Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
The evolution of molecular chaperones and their clients

Thesis submitted to the
Department of Genetics, Faculty of Science, Trinity College Dublin in fulfillment of
the requirements for the degree of Doctor of Philosophy

2011

Tom Williams, B.A.

Supervisor: Dr. Mario A. Fares
Declaration

This thesis is submitted by the undersigned for the degree of Doctor of Philosophy at the University of Dublin and has not previously been submitted as an exercise for a degree at this or any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College.
Summary

Molecular chaperones are ancient and ubiquitous proteins that help other proteins in the cell to fold. In this thesis, we investigate the evolution of assisted protein folding from two perspectives: the evolution of the chaperone molecules themselves, and the evolution of the client proteins that depend upon them for folding.

First, we study the evolution of bacterial chaperonins, which make up one family of molecular chaperones. Given the large number of complete bacterial genomes now available, we develop a new, fast method for identifying functional divergence in the chaperonin genes of different bacterial lineages. This analysis reveals that intracellular pathogens have accumulated the most radical change in chaperonin genes, and also highlights a group of chaperonins that were acquired by lateral gene transfer from archaea. Further bioinformatic analyses suggest that these archaeal-type chaperonins are functional in bacteria, raising the question of what their function might be.

Next, we scale up the functional divergence method introduced for studying chaperonin genes to handle large datasets consisting of hundreds of complete proteomes. In order to avoid the difficulties in scaling up from gene trees to species trees, we introduce the clustering of species by shared patterns of functional divergence, rather than phylogeny. This enables us to extract useful information about high-level patterns of functional divergence from groups of organisms where the species tree is unknown, and perhaps even unknowable. To demonstrate the utility of our method, we apply it to 750 complete bacterial proteomes and visualize the most important patterns of
functional divergence. By breaking genes down by functional category, we recover a split between a conserved core set of functions and a more variable metabolic layer, in agreement with a number of previous reports. Unexpectedly, we also find that host-associated bacteria (such as pathogens and symbionts) generally experience less functional divergence than their free-living relatives, perhaps because they are constrained by their specific host environment.

In the oomycete *Achlya klebsiana*, the chaperone gene Hsp70 has a long, antisense open reading frame on the opposite strand of the DNA molecule. This antisense open reading frame has been proposed to encode a glutamate dehydrogenase, a proposition which, if true, would demonstrate the feasibility of sense/antisense coding of structurally dissimilar proteins. Using phylogenetic evidence, along with an analysis of selective constraints, we show that the antisense open reading frame is not a real gene, but is most likely an artifact of the unusually high level of sequence conservation in Hsp70. This result has significant implications for the evolution of the translation system, since it substantially weakens support for one of the leading hypotheses in the field.

Finally, we evaluate the effect, if any, of chaperone-assisted protein folding on the evolution of client proteins. By comparing the evolutionary rates of client and nonclient proteins in the gamma-proteobacteria, we present evidence that clients evolve faster than nonclients once other factors (including protein-protein interactions and expression levels) are accounted for. This result provides the first evolutionary evidence that chaperones influence the evolution of their clients, a topic that has been the subject of considerable speculation and experiment.
Acknowledgements

Firstly, I would like to thank IRCSET for funding my work over the last three years.

A number of people have played an important role in the development of this thesis. Chief among these is, of course, my supervisor, Mario Fares. Credit is also due to Ken Wolfe, an early co-author, who taught me quite a bit by example.

I’ve learned a lot working closely with various past and present members of the Fares lab—especially Brian Caffrey, Xiaowei Jiang, Orla Keane, Valentín Ruano-Rubio, and Christina Toft—and had an enjoyable time too. Karsten Hokamp, Chris Cotter and Xiaowei Jiang provided much-appreciated comments on earlier drafts of the manuscript.

Finally, I would like to thank my parents for providing encouragement and support; and also Mieke Zwart, who helped out with several of the figures in this thesis, among other things.

Tom Williams
Trinity College Dublin
2011
Contents

List of Tables xi

List of Figures xii

Chapter 1 Introduction 1
  1.1 Thesis structure and aims ..................................................... 1
  1.2 The context: evolutionary biology ........................................... 3
  1.3 Molecular chaperones ............................................................. 11
  1.4 Chaperones and evolvability .................................................... 17

Chapter 2 Two chaperonin systems in bacterial genomes with distinct ecological roles 21
  2.1 Related manuscript ............................................................... 21
  2.2 Introduction ........................................................................... 21
  2.3 Materials and Methods .......................................................... 23
  2.4 Results and Discussion .......................................................... 27

Chapter 3 Whole-proteome analysis of functional divergence with Clusterfunc 43
  3.1 Related manuscripts ............................................................... 43
  3.2 Introduction ........................................................................... 44
  3.3 Materials and Methods .......................................................... 47
List of Tables

2.1 Evidence for horizontal gene transfer (HGT) from archaea into bacterial that possess a group II chaperonin gene .................................................... 32
2.2 Sites under functional divergence in mtHsp60 ........................................ 39
3.1 Effect of organism lifestyle on functional divergence .......................... 59
3.2 Effect of bacterial lifestyle and genome size on functional divergence . 61
3.3 Sites under functional divergence in E. coli UMN026 VirB8 ............... 67
5.1 Main effects in the ANCOVAs evaluating influences on evolutionary rate 96
5.2 Effect of deleting ANCOVA terms on remaining probabilities .......... 99
5.3 Loss of GroEL clients and nonclients from Mycoplasma genomes ........ 103
5.4 Amino acid composition of GroEL client homologs in Mycoplasmas with and without GroEL ................................................................. 107
A.1 Chapter 2: Two chaperonin genes in bacterial genomes with distinct ecological roles ................................................................. 120
A.2 Chapter 3: Phylogenomic inference of functional divergence with Clusterfunc ................................................................. 120
A.3 Chapter 4: No Rosetta Stone for a sense-antisense origin of amino-acyl tRNA synthetase classes ......................................................... 120
A.4 Chapter 5: The effect of chaperonin buffering on protein evolution .... 121
## List of Figures

1.1 Pathways of chaperone-assisted protein folding .................. 12

2.1 Branch-specific detection of functional divergence .............. 26

2.2 GroEL lineages experiencing the greatest functional divergence . 28

2.3 Phylogenetic tree of group II chaperonin sequences from archaea and bacteria ....................................................... 29

2.4 Phylogenetic tree of mitochondrial Hsp60 sequences ............. 37

2.5 Functional divergence in mitochondrial Hsp60 ................... 38

3.1 Clusterfunc program workflow ........................................ 48

3.2 Number of COG assignments for each group of homologs ........ 51

3.3 Different kinds of genes experience different levels of functional divergence 57

3.4 Visualizing high-level patterns of functional divergence ........ 63

3.5 Amino acid residues under functional divergence in *E. coli* .... 65

4.1 Map of the *A. klebsiana* genomic region ......................... 76

4.2 Protein sequence alignment of *Neurospora crassa* NAD-GDH with the partial putative NAD-GDH inferred from *Aphanomyces euteiches* EST data, and the translation of the *A. klebsiana* AS-ORF .......... 78

4.3 Protein phylogeny of DnaK sequences with and without an antisense glutamate dehydrogenase ORF ................................. 81
5.1 The relationships between chaperonin buffering, gene essentiality, protein-protein interactions, residue contact density, and expression levels. . . . 101
5.2 Phylogeny of *Mycoplasma* genomes with and without GroEL . . . . . . 105
6.1 Thesis timeline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 117
Chapter 1

Introduction

1.1 Thesis structure and aims

This thesis presents an evolutionary analysis of molecular chaperones—particularly GroEL/Hsp60—and the proteins that they help fold, which are called clients. The molecular biology and evolution of chaperones is intensively studied, both because of their medical importance (Stebbins et al., 1997; Daugaard et al., 2007) and the intrinsic interest associated with any class of ancient, important molecules. We have learned much about chaperone biology since the original discovery of Ritossa (1962) that gene expression in *Drosophila* changes during heat shock. We now know that chaperones of one kind or another are found in every cellular lifeform (Fink, 1999) where they play a key role in protein folding—preventing the aggregation of newly-synthesized polypeptides, helping them to fold into their functional three-dimensional conformation, and assisting with the assembly of proteins and protein complexes (Hartl & Hayer-Hartl, 2009). During heat shock and other stresses, chaperone expression is upregulated in order to cope with the increased numbers of denatured and aggregation-prone proteins in the cell, explaining Ritossa’s observations and the other name for chaperones—“heat shock proteins” (Becker & Craig, 1994). More recently, it has become apparent that chaperones have evolved new functions in certain groups of
organisms, including roles in bacterial virulence (Morrison et al., 1989) and animal
development (Picard, 2002) and immunity (Asea & Maio, 2007).

If the basic outlines of this field are already well established, what can a new
thesis contribute—particularly one with a very general title? One answer is to explore
unusual or previously unconsidered aspects of chaperone evolution, as will be attempted
in Chapter 4. The main contribution of this thesis, however, is an examination of the
connection between protein folding and molecular evolution—as it is modulated by
chaperones—and an evaluation of the role of chaperones in the evolution of novelty
and organismal adaptation. In the last decade, work on experimental evolution has
suggested that chaperones can buffer the effects of genetic drift in endosymbiotic
bacteria, as well as facilitate the adaptive evolution of their client proteins (Fares
et al., 2002a; Tokuriki & Tawfik, 2009a). The aim of this thesis is to develop and test
these ideas using the end-products of natural evolution—DNA and protein sequences—
combined with new and existing statistical tools. The results have the potential to
contribute to ongoing debates about protein folding, chaperone buffering, and the
relative importance of the many factors that influence molecular evolution.

This thesis contains six chapters: four central chapters presenting original research,
bookended by a general introduction and discussion. Each of the four research chapters
is organized like a journal manuscript, and indeed all four are adaptations of papers
either published or in preparation. With the exception of Chapter 3 (where the
manuscript is still under preparation at the time of writing), these chapters should
probably be considered the “definitive” version of a report, since they have been
freed from journal length constraints. In particular, Chapter 2 has been substantially
expanded from the main text of the original article.

Chapter 1 sets out the structure and aims of the thesis, locates it in the field
of evolutionary biology, and provides an introduction to the biology of the molecular
chaperones that form much of the specific content of the following chapters. Chapter
1.2. The context: evolutionary biology

2 presents an analysis of evolution and functional divergence in bacterial chaperonin (Hsp60) genes, revealing the presence of archaeal-type chaperonins in certain bacteria, as well as implicating chaperonins in the adaptation of pathogenic bacteria to the intracellular environment. In Chapter 3, the new statistical method for detecting functional divergence introduced in the previous chapter is further developed and scaled-up for the analysis of large datasets, and applied to a set of 750 complete bacterial proteomes. In Chapter 4, we investigate the evolution of an unusual set of Hsp70 genes which have long antisense open reading frames on the opposite strand. The results contribute to an ongoing debate about the origin and early evolution of the protein translation system. The hypothesis that chaperones buffer the accumulation of structurally-destabilizing mutations in their client proteins is tested in Chapter 5, while Chapter 6 draws the findings of the thesis together and identifies several possible avenues of future research.

As a final “housekeeping” note, supplementary material (in the form of accession numbers and other potentially re-usable content) is provided as tab-delimited text files on the accompanying CD, and is also available online at

http://bioinf.gen.tcd.ie/~faresm/tw_thesis_suppl.zip

An index to this material is provided in Appendix A. Hopefully, this presentation of supplementary material will prove more useful for future analysis than providing printed copies of the data.

1.2 The context: evolutionary biology

1.2.1 Historical overview

We begin with an outline of the development of evolutionary biology, before considering some relevant modern topics in more detail. The seeds of modern
evolutionary thought were sown in the 17th and 18th centuries, as scientists began to realize that the Earth was much older, and more changeable, than had previously been appreciated. Nicholas Steno, among others, recognized that fossils were the remains of living things, and that rocks were built up over time in a series of layers (Steno, 1669). Hutton and, later, Lyell, argued that modern geological features were the result of gradual, continuous processes (such as erosion) that continue today (Lyell, 1830). An important consequence of this idea was that the Earth must be substantially older than the roughly 6,000 years suggested by the Bible. At around the same time, Cuvier showed that some fossils did not correspond to any living species, demonstrating that life on earth changes over time (Rudwick, 1998). Lamarck was among the first to propose a coherent theory of organic evolution as the process by which this change took place: in his scheme, simple lifeforms appeared by spontaneous generation, gradually climbing a ladder of complexity in response to an intrinsic natural law (Graur et al., 2009).

Darwin's *On the Origin of Species* (Darwin, 1859) was a remarkable advance on previous theories of evolution in several respects: it was dense with evidence, closely-argued, and provided a simple, testable mechanism for evolutionary change which did not rely on any ad-hoc "natural laws". In the *Origin*, Darwin argued for descent with modification: (i) that modern organisms are not specially created, but rather related to each other through descent from a common ancestor; and (ii) that changes in the characteristics of organisms (i.e., evolution) are driven by natural selection. This concept of natural selection, which was independently arrived at by Darwin's contemporary Wallace (Bulmer, 2005), is based on the idea of Malthus that populations over-produce offspring. Given limited resources, this over-abundance of individuals must compete to survive and reproduce. If individuals vary in terms of their competitiveness (fitness), and a portion of that variation is heritable, then each subsequent generation will be made up disproportionately of the descendants of
more competitive individuals. Therefore, the population as a whole will become more competitive, or better adapted to their environment.

In 1859, the mechanism of heredity was not well understood. The consensus was that offspring resulted from a blending of parental characters (Rose & Oakley, 2007), which ought to reduce variation in the population over time. This was perceived as a major flaw in Darwin’s argument for natural selection, which depends on the presence of heritable phenotypic variation. The solution to this problem was to be found in Mendel’s particulate inheritance (Mendel, 1866), but his pioneering work was not fully appreciated until a 1901 English translation by Bateson. The early decades of the 20th century saw the integration of Darwin’s work with the new sciences of genetics and cytology, culminating in the “modern evolutionary synthesis” in the 1940s (Rose & Oakley, 2007). The modern synthesis emphasized mutation as a random process that generated the genetic diversity upon which gradual natural selection acted. It was heavily influenced by the nascent field of population genetics, which viewed evolution as changes in the frequencies of alleles (gene variants) in a population. Major contributors to the modern synthesis such as Wright, Fisher and Dobzhansky recognized that gene frequencies could change significantly by chance in finite populations, through sampling error during the transmission of alleles from one generation to the next (Wright, 1931). This phenomenon, termed genetic drift, is now established as a major process in evolution. However, it was largely sidelined in favour of selection during the decades following the introduction of the synthesis (Rose & Oakley, 2007). Although some early work on experimental evolution suggested that drift might be a significant evolutionary force (Dobzhansky & Pavlovsky, 1957), the “selectionist” perspective was not seriously challenged until molecular sequence data began to accumulate (Zuckerkandl & Pauling, 1965a,b). This data contradicted key elements of the synthesis, including its reliance on selection. In particular, the discovery that the number of differences between two protein sequences was often proportional to the amount of time since their common
ancestor (Zuckerkandl & Pauling, 1965a)—that is, the existence of a “molecular clock”—was difficult to reconcile with the presumably inconsistent rate of adaptive evolution, although this was not immediately appreciated. Significant sequence homologies across large evolutionary distances also proved problematic for selectionism, with its requirement that the genes of each organism ought to be highly adapted to its ecological niche; indeed, the most functionally important regions within genes tended to be the most highly similar, even though these ought to be the regions which should evolve most rapidly under pure Darwinian selection (King & Jukes, 1969; Hughes, 2008). These were among the considerations that motivated the neutral theory of Kimura (Kimura, 1968, 1983).

1.2.2 The neutral and nearly-neutral theories

Under Kimura’s neutral theory, most mutations fixed during evolution have no effect on fitness (hence “neutral”), but simply drift to fixation. A remarkable consequence of this model is that the rate of evolution is equal to the mutation rate. To see why this is so, consider that the rate of evolution must equal the rate at which new mutations enter a population multiplied by the probability that any of one of the mutations will become fixed (that is, reach an allele frequency of 1):

\[ R_{\text{evolution}} = R_{\text{new mutations}} \cdot P_{\text{fixation}} \]

The rate at which new mutations arise is simply the mutation rate per allele \( \mu \) multiplied by the number of alleles in the population, which will be \( 2N \) in a diploid population of \( N \) individuals:

\[ R_{\text{new mutations}} = 2N \mu \]

What is the probability of fixation of a neutral allele? Since genetic drift is a kind of sampling error, it will ultimately lead to the loss of all but one of the alleles in the
population, which will then reach fixation. Each allele in the population has an equal chance of being the “surviving” allele, since they are all neutral and so unaffected by natural selection. This implies that the probability of fixation for a neutral allele is equal to its frequency. For a neutral allele $a$ that has just arisen by mutation,

$$f(a) = P_{fixation}(a) = \frac{1}{2N}$$

Therefore,

$$R_{\text{evolution}} = 2N\mu, \frac{1}{2N} = \mu$$

which, assuming the mutation rate is roughly constant, explains the clock-like evolutionary rate of many proteins. This derivation was provided to illustrate the simplicity and profundity of many of the basic population genetic results of the neutral theory. For an accessible but thoroughly quantitative treatment of these topics, see Gillespie (2004).

The neutral theory did an excellent job of explaining the “molecular clock” and a reasonable job of explaining the conservation of functionally important sites within proteins (most mutations at such “selectively constrained” sites are unlikely to be neutral and so only rarely drift to fixation). However, the theory quickly encountered a serious difficulty in comparisons of evolutionary rates between protein-coding and noncoding DNA, as well as rate comparisons between organisms with different generation lengths (Ohta, 1972b, 1973). The most natural unit of the mutation rate $\mu$ in the calculations above is number of generations (because most germline mutations occur during DNA replication associated with gamete production, and the number of these replications does not vary greatly with generation length (Ohta, 1972b)). If the rate of evolution depends only on the mutation rate, organisms with short generation lengths (e.g. mice) should evolve more quickly than organisms with long generations (e.g. elephants). This prediction of the neutral theory is met for
Chapter 1. Introduction

noncoding DNA but not for protein-coding DNA, where there is little or no effect of generation length (Ohta, 1972a). Therefore, it appears that the fixation of neutral mutations is not the predominant mode of protein sequence evolution. The solution to this problem was provided by Ohta (Ohta, 1972a, 1973), who realized that generation length tends to be inversely proportional to population size. The strength of both positive and negative selection (that is, selection for advantageous alleles and selection against deleterious ones) increases with population size, and stronger negative selection will reduce evolutionary rate (by eliminating deleterious alleles from the population). Therefore, the lack of generation time effect could be explained if most amino acid mutations are deleterious and experience negative selection. In large populations (with short generation times), large numbers of such mutations arise, but most are eliminated by strong negative selection. In small populations (with longer generation times), fewer mutations arise, but more reach fixation because of weakened negative selection, so that the effect of generation length tends to be cancelled out. This idea that most amino acid mutations are slightly deleterious is central to Ohta’s extension of the neutral theory, called the “nearly-neutral theory”. The original neutral theory adequately describes the evolution of noncoding DNA because most of it is nonfunctional.

With this theory in hand, we turn to its applications in the contemporary study of molecular evolution.

1.2.3 Selective constraint, gene duplication, functional divergence

In molecular evolution, most new genes and functions (that is, biochemical activities, cellular roles, and so on) evolve from existing genes, in a process called functional divergence (for a recent review, see Conant & Wolfe (2008)). A major impediment to the evolution of new functions is the requirement to maintain the pre-existing ones. In general, single-copy genes are performing a useful function, so that mutations that interfere with that function will be subject to negative selection—
in other words, it is difficult for a gene that is already under selective constraint to acquire a new function. Ohno (1970) realized that this selective constraint could be alleviated through gene duplication, after which one or the other gene copy would be free to evolve in a new direction, so long as the other copy continued to maintain the ancestral function (a process called neofunctionalization). Gene duplications can also be fixed through subfunctionalization, in which the two copies accumulate degenerative mutations in a complementary way, so that each duplicate performs a different subset of the ancestral gene's function (Force et al., 1999; Lynch & Conery, 2000). It is commonly suggested that these mechanisms could be complementary, since an initial period of subfunctionalization would provide more time for neofunctionalizing mutations to arise (Conant & Wolfe, 2008).

Another way in which selective constraint can be relaxed is through major changes to the environment of an organism. In endosymbiotic bacteria, for example, this relaxation of constraint is manifested in the loss of many of the genes free-living bacteria need to obtain nutrients from the environment (Moran, 2002; McCutcheon & Moran, 2007), but also in episodes of functional divergence (Toft et al., 2009). This is a topic that will be investigated in more detail in Chapter 3.

There is great interest in developing methods to identify events of neo- and sub-functionalization (essentially, functional divergence) from sequence data, and to distinguish these from the fixation of functionally-identical duplicates either neutrally or through dosage effects (fitness benefits due to the possession of two copies of the gene—for instance, higher expression levels (Kondrashov & Kondrashov, 2006)). Many of these approaches exploit the redundancy of the genetic code. Since multiple triplet codons specify the same amino acid during translation, not all DNA mutations lead to amino acid changes. Instead, mutations in protein-coding DNA sequences can be divided into those that alter the protein sequence (nonsynonymous changes) and those that do not (synonymous changes). Since most amino acid substitutions are deleterious
under the nearly-neutral theory, the ratio of nonsynonymous-to-synonymous changes (dN/dS) should typically be much less than 1 in a functional protein. In a nonfunctional sequence such as a pseudogene, dN/dS should be about 1, and in a gene experiencing strong positive selection, dN/dS should be greater than 1 (because the rate of fixation of positively-selected mutations is greater than that of neutral synonymous mutations). These straightforward considerations have led to the development of several methods for calculating dN/dS as a way of identifying positive selection (Miyata & Yasunaga, 1980; Nei & Gojobori, 1986; Hughes & Nei, 1988; Yang, 2006). However, dN/dS ratios must be interpreted with caution. One problem is that requiring dN/dS > 1 over the full length of the protein is extremely conservative, because positive selection is thought to act only on a subset of amino acids in most cases. This problem has been addressed with the development of more sophisticated statistical methods for inferring per-site or per-lineage dN/dS ratios (for review, see Yang (2006)). A potentially more serious problem with these ratios is that synonymous substitutions may not be neutral because different codons are translated at different speeds and with different error rates. Indeed, selection for protein sequences which can still fold and function correctly despite translation errors (translationally-robust proteins) may also constrain the nonsynonymous substitution rate in a way that is not directly related to protein function (Drummond & Wilke, 2008).

One way to reduce the impact of these problems is to compare site-specific evolutionary rates between two phylogenetic clades of homologous proteins, as was attempted by Bielawski & Yang (2004). In this framework, differing dN/dS ratios at homologous sites in two proteins indicate functional divergence. A method very similar in philosophy has been implemented by Gu & Velden (2002), which uses changes in the rate of amino acid substitution to detect functional divergence between paralogous proteins. As was discussed above, one consequence of the neutral theory is that highly-conserved sites are under strong functional constraint. Therefore, homologous positions
which are highly conserved within one clade but variable in another—or conserved within but not between clades—are suggestive of changes in functional constraint and therefore functional divergence. A fast, approximate method for detecting functional divergence heavily influenced by both these approaches is developed in Chapters 2 and 3 of this thesis.

1.3 Molecular chaperones

As discussed above, chaperones are ubiquitous and important molecules involved in the proper folding and regulation of newly-synthesized proteins, among a number of other roles. The overview of chaperone biology that follows focuses on chaperones that play an important role in this thesis—Hsp70, Hsp90 and, in particular, the Hsp60/GroEL family of chaperonins.

Figure 1.1 presents a “consensus” view of the involvement of chaperones in protein folding in modern cells (the protein structure images in the figure were created by David Goodsell for the PDB “Molecule of the Month” feature; see http://www.pdb.org/ for more details). Many newly-synthesized proteins can fold independently after being released from the ribosome, but others need the help of molecular chaperones. Trigger factor (not shown) and Hsp70 bind the exposed hydrophobic regions of nascent polypeptide chains—in some cases while they are still being translated—and prevent their aggregation (Hartl & Hayer-Hartl, 2009). Hsp70 consists of an N-terminal ATP-binding and C-terminal peptide-binding domain whose activities are allosterically coupled (Mayer et al., 2000): when ATP is bound in the N-terminal domain, the C-terminal domain has low substrate affinity, but ATP hydrolysis causes a conformational change in the C-terminal domain that greatly increases substrate affinity. Substrates are delivered to Hsp70 by Hsp40 (DnaJ), a co-chaperone which also stimulates ATP hydrolysis, resulting in a tightly-bound chaperone/substrate complex (Young et al., 2003). The release of substrate from Hsp70 is promoted by another set
Fig. 1.1: Pathways of chaperone-assisted protein folding. Many newly-synthesized proteins can fold independently, but some require the assistance of chaperones. The first line of defense against misfolding and aggregation are Trigger factor (not shown) and Hsp70, which associate closely with the ribosome and chaperone emerging polypeptides. For a subset of proteins, the activity of these chaperones is not enough to reach their final, functional conformation. These are delivered by Hsp70 to the GroEL/Hsp60 protein-folding cage, or to Hsp90 for fine-tuning and maturation.
of accessory proteins called nucleotide exchange factors that exchange ADP for ATP in the N-terminal domain, switching the C-terminal domain to the low-affinity state and prompting substrate release. This cycling of substrates on and off Hsp70 may promote proper folding in two ways. Firstly, Hsp70 binding prevents the aggregation of substrates by protecting the bound hydrophobic segments from interactions with other newly-synthesized polypeptides. On release, fast-folding substrates will quickly bury their hydrophobic regions, folding correctly. Slow-folding proteins will still have some exposed hydrophobic residues and will be bound again by Hsp70, which may deliver them to the GroEL/Hsp60 system for further folding (Hartl & Hayer-Hartl, 2009). Secondly, Hsp70 may directly catalyze the isomerization of certain peptide bonds in bound substrates, helping the polypeptide overcome difficult kinetic barriers (Schiene-Fischer et al., 2002).

A subset of proteins which cannot complete folding with the aid of Trigger factor and Hsp70 alone are delivered to the GroEL/Hsp60 folding complex (Frydman et al., 1994) or an Hsp90 dimer (Wandinger et al., 2008) to complete folding. We first outline the cellular role of Hsp90, which is of only tangential relevance to the research reported in this thesis, before providing a more detailed review of Hsp60 structure, function and evolution.

Unlike the other chaperones discussed here, Hsp90 has been most extensively studied in eukaryotes (see Pearl & Prodromou (2006) for a thorough review). Hsp90 is of great interest to biomedical researchers because its short list of client proteins includes a number of medically-relevant transcription factors and protein kinases involved in the regulation of development (Nathan et al., 1997; Nathan & Lindquist, 1995; Pearl & Prodromou, 2006). Some of these clients are proto-oncogenic (Hawle et al., 2006), and Hsp90 is upregulated in many cancers (Sangster et al., 2004). In fact, it appears that the oncogenic mutants of these genes are especially reliant on Hsp90 for stability and maturation, so that cancer cells have a heightened vulnerability to
Chapter 1. Introduction

Hsp90 inhibition (Chiosis et al., 2003). These findings have led to research into Hsp90 inhibitors as anti-cancer agents (Whitesell & Lindquist, 2005).

As with Hsp70, the folding activity of Hsp90 is coupled to a cycle of ATP binding and hydrolysis which is mediated by a set of accessory co-chaperone proteins (Hessling et al., 2009). The physiologically-relevant form of Hsp90 is a dimer, with each monomer consisting of three domains. In the absence of ATP, the monomers are connected only at the C-terminal domain. The binding of ATP to a pocket within the N-terminal domain (Prodromou et al., 1997; Stebbins et al., 1997) induces conformational changes which reveal a hydrophobic patch in both N-terminal domains (Ali et al., 2006). The interaction between these patches drives the dimerization of the N-terminal domain and a series of conformational changes in the middle domain, leading to ATP hydrolysis (Ali et al., 2006). The client binding interface is spread over the N-terminal and middle domains, so that changes in the relative positions of these domains brought about by ATP/ADP cycling induce conformational changes in the client.

Importantly, Hsp90 was the first chaperone that was investigated for its ability to buffer the accumulation of mutations in client proteins (Rutherford & Lindquist, 1998). This topic will be revisited in the final section of this introduction, and forms the basis of Chapter 5.

1.3.1 GroEL and the Hsp60 family of chaperonins

GroEL and GroES of Escherichia coli were the first chaperone genes to be characterized, being identified during screens for E. coli mutants which block bacteriophage growth (Georgopoulos et al., 1972; Georgopoulos, 2006). It was later discovered that these proteins, required for bacteriophage capsid assembly, were homologous to a chloroplast protein which assisted in the assembly of the enzyme ribulose bisphosphate carboxylase-oxygenase (Hemmingsen et al., 1988). This landmark discovery established the concept of "chaperonins" (one class of what we now call molecular chaperones)
1.3. Molecular chaperones

and the general requirement for assisted protein folding. The GroEL/GroES system remains the most well-studied chaperonin, and the following paragraphs outline our current understanding of its cellular role, structure, mechanism, and evolution.

GroEL is found in nearly all bacteria sequenced so far, with the exception of several *Mycoplasmas* (Lund, 2009). It is the only chaperone which is essential under all studied conditions in *E. coli* (Fayet et al., 1989; Genevaux et al., 2004), and is a "generalist" chaperone, with at least 252 experimentally-confirmed client proteins (Kerner et al., 2005) which it helps to fold. 13 of these clients are themselves essential proteins in *E. coli* (Kerner et al., 2005), which explains the stringent cellular requirement for GroEL.

The functional form of GroEL is a 14-mer composed of 60kDa monomers (see Figure 1.1, (Xu et al., 1997), hence the alternative name "Hsp60" for this family of proteins. The complex consists of two 7-mer rings which are joined back-to-back at the centre, each with a central chamber of 85,000 cubic Angstroms (Lund, 2009). The protein folding activity of GroEL occurs within these chambers in a tightly coupled way, with allosteric communication between the two halves of the complex being essential for activity (Weissman et al., 1996). The reaction proceeds as follows: hydrophobic residues on the surface of unfolded proteins bind to a hydrophobic region in the apical domain of one ring (the cis-ring) of GroEL, near the opening of the complex in Figure 1.1 (Fenton et al., 1994; Stan et al., 2006). Subsequently, ATP and the GroES cap bind to the same ring, bringing about conformational changes that dislodge the unfolded substrate into the cavity where it can fold (Weissman et al., 1995, 1996; Xu et al., 1997). The precise mechanism by which folding occurs within the GroES-covered cavity is a matter of some debate, but it may be that the cavity protects the unfolded protein from nonproductive interactions with other proteins (Lund, 2009). Folding continues until hydrolysis of the ATP bound to the cis-ring is complete, which takes about 10 seconds (Burston et al., 1995). ATP hydrolysis causes conformational changes that both weaken the association with GroES and substrate in the cis-ring and prime
Chaperonins are divided into two groups: the type I chaperonins of bacteria (GroEL and its homologs) and bacteria-derived organelles, and the distantly-related type II chaperonins of Archaea and the eukaryotic cytosol (Lund et al., 2003). There are two main differences between the two classes of chaperonins, as revealed by a series of structural (Ditzel et al., 1998) and evolutionary (Archibald et al., 2000, 2001; Ruano-Rubio & Fares, 2007) analyses. The first is that, despite a high level of overall structural homology to GroEL, type II chaperonins do not require an accessory GroES-like protein for activity. Remarkably, however, they possess a helical extension at the top of the apical domain which performs an analogous function (Ditzel et al., 1998). The second major difference lies in quaternary structure. While GroEL forms a 14-subunit homooligomer, type II complexes are formed from 16- or 18-subunit hetero-oligomers (Ditzel et al., 1998). Both Archaea and eukaryotes contain multiple subunit-encoding genes, although they show contrasting modes of evolution. Archaea contain between one and five different subunit genes, and duplications and losses of these genes have been common throughout archaeal evolution (Lund et al., 2003). Further, the heteromeric complexes of archaeal chaperonins appear to have arisen neutrally, with only limited subfunctionalization or specialization of individual subunits (Ruano-Rubio & Fares, 2007). The evolution of the eukaryotic type II chaperonin complex, termed Chaperonin Containing TCP1 (CCT), has been quite different, with initial duplications early in eukaryotic evolution resulting in a set of eight specialized subunits which are conserved across all known eukaryotic lineages (Archibald et al., 2001; Fares & Wolfe, 2003). The full range of client proteins which depend upon CCT complexes is not known, but the list includes actin and tubulin (Llorca et al., 2000), which are essential components of the eukaryotic cytoskeleton.
1.4 Chaperones and evolvability

We conclude this introduction with a review of the role chaperones have been proposed to play in facilitating adaptive evolution, a subject which forms the basis of Chapter 5. These proposals form part of the emerging field of "evolvability", which might be defined as the capacity of populations to respond to selective pressures—that is, to adapt. The goal of the following discussion is not to give a full account of this controversial new field, but simply to introduce the context necessary for understanding the research on chaperone buffering in Chapter 5. For a wide-ranging, up-to-date take on the field, the reader is directed to Masel & Trotter (2010) (but see also Lynch (2007) for an opposing view).

Research into evolvability aims to understand how the adaptive features of organisms evolve, particularly "difficult" adaptations that are challenging to explain in terms of traditional evolutionary theory. The classical problem of evolvability—originally noted by Wright (1932)—occurs when a combination of several mutations are needed for an increase in fitness, but each of the mutations individually lowers fitness. The mutations are often visualized as an "adaptive valley" running between two "adaptive peaks", representing the phenotypes of the current and potentially more fit genotypes. Since selection is thought of as a gradual, step-by-step process, are such "adaptive leaps" simply off-limits in evolution, or are there mechanisms by which adaptive valleys can be crossed? This question can be made more manageable by dividing it into two parts: firstly, are there features of organisms which promote evolvability (i.e., facilitate adaptive leaps); and secondly, were these features selected during evolution because they promote evolvability? The answer to the first half of the question is a resounding yes, but both theory and evidence are somewhat equivocal on the second half.

Modern lifeforms have a number of features that promote evolvability, many of which were described long before the uptick in evolvability research within the
Chapter 1. Introduction

last 10 years. Perhaps the least controversial of these are the mechanisms which promote the generation of genetic variation in bacteria under stress conditions. During starvation, some bacteria upregulate their mutation rate in a controlled way, perhaps in order to increase the rate of production of new adaptive variants (Galhardo et al., 2007). The competence (DNA uptake) systems of bacteria are also activated during nutrient limitation, presumably in an attempt to obtain useful DNA sequences from the environment (Poole et al., 2003). These sequences might be niche-relevant metabolic genes from distantly-related organisms which could help the competent cells adapt to their environment, an example of which will be seen in Chapter 2. Sexual reproduction also falls into this category of evolvability-promoting mechanisms, since it enables independently-arising mutations to be recombined into a single genome. Indeed, sexual reproduction is a response to stress in facultatively sexual eukaryotes such as *Saccharomyces cerevisiae* (Poole et al., 2003).

More recent research has focused on the relationship between robustness, or the ability of organisms to withstand the accumulation of deleterious mutations, and evolvability (Wagner, 2008). Briefly, organisms are considerably more robust (i.e., more able to keep functioning even after picking up random mutations) than human-engineered systems (such as computer programs, cars, etc.). The best explanation for this observation is that organisms have evolved mechanisms to protect against the phenotypic effect of mutations which, as discussed earlier in this chapter, are mostly deleterious. The first line of defense against new mutations is the DNA proofreading and repair apparatus (Masel & Trotter, 2010), but mutations which slip through may persist in populations for long periods of time, or even reach fixation, when their deleterious effects are slight in finite populations (Ohta, 1973). In principle, the robustness of biological systems ought to act against evolvability, because robustness reduces the heritable phenotypic variation in a population on which selection can act. This theoretical difficulty was partially overcome by postulating
1.4. Chaperones and evolvability

phenotypic "capacitors", molecules that buffer (suppress the phenotypic effects of) and then simultaneously release large quantities of genetic variation. This sudden release of variation can be acted on by selection, perhaps leading to the kind of "adaptive leap" discussed above (Wagner, 2008).

One of the first capacitors to be proposed was Hsp90. Rutherford & Lindquist (1998) noted that impairment of Hsp90 activity in Drosophila resulted in the expression of a range of developmental abnormalities that are apparently suppressed by the protein-folding activity of the chaperone in wild-type flies. Later, Queitsch et al. (2002) demonstrated that Hsp90-buffered morphological variation was also present in populations of Arabidopsis thaliana, even though levels of genetic variation are much lower in the inbreeding A. thaliana in comparison to outbreeding fruit flies.

An appealing feature of capacitance by Hsp90 was that buffered variation ought to be released during stressful conditions, when the chaperone might be titrated by misfolded proteins in the cell. This would make buffering and release of variation a "last gasp" attempt at adaptation, in line with the more well-characterized mechanisms in bacteria mentioned above. At about the same time, evidence that GroEL/Hsp60 could also buffer phenotypic variation began to accumulate. GroEL is over-expressed in bacterial endosymbionts, which experience high levels of genetic drift, perhaps as a mechanism for coping with the accumulation of mutations and the concomitant decrease in proteome stability (Moran, 1996). In support of this idea, Fares et al. (2002a) demonstrated that over-expression of GroEL could partially recover the fitness of E. coli cells experimentally exposed to population bottlenecks and therefore increased genetic drift. More recently, Tokuriki & Tawfik (2009a) presented evidence that GroEL can ameliorate the destabilizing effects of radical mutations in its client proteins, with possible implications for the evolution of new enzymatic functions.

A thorny issue is whether evolvability can evolve—whether features of cells, such as phenotypic capacitors, are selected for their ability to promote evolvability, or
Chapter 1. Introduction

whether this is simply a happy side-effect of the molecule's primary function (see Draghi & Wagner (2008) and Lynch (2007) for the two sides of this story). Since chaperones such as Hsp90 and GroEL are known to have critical functions in the cell that have nothing to do with their potential capacity to facilitate adaptive evolution, the simplest explanation is that these molecules were selected for their protein-folding capabilities, with any phenotypic capacitance being a nice example of the role of chance and contingency in evolution. Fortunately, a definitive answer to this question is not required in order to proceed with Chapter 5, in which we test whether GroEL does, in fact, buffer the accumulation of mutations in its client proteins.
Chapter 2

Two chaperonin systems in bacterial genomes with distinct ecological roles

2.1 Related manuscript


2.2 Introduction

As was described in Chapter 1, GroEL is a highly conserved molecular chaperone, and the only one essential under all conditions in *Escherichia coli* (Fayet et al., 1989). This generalist chaperone assists in the folding of at least 252 client proteins (Kerner et al., 2005). All bacterial genomes sequenced so far, with the exception of several *Mycoplasmas*, contain at least one *groEL* homologue (Lund, 2009). These molecules form one division of a protein family called the chaperonins, defined by sequence similarity (Gupta, 1995). The chaperonins comprise the group I sequences (i.e. *groEL* and its homologues) from bacteria and bacteria-derived organelles and the distantly-
related group II chaperonins of archaea and the eukaryotic cytosol (Lund et al., 2003). Although clearly homologous, group I and II chaperonin complexes differ in quaternary structure and client specificity: the group I complex contains an accessory protein GroES, and group I and II clients differ in structure and evolutionary rate (Hirtreiter et al., 2009).

Previous phylogenetic analysis of bacterial \textit{groEL} genes revealed that 30% of sequenced bacterial genomes contain more than one chaperonin gene (Lund, 2009). Several hypotheses have been proposed to explain the fixation of these duplicates. For example, it was suggested that organisms with multiple copies benefit from increased \textit{groEL} expression levels (dosage effects) (Kondrashov & Kondrashov, 2006), and that the increased regulatory control possible with multiple copies is beneficial to the organism. Alternatively, the functional divergence and neo- or sub-functionalization of the duplicate copies might promote retention of both copies in the genome (Lund, 2009).

To address these questions, we decided to perform an analysis of functional divergence in bacterial GroEL, reasoning that the patterns observed—the lineages that were under functional divergence—might shed some light on the driving force, if any, for the retention of these duplicates. Although a number of well-established methods for detecting functional divergence in sequence data currently exist (Gu & Velden, 2002; Kalinina et al., 2004; Pirovano et al., 2006), these are slow, and cannot be applied to a large, complex topology (such as the tree of bacterial GroEL sequences) directly. Rather, individual points on the tree must be chosen for testing, which limits the extent to which large-scale comparisons can be performed (Goyal et al., 2006; McNally & Fares, 2007). In order to assess patterns of functional divergence across the entire GroEL tree, we developed a simple, fast method, which uses BLOSUM scores (Henikoff & Henikoff, 1992) to quantify functional divergence. The method can be applied either to a large, complex phylogeny or a single node on the tree, and we demonstrate both approaches
2.3. Materials and Methods

2.3.1 Sequences and tree-building

To obtain all bacterial homologs of GroEL, we downloaded the complete set of fully sequenced bacterial proteomes from NCBI on 8 July 2008 (622 proteomes) and used them to create a BLAST database. This database was queried using *Escherichia coli* GroEL and all hits with an E-value lower than $10^{-7}$ were retrieved. The list of hits was manually curated to remove spurious sequences by removing proteins of length greater than 1.5 times that of the query sequence, and/or on the basis of sequence annotation at NCBI, resulting in a final set of 840 GroEL homologues. These proteins were aligned with MUSCLE 3.6 (Edgar, 2004) using the default parameters, and a 100-bootstrap maximum likelihood phylogenetic tree was created using RAxML 7 (Stamatakis, 2006) under the WAG model of protein sequence evolution (Whelan & Goldman, 2001), which was suggested by ProtTest (Abascal et al., 2005). The topology of the resulting tree was then used to infer duplication events in GroEL evolution.

Sequences were obtained for the mitochondrial Hsp60 analysis by BLASTing the *Homo sapiens* mitochondrial Hsp60 sequence against the NCBI non-redundant database. The resulting hits were retrieved and both automatically and manually curated. Due to the large number of hits to partial or otherwise suspect sequences, all sequences shorter than 450 residues were discarded (this is 120 residues shorter than the...
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

H. sapiens sequence). The remaining sequences were aligned with the GroEL sequences of the *Rickettsiales*, a group of bacteria that may be closely related to the ancestor of the mitochondrion (Gupta, 1995), and with a putative outgroup, *Bradyrhizobium japonicum*. A 100-bootstrap, maximum-likelihood phylogenetic tree was built from this alignment as described above (see also Figure 2.4). The topology of this tree suggested that the sequences did indeed fall into three clades (eukaryotic mitochondrial sequences, *Rickettsiales, Bradyrhizobium japonicum*). Given the evidence supporting the monophyly of the mitochondrion and the *Rickettsiales* (Fitzpatrick et al., 2006), we proceeded with our use of *Bradyrhizobium japonicum* as the outgroup. These sequences were then used in our analysis of functional divergence as described below.

Archaeal homologs of the bacterial group II chaperonin genes were obtained by BLASTing the *Desulfotomaculum reducens* group II sequence against all archaeal sequences in the NCBI non-redundant database, using the default parameters. Non-chaperonin sequences were manually removed from the list of hits, and the remaining sequences were used to construct a phylogeny as described above for the GroEL genes, using the LG model of protein evolution (Le & Gascuel, 2008).

The names and accession numbers of all sequences retrieved and used in our analyses are provided as Supplementary Material Ch2-S1-S3.

2.3.2 Detection of functional divergence

To investigate lineage-specific functional divergence in our GroEL tree, we developed a new method, which is outlined below. The method has since been improved and scaled-up for analysis of larger datasets, as will be discussed in Chapter 3.

Starting with the tips of a phylogenetic tree, the software walks down to the root, identifying all interior nodes at which a branch-specific analysis of functional divergence can be performed (the analysis is performed if there are at least four sequences in each descendant clade). A test for functional divergence at each of these
nodes is then carried out. A node is defined by two clades (clusters) and one or more outgroup sequences, and our method uses BLOSUM scores (Henikoff & Henikoff, 1992) to identify conserved unexpected substitutions—that is, substitutions associated with large, negative BLOSUM scores—in one protein cluster which are not observed in the homologous cluster, or in an outgroup cluster. These sites are then taken to be under functional divergence on the branch leading to that cluster. The procedure is illustrated in Figure 2.1, which presents two examples to clarify the method.

For each column of a protein multiple sequence alignment for which the three clusters are already defined, we calculate four quantities: the mean pairwise BLOSUM score between sequences in the first cluster and the outgroup cluster ($C_1$); the mean pairwise BLOSUM score between sequences in the second cluster and the outgroup cluster ($C_2$); and the combined standard error ($SE_{C_1,C_2}$). We then calculate a test score for that column as

$$FD = \frac{(C_1 - C_2)}{SE_{C_1,C_2}}$$

and compare to a normal distribution to assess significance.

Here, we present both a branch-specific, whole phylogeny test (the analysis of all bacterial GroEL genes), and a test of a single lineage (mitochondrial Hsp60). In the case of analyzing the entire tree, we draw a distribution of numbers of sites under functional divergence on each branch, and examine the top 5% of the distribution in order to isolate the most important patterns of functional divergence in the dataset. With our analysis of mitochondrial Hsp60, we examine the specific sites identified as being under functional divergence.
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

Fig. 2.1: **Branch-specific detection of functional divergence.** The method walks over the phylogenetic tree, testing each node for functional divergence. In these examples, a section of the multiple sequence alignment being scanned by the method shows a column in which functional divergence is detected. The amino acids are colour-coded following the same principle as ClustalX, so that the colours correspond to amino acid properties. In the first example (blue node), we identify a transition from an ancestral F/Y to a P, which is conserved thereafter, indicating functional divergence (FD). In the next example (green node), a charged amino acid (D/E) mutates to a M/L, also indicating functional divergence.
2.4 Results and Discussion

2.4.1 Functional archaeal-like chaperonin genes in bacterial genomes

We screened 622 sequenced bacterial proteomes for homologs of *Escherichia coli* GroEL, resulting in a set of 840 proteins. A maximum likelihood phylogenetic tree of this set of genes identifies a clade of bacteria that possess archaeal-like group II chaperonin genes ('Group II' in Figure 2.2). These genes are mainly encoded by the *Clostridia* (5 species), with one from the cyanobacterium *Gloeobacter violaceus*. A search of the NCBI nr database identified four additional bacterial group II sequences from species not included in our initial dataset. Although some of the group II sequences had previously been identified (Karlin et al., 2005), they have not been studied in detail. Where did they come from, and what function are they performing?

To answer these questions, we performed a phylogenetic analysis of group II chaperonin sequences from bacteria and archaea, the results of which suggest that bacterial group II sequences, with the exception of those from *Oligotropha carboxidovorans*, form a monophyletic group (see Figure 2.3). The existence of two distinct clades of bacterial group II sequences implies the occurrence of at least two horizontal gene transfer (HGT) events from archaea to bacteria. Whereas the group II sequences in all non-*Oligotropha* bacteria are monophyletic, the genomes encoding them are phylogenetically-scattered, encompassing a broad phylogenetic range of bacteria that lack a group II chaperonin gene. Although a single transfer followed by large-scale loss in descendant lineages could explain this, a more parsimonious interpretation is a single transfer from archaea to bacteria, followed by a series of within-bacteria transfers. Our analysis shows a nine-species monophyletic clade that is most closely related to two group II chaperonin sequences from *Methanosarcina acetivorans*, suggesting that this methanogen, or a relative, was the archaeal donor. Given the existence of both group I and II chaperonins in this species and other members of the *Methanosarcina* genus
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

Fig. 2.2: GroEL lineages experiencing the greatest functional divergence. These lineages represent the top 5% of branches in the distribution generated by our branch-specific analysis of functional divergence (coloured lineages) mapped onto a maximum-likelihood protein phylogeny of bacterial GroEL. Excluding the group II chaperonins, four of the remaining five branches comprise intracellular pathogens. The functional divergence of GroEL in these organisms might be to mitigate the deleterious effects of their intracellular lifestyle, as is the case for the unusual GroEL proteins of endosymbiotic bacteria. Alternatively, it might be driven by adaptations that enable GroEL to participate more effectively in pathogenesis, as is known to be the case in the Chlamydiae and Mycobacteria.
2.4. Results and Discussion

Fig. 2.3: Phylogenetic tree of group II chaperonin sequences from archaea and bacteria. Bacterial lineages are highlighted in red; major archaeal lineages are coloured to clearly delineate groups. Nine of the ten bacterial sequences form a strongly-supported monophyletic cluster which has two group II sequences from the methanogen *Methanosarcina acetivorans* as an outgroup. The immediate outgroup of the group II sequence from *Oligotropha carboxidivorans* is encoded by the archaeon *Nanoarchaeum equitans.*
(Klunker, 2003), this conclusion suggests two-way transfer of chaperonins between these organisms and bacteria. But what function, if any, are the group II chaperonins performing in bacteria?

These sequences might be pseudogenes: in bacteria, most horizontally transferred genes become nonfunctional unless they are advantageous to the recipient (Taoka et al., 2004). To test this possibility, we performed a comparison of the selective constraints operating on archaeal and bacterial group II sequences. We used the method of Yang and Nielsen (Yang & Nielsen, 2000) implemented in the yn00 program, which is part of the PAML package version 4 (Yang, 2007), to compare the ratios of nonsynonymous and synonymous substitution rates (dN/dS) between these two groups of sequences. The difference between the nonsynonymous and synonymous rate is mainly due to selection, because nonsynonymous substitutions change the encoded protein sequence whereas synonymous substitutions do not. Most protein-changing mutations are deleterious (Kimura, 1983), so that the nonsynonymous rate is usually much lower than the synonymous rate in proteins under selective constraint—that is, functional proteins. If the group II chaperonin sequences in bacteria have become nonfunctional, then the mean pairwise dN/dS ratio among bacterial sequences should be substantially higher than that among archaeal sequences (and probably close to 1). In fact, these ratios were very similar (0.204 for bacteria; for archaea: 0.194), suggesting that the group II chaperonins are evolving under comparable selective constraints in both domains. Under the assumption that the archaeal group II chaperonin sequences are functional, this result suggests the bacterial sequences are as well.

What function might these genes be performing? Because all ten group II chaperonin-possessing bacterial genomes also contain endogenous group I chaperonins, this role must be in addition to the broad-spectrum protein-folding activity of GroEL. Recent work on the two chaperonin systems of Methanosarcina mazei (Hirtreiter et al., 2009) has revealed that group I and II proteins have overlapping but distinct
specificities. In particular, several essential metabolic proteins require the group II system, raising the possibility that a requirement to fold niche-relevant proteins could promote retention of the group II chaperonins in bacteria inhabiting archaeal niches. Existing evidence from a variety of sources supports a substantial level of horizontal gene transfer between the 10 bacteria possessing a group II chaperonin gene and the methanogenic archaea (summarized in Table 2.1). The DarkHorse database, which attempts to catalogue horizontally transferred genes (Podell et al., 2008), suggests that horizontal transfer is ongoing between these species and archaea, with a mean of 46 recently-transferred (DarkHorse LPI score < 0.25) archaeal genes per bacterial genome which contains a group II chaperonin gene. There is certainly the ecological opportunity for such transfers to take place: seven of the bacterial species with a group II gene are involved in carbon monoxide metabolism or sulfur reduction, which occur in association with methanogenic archaea, and three are confirmed archaeal syntrophs (see Table 2.1). Cross-domain transfers in this environment would also be consistent with the methanogenic origin of all but one of the bacterial group II chaperonins (see Figure 2.3). Presumably, putative transferred group II clients would be proteins present in bacteria with a group II chaperonin gene and archaea, but not other bacteria. However, we were unable to detect any proteins in these ten bacterial genomes fulfilling this strict criterion. Future experimental analysis of these unusual chaperonins might help clarify their function, potentially providing important insights into the role of these proteins in functional innovation.

2.4.2 Intracellular pathogens have accumulated the most radical changes in GroEL

Is the evolution of endogenous bacterial chaperonins correlated with the ecological properties of bacteria? To address this question, we analysed lineage-specific functional divergence in GroEL using our novel, distance-based method (see Materials and
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

Table 2.1: Evidence for horizontal gene transfer (HGT) from archaean into bacterial that possess a
between this species and archaea.

<table>
<thead>
<tr>
<th>Species</th>
<th>Lifestyle</th>
<th>Archaeal Association</th>
<th>Other HGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO metabolism (Wu et al., 2004)</td>
<td>Thermophile, hydrophobicogenic</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sulphur-reducing (Chu et al., 2008)</td>
<td>Thermophile, hydrophobicogenic</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Photosynthetic (Klassen et al., 2005)</td>
<td>Thermophile</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Photosynthetic (Satley et al., 2008)</td>
<td>Thermophile</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sulphur-reducing (Imami et al., 2009)</td>
<td>Syntrophic (Klassen et al., 2008)</td>
<td>Thermophile</td>
<td>-</td>
</tr>
<tr>
<td>Sulphur-reducing (Imami et al., 2009)</td>
<td>Syntrophic (Klassen et al., 2008)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sulphur-reducing (Imami et al., 2009)</td>
<td>Syntrophic (Klassen et al., 2008)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CO metabolism (Sokolova, 2008)</td>
<td>Thermophile</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CO metabolism (Paul et al., 2004)</td>
<td>Thermophile</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Geobacillus sp. Y42MIC0</td>
<td>-</td>
<td>% HGT</td>
<td>-</td>
</tr>
</tbody>
</table>

HGT: The presence of genes estimated to be obtained horizontally in that species (not necessarily from archaean).
2.4. Results and Discussion

Methods this chapter, and also Chapter 3). In total, our test could be performed on 282 branches on the tree. When we examined the 5% of branches with the strongest evidence of functional divergence, we found that 4/5 of these branches belonged to intracellular pathogenic bacteria: the Chlamydiae, Anaplasmataceae, Mycobacteria and Mycoplasmas (see Figure 2.2).

The Chlamydiae are a group of intracellular pathogens that possess three highly divergent groEL genes, at least one of which is involved in pathogenesis (Morrison et al., 1989). There is also evidence that differing rates of evolution among the two groEL genes of the Mycobacteria are due to differential interaction with the host immune system (Hughes, 1993). The high-level pattern revealed by our functional divergence analysis is consistent with these results, but extends them to suggest that radical change in GroEL is a general symptom of the intracellular pathogenic lifestyle.

Why do intracellular pathogens have the most-divergent groEL genes among the bacteria surveyed here? Two distinct, though not mutually exclusive, processes might have driven this functional divergence: adaptation to the higher mutational load of the intracellular lifestyle, as with endosymbionts (Moran et al., 2008); and/or the fixation of mutations underlying GroEL-mediated pathogenicity (Morrison et al., 1989; Henderson et al., 2006). Several species of Mycoplasmas that entirely lack groEL genes, however, form an apparent exception to this trend: it has been proposed that these organisms rely increasingly on proteases, rather than chaperones, to cope with misfolded protein (Wong & Houry, 2004). This matter will be explored further in Chapter 5.

These results highlight the interesting parallels that exist between endosymbiotic and pathogenic bacteria: one of the lineages enriched for functional divergence is the Anaplasmataceae, which contains the Wolbachia genus. Though many Wolbachia species are parasites of arthropods (Wu et al., 2004), some nematodes have grown to depend upon them (Foster et al., 2005). This blurred line between intracellular pathogens and symbionts is most famously exemplified by the Rickettsiales, a
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

group that contains both virulent human pathogens and the mitochondrion. These observations motivate the hypothesis that adaptation of GroEL to the host immune system might be an early step in the evolution of endosymbiosis.

2.4.3 Do GroEL and GroES co-duplicate?

Is the duplication and functional divergence of GroEL accompanied by similar changes in its essential accessory protein, GroES? The interaction between GroEL and GroES is required for the functioning of the chaperonin complex (Weissman et al., 1995), implying that the binding surfaces on the two molecules ought to co-evolve. If this interface changes in one or the other of a pair of duplicated GroEL genes, two divergent GroES genes—one to interact with each copy of GroEL—might be required to maintain the function of the complex. To explore this possibility, we examined the phylogenetic distribution of GroES, and the co-occurrence of GroEL and GroES duplicates, in sequenced prokaryote genomes.

We used BLASTP searches against each bacterial and archaeal genome containing one or more \textit{groEL} genes to check for the co-occurrence of \textit{groES}. We first performed our search using \textit{E. coli} GroES, re-running manual BLASTs using a GroES query sequence from a more closely-related species if the initial search did not identify any significant hits (defined as an E-value lower than $10^{-7}$, coupled with NCBI database annotation). The accession numbers for the hits obtained are provided in Supplementary Table Ch2-S4. All but four \textit{groEL}-encoding genomes were shown to contain at least one \textit{groES} gene with this approach. We reasoned that the remaining four species—the bacteria \textit{Escherichia coli APEC O1, Clostridium tetani E88, Magnetospirillum magneticum AMB-1, and Lysinibacillus sphaericus C3-41}—might contain an unannotated \textit{groES} gene, because all four possess typical \textit{groEL} genes as well as having relatives possessing both \textit{groEL} and \textit{groES}. Indeed, TBLASTN searches using the \textit{E. coli} GroES sequence revealed significant hits in all four genomes. In the case of \textit{Lysinibacillus}, the hit to
the *E. coli* sequence is at the limit of our criterion (E-value = 10^{-7}), but the genomic location of the hit—directly adjacent to *groEL*—suggests that this is a genuine *groES* gene. Among the bacteria and archaea we surveyed, then, genomes containing one or more *groEL* genes also contained at least one *groES*.

Substantially more species contain duplicate *groEL* (163) than *groES* (69) genes. Although some groups of bacteria possess elevated copy numbers of both genes (the *Rhizobiales* and the *Burkholderiales*), there is no clear relationship between the number of *groEL* and *groES* genes, and many species with multiple, functionally divergent GroELs have only a single GroES (e.g. the *Chlamydiaceae*, *Mycobacteria*, and *Wolbachia* species; see Supplementary Table Gh2-S4). These results indicate that a single GroES can perform its co-chaperonin function for multiple, functionally divergent GroEL molecules.

### 2.4.4 Radical substitutions in mitochondrial Hsp60 underlying profound structural divergence

To conclude our survey of functional divergence in bacterial *groEL* genes, we investigated the specific amino acid sites under functional divergence in mitochondrial Hsp60. A highly-derived relative of the *Anaplasmataceae* GroEL proteins discussed above (Gupta, 1995), mtHsp60 has experienced substantial functional change since its divergence from the GroEL homologs of free-living bacteria. Key among these is the differing quaternary structure of the chaperonin’s functional state (Nielsen & Cowan, 1998). Bacterial GroEL functions as two rings of seven subunits each, connected back-to-back at the base (Weissman et al., 1995). Allosteric communication between the two rings is essential for the folding activity of the chaperonin, as the binding of ATP to one ring results in the dissociation of the GroEL cap, and the folded substrate, from the opposite ring (Sigler et al., 1998). mtHsp60, in contrast, is functional as a single ring, although it may adopt a double-ring structure under certain conditions.
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

(Levy-Rimler et al., 2001), and does not require the binding of ATP for GroES or substrate dissociation (Nielsen & Cowan, 1998). We sought to identify the molecular basis of these functional shifts in mtHsp60. The extensive body of experimental work on GroEL, especially from *E. coli*, was used to gauge the function of the residues which had undergone radical change on the lineage leading to mtHsp60 (Sun et al., 2003; Zheng et al., 2007; Sot et al., 2003).

The strategy we employed to identify sites under functional divergence on the mtHsp60 lineage was to compare these sequences, as a clade, to their closest homologs from free-living bacteria (the *Rickettsiales*) and a suitable outgroup (*Bradyrhizobium japonicum*). To confirm that these sequences fall into three distinct clades, as predicted by previous work on the origins of the mitochondrion (Fitzpatrick et al., 2006), we built a maximum-likelihood phylogenetic tree (see Figure 2.4). The topology of the tree confirmed the expected relationship between the clades, allowing us to proceed with our analysis.

We identified 19 sites under functional divergence on the branch leading to mtHsp60 (Table 2.2), which cluster within the equatorial domain (Figure 2.5). Two of these sites could not be assigned a homologous position in *E. coli* GroEL due to indels, but of the remaining 17 sites, 15 are either at, or within 4 Angstroms of, residues which, when previously mutated in *E. coli*, result in phenotypes resembling that of wild-type mtHsp60. Three of the sites under functional divergence are close to positions that maintain the inter-ring interaction in GroEL (marked in red in Figure 2.5). V438 is 2.9A away from E434, L444 is directly beside (1.3A) R445, and Q453 is again directly beside (1.3A) R452, all three of which form the interface with the other ring of GroEL (Sot et al., 2003). We also identified two other residues in this region, V14 and K28, of currently unknown function. Radical change at the inter-ring interface makes sense, since bacterial GroEL operates as a double-ring complex while mtHsp60 does not.

Sun et al. (2003) isolated mutants of a single-ring GroEL variant (SR1) which
Fig. 2.4: Phylogenetic tree of mitochondrial Hsp60 sequences. Branches are coloured according to the major clades that were compared in our analysis. In line with previous reports, the mitochondrial and Rickettsial groups receive strong bootstrap support.
Chapter 2. Two chaperonin systems in bacterial genomes with distinct ecological roles

Figure 2.5: Functional divergence in mitochondrial Hsp60. (a) A section through the E. coli GroEL/GroES complex (Xu et al., 1997) showing the
across the interface to function, whereas multiple operators as a single site

Figure 2.5: Functional divergence in mitochondrial Hsp60. (a) A section through the E. coli GroEL/GroES complex (Xu et al., 1997) showing the
across the interface to function, whereas multiple operators as a single site
### Table 2.2: Sites under functional divergence in mtHsp60.

- "Consensus" denotes the most frequent amino acid in one of the clades. All sites significant at $P < 0.01$.

<table>
<thead>
<tr>
<th>Escherichia coli site</th>
<th>mtHsp60 (consensus)</th>
<th>Rickettsiales (consensus)</th>
<th>Bradyrhizobium</th>
</tr>
</thead>
<tbody>
<tr>
<td>V14</td>
<td>A</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>K28</td>
<td>A</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>D121</td>
<td>M</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>T210</td>
<td>G</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>G211</td>
<td>Q</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>A239</td>
<td>L</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A241</td>
<td>N</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>A243</td>
<td>Q</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>R284</td>
<td>N</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>K315</td>
<td>H</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>K390</td>
<td>N</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>E397</td>
<td>T</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>V438</td>
<td>I</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>L444</td>
<td>K</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>Q453</td>
<td>T</td>
<td>Q</td>
<td>Q</td>
</tr>
<tr>
<td>Q505</td>
<td>V</td>
<td>Q</td>
<td>Q</td>
</tr>
<tr>
<td>1515</td>
<td>T</td>
<td>I</td>
<td>V</td>
</tr>
<tr>
<td>-(179 aln)</td>
<td>-</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>-(1016 aln)</td>
<td>T</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>
restored chaperone function but did not restore the double-ring quaternary structure. That is, they isolated mutations resulting in active, single-ring variants of GroEL. Four of the sites detected as being under functional divergence were close to several of these mutants: K390 (under functional divergence) was directly beside (1.3A) M389, E397 was close to both R395 and A399 (3.1A in each case), Q505 is close to A503 (3.3A) and A507 (3.1A), and I515 is directly between M514 (1.3A) and T516 (1.2A). One or more mutations at each of these positions restore single-ring GroEL function (Sun et al., 2003). Again, the proximity of these sites under functional divergence to sites which, when mutated, result in an active single-ring GroEL mutant makes biological sense in light of the functional differences between mtHsp60 and GroEL.

Six additional functionally divergent sites are in positions which suggest functional importance. T210 forms the hinge of a substrate-binding loop in the GroEL apical domain, and also makes contact with other subunits in the same ring (Zheng et al., 2007). Both this site and the adjacent G211 are under functional divergence in mtHsp60. A239 is beside (1.3A) E238, which is involved in GroES binding (Motojima, 2000). A241 is involved in substrate binding (Tanaka & Fersht, 1999). R284 forms a salt bridge, which plays an important role in the conformational changes which occur in one half of the GroEL complex upon ATP binding (Hyeon et al., 2006). E315 is beside (1.3A) L314, which is involved in substrate binding (Zheng et al., 2007).

The mutagenesis experiments referenced here were performed on *E. coli* GroEL, and so cannot provide direct information about the function of the homologous positions in mtHsp60. However, the location of these functionally divergent sites in important regions of the GroEL protein suggest that our method has captured, at least partially, the amino acid-level changes responsible for the observed differences in function between mtHsp60 and GroEL. In particular, residues at the inter-ring interface in *E. coli* have undergone radical substitutions in mtHsp60, providing an explanation for the observed divergence in quaternary structure at the molecular level.
2.4.5 Conclusions

Pioneering work by Rutherford & Lindquist (1998) has previously established a mechanistic link between the modulation of protein folding, a fundamental cellular process, and morphological adaptation. A major leap would be to prove an analogous link between that same buffering process and ecological adaptation. Our analysis has revealed two mechanisms of ecological adaptation that depend on chaperones: the horizontal transfer of chaperonins between ecologically close—but phylogenetically distant—organisms might be an essential first step that enables the transfer of other ecologically-relevant functions; and radical change in endogenous chaperonin genes might better adapt free-living bacteria to the fundamentally different intracellular lifestyles of pathogens and symbionts. These results establish the framework for a broader exploration of the role chaperonins might play in ecological innovation.
Chapter 3

Whole-proteome analysis of functional divergence with Clusterfunc

3.1 Related manuscripts


* denotes joint first authors.


This chapter describes the development of the method for detecting functional divergence introduced in the previous chapter. This project was a group effort, with each author on the two manuscripts listed above making an important contribution. The initial method, as described in Chapter 2, was designed by TAW, CT and MAF. TAW wrote the software implementation used in Chapter 2 and in the PLoS Comput Biol paper, in which CT applied the method to the analysis of endosymbiotic bacteria.
Chapter 3. Whole-proteome analysis of functional divergence with *Clusterfunc*

The scaling-up of the method, and the substantial revisions detailed in this chapter, were developed by BEC and TAW. BEC wrote the optimized C++ implementation which is now being released as “Clusterfunc”. The analysis of bacterial proteomes was carried out by TAW, BEC and XJ. MAF and KH provided advice and direction along the way.

### 3.2 Introduction

Most new genes, functions, and activities originate through the modification of existing ones. The evolutionary process that gives rise to functional differences between related genes is called functional divergence (Lynch & Conery, 2000; Conant & Wolfe, 2008). At the species level, functional diversification is primarily associated with adaptive radiations, when a single ancestor differentiates into multiple descendant species, each adapting by natural selection to one of a new set of ecological niches (see Schluter (2000) and references within). Following this theory, environmental variation triggers divergent natural selection, leading to the emergence of niche specialists. In many cases, species under the same ecological conditions differ in their ability to adapt to new niches, even when they stem from the same ancestor (Gillespie, 2004; Pinto et al., 2008). Therefore, other factors such as genetic constraints may also play an important role in the process of functional divergence.

The process of functional divergence, or departure of a gene from its ancestral function, is constrained by the requirement to maintain the original function: mutations that confer a new function are likely to interfere with the ancestral function and therefore are eliminated by negative selection. This constraint can be relaxed when selection for the ancestral function is weakened, either through gene duplication (and therefore redundancy), or through changes to the environment inhabited by the organism. After gene duplication, one copy of the gene may be free to evolve in a new direction if the other continues to perform the ancestral function (neofunctionalization);
alternatively, ancestral functions can be partitioned between the two gene copies, potentially leading to later specialization or subfunctionalization (Ohno, 1970; Lynch & Conery, 2000; Lynch & Katju, 2004). Major changes in the environment or ecological niche can also lead to a relaxation of selective constraint on ancestral functions, although this process is less well characterized. For example, endosymbiotic bacteria have lost many of the genes their free-living relatives need to obtain nutrients from the environment (Moran, 2002), but have also experienced functional divergence in certain genes (Toft et al., 2009).

Prokaryotes are extraordinarily rich in biological diversity, whether measured in terms of number of species (Dykhuizen, 1998; Gans et al., 2005), habitat range (Pikuta et al., 2007), or the breadth of energy sources and biochemical pathways they can exploit in order to survive (Pace, 1997). Even photosynthesis and oxidative phosphorylation—the mainstays of eukaryotic energy metabolism—are bacterial inventions acquired by endosymbiosis during early eukaryote evolution (Dyall et al., 2004). How did this prokaryotic diversity evolve, particularly when the fixation of gene duplications appears to be somewhat more frequent in eukaryotes (Zhang, 2003)? Adaptive evolution in prokaryotes is promoted by at least three main factors: first, a high strength of selection relative to eukaryotes, on account of their generally large population sizes (Lynch, 2003); second, their ability to obtain genes by lateral gene transfer (LGT), which enables the sharing of niche-relevant functions between distantly-related microbes living in the same environment (Ochman et al., 2000); and third, their use of stress-induced hypermutation (McKenzie et al., 2000), which may increase the production of adaptive variants as a “last gasp” response to a challenging environment.

Although we know that these processes can drive ecological adaptation in prokaryotes, identifying the fraction of genetic variation that is associated with these functional changes remains a challenging problem. In the case of bacteria,
Chapter 3. Whole-proteome analysis of functional divergence with *Clusterfunc*

genome analyses must take into account widespread LGT, which means that different genes may not agree on an overall species tree (Dagan & Martin, 2006). This is a considerable problem for analyses of functional divergence, which require a tree in order to determine the branch upon which a particular trait arose. A variety of DNA sequence-based methods have been developed for identifying changes in selective constraints on protein-coding genes and on single amino acid sites and lineages in a phylogenetic tree (Goldman & Yang, 1994; Nielsen & Yang, 1998; Suzuki & Gojobori, 1999; Yang & Bielawski, 2000; Fares et al., 2002b; Yang & Nielsen, 2002; Suzuki, 2004a,b; Zhang, 2004; Berglund et al., 2005; Zhang et al., 2005). At the protein level, Gu (Gu, 1999; Gu & Velden, 2002; Gu, 2006) developed a Bayesian approach to identify functional divergence, which has become the most widely used. The rationale for these methods, and indeed the new method described here, derives from the neutral theory of Kimura (1983), which predicts that residues important for the function of a protein will be under strong functional constraint and therefore evolve slowly. Comparisons of site-specific evolutionary rate or residue conservation between two homologous clades can therefore be used to identify amino acid sites at which selective constraints have changed, potentially indicating functional divergence. Gu defined two types of functional divergence: one in which the site-specific evolutionary rate differs between two groups of homologous sequences (type I), indicating a change in selective constraints; and a second type in which a residue is conserved within each of two groups, but is biochemically different between them. In this chapter, as in the previous one, we employ a test which accounts for both changes in biochemical properties and selective constraints simultaneously.

In the previous chapter, we developed a new distance-based method which explores a bifurcating phylogenetic tree, testing for functional divergence at each node by comparing the two downstream clades to an outgroup in order to identify sites at which selective constraints have shifted. Like other methods, this approach was limited
to tests of one gene at a time, unless the phylogeny of all genes could be fixed in advance. In this chapter, we further develop our method to (i) handle analyses of functional divergence that include hundreds of complete proteomes, (ii) address the fact that the phylogenies of individual proteins do not necessarily agree with the true phylogeny, as is often the case with organisms that acquire genes through LGT and (iii) provide an intuitive probability score for each test which takes the underlying phylogeny of the sequences into account. We infer patterns of radical change for each protein individually, and then cluster species according to the functional categories (derived from COG (Tatusov et al., 2003)) in which they exhibit significant functional divergence. We provide a fast, open source implementation of our method in the C++ program Clusterfunc (Clusters of functional categories). We perform an analysis of functional divergence on 750 bacterial proteomes. This set includes bacteria from various different ecological niches and therefore provides a good dataset for identifying ecology-related functional divergence. Our approach (i) reveals striking patterns of convergent evolution in phylogenetically distinct but ecologically related groups of bacteria, including pathogens, endosymbionts, and thermophiles, (ii) provides additional support for the view that bacteria have a conserved set of core functions, with a more variable metabolic layer and (iii) provides a detailed picture of how individual species of unusual bacteria have diverged from their closest relatives.

3.3 Materials and Methods

Our analysis of functional divergence, the individual steps of which are detailed below, is summarized in Figure 3.1.
Chapter 3. Whole-proteome analysis of functional divergence with *Clusterfunc*

Fig. 3.1: *Clusterfunc* program workflow. After alignments have been built for each gene in the analysis, the alignments are sorted by functional category. In this case, the COG system was used (Tatusov et al., 2003), but the user is free to pick an ontology of their choice. Trees are built for each gene using BIONJ (Gascuel, 1997) and the JTT substitution model (Jones et al., 1992), and sites are scored for functional divergence on each branch. Significance is assessed by simulating a distribution of test scores under a model of neutral evolution, taking the real phylogeny into account and using the False Discovery Rate approach to correct for multiple testing. For each species and functional category, we use chi-squared tests to evaluate whether the species is enriched or impoverished for functional divergence in that category, or not enriched at all. We then cluster species according to similarities in their profile across all 19 categories. This approach enables us to account for LGT while identifying interesting and atypical patterns of functional change in the data, as discussed in the main text.
3.3.1 Sequences, orthology, and alignment

The first step in a whole-proteome analysis of functional divergence is the grouping of orthologs within the species of interest. We leave orthology assignment for which a number of tools are already in use (Altenhoff & Dessimoz, 2009), to the users' choice according to their own needs. For the present analysis, we retrieved pairwise orthology assignments for 750 completely-sequenced bacterial genomes from the OMA database (Schneider et al., 2007; Roth et al., 2008), representing all bacterial data in the October 2009 revision of the database. We chose the OMA project for its very broad phylogenetic coverage, as well as the favourable performance of its algorithm against other current orthology assignment methods (Altenhoff & Dessimoz, 2009). In addition to providing pairwise orthology calls, the OMA algorithm assembles strict orthologous groups in which every member is directly orthologous to every other. The rationale for this strict approach to grouping is the exclusion of paralogs, which is important for a number of potential applications of the OMA database, such as phylogenetic analysis. Unfortunately, these groups are unsuitable for functional divergence analysis across large phylogenetic distances because lineage-specific gene duplications tend to break up genuine orthologs into multiple, overlapping groups (that is, clustering problems arise because pairwise orthologies are not necessarily transitive). Using these groups in our analysis would result in multiple testing of the same clade, each time with overlapping but incomplete sampling of downstream sequences. The inclusion of both orthologs and lineage-specific paralogs in the same group is, however, of no concern in our per-species comparison of divergence between different functional categories of genes, because our method relies on individual gene trees—and not a single "species tree"—to detect functional divergence (see below). Therefore, we decided to build our own groups from the pairwise orthology assignments in OMA, with the less stringent requirement that a chain of pairwise orthologies connect all members of a group. This strategy produces the most appropriate groups for our analysis, where all genuine
orthologs (and possibly lineage-specific paralogs) for a given gene are included in the same group. However, the approach is vulnerable to erroneous orthology calls in the original database, because a single false call will cause two unrelated groups of sequences to be merged. To assess the possible effect of false OMA orthology assignments on our dataset, we used the relevant genomic data at NCBI to assign COG (Tatusov et al., 2003) ontology tags to each sequence. We then calculated the frequency of the modal COG tag in each group (see Figure 3.2). The largest group (4,788 alignments) contained only one COG tag each, validating our approach to grouping homologs. To avoid ambiguity in the clustering of functional categories, we only analyzed these alignments. We then filtered out poorly-characterized groups (annotated with the ambiguous R or S COG categories) and any group containing less than 9 sequences, which we chose as the minimum number required for analysis. This resulted in a final dataset of 3,813 groups which were then analyzed with Clusterfunc. Sequence alignments were built for each group with MUSCLE (Edgar, 2004), using the default parameters. Data on the ecological niches occupied by the species included in the analysis was retrieved from HAMAP (Lima et al., 2009) and from the Genome database at NCBI. A typical alignment of 78 sequences takes 2 minutes and 40 seconds to analyze on a standard desktop computer, including NJ tree-building. At the other extreme, the large-scale analysis reported below (44,416 tests of functional divergence/3,813 alignments) took 60 hours on a 40-node cluster.

3.3.2 Building gene trees

When analyzing entire proteomes for functional divergence, the use of a species tree to infer events on each branch is problematic: extensive horizontal gene transfer (HGT), particularly among prokaryotes, means that genomes may not be related in a tree-like way (Dagan & Martin, 2006). We therefore calculated a tree for each gene (set of homologous sequences) in the dataset using BIONJ (Gascuel, 1997), under the
Fig. 3.2: Number of COG assignments (tags) for each group of homologous sequences. We included both orthologues and paralogues in our sequence groups, because we are interested in functional divergence. The majority of groups consisted of sequences that had all been assigned to the same COG category, suggesting our grouping strategy did not lump together unrelated sequences due to spurious orthology calls. Since COG categories are relatively broad, we do not generally expect functional divergence to cause a sequence to shift from one category to another, an hypothesis that is also borne out by the clustering of related sequences within the same category.
Chapter 3. Whole-proteome analysis of functional divergence with *Clusterfunc*

JTT model of protein sequence evolution (Jones et al., 1992). Calculations for that gene were then made exclusively using the resulting tree.

### 3.3.3 Scoring functional divergence

The first step of scoring functional divergence is the same as that described in the previous chapter: our method steps through the phylogenetic tree and calculates functional divergence scores at each of the inner nodes. Clades on either side of the bifurcation are compared to the closest available outgroup with respect to the BLOSUM62 substitution matrix (Henikoff & Henikoff, 1992). Scores for each column are given by:

\[
FD = \frac{(C_1 - C_2)}{SE_{C_1,C_2}}
\]

where \(C_1\) and \(C_2\) are the mean substitution scores for the transition from clades on either side of the bifurcation in the phylogenetic tree relative to the outgroup and \(SE_{C_1,C_2}\), the standard error for unequal sample sizes with unequal variances is given by:

\[
SE_{C_1,C_2} = \sqrt{\frac{(s^2)_1}{n_1} + \frac{(s^2)_2}{n_2}}
\]

where \(s^2\) and \(n\) are the variances and sample sizes of \(C_1\) and \(C_2\).

### 3.3.4 Significance testing

For each inner node tested, a simulated sequence alignment was created using the JCprot model according to the gene-specific phylogenetic tree calculated above. That is, the distance matrix for the simulated data was equal to that of the real data, but sequences were evolved under a model that represents molecular evolution without selection constraint, as all amino acid transitions are equally likely. The simulated alignment was tested in the same way as the real data, resulting in a null distribution.
of the test score against which P-values for the real data were calculated. These values were then corrected for multiple testing by the False Discovery Rate method (Benjamini & Hochberg, 1995), with an alpha value of 5%. Following this procedure, branches on the tree that still possess at least one significant site were considered to be under functional divergence for the purposes of enrichment and clustering.

3.3.5 Enrichment analysis

Once all alignments were analyzed, we performed three different enrichment tests to ask three different biological questions of the data. We used a chi-squared test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where $O_i$ is the observed frequency of genes/alignments under functional divergence, $E_i$ is the expected frequency and $n$ is the number of possible outcomes of each event. We used the enrichment tests to identify species and categories of genes that experienced significantly more (enriched: $O_i - E_i > 0$) or significantly less (impoverished: $O_i - E_i < 0$) functional divergence when compared to the background level (that is, $P < 0.05$ in a chi-squared test). We then calculated the enrichment status of each category within each species, in order to identify lineage-specific shifts in the pattern of functional divergence.

3.3.6 Hierarchical clustering

We created a heatmap from the enrichment status of functional categories within species to help visualize the structure in our large dataset. To do this we used the heatmap.2 function from the gplots library in R. This function performs two-dimensional hierarchical clustering according to Euclidean distance and outputs a heatmap together with a corresponding dendrogram. Visualizing the results of the analysis in this way allows identifying unusual patterns of functional divergence—in
Chapter 3. Whole-proteome analysis of functional divergence with Clusterfunc

particular, functional categories or convergent functional divergence among phylogenetically unrelated sequences.

3.3.7 Implementation

Clusterfunc was implemented in C++ and is available under the GNU General Public License v.3 for Linux, Mac and Windows. The code was written using the GNU Scientific Library and the Bio++ libraries (Dutheil et al., 2006). The program is accompanied by full documentation and enables the user to perform several different kinds of analysis, including the identification of lineage-specific functional divergence in a gene-of-interest (such as that reported in the previous chapter) and the kind of multi-proteome investigation reported here. The latest version of the code and documentation is available at http://www.bioinformatics.org/clusterfunc.

3.4 Results and Discussion

3.4.1 Asymmetric functional divergence in bacteria

An obvious sign of functional divergence would be a set of homologues that spans multiple COG categories, e.g. change of a ribosomal protein into a molecular chaperone. In this study we focus only on those alignments where all sequences have the same COG annotation. This represents the majority of homologues and is a reflection of the relatively broad character of the COG categories. The kinds of functional shifts that we detect on the basis of conserved, radical amino acid substitutions are therefore more subtle and not noticeable from simply comparing the COG classifications across homologous sequences. We used chi-squared tests to evaluate the differences in functional divergence between COG gene categories in our dataset (see Figure 3.3). We compared the proportion of positive tests for functional divergence within each of the 19 COG categories to the background expectation, which was calculated by combining
all categories. If genes in different functional categories have similar propensities to undergo functional divergence, we would expect the proportion of positive tests in each category to be similar to the mean, resulting in few significant cases of enrichment. However, fourteen of the nineteen categories were either enriched or impoverished for functional divergence, while only five categories did not deviate significantly from the background expectation. To test whether this polarization of our dataset was simply due to an artifact—for instance, the use of a non-conservative enrichment test—we performed simulations in which the genes in our original dataset were randomly assigned to one of the 19 COG categories before testing for enrichment. In these simulations, events of functional divergence were much more evenly distributed among the categories, so that the great majority (93%) of categories were neither enriched nor impoverished for divergence relative to the background level. This result indicates that the probability of functional change is not evenly distributed among the real categories: there is a stark division between enriched and impoverished categories. This supports the idea that bacterial proteomes comprise a relatively unchanging core (that is, genes in impoverished categories) coupled with a set of more variable functions (enriched categories) (Mushegian & Koonin, 1996; Lake et al., 1999; Makarova et al., 1999).

The impoverished categories are almost exclusively those involved with information storage and processing, including DNA replication, recombination, and repair (L); transcription (K), ribosome biogenesis (J); and cell division (D). Metabolic genes were among those enriched for functional divergence, including genes involved in the metabolism of coenzymes (H), secondary metabolites (Q), carbohydrates (G), amino acids (E) and nucleotides (F). Along with these metabolic categories, cell wall and envelope genes (M) and cellular defense mechanisms (V) were the most enriched categories in our analysis, highlighting the critical role of the environment in directing lineage-specific episodes of functional change. Taken together, our results agree with a number of previous reports indicating that proteins involved in information
processing are more conserved across large evolutionary distances than those involved in metabolism (Azuma & Ota, 2009). An additional point bears emphasizing here: since our method controls for the level of conservation at each node on the tree, the significance of a particular substitution pattern depends on the background evolutionary rate: in slow-evolving proteins, relatively conservative substitutions may be detected as significant events of functional divergence, whereas only very unusual substitution patterns will attain significance in fast-evolving proteins. Therefore, our results indicate that information processing genes are not only more conserved than others purely in terms of evolutionary rate, but that they also experience less functional change even taking this low rate of sequence evolution into account. Why are informational genes under greater functional constraint than the rest of the proteome? One possibility, which follows Crick’s concept of the “frozen accident” (Crick, 1968), is that too many other genes depend on the basic functions of translation, transcription, and repair: functional changes in these genes would disrupt many other systems in the cell. This hypothesis is supported by the observation that the COG category containing protein trafficking and chaperones (O) is also impoverished: the core activities of generalist chaperones such as GroEL and DnaK are required for the proper folding of many different proteins in bacterial cells.

3.4.2 Host interactions constrain functional change in pathogenic and symbiotic bacteria

Does the ecological niche of an organism influence the pattern of functional change it experiences? To answer this question, we evaluated the enrichment of functional divergence in each species relative to the others in our dataset. To calculate the enrichment status of each species, we used the same statistical strategy as employed for enrichment by functional category: we calculated a background proportion of successful tests for functional divergence over all species, and then compared this to the proportion
3.4. Results and Discussion

Fig. 3.3: Different kinds of genes experience different levels of functional divergence. Proportion(FD) is the proportion of tested branches with at least one functionally divergent site across all gene trees in a particular functional category. Categories are labeled according to the COG ontology system (Tatusov et al., 2003). Fourteen of the nineteen categories fall into two groups: significantly enriched and impoverished. Most information processing genes (K, J, L) fall into the latter group, while metabolic functions (E,F,G,H,P,Q) and genes involved in defense (V) or found on the cell surface (M) are enriched for radical change. RNA processing and modification (A) genes are not significantly impoverished, despite a relatively low proportion of positive tests for functional divergence; this may be due to the small number of tests in this category (104).
for each species individually using chi-squared tests (the individual enrichment status for each of the 750 species is provided as Supplementary Table Ch3-S1). We also used chi-squared tests to identify associations between these three enrichment patterns (enrichment, impoverishment, or neither) and organism lifestyle, as is summarized in Table 3.1. While there was no significant difference between psychrophiles and mesophiles in terms of functional divergence (chi-squared = 0.9762, P = 0.6138), thermophilic bacteria were significantly more likely to have proteomes enriched for functional divergence in comparison to mesophiles (chi-squared = 38.1, P = 1.2 \times 10^{-7}). This enrichment in the set of phylogenetically scattered thermophiles may reflect the convergent adaptation of their proteomes to higher temperatures, a process that requires changes in amino acid composition (Singer & Hickey, 2003; Zeldovich et al., 2007).

Interestingly, we found that all bacteria that interact with a host as an integral part of their lifestyle (including pathogens, parasites, symbionts and commensals) were significantly impoverished for functional divergence in comparison to their free-living relatives (see Table 3.1). This result is somewhat surprising because pathogens and symbionts generally experience higher rates of evolution than free-living bacteria, although much of the increase can be attributed to heightened genetic drift (Moran, 2002). Our results suggest that once the overall conservation level of proteins is accounted for, these bacteria have undergone less functional change than their free-living relatives. This result can be explained by greater ecological constraints on host-associated bacteria, which must adapt to the highly specific environment of their host. In particular, pathogenic and symbiotic bacteria preferentially lose metabolic genes as they no longer require the capacity to exploit as wide a range of nutrient sources as free-living bacteria (Moran, 2002). Since these are precisely the kind of genes that are most amenable to functional change (Figure 3.3), their loss from host-associated bacteria may explain the relative impoverishment of functional divergence in these proteomes.
Table 3.1: Effect of organism lifestyle on functional divergence. Associations between lifestyle and enrichment for functional divergence: the numbers of genomes in each category are given in the form Lifestyle/Comparison. Significance was assessed with Yates-corrected chi-squared tests, or Fisher tests when the expected count was lower than 5 for any one cell in the contingency table. Significance codes: N.S. = P > 0.05, * = P < 0.05, ** = P < 0.01, *** = P < 0.001. If an association was significant, "under" or "over" denote the direction of the shift associated with the lifestyle being tested. For instance, interactors are significantly impoverished compared to free-living bacteria.
Chapter 3. Whole-proteome analysis of functional divergence with Clusterfunc

Also, the remaining genes may be under strong constraints imposed by the specialized environment they are in, limiting therefore any opportunity for functional divergence (Toft & Fares, 2008). However, variability in genome size is a complicating factor in this analysis because host-associated bacteria tend to have smaller genomes than their free-living relatives (For instance, endosymbiotic bacteria of insects underwent substantial reduction in the gene content, with genomes sizes ranging between 144kb and 792 kb depending on the host (for example, Moran (2002); McCutcheon & Moran (2007); McCutcheon et al. (2009)).

Since functional divergence often follows gene duplication (Conant & Wolfe, 2008), it might be expected that larger genomes would be enriched for new functions in comparison to smaller ones. Does genome size alone account for the observed differences between host-associated and free-living bacteria? To test this possibility, we modeled genome enrichment and impoverishment for functional divergence as a function of lifestyle (host-associated vs. free-living) and genome size (in nucleotides) using a generalized linear model (see Table 3.2). Both terms were significant, with host-associated bacteria significantly more likely to be impoverished (P = 1.28 x 10^{-13}) and, perhaps surprisingly, a modest tendency towards impoverishment in larger genomes (P = 0.0212). Therefore, variation in genome size does not account for the observed differences in functional divergence between host-associated and free-living bacteria.

To better define the effect of lifestyle on functional divergence, we identified the functional categories with the greatest consistent differences in enrichment status between host-associated and free-living bacteria. Interestingly, genes involved in vesicular transport and secretion systems (U) were enriched for functional divergence in host-associated bacteria but neither enriched nor impoverished in free-living bacteria, while signal transduction genes (T) were impoverished in host-associated bacteria but enriched in their free-living relatives. This pattern can be readily understood in terms of the lifestyles of host-associated bacteria, as pathogens use elaborate secretion systems
3.4. Results and Discussion

<table>
<thead>
<tr>
<th>Term</th>
<th>Coefficient</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.265</td>
<td>8.23 x 10^{-7}</td>
</tr>
<tr>
<td>Lifestyle (Host-assoc.)</td>
<td>-2.536</td>
<td>1.28 x 10^{-13}</td>
</tr>
<tr>
<td>Genome size (bp.)</td>
<td>-1.799 x 10^{-7}</td>
<td>0.0212</td>
</tr>
</tbody>
</table>

Table 3.2: Effect of bacterial lifestyle and genome size on functional divergence. We used a generalized linear model with binomial errors to assess the impact of lifestyle and genome size on the enrichment or impoverishment of genomes for functional divergence. The saturated model was fit with the glm function in R, and simplified to a minimal adequate model with the step function, which determined that the interaction was not significant. Both lifestyle and genome size have a significant impact on enrichment status, with host-associated bacteria and bacteria with larger genomes more likely to be impoverished for functional divergence.

for delivering toxins and other virulence factors to their host (Baron, 2010), while symbionts provision their hosts with nutrients as part of their mutually beneficial relationship (Douglas, 1998; Sandstrom et al., 2000). In addition, the impoverishment in host-associated signal transduction genes may reflect their adaptation to a relatively constant host environment, which is considerably more stable than the fluctuating conditions experienced by their free-living relatives.

3.4.3 Unusual lineage-specific patterns of functional divergence

In order to visualize the results of our functional divergence analysis, we performed two-dimensional hierarchical clustering on the enrichment status (enriched, impoverished, or neither) associated with each species and functional category—that is, we clustered species according to similarities in their enrichment status across the 19 functional categories, resulting in the heatmap and dendrogram in Figure 3.4. This is a powerful and intuitive way to represent our results because it reveals the overall patterns in the data—such as the extreme conservation among informational genes, particularly those involved in ribosome biogenesis (J)—while also highlighting
individual, lineage-specific exceptions to the general trends. In this section, we demonstrate the utility of this approach by using the heatmap to identify species that have undergone major functional shifts.

Although top-level bacterial groups (such as the divisions of the proteobacteria, the Firmicutes, Actinobacteria, and so on) are not resolved in our dendrogram of functional divergence (Figure 3.4), family and genus-level relationships often are, perhaps because of close phylogenetic relatedness, shared gene content, and similarity of ecological niche. This allows us to identify individual species with atypical patterns of functional divergence. A particularly striking case is that of the *Bartonella* genus, which are a group of intracellular parasites that infect and replicate in erythrocytes (Anderson & Neuman, 1997). Of the four *Bartonella* species in our dataset, only one—*Bartonella bacilliformis*—is enriched for functional divergence in cell motility genes (N), with the others being either impoverished (2 species) or neither enriched nor impoverished (1 species). Remarkably, this is the only member of the genus that possesses flagella (Brenner et al., 1991). Since erythrocytes lack an active cytoskeleton, they cannot be induced to take up external bacteria by invagination (Dramsi & Cossart, 1998). Instead, erythrocyte invasion by *Bartonella* species is an active process (Dehio, 2001). The mechanism employed by *Bartonella bacilliformis* involves the use of its flagella (Scherer et al., 1993) and is more efficient than that of other *Bartonella* species, with up to 80% of erythrocytes infected (Ihler, 1996; Dehio, 2001). This appears to be a clear case where our approach has identified an interesting, lineage-specific case of adaptation to a specialized ecological niche.

Our heatmap turns up surprises even among relatively well-characterized species. As expected, closely related *E. coli* and *Shigella* strains cluster together at the bottom of the dendrogram (Figure 3.4), with one exception: *E. coli* SMS 3-5, a multidrug-resistant, heavy-metal tolerant strain isolated from a polluted industrial environment (Fricke et al., 2008). This bacterium is distinguished from other *E. coli* strains on the
3.4. Results and Discussion

Fig. 3.4: Visualizing high-level patterns of functional divergence. We used hierarchical clustering to reveal the main patterns of functional divergence in our dataset of 750 bacterial proteomes. (a) The complete heatmap, with a dendrogram corresponding to category clustering across the top, and species clustering on the left hand side. Visualizing the data in this way reveals the extreme impoverishment of proteins involved in ribosome biogenesis (J), as well as the enrichment of categories involved in interaction with the environment (P, E, M, G) across all species. (b,c) Lineage-specific events of functional divergence picked out from the heatmap (dendrogram colors denote the regions expanded upon—a larger version of the complete heatmap is available as Supplementary Material).
basis of relaxed functional constraint in the categories of cell motility (N) and protein modification, turnover and chaperones (O): the others are largely impoverished for functional divergence, while SMS 3-5 is not. This profile correlates well with what is known about the biology of this strain, which is unique among sequenced *E. coli* genomes in possessing a second, intact lateral flagellar system called Flag-2, in addition to the normal peritrichious flagella found in other *E. coli* strains (Ren et al., 2005; Fricke et al., 2008). This system was originally characterized in a different strain, 042, where it has been rendered nonfunctional by a frameshift mutation in one of the component genes (Ren et al., 2005), although it appears to be complete in SMS 3-5 (Fricke et al., 2008). Interestingly, the Flag-2 gene cluster also contains flagellum-specific chaperones and proteins involved in post-translational modification, perhaps explaining the functional shift in the O category.

Another *E. coli* proteome with an unusual pattern of functional divergence is 017:K52:H18 (strain UMN026), a multidrug-resistant strain that causes urinary tract infections (Manges et al., 2001). Unique among *E. coli* and *Shigella* species, this strain is enriched for functional divergence among genes involved in secretion (U). Investigation of the genes underlying this enrichment revealed functional divergence in the VirB8 and VirB9 genes, which encode core proteins in a Type IV secretion system found only in two *E. coli* strains—UMN026 and 018 (ED1a), although the latter species is neither enriched nor impoverished in this category. In other bacteria, Type IV systems are involved in the exchange of DNA with the environment, as well as the delivery of effector proteins to host cells (Cascales & Christie, 2003). Since these two proteins are important components of the Type IV secretion systems of other bacteria, functional divergence in these genes may be involved in adapting the system to an UMN026-specific role (see Figure 3.5).

To gain further insight into the possible implications of the UMN026-specific changes in these proteins, we mapped the specific residues under functional divergence
Fig. 3.5: Amino acid residues under functional divergence in *E. coli* UMN026 VirB8. **Right:** the structure of a Type IV secretion system found only in two strains of *E. coli*. CM = cytoplasmic membrane, OM = outer membrane. The complex structure is based on that of Baron (2006). In UMN026, the central complex proteins VirB8 and VirB9 are under functional divergence. **Left:** Of the six sites detected by *Clusterfunc* that could be mapped to the VirB8 crystal structure (Bailey et al. 2006), there is evidence that five are involved in forming the VirB8 homodimer, so that functional divergence at these positions might alter the quaternary structure of the complex.
in VirB8 (also output by Clusterfunc) onto the *Agrobacterium tumefaciens* crystal structure (see Figure 3.5, Bailey et al. (2006)). Of the 15 sites under functional divergence, 6 could be mapped onto the crystallized region of the protein (see Table 3.3 for the others). Of these 6, 5 are at or close to positions previously shown to be of functional importance. Thr-194 (residues numbered according to the *A. tumefaciens* sequence), detected as being under functional divergence, is 3.8 Angstroms from Thr-192. Mutating Thr-192 to Met results in a variant that is stable, but cannot complement a VirB8 deletion mutant (Kumar & Das, 2001), indicating that the function of the protein is sensitive to changes at this site. In addition, Thr-194, itself moderately conserved among VirB8 homologues, is directly adjacent to two highly conserved sites, Ala-195 and Thr-196. Thr-196, which Clusterfunc also detected as being under functional divergence in UMN026, is directly involved in the stabilization of the VirB8 homodimer (Bailey et al., 2006), as is Leu-211, another functionally divergent site. Finally, two additional sites identified by Clusterfunc are at positions that suggest they may have an indirect role in dimerization. Val-218 is located between two other residues (Leu-217 and Val-219) that are involved in dimer formation, while Phe-127 is adjacent to Ser-128, a conserved residue that stabilizes the interaction surface on VirB8. The function of the other site detected under functional divergence, Val-183, is currently unknown. Taken together, these results indicate that functional divergence in *E. coli* UMN026 VirB8 has occurred at residues important in forming the homodimer, which may have important implications for the overall structure and function of the complex. With no crystal structure available for VirB9, it is more difficult to evaluate the functional significance of the sites detected there. Further, we detected functional divergence on 19 branches of the VirB9 tree, suggesting that this protein experiences a more general pattern of radical change.
### Table 3.3: Sites under functional divergence in *E. coli UMN026* VirB8

Out of 15 sites under functional divergence, 6 could be mapped to a homologous position in the VirB8 structure from *A. tumefaciens* (Bailey et al., 2006).

<table>
<thead>
<tr>
<th><em>E. coli UMN026</em></th>
<th><em>Agrobacterium tumefaciens</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>R11</td>
<td></td>
</tr>
<tr>
<td>D21</td>
<td></td>
</tr>
<tr>
<td>T25</td>
<td></td>
</tr>
<tr>
<td>V57</td>
<td></td>
</tr>
<tr>
<td>D58</td>
<td></td>
</tr>
<tr>
<td>Y60</td>
<td></td>
</tr>
<tr>
<td>D63</td>
<td></td>
</tr>
<tr>
<td>H71</td>
<td></td>
</tr>
<tr>
<td>T73</td>
<td></td>
</tr>
<tr>
<td>Y116</td>
<td>F127</td>
</tr>
<tr>
<td>L176</td>
<td>V183</td>
</tr>
<tr>
<td>D187</td>
<td>T194</td>
</tr>
<tr>
<td>R189</td>
<td>T196</td>
</tr>
<tr>
<td>E206</td>
<td>L211</td>
</tr>
<tr>
<td>T213</td>
<td>V218</td>
</tr>
</tbody>
</table>

Note: Table 3.3 displays the sites under functional divergence in *E. coli UMN026* VirB8. Out of 15 sites under functional divergence, 6 could be mapped to a homologous position in the VirB8 structure from *A. tumefaciens* (Bailey et al., 2006).
3.4.4 Conclusion

The identification of functional divergence and ecological adaptation from sequence data is an important and interesting goal in evolutionary biology, with the potential to deepen our understanding of the evolution of individual traits and species, as well as the processes of evolution as a whole. Bacteria display an astonishing capacity for adaptation to different lifestyles and ecological niches, but investigating the evolution of these traits is problematic because their phylogenetic context is often unclear. Here, we have circumvented this problem by evaluating functional divergence on gene trees and then clustering species by gene functional category. Our approach revealed overall patterns that have characterized functional change during the evolution of bacteria, including strong constraint on informational genes and host-induced constraints on pathogens and other host-associated bacteria. It also identified lineage-specific events of atypical functional divergence, such as the use of flagella by *Bartonella bacilliformis* to invade host erythrocytes. Although used here to perform a large-scale analysis on bacteria, our method can be applied to any biological system at several different levels of organization, from individual amino acid sites in a single comparison up to metabolic pathways and gene categories in large numbers of proteomes. An open-source implementation of the method, with documentation, is available from the project website at http://www.bioinformatics.org/clusterfunc.
Chapter 4

No Rosetta Stone for a sense-antisense origin of aminoacyl tRNA synthetase classes

4.1 Related manuscript


4.2 Introduction

This chapter presents an evolutionary analysis of the chaperone Hsp70 which, perhaps surprisingly, plays an important role in an ongoing debate about the origin of translation. As will be detailed below, a Hsp70 gene with a long antisense open reading frame (AS-ORF) in the oomycete *Achyla klebsiana* has been proposed as an existence proof—or, more colourfully, a “Rosetta Stone”—suggesting how the two classes of aminoacyl tRNA synthetase genes (aaRS) might have originated during the evolution of the modern translation system. Here, we present several bioinformatic
Chapter 4. No Rosetta Stone for a sense-antisense origin of aminoacyl tRNA synthetase classes

analyses suggesting that the Rosetta Stone is an artifact of biased GC-content and the high conservation of Hsp70, highlighting an unexpected consequence of the functional importance of molecular chaperones.

Translation is the process in which an mRNA molecule is read by the ribosome, which uses the sequence of ribonucleotide bases to direct the synthesis of a specific protein (Stryer et al., 2002). Critical to the specific translation of ribonucleotide codons into amino acids are a group of enzymes called the aminoacyl-tRNA synthetases. These enzymes specifically join amino acids to their cognate tRNAs prior to translation, thereby implementing the genetic code.

AaRS can be divided into two very different structural classes, with ten members each. The evolutionary origin of these enzymes, and therefore the system of translation as we know it, is a tantalizing mystery. Rodin & Ohno (1995) made the dramatic proposal that the two structural classes of aaRS arose on the opposite strands of the same DNA molecule. This proposal was based on the observation that two conserved motifs of Class I aaRS (the HIGH and KMSKS motifs) could potentially be encoded by the complementary strands of DNA sequences coding for two conserved motifs present in Class II aaRS (Motifs 2 and 1, respectively).

The Rodin-Ohno hypothesis received important support from Carter & Duax (2002), who identified a locus from the oomycete Achyla klebsiana as a possible “Rosetta stone” for sense/antisense coding of proteins related to Class I and Class II aaRS. Their work was based on an earlier report that heat shock protein 70 (Hsp70) and an NAD-specific glutamate dehydrogenase (NAD-GDH) are encoded as a sense/antisense pair by a single DNA sequence in A. klebsiana (LeJohn et al., 1994b). The reported overlap between the genes was extensive, with over 1800 basepairs of the proposed NAD-GDH gene being located on the reverse complement of the Hsp70 gene (see Figure 4.1). Since canonical dehydrogenases are structurally similar to Class I aaRS, and Hsp70 has structural homology to Class II aaRS, Carter & Duax (2002) concluded
that the *A. klebsiana* gene proved that such structurally divergent proteins could be encoded on opposite strands of the same DNA molecule. Antisense ORFs were later reported opposite Hsp70 genes in *Drosophila auraria* (Konstantopoulou et al., 1995) and a variety of other organisms (Rother et al., 1997; Silke, 1997), and opposite the Hsp70-related gene Grp78 of *Neurospora crassa* (Monnerjahn et al., 2000). The concept of two genes evolving as a completely overlapping sense/antisense pair is surprising, given the heightened selective constraints that would act on both sequences, and is unprecedented outside of virus genomes.

Although LeJohn and colleagues performed a thorough biochemical characterization of the NAD-GDH activity in *A. klebsiana*, including purification of the enzyme (LeJohn et al., 1994a,b; Yang & LeJohn, 1994), their evidence that the NAD-GDH enzyme is encoded by the ORF opposite Hsp70 is questionable (see Results and Discussion). Here we present evidence that the antisense ORF is spurious, even though it is present in many species, and we identify a different oomycete gene as a more probable candidate for the locus encoding the NAD-GDH enzyme.

### 4.3 Materials and Methods

#### 4.3.1 Analysis of Aphanomyces euteiches EST data

*Aphanomyces euteiches* ESTs (Gaulin et al., 2008) homologous to the *A. klebsiana* Hsp70/AS-ORF genomic locus were identified by BLASTN searches at NCBI, using as a query the sequence of the whole *A. klebsiana* genomic locus (6575 bp made by merging GenBank accession numbers U02504 and U02505; LeJohn et al. (1994b)). The top-scoring 280 ESTs (BLASTN E < 3 x 10^{-15}) were retrieved and assembled into contigs using CAP3 (Huang & Madan, 1999). Two contigs corresponding to probable *A. euteiches* Hsp70 genes were identified. Contigs 1 and 2 had 83% DNA sequence identity to each other, and 87% and 79% identity respectively to the *A. klebsiana* genomic
Chapter 4. No Rosetta Stone for a sense-antisense origin of aminoacyl tRNA synthetase classes

sequence. The proteins encoded by contigs 1 and 2 had 92% amino acid sequence identity to each other, and 97% and 92% amino acid sequence identity respectively to the *A. klebsiana* protein. We reasoned that contig 1 is the ortholog of the *A. klebsiana* Hsp70 gene sequenced by LeJohn et al, and show the locations of the 54 ESTs making up this contig in Figure 4.1.

A contig of *A. euteiches* ESTs coding for the proposed NAD-GDH (Figure 4.1) was identified by TBLASTN searches at AphanoDB (Madoui et al., 2007) using the *Neurospora crassa* protein as a query. The sequence shown in Figure 4.1 was assembled by merging GenBank accession numbers CU354392 and CU354866, AphanoDB contig Ae_15AL7142, and AphanoDB sequence trace file NX0AINT6YK19CM1.SCF.

### 4.3.2 FUGUE searches

FUGUE (Shi et al., 2001) searches were performed using the *A. klebsiana* AS-ORF protein sequence (accession number AAA17563; LeJohn et al. 1994b) and the protein sequence encoded by the *A. euteiches* NAD-GDH contig described above. The program was run on the FUGUE web server at http://tardis.nibio.go.jp/fugue/prfsearch.html, which searches the HOMSTRAD database (Mizuguchi et al., 1998). HOMSTRAD is a collection of high-quality, manually curated structure-based alignments, which FUGUE searches for distant but potentially legitimate homologies to the query sequence.

### 4.3.3 Phylogenetic tree construction

Bacterial homologs of the protein encoded by AS-ORF were identified by a BLASTP search at NCBI, using the AS-ORF sequence as a query, and the Hsp70 genes complementary to these AS-ORFs were retrieved. To assemble a set of *dnaK* sequences without an antisense ORF, we retrieved the protein and nucleotide sequences of the 1,000 BLASTP hits to *Escherichia coli* O157:H7 *dnaK* (accession number NP_285706.1). We removed non-dnaK genes from this set, and then counted the...
number of stop codons in the reverse complement of their coding sequences. We chose
the sequences with the highest number of stop codons, but avoided including numerous
closely-related sequences to increase the phylogenetic coverage of the “without AS-
ORF” gene set. This resulted in 27 dnaK genes with at least 10 stop codons in
the reverse complement of their coding sequences. The accession numbers of these
DnaK proteins are provided in Supplementary Table Ch4-S1. The DnaK sequences
were aligned with MUSCLE (Edgar, 2004) using the default parameters. We used
ProtTest (Abascal et al., 2005) to pick the appropriate model of protein evolution
(RtREV+I+G+F) and the trees were built using phyml (Guindon & Gascuel, 2003),
with 100 bootstraps. A consensus tree was produced using CONSENSE, which is part
of the PHYLIP package (Felsenstein, 1989). The Majority Rule (Extended) method
was used to construct the tree and assign bootstrap values to branches.

4.3.4 Simulations

We identified 8 dnaK genes used in the tree above which had more than 100
nucleotides of C-terminal sequence that did not overlap with the corresponding AS-
ORF. We aligned the protein sequences of these genes with MUSCLE (Edgar, 2004) and
used this alignment to build a codon-based nucleotide alignment. The codon alignment
was split into two sub-alignments, corresponding to the regions of the dnaK genes
which did and did not overlap the AS-ORF. To assess sequence conservation within
each alignment, we used DNADIST from the PHYLIP package (Felsenstein, 1989). DNADIST generated a table of pairwise Jukes-Cantor distances (Jukes & Cantor, 1969)
between the eight sequences, and we took the mean pairwise distance as a measure of
conservation. We performed 100-fold bootstrapping on the shorter non-overlapping
alignment. Significance was tested using a t-test with 1 degree of freedom, based on
the value:
Chapter 4. No Rosetta Stone for a sense-antisense origin of aminoacyl tRNA synthetase

classes

\[
    t = \frac{(Conservation_{overlapping} - Conservation_{non-overlapping})}{SE_{Conservation_{non-overlapping}}}
\]

The t-value was 3.738, resulting in \( P = 0.083 \) for a one-tailed t-test.

4.4 Results and Discussion

Although many Hsp70 genes contain an antisense ORF (AS-ORF) on their opposite strand (Rother et al., 1997), the only reports that this AS-ORF codes for a glutamate dehydrogenase are three consecutive papers published by LeJohn and colleagues about the \textit{A. klebsiana} locus (LeJohn et al., 1994a,b; Yang & LeJohn, 1994). The experimental evidence that the \textit{A. klebsiana} AS-ORF codes for NAD-GDH hinges on the specificity of the polyclonal antibody used in these studies. This antibody was raised against purified \textit{A. klebsiana} NAD-GDH protein, but it was subsequently found to have dual specificity against both NAD-GDH and Hsp70 in cell extracts (LeJohn et al., 1994a; Yang & LeJohn, 1994). When the antibody was used to screen an \textit{A. klebsiana} cDNA expression library in \textit{\lambda}gt11, four cDNA clones were isolated and all of these appeared to be transcripts of the 3' end of Hsp70 (LeJohn et al., 1994a; Figure 1); two of the four cDNAs had poly(A) tails. Even though the recovery of Hsp70 cDNAs from the library is consistent with the antibody's anti-Hsp70 activity (which was demonstrated and commented upon by (LeJohn et al., 1994a)), LeJohn et al. (1994b) pursued the hypothesis that the large AS-ORF on the opposite strand might code for NAD-GDH. Although LeJohn et al. did demonstrate transcription of both strands of the locus (Figure 9 of LeJohn et al. (1994a)), we cannot find any experimental evidence in their papers that the AS-ORF actually codes for the observed NAD-GDH enzyme. The immunological reaction between the antibody and the cDNAs cloned in \textit{\lambda}gt11 cannot be interpreted as proof that the AS-ORF codes for NAD-GDH, because the antibody has been shown (LeJohn et al., 1994a; Yang & LeJohn, 1994) to have
specificity for another protein (Hsp70) that is made by the same cDNAs. Moreover, the genomic structure of 10 exons and 9 introns that was proposed for the complete AS-ORF gene (LeJohn et al., 1994b) is not supported by any cDNA or EST evidence, and the structures of the four cDNAs cloned by LeJohn et al. (1994a) is consistent only with them being derived from Hsp70 mRNA (not spliced AS-ORF mRNAs; Figure 4.1).

The putative AS-ORF protein has 20% amino acid sequence identity to known dehydrogenases when an alignment is forced (LeJohn et al., 1994b; Carter & Duax, 2002). While this level of sequence identity does not preclude a distant but valid homology, protein-protein BLAST searches of the AS-ORF against the NCBI non-redundant database result in no significant hits to any member of the canonical NAD-GDH family. In fact, there are no significant full-length (1063 aa) hits of any kind to the AS-ORF protein, but only to that portion of the sequence that overlaps with Hsp70 on the opposite strand (601 amino acids at the C-terminus of AS-ORF). To investigate the relationship between the AS-ORF and typical dehydrogenases, we used FUGUE (Shi et al., 2001), a program that searches for distant but biologically relevant homologies by fitting a query sequence against a structure database that contains archaeal, eukaryotic and bacterial representatives of the GDH family. Whereas submitting the NAD-GDH sequence of Neurospora crassa (Kapoor et al., 1993) to FUGUE results in a highly significant hit to the dehydrogenase structure (Z-score = 8.44), submitting the AS-ORF sequence results in no significant hits (the best hit has a Z-score of 2.85). FUGUE also provides the option to search using an alignment of PSI-BLAST-derived homologs of the query sequence. Using this option, the AS-ORF sequence returned a hit (Z = 4.67), but this was not to a dehydrogenase structure, and the score was below the recommended cutoff provided by the program (Z = 6). Therefore, it seems that the AS-ORF is not homologous to previously-described dehydrogenases in either sequence or structure.

We then investigated whether the NAD-GDH activity of A. klebsiana could be
Fig. 4.1: Map of the *A. klebsiana* genomic region as annotated by LeJohn et al. (1994b). The Hsp70 gene is oriented from left to right and contains no introns. AS-ORF is a putative gene on the opposite DNA strand, consisting of 10 exons and 9 introns. Exon 10 of AS-ORF overlaps with the Hsp70 gene. LeJohn et al. (1994b) proposed that AS-ORF is the gene coding for NAD-GDH. Grey arrows show the positions and transcriptional orientations of the four *Achlya* cDNAs sequenced by LeJohn et al. (1994a), and of 54 ESTs from the orthologous locus in *Aphanomyces*. There are no ESTs or cDNAs corresponding to transcription from right to left in either species.
encoded by a locus other than that characterized by LeJohn et al. (1994b). Our analysis made use of recent EST data (Gaulin et al., 2008) from Aphanomyces euteiches, a closely related oomycete in the same family (Saprolegniaceae) as A. klebsiana. Using the N. crassa NAD-GDH protein sequence (Kapoor et al., 1993) as the query in a TBLASTN search, we obtained an A. euteiches contig encoding a protein fragment (508 aa). Over the 614-position region where all three sequences (N. crassa NAD-GDH, A. euteiches contig, and A. klebsiana AS-ORF) could be aligned, the N. crassa sequence displayed 52% amino acid identity to the A. euteiches contig and 15.1% identity to the translation of the A. klebsiana AS-ORF (Figure 4.2). The level of identity observed between the N. crassa and A. euteiches sequences is typical of that observed between members of the NAD-GDH protein family (Kersten et al., 1999). In contrast to the AS-ORF, this protein returned highly significant hits to canonical glutamate dehydrogenases in both BLASTP and FUGUE searches (Z-score = 6.61). We suggest that this gene encodes the NAD-GDH enzyme of A. euteiches. We were also able to find an A. euteiches contig orthologous to the Hsp70/AS-ORF locus of A. klebsiana. While there was abundant evidence of transcription of the Hsp70 gene at this A. euteiches locus (54 ESTs), there were no ESTs corresponding to transcription of the complementary strand. Further, there were no ESTs corresponding to exons 1-9 of the AS-ORF, which do not overlap the Hsp70 sequence (Figure 4.1). These data suggest that A. euteiches has an NAD-GDH enzyme that is a typical member of the glutamate dehydrogenase family, and that this enzyme is not encoded by the A. euteiches counterpart of the AS-ORF but at a different locus. We propose that the observed NAD-GDH biochemical activity of A. klebsiana (Yang & LeJohn, 1994) is encoded by an A. klebsiana gene orthologous to the candidate oomycete NAD-GDH gene we identified in A. euteiches (Figure 4.2), but that this A. klebsiana gene has not yet been sequenced.

Our analysis of the A. euteiches EST data revealed two Hsp70 genes in this
Fig. 4.2: Protein sequence alignment of *Neurospora crassa* NAD-GDH with the partial putative NAD-GDH inferred from *Aphanomyces euteiches* EST data, and the translation of the *A. klebsiana* AS-ORF. The *N. crassa* NAD-GDH shows 52% amino acid identity to the *A. euteiches* sequence but only 15.1% identity to the *A. klebsiana* sequence over the region where all three sequences could be aligned. Shading indicates residue identity. Alignment generated by T-Coffee (Notredame et al., 2000), using the default parameters, and visualized in Jalview (Clamp et al., 2004).
species, one orthologous (AeHsp70-1) and the other paralogous (AeHsp70-2) to the Hsp70/AS-ORF locus of *A. klebsiana*. We noticed that the AeHsp70-1 and AeHsp70-2 sequences both also contain long antisense ORFs. Indeed, BLASTP searching the *A. klebsiana* AS-ORF predicted protein sequence against the NCBI protein sequence database reveals a family of proteins with several phylogenetically-scattered eukaryotic members and about 30 bacterial members (Rother et al., 1997; Carter & Duax, 2002). Many of these bacterial sequences have been annotated as NAD-GDH enzymes on the basis of their similarity to the *A. klebsiana* AS-ORF. The presence of database homologs of AS-ORF might suggest that it is a functional gene even if, as argued above, it does not code for NAD-GDH. However, the observation that all the BLAST hits to the AS-ORF sequence are within the region that overlaps Hsp70 raised the possibility that the apparent conservation of this family might be an artifact due to the presence on the opposite strand of sequences coding for Hsp70, one of the most conserved proteins yet described (Gupta & Golding, 1993). Indeed, every one of the apparent AS-ORF homologs in the database contains an intact Hsp70 gene on its opposite strand. We used bacterial sequences to test this hypothesis in two ways.

First, if there exist two kinds of Hsp70 gene—those with a functional AS-ORF and those without—then the two groups must have very different evolutionary histories, which should be reflected in their phylogeny. If the AS-ORFs are functional, there should be phylogenetic separation of Hsp70 sequences with and without the antisense gene. But if the AS-ORFs are artifacts, the Hsp70 sequences should cluster generally according to the species phylogeny. We retrieved 29 DnaK (bacterial Hsp70) sequences from GenBank that were encoded by the complements of the AS-ORF homologs identified by BLAST. We also retrieved 27 DnaK sequences that do not have an intact AS-ORF, based on the presence of at least 10 stop codons in the reverse complement of their coding sequences. A maximum likelihood, bootstrapped protein phylogeny of these sequences revealed extensive mixing between the DnaK sequences with and
without AS-ORFs, a result that is difficult to explain if the antisense sequences are real genes (Figure 4.3). Furthermore, there is a clear correlation between the presence of an intact AS-ORF and the GC-content of the Hsp70 gene (Figure 4.3), as expected on purely statistical grounds (Merino et al., 1994; Silke, 1997).

Second, eight of the dnaK sequences with AS-ORFs that were used to build the phylogeny contained a region of more than 100 nucleotides at the 3' end of dnaK that was not overlapped by the AS-ORF, which permitted a further test of the hypothesis. The AS-ORFs are in the same frame as the dnaK genes, such that the third codon positions in one are base-paired with the first codon positions in the other. We were therefore able to divide a nucleotide alignment of these eight sequences into two sub-alignments containing the third codon positions of dnaK from the regions that (i) overlapped and (ii) did not overlap the antisense ORF. If the AS-ORF sequences are a real gene, the additional selective constraint provided by the first codon positions of this gene should result in higher conservation of codon third positions in the overlapping, but not the non-overlapping, regions of dnaK. We compared sequence conservation between the two sub-alignments, using bootstrapping of the non-overlapping region to increase robustness, and taking mean pairwise Jukes-Cantor differences as a measure of conservation. We find that the difference in conservation between these regions is not statistically significant (P = 0.083, one-tailed t-test). This suggests that the apparent conservation of the bacterial AS-ORF homologs is due to the high conservation of Hsp70 on the other strand.

4.4.1 Conclusions

Our results point to two conclusions. First, the family of ORFs that has been identified antisense to Hsp70 genes in many species is unlikely to be a real gene family because these ORFs do not show the patterns of phylogenetic distribution, sequence constraint, or even transcription that would be expected if they code for functional
4.4. Results and Discussion

Wigglesworthia glossinidia
Buchnera aphidicola str C. Cinara codri
Salmonella enterica subsp enterica serovar K
Salmonella enterica subsp enterica serovar D
Enterobacter sakazakii ATCC BAA-894
Citrobacter koseri ATCC BAA-895
Escherichia coli B
Pasteurella multocida subsp multocida
Actinobacillus pleuropneumoniae serovar 1
Vibrio Harveyi ATCC BAA-1176
Vibrio parahaemolyticus RIMD 2210633
Coelella psychrophilus 34H
Shewanella putrefaciens ATCC 700545
Candidatus Carsonella ruddii PV
Acinetobacter baumanni ATCC 17978
Francisella tularensis subsp tularensis WRC-3
Pseudomonas putida W619
Martinoana sp MED121
Costella endosymb. Amblyopia americana
Nitrosomonas eutrophii C91
Azotobacter sp
Burkholderia phylotype PsJN
Methylylphilum bacterium HTCC2181
Coelella psychrophilus 34H 2
Roseovarius sp HTCC2601
Sapillus stefata E-37
Rhodobacteraceae bacterium HTCC2654
Phaeobacter galaxiaciensis 210
Roseovarius rubrifaciens JSM
Oceanisola granoviae HTCC3611
Loktanella vestfolisovina SKAS3
Rhodobacteraceae bacterium HTCC2150
Paracoccus denitrificans PDO-2Z2
alpha proteobacterium HTCC2255
Methylbacterium populi B0097
Methylbacterium chloromethanicum CM4
Rhodopseudomonas palustris 769-1
Brucella sp. ATCC 11846
Brucella canis ATCC 23365
Bacteriella bacilloformis KS583
Agarimonas sp SBS-9A1
Oceanisola alvandina HTCC2833
Pseudomonas sp SKAS5
Anaplasma marginale str St Maries
Enricha canis str Lake
Wolbachia endosymb. Drosophila melanogaster
Orientia tsutsugamushi Borying
Ricktetsia typhi str Million
Lawsonia intracellularis PHE/MN-00
delta proteobacterium MLMS-1
Anaeromicrobacter dehalogenans ZCP-C
Collinsella aerofaciens ATCC 25988
Opitutaceae bacterium TAV2
Akkrarmania muzzinii ATCC BAA-835
Flavobacteria bacterium BALJ
Parabacteroides distasonis ATCC 8503

Fig. 4.3: Protein phylogeny of DnaK sequences with and without an antisense glutamate dehydrogenase ORF. An asterisk denotes the presence of an AS-ORF. The phylogeny recapitulates the standard view of bacterial relationships and reveals extensive mixing between genes with and without an AS-ORF, as expected if the AS-ORFs are artifactual. Branches are colored by third-position GC-content: high (66.7-100%, red), intermediate (33.4-66.6%, yellow), and low (0-33.3%, blue).
proteins. We therefore doubt that any of the ORFs antisense to Hsp70 genes in any species are functional. Second, the evidence that one particular member of this putative family—the AS-ORF of *A. klebsiana*—codes for the NAD-GDH enzyme in this species is tenuous. We do not doubt the biochemical evidence that an NAD-GDH enzyme exists in *A. klebsiana*, but we do question LeJohn et al.’s claim that the gene coding for this enzyme is located antisense to Hsp70 because (i) their papers do not provide any biochemical evidence linking the NAD-GDH enzyme to the Hsp70 locus, and (ii) we have found a different oomycete gene that seems likely to code for NAD-GDH. Although proof of our proposal would require further biochemical experiments such as the purification and direct amino acid sequencing of an oomycete NAD-GDH enzyme, the results presented here cast substantial doubt on the proposed overlap between the NAD-GDH and Hsp70 genes, on which the Rosetta stone idea depends (Carter & Duax, 2002). The validity of the Rodin-Ohno hypothesis for the proposed dual-strand origin of the two classes of aaRS therefore rests on the original evidence presented for a statistically significant sense/antisense relationship between the conserved aaRS motif sequences (Rodin & Ohno, 1995), and on the demonstration that an equal spacing between conserved motifs of Class 1 and 2 aminoacyl synthetases is compatible with function (Pham et al., 2007).
Chapter 5

The effect of chaperonin buffering on protein evolution

5.1 Related manuscript


5.2 Introduction

In previous chapters, the focus has been on the evolution of chaperone molecules themselves. This focus is now shifted to examine the evolution of client proteins—the proteins which chaperones help fold. Through their modulation of the relationship between a protein’s primary sequence and final structure—that is, between genotype and phenotype—chaperones have been proposed to facilitate the adaptive evolution of their client proteins (Todd et al., 1996; Rutherford & Lindquist, 1998; Fares et al., 2002a; Queitsch et al., 2002; Tokuriki & Tawfik, 2009a,b; Lindquist, 2010). Based on the protein-folding mechanism of GroEL, Todd et al. (1996) suggested that chaperones might influence the evolution of their clients by ameliorating the structurally-
Chapter 5. The effect of chaperonin buffering on protein evolution

destabilizing effects of some new mutations. However, the first experimental evidence of a role for chaperones in shaping client evolution came from studies of Hsp90, an unrelated chaperone. The pioneering work in this area was that of Rutherford & Lindquist (1998), who demonstrated that Hsp90 suppresses (or buffers) the phenotypic effect of deleterious mutations in its clients, which are mainly signalling proteins. They found that the reduction of Hsp90 activity resulted in the expression of underlying developmental abnormalities in *Drosophila*. When subject to selection, these variants could be enriched in the population to the point where, combined in a single genome, they could no longer be suppressed by restored Hsp90 function. The fixation of a set of mutations in this way might cause an "adaptive leap" from one developmental pathway to another, explaining the phenomenon of "genetic assimilation" that had previously been observed by Waddington (1953). Since this initial discovery, Hsp90-buffered variation has been documented in other eukaryotes including *Saccharomyces cerevisiae* (Cowen & Lindquist, 2005) and *Arabidopsis thaliana* (Sangster et al., 2007, 2008). This latter case is particularly striking because of the plasticity of plant development to environmental cues: despite the flexibility of the wild-type *Arabidopsis* developmental program, as well as the low levels of heterozygosity in this inbreeding species, pharmacological inhibition of Hsp90 resulted in the expression of significant morphological variation, consistent with the idea that a single chaperone molecule can buffer a substantial load of phenotypic mutations.

Work on the chaperonin GroEL/GroES of *Escherichia coli* has provided evidence for another mechanism by which chaperone buffering affects client protein evolution. Moran (1996) suggested that overexpression of GroEL/GroES in endosymbiotic bacteria was an evolutionary response to the high levels of genetic drift—and therefore high mutational load—experienced by these intracellular organisms, the idea being that higher levels of GroEL would enable the cell to continue functioning as deleterious mutations accumulated in the proteome. This hypothesis was supported by Fares
et al. (2002a), who showed that overexpression of GroEL recovered the fitness of *E. coli* strains exposed to strong genetic drift. An important caveat was that GroEL buffering could not entirely recover ancestral fitness, which implies that there is a limit to the effectiveness of a single chaperone. Potentially, this limit could arise either from the eventual titration of the GroEL/GroES system by the continuously increasing mutational load, or the specificity of certain destabilized proteins for other chaperone molecules. The phenotypic recovery induced by over-expression of GroEL alone was, nonetheless, remarkable. Additional evidence for the importance of GroEL buffering comes from a recent bioinformatic analysis, which suggested that GroEL clients experience weaker selection for translationally-optimal codon usage in comparison to nonclients, perhaps due to a reduced need to prevent mistranslation (Warnecke & Hurst, 2010). Rather than being contradictory mechanisms, the authors suggested that GroEL buffering and codon usage may represent two complementary ways by which organisms can limit protein misfolding errors (Warnecke & Hurst, 2010).

The first concrete evidence that chaperonin buffering might act as more than a coping mechanism was provided by (Tokuriki & Tawfik, 2009a), who performed experimental evolution on four enzymes in *E. coli* with and without GroEL/GroES overexpression. Their results showed that GroEL/GroES could maintain the function of enzymes that had accumulated highly destabilizing mutations in their core. Even more interesting was their attempt to enhance the inefficient esterase activity of one of the enzymes, *Pseudomonas* phosphotriesterase, by artificial selection in the presence and absence of GroEL/GroES. The esterase activity that evolved in the presence of GroEL was far more efficient than that which could be obtained without GroEL, because it depended upon a destabilizing mutation that reduced the rate of folding and greatly reduced enzyme activity in the absence of chaperonin buffering. Combined with some existing evidence that functionally important mutations are often destabilizing (Wang et al., 2002; Tokuriki et al., 2008), this result provides a straightforward explanation
for how chaperone buffering of deleterious mutations could be involved in the evolution of new functions in client proteins.

Despite this experimental evidence, the extent to which chaperones facilitate the evolution of their client proteins in nature remains unclear. In particular, chaperones may not only buffer deleterious variants, but also expose them to proteolysis (Kandrro et al., 1994; Tomala & Korona, 2008). Tokuriki & Tawfik (2009a) performed their experimental evolution combining GroEL/GroES overexpression with strong purifying selection during each round of evolution—but if chaperones really do buffer phenotypic variation in their clients, then the strength of selection acting on clients should be weaker than that acting on nonclients. Here, we evaluate the effect of chaperonin buffering on client protein evolutionary rate, using data from 85 gamma-proteobacterial genomes. This question can be approached bioinformatically due to two recent, systematic classifications of the *E. coli* proteome into client and nonclient portions (Kerner et al., 2005; Fujiwara et al., 2010a). Kerner et al. (2005) identified 252 proteins that were repeatedly isolated from GroEL/GroES complexes, of which 85 were found so frequently as to suggest all copies of that protein required assistance from the chaperonin complex in order to fold (obligate clients). Fujiwara et al. (2010) examined the solubility of these clients in GroEL/GroES-depleted cells, and found that 49/85 of the obligate clients of Kerner et al., along with another 8 proteins, were absolutely dependent on the chaperonin complex for folding. After controlling for several factors known to influence evolutionary rate, we compare the evolution of clients and nonclients under all these classifications.

We then examine the evolutionary fate of GroEL client proteins in the *Mycoplasmas*, a group of highly-derived bacteria with small genomes that contains the only organisms lacking GroEL/GroES yet described (Woese, 1987; Lund, 2009). We examine whether the loss of GroEL has lead to a loss of obligate client proteins, or whether *Mycoplasma* client homologs have adapted to life without GroEL, as has been reported
5.3. Materials and Methods

5.3.1 Gamma-proteobacterial structures and alignments

All available crystallized protein structures for the gamma-proteobacteria (mostly from Escherichia coli) were downloaded from the Protein Data Bank (PDB). The resulting dataset contained 1000 PDB entries (and 1075 protein chains—see Supplementary Tables Ch5-S2), representing 20-25% of the *E. coli* proteome and half (126/252) of known GroEL clients (Kerner et al., 2005), although it was not over-enriched for any of the functional categories in the Clusters of Orthologous Groups ontology system (Tatusov et al., 2003). Protein sequences homologous to the structure-associated sequences were retrieved by reciprocal BLAST searching of 85 complete gamma-proteobacterial proteomes (see Supplementary Table Ch5-S1), only considering reciprocal hits with E-values < $10^{-4}$ where the length of the whole protein was within the range of ± 25% of the structure sequence. We limited the set of sequences to this range of lengths in order to ensure that only proteins with the same structure and function would be included. Sets of homologs were aligned with ClustalW using the default parameters (Thompson et al., 1994), and the quality of the alignments was inspected manually. Only those alignment columns that could be aligned to the structure sequence were used in our subsequent analyses.

5.3.2 Analysis of protein evolutionary rate

Classification of the *E. coli* proteome into clients and nonclients was carried out on the basis of the system developed by Kerner et al. (2005), who performed a proteome-wide screen for GroEL clients by trapping and then characterizing proteins encapsulated within GroEL/GroES complexes. GroEL interactors were further
subdivided into facultative (class I and II) or obligate (class III) clients depending on the proportion associated with GroEL/GroES complexes versus the total amount of that protein in the cell. Recently, another study screened for obligate GroEL clients by identifying proteins that aggregate or are degraded in GroEL/GroES-depleted cells (Fujiwara et al., 2010a). Their results overlap with, but do not exactly match, those of Kerner et al. (2005), because about 40% of Class III clients remain soluble during GroEL/GroES underexpression. In our analysis, we use both classifications when assessing the effect of chaperonin buffering. It is possible that these screens failed to identify all GroEL clients in the \emph{E. coli} proteome; however, we do not think that a (presumably small) proportion of unclassified clients among our set of nonclient proteins will have a serious effect on the analyses reported below—if anything, they ought to make the results more conservative.

Data on gene essentiality were downloaded from the SHIGEN Profiling of \emph{E. coli} Chromosome database (Hashimoto et al., 2005; Kato & Hashimoto, 2007). A gene is defined as essential if strains carrying a null mutation cannot grow under any conditions. Protein-protein interactions were quantified using the combined interaction dataset from Bacteriome.org, which contains 7613 experimentally-determined interactions between 2283 \emph{E. coli} proteins (Peregrin-Alvarez et al., 2009). In order to avoid trivially biasing our results towards a greater number of client interactions, we removed all interactions involving GroEL/GroES from the dataset. We used gene expression data from the genome-wide study of Covert et al. (2004), using the dChip-normalized mean mRNA expression value across three replicates for wild-type \emph{E. coli} cells growing in aerobic conditions. In the analyses reported below, we only used expression data when all three replicates were called as present on the array (resulting in data for 226/252 clients and 2889/3892 nonclients in the \emph{E. coli} genome). Repeating the analyses using all expression data (regardless of quality) gave results which were qualitatively the same.
5.3. Materials and Methods

Per-residue estimates of evolutionary rate were calculated as follows: for each column in a protein sequence alignment, we counted the number of pairwise differences between residues \(x\) and the total number of comparisons \(n\). To account for multiple substitutions, we applied the Poisson correction to the proportion of differences to obtain a distance \(d\) for that column:

\[
d = -\frac{19}{20} \log(1 - \frac{20}{19} \frac{x}{n})
\]

Per-residue amino acid contact density was defined as the number of other residues within 4 Angstroms of the site of interest (Toft & Fares, 2010). For each atom in an amino acid, we calculated the Euclidean distance between it and all atoms in the other amino acids in the crystal structure. The distance between two amino acids was taken to be the minimum of the atomic distances between the two residues:

\[
\min(\sqrt{(x_{i1} - x_{2j})^2 + (y_{i1} - y_{2j})^2 + (z_{i1} - z_{2j})^2})
\]

where \(i\) and \(j\) represent all atoms in amino acids 1 and 2, respectively.

To evaluate the effect of chaperonin buffering on evolutionary rate after accounting for essentiality, amino acid contact density, gene expression level, and protein-protein interactions, we performed an analysis of covariance (ANCOVA) using the statistical software R (http://www.r-project.org). The ANCOVA was fit using the `lm` (linear model) function. We used this approximation because this function represents a conservative relationship between the different factors and because modeling the relationships between more than two factors is both computationally expensive and combinatorially prohibitive. We are, however, aware of the fact that this linear modeling might represent a simplistic view of the interaction between factor effects, although a systematic bias towards the covariance of two particular factors due to the model is unlikely. We compared the fit of models including (i) all main effects and interactions and (ii) just main effects with an ANOVA. The model without interactions fit the data significantly worse \((p < 10^{-15})\), prompting the retention of the more
complex model. To circumvent the problem of model over-fitting, we assessed the significance of individual terms in the ANCOVA using the step function implemented in R. This function uses Akaike's information criterion (AIC) to remove terms that do not significantly improve model fit—that is, models that increase the AIC value—resulting in the set of minimal adequate models discussed below.

In order to evaluate whether our results were due to bias introduced by the phylogenetic non-independence of the 85 gamma-proteobacterial genomes used, we re-calculated Poisson distances using a reduced subset of our data comprising one species per genus and re-analyzed the data as described above. The representative sequence from each genus was chosen at random because none of the within-genus sequences presented distinctive characteristics regarding genome size, codon composition, etc. The results were qualitatively very similar (see Supplementary Tables Ch5-S9-12), suggesting that the effects discussed below are not an artifact of biased phylogenetic coverage. The numbers reported below are from the original analysis, which uses all of the available data. We would also like to stress that biases in our results due to the phylogenetic non-independence of sequences should affect clients and nonclients equally, and should not, therefore, bias tendencies systematically one way or the other. In other words, sequences are clearly related by a phylogeny but the proteins divided into the “client” and “nonclient” portions of the proteome are not.

5.3.3 Mycoplasma sequences and analysis

Four Mycoplasma genomes that contain a GroEL homolog (Mycoplasma penetrans HF-2, Mycoplasma genitalium G37, Mycoplasma gallisepticum R, and Mycoplasma pneumonae M129) and seven that do not (Mycoplasma pulmonis UAB CTIP, Mycoplasma capricolum subsp. capricolum ATCC 27343, Mycoplasma mobile 163K, Mycoplasma arthritidis 158L3-1, Mycoplasma mycoides subsp. mycoides SC str. PG1, Mycoplasma hyopneumonae 232, and Mycoplasma synoviae) were downloaded from
NCBI (accession numbers provided in Supplementary Table Ch5-S3), with the presence or absence of GroEL being assessed manually using NCBI Web-BLAST, using the *M. penetrans* HF-2 protein sequence (NP_757486.1) as the initial query. The *E. coli* proteome was divided into clients and nonclients as described above, and each set of genes was BLASTed against these 11 genomes. Only 29 of the 252 *E. coli* GroEL clients had significant hits against all 11 Mycoplasma genomes (defined as an E-value < $10^{-7}$, which we found by manual experimentation to be a good trade-off between false positive and false negative presence/absence calls). In order to increase the size of our dataset, we also included genes which were present in at least 3/4 Mycoplasma genomes with GroEL and 6/7 genomes without. This resulted in a set of 57 Mycoplasma homologs of *E. coli* GroEL clients and 282 homologs of non-clients, with 9-11 Mycoplasma sequences per gene.

To evaluate whether GroEL client proteins in *E. coli* have been preferentially lost from *Mycoplasmas* that lack GroEL, we used an analysis of covariance (fitted with the glm function in R, with binomial errors) in which a binary response variable reports the presence or absence of a homolog to each protein in the *E. coli* proteome in a given *Mycoplasma* species (where presence is defined as a BLASTP hit at E < $10^{-7}$) and with client/nonclient status, essentiality, number of protein-protein interactions and mRNA expression level as the explanatory variables.

To investigate the evolution of GroEL client proteins within the *Mycoplasmas*, we built protein sequence alignments from the 57 genes homologous to *E. coli* GroEL clients using MUSCLE 3.7 (Edgar, 2004) under the default parameters. These alignments were used to build 100-bootstrap maximum likelihood phylogenetic trees with RaxML 7.04 (Stamatakis, 2006), using a substitution model chosen by ProtTest (Abascal et al., 2005) in each case. 49/57 consensus trees suggested the topology shown in Figure 5.2, in which a single loss of GroEL occurred within the *Mycoplasmas*. This consensus topology was then used for comparison of selective constraint between
Chapter 5. The effect of chaperonin buffering on protein evolution

Mycoplasmas with and without GroEL.

For each client and nonclient alignment, we calculated the nonsynonymous-to-synonymous substitution ratio (dN/dS) under maximum likelihood using the program codeml, from the PAML package version 4.0 (Yang, 2007). In each case, we compared two models: one in which a single dN/dS ratio applies across the tree, and one in which the genomes with- and without GroEL evolve under different ratios. These models were compared with a likelihood ratio test for which the null distribution is a chi-squared distribution with one degree of freedom. The numbers of client and nonclient homologs which were evolving significantly faster in GroEL-lacking Mycoplasma were then compared with a chi-squared test.

Mycoplasma genomes lacking GroEL were not impoverished for GroEL clients when compared to Mycoplasmas with GroEL, raising the possibility that intrinsic changes in these proteins occurred in non-GroEL Mycoplasmas that made them independent from GroEL. To evaluate this possibility, we tested whether the amino acid compositions or molecular weights of the proteins from Mycoplasmas with GroEL differed significantly from those in Mycoplasmas without GroEL. The molecular weights of client homologs in Mycoplasmas with- and without-GroEL were calculated by summing the weights of their constituent amino acids and, for each protein, calculating a mean protein molecular weight for Mycoplasmas with- and those without-GroEL. Weights were compared with a Wilcoxon two-sample paired signed rank test. Amino acid compositions were compared in a similar way, with mean proportions for each amino acid in each protein in Mycoplasmas with- and without-GroEL being compared with Wilcoxon two-sample paired signed rank tests, using the Bonferroni correction to account for multiple testing.
5.4 Results and Discussion

5.4.1 The functional importance of GroEL client proteins

As outlined in the introduction to this chapter, the idea that molecular chaperones buffer the phenotypic effects of mutations in their clients is critical to the hypothesis that chaperones facilitate adaptive evolution (Rutherford & Lindquist, 1998; Tokuriki & Tawfik, 2009a). Assuming that most mutations affecting phenotype are—at least individually—neutral or deleterious (Kimura, 1983), if selection against such mutations is weaker in GroEL clients than nonclients due to a buffering effect (Tokuriki & Tawfik, 2009a), then clients ought to evolve faster than nonclients. However, precisely the opposite trend has been reported (Hirtreiter et al., 2009), with GroEL preferentially chaperoning slow-evolving proteins. The same trend was apparent in our dataset of 1,075 gamma-proteobacterial proteins, with clients evolving significantly more slowly than nonclients (mean Poisson distance in clients = 0.147, nonclients = 0.178, p < 10^{-15}, Mann-Whitney U test). Does this result falsify the chaperone buffering hypothesis? No, because it does not take into account the many factors that influence evolutionary rate. For instance, if clients are enriched for characteristics that constrain evolution, these might mask a buffering effect. We compared the functional importance of clients and nonclients in terms of essentiality, number of protein-protein interactions, and mRNA expression levels. All these factors have previously been observed to influence evolutionary rate (Krylov et al., 2003; Drummond & Wilke, 2008; Wolf et al., 2010), although their relative importance is a matter of some debate (Bloom & Adami, 2003; Jordan et al., 2003; Pal et al., 2003). We found striking differences between clients and nonclients in terms of essentiality and protein-protein interactions, with clients significantly more likely to prove essential upon single-gene knockout (43/248 essential clients, 242/3900 essential nonclients, P < 10^{-3}, chi-squared test), and participating in significantly more protein-protein interactions than nonclients (mean 13.7 for clients,
5.9 for nonclients, $P = 1.6 \times 10^{-14}$, Mann-Whitney U test), even after interactions with GroEL/GroES are removed from the dataset. At the level of mRNA expression, clients are expressed at a significantly higher level than nonclients in wild-type *E. coli* cells growing aerobically (mean client probe intensity 2186, nonclient 1290, $P < 10^{-15}$, Mann-Whitney U test), although obligate clients were expressed at a lower level than facultative clients (1683 vs. 2425, $P = 0.0008328$, Mann-Whitney U test).

Taken together, these results suggest that client proteins are, on average, of greater functional importance than nonclients. Since a higher proportion of essential genes, a higher number of protein-protein interactions, and higher mRNA expression levels are all either weakly or strongly associated with a decrease in evolutionary rate (Krylov et al., 2003; Drummond & Wilke, 2008; Wolf et al., 2010), their influence must be accounted for when evaluating the effect of chaperonin buffering on client protein evolution.

### 5.4.2 GroEL buffers the evolution of its obligate clients

To evaluate the relative contributions of chaperone buffering, essentiality, network connectivity (in terms of protein-protein interactions) and expression level on evolutionary rate, we performed an analysis of covariance with one response variable, per-residue Poisson distance, and five explanatory variables: two categorial (client/nonclient, essential/nonessential), and three continuous: number of protein-protein interactions, mRNA expression level (in mean probe intensity across three replicates), and amino acid contact density. This final covariate, which quantifies the number of other residues within a 4 Angstrom radius of a particular amino acid site, has previously been shown to correlate negatively with evolutionary rate: that is, amino acids surrounded by large numbers of other residues (such as in the protein core) evolve relatively slowly (Thorne et al., 1996; Goldman et al., 1998; Bustamante et al., 2000; Mintseris & Weng, 2005; Bloom et al., 2006; Conant & Stadler, 2009; Toft & Fares, 2010).
We performed four different analyses, in which GroEL clients were defined in four different ways: (i) all 252 GroEL/GroES interactors identified by Kerner et al. (2005)—that is, both facultative and obligate clients; (ii) 85 obligate clients only (as defined by Kerner et al. (2005)); (iii) 57 obligate clients as defined by Fujiiwara et al. (2010); and (iv) the 34 obligate clients classified by Kerner et al. (2005) that do not depend on GroEL/GroES for solubility. An important difference exists between categories (ii) and (iii). Kerner et al. (2005) classified clients according to their enrichment in GroEL/GroES complexes. If more than 4% of the total cellular content of a particular protein was associated with GroEL/GroES, they inferred that all copies of that protein needed to interact with the chaperonin complex in order to reach their native conformation, making it an obligate “Class III” client. Proteins which were reliably isolated from GroEL/GroES complexes at lower levels of enrichment were assigned to two classes of facultative clients. Fujiiwara et al. (2010) took a more direct approach, measuring the solubility of Class III clients in GroEL/GroES-depleted cells. They found that 34/85 of the Class III clients did not depend on GroEL/GroES for solubility (“Class III-” clients), suggesting that the enrichment of a protein in GroEL/GroES complexes is correlated with, but does not exactly predict, obligate dependency. Combining the 49/85 (60%) of Class III clients that are dependent on GroEL/GroES for solubility with another 8 proteins not previously included in Class III, these authors proposed a new class of obligate GroEL/GroES clients (Class IV). The results of our analyses are summarized in Table 5.1, which shows the effect of chaperonin buffering on the evolution of each of these four groups of clients (all clients, Class III, Class IV, and Class III-).

Regardless of the way in which GroEL clients and nonclients are defined, our analysis recovers the well-documented negative correlations between expression levels, numbers of protein-protein interactions, residue contact density, and evolutionary rate (see Table 5.1), with two exceptions. Firstly, when the Class III- proteins are compared
are provided as Supplementary Tables S2-3. 

<table>
<thead>
<tr>
<th>Class II</th>
<th>Class III</th>
<th>Class IV</th>
<th>All classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.180 X 10^-5</td>
<td>1.317 X 10^-5</td>
<td>3.981 X 10^-6</td>
<td>6.44 X 10^-7</td>
</tr>
<tr>
<td>5.445 X 10^-6</td>
<td>1.178 X 10^-5</td>
<td>3.049 X 10^-6</td>
<td>6.03 X 10^-7</td>
</tr>
</tbody>
</table>

Expression level

Protein-protein interactions

Residue contact density

Nonessential
to the rest of the proteome, the main effect of expression level changes sign, with higher expression levels associated with a moderate increase in evolutionary rate. Deletion of the client/nonclient term recovers the negative correlation between expression level and evolutionary rate observed with all other client/nonclient classifications, suggesting that this effect is due to the interaction between these two terms. Class III- proteins are highly enriched in GroEL/GroES complexes, but do not depend on the chaperonin for solubility. Fujiwara et al. (2010) noted that half of the Class III- proteins bind RNA or DNA, and overall the class is enriched for positively-charged amino acids. On the basis of this evidence, they proposed that these proteins are frequently recovered from GroEL/GroES complexes because they can bind the negatively-charged interior surface of the GroEL/GroES cavity, not because they required GroEL/GroES for folding—an hypothesis that is supported by the lack of any significant chaperonin buffering effect in this class (see below). The division of our dataset into this group of proteins on the one hand, and a mix of the "genuine" clients and nonclients on the other, may have produced the interaction giving rise to the change in sign of the expression level main effect.

Secondly, we find that in two of our four analyses essential genes are evolving faster than nonessential ones when these other factors are taken into account. To explore the reason for this unexpected result, we compared essential and nonessential genes in several ways. A simple comparison of mean evolutionary rate recovers a moderate but statistically-significant reduction in rate in essential genes, as has previously been reported (mean Poisson distance in essential genes = 0.160, nonessential = 0.174, $P < 10^{-15}$, Mann-Whitney U test); (Koonin, 2005; Wolf et al., 2010). Essential genes participate in more protein-protein interactions (16.1 vs 5.4, $P < 10^{-15}$, Mann-Whitney U test) and have higher expression levels (2500 vs. 1239, $P < 10^{-15}$, Mann-Whitney U test) than those that are nonessential, which may go some way to explaining why they are essential in the first place. To identify the factor(s) underlying the effect
of essentiality in our analyses of covariance, we re-analyzed the data while dropping each one of the other factors in turn. Failing to account for residue contact density or client/nonclient status resulted in no change in sign or significance of the essentiality term, but its significance was abolished when either of the terms modelling the number of protein-protein interactions or expression level were dropped. In comparison, dropping any one of protein-protein interactions, expression levels, or essentiality from the ANCOVA neither changed the sign nor abolished the significance of the client/nonclient term (see Table 5.2).

The method used to classify GroEL clients had a striking effect on the analysis: considering both facultative and obligate clients together, there was a marginally significant effect of chaperonin buffering on evolutionary rate, with an increase in rate associated with clients—although, unlike the effects discussed below, the significance of this term was abolished when we re-analyzed a nonredundant subset of our data to test for biases arising from phylogenetic non-independence (see Methods). We note, however, that this could also have resulted from the reduction in statistical power when decreasing the number of sequences in our analyses. When only Class III clients were considered, the significance and effect size of this rate shift was greatly increased ($P = 0.00301$), and became even more striking among Class IV clients which absolutely depend on GroEL/GroES for solubility ($P < 10^{-15}$). The analysis suggested that once other factors are accounted for, these obligate clients show an increase in mean per-residue Poisson distance of 0.4263 relative to the rest of the gamma-proteobacterial proteome. As discussed above, there is no significant effect of chaperonin buffering when only Class III- proteins are considered—that is, proteins enriched in GroEL/GroES complexes but that remain soluble in GroEL/GroES-depleted cells. These results lead to two conclusions: (i) at least among the gamma-proteobacteria, GroEL/GroES facilitates the accumulation of amino acid substitutions in its obligate clients, but not in all proteins with which it regularly interacts; and
<table>
<thead>
<tr>
<th>Dropped term</th>
<th>Essentia lity</th>
<th>Client/nonclient Density</th>
<th>Interactions</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residue contact density</td>
<td>-3.553 x 10^{-1} (&lt; 10^{-15})</td>
<td>-1.574 x 10^{-1} (6.78 x 10^{-8})</td>
<td>-2.540 x 10^{-2} (2.33 x 10^{-8})</td>
<td>-6.747 x 10^{-5} (1.25 x 10^{-12})</td>
</tr>
<tr>
<td>Protein-protein interactions</td>
<td>-4.296 x 10^{-2} (0.43291)</td>
<td>-1.559 x 10^{-1} (0.00271)</td>
<td>-1.926 x 10^{-2} (0.00113)</td>
<td>-4.522 x 10^{-5} (0.01686)</td>
</tr>
<tr>
<td>Expression level</td>
<td>4.327 x 10^{-1} (0.254376)</td>
<td>-1.068 x 10^{-1} (3.35 x 10^{-6})</td>
<td>-1.622 x 10^{-2} (5.61 x 10^{-10})</td>
<td>-4.104 x 10^{-3} (0.003124)</td>
</tr>
<tr>
<td>Client/nonclient status</td>
<td>-2.736 x 10^{-2} (1.42 x 10^{-11})</td>
<td>-3.267 x 10^{-3} (&lt; 10^{-15})</td>
<td>-1.215 x 10^{-3} (3.49 x 10^{-6})</td>
<td>-2.153 x 10^{-5} (&lt; 10^{-15})</td>
</tr>
<tr>
<td>Essentia lity</td>
<td>-1.680 x 10^{-1} (0.000169)</td>
<td>-3.114 x 10^{-2} (1.98 x 10^{-9})</td>
<td>-1.026 x 10^{-2} (0.295777)</td>
<td>-3.590 x 10^{-5} (0.0198)</td>
</tr>
</tbody>
</table>

**Table 5.2: Effect of deleting ANCOVA terms on remaining probabilities.** This table shows the effect size and P-value (in parenthesis) for each main term after removing one of the other main terms and its interactions. Connections in Figure 5.1 are created if deleting one term changes the probability of the other term by at least one order of magnitude; the width of the line is proportional to the number of orders of magnitude brought about by the deletion. Clients/nonclients are classified according to Fujiwara et al. (2010).
(ii) this buffering effect is most pronounced in client proteins that depend on the GroEL/GroES system for solubility (Class IV clients), as opposed to all proteins which are highly enriched in GroEL/GroES complexes. This relationship is masked in simple comparisons of client and nonclient evolutionary rate due to the increased functional importance of clients. To identify the factors that most directly interfere with the buffering effect, we deleted individual factors from our Class IV ANCOVA and evaluated the effect upon the remaining terms (see Figure 5.1). This approach suggested that gene essentiality and the number of protein-protein interactions were the most important confounding factors.

The increase in evolutionary rate that we observed among GroEL clients might be taken as evidence in favour of the "chaperonin-facilitated adaptation" model of (Tokuriki & Tawfik, 2009a), but we note that this will only hold if mutations which confer new functions are disproportionately likely to interfere with protein folding—that is, to make folding intermediates more difficult to reach. If the effect of chaperonin-mediated buffering is simply to broaden the spectrum of neutral mutations in clients, then the ability of positive selection to promote the fixation of adaptive mutations will be weakened—that is, buffering will mainly act to increase the strength of genetic drift operating on clients. If, however, neofunctionalizing mutations tend to be destabilizing—a proposition for which there is some evidence (Wang et al., 2002; Tokuriki et al., 2008)—then buffering could maintain such variants in the population, making them accessible to positive selection if they confer an advantageous phenotype.

5.4.3 Neutral evolution of GroEL clients in Mycoplasmas?

Certain species of *Mycoplasma* and *Ureaplasma* are unique among sequenced genomes in lacking a chaperonin homolog of any kind (Lund, 2009). Although these bacteria have experienced extensive genome reduction (Woese, 1987), the loss of GroEL is surprising. GroEL is an essential gene in *E. coli* at least in part because several other
Fig. 5.1: The relationships between chaperonin buffering, gene essentiality, protein-protein interactions, residue contact density, and expression levels. The effect of deleting each main term and its interactions on the remaining terms: the arrows point away from the term being deleted, with the width of the arrow proportional to the change in significance. Colors denote the direction of the change: blue indicates a decrease in the P-value, while orange indicates an increase. The raw data used to generate this figure is provided in Table 5.2. The effect of chaperonin buffering becomes less significant when numbers of protein-protein interactions and gene essentiality are taken into account, suggesting that these are the most important confounding factors in simple comparisons of client and nonclient evolutionary rate.
essential proteins depend on it for proper folding (Lund, 2009). Presumably, the loss of GroEL in Mycoplasmas must have been accompanied by either the loss of client homologs or the loss of their dependency on GroEL for folding. One possibility is that Mycoplasmas invest more in protein degradation in order to prevent aggregation (Wong & Houry, 2004). However, there is now experimental evidence (Fujiwara et al., 2010b) that at least some homologs of E. coli GroEL clients have lost their obligate chaperonin dependency in these bacteria, folding independently when expressed in E. coli. In the present study, our aim was to assess the effect of GroEL loss on the evolution of chaperonin clients in those Mycoplasmas that have lost GroEL. First, we used BLASTP to identify homologs of E. coli clients in 11 complete Mycoplasma genomes and Ureaplasma, comprising 4 genomes which retain a copy of GroEL and 8 which have lost it. Perhaps surprisingly, there was no significant difference in the retention of obligate (Class III/IV) clients and nonclients in 9/12 of these genomes, and in the 3 genomes where the difference was significant (M. capricolum, M. mycoides, and M. synoviae), it reflected preferential retention of client proteins, even though these species have all lost GroEL (see Table 5.3). How can the loss of GroEL have no effect, or even a positive effect, on the retention of obligate clients? A simple comparison of the numbers of retained clients and nonclients does not take into account other factors that might influence the loss of genes in Mycoplasmas. To account for these, we performed an analysis of covariance that indicated that the heightened functional importance of client proteins (discussed above) plays some role in their retention in Mycoplasmas (Table 5.3). In particular, proteins involved in higher numbers of interactions and essential proteins are significantly more likely to be retained in Mycoplasma genomes. Controlling for these covariates, client/nonclient status did not in itself have a significant effect on retention in any of the 12 genomes we analyzed, suggesting that the uncoupling of obligate client folding from GroEL reported in Ureaplasma may also apply in the related Mycoplasmas.
<table>
<thead>
<tr>
<th>Species</th>
<th>Class IV</th>
<th>NonC</th>
<th>P</th>
<th>C/NonC</th>
<th>Essen.</th>
<th>P-P Ints</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. genitalium</em></td>
<td>8/57</td>
<td>424/4087</td>
<td>0.3691</td>
<td>-2.655 x 10^{-1}</td>
<td>1.289 (***)</td>
<td>3.737 x 10^{-2}</td>
<td>1.017 x 10^{-1}</td>
</tr>
<tr>
<td><em>M. penetrans</em></td>
<td>11</td>
<td>600</td>
<td>0.3288</td>
<td>-2.867 x 10^{-1}</td>
<td>1.201 (***)</td>
<td>3.127 x 10^{-2}</td>
<td>6.959 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. gallisepticum</em></td>
<td>10</td>
<td>465</td>
<td>0.1467</td>
<td>-1.812 x 10^{-1}</td>
<td>1.352 (***)</td>
<td>3.828 x 10^{-2}</td>
<td>9.936 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. pneumoniae</em></td>
<td>8</td>
<td>463</td>
<td>0.5226</td>
<td>-3.519 x 10^{-1}</td>
<td>1.191 (***)</td>
<td>3.728 x 10^{-2}</td>
<td>7.745 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. pulmonis</em></td>
<td>10</td>
<td>515</td>
<td>0.2652</td>
<td>7.775 x 10^{-1}</td>
<td>3.723 x 10^{-2}</td>
<td>4.373 x 10^{-2}</td>
<td>4.452 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. capricolum</em></td>
<td>17</td>
<td>572</td>
<td>0.0006768 (***)</td>
<td>5.784 x 10^{-1}</td>
<td>1.074 (***)</td>
<td>3.907 x 10^{-2}</td>
<td>9.791 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. mycoides</em></td>
<td>17</td>
<td>581</td>
<td>0.0008672 (***)</td>
<td>5.535 x 10^{-1}</td>
<td>9.998 x 10^{-1} (***)</td>
<td>3.622 x 10^{-2}</td>
<td>1.026 x 10^{-4}</td>
</tr>
<tr>
<td><em>M. mobile</em></td>
<td>11</td>
<td>483</td>
<td>0.07522</td>
<td>-3.995 x 10^{-2}</td>
<td>1.226 (***)</td>
<td>3.745 x 10^{-2}</td>
<td>1.148 x 10^{-4} (0.00434)</td>
</tr>
<tr>
<td><em>M. arthriditis</em></td>
<td>8</td>
<td>410</td>
<td>0.3189</td>
<td>-7.284 x 10^{-1}</td>
<td>7.322 x 10^{-1} (*)</td>
<td>5.05 x 10^{-2}</td>
<td>5.498 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. hyopneumoniae</em></td>
<td>10</td>
<td>476</td>
<td>0.1694</td>
<td>1.187</td>
<td>4.190 x 10^{-1} (***)</td>
<td>4.258 x 10^{-2}</td>
<td>-5.908 x 10^{-5}</td>
</tr>
<tr>
<td><em>M. synoviae</em></td>
<td>13</td>
<td>452</td>
<td>0.00526 (***)</td>
<td>2.004</td>
<td>9.391 x 10^{-1} (***)</td>
<td>4.145 x 10^{-2}</td>
<td>1.802 x 10^{-5}</td>
</tr>
<tr>
<td>Ureaplasma</td>
<td>10</td>
<td>415</td>
<td>0.0678</td>
<td>1.581 x 10^{-1}</td>
<td>1.321 (***)</td>
<td>4.315 x 10^{-2}</td>
<td>6.825 x 10^{-5}</td>
</tr>
</tbody>
</table>

**Table 5.3: Loss of GroEL clients and non-clients from Mycoplasma genomes.** The Chi-square P-value reported is for a test of association between retention in *Mycoplasma* genomes and client/non-client status in *E. coli*. Proteins that are essential or involved in a high number of interactions are preferentially retained in *Mycoplasma* genomes, with higher mRNA expression levels in *E. coli* also being associated with retention in some cases. Client/non-client status has no significant effect on retention in any species. Significance levels: * = P < 0.05; ** = P < 0.01; *** = P < 0.0001. The numbers reported here are for the Class IV clients of Fujiwara et al. (2010), but the results are qualitatively similar for Class III clients (see Supplementary Table Ch5-S13).
Did the loss of GroEL dependency in client proteins occur before or after the loss of GroEL in *Mycoplasmas*? Although the two events might be expected to be coupled, there is evidence that GroEL is not essential in *M. genitalium* and *M. pneumoniae* (Hutchison et al., 1999; Wong & Houry, 2004), two of the four species which still possess the chaperonin. We addressed this question from an evolutionary perspective, asking whether the loss of GroEL had an effect on the nonsynonymous-to-synonymous substitution ratio ($dN/dS$) in GroEL-lacking *Mycoplasmas*. Our set of *Mycoplasma* homologs of *E. coli* GroEL clients contained 57 proteins, which we used to build 100-bootstrapped, maximum likelihood phylogenetic trees. Interestingly, 49/57 of these unrooted trees had a topology in which the *Mycoplasmas* with GroEL were separated from those without GroEL, the most parsimonious interpretation of such an arrangement being a single loss of GroEL within the *Mycoplasmas* (see Figure 5.2).

If former GroEL clients have accumulated mutations that enable them to fold independently, then this process might be detectable as an elevated $dN/dS$ ratio among client proteins in the *Mycoplasmas* that lack GroEL. We tested this hypothesis using maximum likelihood estimates of $dN/dS$ calculated using codeml (Yang, 2007) on the consensus tree obtained from our client phylogenies. In order to increase the size of our dataset, we considered any *E. coli* homolog, client or nonclient, if it was present in at least 3/4 of the *Mycoplasma* genomes with GroEL and 6/7 without. 28/57 client homologs and 100/282 nonclient homologs experienced significantly relaxed selective constraints in the *Mycoplasma* without GroEL (that is, a two-$dN/dS$ model, with a higher value on the branches without GroEL, was a significantly better fit to the data), but the difference in these proportions did not attain statistical significance ($P = 0.0522$, chi-square test). Although this $P$-value exceeds the standard alpha value, we suggest that the analysis provides weak support for the idea of increased $dN/dS$ in the client proteins of *Mycoplasma* that have lost GroEL, which might represent an evolutionary signature of adaptation to a GroEL-independent folding pathway. A
Fig. 5.2: Phylogeny of *Mycoplasma* genomes with and without GroEL.

49/57 *Mycoplasma* homologs of *E. coli* GroEL clients support a topology in which the *Mycoplasma* species that have retained GroEL (red) cluster to the exclusion of those that have lost it (black), suggesting a single loss of GroEL within this group of organisms. Each client-protein maximum-likelihood tree was built using RaxML, using 100 bootstraps.
plausible alternative explanation, however, is simply that GroEL-lacking *Mycoplasmas* experience a higher rate of genetic drift, which is supported by the remarkable observation that of all nonclient genes for which the two-ratio model fit better than the one-ratio, 100 showed a higher dN/dS on the GroEL-lacking branches, versus only 4 in the GroEL-possessing *Mycoplasmas*. Additional support for this neutral explanation comes from comparisons of amino acid composition and molecular weight between client homologs in *Mycoplasmas* with- or without-GroEL. Fujiwara et al. (2010) reported an enrichment of alanine and glycine residues in Class IV (obligate) clients versus the rest of the *E. coli* proteome, suggesting that this property might distinguish independently-folding from chaperonin-buffered proteins. Such a bias, if also present in the client homologs of GroEL-possessing but not GroEL-lacking *Mycoplasmas*, would provide additional evidence for the acquisition of independent folding exclusively in GroEL-lacking *Mycoplasmas*. A comparison of amino acid frequencies in the client homologs of these two sets of genomes, however, revealed no such pattern. Although we did detect significant differences in the frequencies of certain amino acids (see Table 5.4), there was no systematic bias in the biochemical properties of those amino acids enriched in one group or the other: for instance, valine was enriched in the clients of GroEL-possessing *Mycoplasmas* while isoleucine was enriched in those of GroEL-lacking species. We also compared the molecular weights of client homologs between the two sets of *Mycoplasma* genomes, with the idea that the acquisition of independent folding might lead to increases in the mass of proteins no longer constrained by the volume of the GroEL protein-folding cavity. This test also allowed us to determine whether small changes in the frequencies of multiple amino acids in *Mycoplasma* client homologs might have added up to a significant change in mass, with potential implications for the interaction of the proteins with GroEL. However, we failed to detect a significant difference (P = 0.7771, Wilcoxon two-sample paired signed rank test).

Taken together, these results suggest that the folding of *E. coli* client homologs has
5.4. Results and Discussion

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Proportion (M+)</th>
<th>Proportion (M-)</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.06705602</td>
<td>0.06412697</td>
<td>0.01760264</td>
</tr>
<tr>
<td>R</td>
<td>0.03198514</td>
<td>0.03057579</td>
<td>0.1462221</td>
</tr>
<tr>
<td>N</td>
<td>0.06347089</td>
<td>0.06475842</td>
<td>0.2787419</td>
</tr>
<tr>
<td>D</td>
<td>0.05450019</td>
<td>0.05703637</td>
<td>0.009520734</td>
</tr>
<tr>
<td>C</td>
<td>0.01077632</td>
<td>0.006480572</td>
<td>1.78869 x 10^{-9} (***))</td>
</tr>
<tr>
<td>E</td>
<td>0.06659824</td>
<td>0.0746438</td>
<td>4.331984 x 10^{-7} (***))</td>
</tr>
<tr>
<td>Q</td>
<td>0.03920396</td>
<td>0.03260812</td>
<td>7.853544 x 10^{-8} (***))</td>
</tr>
<tr>
<td>G</td>
<td>0.05637234</td>
<td>0.05740576</td>
<td>0.2034257</td>
</tr>
<tr>
<td>H</td>
<td>0.02010727</td>
<td>0.01717087</td>
<td>1.902904 x 10^{-5} (**)</td>
</tr>
<tr>
<td>I</td>
<td>0.08637445</td>
<td>0.0946842</td>
<td>2.57564 x 10^{-5} (**)</td>
</tr>
<tr>
<td>L</td>
<td>0.09317583</td>
<td>0.0896082</td>
<td>0.01208208</td>
</tr>
<tr>
<td>K</td>
<td>0.09985067</td>
<td>0.1057008</td>
<td>0.001125972 (*)</td>
</tr>
<tr>
<td>M</td>
<td>0.01986343</td>
<td>0.01961691</td>
<td>0.5534994</td>
</tr>
<tr>
<td>F</td>
<td>0.0424104</td>
<td>0.04271334</td>
<td>0.9368301</td>
</tr>
<tr>
<td>P</td>
<td>0.03156366</td>
<td>0.02916403</td>
<td>0.000935111 (*)</td>
</tr>
<tr>
<td>S</td>
<td>0.05786761</td>
<td>0.06298605</td>
<td>8.017592 x 10^{-5} (**)</td>
</tr>
<tr>
<td>T</td>
<td>0.05121503</td>
<td>0.04919868</td>
<td>0.0818883</td>
</tr>
<tr>
<td>W</td>
<td>0.006601632</td>
<td>0.006415254</td>
<td>0.5435505</td>
</tr>
<tr>
<td>Y</td>
<td>0.03264973</td>
<td>0.03365232</td>
<td>0.2528167</td>
</tr>
<tr>
<td>V</td>
<td>0.06992637</td>
<td>0.06387168</td>
<td>1.098794 x 10^{-5} (**)</td>
</tr>
</tbody>
</table>

Table 5.4: Amino acid composition of GroEL client homologs in *Mycoplasmas* with and without GroEL. We used Wilcoxon two-sample paired signed rank tests to compare amino acid compositions of GroEL client homologs in *Mycoplasma* genomes with (M+) and without (M-) a GroEL homolog. P-values were corrected with the Bonferroni correction (for 20 amino acids and alpha = 0.05, P < 0.0025 denotes significance). Significance levels: * = P < 0.0025; ** = P < 0.0005; *** = P < 0.000005.
Chapter 5. The effect of chaperonin buffering on protein evolution

become uncoupled from GroEL in the *Mycoplasmas*, perhaps even in the species that have retained GroEL. Our conclusions are in agreement with those of Clark & Tillier (2010), who recently reported no differences in the folding properties of *Mycoplasma* client and nonclient homologs as predicted by the FoldIndex program (Prilusky et al., 2005). We also note that the results presented here do not exclude the possibility that GroEL clients in *E. coli* acquired chaperonin dependency after the divergence of the *E. coli* and *Mycoplasma* lineages: in this case, the equal retention of client and nonclient homologs in *Mycoplasma* genomes would not reflect the gain of independent folding in former clients, but rather retention of the ancestral state.

5.4.4 Conclusions

Although the models of chaperone-facilitated adaptive change proposed by Rutherford & Lindquist (1998) and Tokuriki & Tawfik (2009a) suggest that chaperone clients should evolve faster than nonclients, the opposite is observed in the case of the *E. coli* chaperonin clients and their homologs. Here we have shown that this pattern is due to the increased functional importance of clients, and that once this is accounted for, client proteins are evolving faster than nonclients. As discussed above, our results support the hypothesis that chaperones facilitate adaptive evolution under the condition that functionally innovative mutations tend to interfere with protein folding. But why do clients tend to be more functionally important? We propose two hypotheses, based on the observation of increased evolutionary rates in clients. Firstly, proteins that are buffered by chaperones might be able to more easily fix functionally innovative mutations despite their structurally destabilizing effects. The acquisition of new functions by these proteins would then lead them to take on a more important role in the cell. Alternatively, proteins that are already performing important functions are highly constrained and therefore might have more need of chaperone-assisted folding following the fixation of functionally innovative mutations. However, if there is no
connection between functional innovation and structural stability, then the effect of chaperonin buffering observed here would simply act to increase the strength of genetic drift acting on clients.
Chapter 6

Discussion

6.1 Chaperoning evolution

Chaperones are a class of ubiquitous, ancient and important molecules that help other proteins in the cell to fold. In this thesis, we have investigated the evolution of molecular chaperones in bacteria, and evaluated the influence they exert on the evolution of their protein clients. These analyses had two main conclusions.

First, we reported the discovery of group II chaperonin genes in ten species of bacteria, demonstrating that the phylogenetic distributions of group I and group II chaperonins overlap within the prokaryotes to a much greater extent than was previously appreciated. Although archaea with both group I and group II chaperonin genes had previously been reported (e.g. Hirtreiter et al. (2009)), this was the first report of bacteria possessing group II genes. A bioinformatic analysis of selective constraints indicated that these group II chaperonins are functional in bacteria, suggesting that organisms from all three domains of life can benefit from the presence of two highly-divergent protein-folding cages in their cells.

This unexpected finding raises a number of questions which can best be addressed experimentally: most importantly, what function are these group II chaperonins fulfilling for the bacteria which possess them? There are several possibilities which
we were unable to distinguish bioinformatically. One possibility is that the group II chaperonins might be helping to fold one or more niche-relevant proteins (such as a locally-useful enzyme) that were also horizontally transferred from methanogenic archaea. We failed to identify any archaeal genes that were found only in bacteria with group II chaperonins, although it could be that the bacterial group II chaperonins have different archaeal clients in different species of bacteria. Alternatively, these extra chaperonins might be helping “endogenous” bacterial proteins to fold. The phylogenetic origin of a chaperonin does not have to match the origin of a client for productive folding to occur, as was observed by Hirtreiter et al. (2009): in *Methanosarcina mazei*, an archaeaon that has both group I and group II chaperonin systems and which has experienced extensive cross-domain gene transfer, there is no correlation between the source of the chaperonin (bacterial or archaeal) and the source of the client protein. Instead, the preference of client proteins for one chaperonin complex or the other is determined by structural features. In principle, this could allow bacteria with both systems to chaperone a much broader range of proteins. To resolve this question, the laboratory of Peter Lund is currently in the process of characterizing these chaperonins experimentally. Whatever the answers, it seems that the transfer of group II chaperonins from methanogenic archaea to syntrophic bacteria is an example of lateral gene transfer as a mechanism of niche adaptation and evolvability, in line with the other cases discussed in Chapter 3.

The second major finding of this thesis is that chaperonins buffer the accumulation of mutations in their client proteins over evolutionary time. The idea that chaperones might act as phenotypic capacitors is, of course, not new: as described above, it was first proposed by Todd et al. (1996) and has been supported by several pieces of evidence over the past decade (Rutherford & Lindquist, 1998; Queitsch et al., 2002; Fares et al., 2002a; Tokuriki & Tawfik, 2009a). This evidence has indicated that GroEL can mitigate the fitness effects of mutations fixed during population bottlenecks (which
may explain our observation of increased functional divergence in the GroEL proteins of intracellular bacteria), as well as buffer the structurally-destabilizing effect of mutations conferring new protein functions. Our contribution has been to show that this buffering activity actually accelerates the accumulation of new mutations in client proteins during millions of years of natural (as opposed to experimental) evolution. Interestingly, the same conclusion has been reached more-or-less simultaneously by Bogumil & Dagan (2010) using a rather different statistical approach, in a paper published in the same issue of *Genome Biology and Evolution*.

These results are interesting and raise a number of questions which could be very worthwhile to pursue. Perhaps the most fundamental question is simply what determines whether a protein can fold independently, or whether it needs help from a molecular chaperone. Some general trends have been described—for instance, small proteins tend to fold independently, whereas larger proteins or proteins with particular folds tend to be enriched among obligate clients (Kerner et al., 2005). Nonetheless, the process by which an independently-folding protein acquires destabilizing mutations to become a chaperone client (or vice-versa) is only poorly understood. A tempting model that would be consistent with our results is "chaperone addiction", in which the transition from independent protein to client is basically one-way: after the fixation of one or a small number of destabilizing mutations, the protein becomes permanently dependent on chaperone buffering, which would weaken selection against the accumulation of further structurally-destabilizing mutations. However, this model is contradicted by the results of Tokuriki & Tawfik (2009a), which suggest that, just as small numbers of mutations can induce GroEL dependence, subsequent individual mutations can significantly re-stabilize a protein that was dependent on chaperone-assisted folding. The strength of selection in favour of such mutations in natural populations is unclear, but in principle the re-acquisition of independent folding by chaperone clients might "free up" the limited number of GroEL complexes in the cell to
help with the folding of other proteins. If, as was suggested by the results of Fares et al. (2002a), there are limits to the buffering capacity of GroEL, then independently-folding proteins might be favoured by selection. One way in which a chaperone-dependent protein might become independently-folding once more would be through the fixation of compensatory mutations in the structural region surrounding the original destabilizing mutation, leading to co-evolution in the functionally-important regions of current and former GroEL clients.

The obvious way to distinguish between these models of long-term and short-term dependency of proteins on GroEL would be to screen for client proteins in several lineages of distantly-related organisms. At present, these experiments have only been carried out in *E. coli*, although the observed increase in evolutionary rate across the gamma-proteobacteria suggests that the list of clients is largely conserved among these organisms. Our finding that *Mycoplasma* client homologs have not been preferentially lost in species that have lost GroEL, however, is most simply explained if these proteins can fold independently. This finding suggests that the identities of chaperone clients might shift over larger evolutionary distances than those considered either in our study, or in that of Bogumil & Dagan (2010). In the absence of this kind of exhaustive experimental characterization, it may be possible to learn more about the evolution of chaperone dependency and independent folding from systems such as the steric chaperone (foldase) LimA of *Burkholderia cepacia* and related bacteria (Hobson et al., 1993). This chaperone has only a single client, the lipase LipA. In species where both genes occur together, the lipase depends on its chaperone for activation, but the lipase is also found in a number of species that lack LimA, where it is capable of folding independently (Pauwels et al., 2007). A phylogenetic analysis of limA-dependent and independent lipases might permit the specific molecular changes underlying the transition from one state to the other to be identified, although the generality of any results obtained from this special case is unclear.
6.2 Functional divergence

The new methodology introduced in this thesis is a piece of software, Clusterfunc, that implements a new, fast approach for identifying amino acid sites under functional divergence on a gene tree. The current advantages of this method over others are (i) speed, permitting analyses of thousands of sequence alignments on a desktop computer in a reasonable amount of time, and (ii) a novel clustering approach that enables the user to extract meaningful patterns of functional divergence from organisms where the species tree is not known (and may not even exist)—which includes the great majority of life as we know it. The method was originally developed in order to analyze a very large tree of chaperonin genes, but was soon exapted to identify divergence associated with endosymbiosis and—on a grander scale—to identify high-level patterns in the evolution of 750 bacterial genomes. In this final form, our analysis of functional divergence identified both previously-described patterns (such as the existence of “core” and “variable” functions in bacterial genomes) and some surprises (including the constrained nature of host-associated evolution).

There are plans to continue our analyses with this method, which have the potential to produce interesting results. We are going to apply the same approach used in Chapter 3 to eukaryotes and archaea, in order to compare the evolution of genes within each functional category across the three domains. One prediction for these analyses would be that the proportion of successful tests for functional divergence will be lower in eukaryotes than in prokaryotes, but differences in the relative rates of change in each gene category may also prove interesting. There is also room for improving the scope and flexibility of the Clusterfunc software: for instance, allowing the user to generate gene trees under maximum likelihood, or allowing the output from similar packages—such as DIVERGE—to be used as input for our clustering analysis. Such enhancements should enable our method to take advantage of continuously improving computer hardware.
Chapter 6. Discussion

6.3 Conclusion

The most important findings of this thesis relate to chaperonins in bacteria: both group I and group II sequences are present in certain bacteria; GroEL and its homologues buffer the accumulation of mutations in their client proteins, at least among the gamma-proteobacteria; and intracellular pathogens have experienced the greatest functional divergence in chaperonin genes, perhaps as an adaptation to the intracellular lifestyle. In addition, we have introduced a new method for detecting functional divergence which can be applied to any gene or set of genes, not just chaperonins; and we have demonstrated that the apparent sense/antisense coding of genes in *Achyla klebsiana* is an artifact of the unusual evolutionary properties of the chaperone Hsp70.

6.4 Another perspective

In writing this thesis, I have tried to explain clearly and concisely the rationale for each project, the details of its execution, and the way in which it connects to the central theme of the work, which is the link between protein folding and evolvability. In structuring the final version of the report, I decided to begin and end the four research chapters with two different perspectives on chaperonin evolution—the evolution of the chaperonins themselves, and of their clients—but other "thesis-like" arrangements would have been possible, and would have cast the connections between the chapters (and the development of the ideas contained within) in a different light. To provide another perspective on the development of the research in this thesis, I conclude with a chronological ordering of the most relevant events (project beginnings, first meetings with people, and so on) which formed part of it.

The timeline in Figure 6.1 provides an overview of a research program in which different projects proceeded in parallel, and also hints at several projects which died out without becoming fossilized as a thesis chapter. Compiling this list of dates revealed
Fig. 6.1: Thesis timeline. A plot of the key dates associated with the work in this thesis, as inferred from labbook entries and email timestamps. Numbers in brackets indicate the thesis chapter which ultimately resulted from a project. The timeline illustrates the chronological development of the main ideas of the thesis, as well a few extinction events (denoted by asterisks).
both expected and surprising patterns. Despite great industriousness, nothing tangible emerged from the first five months of Ph.D. work. More surprising was the identification of a period of three days (18-20 May, 2008), during which much of the work reported in this thesis was initiated. It is interesting to note that this fertile period was immediately preceded by a visit to Cold Spring Harbor for a chaperone conference which, at the time, was considered something of a mismatch by all concerned. Whether this burst of activity was inspired by, or a reaction to, that event, remains unclear; in any case, it determined much of the direction for the following two years. While the thesis as a whole provides the final word on this body of work, it is hoped that the timeline gives some sense of the structure of the research as it unfolded over the last three years.
Appendix A

Index to Supplementary Material

In order to conserve paper—and also make them far more useful for future analysis—large tables of sequence accession numbers, database annotations, and so on that were used in the thesis are provided as supplementary text files. These files are available on the accompanying CD, or on the web at

http://bioinf.gen.tcd.ie/~faresm/tw_thesis_suppl.zip

The files are presented in a consistent tab-delimited format, with the first line of the file providing column headings. This appendix contains an overview of the available data.
Appendix A. Index to Supplementary Material

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch2_s1.txt</td>
<td>GroEL homologs used to build the ML phylogeny.</td>
</tr>
<tr>
<td>ch2_s2.txt</td>
<td>Group II chaperonin sequences used to build the ML phylogeny of group II genes.</td>
</tr>
<tr>
<td>ch2_s3.txt</td>
<td>mtHsp60 sequences used in the mtHsp60 analysis.</td>
</tr>
<tr>
<td>ch2_s4.txt</td>
<td>GroES sequences (including number per genome) in bacteria with at least one GroEL gene.</td>
</tr>
</tbody>
</table>

**Table A.1:** Chapter 2: Two chaperonin genes in bacterial genomes with distinct ecological roles.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch3_s1.txt</td>
<td>Individual enrichment status of all 750 bacterial proteomes.</td>
</tr>
</tbody>
</table>

**Table A.2:** Chapter 3: Phylogenomic inference of functional divergence with Clusterfunc.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch4_s1.txt</td>
<td>Protein accession numbers for DnaK sequences used to build the protein phylogeny.</td>
</tr>
<tr>
<td>ch4_s2.txt</td>
<td>Nucleotide accession numbers and base ranges of <em>dnaK</em> sequences used in the analysis of selective constraints.</td>
</tr>
</tbody>
</table>

**Table A.3:** Chapter 4: No Rosetta Stone for a sense-antisense origin of amino-acyl tRNA synthetase classes.
<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch5_s1.txt</td>
<td>Genomes used in building the gamma-proteobacterial alignments.</td>
</tr>
<tr>
<td>ch5_s2.txt</td>
<td>PDB chains used in the calculation of residue contact densities.</td>
</tr>
<tr>
<td>ch5_s3.txt</td>
<td><em>Mycoplasma</em> genomes used to obtain homologs of <em>E. coli</em> client proteins.</td>
</tr>
<tr>
<td>ch5_s4.txt</td>
<td>ANCOVA summary (All clients).</td>
</tr>
<tr>
<td>ch5_s5.txt</td>
<td>ANCOVA summary (Class III clients).</td>
</tr>
<tr>
<td>ch5_s6.txt</td>
<td>ANCOVA summary (Class IV clients).</td>
</tr>
<tr>
<td>ch5_s7.txt</td>
<td>ANCOVA summary (Class III- clients).</td>
</tr>
<tr>
<td>ch5_s8.txt</td>
<td>Genomes used in the &quot;one-genome-per-genus&quot; re-analysis of the original data.</td>
</tr>
<tr>
<td>ch5_s9.txt</td>
<td>ANCOVA re-analysis summary (All clients).</td>
</tr>
<tr>
<td>ch5_s10.txt</td>
<td>ANCOVA re-analysis summary (Class III clients).</td>
</tr>
<tr>
<td>ch5_s11.txt</td>
<td>ANCOVA re-analysis summary (Class IV clients).</td>
</tr>
<tr>
<td>ch5_s12.txt</td>
<td>ANCOVA re-analysis summary (Class III- clients).</td>
</tr>
<tr>
<td>ch5_s13.txt</td>
<td>Loss of Class III client homologs from <em>Mycoplasma</em> genomes.</td>
</tr>
</tbody>
</table>

**Table A.4:** Chapter 5: The effect of chaperonin buffering on protein evolution.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES

3693–3706.


REFERENCES


REFERENCES


REFERENCES


REFERENCES


137
REFERENCES


Steno, N. (1669). *Concerning solids naturally contained within solids*.


REFERENCES


