2–4.2.5, methods § 4.4.1–4.4.2) and characterised via a series of spectroscopic assays (methods § 4.4.3).

![Figure 4.2.](image)

Representative aliquots of the various enzymes used in this study. All wells contain ~1 µg of protein, except for the ladders (A, G, sizes in kDa). The lanes contain *E. coli* MetC (B), *Wolbachia* MetC (C), *P. ubique* MetC (D), *T. maritima* MetC (E) and *T. maritima* MetC S86T/S305C (F, viz. § 5.2.4).

Unless otherwise noted, the assays were done in biological triplicates; that is, different batches of each protein were purified from three independent cultures of the expression strain. For each biological replicate, most of the data points were acquired in technical triplicates. All the data points were used for the non-linear regression in Prism (appendix § I.II.VII).

The main assays were:

- Cystathionine elimination (free thiol detected via Ellman’s reagent, methods appendix § I.II.VIII)
- Cysteine elimination (pyruvate detected via its reduction by lactate dehydrogenase with NADH, methods appendix § I.II.IX)
• D→L alanine racemisation (L-alanine detected via its oxidation by alanine dehydrogenase with NAD\(^+\), methods appendix § I.II.X)

• L→D alanine racemisation (D-alanine detected via the coupling with D-amino acid oxidase, which deaminates it to pyruvate, and lactate dehydrogenase, which is in turn reduces pyruvate with NADH, appendix § I.II.XI)

• D→L glutamate racemisation (L-glutamate detected via its oxidation by glutamate dehydrogenase with NAD\(^+\), methods appendix § I.II.XII)

As cysteine is a reducing agent, it is incompatible with some redox assays (e.g. with iodonitrotetrazolium chloride; Li et al., 2010); however, in my control assays cysteine did not interfere with the reduction of pyruvate by LDH, even at high concentrations, and had only a near negligible effect on the background in the presence of NAD\(^+\) (data not shown).

### 4.2.2 E. coli MetC kinetics

Previous work has determined the cystathionine and cysteine elimination activities and the D-alanine racemisation activity of E. coli MetC, which were used for comparison with the other MetC enzymes (Table 4.1).

The presence of hexahistidine tags is known to affect MetC activity (Farsi et al., 2009), consequently the compiled values differ slightly: in Lodha et al. (2011) the E. coli MetC has an N-terminal hexahistine tag and a linker cleavable with factor Xa, whereas in Soo (2012) the E. coli MetC was expressed from pCA24N/metC which had a N-terminal hexahistine tag with a different linker (MRGSHHHHHHTDPALRA).

<table>
<thead>
<tr>
<th>Table 4.1. E. coli MetC activities measured in other studies. Values are reported with standard error when available.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Reaction</th>
<th>( k_{cat} ) [s(^{-1})]</th>
<th>( K_M ) [( \mu )M]</th>
<th>( k_{cat}/K_M ) [s(^{-1})·M(^{-1})]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystathionine elimination</td>
<td>( 89.5 \pm 13.9 )</td>
<td>( 39 \pm 7 )</td>
<td>( 2.3 \cdot 10^6 )</td>
</tr>
<tr>
<td>(Soo, 2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystathionine elimination</td>
<td>( 34.1 \pm 0.6 )</td>
<td>( 180 \pm 10 )</td>
<td>( (1.9 \pm 0.1) \cdot 10^5 )</td>
</tr>
<tr>
<td>(Lodha and Aitken, 2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine elimination</td>
<td>( 0.49 \pm 0.02 )</td>
<td>( 240 \pm 20 )</td>
<td>( (2.0 \pm 0.1) \cdot 10^3 )</td>
</tr>
<tr>
<td>(Lodha and Aitken, 2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alanine racemisation (L→D)</td>
<td>( 3.3 \pm 0.6 )</td>
<td>( 51,000 \pm 4,000 )</td>
<td>( 65 )</td>
</tr>
<tr>
<td>(Soo, 2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In order to fully compare *E. coli* MetC with the other enzymes, two more assays were performed: D-alanine to L-alanine racemisation and D-glutamate to L-glutamate racemisation (Table 4.2, Fig. 4.3A). In these, I utilised the protein expressed via the pBAD/metC construct designed by Soo (2012).

### Table 4.2. *E. coli* MetC activities measured in this study. Measurements (with standard error) were done in technical triplicate with one biological replicate, whereas no glutamate racemisation could be detected (N.D.).

<table>
<thead>
<tr>
<th>Reaction</th>
<th>$k_{cat}$ [s$^{-1}$]</th>
<th>$K_M$ [µM]</th>
<th>$k_{cat}/K_M$ [s$^{-1}$ · M$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine racemisation (D→L)</td>
<td>1.32 ± 0.05</td>
<td>4,700 ± 500</td>
<td>280 ± 20</td>
</tr>
<tr>
<td>Glutamate racemisation (D→L)</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
</tbody>
</table>

**Figure 4.3.** Michaelis–Menten plots of D→L alanine racemisation reaction in *E. coli* MetC. The enzyme was assayed as described in methods § 4.4.3 and in § I.II.XI. Red dots are values removed by the automatic outlier function in Prism 6.0. Each point is represented with error bars.

Due to the fact that there is no difference in free energy between the two stereoisomers of alanine, the Haldane relationship between the two directions of racemisation should be one, or, in other words, the catalytic efficiencies should be identical (Cleland, 1982). However, this does not appear to be the case with the measured values of *E. coli* MetC. This is most likely due to the high concentration of substrates used in the assays. Regardless, the parameters are still indicative that the Michaelis constant ($K_M$) is lowest for D-alanine, while the turnover ($k_{cat}$) is highest for L-alanine racemisation.

Glutamate racemisation could not be detected, even when the concentration of enzyme was as high as 5 µM.
4.2.3 Wolbachia MetC kinetics

Despite being codon-optimised for expression in *E. coli* (§ 2.2.2), *Wolbachia* MetC did not express with high yields (~0.5 mg purified *Wolbachia* MetC per 1 ℓ LB culture) and a large fraction of the protein was insoluble. Consequently, several conditions were tested for *Wolbachia* MetC expression (methods § 4.4.1), including different time and arabinose concentration, co-expression with helper chaperones or heat-shock chaperone induction with 3% ethanol. None of these was successful (data not shown), therefore the protein was expressed without chaperone induction and instead the volume of culture was scaled up (4 ℓ LB per batch).

In *Wolbachia* MetC, the most efficient reactions were alanine racemisation and cysteine elimination (Table 4.3, Fig. 4.4). It was not unexpected that *Wolbachia* MetC could eliminate cysteine due to the presence of this activity in *E. coli* MetC (Lodha and Aitken, 2011), however, it was unexpected that this activity was ten times stronger than that for cystathionine. In comparison, *E. coli* MetC eliminates cysteine three orders of magnitude less efficiently than cystathionine, despite the former being a thioether of the latter.

In the case of *Wolbachia* MetC alanine racemisation, the standard errors of the catalytic efficiencies for the two activities overlap, indicating that the two values are not significantly different and therefore the Haldane relationship holds.

Due to the low turnover, the errors associated with the glutamate racemisation and cystathionine elimination were high, a phenomenon also seen with *T. maritima* MetC (§ 4.2.5). To reduce the standard errors further replicates could have been done; however, the standard errors are small enough that meaningful conclusions can still be clearly drawn.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>$k_{cat}$ [s$^{-1}$]</th>
<th>$K_M$ [μM]</th>
<th>$k_{cat}/K_M$ [s$^{-1}$ · M$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystathionine elimination</td>
<td>0.0011 ± 0.0002</td>
<td>20 ± 20</td>
<td>60 ± 50</td>
</tr>
<tr>
<td>Cysteine elimination</td>
<td>0.039 ± 0.005</td>
<td>70 ± 40</td>
<td>600 ± 300</td>
</tr>
<tr>
<td>Alanine racemisation (D→L)</td>
<td>0.45 ± 0.04</td>
<td>980 ± 360</td>
<td>500 ± 100</td>
</tr>
<tr>
<td>Alanine racemisation (L→D)</td>
<td>2.3 ± 0.1</td>
<td>3,800 ± 500</td>
<td>580 ± 60</td>
</tr>
<tr>
<td>Glutamate racemisation (D→L)</td>
<td>0.017 ± 0.002</td>
<td>800 ± 410</td>
<td>21 ± 9</td>
</tr>
</tbody>
</table>

Table 4.3. *Wolbachia* MetC activities. Measurements were performed with three biological replicates with multiple measurements per concentration. The values are reported with standard errors.
4.2.4 *P. ubique* MetC kinetics

A Bsc (Hons) student, Natasha le Roux, worked on *P. ubique* MetC and found it to be highly unstable. It was expressible solely as a fusion protein with maltose-binding protein (le Roux, 2012). Consequently, I purified (§ 4.4.2 and § I.II.VI) the enzyme as a fusion protein with maltose-binding protein. Due to the different purification method, amyllose affinity chromatography, some contaminants were present (Fig. 4.2, lane E). Due to the different expression construct, the yield was larger (over 5 mg *P. ubique* MetC per 500 ml culture). Assays revealed the balance of activities to be different from *Wolbachia* MetC (Table 4.4, Fig. 4.5). Similarly to *E. coli* MetC, no glutamate racemisation could be
detected. The cystathionine elimination and the cysteine elimination were equally efficient and were more efficient than alanine racemisation. The Michaelis constants for the various reactions were much higher than those in the other MetC enzymes measured here and the Haldane relationship was not close to one, a peculiarity discussed in section 4.3.2.

**Table 4.4.** *P. ubique* MetC activities. N.D.: not detected. These measurements are from one biological replicate due to expression problems. The measurements were done in technical triplicates. The values are reported with standard errors.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>$k_{cat}$ [s⁻¹]</th>
<th>$K_M$ [µM]</th>
<th>$k_{cat}/K_M$ [s⁻¹ · M⁻¹]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystathionine elimination</td>
<td>1.01 ± 0.09</td>
<td>2,100 ± 500</td>
<td>470 ± 60</td>
</tr>
<tr>
<td>Cysteine elimination</td>
<td>0.079 ± 0.008</td>
<td>190 ± 60</td>
<td>400 ± 100</td>
</tr>
<tr>
<td>Alanine racemisation (D→L)</td>
<td>0.043 ± 0.005</td>
<td>23,000 ± 6,000</td>
<td>1.9 ± 0.3</td>
</tr>
<tr>
<td>Alanine racemisation (L→D)</td>
<td>0.15 ± 0.02</td>
<td>12,000 ± 4,000</td>
<td>12 ± 3</td>
</tr>
<tr>
<td>Glutamate racemisation (D→L)</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
</tbody>
</table>

**Figure 4.5.** Michaelis–Menten plots of the various activities measured of *P. ubique* MetC. A. cystathionine elimination. B. cysteine elimination. C. L→D alanine racemisation. D. D→L alanine racemisation. Each point is represented with vertical bars representing standard error.
4.2.5 *T. maritima* MetC kinetics

The MetC from *T. maritima* was purified (4 mg protein per 1 ℓ culture). Similarly to *P. ubique* MetC, could eliminate cysteine with a similar efficiency to its elimination of cystathionine with a lower $K_M$ towards cysteine elimination, but a higher $k_{cat}$ towards cystathionine elimination (Table 4.5, Fig. 4.6). The catalytic efficiencies for the two directions of alanine racemisation do not differ significantly (overlapping error bars) and similarly to *P. ubique* MetC there is a stronger affinity for the L-isomer of alanine, which was the opposite preference to *Wolbachia* MetC and *E. coli* MetC. In contrast to any of the other enzymes, *T. maritima* MetC could racemise glutamate more efficiently than alanine.

| Table 4.5. *T. maritima* MetC activities. Data obtained using biological triplicates and two or more technical replicates. The values are reported with standard errors. |
|---------------------------------|-----------------|-----------------|-----------------|
| **Reaction**                    | $k_{cat}$ [s⁻¹] | $K_M$ [µM]      | $k_{cat}/K_M$ [s⁻¹·M⁻¹] |
| Cystathionine elimination       | 2.3 ± 0.1       | 500 ± 60        | (5.8 ± 0.6) · 10³ |
| Cysteine elimination            | 0.30 ± 0.02     | 40 ± 10         | (6.8 ± 5.8) · 10³ |
| Alanine racemisation (D→L)      | 0.106 ± 0.007   | 2,500 ± 700     | 41 ± 9           |
| Alanine racemisation (L→D)      | 0.0009 ± 0.0002 | 50 ± 80         | 20 ± 20          |
| Glutamate racemisation (D→L)    | 0.024 ± 0.001   | 30 ± 6          | 700 ± 200        |
Due to the temperature dependence of enzymatic properties and the thermophilic lifestyle of *T. maritima* (55–90°C, Huber *et al.*, 1986), the cystathionine elimination activity of *T. maritima* MetC was assayed across several temperatures (20–80°C). Tris pH 8.8 was substituted with CHES pH 8.8 as the pH of Tris buffer is much more sensitive to temperature (Sambrook and Russell, 2001a), furthermore, thermal shift assays (methods § 4.4.5) showed that the enzyme was much more stable in CHES (*T_M* = 85°C) than in Tris (*T_M* = 48°C, possibly for the same reason). Despite the fact that cystathionine elimination was chosen because it does not involve mesophilic coupled enzymes, at high temperatures Ellman’s reagent did show a high spontaneous cleavage rate (for example, 0.1 µM/s at 80°C without *T. maritima* MetC or cystathionine). Consequently, the values at higher

---

**Figure 4.6.** Michaelis–Menten plots of the various activities measured of *T. maritima* MetC. Bars represent with standard error. Red dots are values removed by the automatic outlier function in Prism 6.0.

A. cystathionine elimination

B. cysteine elimination

C. L→D alanine racemisation

D. D→L alanine racemisation

E. D→L glutamate racemisation
temperatures were lower than values predicted by the Arrhenius equation (discussed in § 4.3.6). Nevertheless, whereas the turnover increased with temperature, the Michaelis constant did not (Fig. 4.7).

![Figure 4.7](image_url)

**Figure 4.7.** The effect of temperature on the parameters of *T. maritima* MetC. The Michaelis constant and the turnover number at each temperature were determined by assaying the enzyme with eight concentrations of cystathionine (no replicates). At temperatures of 50°C and upwards, the uncatalysed rate was subtracted from the catalysed rates due to increased background cleavage. The predicted turnover number was calculated using the Arrhenius equation (§ 4.3.6).

### 4.2.6 Background control

In addition to standard negative controls without each of the components (MetC, coupled enzymes or NAD(H)), a further control was done. It was possible that the activities detected were from background contaminants due to the similarity between the Michaelis constant of the three enzymes with the dedicated enzymes in *E. coli* (MurI, Alr and MetC). As a sample composed solely of background would have these activities, a buffer exchanged *E. coli* lysate was assayed. The lysate was assayed up to 0.2 mg/ml (5 µM if it were pure MetC) in triplicate at multiple substrate concentrations, but no activity was detected with the five assays. Consequently, it could be concluded the activities detected for the three MetC enzymes were not a result of contaminants. This is consistent with the fact that only one band of a size expected for MetC was seen on SDS-PAGE gels for each sample (Fig. 4.2) and that these fractions absorbed at 420 nm due to the bound PLP.
4.2.7 Analysis of cystathionine elimination product by mass spectrometry

The elimination assay could not discern whether cystathionine was eliminated on the cysteine side (β-elimination) or the homocysteine side (γ-elimination). Both reactions are found among enzymes in the PF01053 family, often in closely related sequences (viz. chapter 2). Consequently, the products of the cystathionine elimination were analysed. In order to obtain the elimination products, *T. maritima* MetC was incubated with 1 mM cystathionine in ammonium-bicarbonate buffer (methods § 4.4.4). The incubation was brief (20 min) in order to minimise any degradation of the products by MetC. After the removal of the protein from the solution with an Amicon centrifugal filter unit, the free thiols were alkylated with iodo-acetamide and the buffer and water were subsequently removed by centrifugal evaporation.

The prepared samples were analysed by mass-spectrometry by Dr Torsten Kleffmann (Centre for Protein Research, University of Otago).

Analyses revealed that there was at least twelve times more homocysteine than cysteine produced (Fig. 4.8). This showed that *T. maritima* MetC was specific towards the β-elimination reaction.
Fig. 4.8 (A–F). Mass spectra of cystathionine elimination products of *T. maritima* MetC. The alkylated forms of the two possible elimination products (cysteine, γ-elimination product, structure in panel A, data in top half of figure; homocysteine, β-elimination, structure in D., data in bottom half of figure) could be detected but at different concentrations (mass spectra showing alkylated cysteine and alkylated homocysteine in panels B and E respectively; predicted spectra in panel C and F). Specifically, the peak for homocysteine had 12-fold more counts than that for cysteine (cf. panel B with E; NL: normalised intensity levels).
Fig. 4.8 (G–H). Mass spectra of cystathionine elimination products of *T. maritima* MetC. The identity of these could be further confirmed thanks to the presence of peaks from natural isotopes, e.g. $^{33}$S and $^{34}$S (panel G). The ratio of the area under the curves (AA) in the elution spectra between homocysteine and cysteine could not be calculated due to background interference in the cysteine peak (panels H).
Fig. 4.8 (I–J). Mass spectra of cystathionine elimination products of T. maritima MetC. The identity of these could be further confirmed thanks to the presence of peaks from natural isotopes, e.g. $^{33}$S and $^{34}$S (panel G). The ratio of the area under the curves (AA) in the elution spectra between homocysteine and cysteine could not be calculated due to background interference in the cysteine peak (panels H).

4.2.8 T. maritima MetC structure

The determination of the T. maritima MetC structure was done in collaboration with Dr Chris Squire and Yuliana Yosaatmadja from the University of Auckland. I purified the protein and performed two preliminary crystallisation trials, while Dr Squire and Ms. Yosaatmadja did two further trials and diffracted the obtained crystal at their facilities, which allowed them to resolve the structure at a resolution of 2.0 Å by molecular replacement with the E. coli MetC structure (Fig. 4.9). While I did not solve the structure, the analyses of the structure, presented here, are my own.
As expected from *E. coli* MetC data (Clausen et al., 1996), *T. maritima* MetC crystallised as a tetramer with 2-fold symmetry, where each monomer forms a Schiff base with a PLP molecule and forms most of the active site. The remainder of the active site is formed by a few residues from the N-terminus of the adjacent monomer that are involved in hydrogen bonding the phosphate and in forming part of the entrance to the active site.

In MetC structures, the monomer is composed of three domains (Fig. 4.10): (i) an N-terminal domain, which is predominantly helical with a disordered section, (ii) a central domain with an seven-fold α/β pattern forming a two-layered sandwich, and (iii) a C-terminal domain with a two-layered α/β sandwich (Clausen et al., 1996). The central domain and the C-terminus flank the PLP moiety and are connected by a long, kinked α-helix (helix 12) that spans the length of the monomer (Clausen et al., 1996). For consistency, the nomenclature of the structural elements in Clausen et al. (1996) is used herein.

**Figure 4.9.** *T. maritima* MetC crystal obtained by our collaborators (left) and the resolved structure of the tetrameric assembly of *T. maritima* at 2 Å resolution (right).
Figure 4.10. Domains of the *T. maritima* MetC monomer. The annotations of the features are as they appear in Clausen et al. (1996); specifically helices are numbered from the N to the C terminus, while sheets in the central domain are listed alphabetically in lower case in the N to C direction and sheets in the C-terminal domain are in upper case. Helix 2 is absent in *T. maritima* MetC, while helices 6 and 7 are conjoined.

4.2.9 Structural comparison (*T. maritima* MetC – *E. coli* MetC)

The structures of *T. maritima* MetC and *E. coli* MetC were compared. The alignment of the two structures showed few differences in most secondary structural elements. The three major differences in *T. maritima* MetC are increased oligomerisation contacts at beginning of the N-terminus, a loop that occludes part of entrance to the active site and a tryptophan close to the PLP.

The N-terminus is one of the segments involved in assembling the MetC tetramer. *T. maritima* MetC forms a dimer of dimers, where two monomers interact by completing the active site of the other due to α-helix 3 of the N-terminal domain, while this pair of monomers binds to another pair due to several surface residues. In particular, several residues preceding part of the loop before α-helix 3 (last: Y31) interact with the other dimer, while several residues of α-helix 3 and part of the adjacent loops (first: D29) interact with the adjacent monomer to complete its active site. The structures from α-helix 1 to α-
helix 2 between *T. maritima* MetC and *E. coli* MetC overlap poorly because α-helix 2 is absent in *T. maritima* MetC. Despite these differences there are a large number of interactions between the monomers in *T. maritima* MetC.

The α-helix 3 forms part of a large series of differences surrounding the entrance to the active site of the other monomer (Fig. 4.11). In *E. coli* MetC the entranceway of the active site starts wide and narrows towards the PLP, while in *T. maritima* MetC the entranceway starts narrow due to the inwards tilt of α-helix 3* (other monomer) and α-helix 14, as well as the different conformation adopted by α-helix 15 and its associated loop (from P341 to N354), which is oriented over part of what is the entrance in *E. coli* MetC.

Amongst the differences two residues stand out, namely D347 from the occluding loop and R43* from α-helix 3*. D347 protrudes across a space that in *E. coli* MetC is part of the active site entrance, yet in the crystal structure D347 does not hydrogen bond with residues from α-helix 6 on the other side, despite the closest residue being a lysine (K111). The other noteworthy residue is R43*: its side chain extends towards the loop and under

---

**Figure 4.11.** Difference in the entranceway. *E. coli* MetC monomers are in blue and lavender blue, while *T. maritima* MetC monomers are in red and lilac. The monomer with the relevant active site is in a primary colour, while the adjoining monomer in duller colour. The segments not involved in the differences are more transparent; this includes those in the foreground.
certain conformations could be within hydrogen bonding distance of the backbone keto groups of Y345, A344 and V343.

Furthermore, a small cavity above an active site arginine is absent in *T. maritima* MetC due to the bulky sidechains of Y345 (from the loop) and M153*.

As a consequence of these differences, the entrance to the active site is much smaller in *T. maritima* MetC (Fig. 4.12).

**Figure 4.12.** Difference in active site entrance. The *E. coli* MetC entrance (left) is funnel shaped, while the *T. maritima* MetC entrance (right) is much narrower.

The roles of these differences is unclear, but three features stand out. Firstly, the loop corresponding to α-helix 15 in *E. coli* MetC is not conserved and is unique to MetC from the *T. maritima* group (Fig. 4.13 inset). Secondly, the B-factors (*i.e.* a measure of disorder in the structure) of the terminal atoms of R43 and D347 are amongst the highest in the structure and the backbone atoms of part of the loop (Y345 to D352) have B-factors over 35 Å² (mean 24 Å²) (Fig. 4.13). Thirdly, the solvent exposed side of the loop is concave and could fit a molecule shorter than 7.5 Å with two partial positive charges on each end (pocket labelled in Fig. 4.12).
Although the entrances differ between \textit{E. coli} MetC and \textit{T. maritima} MetC, the internal active site differences are subtler. As expected, throughout the MetC tree, the active site is highly conserved (Fig. 4.14), consequently it is unsurprising that in both structures the PLP is held in place by several common residues.

Several residues are involved in shaping the active site. The PLP forms a Schiff base with a lysine (\textit{E. coli} MetC: K210; \textit{T. maritima} MetC: K202), $\pi$-stacks with a tyrosine (\textit{E. coli} MetC: Y111; \textit{T. maritima} MetC: Y106) and forms several hydrogen bonds to surrounding residues, especially via its $5'$-phosphate group (Fig. 4.15). When the catalytic lysine is not bound to PLP it hydrogen bonds to a serine (\textit{E. coli} MetC: S339, \textit{T. maritima} MetC: S331) and a tyrosine (\textit{E. coli} MetC: Y56*, \textit{T. maritima} MetC: Y51*). The $\pi$-stacking tyrosine hydrogen bonds to an arginine (\textit{E. coli} MetC: R58*, \textit{T. maritima} MetC: R53*) that also interacts with the phosphate.

\textbf{Figure 4.13}. The occluding loop. The loop is not rigid (B-factor colouring from blue (30 Å²) to red (60 Å²)) or conserved (inset: alignment of the loop and nearby residues of \textit{T. maritima} MetC and its closest homologues, the residues with numbers above alignment are also marked in the structure and the occluding aspartate is marked in bold).
Figure 4.14. Conservation of residues across the protein. Conserved residues are in blue, while unconserved residues are in red (and as thick lines). This was calculated with the Consurf server (§ 4.4.6). This reveals that the majority of the differences are on the outside of the protein and not in the active site, despite the kinetic differences between the various MetC enzymes discussed.

Despite the common residues the position of the PLP is not identical in the overlay, even when only either the PLP or a neighbouring residue is aligned with its counterpart in the other structure. In the T. maritima structure the PLP is shifted towards the π-stacking tyrosine (also shifted, albeit by only 0.6–0.9 Å), and is tilted 11° along the axis formed by the nitrogen heteroatom and the carbon in para position (C4, whose substituent forms the Schiff base) resulting in a 0.4 Å shift for the ring nitrogen from its E. coli MetC counterpart, while 1.0 Å for carbon 4 (Fig. 4.15).

Similar subtle differences are seen in the residues involved with substrate binding. The α-carboxyl group of the bound substrate is held in place by hydrogen bonds to an arginine (E. coli MetC: R372; T. maritima MetC: R357), to the nitrogen heteroatom of a tryptophan (E. coli MetC: W332; T. maritima MetC: W340) and to the secondary amine of the peptidic backbone of a serine (E. coli MetC: S339; T. maritima MetC: S331). Like the PLP and tyrosine, R357 in T. maritima MetC is shifted by 1 Å compared to R372 in
These minor changes might influence the angle of the α-carbon of the bound substrate, which might have an effect on racemisation efficiency.

Farther away from the PLP, some differences are present. The tunnel that leads to the active site is at an angle compared to the plane of the pyridine ring and differs in the *T. maritima* MetC structure in the presence of an extra pocket and the protrusion of W108.

In the *E. coli* MetC structure R59 (C-terminus of α-helix 2) has been implicated in stabilising the distal carboxyl group of an unbound cystathionine and E235 (α-helix 11) in stabilising the nearby amino group of the substrate (Clausen et al., 1996). Neither residue is conserved in *T. maritima* MetC, as R59 corresponds to G54, while E235 corresponds to V227. However, in *T. maritima* MetC two other residues, K231 and G45, have similar interactions to the *E. coli* MetC R59–E235 pair (Fig. 4.16). Specifically, the ε-amino group of the *T. maritima* MetC K231 occupies the same space as part of the guanidino group of the *E. coli* MetC R59, and hydrogen bonds instead to the backbone ketone of G45 (2.8 Å). In addition to the lack of conservation, the termini of *E. coli* MetC E235 and R59 side chains are 16.8 Å and 13.6 Å away, respectively, from the bridging nitrogen of the phospho-pyridoxyl–substrate aldimine. Cystathionine is much smaller than that (the amino groups of cystathionine are up to 8.6 Å apart). Therefore, these various differences appear to
be too distant to play the previously hypothesised role in stabilising the distal carboxylate of cystathionine. A more likely candidate to hydrogen bond with the terminal carboxyl of a bound cystathionine is the conserved arginine that also coordinates the PLP phosphate (*E. coli* MetC: R58*; *T. maritima* MetC: R53*).

![Fig. 4.16. R59–E235 interaction in *E. coli* MetC (blues) and G45–K235 in *T. maritima* MetC (reds). The monomer associated with the PLP moiety shown is in a primary colour, while the other monomer is in a duller colour. The hydrogen bond between the terminal amine of K231 and the keto backbone of G45 is in yellow. Residues R43*, D347, Y106 and W108 are illustrated for reference purposes and are described elsewhere.](image)

Possibly the most important difference in this side of the active site is the protrusion of a tryptophan (W108) in the *T. maritima* MetC structure, which is replaced by a proline (P103) in *E. coli* MetC. The indole nitrogen of W108 is 8.9 Å away from the catalytic aldimine group, which suggests that, under some conformations, the γ-carboxyl group of the bound substrate glutamate might hydrogen bond with the tryptophan (Fig. 4.17).

The equivalent residues to W108 are conserved within subgroups of the MetC tree (Fig. 4.18). In the group with *E. coli* and *P. ubique* the residue is a proline, while in *Wolbachia* it is an arginine, which is a flexible positively charged residue. In the ancestral sequence determined by maximum likelihood, the residue is a proline.
Figure 4.17 Potential glutamate position in the active site. Glutamate was docked manually into the cavity using the inhibitor bound *E. coli* MetC (PDB: 1cl2) structure as an additional guide.

Figure 4.18. Residues at positions equivalent to 108 across the MetC tree. The residues at the ancestral nodes were inferred by maximum likelihood (RAxML) using the tree and dataset obtained in chapter 2 (§ 2.2.9).
4.3 Discussion

4.3.1 Summary of results

In chapter 2, the knowledge that *E. coli* MetC can promiscuously catalyse the racemisation of alanine was utilised to search among sequenced genomes for organisms with *metC*, but without *alr*, the gene that encodes the enzyme that normally catalyses the racemisation of alanine. This led to the identification of *metC* homologues that have most likely displaced *alr* in three species—*Wolbachia*, *T. maritima* and *P. ubique*.

The bioinformatics analysis in chapter 2 suggested that in *T. maritima* and *P. ubique*, I hypothesised that the MetC enzyme is involved in one of two routes for methionine biosynthesis, which allow this amino acid to be produced either either hydrogen sulfide or cysteine as the source of sulfur (§ 2.3.2). Consequently, in *T. maritima* and in *P. ubique*, the MetC enzymes, in all likelihood, physiologically catalyse two different reactions, making them catalytically multitasking enzymes.

In this chapter, the *Wolbachia*, *P. ubique* and *T. maritima* MetC enzymes were characterised biochemically and the structure of one, *T. maritima* MetC, was analysed (after it was solved by our collaborators, Dr C. Squire and Y. Yosaatmadja). This not only strengthened the conclusions from chapter 2, that the three enzymes can catalyse both alanine racemisation and cystathionine elimination, but also enriched them by revealing the balances between the two and the presence of other activities that were not tested in chapter 2.

4.3.2 Glutamate racemisation

The three enzymes possess several differences and several similarities. In chapter 2, *murI*, the gene encoding glutamate racemase, was noted to be absent in *T. maritima* and *Wolbachia*. This raised the question of whether MetC is catalysing this reaction, but this question could not be tested by complementation due to a high reversion rate in the available *murI* mutant strain (*E. coli* WM335) and the unsuccessful attempts at constructing a better knockout (detailed in appendix V).

In this chapter, glutamate racemisation was assayed. Whereas *T. maritima* MetC and *Wolbachia* MetC could racemise glutamate, *E. coli* MetC and *P. ubique* MetC could not.
This is in agreement with chapter 2, where *E. coli* and *P. ubique* possess a gene encoding a glutamate racemase, while *T. maritima* and *Wolbachia* do not, and where the *metC* gene from *P. ubique* is more closely related to *E. coli metC* than to *Wolbachia metC* or *T. maritima metC*.

MurI, the enzyme that racemises glutamate in *E. coli* and in most bacteria, catalyses the reaction using acid–base catalysis performed by two cysteine residue on opposite sides of the α-carbon. These are aided on each side by a tethered aspartate or glutamate residue, which also orient water molecules to possibly stabilise the carbanionic intermediate (Hwang et al., 1999). This is in contrast to the alanine racemisation in Alr or MetC, where the acid–base catalysis is performed by tyrosine and lysine residues, while the stabilisation of the intermediate is done by PLP (Watanabe et al., 2002). Consequently, the ability of MetC to racemise glutamate is via a different mechanism than MurI and this is the first reported case of PLP-dependent glutamate racemisation.

In the case of *T. maritima* MetC, the ability to racemise glutamate may be due to the presence of a tryptophan residue (W108), which may stabilise the terminal carboxyl group of glutamate, as opposed to the proline that is present in *E. coli MetC* and *P. ubique MetC*. This untested hypothesis is based on crude manual placement of the glutamate in the active site: this could be improved by *in silico* docking or by determining the structure of *T. maritima MetC* in the presence of one of the two isomers of glutamate. While this is the best structural explanation, there are some minor issues with the hypothesis. In *Wolbachia* MetC the equivalent residue (R119) is an arginine, which is positively charged. *Wolbachia* MetC can catalyse glutamate racemisation suggesting that any residue able to hydrogen bond with the terminal carboxyl of glutamate might be sufficient to confer glutamate racemase activity.

A peculiar corollary of this hypothesis is presented by the fact that the enzyme encoded by the ancestral *metC* probably had a proline at that position (Fig. 4.17), presumably rendering it unable to racemise glutamate. Instead, the glutamate racemisation might have arisen fortuitously at least twice in the *metC* tree. Specifically, it occurred once (proline → arginine) in the *metC* of the ancestor of the *Anaplasmataceae* and once (proline → tryptophan) in a *metC* gene that may have been horizontally transferred to both the ancestor of the *Thermotoga* genus and the ancestor of the *Cytophagaceae* (a family in the
Further evidence, such as a *T. maritima* MetC W108F or W108P mutant, is required to fully support the role of tryptophan in glutamate racemisation.

### 4.3.3 Alanine racemisation

The major difference between the alanine racemisation activities in *P. ubique*, *Wolbachia* and *T. maritima* MetC, and that in *E. coli* MetC, is lower Michaelis constants for the former enzymes (discussed further in § 4.3.8). Comparing the structure of *T. maritima* MetC with that of *E. coli* MetC does not give a strong indication of why this may be. Most residues in the active site of *T. maritima* MetC and *E. coli* MetC are conserved and differ only in that the various residues do not align perfectly, resulting in a tilt of the PLP moiety in *T. maritima* MetC compared to *E. coli* MetC. Consequently, further experiments, such as in silico docking or diffraction of MetC crystals soaked in substrate, are required to determine the source of the difference. As alanine is a substrate, which might be racemised even in the crystals, 1-aminoethylphosphonate (Ala-P) could be used instead as it forms a stable aldimine that is not racemised or hydrolysed (Stamper et al., 1998). Furthermore, a structure with a bound substrate might not only reveal how the substrate binds, but might reveal if there are any conformational changes to the entranceway to allow substrates to bind.

While the ability to racemise glutamate in MetC may have arisen multiple times, it is likely that the ability to racemise alanine is ancestral, for various reasons. On one hand, the three enzymes assayed were chosen because they were from organisms without *alr*, but with peptidoglycan, and consequently are a biased representation of MetC diversity. On the other hand, the only other MetC that has been tested for alanine racemisation is that of *E. coli*, which possesses it as a promiscuous activity. The closest homologue that is known to lack this activity is *E. coli* MetB (Soo, 2012).

For racemisation to occur, the acid–base catalysis needs to be performed by residues on opposite faces of the PLP-bound amino acid; in the case of MetC, lysine (*T. maritima* MetC K202) abstracts the Ca proton from an L-amino acid, while tyrosine (Y106) abstracts the proton from D-amino acids. The phospho-pyridoxyl lysine is conserved throughout the whole PLP fold I family (CL0061) and the tyrosine is conserved throughout the Cys/Met metabolism subdivision (PF01053). Consequently, it is reasonable to assume that the ability to racemise alanine was present in the enzyme encoded by the ancestral *metC* gene.
Given that the other closely related homologues of *metC* encode enzymes involved in elimination or replacement reactions involving sulfur, and Alr is highly conserved, it seems likely that the hypothetical alanine racemisation activity in the ancestral MetC was a promiscuous activity, while cystathionine/cysteine β-elimination was the main activity.

### 4.3.4 Cysteine β-elimination

A further assay for an activity that could not be tested in chapter 2 was the β-elimination of cysteine. In chapter 2, no physiological role could be inferred for the cystathionine elimination activity of *Wolbachia* MetC, which is partially at odds with the fact that in this chapter its cystathionine elimination activity was found to have a very low Michaelis constant, albeit with a barely detectable turnover number. The most likely explanation involves the closely related cysteine elimination activity.

In *E. coli* MetC, cystathionine elimination is catalysed much more efficiently than cysteine elimination, whereas in *T. maritima* MetC and *P. ubique* MetC the catalytic efficiencies are nearly equal; in *Wolbachia* MetC cysteine elimination is ten-fold more efficient.

Cysteine elimination produces hydrogen sulfide, pyruvate and ammonia. Free hydrogen sulfide could be advantageous for certain reactions, in fact several distinct homologues of MetC are found fused to rhodanese-like domains (PF00581), which are generally sulfurtransferases (pfam.sanger.ac.uk/family/PF01053#tabview=tab1, Punta et al., 2012). However, the precise role of cystathionine or cysteine elimination in *Wolbachia* MetC cannot be ascertained nor can it be concluded whether cysteine elimination in *E. coli* MetC, *T. maritima* MetC and *P. ubique* MetC is an unavoidable side effect of cystathionine elimination due to the similarity of the two compounds.

### 4.3.5 Cystathionine elimination specificity

In chapter 2, it was found that the *metC* genes were unable to rescue a Δ*metB* auxotroph. This was supported in this chapter, where cystathionine was cleaved by β-elimination and not by γ-elimination (§ 4.2.7 and Fig. 4.4). Additionally, whereas cysteine was eliminated, methionine γ-elimination or acetyl-homoserine γ-replacement could not be detected (data not shown). Several studies have been done to determine the factors that favour β-elimination as opposed to γ-elimination (Farsi et al., 2009, Lodha et al., 2010, Aitken et al., 2011, Lodha and Aitken, 2011, Jaworski et al., 2012, Manders et al., 2013),
but the picture is incomplete. As sequence-based predictions of specificity cannot be made, the boundary between homologues with MetC activity and homologues with MetB activity cannot be determined bioinformatically (v. § 2.2.9). Nevertheless, the results from this chapter and chapter 2 move back the boundary between the MetC group and the MetB group in favour of MetC, by determining that the *T. maritima* homologue is a MetC.

### 4.3.6 Temperature dependence of enzyme kinetics

The values of the various activities reported in Tables 4.3, 4.4 and 4.5 were measured at 37°C for consistency, even though the kinetic parameters of enzymes are temperature dependent and enzymes from different organisms adapted to different temperatures have different behaviours (Somero, 1995). *P. ubique* is cultivated at 15°C (Rappe et al., 2002), *Drosophila melanogaster* (the host of *Wolbachia* wMel) can grow at room temperature (Trotta et al., 2006), while *T. maritima* has an optimal temperature of 80°C (Huber et al., 1986). Temperature affects the denaturation of the enzyme and also the turnover number, which within a certain range is dependent on the Arrhenius equation for a transition state, 

$$k_{cat} = k_B T \frac{e^{-\Delta G^*/RT}}{e^h}$$

where $\Delta G^*$ is the Gibbs free energy of the intermediate, $T$ the temperature, $k_b$ the Boltzmann constant, $h$ the Planck constant and $R$ the universal gas constant (Daniel et al., 2001). Due to the fact that the Michaelis constant ($K_M$) is the ratio of the substrate off-rate ($k_i$) plus the turnover rate ($k_{cat}$) over the substrate on-rate ($k_f$), the Michaelis constant is also affected by temperature. Consequently, the physiological values for *T. maritima* MetC are likely to be higher, while the physiological values of *P. ubique* MetC and *Wolbachia* MetC are likely to be lower. Data for *T. maritima* MetC suggest that the effect of turnover number on Michaelis constant is negligible, so only turnover number changes significantly (Fig. 4.3). This is consistent with a study of a set of enzymes with Michaelis constants around 1 mM, which showed only a minor change in Michaelis constants until the enzymes approached their melting temperature (Thomas and Scopes, 1998), so it may be similar for the other two MetC enzymes.

### 4.3.7 Balance of activities in *P. ubique* MetC

The balances between the activities of the three enzymes differ. In *P. ubique* MetC, cystathionine elimination had the highest turnover, but a similar catalytic efficiency to cysteine elimination, as the latter possessed a Michaelis constant and turnover number
that were both an order of magnitude lower. Alanine racemisation had a lower catalytic efficiency due to both a lower turnover and higher Michaelis constant, for both directions. The Michaelis constant was lower in the L $\rightarrow$ D direction, a trait shared by several Alr enzymes (Esaki and Walsh, 1986, Inagaki et al., 1986, Strych et al., 2001, Noda et al., 2004a, di Salvo et al., 2012).

Of the three enzymes, P. ubique MetC was the most problematic to assay. The enzyme precipitates when the N-terminal fusion partner (maltose binding protein) is cleaved off (le Roux, 2012), consequently it was assayed with an unavoidable 44 kDa fusion partner close to the entrance to the active site of each monomer, which most likely altered the kinetic parameters. One peculiarity was that the catalytic efficiencies for the two directions of alanine racemisation were not equal. The Haldane equilibrium for some alanine racemases in the literature differs from one and this has been attributed to the coupled enzyme assays (Noda et al., 2004b, Wu et al., 2008). In this case, the possibility that it may be artefactual due to the fusion partner cannot be excluded.

In hindsight, a more thermostable MetC from the Pelegibacter genus could have been characterised. Herein, the MetC that was characterised was from HTCC1062, the type-strain of P. ubique, which has an optimal temperature of 16°C. A better choice may have been the MetC from Pelagibacter sp. HTCC7211 (which will be shortly named Pelagibacter bermudensis), as this strain has an optimal temperature of 23°C (Dr J. Cameron Thrash, personal communication).

Due to this issue, the main focus of this study is T. maritima MetC. Nevertheless, the data from P. ubique MetC are informative for the purpose of comparisons, despite the potentially shifted parameters.

### 4.3.8 Balance of activities in Wolbachia MetC

As mentioned, no physiological role for cystathionine elimination could be ascribed in Wolbachia, therefore it is unsurprising that Wolbachia MetC possesses a much higher turnover for alanine racemisation than for cystathionine elimination, which is probably a vestigial activity that is neutrally maintained. This indicates that Wolbachia MetC differs from other MetC enzymes in that it is not a physiological cystathionine $\beta$-lyase. The highest turnover in Wolbachia MetC is for the two directions of alanine racemisation, followed by cysteine elimination and then by glutamate racemisation.
Whereas in the *P. ubique* MetC and *T. maritima* MetC enzymes, the efficiencies for the elimination of cystathionine and of cysteine are similar, in *Wolbachia*, MetC cysteine elimination is an order of magnitude more efficient. A further difference between *Wolbachia* MetC and the two other enzymes is the lower Michaelis constant for alanine racemisation in the D → L direction.

### 4.3.9 Balance of activities in *T. maritima* MetC

The MetC from *T. maritima* has a different balance of activities from the other two enzymes. Similar to *P. ubique* MetC, the elimination of cystathionine and of cysteine are the most efficient reactions, and the latter has turnover and Michaelis constants that are an order of magnitude lower than the former. However, the third most efficient reaction is glutamate racemisation, which is an order of magnitude lower than cysteine and cystathionine elimination, while alanine racemisation is even lower. Unlike for the other two enzymes, the Michaelis constant for glutamate racemisation is the lowest. In *T. maritima* MetC, there is a very pronounced difference between the two directions of alanine racemisation, which may be less sharp at physiological temperatures.

A unique feature of *T. maritima* MetC is the presence of a latch-like loop (345–352), which in the crystal structure was partially closing the entrance (Fig. 4.11). The high B-factors of the loop and the ability of aspartate residues to form salt bridges to arginine residues raises the possibility that this loop might form a flap that is closed in the crystal structure, but could be latched open due to a potential D347–R43* interaction. The enzyme was evolved to operate at high temperature (~80°C), which may allow this loop to frequently switch between the two configurations and the time spent in the closed configuration could potentially be modulated by a certain compound. To conclusively ascertain the role of this loop, future studies could probe its role on the kinetics by mutagenesis (e.g. D347E or R43E) or by molecular dynamics.

### 4.3.10 Physiological Michaelis constants

In general, the three MetC enzymes have poor catalytic efficiencies for their various main activities. It is likely that these enzymes have evolved to balance the various activities in such a way that the turnover numbers meet physiological requirements for the products, while keeping the Michaelis constants close to the physiological concentrations of the substrates.
Firstly, low turnover does not indicate that the activity is not physiological. While the catalytic efficiency of some enzymes is close to the diffusion limit of the substrate, this is rarely the case. In fact, the trends for the turnover number of enzymes studied to date differ depending on the cellular role, such as energy production (median $k_{cat} = 79$ s$^{-1}$), high-demand primary metabolism (18 s$^{-1}$), cofactor biosynthesis (5.2 s$^{-1}$) and secondary metabolism (2.5 s$^{-1}$) (Bar-Even et al., 2011). This indicates that not all enzymes are evolved to be maximally catalytic and are instead tailored towards physiological requirements (Bar-Even et al., 2011). Consequently, there is no such thing as a catalytic efficiency cut-off between main activities and promiscuous activities.

As mentioned in chapter 2, the demand for the products of the various activities is likely to differ in P. ubique, T. maritima and Wolbachia, and most likely be lower than in E. coli due to various reasons, such as slower growth rate (D-alanine in P. ubique), lower demand for peptidoglycan (D-alanine and D-glutamate in Wolbachia), different ratio of peptidoglycan components (D-alanine and D-glutamate in T. maritima) or pathway redundancy (homocysteine in P. ubique and T. maritima).

In contrast to turnover numbers, it has been suggested that Michaelis constants correlate with substrate concentration (Benner, 1989), and are often lower than the intercellular substrate concentration (Bennett et al., 2009). In E. coli there is 2.6 mM alanine, 92 mM glutamate and 0.37 mM homocysteine (cystathionine data not available) (Bennett et al., 2009). Assuming that the concentrations are similar in other bacteria, this indicates that despite the constraint resulting from the multiple activities, T. maritima MetC and Wolbachia MetC possess Michaelis constants for the various activities that are lower than the intracellular substrate concentrations. This may also hold true for P. ubique MetC, if the assumption that fusion partner is affecting the kinetic parameters is correct (§ 4.3.7).

This situation is in stark contrast to the promiscuous alanine racemisation activity seen in E. coli MetC, which has a higher turnover, but also a much higher Michaelis constant (4.7 mM for L-alanine and 51 mM for D-alanine). The Michaelis constants for promiscuous activities are varied depending on the relationship between the physiological substrate and the promiscuous one, but are generally much higher than those of the dedicated enzyme. For example the dedicated enzyme for the dephosphorylation of phosphoserine (SerB) has a Michaelis constant of 0.2 mM, while the Michaelis constants for the promiscuous
phosphoserine dephosphorylation activities in three other enzymes are 5 mM (Gph), 0.6 mM (HisB) and 8 mM (YtjC) (Yip and Matsumura, 2013). In several cases of promiscuous activities, the Michaelis constant is larger than the solubility of the substrate (Aharoni et al., 2004).

Consequently, it is likely that the biochemical difference between main activities and promiscuous activities is that the former are evolved to meet both the physiological requirements and the physiological concentrations of the substrate, while the latter are not. Because turnover numbers of physiological activities can be low if the product is not required in abundance, the Michaelis constants are better suited for probing whether an activity is physiological or not.

4.3.11 Interplay of activities

Mutants of E. coli MetC with an improved alanine racemisation activity have a decreased cystathionine β-elimination activity (Soo, 2012). This indicates that there probably is an unavoidable trade-off in activity between the two activities and that consequently it may not be possible to perfectly balance low Michaelis constants (in order to meet the intracellular substrate concentration) and the appropriate turnover numbers (in order to meet the demand of products) on the MetC scaffold. Therefore, if such a constraint were in place, the activities would be dramatically altered in mutant MetC enzymes obtained by directed evolution with a selection for only one of the activities. Alternatively, the constraint may be negligible and the other activities may be robust against change.

To investigate the presence of a constraint and to tease apart the various activities, in chapter 5 the T. maritima metC was evolved to encode an enzyme with an improved cystathionine β-elimination activity.
4.4 Methods

Materials are present in appendix I, including strain genotypes (§ I.I.11) and plasmids (I.I.III) used.

4.4.1 Expression of the MetC enzymes

For cell culture protocol see appendix § I.I.1. For electroporation see § I.II.11. Expression of the genes was done using the recommended expression strain. The pBAD plasmids bearing the metC genes from Wolbachia (construct without KpnI, § 2.4.3) and T. maritima described in chapter 2 (§ 2.4.3) were used to transformed E. coli strain LMG194 (described in Guzman et al., 1995, distributed by Invitrogen, San Diego, CA); while the pMALc5X/metCpub–malE.E.co plasmid was transformed into strain ER2523 (NEB, Ipswich, MA).

Due to the different yields, generally four 2 ℓ baffled flasks with 1 ℓ LB broth each were used for Wolbachia MetC expression (2 mg yield), 1 ℓ LB broth was used for T. maritima MetC expression (4 mg yield) and 500 mℓ LB broth in a 1 ℓ baffled flask was used for P. ubique MetC expression (over 5 mg yield).

Cells were grown in lysogeny broth (LB) (Sambrook and Russell, 2001b)) at 37°C until OD₆₀₀ ≅ 0.6. Afterwards the cultures were induced with either 200 µg/mℓ arabinose (T. maritima and Wolbachia) or 300 µM IPTG (P. ubique). Additionally, 10 µg/mℓ pyridoxine was added because this PLP precursor may potentially bolster available PLP. The cultures were then incubated at either 28°C (T. maritima and Wolbachia) or 18°C (P. ubique) overnight.

To optimise expression of Wolbachia MetC, different chaperones were tested, namely stress-response chaperone induction with 3% v/v EtOH (Thomas and Baneyx, 1997) or over-expression with the chaperone plasmids (ClonTech®, TaKaRa Bio Inc., Shiga, Japan).

4.4.2 Purification of the MetC enzymes

The cells were harvested and lysed with the same method (§ I.II.1v), and purified depending on the tag present. T. maritima MetC and Wolbachia MetC were purified by metal affinity chromatography (§ I.II.v), while P. ubique MetC was purified by amylose affinity chromatography (§ I.II.vi). Whereas the T. maritima MetC and Wolbachia MetC
samples were cleaned of imidazole by three or more iterations of concentration with an Amicon centrifugal concentrator, *P. ubique* MetC could not be concentrated by centrifuge or by dialysis against hygroscopic PEG10000 as it precipitated.

The MetC-bearing fractions were lemon yellow due to the bound PLP and were partially verified by *SDS-PAGE* (protocol in § I.II.V) due to the presence of a band of the expected size (~45 kDa for *Wolbachia*, *E. coli* and *T. maritima* MetC, while ~90 kDa for *P. ubique* MetC–MBP fusion).

The concentration was determined by absorbance at 280 nm (§ I.II.VI).

For the background control experiment, a dilution of the lysate from 20 ml of LMG194 in exponential phase was equilibrated against the protein buffer with a 50-kDa Amicon® centrifugal concentrator.

### 4.4.3 Spectrophotometric assays

The general assay protocol and analysis is described in appendix § I.II.VII, while specific reactions are described in the following sections:

- § I.II.VIII for cystathionine elimination
- § I.II.IX for cysteine elimination
- § I.II.X for alanine racemisation from L-alanine
- § I.II.XI for alanine racemisation from D-alanine
- § I.II.XII for glutamate racemisation from D-glutamate

In the case of *Wolbachia* and *T. maritima* MetC the assays were done in biological triplicates, while for *E. coli* MetC and *P. ubique* MetC a single biological set was used. Ten or more concentrations of substrate were tested in two to four technical replicates per biological replicate depending on the consistency of the results. The concentration of enzyme used varied between 0.1 µM and 0.5 µM depending on the strength of the activity.

For the temperature-dependent assays of cystathionine elimination by *T. maritima* MetC, assays were done at 20, 30, 40, 50, 60, 70 and 80°C according to the protocol § I.II.VIII, but with some differences. Firstly, 50 mM Na·CHES buffer pH 8.8 instead of Tris·Cl. Secondly, each slope of Ellman’s reagent cleavage over time of samples with enzymes was subtracted by the slope of samples without enzyme at the same concentration of
substrate due to the chromophore's spontaneous concentration-dependent cleavage. The predicted \( k_{cat} \) was calculated with the formula \( k_{cat} = k_B \frac{T}{b} e^{-\Delta G^\ddagger/RT} \) (Daniel et al., 2001), where the \( \Delta G^\ddagger \) was calculated from the \( k_{cat} \) at 40°C and the physical constants (Boltzmann constant \( (k_B) \), Planck constant \( (h) \) and gas constant \( (R) \) were taken from Voet and Voet (2011).

### 4.4.4 MS of cystathionine elimination product

To determine whether cystathionine elimination reaction was a \( \beta \)-elimination or a \( \gamma \)-elimination, a 200 \( \mu \ell \) reaction with 50 nM pre-equilibrated \( T. maritima \) MetC and 1 mM cystathionine · (HCl)_2 in 50 mM ammonium-bicarbonate buffer was left at 37°C for 20 min, before being brought up to 6 m\ell with water and separated from the protein with a prewashed 10 kDa cut-off Amicon® centrifugal filter unit. To distinguish between trace carryover of Tris base and cysteine (both 121 Da), The flow-through was concentrated with a Speedvac and the pellet was resuspended in 5 \( \mu \ell \) with 20 nmol iodoacetamide and incubated at 4°C for 30 min to derivatise the free thiols, before being resuspended in 2 m\ell and re-concentrated. Dr Torsten Kleffmann (Centre for Protein Research, University of Otago) analysed the sample by reverse-phase LC/MS using a C18 column that was eluted isocratically to the emitter tip of the mass spectrometer, which was in positive ion mode.

### 4.4.5 Thermal shift assay

Thermal shift assays were preformed with a LightCycler®480 (Roche Applied Biosciences, Penzberg, Germany) in a 96 well plate. Each well contained a 100 \( \mu \ell \) 1:1 mix of 100 \( \mu \ell \) \( T. maritima \) MetC stock and 5× SYPRO® Orange (Invitrogen, San Diego, CA) master stock and a series of conditions from a library prepared by Sylvia Luckner.

The LightCycler®480 software was used with a 1°C/30 s ramp from 30 to 90°C. The exported data was retabulated in Excel 2008 and the \( T_m \) determined by the temperature at which the upwards linear slope reaches its midpoint.

### 4.4.6 Determination of \( T. maritima \) MetC structure

Dr Chris Squire from the the University of Auckland crystallised \( T. maritima \) MetC by seeding with smaller crystals in a hanging drop suspension containing 1.2 mg/mL \( T. maritima \) MetC in 0.8M ammonium citrate, 2% PEG 8000, 0.1M Tris · HCl pH 8.5. The space group was orthorhombic (I222) and was resolved by molecular replacement with
1CL1 as a template at a resolution of 2.04–89.19 Å with an R-factor of 0.175 and an R-free of 0.208.

Structures were visualised in PyMol (PyMOL Molecular Graphics System, Schrödinger, LLC) and annotated in Photoshop CS4 (Adobe Systems, San Jose, CA) if necessary.

The alignment of the *T. maritima* MetC and the *E. coli* MetC structure was done by aligning only the chain of interest as the oligomerisation differs slightly due to differences at the N-terminus (*viz.* 4.2.10).

The MetC alignments and trees obtained in chapter 2 (Fig. 2.9) were used to map conservation onto the structure of *T. maritima* MetC thanks to the Consurf server (consurf.tau.ac.il, Ashkenazy *et al.*, 2010). The ancestral sequence was reconstructed in RAxML 7.7.8 using the marginal ancestral states reconstruction feature (-f A) with a WAG+Γ + F model (Stamatakis, 2006).
Chapter 5

Directed Evolution of
*T. maritima* MetC


5.1 Introduction

5.1.1 Evolvability of T. maritima MetC

In chapter 4, the biochemistry of the Wolbachia, Pelagibacter ubique and Thermotoga maritima MetC enzymes was explored, revealing different balances of their activities. There is a strong cystathionine elimination activity in P. ubique and T. maritima MetC and a weak one in Wolbachia MetC, while Wolbachia and T. maritima MetC can racemise glutamate, unlike P. ubique MetC. The ability to catalyse two different types of reaction, a phenomenon that was termed catalytic multitasking in chapter 1 (§ 1.6), is rare as most enzymes catalyse a single reaction on a single substrate to avoid having conflicting biochemical requirements that may otherwise arise in the shared active site. One example discussed in chapter 1 (§ 1.7) was the strong switch in specificity from tyrosine to phenylalanine seen with a H89F mutant of tyrosine ammonia-lyase (Watts et al., 2006).

In chapter 4 (§ 4.3.12), it was hypothesised that multitasking in these MetC enzymes was achieved due to the low requirement for products, so the turnovers could be low; while Michaelis constants had to be similar to the physiological concentrations of substrate. As a result, the Michaelis constants are comparable to those of the enzymes in other organisms that perform a single reaction, while the turnover numbers are much lower. Nevertheless, this does not address the question of how the biochemical requirements for one activity affect the others. Consequently, this chapter will attempt to address how these activities behave when T. maritima metC is evolved.

5.1.2 Broad substrate ambiguity and primordiality

As mentioned in chapter 1 (§ 1.7), it has been conjectured that primordial enzymes must have performed a plethora of reactions on a broad range of substrates in order for metabolic networks to be patched together (Jensen, 1976). Primordial enzymes were presumably very rudimentary and possibly even constructed from fewer than twenty amino acids (Müller et al., 2013); consequently a large number of differences are present between primordial enzymes and contemporary enzymes, but one key difference that is of interest here is the breadth of reactions that they could catalyse.
The current models that are used as a comparison to infer the properties of primordial enzymes are the generalist intermediates from directed evolution studies (§ 1.7). Specifically, when a gene is evolved to encode an enzyme that catalyses a different reaction, in many cases an initial weak trade-off is seen with the ancestral activity.

This was seen in a study where three genes were evolved for new hydrolytic functions and the resulting mutants encoded enzymes that had small losses in their ancestral activity, but large gains in the evolved hydrolytic activity (i.e. weak trade-off) (Aharoni et al., 2005). Consequently, the mutants are generalist intermediates as they can perform both the ancestral and the evolved activity. These enzymes also possessed other promiscuous activities that were not put under selection, but that displayed large fluctuations, either increases or decreases (Aharoni et al., 2005). Therefore, these unselected activities were highly plastic in contrast to the ancestral activity, which was robust to change. In general, as these changes are the result of biochemical differences in the enzymes, the activities for chemically similar compounds vary together, such as the fortuitous shifts in preference in promiscuous substrates seen with a P450 peroxygenase drifting under purifying selection (Cirino and Arnold, 2003, Bloom et al., 2007).

Additionally, generalist intermediates often can act on a large breadth of substrates, often unseen in the ancestor. For example, the generalist intermediates encountered during the evolution of a gene encoding a β-glucuronidase into one encoding a β-galactosidase were found to be able to better act on a broad variety of substrates. This included one substrate that could not be recognised by the ancestor (Matsumura and Ellington, 2001). Upon further evolution, the breadth of activities was decreased in favour of high specificity for the activity under selection; the mutants were specialists again (Matsumura and Ellington, 2001). This has been seen even when evolution was towards a smaller substrate. For example, the directed evolution of a P450 propane monooxygenase from a long-chain fatty acid hydroxylase progressed from a specialist, through an intermediate that possessed a broad substrate ambiguity for most alkanes and even terpenes, before losing this in favour of a higher catalytic efficiency for the selected substrate, propane (Fasan et al., 2008).

The breadth of accepted substrates and the non-specific nature of these generalist intermediates led to the hypothesis that these enzymes may be similar to primordial enzymes. This, however, raises the question of how do multitasking enzymes compare. Multitasking enzymes can also catalyse multiple reactions, but, unlike generalist
intermediates, they are shaped by multiple evolutionary pressures. It is likely that primordial enzymes were subjected to several evolutionary pressures and consequently may be more similar to multitasking enzymes than generalist intermediates.

In the case of *T. maritima* MetC, the enzyme can catalyse reactions on alanine, cysteine, glutamate and cystathionine, which share only the core amino acid structure, yet are tightly bound by the enzyme. In all likelihood this enzyme can catalyse other reactions on a broad range of similar compounds that are not physiologically relevant, such as d-cysteine elimination, aspartate racemisation and in the case of serine potentially a mix of elimination and racemisation. One activity of the MetC enzymes for which no physiological role could be ascribed was cysteine elimination. *P. ubique*, *Wolbachia* and *T. maritima* MetC have a cysteine elimination activity with a low Michaelis constant and a catalytic efficiency par (in the case of *P. ubique* and *T. maritima* MetC) or better (*Wolbachia* MetC) than the cystathionine elimination activity. This is in contrast to *E. coli* MetC where cystathionine is eliminated three orders of magnitude more efficiently than cysteine.

### 5.1.3 Mutagenesis

In this chapter, directed evolution was used to improve the cystathionine elimination activity of *T. maritima* MetC. The aim was to address the question of how the various activities change when one activity is under a strong evolutionary selection. I hypothesised that this would allow the inference of which activities are robust to mutations when another is mutated and which activities are poorly robust.

### 5.1.4 Methodologies and requirements for error-prone PCR

From a technical point of view, there are several ways to introduce errors into a sequence via PCR. One is to decrease fidelity of the polymerase by using manganese instead of magnesium as a cofactor (El-Deiry *et al.*, 1984), another is to use a nucleotide mix with mismatch-inducing nucleotide analogues, such as 8-oxo-2′-deoxyguanosine (Zaccolo *et al.*, 1996), and another still is to use an error-prone DNA polymerase, such as Mutazyme® (Hanson-Manful and Patrick, 2013). Each of these methods has its advantages and disadvantages; in this study an error-prone DNA polymerase was chosen as it does not suffer as strongly from substitution biases compared to other methods (Hanson-Manful and Patrick, 2013).
For a directed evolution experiment to yield mutants with an improved activity (“winners”), several requirements need to be met. Firstly, the selection system needs to be stringent: there should be a large margin for improvement in order to differentiate winners from wild-type. Secondly, the library needs to be diverse, which reflects both the average number of mutations per clone and the number of clones in the library. Regarding the number of mutations, on one hand more mutations result in larger numbers of unique mutants, but on the other the number of viable mutants decreases exponentially as the number of mutations per gene increases. For example, with PSE-4 β-lactamase the fraction of active enzymes decreased by 46% per mutation (Drummond et al., 2005). Consequently, a library with 4–5 mutations per gene is ideal (Hanson-Manful and Patrick, 2013).

5.1.5 Effect of directed evolution on *T. maritima* MetC

By performing error-prone PCR on *T. maritima metC* and selecting for mutants with altered activities, the evolutionary behaviour of this catalytic multitasking enzyme can be gleaned. This question is of interest as the presence of multiple physiological activities is a property potentially shared by primordial enzymes. Consequently, evolution of *T. maritima* MetC may provide a further insight into the nature of primordial enzymes, including novel properties that differ from those of generalist intermediates, the current comparative model.
5.2 Results

5.2.1 Library construction and analysis of diversity

The *T. maritima* metC gene was mutated with an error-prone polymerase (for methods see § 5.4.1). It was amplified with Mutazyme® II and then cloned back into the pBAD vector. The purified ligation mixture was used for transformations. The *E. coli* strains transformed were: (i) *E. coli* DH5α, in order to analyse the diversity (*vide infra*); (ii) *E. coli* MB2795, in order to select for mutants with improved alanine racemisation activity (§ 5.2.2); and (iii) *E. coli* JW2975, in order to select for mutants with improved cystathionine elimination activity (§ 5.2.3).

To estimate the library diversity, 20 clones were checked (methods § 5.4.2). Several mutants were found (Fig. 5.1), including a sequence with six mutations and sequence with the same six mutations plus one extra, which distorted the average mutation frequency. In fact, without imposing any assumptions on the distribution of the mutations, an average of 1.85 ± 2.7 mutations per sequence was found. While this is the standard approach, the accuracy of this value can be improved by fitting the distribution to a Poisson distribution, which is an acceptable approximation of a PCR-distribution (Sun, 1995) especially in light of the limited sampling. Under this assumption, the mean was found to be 0.9 ± 0.1 mutations per gene. This low rate is most likely to be due the fact that the proportion of sequences that were amplified during each PCR cycle (PCR efficiency) was low (10%). A mutation jackpot is the phenomenon where one or more mutations arise early during the amplification stage and are propagated to a large number of clones, therefore skewing the estimated mutational frequency (Pope et al., 2008); the sampling of both a sextuple and a septuple mutant here is such a case. Several studies have nevertheless had successful outcomes with similar mutational frequencies (Soo, 2012).

Different polymerases have specific error rates depending on the buffer, for example *Taq* polymerase has a reported error rate varying from $1.2 \cdot 10^{-4}$ to $2 \cdot 10^{-5}$ mutations per base copied (Eckert and Kunkel, 1990). In this study, the error prone PCR resulted in an average 0.9 mutations across 1.1 kb and 4 doublings on average (3.6 µg product from 0.2 µg template), which means that Mutazyme® II had an error rate of $2 \cdot 10^{-4}$ mutations per base, which is comparable to the lower values of wild-type *Taq* polymerase. The low
mutagenicity of Mutazyme® is a known issue and in one study a second round of error-prone PCR was performed with *Taq* polymerase and MnCl$_2$ to increase the mutagenicity (Patrick and Matsumura, 2008).

Figure 5.1. Distribution of the number of mutations present in the sampled population.

Error-prone PCR via Mutazyme polymerase was used (§ 5.1.4) as it one of the least biased methods, which, due to the nature of the genetic code, is important to achieve a broad range of mutations that are not biased towards certain amino acids (Hanson-Manful and Patrick, 2013). Consequently, the types of mutation were tallied and used to verify that there were no unexpectedly strong biases (Table 5.1). The values reported for the enzyme in the Stratagene GeneMorph® II manual (www.chem-agilent.com/pdf/strata/200550.pdf) and those found in Hanson-Manful and Patrick (2013) are similar to those that were found here. While there was some variation in the ratio of transitions to transversions, this was nevertheless much lower than the bias seen in epPCR with manganese ions (Shafikhani *et al.*, 2007). Consequently, no unexpected biases were detected.
Table 5.1 Mutational frequencies in the sampled population compared to others in the literature. The Stratagene values were copied from the Stratagene GeneMorph® II manual. *cynT* values are from Hanson-Manful and Patrick (2013), where a gene, *cynT*, was mutated with the Stratagene GeneMorph® II kit. Ts: transition, Tv: transversions.

<table>
<thead>
<tr>
<th>Bias indicators (ratios)</th>
<th>This study</th>
<th>Stratagene values</th>
<th><em>cynT</em> values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ts/Tv</td>
<td>1.3</td>
<td>0.9</td>
<td>0.86</td>
</tr>
<tr>
<td>AT→GC/GC→AT</td>
<td>0.9</td>
<td>0.6</td>
<td>1</td>
</tr>
<tr>
<td>A→N, T→N</td>
<td>60%</td>
<td>51%</td>
<td>57.3%</td>
</tr>
<tr>
<td>G→N,C→N</td>
<td>40%</td>
<td>44%</td>
<td>40.3%</td>
</tr>
<tr>
<td>Total</td>
<td>57%</td>
<td>43%</td>
<td>45.2%</td>
</tr>
<tr>
<td>A→G, T→C</td>
<td>30%</td>
<td>18%</td>
<td>22.6%</td>
</tr>
<tr>
<td>G→A, C→T</td>
<td>28%</td>
<td>26%</td>
<td>22.6%</td>
</tr>
<tr>
<td>Total</td>
<td>43%</td>
<td>51%</td>
<td>52.4%</td>
</tr>
<tr>
<td>A→T, T→A</td>
<td>25%</td>
<td>29%</td>
<td>27.4%</td>
</tr>
<tr>
<td>A→C, T→G</td>
<td>5%</td>
<td>5%</td>
<td>7.3%</td>
</tr>
<tr>
<td>G→C, C→G</td>
<td>3%</td>
<td>4%</td>
<td>3.2%</td>
</tr>
<tr>
<td>G→T, C→A</td>
<td>10%</td>
<td>14%</td>
<td>14.5%</td>
</tr>
</tbody>
</table>

As several nucleotide mutations result in a synonymous codon, the proportion of amino acid mutations is expected to be lower. This was estimated using the online tool PEDEL-AA (Firth and Patrick, 2008). The rate of 0.9 nucleotide mutations per gene was equivalent to 0.6 amino acid mutations per peptide; this meant that 52% of the library was not mutated (wild-type), which makes the library less than ideal.

5.2.2 Attempts at racemisation selection

Wild-type *T. maritima metC* rescues *E. coli* MB2795 (*alr::frt dadX::frt*) more slowly that when the strain is supplemented with D-alanine on LB agar. When plated on LB agar from the same batch and in equal amounts, the latter grows colonies overnight and to sizes similar to other strains, while the former provided colonies that are half as large and fewer in number. This difference, however, was not sufficient for selection. Rescue experiments with the mutant pBAD/metC<sub>Mar</sub> pool were done under several conditions (methods § 5.4.3); these were different temperatures (28°C and 37°C), different numbers of cells plated (10<sup>8</sup>–10<sup>10</sup> colony forming units, CFU) and inducer concentrations (130 µM, 13 µM and 1.3 µM arabinose; the latter value was used for subsequent experiments). None of these conditions gave positive results and instead either a slow-
growing lawn or nothing was seen. In many cases larger colonies were seen, but bore wild-type genes and were artefacts resulting from an issue where colony size varied depending on the proximity to other colonies.

To increase the selective pressure, growth on L-alanine as the sole carbon source was attempted (rationale discussed in § 5.3.1). Simply growing the *E. coli* MB2795 strain on media with M9 salts and alanine as a carbon source was not possible as it is a derivative of the MC1000 strain, which is reported to be deficient in thiamine and leucine biosynthesis (Strych *et al.*, 2001). Despite this, MB2795 did not grow on D-alanine–containing minimal medium even when these supplements were added (data not shown). However, supplementation of LB-agar with thiamine, leucine and uridine greatly diminished the issue of colony size variation due to nearby colonies.

The *E. coli* MB2795 strain did grow on minimal medium in the presence of 0.05% or more casamino acids and therefore the selective minimal media was made to include 0.05% w/v casamino acids, 50 µg/ml leucine, 20 µg/ml uridine and 10 µg/ml thiamine. However, a further peculiarity of the strain made the selection process problematic, namely all plates had fewer surviving cells than expected, with the exception of cases where the number of cells was very high, possibly as this would allow the cells to gain D-alanine from neighbouring dead cells. Consequently, on several selection plates, depending on the number of cells plated, there would either be a very small number of colonies, or a lawn. Twelve colonies were tested and were found to contain wild-type *T. maritima* metC or metC mutants with synonymous mutations; these clones lost their advantage when retransformed.

Therefore, no successful selection of improved variants was obtained due to the low fitness of *E. coli* MB2795.

### 5.2.3 Cystathionine elimination selection

_**For**_ the transformation of *E. coli* JW2975 (metC::kan) with the mutant pool, 2 · 10⁶ CFU (LB agar control plate with ampicillin, for methods see § 5.4.4) were plated in total on five square dishes. In light of the estimated average number of nucleotide mutations (0.9 mutations per variant), this library in terms of amino acid sequences can be
expected to be composed of (PEDAL-AA with a PCR distribution, § 5.4.2):

- $1.0 \cdot 10^6$ wild-type sequences
- $5.6 \cdot 10^5$ single non-synonymous mutants, 2518 of which are unique out of a potential 7201 amino acids variants (the remainder are only accessible via more than one mutation per codon), which translates to 34% coverage of the single mutants
- $2.2 \cdot 10^5$ double mutants, $2.0 \cdot 10^5$ of which are unique for a 0.8% coverage
- $6.9 \cdot 10^4$ triple mutants (only one in a million coverage)
- $1.9 \cdot 10^4$ quadruple mutants
- $4.5 \cdot 10^3$ quintuple mutants
- fewer than a thousand mutants with more than five substitutions

To monitor whether a colony observed on the library plate was larger than expected, a reference Petri dish with *E. coli* JW2975 expressing wild-type *T. maritima* metC was grown in parallel. Several colonies from the library grew faster than the wild-type on the control plate. When replated, 19 of these colonies grew faster than wild-type cells, while another 22 did not (methods § 4.4.5). However, 13 proved to contain no mutations in the *T. maritima* metC gene and did not retain their advantage when the plasmid was used to retransform a fresh aliquot of *E. coli* JW2975. One winner could not be sequenced (*i.e.* the plasmid appeared to lack the priming site), while two winners from two separate plates contained the same insert: the *E. coli* malY gene (function discussed in § 5.3.4). In these winners the *malY* gene inserted itself upstream of *metC*, which was frameshifted; the site of the insertion was a stretch of six bases that are identical in both *T. maritima* metC and *E. coli* malY and that are flanked at one side by two mismatches and three other identical bases. One winner grew faster than wild-type by only one–two days and possessed an A131V mutation, which is not close to the active site. Two winners rescued the auxotroph within a week (as opposed to ten days) and both possessed S86T/S305C mutations.

### 5.2.4 Kinetics of the S86T/S305C mutant

The best mutant for cystathionine elimination in the library therefore appeared to be *T. maritima* MetC with the S86T/S305C mutations. The S86T/S305C mutant was expressed and purified in the same manner as the wild-type *T. maritima* MetC, resulting in
a similar yield (lane F in Fig. 4.2, methods § I.II.IV). The enzyme was assayed for cystathionine and cysteine elimination, alanine racemisation (D→L direction only) and glutamate racemisation in the same way as the wild-type (Table 5.3, Fig. 5.4, methods in § 5.4.6).


The simplest and most robust way to test for statistical significance between the each value of the wild-type against the mutant is to ascertain whether the 95% confidence intervals overlap. They do not for the Michaelis constant for cysteine cleavage, the turnover number for alanine racemisation, the turnover number for the glutamate racemisation and the catalytic efficiency of cysteine, implying significant differences. Conversely, an overlap between standard errors implies a lack of statistical significance. There is overlap for the
Michaelis constants for cystathionine cleavage and glutamate racemisation and for the turnover numbers of cysteine cleavage.

<table>
<thead>
<tr>
<th>Table 5.2. Kinetics of <em>T. maritima</em> MetC S86T/S305C.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wild-type</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Cystathionine elimination</td>
</tr>
<tr>
<td>Cysteine elimination</td>
</tr>
<tr>
<td>Alanine racemisation (<strong>D</strong>→<strong>L</strong>)</td>
</tr>
<tr>
<td>Glutamate racemisation (<strong>D</strong>→<strong>L</strong>)</td>
</tr>
</tbody>
</table>

The different activities varied (Table 5.3) and three changes occurred. Firstly, there was an increase in the turnover numbers for both elimination reactions; secondly, there was a reduction in Michaelis constant for all activities, especially for the smaller substrates (cysteine and alanine) and, thirdly, there was a decrease in turnover numbers for the two racemisation reactions.

<table>
<thead>
<tr>
<th>Table 5.3. Differences between the mutant and wild-type.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ratio (mutant/wild-type)</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Cystathionine elimination</td>
</tr>
<tr>
<td>Cysteine elimination</td>
</tr>
<tr>
<td>Alanine racemisation (<strong>D</strong>→<strong>L</strong>)</td>
</tr>
<tr>
<td>Glutamate racemisation (<strong>D</strong>→<strong>L</strong>)</td>
</tr>
</tbody>
</table>
5.2.5 Structural predictions for the S86T/S305C mutant

Serine 305 is on β-sheet B of the C-terminal domain and is 11 Å away from the nearest cysteine (i.e. too far to form a disulfide) (Fig. 5.3). A serine to a cysteine is a fairly conservative mutation as cysteine differs from serine by having a sulfur as opposed to an oxygen, making it sterically similar, but more hydrophobic (2.5 versus −0.8 on the hydrophathicity scale from Kyte and Doolittle (1982)). The hydroxyl group of S305 forms two hydrogen bonds. One is with the side chain amide of N280, which is on β-sheet A and partially forms the exterior of the protein. The other bond is with the backbone keto group of M153, which is on a loop surrounding the active site. The lack of this hydrogen bond in the S86T/S305C mutant may shift the position of the loop. Neither the loop nor β-sheet B contain any residues involved in substrate binding, but the S-methyl group of M153 is 3.5 Å away from Y345, which is a residue from the occluding loop discussed in chapter 4.

Figure 5.3. Location of S305 (left) and S86 (right) residues in the structure of T. maritima MetC. In the main figure, S305 interacts with M153, while S86 interacts with several residue of helix 5. Residues from the other monomer are in dark green. The loop that occludes part of the active site is marked as a grey outline as it is directly above M153. The location of D347 from the loop is marked with an asterisk. In the insets are the expected effects of the mutations with the red disks representing potential steric clashes (PyMol mutagenesis).

In the E. coli crystal structure, the residue equivalent to S305, S312, is conserved and interacts with the backbone of an isoleucine (159), the side chain of which is oriented slightly differently to its equivalent (M153) in T. maritima MetC. Whereas the α, β, γ carbons closely overlay, the terminal δ-carbon is rotated compared to the sulfur of
methionine. This leaves instead a cavity above R372, which hydrogen-bonds with the \( \alpha \)-

The other mutated serine (86) is on \( \alpha \)-helix 5 and is at the dimerisation interface close to the opening of the active site (Fig. 5.3). Threonine differs from serine by having a methyl group on the \( \beta \)-carbon in addition to a hydroxyl group. If the threonine of the S86T mutation adopts the rotamer with the hydroxyl in the same location (Fig. 5.3 inset), it continues to hydrogen bond to nearby residues of the \( \alpha \)-helix, therefore maintaining cohesion, whereas the methyl group disrupts the keto group of the peptide backbone of N232* (other monomer). N232* is the penultimate residue in the C-terminus of \( \alpha \)-helix 11*, which plays a role in shaping the active site in both \textit{E. coli} (Clausen et al., 1996) and in \textit{T. maritima} MetC (§ 4.2.9), with the difference that in the latter enzyme K231* blocks off the side of the entranceway by interacting with the backbone of G45* (Fig. 4.12). At least three possible effects of the steric clash from the methyl group can be envisioned. One possibility is that \( \alpha \)-helix 5 is shifted slightly, including G81 and M82, whose backbone amides hydrogen bond to the phosphate of the PLP. A second possibility is that \( \alpha \)-helix 11* is shifted and affects the nearby residue W108, which may be involved in glutamate racemisation. A third option is that a shift in \( \alpha \)-helix 11* also shifts, via the K231*–G45* interaction, the position of R43*, which may be involved with the occluding loop. However, the structural perturbations were not large enough to affect the expression and purification of the protein, which had a yield similar to the wild-type enzyme.

The probability that a mutant with a single mutation, such as S86T or S305C individually, was not sampled is virtually nil (§ 5.2.3). Therefore, these mutations are likely to work synergistically even though they are on opposite sides of the active site (Fig. 5.3).

5.2.6 Structural predictions based on the kinetics of the mutant

The S86T and S305C mutations produce three changes in kinetics (§ 5.2.4). Firstly, the turnover numbers for the racemisation reactions halved, while the turnover for the elimination reactions increased by 13%. Racemisation in the \( \text{D} \) to \( \text{L} \) direction occurs when Y106 deprotonates the \( \text{D} \)-amino acid and K202 reprotonates the intermediate, while elimination of an \( \text{L} \)-amino acid occurs when K202 deprotonates the bound substrate. When K202 is not bound to lysine, it hydrogen bonds with S331* and Y51*. The latter residue is also one of the residues that hydrogen bonds to the phosphate of PLP (§ 2.2.10) and is part
of the same loop as G45* and R43*. However, neither of these two residues is close to either the S86T or the S205C substitutions. In the case of Y106, its hydroxyl group hydrogen bonds to R53*, which is hydrogen bonded additionally to the phosphate of PLP and the keto backbone of K231*, both of which are elements of the enzyme that were speculated to be affected by the S86T mutation (§ 5.2.5). Therefore, there is no obvious structural interpretation for the effects of the mutations on racemisation.

A third kinetic change was the decrease in Michaelis constants, especially for the smaller substrates. In chapter 4, a loop that partially occluded the entrance to the active sites due to D347, was described. This loop was hypothesised to be able to adopt a “latched open” conformation that removed D347 from the entrance of the active site, therefore expanding it, thanks to a hypothetical D347–R43* salt bridge (§ 4.3.10). As both mutations (i.e. S86T and S305C) may be required for the improved activity, it is possible that they improve the accessibility of the active site by affecting the propensity of the loop to adopt this open conformation.
5.3 Discussion

5.3.1 Selection systems

In chapters 2 and 4 it was established that *T. maritima* MetC is able to catalyse the elimination of cystathionine and of cysteine, and the racemisation of alanine and of glutamate at physiological concentrations of substrates. Here the question asked was how these activities change when the enzyme is mutated to improve one function. To investigate this question, directed evolution by error-prone PCR was performed.

In order for a directed evolution experiment to be successful the selection system has to be stringent and the library sufficiently diverse.

Of the four activities, cysteine elimination is the one that it is least plausible to construct a selection system for. This is because the *E. coli* strain (JM39) without the various genes known to be able to eliminate cysteine in *E. coli* (*tnaA*, *metC*, *cysK*, *cysM*, *malY*) is still able to eliminate cysteine, albeit threefold less rapidly (Awano *et al.*, 2003, Awano *et al.*, 2005). Consequently, no screen for cysteine elimination can be easily devised.

Another activity for which no selection system is available was glutamate racemisation. The D-glutamate auxotroph, *E. coli* WM335, possesses a deactivating point mutation in glutamate racemase (*murI*) and two mutations in the glutamate transporter (*gltS*) to enable uptake of D-glutamate. Due to the point mutation in *murI*, revertants are too frequent for selection experiments to be successful and the construction of a knockout with a more suitable genotype was unsuccessful as mentioned previously (§ 4.3.2).

In the case of the alanine racemase activity, the selection system proved more taxing than expected, due to the low fitness of *E. coli* MB2795. Despite the fact that the *T. maritima* MetC enzyme has a turnover number approximately 500-fold smaller than the *E. coli* Alr enzyme (Soo, 2012), the difference in growth rate between *E. coli* MB2795 over-expressing *T. maritima* MetC and *E. coli* MB2795 over-expressing *E. coli* Alr was insufficient to accurately discriminate between winners and wild-type. To overcome this, I attempted to devise a selection system based on alanine catabolism. Alanine catabolism occurs via the racemisation of alanine by the *dadX*-encoded racemase and the subsequent oxidative demamination of D-alanine by D-amino acid dehydrogenase (*dadA*-encoded) (Franklin and Venables, 1976). However, I was unable to grow *E. coli* MB2795 on minimal
media supplemented with several nutrients (leucine, uridine and thiamine) that were expected to overcome various auxotrophies. Instead, I could only grow the strain in the presence of at least 0.05% casein-derived amino acids. This culture condition increased the inconsistencies due to colony separation seen with MB2795. Consequently, for these reasons, no winners were found. The only way to overcome these obstacles will be to construct a healthier \( E.\ coli \) \( D\)-alanine auxotrophic strain (e.g. BW25113 derived) that is able to grow on minimal media supplemented with \( D\)-alanine.

Unlike the alanine racemisation selection, the selection for a better cystathionine elimination was robust due to a large window for selection, namely at a high-induction (200 ng/\( \text{mL} \) arabinose), \( T.\ maritima \) \( \text{metC} \) took two weeks to rescue the \( E.\ coli \) JW2975 strain (\( \text{metC::kan} \)). In fact, the sole downside to the lengthy growth time was fungal contamination arising from the humidifying measures taken to prevent the plates from shrinking.

5.3.2 Mutational load

In error-prone PCR the number of amplifications achieved dictates the mutational frequency. In the library I constructed, the PCR efficiency was low (10%) and consequently the number of mutations was low (0.9 per gene). As a consequence only half of the library contained mutated variants.

Nevertheless, one winner was found (twice) amongst the 500,000 unique variants that were subjected to selection for improved cystathionine elimination. This clone possessed two non-synonymous mutations that corresponded to S86T and S305C substitutions in the encoded enzyme. Ideally, several winners would have endowed a richer picture thanks to commonalities and differences, but one winner can still provide a wealth of information.

Conversely, information can also been gleaned from the false positives. The number of plasmids screened means that unlikely events may be seen, such as the insertion of the gene \( \text{malY} \) into the pBAD plasmid. The \( \text{malY} \) was of \( E.\ coli \) (K-12 or B strains) origin and may have been inserted into the pBAD plasmid either at the epPCR step from contamination in the pBAD/\( \text{metC}_{Tma} \) template or at the screening step via recombination.

\( \text{MalY} \) is a homologue of \( \text{MetC} \) and can both regulate the \( \text{mal} \) operon and eliminate cystathionine, which is believed to be a promiscuous activity, while its physiological substrate is unknown (Clausen \textit{et al.}, 2000). Patrick \textit{et al.} (2007) found that overexpressed
malY was able to rescue the metC knockout strain as quickly as metC. The malY gene is actually a distant homologue of metC and is more similar to aminotransferases than the enzymes that act on thiol groups (data not shown). This suggests a case of convergent evolution rather than an ancestral activity.

In addition to malY, several false positives were found, all of which contained wild-type *T. maritima* metC and lost their advantage when the gene was retransformed. Therefore a chromosomal event occurred to give the advantage.

Due to the layout of the genetic code, not all substitutions are equally likely, but most substitutions requiring a single mutation conserve some physico-chemical property (Pelc, 1965), consequently the 35% coverage of all possible single non-synomous substitutions is heavily biased towards substitutions that differ subtly in certain physio-chemical properties as opposed to disruptive substitutions. Consequently, a high proportion of subtle amino acid single substitution mutations have been explored in the screen, undoubtedly including residues involved in the occluding loop, such as R43, W108, K111, Y345 and D347 (§4.2.9). Nevertheless, the mutant found may affect the loop via more subtle changes.

### 5.3.3 The S86T/S305C mutant

The winner found had two mutations, both of which were conservative. One was a serine to a cysteine (S305C), a mutation that removes the hydrogen bonding potential of the hydroxyl group of the serine. The other mutation was a serine to a threonine (S86T), a mutation that conserves the hydrogen bonding of the hydroxyl group, but that sterically affects a nearby residue due to the extra methyl group. Statistically each of the two mutations were inevitably present multiple times, yet were not found in separate mutants by themselves, which indicates that the mutations are likely to work together in reshaping the enzyme. Testing this hypothesis in the future will be straightforward, because a *Hind*III site is present in the nucleotide sequence between the two mutations that would make it easy to construct plasmids bearing the mutations separately.

Curiously, both residues are conserved as serine in *E. coli* MetC, which indicates that natural evolution for the homologues that are cystathionine β-lyase specialists did not progress down the path taken by *T. maritima* MetC S86T S305C.

Several possible effects of these mutations can be speculated upon (§5.2.5), but the picture is greatly improved when the kinetic changes are considered (§5.2.6).
5.3.4 Activity balance: overall gains in activities

As expected in a winner of a selection experiment, the cystathionine elimination had an increased turnover by +13% (significance not determined) and possibly a decreased Michaelis constant (−22%, not significant). In addition to changes in the cystathionine elimination activity, changes were seen for all activities. On one hand, in terms of turnover numbers there was a slight increase for the elimination reactions (13%) at the expense of the turnover numbers for the racemisation reactions, which halved, i.e. a trade-off in terms of turnover numbers. On the other hand, all the Michaelis constants decreased, especially for the smaller substrates, i.e. no trade-off.

To simplify comparisons, turnover numbers and Michaelis constants are frequently reported as their ratio, i.e. the catalytic efficiency. Despite the simplification, this value can be misleading as enzymes with profoundly different kinetics can have similar catalytic efficiencies (Eisenthal et al., 2007). This is the case here, where the decrease in turnover numbers of the racemisation reactions is partially masked by the decrease in Michaelis constants. This results in an increase in catalytic efficiency for all reactions, except for glutamate racemisation. The largest gain in terms of catalytic efficiency was not cystathionine elimination (47% increase), the activity under selection, but cysteine elimination, which increased by 3.5-fold (250% increase).

It is unlikely that cysteine β-elimination is beneficial for methionine biosynthesis. Cysteine elimination is only required for methionine biosynthesis by organisms that synthesise homocysteine directly from sulfide. However, this route is not known to exist in E. coli. If it did, one would predict that overexpressing a known cysteine β-lyase, such as tnaA, would rescue E. coli JW2975 (metC::kan); however, this has not been seen (Patrick et al., 2007). Consequently, the increase in cysteine elimination was not a direct result of selection, but was a biochemical consequence of the increase in cystathionine elimination, which did confer a selective advantage.

5.3.5 Activity balance: deceased racemisation turnover

Overall there were several gains in the activity of T. maritima MetC S86T/S305C, but not all parameters improved. The turnover numbers of both racemase activities
in the D to L direction halved (significantly), a much larger change that the upward change in cystathionine elimination.

Elimination reactions of L-amino acids require the substrate’s deprotonation by K202, while racemisation of D-amino acids requires their deprotonation by Y106 and subsequent protonation by K202. One possible explanation of the changes is that S86T may cause an indirect shift of several residues, such as W108, R53* or K231*; of these, R53* hydrogen bonds to K202 and to the phosphate of PLP. Consequently, a slight shift in position could potentially change the pKᵢ of K202 and therefore improve its ability to deprotonate L-amino acids and decrease its ability to reprotonate ketimine intermediates. In order to definitely determine the structural changes, the \textit{T. maritima} MetC S86T/S305C mutant would have to be crystallised.

### 5.3.6 Activity balance: decreased Michaelis constants

One unexpected change in the S86T/S305C mutant was the decrease of all the Michaelis constants. Of the three physiological activities of wild-type \textit{T. maritima} MetC, cystathionine elimination is the one with the largest turnover number and, due to its low Michaelis constant, the catalytic efficiency was also the largest. Consequently, I was expecting an increase in specificity at the cost of the other activities. This is especially true because the intracellular concentrations of the other substrates are higher and they will act as competitive inhibitors of the cystathionine elimination activity.

Instead the Michaelis constants decreased for all compounds, especially alanine and cysteine (§ 5.3.4). These two compounds share the commonality of being small, which could be indicative that the active site has become more accessible. One possible explanation could be that the S305C and the S86T substitutions shifted other structural elements towards the occluding-loop, which may more frequently adopt a hypothetical latched-open conformation. One way to test this would be to crystallise the mutant.

The most clear example of the permissive nature of the mutant comes from cysteine elimination. This is a potentially promiscuous or even possibly deleterious activity and is much larger in \textit{T. maritima} MetC than in \textit{E. coli} MetC, relative to the equivalent cystathionine elimination activities. Cysteine is hyposteric to cystathionine and consequently the two activities are linked, but the difference in activities may indicate that \textit{T. maritima} MetC is more accommodating than \textit{E. coli} MetC, possibly as a side effect of the broadened
specificity of *T. maritima* MetC for multiple substrates. This potentially promiscuous activity increases in the mutant to the point that it is 3-fold larger in terms of catalytic efficiency than the cystathionine elimination activity (Table 5.2).

### 5.3.7 *T. maritima* MetC S86T/S305C in a physiological perspective

With the exception of the decrease in the turnover numbers for the racemisation activities, all parameters tested improved. However, this does not mean that *T. maritima* MetC S86T/S305C is a better enzyme than wild-type *T. maritima* MetC in its physiological context. The enzyme was adapted for *T. maritima*, an extreme thermophile, not for *E. coli*, a mesophile. The high melting temperature of *T. maritima* MetC was therefore not under selection and may have dropped in the mutant. This remains to be determined experimentally. Alternatively, the non-specific nature of *T. maritima* MetC S86T/S305C may catalyse reactions that are deleterious because they are unnecessarily wasting cellular metabolites.

### 5.3.8 Trade-offs

Several studies have investigated the evolution of new functions from promiscuous activities. The transition from one specialist to another specialist can either have a weak negative trade-off or a strong negative trade-off. The former means that a large increase in the catalytic efficiency for the promiscuous activity under selection comes at a small cost to the native activity, therefore forming a generalist. The latter means that in some cases the increase of a new activity amounts to the abolition of the native activity with a single mutation (Khersonsky and Tawfik, 2010b).

In the case of weak trade-off, the intermediate is a generalist that is able to catalyse both reactions as a result of a loss of specificity towards its ancestral activity (§ 5.1.2). Often an additional effect of this non-specific nature is that the catalysis of activities that were not under selection increases, including substrates that were not recognised by the ancestor, as was seen in Matsumura and Ellington (2001). Upon further mutation, the generalist acquires specificity for the substrate of the activity under selection, at the cost of the other activities.

Due to its breadth of activities and the relative strength of its cystathionine elimination activity, *T. maritima* MetC was expected to behave like a generalist
intermediate and lose its broad specificity. Instead, it behaved like a specialist and progressed through a generalist intermediate with an even more permissive active site.

Consequently, as wild-type *T. maritima* MetC is a multitasking enzyme that is able to catalyse several reactions, it differs from generalist intermediates in terms of evolution, at least in this mutant. Wild-type *T. maritima* MetC is not non-specific, but possesses a constrained balance of specificity towards three physiological substrates (alanine, glutamate and cystathionine).

In chapter 6, the properties gleaned from the evolution of *T. maritima* MetC are extrapolated in order to conjecture on how primordial enzymes may have evolved.
5.4 Methods

Materials are present in appendix I, including strain genotypes (§ I.I.I), plasmids (I.I.I.III) and primers (§ I.I.III) used.

5.4.1 Error-prone PCR

Error-prone PCR was performed using Mutazyme® II (GeneMorph® II kit, Stratagene, La Jolla, CA) according to manufacturer’s instructions, using the *T. maritima* MetC specific primers (Tma_metC_kpnI_F, Tma_metC_xbaI_R, § I.I.II) and pBAD/metC* as a template.

The conditions used were those recommended by the manufacturer to achieve a low to medium mutation rate (1–5 mutations per gene).

Thermocycling condition:

<table>
<thead>
<tr>
<th>Step</th>
<th>Temperature</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial denaturation</td>
<td>95°C</td>
<td>120</td>
</tr>
<tr>
<td>Denaturing step</td>
<td>95°C</td>
<td>30</td>
</tr>
<tr>
<td>Annealing step</td>
<td>62°C</td>
<td>30</td>
</tr>
<tr>
<td>Extension step</td>
<td>72°C</td>
<td>70 (1.5 kb)</td>
</tr>
<tr>
<td>Final extension step</td>
<td>72°C</td>
<td>10</td>
</tr>
</tbody>
</table>

50 µl Reaction mix:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>nuclease-free water (Ambion®)</td>
<td>35.5 µl</td>
</tr>
<tr>
<td>10× Mutazyme buffer</td>
<td>5 µl</td>
</tr>
<tr>
<td>40 mM dNTP stock</td>
<td>1 µl</td>
</tr>
<tr>
<td>10 µM primer mix</td>
<td>2.5 µl</td>
</tr>
<tr>
<td>200 ng equivalent of metC<em>Tma</em></td>
<td>5 µl</td>
</tr>
<tr>
<td>Mutazyme® II polymerase</td>
<td>1 µl</td>
</tr>
</tbody>
</table>

*) Only a portion of pBAD/metC*Tma* acted as a template. The plasmid itself was at 180 ng/µl, which corresponds to 40 ng/µl of the amplified region if this were free from the plasmid backbone.

After the confirmation by gel electrophoresis (§ I.I.I in appendix I), the amplicon was purified (§ I.I.I.III). The cleaned product and the pBAD bearing wild-type *T. maritima* metC were digested to become the insert and vector, respectively, for a sticky-ended ligation (§ I.I.I.III). The mutated amplicon was digested in 50 µl reactions (§ I.I.I.III) with *KpnI, XbaI*
and *Dpn*I — the first two to generate sticky ends, while the latter to degrade the Dam-methylated template — while the plasmid was digested with *Kpn*I, *Xba*I and *Hind*III — the latter to fragment the *T. maritima metC* to further reduce the risk of carry over of template. After a 3-hour incubation at 37°C and heat-killing at 60°C for 20 min, the digests were run on a SYBR® Safe 1% agarose gel and 1.5 kb insert and the 4.5 kb vector bands were excised from the corresponding digests. The DNA was purified out of the agar slabs (§ I.II.III) and ligated overnight at 22°C with 1 U/20 µl Fermentas T4 ligase (§ I.II.III) with a two to one insert to vector molar ratio (250 ng vector to 150 ng insert in a 100 µl reaction mix).

### 5.4.2 Library construction

The plasmid pool was used to transform (~10 ng/µl) competent *E. coli* DH5α, JW2975 or MB2795 cells (§ I.II.II). The cells were made fresh and electroporated with 2-mm–gapped electroporation cuvette as normal (§ I.II.II). Forty-eight 50 µl aliquots of *E. coli* JW2975 (for five square dishes, *vide infra*) and seventeen 50 µl aliquots of *E. coli* MB2795 (for seven square dishes and twelve petri dishes) were used in total.

The pool was transformed into *E. coli* DH5α to form a test library, in order to determine the mutational load.

The test library was plated on LB with 100 µg/ml ampicillin. The colonies were picked and sent for sequencing in both directions (primers pBAD_F, pBAD_R, § I.IV). The sequences were checked in MacVector 12.1 (MacVector Inc, Cary, NC). The data was analysed in Excel 2008 (Microsoft, Redmond, WA) and the distribution fitted to a Poisson in Prism 4.1 (GraphPad, La Jolla, CA). PCR efficiency was obtained from the yield, cycle number and template used, namely

\[
\text{efficiency} = 2^{\frac{\log(yield/tape\text{mple})}{\text{cycles}}} - 1 \quad \text{(Firth and Patrick, 2005)}
\]

The library diversity was calculate using PEDEL-AA (guinevere.otago.ac.nz/cgi-bin/ae/pea/pe/AA.pl, Firth and Patrick, 2008).

The selection experiments were performed on square bioassay dishes (245 mm × 245 mm, Corning Inc., Corning, NY), each of which contain 200–250 ml agar and are eight-fold larger in terms of surface than a Petri dish.
5.4.3 *E. coli* MB2795 rescue

For the selection with *E. coli* MB2795 (Δ*alr ΔdadX*), a variety of cell concentrations and media compositions were used. The majority of media were tested in Petri dishes with $10^5$ CFU *E. coli* MB2795 (determined by OD$_{600}$). All plates contained 1.5% agar, 100 µg/ml ampicillin and 200 ng/ml arabinose (0.00002% w/v = 200 ppb w/v; 1.3 µM), while the carbon sources and amino acid stocks varied (*vide infra*). After 1 h recovery (500 ml SOC media per 50 µl cell aliquot), cells were washed before plating (further detail in appendix I at § I.II.1). To minimise colony size differences due to proximity to other colonies, carbenicillin was substituted for ampicillin, and 50 µg/ml leucine, 20 µg/ml uridine and 10 µg/ml thiamine were added to the media. With the exception of plates with 1× LB, all plates contained 1× M9 salts (§ I.II.1 for recipes). Different dilutions of LB or casamino acids were tested. In addition to these dilutions, some plates had an extra carbon source, either 0.4% glucose or 0.2% L-alanine. The final plates contained: 1.5% agar, 0.2% L-alanine, 1× M9, 0.05% w/v casamino acids, 200 ng/ml arabinose, 100 µg/ml ampicillin, 50 µg/ml leucine, 20 µg/ml uridine and 10 µg/ml thiamine.

5.4.4 *E. coli* JW2975 rescue

For *E. coli* JW2975 (*metC::kan*) upwards of $2 \cdot 10^{10}$ CFU (determined by OD$_{600}$) were plated on media containing 1.5% agar, M9 salts, 0.4% w/v glucose, 20 µg/ml arabinose (0.002% w/v = 20 ppm w/v; 130 µM), 30 µg/ml kanamycin, 100 µg/ml carbenicillin.

5.4.5 Mutant verification

Winners were replated and if the colonies still grew faster than wild-type, the plasmid was purified and sequenced (Otago Genetic Analysis Service). The true winners were then retransformed. For probability calculations, Excel 2008 (Microsoft, Redmond, WA) was used and in the case of geometric distribution the negative binomial distribution function (*negbinomdist*) was used as the former is a special case of the latter when the number of successes is 1.

5.4.6 Assays of *T. maritima* MetC S86T/S305C

The *T. maritima* MetC S86T/S305C mutant was assayed biochemically similarly to chapter 4. The protocols for expression (§ I.II.IV–I.II.V), for purification (§ I.II.V–
I.II.VI) and for the four assays performed were as described previously (§ I.II.VII–I.II.XII).
Chapter 6

Concluding Remarks
6.1 Cellular roles of MetC

Most enzymes are highly specialised in that they have evolved to perform a single reaction on a single substrate. This, however, has not always been the case. Primordial enzymes have been hypothesised to have been able to catalyse a large breadth of reactions, allowing primordial metabolic pathways to assemble (introduced in § 1.1). Some modern enzymes are capable of performing the same reaction on a range of substrates (broad specificity) and a few enzymes are capable of performing more than one type of reaction (catalytic multitasking, introduced in § 1.6). Characterised members of the latter group are rare, but in this thesis three such enzymes were identified and studied in order to investigate whether they could shed light on the characteristics of primordial enzymes.

I discovered that in three distinct clades — the genus *Pelagibacter*, the genus *Thermotoga* and the family *Anaplasmataceae* — one enzyme (MetC) has taken over the role of the absent alanine racemase (Alr, discussed in § 2.3.4). A MetC from each group was tested: these were the enzymes from *Pelagibacter ubique* HTCC1062, *Thermotoga maritima* MSB8 and Wolbachia from *Drosophila melanogaster* (analysed in § 2.2.2). These MetC enzymes were shown to possess both alanine racemisation activity and cystathionine elimination activity (analysed in § 2.2.3). Furthermore, the respective genomes lacked strong candidates for dedicated alanine racemase enzymes (analysed in § 2.2.3–5). In contrast, the sister taxa that possessed *alr* lacked *metC*, which is consistent with the fact that the two genes perform the same role, albeit with additional roles in the case of *metC* (discussed in § 2.3.4). Moreover, two of the three MetC enzymes have taken over the role of the absent glutamate racemase (MurI, discussed in § 4.3.2), which, along with alanine racemase, provides a D-amino acid for peptidoglycan biosynthesis.

In *Escherichia coli*, MetC catalyses only the β-elimination of cystathionine, a step in methionine biosynthesis (introduced in § 2.1.2). The *E. coli* MetC has a small promiscuous activity for alanine racemisation, but no detectable activity for glutamate racemisation (analysed in § 4.2.2). The three enzymes investigated are located in different groups along the MetC tree and possessed different properties (§ 4.3.8–10). The enzyme most similar to *E. coli* MetC was the MetC from *P. ubique*, which could eliminate cystathionine and racemise alanine, but could not racemise glutamate (§ 4.3.8). This is consistent with the fact that the genome of *P. ubique* possesses *racX*, a homologue of the gene that normally
encodes glutamate racemase (\textit{murI}, § 2.2.5). Further away on the tree, \textit{Wolbachia} MetC had a glutamate racemisation activity and an alanine racemisation activity that were both stronger than its cystathionine $\beta$-elimination activity in terms of turnover numbers (§ 4.3.10). In a basal group, \textit{T. maritima} MetC possessed the strongest glutamate racemisation activity of the enzymes studied (§ 2.3.10). The low cystathionine elimination in \textit{Wolbachia} MetC was expected, due to the fact that this activity is vestigial and no other methionine biosynthetic genes were present in \textit{Wolbachia} (§ 2.3.2). This situation was different in \textit{P. ubique} and \textit{T. maritima}.

In \textit{E. coli}, MetB produces cystathionine, the substrate for the elimination reaction, from \textit{O}-succinyl-homoserine and cysteine (§ 2.1.2). In \textit{P. ubique} and \textit{T. maritima}, metB is absent and instead \textit{metY} is present (§ 2.3.2). MetY synthesises homocysteine, the same product as the MetC-catalysed elimination reaction, directly from \textit{O}-acetyl-homoserine and free sulfide without the need of MetB or MetC. However, it is likely that in \textit{P. ubique} and \textit{T. maritima} MetY is producing cystathionine (from cysteine) in addition to homocysteine (from free sulfide) in light of the following three observations. Firstly, a closely related enzyme, MetI from \textit{Bacillus subtilis}, can catalyse the replacement of the acetyl group of \textit{O}-acetyl-homoserine with either free sulfide or cysteine. Secondly, several MetY homologues can utilise methanethiol in addition to free sulfide. Thirdly, a high number of sequenced species possess \textit{metY} and \textit{metC}, without \textit{metB}. This hypothesis remains to be tested, but if this dual specificity of MetY is correct, there are two routes of methionine biosynthesis starting from different metabolites (cysteine or sulfide), one of which involves MetC. This would confer a metabolic flexibility depending on availability of sulfide (discussed in § 2.3.2).

### 6.2 Presence of multitasking enzymes in streamlined genomes

The common feature between \textit{P. ubique}, \textit{Wolbachia} and \textit{T. maritima} is the possession of genomes that have been under selection to possess fewer genes. I predicted that in order to reduce genome size, several specialized enzymes were lost and in their place multitasking enzymes arose. This scenario is in contrast to the norm, where evolution favours duplication and subfunctionalisation over the maintenance of a single enzyme, the activity and regulation of which are under two or more selective pressures (§ 2.3.5).
A related question regards the evolutionary potential of organisms with streamlined genomes. Increasingly, it is being realised that prokaryotic genome evolution is likely to involve prolonged periods of genome simplification and streamlining, interspersed with bursts of genome complexification, in which innovations arise (Wolf and Koonin, 2013). However, the extent to which an evolutionary lineage can undergo genome streamlining and still be able to revert to a complexification mode is unclear. In fact, genome reduction is often favoured by an unchanging environment and therefore it can be see as a form of specialisation. Cope's law of unspecialised descent states that generalists are more likely than specialists to diverge into new species (Cope, 1896). Consequently, one might expect that organisms with a strongly reduced genome will be an evolutionary dead-end.

In chapter 3, the phylogeny of the Alphaproteobacteria was addressed. The genomes of members of the Pelagibacteriales, Rickettsiales and Holosporales, and the mitochondrial organelles, have all undergone genome reduction (§ 3.1.3, 3.3.4). The phylogenetic positions of these orders would suggest that the ancestor of the Rickettsiidae and the Caulobacteridae may have had a streamlined genome, a phenomenon that was maintained in various lineages but that was reversed in the crown group of the Caulobacteridae, which reacquired several genes, including murI, by horizontal gene transfer. If this hypothesis is correct, the streamlined genome of the Rickettsiidae–Caulobacteridae ancestor shows that strong streamlining is not an evolutionary dead-end, although it appears to be hard to reverse based on the fact that only one lineage (the clade of Caulbacteridae orders without Holosporales) has given rise to extant species with large genomes.

In primordial protocells, before the advent of chromosomes, there would have been a very strong pressure to possess a small amount of genes (§ 1.1), which means that the evolution of streamlined organisms can, to some degree, shed some light on the evolutionary processes shaping primordial genomes.

### 6.3 Catalytic promiscuity vs. catalytic multitasking

The most striking difference between the P. ubique, Wolbachia and T. maritima MetC enzymes and E. coli MetC is that they have Michaelis constants ($K_M$) for the various substrates that are comparable to those of the dedicated enzymes (i.e. MetC, Alr and MurI) in other organisms, while having turnover numbers ($k_{cat}$) that are much lower than those of the dedicated enzymes. I conjectured that this is due to the physiological
necessity to operate with low concentrations of each substrate, in order to produce a small amount of products to meet the low cellular demand for these metabolites (§ 4.3.11).

MetC is not homologous to either Alr or MurI, but relies on the cofactor pyridoxal-5'-phosphate (PLP) for catalysis, similar to Alr. However, MurI relies on a different mechanism to MetC and my data are the first reported case of PLP-dependent glutamate racemisation (§ 4.3.2). The crystal structure of *T. maritima* MetC suggests that a tryptophan residue (W108) may hydrogen bond with the terminal carboxyl group of glutamate. This tryptophan residue is an arginine in Wolbachia MetC (which can also racemise glutamate), while it is a proline in *P. ubique* MetC and *E. coli* MetC (which cannot racemise glutamate). Unlike proline, tryptophan and arginine have a partial positive charge that enables non-covalent interactions to form with the terminal carboxyl group (§ 4.3.2).

Unexpectedly, a disordered loop close to entrance was found in the structure of *T. maritima* MetC. This loop might be able to adopt a different conformation to that present in the crystal structure and make the active site more accessible, possibly for some role in regulating active site access for the various substrates of *T. maritima* MetC (§ 4.3.10).

### 6.4 Catalytic multitasking vs. generalist intermediates

Primordial enzymes are believed to have been able to catalyse multiple reactions on multiple substrates. Another instance where a variety of activities is seen is in the generalist intermediates from directed evolution studies that changed the main activity of an enzyme (§ 1.7, Matsumura and Ellington, 2001, Hoffmeister *et al.*, 2003, Cohen *et al.*, 2004, Aharoni *et al.*, 2005, Fasan *et al.*, 2008). In some cases these mutants were found to act on a broad range of substrates, including substrates that were not under selection. This led to the hypothesis that these enzymes may be used as models to infer the properties of primordial enzymes. By definition, catalytically multitasking enzymes can catalyse multiple reactions and consequently could also be good models to infer the properties of primordial enzymes.

The two different types of reaction catalysed by *T. maritima* MetC, elimination and racemisation, start with the same mechanistic step: abstraction of a proton from Cα of the amino acid substrate. Subsequently, they differ due to the departure of a leaving group
(homocysteine) from cystathionine in the elimination reaction, as opposed to reprotonation on the other face in the racemisation reaction. The ability of *T. maritima* MetC to bind three different substrates (alanine, glutamate and cystathionine) has the side effect that the active site may be more permissive to other substrates than the active site of *E. coli* MetC, as seen with the proportionately larger cysteine elimination activity, which is most likely to be promiscuous (rather than physiological).

Directed evolution to improve the cystathionine elimination activity of *T. maritima* MetC resulted in two phenomena. On one hand, the turnover number for the racemisation activities halved, while the turnover number for the elimination activities increased by ~13%, which indicates a trade-off in terms of turnover numbers. On the other hand, all of the Michaelis constants decreased and the strongest decreases were for the two smaller substrates, cysteine (a substrate for elimination) and alanine (a substrate for racemisation). In terms of catalytic efficiency ($k_{cat}/K_M$), the two phenomena did not mask each other and instead an increase was seen for all activities, bar possibly for glutamate racemisation.

In the landscape of the enzymatic activities (§ 1.7 and Fig. 1.2), the difference in the balances of activities between the wild-type *T. maritima* MetC and the mutant might suggest that the wild-type enzyme is in a local minimum that meets the balance required. This indicates that the evolutionary paths that lead to and from a multitasking enzyme are different paths to those that progress to and from a generalist intermediate.

Consequently, the evolution of *T. maritima* metC progressed to an improved generalist intermediate, similar to previous evolution studies that began with specialised enzymes. This was in contrast to my original hypothesis (§ 5.3.6), that *T. maritima* MetC would specialise, especially in light of the fact that cystathionine elimination was already the activity with the largest catalytic efficiency.

### 6.5 Future directions

In previous chapters, many hypotheses were addressed, but several remain untested. In terms of *T. maritima* MetC, the crystal structure (solved by our collaborators) was informative in many respects; however, not all information was extracted. The ability to racemise glutamate was hypothesised to be a result of the W108 residue in light of a crude manual docking, so for conclusive proof more accurate results should be obtained either by
computationally docking the two isomers of glutamate or by soaking T. maritima MetC crystals with glutamate. This strategy could be extended to other substrates, such as alanine, to similarly determine the underlying mechanism. Further proof of the mechanism of glutamate racemisation could come from the prediction that a T. maritima MetC W108F mutant would be unable to racemise glutamate, while retaining its other activities. Additionally, the structure showed the presence of an occluding loop, which may open up. This is a possibility that could be tested by molecular dynamics simulation or by mutagenesis. For example, an R43A substitution might disfavour the hypothetical latched-open conformation, resulting in increased Michaelis constants. A D347E substitution might strengthen the salt bridge with K111 or K110, resulting in similar changes. On the other hand, a D347A or D347A/K111A substitutions would lack the occlusions and the more accessible active site might therefore result in lower Michaelis constants. In terms of the filled space between the occluding loop and R357 that corresponds to a cavity in E. coli MetC (§ 5.2.5), a Y345A or Y345A/M153A substitutions might result in a similarly permissive active site.

From a kinetic perspective, T. maritima MetC is most likely to racemise or eliminate several other compounds in a promiscuous manner, including the elimination of sulfur-containing D-amino acids, such as D-cysteine, and a mix of elimination and racemisation for serine. These possible promiscuous racemisation activities could be probed if suitable assays were developed. This breadth of activities might give a more precise picture into the balance of activities in wild-type and in the mutant.

A further question is what are the future steps in the evolution of T. maritima metC towards the encoding of a specialist cystathionine β-lyase. Further rounds of directed mutagenesis via error-prone PCR could be done using the T. maritima metC mutant as a template. Additionally, other activities could be targeted for directed evolution, once a selection system is established. For example, for an improved alanine racemisation activity for alanine catabolism, an alr dadX double knockout strain able to grow on minimal media would be required (§ 5.3.2).

In this thesis, one metC gene from each of the taxa of interest was studied, however, several other species in those taxa are present and may give more information on the balance of activities in multitasking enzymes. Examples of these include the metC from T. lettingae, the metC from “Pelagibacter bermudensis”, a species with a higher optimal
temperature than *P. ubique* (§ 4.3.8) and the metC from *Anaplasma marginale*, as the encoded enzyme possesses a lysine at the residue equivalent to W108.

A broad hypothesis that remains to be addressed is that organisms with streamlined genomes may possess an abundance of multitasking enzymes; this line of enquiry opens up a gold mine of possible candidate multitasking enzymes. Despite the issues associated with investigating the metabolism of an organism that is not *E. coli* (cf. chapter 2), the identification and characterisation of novel multitasking enzymes, especially those with different folds to MetC, would greatly enrich our understanding of this biochemical phenomenon and potentially aid us in better understanding primordial enzymes.

### 6.6 Implications of this study

Due to the sheer amount of evolution that has occurred since the beginning of life on Earth, traces of early events have all but completely been erased. This erasure of early events is so strong that there is a limited amount of evidence that RNA sequences were even present in the last common universal ancestor, despite the fact that RNA is the strongest candidate as the genetic element of early life (Hoeppner *et al.*, 2012). Even though protein metabolism is slightly more recent as it is a palimpsest of prebiotic/RNA-world chemistry (Benner *et al.*, 1989), the properties of early enzymes can only be speculated upon.

The advent of life required various steps and consequently the model we have of how life began is an amalgamation of various hypotheses. The initial process in the origin of life may have been the rise of prebiotic organic compounds. Stanley Miller’s spark discharge experiments and the presence of organic compounds, such as amino acids, in meteorites give support to the most commonly accepted theory of how life began. This hypothesis conjectures that geological and atmospheric conditions present in the Hadean eon resulted in the abiotic generation of a large diversity of simple organic compounds (Caetano-Anollés *et al.*, 2009).

In this environment, dubbed the primordial soup, some biochemical reactions could take place. These reactions are presumed to have given rise to ribonucleotides, among other metabolite-like molecules (Bada, 2013). The presence of ribonucleotide-based cofactors in modern enzymes, the catalytic ability of RNA and other factors give credence to the RNA
world hypothesis, which is a stage where RNA or RNA-like polymers were able to replicate themselves (Neveu et al., 2013). The number and complexity of the metabolic processes that arose during the RNA world is debated, but the most commonly accepted model is a brief RNA world in which a few processes, such as translation, arose, while most of central metabolism arose later (Fox, 2010).

The first genetically-encoded proteins are likely to have been short, partially structured and composed of fewer than 20 amino acids (Bernhardt and Patrick, 2014). In spite of these properties, weak catalysis must have been possible, allowing the encoding genes to be positively selected. It has been shown that molten globules (proteins lacking defined tertiary structures) can catalyse reactions, albeit less efficiently and with less specificity that structured enzymes (Vamvaca et al., 2004). Similarly, functional enzymes can be made with reduced amino acid alphabets (Müller et al., 2013).

As the encoding genes were evolved, the enzymes became more stable and more efficient, but they might have retained the broad specificity. In fact, chromosome segregation is a complex process, which is believed to have originated relatively late and as a result before this innovation protocells were forced to have a small number of multicopy unlinked genes in order for inheritance to be possible. Consequently, both early and late primordial enzymes must have had a broad range of activities. In the former group multitasking is a result of the multiple conformations adopted due to the lack of a rigid tertiary structure, while in the latter group multitasking is a result of selective pressures.

In both groups, as a result of the combined activities of various enzymes with broad specificities might have formed a patchwork network, which might have produced several useful metabolites, which would result in the steps that produced them to be selected for, even though each step in isolation would have not have conferred a selective advantage (§ 1.1).

In light of the difference between catalytic multitasking and catalytic promiscuity (i.e. physiological Michaelis constants, § 6.3) and the difference between catalytic multitasking enzymes and generalist intermediates (i.e. constrained specificity as opposed to a non-specific nature, § 6.4), the question of what is the best modern model for primordial enzymes arises.

Multitasking enzymes have been shaped by multiple evolutionary pressures. This is in contrast to generalist intermediates, which are the result of a single pressure: a pressure
towards the gain of a new specificity, which requires that the ancestral specificity be lost. In *T. maritima* MetC, and possibly multitasking enzymes in general, in order to acquire a new specificity, the constraints on the ancestral specificities need to be relaxed. The constrained specificity of multitasking enzymes may be a way to achieve a better balance of activities and as a result is the product of selective pressure and a safeguard against wasteful activities.

It can be envisaged that unbridled catalytic activities would most likely result in the wasteful degradation of many metabolites, which would make the production of complex metabolites impossible. More precisely, without negative selection, certain reactions, such as those with low activation energies or with more entropically favourable products, would be more common than more complex reactions. For example, in many ping-pong reactions, such as the transsulfurylation reaction of MetB or MetY, the premature release of the intermediate would be more favourable in terms of entropy (two or more products from a single substrate, as opposed to two products from two substrates). Rapid selection would have been required to prevent the premature release of the intermediate, otherwise it would waste the substrate that took several reactions to produce. Unconstrained reactivity may have been acceptable for early primordial enzymes – the poorly selected molten globules encoded by a genetic code that was still in flux – as they most likely performed simple reactions. This, however, would not have been acceptable for late primordial enzymes – the selected and more stable enzymes that were encoded by unlinked genes and with a fixed genetic code. If the late primordial enzymes did not inhibit wasteful reactions, but were highly permissive, few catabolic reactions would be possible. This repression of wasteful activities is a trait that was also seen in the multitasking enzymes studied here.

If primordial enzymes were like multitasking enzymes, the multitasking primordial enzymes might have diverged directly into specialists, as the current comparative model (generalist intermediates) would suggest. However, the results of the directed evolution experiments presented in this thesis (§ 5.3.3–5.3.8) suggest a novel step in which the multitasking enzymes evolved into more permissive generalists. Metabolic pathways are unlikely to have appeared simultaneously, but rather they may have appeared in bursts (Caetano-Anollés *et al.*, 2009) and therefore the broad range of activities of primordial enzymes was maintained. This would indicate that primordial enzymes might have switched several times between a permissive generalist form and a multitasking form. I argue that a similar situation is seen in organisms with streamlined genomes, which are likely to possess many multitasking enzymes.
Consequently, in this thesis the properties of three catalytic multitasking enzymes have been explored, revealing a complex balance of specificities. This affects how the enzymes evolve, which can be used to gain new insights into the likely nature of primordial enzymes.
References


References | 178


Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., Bhattacharyya, A., Kapatral, V., Kumar, S., Posfai, J., Vincze, T., Ingram,


Kurahashi, M., Fukunaga, Y., Harayama, S. & Yokota, A. (2008). *Sneathiella glossodoripedis* sp. nov., a marine alphaproteobacterium isolated from the
nudibranch *Glossodoris cincta*, and proposal of *Sneathiellales* ord. nov. and *Sneathiellaceae* fam. nov. *Int J Syst Evol Microbiol*, 58 (Pt 3): 548-552.


**Le Roux, N.** (2012). Characterisation of a multifunctional enzyme from the marine bacterium *Pelagibacter ubique*. BSc (Honours), Massey University.


Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., Mcdonald,
M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D.,
Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O.,
for lateral gene transfer between archaea and bacteria from genome sequence of

Neveu, M., Kim, H. J. & Benner, S. A. (2013). The "strong" RNA world hypothesis:


Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-
GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen,

Noda, M., Kawahara, Y., Ichikawa, A., Matoba, Y., Matsuo, H., Lee, D. G.,
cycloserine-producing Streptomyces lavendulae. Gene cloning, characterization, and

that alanine racemase from a D-cycloserine-producing microorganism exhibits


Ono, B., Ishii, N., Naito, K., Miyoshi, S., Shinoda, S., Yamamoto, S. & Ohmori,


catalytic efficiency effects caused by random single amino acid substitutions. Mol


Ladner, J. E. (2002). Structure of Escherichia coli aminodeoxychorismate synthase:


Appendix I

General Materials and Methods
I.1 Materials

I.1.1 Chemicals

Chemicals used were purchased from Sigma–Aldrich (St. Louis, MO), unless otherwise stated. Whereas molecular cloning was performed with nuclease-free (not DEPC-treated) water (Ambion®, subdivision of Life Technologies, Carlsbad, CA), all other solutions, including rich media, were prepared using filtered deionised water (18 MΩ/mℓ) from a MilliQ® system (Millipore, subdivision of Merck Darmstadt, Darmstadt, Germany).

I.1.2 Strains

E. coli DH5α

**Purpose:** Common high efficiency transformation strain

**Genotype:** K-12 λ⁻ F⁻ hsdR17 hsdM(wt) (r⁻ m Exploration) Δ(lacZYA–argF)205U169 Φ80(ΔlacZΔM15) endA1 relA1 recA1 thi⁻1 supE44 gyrA96 phoA deoR

**Note:** phoA deoR were reported to be mutated, but whole-genome sequencing has disproven it (Xia et al., 2011).

**Source:** Invitrogen (San Diego, CA), now Life Technologies (Carlsbad, CA)

E. coli BW25113

**Purpose:** “wild-type”, a Clifton’s K-12 derivative without λ phage, F plasmid, ara operon and hsdR.

**Genotype:** K-12 λ⁻ F⁻ hsdR514 hsdM(wt) (r⁻ m Exploration) Δ(araD–araB)567 Δ(lacA–lacZ)514::kan) Δ(rhaD–rhaB)568 rph⁻1

**Source:** G. Cook. Constructed in (Datsenko and Wanner, 2000)

E. coli BW25113 derivatives

**Purpose:** complementation tests

**Strains and genotypes:** JW2975 {BW25113 metC::kan}

JW3973 {BW25113 metB::kan}

JW1082 {BW25113 pabC::kan}

**Source:** Keio collection (Baba et al., 2006)

E. coli MB2795

**Purpose:** complementation tests

**Genotype:** MC1000 dadX::frt alr::frt
Genotype of MC1000: K-12 λ⁻ F⁻ hsdR( wt) hsdM( wt) (rK mK) araD139b/s Δ(araA–leu) 7679 Δlac174 Δ(codB–lacI) 3 galE15 u38 galK16 rpsL150? thi⁻ 1 spoT1

Note: rpsL150 confers streptomycin resistance, but I found no indication of such with MB2795 (data not shown).

Source: K. Krause (Strych et al., 2001)

E. coli WM335

Purpose: complementation tests

Genotype: WM301 gltS' mur355

Genotype of WM301: B/r F⁻ hsdS K-12 (rK mK) arg⁻ 28 (dra or drm) gal⁻ 11 his⁻ 47 lac⁻ 11 leuB19 met⁻ 55 pro⁻ 19 rpsL150 thyA59 trp⁻ 25 sulA1 deoB23

Source: D. Mengin-Lecreulx (Doublet et al., 1992, Dougherty et al., 1993)

E. coli ER2523

Purpose: protein over-expression using pBAD

Genotype: DL21 {B F⁻ dcm⁻ ompT⁻ lon⁻ hsdS⁻ (r⁻ m⁻) gal⁻ malB K-12 λ⁻ } fhuA2 sulA11 endA1 mcr-73::miniTn10(Tet+) zgb-210::Tn10(Tet+) (mcrC–mrr)114::IS10

Source: NEB (New England Biolabs, Ipswich, MA)

E. coli LMG194

Purpose: protein over-expression using pMAL

Genotype: KS272 Δ(ara–leu):Tn10

Note: Life Technologies states that KS272 {K-12 λ⁻ F⁻ lacX74 galE⁻ glaK thi⁻ 1 rpsL150 phoA(PvuII)} is a derivative of the MC1000 strain, but the genotypes do not match (viz. MB2795).

Source: Invitrogen

I.I.III Plasmids

The antibiotic selection system is written in the description, namely kan-30 (kan gene encoding aminoglycoside phosphotransferase, 30 µg/ml kanamycin, powder purchased from Melford, Ipswich, UK), amp-100 (bla gene, 100 µg/ml ampicillin, powder from Roche Applied Biosciences, Penzberg, Germany), cm-34 (cat gene, 34 µg/ml chloramphenicol, powder from Duchefa Biochemie, Haarlem, Netherlands). All antibiotics were made as ×1,000 stocks in water, except chloramphenicol, which was dissolved in ethanol.
<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pMK-RQ/metc&lt;sub&gt;Wmel&lt;/sub&gt;</td>
<td>Gene synthesis (11AAHZEP). N-terminal hexa-histidine tag and TEV cut site. Kan-30</td>
<td>GeneArt (Regensburg, Germany)</td>
</tr>
<tr>
<td>pBAD/myc-his(B)</td>
<td>Empty pBAD plasmid with C-terminal myc epitope and hexa-histidine tag. Amp-100</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>pBAD/metc&lt;sub&gt;Wmel&lt;/sub&gt;(Kpn&lt;sup&gt;−&lt;/sup&gt;)</td>
<td>wMel metC cloned into pBAD with a stop codon before the C-terminal tags. Amp-100</td>
<td>This study</td>
</tr>
<tr>
<td>pBAD/metc&lt;sub&gt;Wmel&lt;/sub&gt;(Kpn&lt;sup&gt;+&lt;/sup&gt;)</td>
<td>Above, but with KpnI site downstream of the sequence encoding the TEV protease cut site. Amp-100</td>
<td>This study</td>
</tr>
<tr>
<td>pBAD/metc&lt;sub&gt;Tma&lt;/sub&gt;</td>
<td>Tma metC cloned replacing wMel metC in pBAD/metc&lt;sub&gt;Wmel&lt;/sub&gt;(Kpn&lt;sup&gt;+&lt;/sup&gt;). Amp-100</td>
<td>This study</td>
</tr>
<tr>
<td>pMAL/metc&lt;sub&gt;Pub−malE&lt;/sub&gt;</td>
<td>pMAL-c5X derivative encoding Pub MetC fused to MBP. Amp-100</td>
<td>Natasha le Roux</td>
</tr>
<tr>
<td>pCA24N/metc−gfp&lt;sub&gt;uv4&lt;/sub&gt;</td>
<td>ASKA Collection plasmid carrying E. coli metC fused to gfp. Cm-34</td>
<td>(Kitagawa et al., 2005)</td>
</tr>
<tr>
<td>pCA24N/alr−gfp&lt;sub&gt;uv4&lt;/sub&gt;</td>
<td>ASKA Collection plasmid carrying E. coli alr fused to gfp. Cm-34</td>
<td>(Kitagawa et al., 2005)</td>
</tr>
<tr>
<td>pCA24N/murI−gfp&lt;sub&gt;uv4&lt;/sub&gt;</td>
<td>ASKA Collection plasmid carrying E. coli murI fused to gfp. Cm-34</td>
<td>(Kitagawa et al., 2005)</td>
</tr>
<tr>
<td>pBAD/metc&lt;sub&gt;Eco&lt;/sub&gt;</td>
<td>E. coli metC in pBAD. Amp-100</td>
<td>(Soo, 2012)</td>
</tr>
<tr>
<td>pCA24N/gfp&lt;sub&gt;uv4&lt;/sub&gt;</td>
<td>ASKA Collection plasmid with gfp. Cm-34</td>
<td>(Soo, 2012)</td>
</tr>
<tr>
<td>pKD46</td>
<td>Temperature-sensitive plasmid with arabinose-inducible λ red genes for homologous recombination. Amp-100</td>
<td>(Datsenko and Wanner, 2000)</td>
</tr>
<tr>
<td>pCP20</td>
<td>Temperature-sensitive plasmid encoding FLP recombinase to remove genes flanked by FRT sites. Amp-100</td>
<td>(Datsenko and Wanner, 2000)</td>
</tr>
</tbody>
</table>
### I.1.IV Primers

<table>
<thead>
<tr>
<th>Name (collection number)</th>
<th>Sequence (5’→3’)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>tev_kpnI_F (512)</td>
<td>5’-Phos-GAGAACCTGTATTTCCAAGGAAAAGAGAAAG</td>
<td>This study</td>
</tr>
<tr>
<td>tev_kpnI_R (513)</td>
<td>5’-Phos-GGTACCTGTAGGAATGTCGTAGTCGTGATGG</td>
<td>This study</td>
</tr>
<tr>
<td>Tma_metC_kpnI_F (514)</td>
<td>CAGGTACCGAGAACCTGTATTTCCAAGGAAAACACAGACATTCTGTTTTC</td>
<td>This study</td>
</tr>
<tr>
<td>Tma_metC_xbaI_R (515)</td>
<td>CTTGTTCTAGATTTAGTTTTTTTGGATGCCTGATCCAGATTTCC</td>
<td>This study</td>
</tr>
<tr>
<td>pBAD-60 (108)</td>
<td>ATCGCAACTCTCTACTGTTTCTCCATAC</td>
<td>(Soo, 2012)</td>
</tr>
<tr>
<td>pBAD.rev (325)</td>
<td>GAAACTTCTCTCTACTGTTTTTCTCCATAC</td>
<td>(Soo, 2012)</td>
</tr>
<tr>
<td>metC-upstr.for (231)</td>
<td>CCGGAATATTCTAGCTAGTTTTAGACATCCAGAC</td>
<td>(Soo, 2012)</td>
</tr>
<tr>
<td>metC-dwnstr.rev(232)</td>
<td>GACTTTTCACAATAAAAATGTCTGCAAATTGTCCAAAG</td>
<td>(Soo, 2012)</td>
</tr>
</tbody>
</table>
### I. Methods

All laboratory equipment was used according to manufacturer's instruction, unless otherwise stated. All laboratory equipment was commercially available, except for the bucket of light, which was constructed by Gary Shriffer (workshop manager, Department of Biochemistry, University of Otago) based upon Parkinson (2007).

#### I.II.1 General microbiology

RABINOSE was from GoldBio (St. Louis, MO), LB medium pre-mixed powder, TB medium pre-mixed powder, casein-derived peptone and agar (various grades) were from ForMedium (Huntstanton, UK), whereas M9 salts mix was from Sigma–Aldrich.

The following media were used in this study:

- LB medium (Lysogeny broth; 10 g/ℓ tryptone, 5 g/ℓ yeast extract and 10 g/ℓ NaCl)
- Buffered TB medium (12 g/ℓ tryptone, 24 g/ℓ yeast extract, 9.4 g/ℓ KH₂PO₄ and 2.2 g/ℓ KH₂PO₄, without glycerol)
- SOC medium (Super-optimal broth; 20 g/ℓ tryptone, 5 g/ℓ yeast extract, and 0.5 g/ℓ NaCl, 2.5 mM KCl, 10 mM Mg₂Cl, 20 mM glucose), prepared according to Sambrook and Russell (2001a, p. A2.3)
- M9 medium (M9 salts: 12.8 g/ℓ Na₂HPO₄·7H₂O, 3 g/ℓ KH₂PO₄, 0.5 g/ℓ NaCl, 1 g/ℓ NH₄Cl; 2mM MgSO₄, 100µM CaCl₂), prepared according to Sambrook and Russell (2001a, p. A2.2)
Unless otherwise stated, these media were sterilised by autoclaving for 20 min on a liquid cycle. Individual amino acids, cofactors, antibiotics, salts inducers and glucose were sterilised by filtration through 0.22 µm pores (either from Millipore or from Whatman, a subdivision of General Electric, Little Chalfont, UK). Disposable polystyrene Petri dishes, 50 ml and 15 ml tubes were pre-sterilised by the manufacturer (vv. distributors), as were cryotubes® (Nunc, Roskilde, Denmark), while glassware (mostly Schott, Jena, Germany), tips (vv. brands), 1.5 ml and 2 ml tubes (vv. brands) were sterilised by autoclaving for 20 min.

_E. coli_ were routinely grown in LB medium at 37°C with agitation of 180 rpm and with about ten-fold more headspace than broth. Freezer stocks were made from 1 ml overnight cultures mixed with 250 µl 50% autoclaved glycerol (10% final) and stored at −80°C.

In the case of solid media, 1.5% agar was present in the media, with the exception of rescue experiments on M9 with a cofactor auxotrophy, in which case 1.5% agarose was used.

### I.II.II Electroporation

Electrocompetent cells were prepared according to Sambrook and Russell (2001b, pp. 1.119–1.122) with the exception that 10% glycerol was used instead of GYT (glycerol, yeast extract and tryptone) medium as the solution in which the cells were electroporated as preliminary data suggested it reduced arcing.

Briefly, this protocol entailed the following steps: the culture was grown to an OD<sub>600</sub> of 0.4–0.6, the culture was cool for 20 min on ice before being distributed among 50 ml tubes in order to sediment the cells on a bench-top centrifuge at over 1,000 g for 10–15 min. This process was repeated once with a wash with pre-chilled sterile water, twice with pre-chilled 10% sterile glycerol, before being precipitated and resuspended to a concentration of 2–3 · 10<sup>10</sup> CFU/ml determined by absorbance at a wavelength (λ) of 600 nm, where 1.0 AU = ~2.5 · 10<sup>8</sup> CFU/ml for most _E. coli_ strains. The suspension was aliquoted as 50 µl samples and stored at −80°C.

In the cases where a limited amount of colonies were required, a quicker protocol was used (adapted from Choi _et al._, 2006, Soo, 2012). Namely, a single aliquot of cells was
prepared from 6 mℓ of saturated culture that was first aliquoted in 2 mℓ tubes as microcentrifuges can attain higher g-forces than tabletop centrifuges; then the cells (not prechilled) were washed twice in either 10% glycerol or 330 mM filtered sucrose with spins at 13,000 g for 1 min. The pellets were then pooled, resuspended and centrifuged. Most of supernatant was removed, leaving a final pellet of ~50 µℓ, which was used for electroporation.

Electroporation was performed with a BioRad MicroPulser® (BioRad, Hercules, CA). 1–100 ng of plasmid or ligation mixture was added to the thawed cell suspension and transferred to an electroporation cuvette with a 2 mm gap in order to subject it to a 3–6 ms pulse at 2.5 kV. After electroporation, 1 mℓ SOC (§ I.II.I) was added to the suspension, which was subsequently transferred to a test tube in order for a 1h recovery at 37°C (or 28°C if temperature sensitive), before being spread on plates with the appropriate markers.

I.II.III Cloning

Routine PCR was performed either with iTaq™ (iNTrON Biotechnology, Seongnam, South Korea) or KAPA Taq™ (KAPA, Boston, MA) according to manufacturer’s instructions in a standard thermocycler with a heated lid. When high fidelity amplicons were required, Phusion® (NEB) or KAPA HiFi Taq™ (KAPA, Boston, MA) were used according to manufacturer’s instructions.

Plasmids were purified from 5 mℓ overnight culture with either the QIAprep® Spin Mini Kit (QIAGEN, Limburg, Netherlands) or the E.Z.N.A.* plasmid mini kit (Omega BioTek, Norcross, GA). Amplicons or ligated fragments were purified with either the QIAquick® PCR clean-up kit (QIAGEN, Limburg, Netherlands) or the E.Z.N.A.* PCR clean-up kit (Omega BioTek, Norcross, GA). When a concentrated amount was required the MinElute® Reaction Cleanup Kit was used (QIAGEN, Limburg, Netherlands). DNA was quantified by absorbance at 260 nm with an Eppendorf BioPhotometer (Eppendorf, Hamburg, Germany), which uses UVettes™, special cuvettes that handle 50–100 µℓ of liquid.

Agarose gel electrophoresis was performed according to Sambrook and Russell (2001b, pp. 5.6–5.17) in 7 × 10 cm gel tray in a Mini-Sub® Cell GT System (Bio-Rad Laboratories, Hercules, CA) with 1–1.5% agarose (various distributors), 0.5 µg/mℓ
ethidium bromide (Bio-Rad Laboratories) and with 1× TAE buffer (prepared according to Sambrook and Russell, 2001a, p. A1.17).

As a comparison one lane contained a regularly spaced molecular weight ladder either from NEB (Ipswich, MA), Fermentas (Thermo Fisher Scientific, Waltham, MA) or KAPA (Boston, MA).

The gels were visualised with UV light and a Gel Doc™ system (Bio-Rad Laboratories).

In the case of DNA fragment separation and purification, the stain used was SYBR® Safe (Invitrogen) instead of ethidium bromide and the bands were excised with a scalpel and a Safe-Imager® blue-light trans-illuminator (Invitrogen) as the UV light would otherwise damage the DNA. The DNA was isolated with either the QIAquick® gel extraction kit (QIAGEN, Limburg, Netherlands) or the GenElute® gel extraction kit (Sigma–Aldrich).

DNA restriction reactions were done with reagents from NEB. A typical reaction contained 50 µl water buffered with NEB-buffer 4, 2–5 µg of purified DNA (plasmid or amplicon), 100 µg/ml BSA if necessary (e.g. for XbaI) and 1 µl (20 U) of the required enzyme(s). In the case of KpnI and XbaI double digest, 2 µl (40 U) KpnI and 1 µl (20 U) XbaI were used in NEB buffer 4, as the former enzyme had a subprime activity (50%) in the preferred buffer of the latter. Reactions were done at 37°C in a thermocycler for 20–60 min.

Sticky-end ligations were done with T4 ligase (Fermentas) according to manufacturer’s instructions, with the modification that the incubations were at 16°C from one hour to overnight, to favour the annealing of the sticky-ends. The reactions generally consisted of a 2:1 molar ratio of purified cut insert and vector, where the vector was ~5 ng/µl, and the total volume was between 20 µl and 100 µl.

I.II.IV Cell lysis and fractionation

Cells were grown and induced according to the specific conditions for each gene as discussed in § 4.4.1 and § 5.2.3.

The cells were harvested by centrifugation at 7,000 g for 15 s at 4°C and either lysed immediately or the pellets were scraped into a 50 ml tube and stored at −80°C.
The cells were resuspended (2–5 mL per g of pellet) in the relevant column buffer (viz. § I.11.V for IMAC chromatography or § I.11.VI for amylose affinity chromatography) with 0.5 mg/mL lysozyme (vv. distributors), 0.1 µg/mL DNase I (Thermo Fisher Scientific), 1 mM β-mercaptoethanol, 10 µM PLP and incubated on a roller at 4°C for 30 min to an hour. The cells were then sonicated (Ultrasonic processor S-4000, Misonix, Farmingdale, NY, or Vibra-Cell sonicator, Sonics & Materials, Inc., Newtown, CT) on ice at 40% amplitude with ten 10 s pulses interspersed with 30 s pauses. The lysate was then fractionated by centrifugating at 20,000 g for 45 s at 4°C. The supernatant was further clarified through a 0.44 µm filter, either using a syringe with an in-line filter or a vacuum-assisted filtration system.

**I.11.V Gravity flow IMAC purification**

For gravity flow metal affinity chromatography, the column buffer contained 50 mM potassium phosphate buffer pH 8.0, 300 mM NaCl, 10% v/v glycerol and 1 mM imidazole pH 8.0.

The clarified lysate (§ I.11.IV) was mixed with 0.5 mL bed volume prewashed cobalt-carboxylmethylaspartate-functionalised agarose beads (Talon resin®, ClonTech, TaKaRa Bio Inc., Shiga, Japan) for 20 s at 4°C. The beads were then washed in column buffer with 1–3 mM imidazole (pH 8.0) by centrifugation 2–5 times at 800 g for 5 s and subsequently resuspended in column buffer before being packed in an empty column (Econo-Column® or Bio-Spin®, BioRad, Hercules, CA) and further washed by gravity flow for 20–30 bed volumes. The protein was eluted from the beads with 3 mL of 100 mM imidazole in column buffer and then again with 3 mL buffer with 500 mM imidazole pH 8.0.

Aliquots of the elution fractions along with an aliquot of the total, supernatant, pellet, first and last wash were run on an SDS-PAGE gel.

This fraction was then concentrated and buffer exchanged four or more times using an Amicon centrifugal filter unit with a 50 kDa cut-off (Millipore) for an expected imidazole concentration <10 µM.

The enzymes were stored at −80°C as either 100 µM or 50 µM aliquots.
I.II.VI Gravity flow amylose affinity purification

For gravity flow amylose affinity chromatography of *P. ubique* MetC, the column buffer contained 20 mM Tris · HCl (pH 8.0), 600 mM KCl, 1 mM EDTA, 10 mM β-mercaptoethanol, 10 µM PLP, while the lysis buffer additionally contained 0.5 mg/ml lysozyme, 0.2 µg/ml DNase I and 50 µl protease inhibitor cocktail (P8849 from Sigma–Aldrich).

The washing was performed as described for metal affinity chromatography (§ I.II.V). The protein was eluted with 5 ml 10 mM maltose and dialysed with a 10 kDa dialysis cassette (Pierce Biotechnology, Rockford, IL) overnight against storage buffer, which contained 600 mM KCl, 50 mM Tris · HCl pH 7.5, 10% glycerol. The enzymes were stored at −80°C as 50 µM aliquots.

I.II.VII SDS-PAGE

SDS polyacrylamide gel electrophoresis was performed according to Sambrook and Russell (2001a, p. A8.40–49) using a Mini-Protean® Tetra Cell set-up (Bio-Rad Laboratories) on 12% acrylamide home-cast gels in Tris · glycine buffer prepared as a 5× stock according to Sambrook and Russell (2001a, p. A1.17). The gels were either 0.75 mm or 1 mm thick with 10 or 20 wells depending on the requirement.

The gels were cast with a variant on the recipe provided by Sambrook and Russell (2001a, p. A8.43), as a 40% acrylamide/bis-acrylamide 29:1 mix (Bio-Rad Laboratories) was used as opposed to a 30% mix. Consequently, for 20 ml of resolving gel, 8.7 ml water, 5 ml 1.5 M Tris · HCl pH 8.8 (Bio-Rad Laboratories), 200 µl 10% w/v SDS, 6 ml 40% acrylamide mix, 200 µl 10% w/v ammonium persulfate (powder from Bio-Rad Laboratories) and 20 µl TEMED (tetramethylethylenediamine) were used. The latter quantity differs from Sambrook and Russell (2001a, p. A8.43, e.g. 8 µl). The top ~2 cm of the gel was composed of 3.8% stacking gel; for 4 ml 3.2 ml water, 2.5 ml 0.5 M Tris · HCl pH 6.8 (Bio-Rad Laboratories), 50 µl 10% w/v SDS, 500 µl 40% acrylamide mix, 25 µl 10% w/v ammonium persulfate and 5 µl TEMED. The samples were mixed with the same volume of ×2 loading buffer with 200 mM β-mercaptoethanol (Sambrook and Russell, 2001a, p. A1.20) and heated to 99°C for 5–10 min before loading. The molecular-
weight ladder was the Precision Plus Protein™ All Blue Standard from Bio-Rad Laboratories, 10 µl of which was added to one or more wells of the gel.

Once run, the gels were stained in a solution of 1.74 M acetic acid and 30% methanol with 0.25 % w/v Coomassie® Brilliant Blue R-250 (vv. manufacturers) by gentle agitation generally overnight, followed by a destaining soak in the same solution without the dye. This is as seen in Sambrook and Russell (2001a, p. A8.47), except for the methanol concentration, which in the book is 50%, a concentration that on one hand is more stringent at removing excess dye, but on the other shrinks the gel.

I.II.VI Quantification

The concentration of each enzyme was determined based on its UV absorbance using a Varian Cary® 100 Bio spectrophotometer (Agilent Technologies, Santa Clara, CA) and fused-quartz cuvettes (6Q grade, 1 cm pathlength, 100 µl volume).

The spectra were measured with 0.2 s reads each nanometre from 200 nm to 800 nm in order to see the absorbance of the aromatic residues and of the PLP cofactor. Due to the switch between the ultraviolet and the visible lamp at 348–349 nm a shift in absorbance was present between the UV and the visible part of the spectrum; to overcome this, the difference between the absorbance at 348 and at 349 nm was subtracted from the UV range, before the whole spectrum was subtracted first by the absorbance at 600 nm, not 350 nm, and then subtracted again by a similarly normalised spectra of a water blank.

The absorbance of the PLP moiety at 420 nm was monitored, but not used to quantify the protein, e.g. with a \( \varepsilon_m = 8,450 \text{ M}^{-1} \cdot \text{cm}^{-1} \) (Inagaki et al., 1986, Sambrook and Russell, 2001a, p. A8.47), as PLP is slightly solvatochromic (data not shown). Instead, the absorbance of the aromatic residues of the protein at 280 nm was estimated by the ProtParam tool of the Expasy server (web.expasy.org/protparam) using the method of Pace et al. (Pace et al., 1995) was used. The molar absorbance coefficients (\( \varepsilon_m \)) were:

- **Wolbachia MetC**: 56,400 M\(^{-1}\) · cm\(^{-1}\)
- **T. maritima MetC**: 44,350 M\(^{-1}\) · cm\(^{-1}\)
- **P. ubique MetC–MBP**: 109,670 M\(^{-1}\) · cm\(^{-1}\)
I.H.VII General procedure for assays

All assays were performed at 37°C (310.15 K) with the aforementioned spectrophotometer and in a 500 µl volume in 1 ml fused-quartz cuvettes (on average 8 in the cell changer) buffered with 50 mM Tris · HCl pH 8.8 and with 10 µM PLP for the enzyme.

In the case of assays using NADH as a reporter, the absorbance at 340 nm was used, as the absorbance of NAD⁺ is negligible at that wavelength, whereas the molar absorption coefficient of NADH at 340 nm is 6,220 M⁻¹ · cm⁻¹ (Esaki and Walsh, 1986).

A master stock was made for each substrate, which was used to make the 20× stocks for each concentration tested (i.e. 25 µl substrate stock was used in the 500 µl reaction solution).

The cuvettes with the assay mixture were left in the machine for at least 5 min to reach the desired temperature, before MetC was added at a final concentration between 0.1–5 µM. The linear portion of the absorbance plot, generally the first 5 min, was used for a linear regression in Excel 2008 (Microsoft, Redmond, WA) with the slope function and the rsq function to verify fit (R²>95%). The AU/min rate was then converted to µM/s. The values were subtracted by the values without substrate and then the data was fitted to the Michaelis–Menten equation in Prism 4.1 (GraphPad, La Jolla, CA) using the preset equation with automatic outlier elimination. The resulting values from the regression were reported with standard errors. In order to obtain the catalytic efficiency (kcat/KM) with standard error, the dataset was fitted to a modified Michaelis–Menten equation, resulting from the division of the numerator and the denominator by the Michaelis constant (Manders et al., 2013).

In all assays eight or more concentrations were assayed and generally were centered around the Michaelis constant. High concentration data points were omitted when over 10 fold larger than the Michaelis constant. For most concentrations of interest, three or more data points were taken. In the case of E. coli MetC and P. ubique MetC only one biological replicate was performed, while in the case of T. maritima MetC and Wolbachia MetC three replicates were performed.
I.II.VIII Cystathionine elimination

Cystathionine β-lyase activity was assessed by measuring the production of free thiols, in this case homocysteine, using Ellman’s reagent (DTNB), a disulfide-bonded inactive dye that absorbs at 412 nm ($\varepsilon_M=14,150 \text{ M}^{-1} \cdot \text{cm}^{-1}$) when the disulfide is reduced by free thiols (Ellman, 1959). The initial concentration of DTNB in each assay was 0.4 mM and, due to its absorbance at that wavelength, PLP was omitted. The cystathionine master stock was 100 mM cystathionine in 200 mM HCl.

I.II.IX Cysteine elimination

Cysteine cleavage into pyruvate was assayed through the reduction of the latter by lactate dehydrogenase (LDH from leporine muscle, L2500-100KU from Sigma-Aldrich) at 120 U/mℓ with 0.2 mM NADH. The master stock was 1 M L-cysteine · HCl.

I.II.X Alanine racemisation (from L-alanine)

D-alanine production was assayed with a coupled system where D-alanine is deaminated by D-amino acid oxidase (DAAO, from porcine kidney, A5222-500UN from Sigma-Aldrich) and the resulting pyruvate is reduced by LDH as above (Esaki and Walsh, 1986). The assay components consequently were 50 mM Tris · HCl pH 8.8, 0.2 mM NADH, 1 U/mℓ DAAO and 120 U/mℓ LDH (Esaki and Walsh, 1986). The master stock was 1 M L-alanine.

I.II.XI Alanine racemisation (from D-alanine)

L-alanine production was assayed with 10 mM NAD$^+$ and 0.2 U/mℓ alanine dehydrogenase (ADH, from B. subtilis, A7189-250UN from Sigma-Aldrich) (Milligan et al., 2007). The master stock was 1 M D-alanine in water.

I.II.XII Glutamate racemisation (from D-glutamate)

L-glutamate production was assayed with 10 mM NAD$^+$ and 1 U/mℓ glutamate dehydrogenase (GDH, bovine liver, G2501-20MG LDH from Sigma-Aldrich). The master stock was 1 M D-glutamate in 1 N NaOH and was diluted into 20× stocks. The glutamate racemisation assay was adapted from Lundqvist et al. (2007) with the omission of BSA and with a reduction in glutamate dehydrogenase concentration to avoid its precipitation at concentrations above 5 U/mℓ (data not shown).
I.II.XIII pH of reaction

To avoid potential changes in pH due to high substrate concentrations, the pH of the reaction mixtures was ascertained using the pH indicator phenolphthalein. 2% w/v phenolphthalein in EtOH was added at 1:1,000 ratio to a cuvette with a 500 µl assay reaction mixture. If the difference in absorbance at 553 nm at 25°C was about 0.1 AU, the solution was considered to be correctly buffered (pH 8.8), whereas if the absorbance was lower the solution was too acidic (e.g. >2 mM cystathionine · (HCl)_2), while higher too alkaline.

I.II.XIV Image creation

Various programs were used to generate the illustrations. Diagrams were drawn in Illustrator CS4 (Fig. 1.1, 1.2, 2.11, 2.12). Diagrams with chemical structures were made in ChemSketch before being imported into Illustrator (Fig. 2.2, 2.3, 4.1). Trees were also edited in Illustrator and they were made beforehand with the iTol server (Fig. 2.4), with FigTree (Fig. 2.9, 3.3, 3.8, 4.18) or directly in Illustrator (Fig. 2.7, 2.10, 3.2, 3.6, 3.7, 3.9). Some graphs were made in with Excel 2008 (Fig. 4.7, 5.1) or with Prism 4.1 (Fig. 3.5, 4.3, 4.4, 4.5, 4.6, 5.2). Photographs were taken with various cameras and edited in Photoshop (Fig. 2.6, 2.8, 4.2). Protein structural images were made with PyMol 1.2 (Fig. 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15, 4.16, 4.17, 5.3).
Known Racemases
The structural information was taken from the Pfam database, a database that structurally classifies protein (Punta et al., 2012). The sequence of lysine racemase was not determined. The origins of the D-serine and D-tyrosine residues of pyoverdine have not been determined fully.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Name</th>
<th>PFAM</th>
<th>Role</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>alr</td>
<td>Alanine racemase</td>
<td>PF01168 (CL0036)</td>
<td>Peptidoglycan, alanine catabolism</td>
<td>(Strych et al., 2000)</td>
</tr>
<tr>
<td>vanT</td>
<td>Serine racemase</td>
<td>PF01168 (CL0036)</td>
<td>Vancomycin-resistant peptidoglycan</td>
<td>(Arias et al., 2000)</td>
</tr>
<tr>
<td>bsrV</td>
<td>Broad specificity racemase</td>
<td>PF01168 (CL0036)</td>
<td>Regulatory signals</td>
<td>(Lam et al., 2009)</td>
</tr>
<tr>
<td>orr</td>
<td>Ornithine racemase</td>
<td>PF01168 (CL0036)</td>
<td>Arginine catabolism</td>
<td>(Chen et al., 2000)</td>
</tr>
<tr>
<td>lyr</td>
<td>Lysine racemase</td>
<td>ND</td>
<td>Lysine catabolism</td>
<td>(Huang and Davisson, 1958)</td>
</tr>
<tr>
<td>murl</td>
<td>Glutamate racemase</td>
<td>PF01177 (CL0399)</td>
<td>Peptidoglycan</td>
<td>(Lundqvist et al., 2007)</td>
</tr>
<tr>
<td>racX</td>
<td>Aspartate racemase</td>
<td>PF01177 (CL0399)</td>
<td>Peptidoglycan</td>
<td>(Yohda, Okada et al., 1991)</td>
</tr>
<tr>
<td>mycF</td>
<td>Aspartate racemase</td>
<td>PF01177 (CL0399)</td>
<td>Microcystin biosynthesis</td>
<td>(Sielaff et al., 2003)</td>
</tr>
<tr>
<td>dapF</td>
<td>Diaminopimelate epimerase</td>
<td>PF01678 (CL0288)</td>
<td>Lysine biosynthesis</td>
<td>(Usha et al., 2009)</td>
</tr>
<tr>
<td>prdF</td>
<td>Proline racemase</td>
<td>PF05544 (CL0288)</td>
<td>Catabolism</td>
<td>(Cardinale and Abeles, 1968)</td>
</tr>
<tr>
<td>dcsC</td>
<td>O-ureido-serine racemase</td>
<td>PF01678 (CL0288)</td>
<td>D-cycloserine biosynthesis</td>
<td>(Dietrich et al., 2012)</td>
</tr>
<tr>
<td>SRR</td>
<td>Serine racemase</td>
<td>PF03313</td>
<td>D-serine neurotransmitter</td>
<td>(Baumgart and Rodriguez-Crespo, 2008)</td>
</tr>
<tr>
<td>ND</td>
<td>Serine and tyrosine racemases</td>
<td>ND</td>
<td>Pyoverdine biosynthesis</td>
<td>(Visca et al., 2007)</td>
</tr>
</tbody>
</table>
Appendix III

Large Trees
Figure III.1. Maximum likelihood cladogram of Thermotogae. This was used to create Fig. 2.7.
Figure III. II. Maximum likelihood cladogram of MetC homologues. The species were picked based on preliminary trees constructed from a Blast search dataset, and rationally, to cover diversity and type species. The colours of the tiles bordering the tree represent the NCBI annotation of the sequence, in particular MetC (CBL) in green and MetB (CGS) in blue.
Figure III. Expanded version of Fig. 2.9. The tree was inferred by maximum likelihood (RAxML) using a manually assembled dataset aligned with MUSCLE (methods § 2.4.5).
Figure III IV. Maximum likelihood cladogram of Alr amino acid sequences. The highlighted four families of Rickettsiales are:

- Anaplasmataceae in yellow
- Rickettsiaceae in orange
- Holosporaceae in green
- Pelagibacteraeaceae in blue

The Rickettsiales cluster together with 20% support, but with the inclusion of a Rizhobiales species, Bartonella bacilliformis. This may be an artifact of long branch attraction or a horizontal gene transfer event to this rhizobacterium.

With the exception of three gammaproteobacteria and the outgroup, all the species present are Alphaproteobacteria. Overall, the tree demonstrates that alr in the Rickettsiales is of alphaproteobacterial origin and is vertically transmitted in this group.
Appendix IV

Perl Scripts
IV. I Subroutine package

package sub;
#17-9-10
my $start_time=time();
my $os = $^O;

sub readfile {
    my ($file)=@_;}
BACK:
    open(FILE,$file) or die {
        print "Cannot find file called $file, what is its real name?\n"; $file=<>; chomp $file; goto BACK;};
    my @line=<FILE>;
    close FILE;
    chomp(@line);
    if @{$line}<2) {
        @temp=spliit(/\r/,@line[0]);
        @tempo=spliit(/\r/,@line[1]); #pretty sure cannot be.
        @line=@temp;
        push(@line,@tempo);
    }
    return (@line);
}

sub get {
    return got($_[0]);
}

sub got {
    my $home=$_[0];
    require LWP::UserAgent;
    my $ua = LWP::UserAgent->new;
    $ua->timeout(200);
    $ua->proxy(['http', 'ftp'], 'http://tur-cache.massey.ac.nz:8080/');
    $ua->env_proxy;
    my $response = $ua->get($home);
    if ($response->is_success) {
        return $response->decoded_content; # or whatever
    } else {
        print "Internet problems again.. Could you check the router for me?\n";
        print $response->status_line;
    }
}

sub clean_array {
    my @array=@_;
    my @tempus=();
    for $n (@array) {
        if (length($n)>0) {
            if (grep($n,@tempus)==0){push(@tempus,$n);}
        }
    }
    return sort(@tempus);
}

sub clean_hash {
    my %hash=%_;}
    my %tempus=();
    for $n (keys %hash) {
        if (length($hash{$n})>0) {
            if (grep($n,@tempus)==0) {push(@tempus,$n);}
        }
    }
    return %tempus;
}

sub taken {
    $now = time - $start_time;
    $string=sprintf("%02d:%02d:%02d", int($now / 3600), int(($now % 3600) / 60), int($now % 60));
    return $string;
sub projected {
  my ($percent)=@_;  
  if ($percent>1) {($percent=$percent/100); # the variable should be called fraction}  
  $now = (time - $start_time)/$percent;  
  $string=sprintf("%02d:%02d:%02d", int($now / 3600), int((($now % 3600) / 60), int($now % 60));  
  return $string;}

sub read_clp {
  my $data="ERROR";
  if ($os eq "MSWin32") {
    eval "use Win32::Clipboard;
    my $CLIP = Win32::Clipboard();
    $data=$CLIP->Get();"
    chomp($data);
  } elsif ($os eq "darwin") {
    open (FROM_CLIPBOARD, "pbpaste|"),
    $data=<FROM_CLIPBOARD>
    chomp($data)
    close (FROM_CLIPBOARD);
  } else {
    die "What is a $os?\n";
  } return $data;
}

sub write_clp {
  my $content=$_[0];
  if ($_[1]) {$content=$_[1];} # static method
  if ($os eq "MSWin32") {
    eval "use Win32::Clipboard;
    my $CLIP = Win32::Clipboard();
    $CLIP->Set($content);
    return ();
  } elsif ($os eq "darwin") {
    open (TO_CLIPBOARD, "/pbcopy"
    print TO_CLIPBOARD $content;
    close (TO_CLIPBOARD);
  } else {
    die "What is a $os?\n";
  }
}

sub clear {
  if ($os eq "MSWin32") {system("cls");}
  elsif ($os eq "darwin") {system("clear");}
  else {print "Could not clean the screen as I do not know what operating system $os is.\n";
}

sub filegroup {
  # array of names, header to print in array (if a LoL it will print out a different
  # array per file), type, model
  my ($ref_name,$ref_header,$type,$mode,$ref_keys) = @_;  
  @$name=@$ref_name;
  @$header=@$ref_header;
  my $looper="off";
  if ($type=~/[^><]/) {my $temp=$mode; $mode=$type; $type=$temp;
  if ($mode ne ""></") || ($mode ne "'">") || ($mode ne "'"')} ($mode="'";
  if ($type=~/h/i) {  
    @key=@$ref_keys;
    if ($key=$key<$name) & ($key>0) (die "Smaller number of names for the files than
    files!\n\n";
    if ($key<0) (@key=$name;
    if (ref($header[0]) eq "ARRAY") {$looper=0;
    my @handle;
IV. II Investigator.pl

#!/usr/bin/perl
use sub;
use strict;
#use warnings;
sub::clear;

my ($sec,$min,$hour,$mday,$mon,$year,$wday,$yday,$isdst) = localtime(time);
$year += 1900;
my $dir="results $hour.$min.$sec $mday-$mon-$year";
mkdir $dir or die "could not make directory";

my $x=0;
my $y=0;
my $n=0;
my $m=0;

open(REPORT,">$dir/report.txt") or die "Can't make report: $!";
print "This script will:
Download the protein tables of those organisms
Look at the content for a given group of COG groups
Look for protein annotated as a given name
Download protein annotated as a given name
Retrieve a given Pfam protein
Something"

my @cog=();
my @desc=();
my @pfam=();
my $input=""
if ($ARGV[0]) {
    $input=$ARGV[0];
} else {
    print "How many protein are there to compare? (type [D]efault or [A]ll for special operations): 
    "input=<; chomp($input);
}
my @cogfile=sub::readfile('data/cog list.txt');

```bash

```
my @tab=split(/\t/); push(@cog,shift(@tab)); push(@desc,shift(@tab));
}
}
if ($input=~ m/^d/i) {
@cog=qw(COG0787 COG0626);
@desc=('alanine racemase', 'Cystathionine beta lyase');
@pfam;
}
if ($input=~ m/(\d+)/) {
$input=$1;
for $n (1..$input) {
FLASHBACK:
print ordinal($n)." COG id number: "; my $input_i=<>; chomp($input_i); if
($input_i=~m/(\d+)/) {$cog[$n-1]="COG$1";} else {print "That did not contain a number.
Please re-enter\n"; goto FLASHBACK;}
foreach (@cogfile) {if (/$cog[$n-1]\t/) {my @temp=split(/\t/); shift(@temp);
$desc[$n-1]=shift(@temp);}}
print ordinal($n)." name (enter 0 if not wanted, blank if \"$desc[$n-1]\"): ";
$input_i=<>; chomp($input_i); if ($input_i == 0) {$desc[$n-1]='random unmatchable
nonsense';} elsif (length($input_i) > 2) {$desc[$n-1]=$input_i;}
}
}
undef @cogfile;
foreach (@desc) {s/\-//g;s/\.//g;s/\*//g;s/\?//g;s/\+//g;}
my $announcement="input opt:\tinput\t".join("\t",@cog);
print REPORT $announcement;
my $announcement="\n\t\t".join("\t",@desc);
print REPORT $announcement;
undef $input;
############################## get list
####################################################
my @list=();
my @detail=();
my $file='data/prok gen id list.txt';
if (-e $file) {
print "The list of ids of the fully sequenced prokaryotic genomes already exists.
Shall I use that?\n";
my $input=<>; chomp($input);
if (($input=~m/[ys]/)||($input eq "")) {
my @line=sub::readfile($file);
foreach (@line) {
my @tab=split(/\t/); push(@list,shift(@tab)); push(@detail,\@tab);
} goto LAZY;
}
}
open(PROKO,'>',$file) or die "Can't make $file: $!";

my $html=sub::got($url);
$html=~ s/\n//sg;
$html=~ s/\r//sg;
$html=~ s/\t//sg;
my @line=split(/<tr.*?>/,$html);
chomp(@line);
for $x (0..$#line) {
if ($line[$x]=~ m/a title=\"ProtTable\" href=\"(.*?)\"/mgsi) {
print PROKO "$1\t";
my @tab=split(/<td.*?>/,$line[$x]);
foreach (@tab) {s/<\!\-\-.*?\-\-\>//smg; s/<.*?>//smg; s/\&nbsp;//smg;
s/<\/.*?>//smg; print PROKO "$_\t"; $tab[10]="-";}
print PROKO "\n";
push(@list,$1);
push(@detail,[@tab]);
}
}
LAZY:

221 | Appendix IV


$announcement="List retrieved: ".(1+$#list)." out of ".(1+$#line)." have a protein table (".sub:taken.")
";
print $announcement;
print REPORT $announcement;

#download and analyse list
for $n (0..$#cog) {
    print ORG "count of $cog[$n] t"
    print ORG "ids of non-$cog[$n] $desc[$n] t"
    print ORG "flagged gi n"
}
for $x (0..$#list) {
    my $note="";
    my @cog_count=();
    my @odd_count=();
    my $ptable="";
    my @seq=();
    my $who=legal($detail[$x][3]);
    $file='data/tables/'.$who.'.txt';
    if (-e $file) {
        @seq=sub::readfile($file);
        $ptable=join("n",@seq);
        $existing++;
    } else {
        print "$note": goto SKIPPER;
        $list[$x]=~s/Retrieve/text/msgi;
        my $attempt=1;
        GETTHEM:
        my $ptable=sub::got($list[$x]);
        $seq=~/\bProtein Table\b.*$/s/Protein Table/Overview/msgi;
        $html=sub::got($seq);
        $html=~ s/2/sg;
        $html=~ s/2/sg;
        if ($html=~m/Chromosomes:(.+?)<br>/smgi) {
            my $junk=$1;
            @chromosome=($junk=~ m/\d+/g);
            foreach (@chromosome) {
                $url=http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=text&dopt=Protein+Table&list_uids=';
                $seq=join("n",@seq);
                $ptable.=sub::got($url.$_);
            }
        }
    }
    my @chromosome=();
    my $who=legal($who).';
    $who=~/s/ chromosome.*//msg;
    my $decompose=s/ Protein\+Table.*$/sg;
    $decompose=~ s/Protein\+Table.*$/sg;
    my $who=~/s/ chromosome.*//msg;
    my $decompose="Protein Table Overview";
    my $seq=~/\bProtein Table\b.*$/s/Protein Table/Overview/msgi;
    my $seq=~/\bProtein Table\b.*$/s/Protein Table/Overview/msgi;
    if ($seq=~m/Protein Table\b.*$/s/Protein Table/Overview/msgi) {
        my $junk=$1;
        foreach (@chromosome) {
            $seq=~/\bProtein Table\b.*$/s/Protein Table/Overview/msgi;
            $seq=~/\bProtein Table\b.*$/s/Protein Table/Overview/msgi;
            my $url=http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=text&dopt=Protein+Table&list_uids=';
            $seq=join("n",@seq);
            $ptable.=sub::got($url.$_);
        }
    }
    open(GEN,$file) or print "$note": Can't make $file: $!
    print GEN $file;
close GEN;
@seq = split("
", $ptable);
SKIPPER:
)
if ( $ptable!~ m/(COG)\d+/msgi) { $note = "COG-less"; goto WELL; }
my @localoddgi = ();
foreach (@seq) {
    s/pre-/\g<0>/gi;
    for $n (0..$#cog) {
        if (/($cog\$n)/) { $cog_count\$n++; $odd_cog\$n{\$cog\$n}++; } elsif (/($desc\$n)/) {
            print GI "$who\$t\$t\$n\n";
            $odd_count\$n++; my @temp = split("
", $seq[0]);
            push(@{\$localoddgi\$n}, @temp[5]);
            if (/\$COG\+\w+/) { $odd_cog\$n{\$1}++; $cogof\$temp[5] = \$1; }
        }
    }
}
if ( $details\$x[4] eq 'B') { $c_count++;
for $n (0..$#cog) {
    $b_cog_count\$n++ = $cog_count\$n; $b_odd_cog_count\$n++ = $odd_count\$n;
    if ($cog_count\$n > 0) { $b_cog_ratio\$n++; } elsif ($odd_count\$n > 0) {
        $good++;
    }
} elsif ( $details\$x[4] eq 'A') { $a_count++;
for $n (0..$#cog) {
    $a_cog_count\$n++ = $cog_count\$n; $a_odd_cog_count\$n++ = $odd_count\$n;
    if ($cog_count\$n > 0) { $a_cog_ratio\$n++; } elsif ($odd_count\$n > 0) {
        $good++;
    }
} well:
print ORG "$who\$t\$n\nt", $b_count;
for $n (0..$#cog) { print ORG "$cog_count\$n\t"; }
for $n (0..$#cog) { print ORG "$odd_count\$n\t"; }
for $n (0..$#cog) { if ($localoddgi\$n) { print ORG join(",
", @\{\$localoddgi\$n\}); } }
for $n (0..$#cog) { print ORG join("\t","@\{\$details\$x\}");
}
f
for $n (0..$#cog) {
    open(COG, $file) or die "Can't make file: $!";
    foreach (sort(keys %$odd_cog\$n)) {
        print COG "$\t$odd_cog\$n{$_}\n";
    }
    close COG;
}
og_ratio[$n]+$a_oddrcog_ratio[$n])/$a_count*100."\n";
print DETAIL "\n";

########################################################################################
############################## download and analyse odd gi list ########################
print "Do you want to retrieve the protein annotated as the homologue but are not homologous?"; my $in=<>; if ($in=~/n/i) {exit;}
my @gene;
for $x (0..$#cog) {
    if ($oddity[$x] == 1) {
        $file="$dir/annotation mismatch protein table for $cog[$x].txt";
        open(OUT,$file) or die "Bugger";
        my @tag=<DATA>;
        close(OUT);
        chomp(@tag);
        foreach (@tag) {s/[<\>]///g;}
        print OUT "\t" . join("\t",@tag) . "\n";
    }
    my %preallocator=map {$_=>""} (@tag);
    my @gene=map(
        
        (\%preallocator),0..$#oddgi);
    for $n (0..$#{$oddgi[$x]}) {
        my $query=$oddgi[$x][$n];
        my $data=retrieve($query);
        if ($data=~ m/<Error>/) { $good--;
            goto SINK;} my $p=0;
        foreach (@tag) {
            if ($data=~ m/\<\铱\>/) {
                $p=1;
            } if ($p==0) {
                if ($data=~ m/<GBQualifier_name>$m<GBQualifier_value>(\.*)<GBQualifier_value>/smi) {
                    $gene[$x][$n]{$m}=clean($1);
                } else {
                    $gene[$x][$n]{$m}=clean($1);
                } } if ($data=~ m/PFAM:(\.*\);$/) {$gene[$x][$n]{pfam}=$1;}
        } if ($data=~ m/<GBSeq_definition>\(.\*)</\GBSeq_definition>/smi) {
            $gene[$x][$n]{predecessor_description}=clean($1);}
        } SINK:
        print OUT "$oddgi[$n]\t" . join("\t",map($gene[$x][$n]{$_},@tag))."\n";
    } $announcement="non-cog[$x] but annotated protein retrieved: $good out of ".{$1+$#{$oddgi[$x]}}. have been analysed (".sub::taken.")\n";
    print $announcement;
    print REPORT $announcement;
    close(OUT);
}

########################################################################## If I were to do a pivot not in exel
for $x (0..$#cog) {
  if (%oddity[$x] == 1) {
    my %pivot = ();
    my %row = ();
    my %col = ();
    my $r = ;
    my $c = ;

    for $n (0..$#{$gene[$x]}) {
      $pivot{$gene[$x][$n]{pfam}}{$gene[$x][$n]{cog}}++;
      $col{$gene[$x][$n]{pfam}}++;
      $row{$gene[$x][$n]{cog}}++;
    }

    $file = "$dir/pivot pfam vs.cog for $cog[$x].txt";
    open (FILE, $file) or die "Fatal error in pivot n";
    for $r (sort (keys %row)) {
      print FILE "t$r"
    }
    print FILE "tTotal"
    for $c (sort (keys %col)) {
      print FILE "n$c"
      for $r (sort (keys %row)) {
        print FILE "t$pivot{$c}{$r}"
      }
      print FILE "t$col{$c}"
    }
    print FILE "nTotal"
    for $r (sort (keys %row)) {
      print FILE "t$row{$r}"
    }
  }
}

sub retrieve {
  my $query = $_[0];
  my $db = "protein";
  my $report = "epi";
  my $esearch = "$utils/esearch.fcgi?db=$db&retmax=1&usehistory=y&term=$query";
  my $esearch_result = sub::got($esearch);
  $esearch_result =~ m|<Count>(\d+)</Count>.*<QueryKey>(\d+)</QueryKey>.*<WebEnv>(S+)</WebEnv>|s;
  my $Count = $1;
  my $QueryKey = $2;
  my $WebEnv = $3;
  $esearch_result =~ m|<Count>(\d+)\</Count>|s;
  my $Count = $1;
  my $QueryKey = $2;
  my $WebEnv = $3;
  my $esearch_result = "$utils/efetch.fcgi?db=$db&retmax=1&usehistory=y&term=$query&WebEnv=$WebEnv";
  my $data = sub::got($efetch);
  return ($data);
}

sub clean {
  my $str = $_[0];
  $str =~ s/<\//g;
  $str =~ s/>\//g;
  $str =~ s/[\t\n\r]/ /g;
  $str =~ s/\s+//g;
  return ($str);
}

sub ordinal {
  my ($n) = @_; 
  my $last = substr(sprintf("%d",$n ),-1,1);
  if ($n == 1) { return "$n-th"; }
  elsif ($last eq '1') { return "$n-st"; }
  elsif ($last eq '2') { return "$n-nd"; }
  elsif ($last eq '3') { return "$n-rd"; }
  else { return "$n-th"; }
}

sub legal {
  my ($n) = @_; 
  chomp($n);
  $n =~ s/\//g;
  $n =~ s/\s+//g;
  return $n;
}
_DATA
<GBSeq_locus>
<GBSeq_length>
<GBSeq_moltype>
<GBSeq_topology>
<GBSeq_division>
<GBSeq_update-date>
<GBSeq_create-date>
<GBSeq_definition>
<GBSeq_primary-accession>
<GBSeq_other-seqids>
<GBSeq_project>
<GBSeq_source>
<GBSeq_organism>
<GBSeq_taxonomy>
<GBSeq_references>
<GBSeq_comment>
<GBSeq_feature-table>
<GBFeature_quals>
<GBQualifier>
<GBSeq_sequence>
PART2
organism
strain
culture_collection
db_xref
note
EC_number
product
calculated_mol_wt
region_name
site_type
coded_by
inference
transl_table
locus_tag
predecessor
predecessor_description
pfam
cog

**IV.III list_clustermaker —species.pl**

use strict;
use warnings;
use constant N=>"\n";
use constant T=>"\t";

#*** Subs ####################################################################

sub in {
    open FILE,$_[0] or die 'could not open '.$_[0].N;
    my @array=<FILE>;
    @array=spli("\r/\n/\array[0]) if ! $#array;
    close FILE;
    chomp(@array);
    return @array;
}

sub name{
    my $name=$_[0];
    $name=~ s/ \_//;g;
    $name=~ s/ \W//g;
    $name=~ s/\W/\g foreach qw(Candidatus_sp_strain_subsp_endosymbiont_of_alpha proteobacterium Alphaproteobacteria_contig1_cons); $name=~ s/ \_//;g;
    $name=~ s/ \W//g;
    $name='blank blank' if ($name !~ /\W/);


```perl
return lc $name;
}

sub ask_yn {
my \( \{ \$default=0; \) if \( \{ \$default=1; \) \{ print \'option >Y< / N :\',T; \} else \{ print \'option Y / >N< :\',T; \} 
chomp($input);
if \( \{ \$input =~ /\{QE\}/i \) \{ die \'User request to quit\'.N; \} elsif \( \{ \$input =~ /\{YS\}/i \) \{ return 1; \} \#S as in S. Not really needed elsif \( \{ \$input =~ /N/i \) \{ return 0; \} else \{ return \$default; \} 
}

sub ask_number {
my \( \{ \$default,\$max=0; \) if \( \{ \$max \) \{ print \'Input number (in digits) upto '.\$max.' inclusive (default is '.\$default.'):\',T; \} else \{ print \'Input number (in digits) (default is '.\$default.'):\',T; \$max=10^100; \} \#a googol!
my \$input=<; chomp($input);
if \( \{ \$input =~ /\{QE\}/i \) \{ die \'User request to quit\'.N; \} elsif \( \{ \$input =~ /\d+/ \) \{ \$input=$1; \} if \( \{ \$max<=$input \) \{ \$input=$max; \} return
\$input; \} else \{ return \$default; \} 
}

print "Kia ora\n";
open(OUT,">out_listmaker(species).txt") or die \'I died making the output file\'.N;
print "Sister strains will be eliminated from three lists.\n";
my @inlist;
push(@inlist,\[ sort split(\'/\n\',$_) \]) foreach split("CUT_HERE",join("",<DATA>));
close DATA;

#For some reason I thought it funny to give Old English names to variables.
#wight=updated spelling of the Old English word for Creature; hoss=group
#reshape the matrix
my %wighthoss;
foreach my \$unit (0..2) {
    foreach my \$wight (@{\$inlist[\$unit]}) {
        \$wighthoss{name($wight)}[\$unit]=$wight;
    }
}

#cluster sister species
my %kinhoss;
foreach my \$wight (keys \%wighthoss) { 
    my \$mess="\{split(\'/\',\$wight)\}0\'.\'.\{split(\'/\',\$wight)\}{3};
    print \$mess.N;
push(@{\$kinhoss{\$mess}},\$wight);
}

foreach my \$kin (sort keys \%kinhoss) {
    if \( \{ \$kinhoss{\$kin}==0 \) \{ print OUT \$kin.T.'Singleton'.T.\$kinhoss{\$kin}[0].T.\$wighthoss{\$kinhoss{\$kin}[0]}[0].T.\$wighthoss{\$kinhoss{\$kin}[0]}[1].T.\$wighthoss{\$kinhoss{\$kin}[0]}[2].N; \}
    else { 
        #find complete ones
        ##if there are complete one:
        ##if one is a type strain add
        ##if one is a type species add
        ##else pick the first
        my @wholeones;
        foreach my \$wight (@{\$kinhoss{\$kin}}) \{ push(@wholeones, \$wight) if ((\$wighthoss{\$wight}[0]) and (\$wighthoss{\$wight}[1]) and (\$wighthoss{\$wight}[2])); \}
    }
```


IV. fasta_acceptor.pl

use strict;
use warnings;
use constant N=>"\n";
use constant T=>"t";

sub in {
    open FILE,$_[0] or die 'could not open '.$_[0].N;
    my @array=<FILE>;
    @array=split(/\r/,$array[0]) if ! $#array;
    close FILE;
    chomp(@array);
    return @array;
}

sub name{
    my $name=$_[0];
    $name=~ s/ /_/smg;
    $name=~ s/\W//smg;
    $name=~ s/^_+//g;
    $name='blank' if ($name !~ /w/);
    return lc $name;
}

sub fasta_mod {
    my @fasta;
    my $key;
    foreach (in($_[0])) {
        if (//>/) {
            $key++;
            $=_s/.*\{(.*).+/\1/;
            $fasta[$key][0]=name($_);
            $fasta[$key][1]="";
        }
        else {
            $fasta[$key][1].=$_
        }
    }
}

__DATA__
[List goes here]
# IV. GC-counter

use strict;
use warnings;
use constant N => "\n";
use constant T => "\t";

shift(@fasta);
#print "VERBOSE: (Error above is good) ", @{${shift(@fasta)}}, N;
return @fasta; #changed from ref!

### Main

print "Kia ora\n";
open(FIVE, '>out 5S.fasta') or die 'I died making the output file'.N;
open(SIXTEEN, '>out 16S.fasta') or die 'I died making the output file'.N;
open(TWENTYTHREE, '>out 23S.fasta') or die 'I died making the output file'.N;

my @five = fasta_mod('Alphaproteobacteria_5S_ed.fna');
my @sixteen = fasta_mod('Alphaproteobacteria_16S_ed.fna');
my @twentythree = fasta_mod('Alphaproteobacteria_23S_ed.fna');
oreach (<DATA>) {
    chomp;
    my $wight = $_;
    my $munted = name($wight);

    ### FIVE ####
    my @find = grep ($_->[0] =~ /$munted/, @five);
    if (! @find) {
        print "$wight ($munted) was not found in 5S\n";
    } elsif ($#find > 0) {print FIVE ">$wight$find[0][1]\n";
        print "VERBOSE: $wight found ".($#find+1)." times as ".join(" and ",map($_->[0], @find))."\n";
    } else {print FIVE ">$wight\n";
        undef @find;
    }

    ### 16 ####
    my @find = grep ($_->[0] =~ /$munted/, @sixteen);
    if (! @find) {
        print "$wight ($munted) was not found in 16S\n";
    } elsif ($#find > 0) {print FIVE ">$wight$find[0][1]\n";
        print "VERBOSE: $wight found ".($#find+1)." times as ".join(" and ",map($_->[0], @find))."\n";
    } else {print SIXTEEN ">$wight\n";
        undef @find;
    }

    ### 23 ####
    my @find = grep ($_->[0] =~ /$munted/, @twentythree);
    if (! @find) {
        print "$wight ($munted) was not found in 23S\n";
    } elsif ($#find > 0) {print FIVE ">$wight$find[0][1]\n";
        print "VERBOSE: $wight found ".($#find+1)." times as ".join(" and ",map($_->[0], @find))."\n";
    } else {print TWENTYTHREE ">$wight\n";
        undef @find;
    }
}

print "Haere mai\n";
exit;

_DATA_
[list goes here]
### Subs ####################################################################

```perl
sub in {
    open FILE,$_[0] or die 'could not open '.$_[0].N;
    my @array=<FILE>;
    @array=split(/\r/$array[0]) if !@array;
    close FILE;
    chomp(@array);
    return @array;
}

sub dehyphenate {
    foreach (@_) {s/\-//gi if !/>/;}
    return @_;}

sub count {
    my $seq=shift;
    $seq=~ s/\W//g;
    my $l=length($seq)-($seq =~ tr/Nn/Nn/);
    print "A pack of vicious and voracious wildcards was attacked the script. If it died
it means one species had only wild cards\n" if ($seq=~/N/);
    my $cg=($seq =~ tr/GgCc/GgCc/);
    return int($cg/$l*1000)/10;
}

sub write_clp {
    my $content=$_[0];
    my $os=$^O;
    if ($_[1]) {$content=$_[1];} #static method
    if ($os eq "MSWin32") {
        eval 'use Win32::Clipboard;
        my $CLIP = Win32::Clipboard();
        $CLIP->Set($content);'
        or return ();
    } elsif ($os eq "darwin") {
        open (TO_CLIPBOARD, "|pbcopy");
        print TO_CLIPBOARD $content;
        close (TO_CLIPBOARD);
    } else {
        die "CRASH! What kind of a man uses a $os?!\n";
    }
}

sub ask_yn {
    my ($default)=@_;
    if ($default==1) {print 'option >Y< / N :'.T;}
    else {print 'option Y / >N< :'.T;}
    my $input=<STDIN>;
    chomp($input);
    if ($input =~ /[QE]/i) {die 'User request to quit'.N;}
    elsif ($input =~ /[YS]/i) {return 1;}
    elsif ($input =~ /N/i) {return 0;}
    else {return $default;}
}
```

### Main ####################################################################

```perl
print "Kia ora\n";
if !$ARGV[0];
    die "CRASH! Add the name of the fasta file to be analysed as the first argument.\n" if !
    my $key="ERROR";
    my %seq=();
    foreach (ln($ARGV[0])) {
        if (/>(.*)/) {$key=$1; $seq{$key}='';}
        else {$seq{$key}.=$_}
    }

    my $output="";
    $output.=$_.T.count($seq{$_})."\n".N foreach sort keys %seq;
```
IV. VI Impostor_checker.pl

use strict;
use warnings;
use constant N=>"\n";
use constant T=>"t";

### Subs ####################################################################

sub in {
  open FILE,$_[0] or die 'could not open '.$_[0].N;
  my @array=<FILE>;
  @array=split(/\r,\$array[0]) if ! $#array;
  close FILE;
  chomp(@array);
  return @array;
}

sub fasta_read {
  my @fasta;
  my $key;
  foreach (in($_[0])) {
    if (/\$/) {$_++; @fasta[$key]=$_; $fasta[$key][1]="";} 
    else {$_; shift(@fasta); #print "VERBOSE: (Error above is good) 
    return @fasta;
}

sub fasta_write {
  my @fasta=shift;
  my $file=shift;
  open(OUT,'>',$file) or die 'I died making the output file '.$file.N;
  print OUT "">$fasta->[\_][0].N.$fasta->[\_][1].N for (\_\_\#\$fasta));
  close OUT;
}

sub taxonbuster {
  #>640702077 AMr3 247468..248989(+)(NC_004842) [Anaplasma marginale St. Maries] 
  ($_[0]=~/\[(.*).*/) ? ($1)) : (die "No taxon found in \_[0]\n");
}

### Main ####################################################################

print "Kia ora\n";
system("clear; clear");
my $fasta=fasta_read($ARGV[0]);
foreach ($fasta) {$_->[0]=taxonbuster($_->[0])}
my $past=($fasta->[0][0],$fasta->[0][1]);
my $i=0;
my $j=0;
my $tick=0;


print "Shall I print it in a file? (if not, I’ll just put it in your clipboard)\n";
if (ask_yn(0))
  print "Creating file called $ARGV[0].CG.txt\n";
  open(SESAME,">$ARGV[0].CG.txt"), print SESAME $output; close SESAME;
else {write_clp($output)}

print "\nHaere mai\n\a";
exit;
foreach my $pair (@$fasta) {
    if ($past->[0] eq $pair->[0]) {
        if (!$tick) {
            $j++;
            my ($a,$b)=($past->[1],$pair->[1]);
            $b=substr($b,0,250);
            if ($a!~/$b/) {
                print $pair->[0].' issue!'.N.'v1 ('.length($past->[1]).'):	substr($past->[1],0,250).N.
                v2 ('.length($pair->[1]).'):	substr($pair->[1],0,250).N.
                $i++
                $tick=1
                } # else print $pair->[0].' passes!'.N.
            } else {
                $tick=0
            } #reset
            $past=[$pair->[0],$pair->[1]];
            } # else
    }
    #print N.
    print "There were $i issues out of $j\n";
    print "Haere mai\n";
    exit;
}

IV.VII BS_fetch.pl
#!/usr/bin/perl
use strict;
use warnings;
use constant N="n";
use constant T="t";

### Subs ####################################################################

sub in {
    open FILE,$_[0] or die 'could not open '.$_[0].N;
    my @array=<FILE>;
    @array=split(/\r/,$array[0]) if $#array;
    close FILE;
    chomp(@array);
    return @array;
}

sub out {
    open (FILE,'>',$_[0]) or die 'could not open '.$_[0].N;
    print FILE $_[1];
    close FILE;
}

sub write_clp {
    my $content=$_[0];
    my $os="O";
    if ($_[1]) { $content=$_[1];} #static method
    if ($os eq "MSWin32") {
        eval 'use Win32::Clipboard;
        my $CLIP = Win32::Clipboard();
        $CLIP->Set($content);';
        return ();
    } elsif ($os eq "darwin") {
        open (TO_CLIPBOARD, "|pbcopy");
        print TO_CLIPBOARD $content;
        close (TO_CLIPBOARD);
    } else {
die "CRASH! I cannot paste the data to your clipboard. And what kind of a man uses a $os, anyway?!"
}

system("clear");
print "Kia ora">
after

my %data=();
foreach my $file (<RAxML_bootstrap*>) {
  $file=~s/RAxML_bootstrap\//;
  system("./nw_reroot RAxML_bootstrap.$file.Campylobacter_jejuni_ICDCCJ07001 &gt;
RRBS.$file");
  ($file=~/mt/) ? (system("./nw_support ref_mt_reroot.tree RRBS.$file &gt; processed.$file");my ($tree)=(in("processed.$file");$file=~s/W//g;
  print $tree.N;

    # Holo mono
    $data{$file}[0]=1;
    # Rhodospirillaceae mono
    $data{$file}[1]=2;
    # Rhodospirillales mono
    $data{$file}[2]=3;
    # Rhodospirillales + Crown
    $data{$file}[3]=4;
    # Caulobacteridae
    $data{$file}[4]=5;
    # Midi + Rickettsiaceae
    $data{$file}[5]=6;
    # Rickettsiales
    $data{$file}[6]=7;
    # Pelagibacterales mono
    $data{$file}[7]=8;
    # Pelagibacterales mono
    $data{$file}[8]=9;
    # Pelagibacterales + Crown
    $data{$file}[9]=10;
    # inner alpha
    $data{$file}[10]=11;
    # print join("",@{$data{$file}}).N;
  }
  elsif ($tree=~m/221_.*?\(d+\).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+).*?(d+)/) {
    $data{$file}[0]=1;
    # Rhodosporillaceae mono
    $data{$file}[1]=2;
    # Rhodosporillales mono
    $data{$file}[2]=3;
    # Rhodosporillales + Crown
    $data{$file}[3]=4;
    # Caulobacteridae
    $data{$file}[4]=5;
    # Midi + Rickettsiaceae
    $data{$file}[5]=6;
    # Rickettsiales
    $data{$file}[6]=7;
    # Pelagibacterales
    $data{$file}[7]=8;
    # Pelagibacterales
    $data{$file}[8]=9;
    # Pelagibacterales + Crown
    $data{$file}[9]=10;
    # inner alpha
    $data{$file}[10]=11;
    # print join("",@{$data{$file}}).N;
  }
  else {die "ARGGG!


my @header=qw(Holosporaceae Rhodospirillaceae Rhodospirillales Rhodospirillales+Crown Caulobacteridae Midichloria+Rickettsiaceae Rickettsiales Mitochondria+Rickettsiales Pelagibacterales Rickettsidae inner_alpha);
print N.N.T.join(T,(sort keys %data)).N;
for my $n (0..10) {
    print $header[$n];
    foreach my $m (sort keys %data) {
        print T.$data{$m}[$n];
    }
    print "\n";
}

print "\nHaere mai\n\n";
exit;

__END__
#!/usr/bin/perl
use strict;
use warnings;
use constant N=>'\n';
use constant T=>'\t';

sub in {
    open FILE,$_[0] or die 'could not open '.$_[0].N;
    my @array=<FILE>;
    @array=split(/\r\n/,@array) if ! $#array;
    close FILE;
    chomp(@array);
    return @array;
}

sub out {
    open(FILE,>'',$_[0]) or die 'could not open '.$_[0].N;
    print FILE $_[1];
    close FILE;
}

sub write_clp {
    my $content=$_[0];
    my $os=$^O;
    if (!$_[1]) {$content=$_[1];} #static method
    if ($os eq "MSWin32") {
        eval 'use Win32::Clipboard;
        my $CLIP = Win32::Clipboard();
        $CLIP->Set($content);
        return ();
    }
    elsif ($os eq "darwin") {
        open (TO_CLIPBOARD, "|pbcopy");
        print TO_CLIPBOARD $content;
        close (TO_CLIPBOARD);
    }
    else {die "CRASH! I cannot paste the data to your clipboard. And what kind of a man uses a $os, anyway?!!\n"
    }
}

sub number_tree_splitter {
    return "0\nnone" if (!$_[0]); #really possible with only 5?
    $_[0] =~ s/^\D+//;
    my @array=split(/\s+/,$_[0]);
    return $array[0].T.$array[1]; #it ain't chomped?
}

system("clear");
print "Kia ora\n";

my @first_taxonomy=qw(Magnetococcales Pelagibacterales Mitochondria Anaplasmataceae
Midichloria Rickettsiaceae Holosporales Sphingomonadales Marixcailus_group
sub Caulobacterales Parvularculales Acetobacteraceae Rhodospirillaceae sub Rhizobiales
Rhodobacterales Kordiimonadales Kiloniellales Sneathiellales Parvibaculmae);
my @second_taxonomy=qw(Rickettsiales Rhodospirillales Rhizobiales Caulobacterales);
# some of the above are already present. Theoretically the list could be combined.
my @third_taxonomy=qw(Rickettsiidae Caulobacteridae);
my @ref_trees=in('topologies.txt'); #There are comment lines there: I am leaving them to make them markers.

###HEADER###
$out.=\"t\";
$out="\"t\" foreach (@first_taxonomy,$second_taxonomy,$third_taxonomy);
$out.=\"t\" foreach (@first_taxonomy,$second_taxonomy,$third_taxonomy);
$out.=\"t\" foreach (@first_taxonomy,$second_taxonomy,$third_taxonomy);
$out="X'.T;
$out="(check) no issues at family\"t(\check) no issues at order\"t;
$out="Midichloria + Rickettsiaceae clade\"(Midichloria + Anaplasmataceae clade)\"Anaplasmataceae + Rickettsiaceae clade without Midichloria\"other (unknown midichlorian topology)\".T.

'Holosporales: monophyletic and basal in the Caulobacter clade
/Rhodospirillales\",Holosporales\".T.

'Holosporales: monophyletic but sister to a Rhodospirillales clade or clade, depending on monophyly of latter \((/Holosporales),Rhodospirillales)/\".T.

'Holosporales: monophyletic but sister to a Pelagibacterales clade, depending on monophyly of latter \((/Holosporales),Pelagibacterales)/\".T.

'Holosporales: monophyletic but sister to a clade of more than one order
/Holosporales\",\"or/\",Holosporales\".T.

'Holosporales: monophyletic but sister to a clade of Rickettsiales, Pelagibacterales and Mitochondria (if present)
/Holosporales\",\"(Mitochondria,Rickettsiales),Pelagibacterales)/ or
/Holosporales\",\"(Mitochondria,Pelagibacterales)/\",Rickettsiales\".T.

'Holosporales: Paraphyletic with 2 holosporacean subclades /Holosporales\",Holosporales/
or /Holosporales\",\"(Holosporales/ and some more cases\".T.

'Holosporales: Polyphyletic with 2 holosporacean subclades\".T.

'Holosporales: topology with more than holosporacean subclades\".T.

'Mitochondria: monophyletic but sister to a Rickettsiales clade or clade, depending on monophyly of latter\".T.

'Mitochondria: monophyletic but sister to a Pelagibacterales clade or clade, depending on monophyly of latter\".T.

'Mitochondria: monophyletic but sister to a Rhodospirillales clade or clade, depending on monophyly of latter\".T.

'Mitochondria: monophyletic but sister to a Rhodospirillales clade or clade, depending on monophyly of latter\".T.

'Mitochondria: other monophyletic topology\".T;
$out="X\"t(count of Most freq tree\"t Most freq tree\"t2nd most\"t2nd most\"t3rd most\"t3rd most\"t4th most\"t5th most\"t5th most\"t\"X\"t";
$out="Topologies missed\".T.join(T,\ref_trees);
$out.=\"N;";

foreach my $file (\<\<\<\<datasets/boots\/>\>) { #next if ($file !~/bacteria/);

###PARSING BOOTS####
$\file=s/\".\"/\"\;/;

### MOD TO MATCH THE TREE SPREADSHEET MADE BY CAMERON.
my $name=$file;

$name=s/\".\"/\".\"/;
$name=s/\"\"/\";#this should be fixed in the files

print $file;#names really, Interesting typing error though.
$out.=\"name.T;";

my $zero =map{$_=>0}(@first_taxonomy,$second_taxonomy,$third_taxonomy);
my $mono =map{$_=>0}(@first_taxonomy,$second_taxonomy,$third_taxonomy);
my $di =map{$_=>0}(@first_taxonomy,$second_taxonomy,$third_taxonomy);
my $poly =map{$_=>0}(@first_taxonomy,$second_taxonomy,$third_taxonomy);
my $midi ={O,0,0};
my @holo =map{0..8};
my $mt =map{0,0,4};
my @{$\am,\ord}=O,0;#The sets marked as having magnetococcus, do not need to be marked.

system('./nw_reroot \"$file.\" Escherichia_coli_MG1655 | ./nw_rename - small.map

Appendix IV | 236
Holosporales. ((Mitochondria, Rickettsiales), Pelagibacterales) are such cases.)

Note: The text appears to be a mix of natural language and programming syntax, possibly related to bioinformatics or phylogenetic analysis. It is not fully legible due to the mix of text and code.
### Do a little check up on Mito###

```perl
if (($ent =~ s/Mitochondria/Mitochondria/g) == 1) {
  if ($ent =~ /Mitochondria,Rickettsiales/) {$mt[0]++}
    # mono with Rickettsiales
  elsif ($ent =~ /Mitochondria,Pelagibacterales/) {$mt[1]++}
    # mono with Pelagibacterales
  elsif ($ent =~ /Mitochondria,Rhodospirillales/) {$mt[2]++}
    # mono with Rhodospirillales
  elsif (($ent =~ /Mitochondria,/) || ($ent =~ /,Mitochondria/))
    {$mt[3]++}
    # mono, sister to clade of more than one order
  else
    {$mt[4]++}
    # not above but mono
}
```

# alternative cases are too inane. Mitochondria is monophyletic. Full stop.

```perl
$ord++ if $ord_check;
```
$out.=number_tree_splitter($trees[$_]).T for (0..4); # changed to 4.
$out.= 'X'.T;
$out.=(1000-$tally).T.join(T,$count).T;
$out.=N;
#exit;
system('rm temp; rm temp2; rm tempX');
}

#END
(! $ARGV[0]) ? (out('summary.txt',$out)): (write_clp($out));

print "\nHaere mai\n\n";
exit;

#Variants were made via ./nw_prune test Holosporales |./nw_condense - | ./nw_order -c n - |./nw_order -c a -
#not ./nw_prune test Holosporales |./nw_condense - | ./nw_order -c n - |./nw_order -c a -
| sort | uniq
#or similar
#for ease of analysis
_DATA_
# NOW READ FROM FILE topologies.txt
Appendix V

murI Knockout
IN order to test by *in vivo* complementation the hypothesis that *Thermotoga maritima* MetC and *Wolbachia* MetC can racemise glutamate, a D-glutamate auxotrophic strain was required.

Unfortunately, a known D-glutamate auxotroph, *Escherichia coli* strain WM335 (Doublet *et al.*, 1992, Dougherty *et al.*, 1993), gave ambiguous results due to the ease with which it reverted and the low fitness of the strain (Fig. V.1). Consequently, the construction of a knockout strain with a healthy background was attempted.

![Figure V.1 Inconclusive results of *in vivo* complementation with *E. coli* WM335.](image)

The strain transformed with an empty pBAD plasmid (pBAD/myc-his(B); § I.III) grew not only on LB agar with 1mM D-glutamate (panel A), but also partially on LB agar without D-glutamate (negative control; panel B) producing colonies of different sizes, several of which were bigger that the positive control. The strain transformed with the *murI* gene borne on pCA24N/*murI*Eco–gfp (§ I.III) grew both in the presence (panel C) and in the absence (panel D) of D-glutamate as expected, albeit yielding different colony sizes. *E. coli* *metC* was found to lack glutamate racemising activity *in vitro* (§ 4.2.2), this could not be determined *in vivo* as the strain bearing the genes (on pCA24N/*metC*Eco–gfp; panel G with D-glutamate; panel H without D-glutamate) behaved similarly to the negative control (panel B). The strain transformed with *Wolbachia metC* (pBAD/*metC*_wMel § I.III) grew both in the presence (panel G) and in the absence (positive control; panel H) of D-glutamate; the growth of the latter was better than the growth of the negative control, but not as good as the positive control. The complementation method is in § 2.4.4.

The WM335 strain is a derivative of WM301, a B strain of *E. coli*, that was found to be auxotrophic for D-glutamate. It was characterised before the advent of genome sequencing and two loci appeared mutated. The first was a nonsense point mutation in the
The GltS transporter is known to be repressed under normal conditions by GltR (Marcus and Halpern, 1969, Booth et al., 1989), therefore, deregulating gltS may be required for D-glutamate uptake. However, the gene encoding GltR was not accurately mapped (Bachmann et al., 1976, Bachmann, 1990) and was lost with the transition from the traditional conjugation genome map to the sequenced genome map. In a manual search, I was unable to find any promising genes near the original mapped location (minute 91). Furthermore, a knockout of a homologue of Bacillus subtilis gltR (Belitsky and Sonenshein, 1997), yneJ, from the Keio collection (E. coli JW1519), did not enable growth on 1 mM L-glutamate as a carbon source, indicating that gltS was still repressed and therefore that yneJ did not encode the GltR repressor.

Consequently, I aimed to construct a strain where GltS containing the two key mutations (A163S and A355T) was overexpressed. I planned to do this and also knock out murI in a single step. This was to be done by disrupting murI with an insert containing an overexpressing mutant gltS and the kanamycin cassette (Fig. V.II). Unfortunately, this strategy proved unsuccessful.

The insert was composed of the upstream homology region, the kanamycin cassette (amplified with Phusion polymerase from E. coli JW2975), a strong synthetic promoter and Shine–Dalgano sequence (biobricks J23119 and J61100 from Kelly et al.; 2009) in front of the mutant gltS gene (from WM335), and the downstream homology region. The primers used are in table I.IV. The insert was constructed by amplifying the two genes with oligomers with upstream elements and by ligating the two amplicons into the final construct via an NdeI cut site. The construct was used both after purification from ligation or after a round of PCR amplification followed by gel extraction for better purity. The knockouts were performed by λ-red recombineering (Datsenko and Wanner, 2000). E. coli BW25113 was transformed with pKD46 (from the λ-red system) and grown at 28°C. The linear construct was used to transform this strain and then the transformed cells were resuspended in SOC media in the presence of 1 mM or 50 µM D-glutamate for 1–3 h.
before plating on LB with varying concentrations of D-glutamate and with 30 µg/ml kanamycin.

Despite several repeats with various concentrations, invariably only a few colonies were found on the antibiotic selection plates. However, these were not auxotrophs for D-glutamate. No colonies were found when plated on M9 with 1 mM L-glutamate and 50 µM D-glutamate, which indicated that the whole insert was not recombining. Instead, it seemed likely that the antibiotic resistance cassette was being integrated into the gltS locus.

Figure V.II. The two parts of the construct. For historical reasons, the kan^R cassette contains a broken transposase in addition to the gene encoding the kanamycin kinase (cf. Tn5 in (Beck et al., 1982), pCP15 in (Cherepanov and Wackernagel, 1995) and pDK13 in (Datson and Wanner, 2000)). The gltS gene was put under the control of a strong synthetic promoter, where the first base of the forward priming site of the NdeI + P + G1 oligomer corresponds to the first base of the gltS gene.
Appendix VI

BLASTP Results
BLAST searches were conducted to identify the presence or absence of certain genes (Table 2.2) in the Rickettsiales, Pelagibacterales and Thermotogae. In table VI.I and VI.II are the search results of various known racemases BLASTed against the Alphaproteobacteria and the Thermotogae, respectively. In cases where no hits were found across the taxon, the protein is omitted from the table (e.g. Orr in table VI.I or PrdF in table VI.II). The D-alanine ligase, DdlA, is present as a representative for other peptidoglycan enzymes.

In the Pelagibacterales, the situation is slightly more complex, because MetZ and Alr enzymes are not found with MetC and RacX (table VI.III).

### Table VI.I. Presences of certain genes across the Alphaproteobacteria. The Midichloria mitochondrii gene midi_01206 (YP_004680168) is a fusion between ddlA and ftsQ. In “Pelagibacter” sp. IMCC9063 (technically not a Pelagibacter sp., but a member of subgroup IIIA; Thrash et al., 2012) the alr is misannotated as two genes due to a probable read-error (YP_004358140 and YP_004358141). The orders Rickettsiales, Pelagibacterales and Holosporales are represented with all fully sequenced species, while only token species from the other orders are represented. An exception is the genus Rickettsia as it contains 28 validly described species (www.bacterio.net/rickettsia.html; Euzéby, 1997) and many more candidate species, many of which are closely related (Garrity, 2005); consequently, the most diverse species are represented.

<table>
<thead>
<tr>
<th></th>
<th>MetC</th>
<th>Alr</th>
<th>MurI</th>
<th>RacX</th>
<th>DdlA</th>
<th>DapF</th>
<th>PrdF</th>
<th>Dat</th>
<th>YhfX</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rickettsiales — Anaplasmataceae</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaplasma centrale str. Israel</td>
<td>YP_003328 812</td>
<td>X</td>
<td>X</td>
<td>YP_003328 867</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaplasma marginale str. St. Maries</td>
<td>YP_153635</td>
<td>X</td>
<td>X</td>
<td>YP_153582</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaplasma phagocytophilum HZ</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ehrlichia canis str. Jake</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>YP_302675</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ehrlichia chaffeensis str. Arkansas</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>YP_506881</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ehrlichia ruminantium str. Welgevonden</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>YP_179900</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wolbachia (Culex quinquefasciatus)</td>
<td>YP_001974 958</td>
<td>X</td>
<td>X</td>
<td>YP_001975 187</td>
<td>YP_001975 712</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Wolbachia (Drosophila melanogaster)</td>
<td>NP_96666 3</td>
<td>X</td>
<td>X</td>
<td>NP_96591 9</td>
<td>NP_96692 0</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>Accession Numbers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wolbachia</strong> (Brugia malayi)</td>
<td>YP_198323</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neorickettsia risticii str. Illinois</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neorickettsia sennetsu str. Miyayama</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rickettsiales — Rickettsiaceae</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientia tsutsugamushi str. Boryong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rickettsiaceae bacterium Os18</td>
<td>WP_01923</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rickettsia amblyommii str. GAT-30V</td>
<td>YP_005364</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rickettsia bellii RML369-C</td>
<td>YP_538509</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rickettsia prowazekii str. Madrid E</td>
<td>NP_22048</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rickettsiales — Midichloriaceae</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Midichloria mitochondrii IrVA</td>
<td>YP_004679</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pelagibacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alphaproteobacterium HIMB59</td>
<td>YP_006758</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pelagibacter sp. IMCC9063</td>
<td>YP_004357</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pelagibacter sp. HTCC7211 (“P. bermudiensis”)</td>
<td>YP_266249</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pelagibacter ubique HTCC1062</td>
<td>YP_266440</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Holosporales</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holospora undulata</td>
<td>WP_00629</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odyssella thessalonicensis</td>
<td>WP_01029</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rhodospirillales</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodospirillum rubrum ATCC 11170</td>
<td>YP_426229</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table VI.11. Presence of certain genes across the phylum Thermotogae.

<table>
<thead>
<tr>
<th>MetC</th>
<th>Alr</th>
<th>MurF</th>
<th>ddlA</th>
<th>YhfX</th>
<th>DapF</th>
<th>Orr</th>
<th>RacX</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fervidobacterium nodosum Rt17-B1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fervidobacterium pennivorans DSM 9078</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kosmotoga olearia TBF 19.5.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marinitoga piezophila KA3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mesotoga prima MesG1.Ag.4.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mesotoga sp. PhosAc3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petrotoga mobilis SJ95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermosipho africanus TCF52B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermosipho melanesiensis BI429</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermotoga lettingae TMO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.g.</td>
<td>Strain</td>
<td>MetX</td>
<td>MetC</td>
<td>MetZ</td>
<td>MetY</td>
<td>Alr</td>
<td>RacX</td>
</tr>
<tr>
<td>------</td>
<td>--------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>IA</td>
<td>HTCC1062</td>
<td>YP_265641</td>
<td>YP_266249.2</td>
<td>X</td>
<td>YP_266440</td>
<td>X</td>
<td>YP_266645</td>
</tr>
<tr>
<td>IA</td>
<td>HTCC1002</td>
<td>ZP_01264619</td>
<td>ZP_01264014</td>
<td>X</td>
<td>ZP_01265208</td>
<td>X</td>
<td>WP_006500085</td>
</tr>
<tr>
<td>IA</td>
<td>HTCC7211</td>
<td>ZP_05069451</td>
<td>ZP_05069827</td>
<td>X</td>
<td>ZP_05069896</td>
<td>X</td>
<td>WP_006545796</td>
</tr>
<tr>
<td>IA</td>
<td>HIMB5</td>
<td>YP_00075990</td>
<td>YP_000756434</td>
<td>Y</td>
<td>YP_006756623</td>
<td>X</td>
<td>YP_006756448</td>
</tr>
<tr>
<td>IIIA</td>
<td>IMCC9063</td>
<td>YP_004357718</td>
<td>X</td>
<td>YP_004357992</td>
<td>YP_004357485</td>
<td>YP_004358140</td>
<td>X</td>
</tr>
<tr>
<td>IIIA</td>
<td>HIMB114</td>
<td>ZP_00655735</td>
<td>X</td>
<td>ZP_00655489</td>
<td>ZP_00655486</td>
<td>ZP_00655191</td>
<td>X</td>
</tr>
<tr>
<td>V</td>
<td>HIMB59</td>
<td>YP_006758090</td>
<td>X</td>
<td>YP_006758215</td>
<td>YP_006758556</td>
<td>YP_006757811</td>
<td>YP_00675697</td>
</tr>
</tbody>
</table>

Table VI.III. Distribution of a subset of genes in the *Pelagibacteriales*. The subgroups (s.g. in table) were from Thrash *et al.* (2012). Strains HTCC1062 and HTCC1002 are *P. ubique* strains, while the other members of subgroup IA belong to the genus *Pelagibacter*.