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.

Investigation of the molecular mechanisms of functional innovation

Thesis submitted for the degree of

Doctor of Philosophy

Brian E. Caffrey

Supervisors

Dr. Mario A. Fares

Dr. Karsten Hokamp

Smurfit Institute of Genetics

Trinity College Dublin

2012

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.

Signed,

A person who never made a mistake never tried anything new.

Albert Einstein

Acknowledgements

Mario Fares

Karsten Hokamp

Tom Williams

Xiaowei Xiang

Eiseart Dunne

Neena Thomas

Elaine Kenny

Summary

The robustness to perturbations and evolvability of genomes are two major principles that govern the emergence of genetic diversity across all forms of life. Functional innovations that occur through the genetic diversity cryptically present themselves in the populations. The mechanisms underlying the emergence of biological innovations from this genetic diversity have been a major question in evolutionary biology. Chief among these mechanisms are gene duplication and coadaptation, which are known to facilitate the divergence from ancestral functions generating novel ones. How these mechanisms interact in a complex biological system remains to be understood. In particular, the interplay between these mechanisms, the complexity of the functional interactions and the population parameters of an organism are unresolved mysteries.

Here we explore the mechanisms of innovations through the development and application of novel approaches and computational tools. In particular, we develop a novel method to identify events of functional divergence, not only between paralogues but also among orthologues. We apply this software to a large set of bacterial genomes which includes bacteria with different lifestyles, *i.e.* pathogens, mutualists and bacteria living in extreme environments. We identify significant signatures of functional divergence among the different bacterial clades and find that these signatures are linked to bacterial lifestyles. We show that micro-events of functional divergence take place even between closely related bacterial strains, indicating that evolutionary innovation is a rather more dynamic phenomenon than previously thought.

To understand how innovative mutations are fixed in the genomes, we explore the interaction between mutations through coevolutionary sources. We add to a previous

method that identifies signatures of coevolution at the molecular level, improving the sensitivity and precision of the identification of structural and functional evolutionary dependencies between amino acid sites in a protein. Our improved method outperforms other widely used models in identifying true signatures of coevolution and shows promising results for future identification of protein-protein interactions.

We explore mechanisms that confer mutational robustness, namely the buffering activity of molecular chaperones. The role played by chaperones in evolvability and functional innovation is a topic of contention in molecular evolutionary research. We use both the software developed for functional divergence analysis and coevolution analysis to investigate the role of a bacterial chaperone, GroEL, in the evolution of its client proteins. Previous work demonstrated that GroEL "buffers" the effects of slightly-deleterious mutations in their clients enabling the fixation of destabilizing mutations, of which a fraction may cause functionally innovative mutations. The mechanism by which GroEL buffers the effects of mutations is studied in the context of functional innovation and coevolution between amino acid sites. We identify lineage specific adaptations to ecological conditions through niche relevant genes, which have lead to the emergence of drug resistance and the regulation of toxicity.

To further understand the mutational dynamics under extreme conditions and with the mechanisms of mutational robustness in action, we conducted a mutation-accumulation experiment in the bacterium *E. coli* and analyzed the distribution of single nucleotide polymorphisms (SNPs) in two strains evolving under strong genetic drift or under heat stress. We reveal that accumulations of mutations occur at greater frequency in GroEL clients, that there is a bias towards accumulation of mutations leading to cysteine amino acids, and that compensatory mutations are more likely to occur in interacting proteins under heat-shock. In summary, this thesis contributes to the understanding of evolvability in biological systems and the description of the mechanisms underlying functional innovations.

Contents

Contents

List of Figures

List of Tables

List of Abbreviations

- DNA: Deoxyribonucleic acid
- RNA: Ribonucleic acid
- COG: Clusters of Orthologous Groups
- HGT: Horizontal Gene Transfer
- NCBI: National Center for Biotechnology Information
- OMA: Orthologous MAtrix project
- HAMAP: High-quality Automated and Manual Annotation of Proteins
- SNP: Single Nucleotide Polymorphism
- GTR: General time reversible
- PAML: Phylogenetic Analysis by Maximum Likelihood
- CAFS: Clustering Analysis of Functional Shifts
- CAPS: Coevolutionary Analysis of Protein Sequences
- PDB: Protein Data Bank

Chapter 1

Introduction

1.1 Historical background

1.1.1 Early evolutionary theory

Scientists have strived to understand the relationships between living organisms for hundreds of years. Current evolutionary theory is built upon early pioneering work carried out most famously by Cuvier, Lamarck and Darwin. As early as 1796 Cuvier announced that his comparisons of fossils and living animals did not represent the same species (Cuvier, 1796). This insightful work implied that there had been evolutionary changes between the species at that time and those of fossils. In 1809 Lamarck developed a theory of evolution which proposed that organic life forms were spontaneously generated from inorganic matter and proceeded to develop into more complex organisms. Lamarck observed that evolutionary change is slow and imperceptible, it occurs through adaptation to the environment and that species are related by common decent (Lamarck, 1809).

Charles Darwin rode upon the *Beagle* on an expedition to the Galápagos islands. While there he collected and documented many samples of animals which helped him develop his theory of evolution which he presented in his book *On the origin of species* (Darwin, 1859). Therein Darwin put forward a theory which advanced evo-

lutionary theory tremendously for multiple reasons. Firstly, Darwin argued strongly for *descent with modification*, that species do not arise simulteously but rather are related to each other by common ancestor. Additionally, he wrote that species undergo natural selection; in an environment with limited resources individuals must compete to survive and reproduce. If there are differences in fitness of the individuals (relative to their environment) and their offspring can inherit this fitness then subsequent generations will have a greater proportion of individuals which are more competitive. These findings were extremely profound and far-reaching not least because of the simplicity of the theory proposed but also because of the extensive levels of evidence that Darwin had accumulated.

1.1.2 From early evolutionary theory to modern molecular evolution

Whilst Darwin's theory was extremely insightful, much was still unknown at the time about the science of heritable traits. This was one of the greatest obstacles for Darwin's theory; at the time most scientists believed that offspring were composed of a mixture of parental characteristics (Lamarck, 1809). The outstanding breakthrough which lead to a greater understanding of heritable traits was made by Gregor Mendel. Mendel conducted a multitude of experiments on peas (Mendel, 1866), noting most famously that one in four were purebred recessive, half were hybrid and one in four were purebred dominant. The importance of Mendel's work was not understood fully for many years until it was translated into English in 1901. The early to mid 1900s saw the incorporation of Darwin and Mendel's seminal works with new sciences of cytology and genetics (Fisher, 1930; Haldane, 1941).

The 1960s saw the development of protein and DNA sequencing technologies, this

lead to accelerated development of the theories of evolution and molecular biology. With these developments came the ability to tackle many fundamental hypotheses which had been largely reliant on subjective traits. The resolution of the tree of life, characterization of population dynamics and quantitative analysis of forces which underpin genome evolution are all concepts which had until then been impossible. This sequence data has many varying properties and can be analyzed in many diverse ways to develop a better understanding of life and its evolutionary history. Below I will describe some of these properties and analyses.

1.2 Molecular sequence data

Organisms can be large complicated entities or smaller, more simple entities; nonetheless all organisms are comprised of one or more cells. A cell is the smallest entity to be considered living and contains the genetic information needed for all of the form and function of the organism. DNA is biological polymer; the genetic information which contains the blueprint for all of the attributes of an organism. There are three biological polymers, DNA, RNA and protein. DNA can be considered the information storage device of organisms. This information is inherited by each organism from its ancestors and is the most stable (Stryer *et al.*, 2002) of the three polymers making it ideal for this role. DNA is composed of four nucleotide structures adenine, cytosine, guanine and thymine with double stranded sugar-phosphate back bones. DNA can be "packaged" in different forms, long continuous blocks which are wrapped about proteins (histones) called chromosomes, in eukaryotes, and circularly as in many prokaryotes.

RNA is a single stranded complement to DNA composed of the same nucleotides, replacing thymine with uracil. RNA can be classified into many categories and has

many uses and roles within the cell. One of these, messenger RNA (mRNA) is a chemical template for proteins. mRNA is transcribed as a complement to genes on the DNA and carries this coded information to the ribosomes which in turn create proteins from amino acids. Whilst DNA and RNA generally act as information storers or carriers this is not generally the case with proteins.

DNA sequences contain comparatively small regions called genes which are blocks of functional DNA. Genes can code for a functional RNA, a protein coding region or in some cases for elements involved in gene regulation, such as enhancer regions. Gene structure contains many elements including promotor regions, exons, introns, enhancers and stop sites. Much of this information does not include protein coding data but is necessary for identification of when a protein is needed or should be transcribed.

Proteins are coded for by *codons*, a set of three nucleotides. The three positions in a codon can be filled with any of the four nucleotides, therefore there are 64 combinations possible for codons. 61 of these codons code for an amino acid and are referred to as *sense* codons and the other three code for a stop codon or *nonsense* codon.

DNA and RNA are somewhat restricted through their involvement in information storage. Additionally, the alphabet of nucleotides is comprised of only four different molecules and therefore these polymers do not have the ideal properties for carrying out all functions of an organism. Proteins on the other hand are polypeptides built from the set of 20 amino acids. These amino acids have a vast set of properties; polar, non-polar, aromatic, acidic and basic in addition to a varying degree of physical size and topology. These compounds and their vast array of properties are better suited to carrying out functions within an organism, not least because of the greater

number of combinations of amino acids but also because of the multiple folding combinations they can take. Proteins will provide the majority of data used in the scientific analysis in the following chapters.

1.3 Mutation, selection, drift and populations

1.3.1 Mutations

While DNA can be seen as the blueprint for life and has to be maintained, there are many conditions under which it can change both in the DNA of a single organism but also in terms of the collective DNA of a population. Mutations in the genetic code are the source of all of the diversification across all life forms. Mutations arise through copying errors during cell division/replication, cross-over events during genetic recombination and insertions of extra nucleotides. These mutations are identifiable changes in the DNA that *can* cause a cell or organism to have a different function to the ancestral organism or cell.

1.3.1.1 Point mutations

Point mutations are those which cause a change in the nucleotide sequence of the genome. These point mutations can arise spontaneously due to chemical changes such as altered hydrogen bonds, slipped strand mispairing and deamination.

- Hydrogen bonds pair nucleotides together on opposite sides of the paired backbones of DNA. These hydrogen bonds pair adenine and thymine together with two bonds and guanine and cytosine together with three bonds. If a hydrogen

atom's position is altered in a strand of DNA then a mismatching of nucleotides can occur.

- Slipped strand mispairing can occur when one strand of DNA is temporarily denatured during the replication process. This denaturing can lead to mismatched base pairs. This process can result in the insertion or deletion of portions of DNA.

- Deamination is a chemical process which results in the removal of an amine group from a molecule. These chemical reactions are spontaneous reactions which can result in a change of nucleotide sequence.

1.3.1.2 Chromosomal crossover

Chromosomal crossover refers to the exchange of genetic material between homologous chromosomes. This process occurs when a break forms in the homologous chromosomes resulting in a reconnection of alternative strands.

1.3.2 Allele frequencies and Hardy-Weinberg

Alleles are alternative versions of a gene in the same location in the genome. In a population of organisms there may be varying frequencies of these alleles. These frequencies are not stationary and can be affected by fitness and hence natural selection, genetic drift, the rate of mutation, migration and recombination.

Given a population of size N and two alleles x_1 and x_2. Allele x_1 has frequency p/N and allele x_2 has frequency q/N where

$$\frac{p+q}{N} = 1 \tag{1.1}$$

and

$$p + q = N \tag{1.2}$$

Then if given a diploid genotype there are three combinations of alleles x_1x_1, x_2x_2 and x_1x_2 (since $x_1x_2 = x_2x_1$). If the mating process in the population is random then the respective frequencies will be $p^2, 2pq$ and q^2. Populations which maintain these frequencies are said to be in Hardy-Weinberg equilibrium.

1.3.3 Selection

Natural Selection is defined as the differential production of genetically distinct individuals or genotypes within a population. Differential reproduction is caused by differences among individuals according to factors, such as mortality, fertility, mating success and the viability of offspring. These factors are said to confer fitness upon the organism. This fitness is a measure of how adept an organism is to its environment. The underlying attributes which can be associated with differences in fitness are alternate copies of a gene. These copies arise due to mutations. In diploid organisms it is necessary to consider the relative affect of the two copies of a gene within the organisms. These copies at the same position of a chromosome are called alleles (as mentioned in section 1.3.2).

When considering selection in terms of allele frequencies we assign a fitness w_i to each of the alleles. The allele frequencies then become $p^2w_{11}, 2pqw_{22}$ and q^2w_{12}. There are two models of dominance to consider, codominance and overdominance, codominance is the state when the heterozygote has a fitness that is the mean of the two homozygotes. Overdominance is when the heterozygote has the highest fitness.

1.3.4 Drift

Changes in allele frequencies attributed solely to randomness is called *random genetic drift.* This genetic drift is due to the randomness of nature. An organism may by chance encounter a rich food source allowing it time to pass on its genetic information. Equally likely is the case of an organism encountering a harsh environment which prevents the passing of its genetic material.

For Example, in a diploid population of N individuals reproducing to create a new generation there are 2N alleles to pick from. Some of the N individuals will be successful in reproducing, these are members of N_e, the effective population, a subset of N. Due to random breading (through all factors, such as: food source, random encounters and harsh environments) there may be unequal frequencies of two alternate alleles in N_e leading to unequal frequencies of alleles in the next generation. One can therefore envisage a situation in which this happens over multiple generations until there are no more variants of one particular allele. This implies that one allele has *drifted* to *fixation.*

1.3.5 Neutral and nearly neutral theory

The theory of neutral evolution was put forward by Kimura in 1968 (Kimura 1968). Kimura postulated that most mutations have no effect on fitness but have simply drifted to fixation. A surprising and far-reaching conclusion made by the neutral theory is that the rate of evolution is equal to the mutation rate. Given a population of size N, the number of new mutations in a genome is equal to the mutation rate μ multiplied by the number of alleles in the population, in this case, 2N.

$$R_{newmutations} = 2N\mu \tag{1.3}$$

Furthermore the rate of evolution must be given by the number of new mutations multiplied by the probability of fixation:

$$R_{evolution} = R_{newmutations}.P_{fixation} \tag{1.4}$$

We discussed in the previous section that if a mutation is neutral then it has a probability of fixation $P_{fixation} = \frac{1}{2N}$. Therefore:

$$R_{evolution} = 2N\mu.\frac{1}{2N} = \mu \tag{1.5}$$

A corollary of the neutral theory pertains to the extremely conserved nature of functionally important sites within proteins. If a site is functional, or moreover has a *very* important function then any mutation that would cause an effect (non-synonymous) to the resulting amino acid would not be neutral and would therefore not drift to fixation in this manner.

This result, that the rate of evolution relies solely on the mutation rate is very profound and elegant. The majority of mutations occur during replication, therefore the greater the number of replications the greater the number of mutations that can be introduced to a genome and hence have a chance to drift to fixation. This implies that organisms with short generation times should evolve more quickly than organisms with long generation times. This assumption holds for noncoding DNA but does not hold for protein-coding DNA(Ohta, 1972a). For protein coding DNA the generation time has little to no effect on the rate of evolution(Ohta, 1972b). This suggests that the theory is not quite satisfactory for protein-coding sequence evolution.

Until this point in time it was understood that most mutations are neutral to

the fitness of an organism. This turned out not to be the case and was illustrated when Ohta realized that generation time is inversely proportional to population size. Since population size increases the effect of selection (both positive and purifying), the lack of effect of generation time on the rate of evolution could be explained if the level of purifying selection is high. This information identified the fact that that the majority of non-synonymous mutations are deleterious. The Kimura model of "neutral" evolution was thus extended to the "nearly neutral theory" of evolution.

1.3.6 Selection vs drift

Given the enormous amount of time that life has existed there have been numerous changes in the alleles and allele frequencies of all organisms to this date. If a mutation arises in a population it is probabilistically likely that it is deleterious. This new allele may take numerous courses: it may be selected out by purifying selection or it may randomly drift to either fixation or extinction. In the small number of cases where a mutation is advantageous it may be positively selected and become fixed in the population if the selection pressure upon that site is strong enough. This mutation which causes an increase in fitness may of course also drift to fixation or drift to extinction depending on the relative effects of selection pressures in the population.

When accounting for selection the probability for fixation of an allele is given by

$$P = \frac{1 - e^{-4N_e sq}}{1 - e^{-4N_e s}} \tag{1.6}$$

With q, the initial frequency, N_e the effective population size and s the selection coefficient. The selection coefficient is a measure of relative fitness of a mutation. A mutation which is completely neutral has a selection coefficient of zero, mutations

which are advantageous are positive and deleterious mutations are negative. The implications of equation 1.6 are very profound and merit further discussion, the following sections discuss the both neutral mutations and mutations causing a change in fitness.

Neutral mutations:

In the case of a neutral mutation the selection coefficient $s \to 0$. Taking the mathematical property:

$$\lim_{x \to 0} e^{-x} \approx 1 - x \tag{1.7}$$

$$\lim_{s \to 0} P \approx q \tag{1.8}$$

$\Rightarrow$ the probability of fixation of a neutral mutation is equal to the initial frequency of the allele, then in a diploid population, N, we have 2N alleles making the probability of fixation of a new mutation:

$$P_N = \frac{1}{2N} \tag{1.9}$$

Therefore the larger the population the lower the probability that a new mutation will become fixed.

Advantageous and deleterious mutations:

For simplicity we will set $N_e = N$, in this case. We substitute $q = 1/2N$ and

when $|s|$ is small equation 1.6 becomes

$$P \approx \frac{2s}{1 - e^{-4Ns}} \tag{1.10}$$

Now when N is large this becomes

$$P \approx 2s \tag{1.11}$$

Therefore when the selection advantage is less than 5% and the population is large then the chance of fixation is roughly twice the selection advantage. As an example, when $s = 0.02$ then $P \approx 4\%$.

1.3.7 Measuring selection in multiple sequence alignments

Under the assumption that mutations occur uniformly across sequences, then, according to the neutral theory of evolution there should be a homogeneous distribution of synonymous mutations across the sequence. This is the case because synonymous mutations do not change the resulting amino acid at the protein level. Conversely, amino acids that are non-synonymous **do** result in a change in amino acid composition and have the possibility of being neutral, advantageous or deleterious. Furthermore we can quantify the selection pressure by analyzing the ratio of the average number of non-synonymous to the average number of synonymous sites (Kimura 1983):

$$\omega = \frac{d_N}{d_S} \tag{1.12}$$

According to equation 1.12 there are three possible outcomes:

- $\omega = 1$ indicates that the gene is under neutral selection since sites which cause a change in amino acid are just as likely as those that do not.
- $\omega > 1$ indicates that the gene is under positive selection since amino acid replacements are fixed more often than neutral mutations.
- $\omega < 1$ indicates that the gene is under purifying selection since amino acid replacements are discarded more often than neutral mutations.

This test is powerful and many groups have developed methods aimed at identifying positive selection (Perler *et al.*, 1980; Miyata and Yasunaga, 1980; Nei and Gojobori, 1986; Hughes and Nei, 1988; Yang and Nielsen, 2000). There are however numerous shortcoming and precautions which should be considered when performing these analyses. Firstly, the estimation of synonymous sites can easily be underestimated if the divergence times between two sequences is large enough. This is referred to as saturation and can be most easily understood by realizing there is an upper bound on the number of synonymous substitutions in any finite sequence. Secondly, when comparing two sequences the observer cannot distinguish which sequence in the pair is undergoing selection; it is possible that both sequences are under selection. Thirdly, the requirement that $\omega > 1$ over the full length of a sequence is conservative in most cases as positive selection can occur on a small subset of amino acids within a sequence.

1.3.8 Models of evolution

The majority of evolutionary models are based on Markov chains. A Markov chain has the property that given a state X at time t, the only state that it depends on is that of state X at time $t-1$. This implies that there is no memory to the evolutionary

model. For biological data we can represent the probability of a substitution as a transition matrix, P, with elements p_{ij} given by:

$$p_{ij} = Pr(x^{t+1} = j | x^t = i) \tag{1.13}$$

where x^t is the molecular state at time t. The matrix P is given by a 4 x 4 matrix in the case of DNA models and a 20 x 20 matrix in the case of amino acid models.

1.3.8.1 DNA models

The first model of evolution was the Jukes and Cantor (JC)(Jukes and Cantor, 1969) model. This is a DNA model with no parameters and assumed that all possible mutations are equally probable. Additionally, the frequencies of the bases are assumed to be equal. This model did not fit with biological data and was amended to account for differing rates in substitution between transitions and transversions (Kimura, 1980). Felsenstein developed a model which accounted for unequal base frequencies (Felsenstein, 1981) and this in turn was joined with the Kimura80 model by Hasegawa to make the HKY model (Hasegawa *et al.*, 1985). The most flexible and currently most popular DNA model is the generalized time reversible model (GTR)(Tavare, 1986). The GTR model has ten parameters, six of those pertain to the probability of substitution from one nucleotide to another and four stationary frequencies of base composition.

1.3.8.2 Amino acid models

When analysing protein sequences it is difficult to do so in the same manner as is carried out in DNA sequences. For example, there are ten (sometimes nine)

parameters in the GTR model; to implement this model in an amino acid framework there would be 190 parameters to estimate. This huge number of parameters makes optimization extremely difficult in terms of computation time. As an alternative empirical protein evolution models were developed. The first amino acid model was developed by Dayhoff and relies on the calculation of transition matrices from large training datasets (Dayhoff *et al.*, 1972). The JTT model (Jones *et al.*, 1992) was calculated using an approximate peptide-based sequence comparison algorithm on a set of sequences clustered at the 85% identity level. This represented a major update on the original PAM matrices calculated by Dayhoff *et al.*, since thirteen years had passed since the last updated PAM matrix. A multitude of matrices have been developed since including BLOSUM (Henikoff and Henikoff, 1992) and WAG (Whelan and Goldman, 2001), the latter being more appropriate for globular proteins.

1.4 Changes in selection pressure

1.4.1 Adaptation and diversification through environmental change

According to Darwin's evolutionary theory individuals within a population undergo natural selection, leading to subsequent generations containing a greater proportion of individuals with a phenotype of higher fitness. Under this theory the scenario continues with each generation slightly fitter than the last, resulting in a population which is highly specific to its environment. In this case, the emergence of a mutation, in an individual which has evolved to become highly specific to its environment is highly likely to be deleterious. There is little scope for change in the organism

as there is little need for change. If, however, there is a sudden change in the environment then this can have extremely detrimental effects upon the population. Furthermore if an environment becomes uninhabitable or the abundance of food is too small to sustain the population then migration may occur.

In an environment where some selection pressure(s) have changed, many populations are decimated by: lack of food, harsh conditions (e.g. lack of oxygen) and even differential advantages for predators. Those individuals which survive are those that have traits which give them a fitness advantage over others. This new population again passes on these heritable traits as before, however these traits differ from those which conferred high fitness under the previous selection pressures. When a trait is no longer beneficial, there are relaxed pressures/constraints upon the region(s) of the genome which contained them. These genes can develop new function or be removed from the genome. An example of the loss of genes through relaxed constraint is that of endosymbiotic bacteria. These bacteria have lost a a large number of genes which are necessary for free-living bacteria to obtain nutrients from the environment (Moran, 2002; McCutcheon & Moran, 2007).

1.4.2 Gene duplications: a precursor to genetic variability

Gene duplications are those events which result in the duplication of a section of DNA containing a gene. Gene duplications can occur as small scale events in the case of unequal crossover events or large scale duplications as in the case of polyploidy. During meiosis complementary chromosomes undergo recombination. Recombination involves the breaking and rejoining of the DNA in both complementary chromosomes. This process shuffles the genes in the resulting cell. When this process is not carried out perfectly it can result in a deletion of a gene on one chromosome and

a duplicated gene on the other chromosome. Polyploidy events are drastic events result in a duplication which alters the number of chromosomes in the cell. Polyploidy events occur regularly in the evolution of plants but are rare in eukaryotes. Polyploidy events occur as a result of errors during cell division.

The long term effects of gene duplication mean that there are relaxed selection pressures upon one or both of the duplicates. This relaxed pressure leads to diversification of genes, however, short term effects can be extreme. First and foremost gene duplication in most cases leads to an increase in expression (Schuster-Bockler *et al.*, 2010). This increase may be a small effect in many cases but there have been studies demonstrating that duplicated genes are selected against due to increased gene expression effects (Zhang *et al.*, 2009). The relative redundancy of one (or more) of these copies in addition to negative selective pressures increase the likelihood of mutations with duplicates. A product of this is the theory that it is easier to produce a novel function from an existing one rather than a *de novo* production of a gene, not least from a purely statistical point of view. Whole genome duplication events are thought to be much more detrimental to the cell and have been observed to be followed by extensive gene loss (Scannel *et al.*, 2006).

1.4.3 Adaptation and diversification through functional divergence

Central to Darwin's and indeed most evolutionary theories is the common ancestry of all diverse organisms. Taking together, the knowledge of common ancestry and the diversity of current organisms there is an implied necessity for functional divergence of genes whilst simultaneously a necessity for homology. Such divergences in function can follow gene duplication or can arise through environmental pressures due to

an opening of a new ecological niche. Gene duplication can lead to the relaxed selective constraint on the genes involved (Stephens, 1951; Nei, 1969; Ohno, 1970; Lynch & Conery, 2000). Two or more exact copies of a gene can be at best neutral to an organism's fitness at origin (Conant & Wolfe, 2008; Lynch & Conery, 2000), therefore the retention of multiple copies of a gene until such a time as new functions can emerge through mutations must itself be a mechanism of evolution.

The emergence of new functions is thought to arise through two processes, subfunctionalization and neofunctionalization (Force *et al.*, 1999). In the process of subfunctionalization both copies of a duplicated gene accumulate mutations in a complementary fashion. Both genes maintain a subset of the functions carried out by the pre-duplicated gene, however, between these two new genes all of the original functions remain intact. In the process of neofunctionalization one of the two duplicated genes accumulates mutations in functional areas which leads to a gene which carries out a different role than that of the original gene. Indeed it is understood that subfunctionalization and neofunctionalization need not be mutually exclusive mechanisms of evolution (He and Zhang, 2005; Rastogi and Liberles, 2005; Conant & Wolfe, 2008). The third possibility when a gene is duplicated is that one becomes a pseudogene. One of the copies of the gene maintains all the function of the ancestral gene and the second is free to accumulate mutations; it does so until such a time as it can be classified as a pseudogene. Figure 1.1 summarizes the emergence of new functions following gene duplication. I will discuss the development of a novel method for analysis of functional divergence in Chapter 2 and, additionally, I will present a case study of proteome-wide functional divergence in chapter 3.

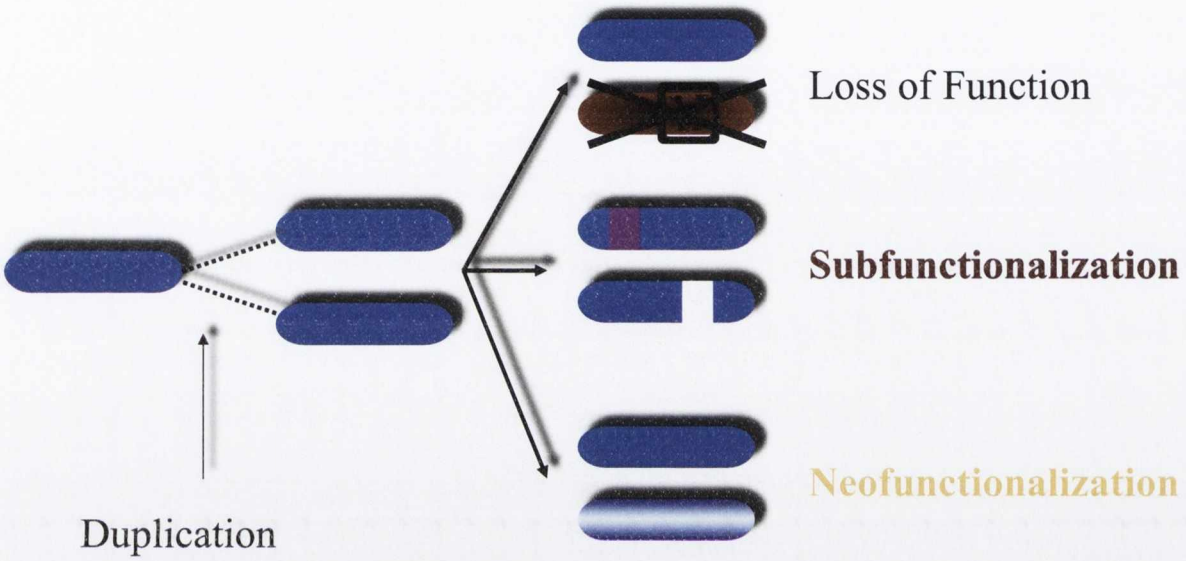

Figure 1.1: Evolution by gene duplication. This Figure shows the processes by which a duplicated gene can lead to a pseudogene, subfunctionalization or neofunctionalization. After a gene is duplicated there is redundancy. Due to this redundancy there is reduced selection pressures upon one of the genes. In the first scenario the gene accumulates mutations including insertions and deletions until it becomes a pseudogene or possibly not even recognizable as a gene. The second scenario sees one of the duplicates maintain some of the ancestral functions and the other duplicate retains the rest of the functions; this is called subfunctionalization. The third scenario sees one of the duplicates retain all of the function(s) and the second copy accumulates mutations until it has obtained new function, resulting in neofunctionalization.

1.4.4 Molecular chaperones

Molecular chaperones are important molecules involved in the proper folding and regulation of newly synthesized proteins. Many proteins can fold correctly independent of molecular chaperones, some are however susceptible to misfolding and aggregation. Aggregates can be toxic and are known to be involved in disease as in the cases of amyloidoses and prion disease. When a protein is released from the ribosome and needs "help" to fold, it is delivered by Hsp40 to Hsp70 and trigger factor. Hsp70 binds to hydrophobic regions of the polypeptide to prevent aggregation (Hartl & Hayer-Hartl, 2009). Hsp70 has two ATP-binding domains which are allosterically coupled (Mayer *et al.*, 2000). The N-terminal domain is activated by ATP hydrolysis when Hsp40 delivers a protein. This increases the substrate binding affinity of the C-terminal domain (Young *et al.*, 2003). Hsp70 releases the peptide when an exchange factor substitutes the ATP with an ADP molecule again lowering the binding affinity of the C-terminal domain. This process is thought to protect the hydrophobic regions from aggregation due to interactions with other newly synthesized proteins until partial folding has occurred; on release the protein folds into place, burying the hydrophobic regions resulting in a correctly folded protein.

Proteins which are not correctly folded by Hsp70 are delivered to Hsp90 (a molecular dimer) or GroEL/Hsp60 (a molecular multimer). The method by which Hsp90 chaperones proteins is analogous to that of Hsp70. Hsp90 is bound by ATP with the assistance of co-chaperone proteins. Hsp90 undergoes many conformational changes due to ATP binding and hydrolysis, these changes impose conformational changes upon the client protein until it folded correctly.

GroEL is a 14-mer composed of 60kDa monomers (shown in Figure 1.2), it is located on an operon (groE) alongside GroES which is also involved in the chaperone

process. The GroEL complex consists of two stacked heptameric rings joined back to back in the centre. Each of the rings has a central chamber of 85,000 cubic Angstroms (Lund, 2009). Protein folding with the assistance of GroEL occurs in a controlled allosteric manner. Hydrophobic regions of unfolded proteins bind to a hydrophobic region of the apical domain of GroEL, close to the entrance to the folding chamber in one of the rings. GroES and ATP bind to this ring causing conformational changes which release the unfolded protein into the folding chamber. The protein now folds within the chamber, possibly making use of the fact that it is completely protected from destablizing interactions with other proteins (Lund, 2009). After roughly 10 seconds (Burston *et al.*, 1995) ATP hydrolysis is complete which causes conformational changes that weaken the interaction between GroES and GroEL and also prepare the other GroEL ring for another cycle of protein folding.

1.4.5 Adaptation and diversification through chaperone buffering

1.4.5.1 Evolvability

In the broad and diverse field that is molecular evolution there have been some challenging features which have arisen that do not (as yet) easily fit with modern evolutionary theory. One such example is evolvability, first proposed by Wright in 1932. Wright noticed that a combination of several mutations could lead to an increase in fitness where any of those mutations alone would cause a fitness *decrease*. These occurrences led to the visualization of sequence space as a series of peaks and troughs or "adaptive valleys" which can be explored through mutations. In this analogy one mutation brings the fitness of the organism down from a fitness "peak"

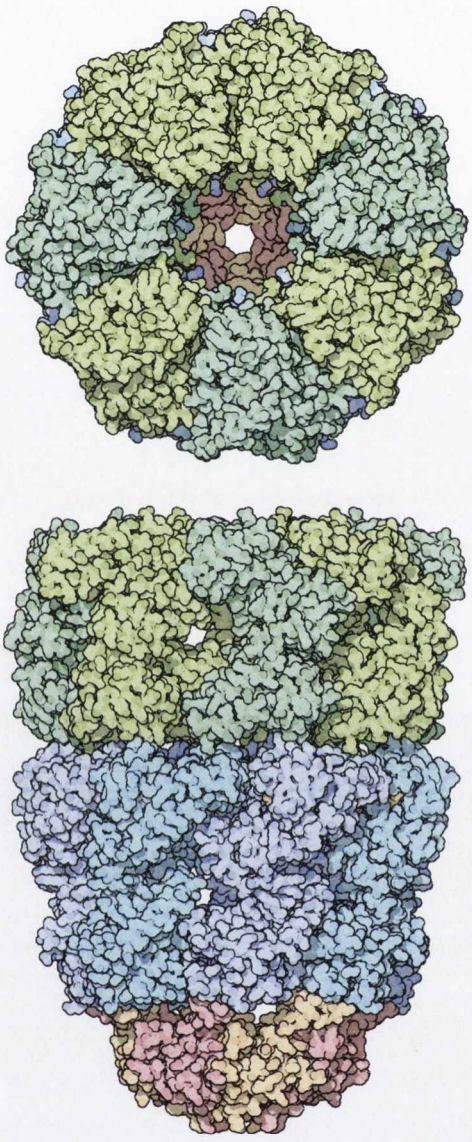

Figure 1.2: GroEL structure. Pictured is the protein structure of GroEL in complex with its co-chaperone GroES. The green section represents one heptameric ring of GroEL proteins and the blue/indigo section represents the second ring. The lower ring is in complex with GroES which acts as a "lid" sealing off the inner chamber to protect the client protein whilst it folds. In the image the lower ring is elongated due to conformational changes brought about by the binding of GroES. This elongation is both to release the client protein from binding the inner wall and also to create extra space for folding to occur. **Note:** This imagine was created by David Goodsell as part of the www.pdb.org molecule of the month series.

into a fitness "valley". A second mutation brings the fitness back up a "peak" which is higher than that of the original fitness.

There are many examples of evolvability; horizontal gene transfer (HGT), upregulated mutation rates and sexual recombination. Horizontal gene transfer is experienced by bacterial strains; during nutrient deficient times, bacteria activate DNA uptake mechanisms to obtain useful DNA from its surroundings (Poole *et al.*, 2003). Upregulated mutation rates occur in some bacteria during starvation; the mutation rate is increased in a controlled way in an attempt to accumulate mutations causing adaptations beneficial to the current surroundings (Galhardo *et al.*, 2007). Sexual recombination *can* result in the accumulation of mutations at a greater rate than that expected according to the mutation rate alone.

1.4.5.2 Robustness

Robustness is the ability of an organism to accumulate deleterious mutations without serious detrimental effects to fitness. As discussed most mutations are deleterious according to the nearly neutral theory of evolution, however many mutations arise in the population, persist and can even drift to fixation. If mutations continue to arise then it is expected that the fitness will slowly decline. At first glance it seems as though robustness is a property which acts against evolvability as it reduces or smoothes down the fitness landscape upon which selection can act. Wagner proposed that there are molecular "capacitors" which facilitate the accumulation of deleterious mutations by "buffering" detrimental fitness effects caused by the accumulation of deleterious mutations (Wagner, 2008). In addition to molecular capacitors it has been proposed that those organisms which are most robust to mutations can do so because as an organism, mutations have little impact upon the change of their fitness

landscape. This has been proposed under the model of "Survival of the flattest", for which evidence has been found (Codoner *et. al.*; 2006).

In 1998 Rutherford and Linquist proposed Hsp90 as a capacitor, they had discovered that if Hsp90 was impaired then many morphological changes can develop which would normally be suppressed due to Hsp90s activity in mediating protein folding. Morphological changes were also discovered in *Arabidopsis thaliana.*(Queitsch *et al.*, 2002). Another molecule which was proposed as a capacitor was GroEL/GroES. It was shown that endosymbiotic bacteria over-expressed GroEL (Moran, 1996). It was proposed that this was to cope with the accumulation of mutations experienced by large levels of genetic drift. Fares *et al..*(2002) demonstrated that *E. coli* exposed to evolutionary bottlenecks (and therefore greater genetic drift) could partially recover fitness with over-expression of GroEL/GroES.

This extensive research into evolvability and robustness raises some difficult questions. Firstly, can robustness lead to adaptation and secondly, whether the phenotypic effects of buffering by capacitors is selected for or is it merely a lucky consequence which has been born out of the useful features of protein folding mediators, such as Hsp90 and GroEL.

1.5 Coevolution

The theory of Coevolution was proposed under the red queen hypothesis (Van Valen, 1973), the analogy of alice in Lewis Carrol's *"Through the looking-glass"* pertained to a race in which alice (the main character) continually ran but always stayed in the same position. This analogy lends well to that of constant battle in the animal kingdom between a predator and its prey. This theory implies that an adaption in

one organism which leads to an increased fitness has a negative impact on another species in the same ecological niche. Subsequently there is a selection pressure on the weakened organism which is intrinsically linked to the increase in fitness of the other species. In the case of the predator and prey, if the predator adapts such that it is fitter athletically then it will have an adverse affect upon the likelihood that the prey can pass on its genetic information; because it may be eaten by the predator. This scenario implies that as a result of selection, organisms must continually change in order to survive. When considering molecules present within organisms there is an analogous argument to be made. These molecules are under a similar set of selection pressures, a change in one molecule which leads to a fitness change can have a detrimental effect on another molecule. This leads to a selection pressure on the second molecule to adapt in order to maintain its function. Since organic molecules are involved in numerous interactions within a molecular network, it is evident that there are coevolutionary signals within this network which act reciprocally to conserve the connections/fitness of the organism (Tillier and Charlebois, 2009) .

A common assumption in current molecular evolution is that functionally important residues are considered to be evolving more slowly than non-functional sites. This prediction is a direct result of Kimura's theory of neutral evolution(Kimura, 1983) and Ohta's nearly-neutral theory(Ohta, 1972; Ohta 1973). This assumption and the work based upon it has proved invaluable to both bioinformaticists and experimental scientists alike. The assumption allows for extremely accurate prediction of homologous genes, phylogenetic inference and a host of other statistical methods. As a result, it also allows for far more direct experimental procedures, such as site-based mutagenesis to be performed to discover the molecular function

of amino acids in a protein.

In contrast to the usefulness of this assumption and subsequent approaches there is a group of residues which are ignored, non-conserved functional sites, *i.e.* coevolving sites. Coevolving sites are those which are observed to have mutated from the original residue but the function has been saved by a corresponding mutation at another site in the protein *or* sites which have mutated causing an increase in the fitness of the organism and a corresponding mutation boosts this fitness increase. In the antagonistic case the initial mutation can be considered deleterious, however may not have been removed from the population by selection as the strength of selection relies on population size (Ohta, 1973). If there is enough time for a corresponding mutation which restores function to the region then this pair is said to be coevolving and can be fixed in the population.

Protein interactions define molecular and cellular function. These interactions are rarely distinct events but are connected in complicated networks which are of huge importance to modern research. Experimental approaches to identify such interactions include high throughput screenings using affinity purification (Gavin *et al.*, 2006), yeast two-hybrid systems (Uetz *et al.*, 2000; Ito *et al.*, 2001) and protein chips (Zhu *et al.*, 2001). These approaches have obvious merits by directly identifying and observing interactions. They do, unfortunately have shortcomings also. First and foremost levels of false positives have been shown to be very high (Sowa *et al.*.). Coupling this with the obvious financial costs associated it is evident that robust computational models for predicting interactions is a not only a worthy research area but also one which can give significant insight into protein-protein interactions and evolution as a whole.

In a review (Lovell and Robertson, 2010) Lovell and Robertson adapt an early

definition of coevolution (Thompson, 1994) to define coevolution as "the reciprocal evolutionary change in evolutionarily interacting loci". This definition encompasses both intra-molecular coevolution and inter-molecular coevolution. Intra-molecular coevolution pertains to correlated evolutionary changes within a gene, these correlated changes may be to preserve function or alternatively lend stability to the complex folds involved in protein structure. Inter-molecular coevolution refers to correlated evolutionary changes between two genes. It is interpreted as coevolution which maintains interactions due to collective selection pressures on all loci involved either directly or indirectly in an interaction or group of interactions. Many studies have identified site-specific intra-molecular and inter-molecular coevolution from historically well understood proteins like myoglobin (Pollock *et al.*, 1999) to proteins with function of a more illusive nature in HIV-envolope proteins (Travers *et al.*, 2007).

1.6 Thesis structure and aims

This thesis presents an evolutionary analysis of the molecular mechanisms of functional innovation through functional divergence, coevolution and the buffering effects of the chaperone GroEL. It is presented in 7 chapters including this introduction and a final conclusions chapter. Above I have discussed some of the extensive research which has been performed in the area of molecular evolution. More specifically in areas which pertain to this thesis, namely functional divergence, coevolution and the buffering of deleterious mutations by chaperones. This thesis aims to not only investigate some of the remaining unanswered questions in this field but also to develop novel methods of analysis that are far-reaching in effectiveness but also faster and more user friendly than existing methods.

Techniques for the analysis of changes in selective constraints at both the DNA (Berglund *et al.*, 2005; Fares *et al.*, 2002; Goldman & Yang, 1994; Nielsen & Yang, 1998; Suzuki, 2004a; Suzuki, 2004b; Suzuki & Gojobori, 1999; Yang & Bielawski, 2000; Yang & Nielsen, 2002; Zhang, 2004; Zhang *et al.*, 2005) and the protein level, Gu (Gu, 1999; Gu, 2001; Gu, 2006) have been developed. The most widely used of these is Gu's method DIVERGE. In chapter two we present a novel method for the detection of site specific functional divergence; therein we compare and contrast our method relative to that of DIVERGE. In chapter 3 we apply our novel method to 750 bacterial proteomes to identify high-level patterns of functional divergence and link these patterns to ecological adaptations.

Whilst the theory of molecular coevolution is not a new topic of research it remains poorly understood. There have been many experimental approaches which have demonstrated its occurrence (discussed above) and many methods developed to detect coevolution in multiple sequence alignments. In chapter 4 we present an updated model and software for the analysis of both intra-molecular coevolution and inter-molecular coevolution. We administer site-specific bootstrapping in order to separate phylogenetic coevolution and functional coevolution.

The effect of chaperone buffering has been studied in some pioneering works but still raises awkward and as yet unanswered questions, such as can buffering of mutations by GroEL cause lineage specific adaptations to ecological niches? In chapter 5 we attempt to address this question with analyses of strains of *E. coli* subjected to evolutionary bottlenecking. In chapter 6 we present an analysis of GroEL and its clients which investigates the effect of buffering on molecular coevolution and functional divergence in a mutation accumulation experiment.

Chapter 2

CAFS: Clustering Analysis of Functional Shifts

2.1 Related poster accepted for publication.

Williams T.A., Caffrey, B.E., Jiang X., Toft, C., and M. A. Fares. (2009). Phylogenomic inference of functional divergence. BMC Bioinformatics 10(Suppl. 13): P4.

This chapter describes the development and expansion of the method for detecting functional divergence first presented by Toft *et al.*. (2009). This chapter represents collaborative work between B.E. Caffrey, T. A. Williams, X. Jiang and C. Toft. BEC and TAW were responsible for expanding the model to incorporate a more accurate test statistic, greater assessment of significance through simulated alignments and functional divergence enrichment analysis. BEC wrote all C++ optimized code which allowed for whole proteome analysis. BEC performed all benchmarking, testing and comparisons contained within this chapter. XJ and CT added visualizations of results.

2.2 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 (Conant & Wolfe, 2008; Lynch & Conery, 2000). 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 (Schluter 2000). 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 (Conant & Wolfe, 2008; Innan & Kondrashov, 2010; Lynch & Conery, 2000; Lynch & Katju, 2004; Ohno, 1970). 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).

Although it is known 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, whole-genome analyses must take into account widespread HGT, 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. The rationale for previous methods, and indeed the new approach 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. These considerations have motivated the development of a number of methods for identifying changes in selective constraints on protein-coding genes and on single amino acid sites and lineages in a phylogenetic tree (Berglund *et al.*, 2005; Fares *et al.*, 2002; Goldman & Yang, 1994; Nielsen & Yang, 1998; Suzuki, 2004a; Suzuki, 2004b; Suzuki & Gojobori, 1999; Yang & Bielawski, 2000; Yang & Nielsen, 2002; Zhang, 2004; Zhang *et al.*, 2005). At the protein level, Gu (Gu, 1999; Gu, 2001; Gu, 2006) developed a Bayesian approach to identify functional divergence, which has become the most widely used. Comparisons of amino acid 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.

Recently, we have 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 (Toft *et al.*, 2009; Williams *et al.*, 2010). Similar to other methods, our approach was limited to tests of one gene at a time, unless the phylogeny of all genes could be fixed in advance. In the present study, we have optimised 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 HGT, (iii) provide an intuitive probability assignment for each test which takes the underlying phylogeny of the sequences into account and (iv) explore all levels of each gene tree, testing for functional divergence at each node. This method and analysis was co-developed with Thomas Williams.

2.3 Materials and Methods

Our analysis of functional divergence, the individual steps of which are detailed below, is summarized in Figure 2.1.

2.3.1 Sequence Input

The input for the software should be a single multiple sequence alignment in Fasta format. Optionally a folder of multiple sequence alignments can be provided as input, again in Fasta format. In addition to this, if a phylogenetic tree calculated using an alternative method to that implemented in the software is desirable then a Newick formatted tree may be used as input.

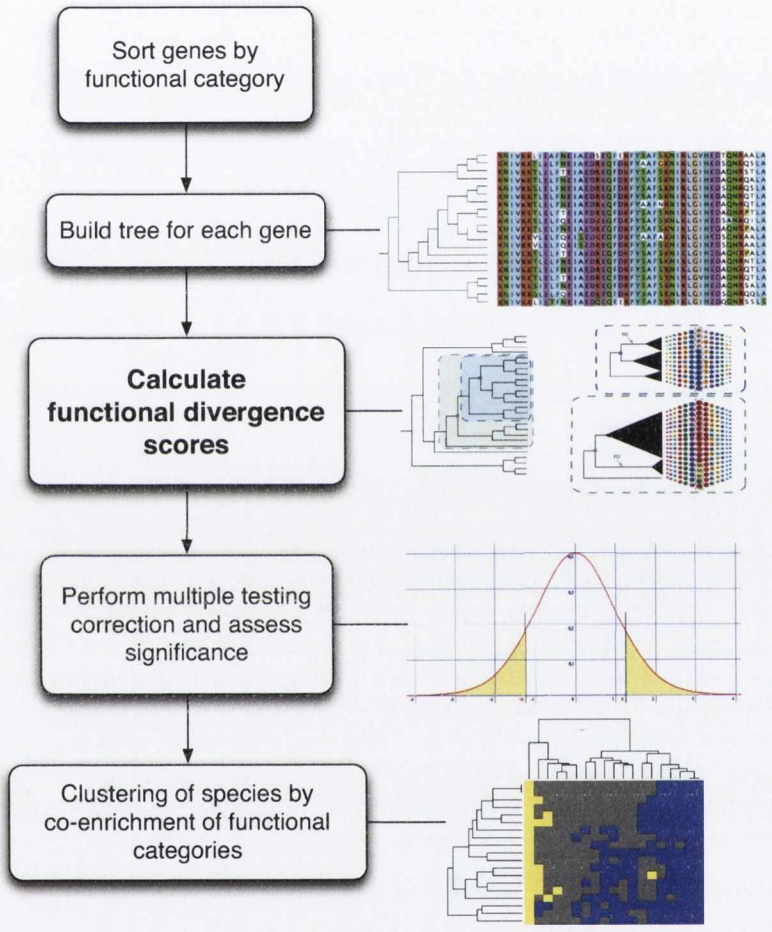

Figure 2.1: CAFS workflow. Above shows a workflow of the CAFS Analysis, in the following sections there is further documentation of each step.

2.3.2 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 calculate a tree for each gene (set of homologous sequences) in the dataset using BIONJ (Gascuel, 1997), under the JTT model of protein sequence evolution (Jones *et al.*, 1992). Calculations for that gene are then made exclusively using the resulting tree. Additional future development will include the option to choose a alternative model to be used for tree construction.

The CAFS software has the ability to use a pre-calculated tree made using likelihood or bayesian methods. For large analyses, such as those in chapter 3. this may not be possible due to the excessive computational time required. It is suggested where possible that rigorous trees be used.

2.3.3 Scoring functional divergence

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), indeed any substitution matrix can be used. Scores for each column are given by:

$$FD_{score} = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 - X_2}} \tag{2.1}$$

where $\bar{X}$ 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 $S_{X_1} - S_{X_2}$, the standard error for unequal sample sizes with unequal variances is given by:

$$S_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2.2}$$

Where s_i^2 refers to the variance in sample i and n_i refers to the size of sample i.

2.3.4 Significance testing

For each inner node tested, a simulated sequence alignment is automatically created using the JTT model (Jones *et al.*, 1992) according to the gene-specific phylogenetic tree calculated above. That is, the distance matrix for the simulated data is equal to that of the real data, but sequences were evolved under a model that represents molecular evolution without the specific selection constraints which are upon each gene tested, as all amino acid transitions are equally likely. The simulated alignment is tested according to equation (1), resulting in a null distribution of the test score against which P-values for the real data were calculated. These values are then corrected for multiple testing by the False Discovery Rate method (Benjamini & Hochberg, 1995) using an alpha value of the users choice as the threshold of significance. Following this procedure, branches on the tree that still possess at least one significant site are considered to be under functional divergence for the purposes of enrichment and clustering. These simulated alignments are created with the use of the Bio++ libraries and are hard coded into the software.

2.3.5 Enrichment analysis

Once all alignments were analyzed, we performed three different enrichment tests to ask three different biological questions. These are based on a chi-squared test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.3}$$

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 use the enrichment tests to identify (i) species and (ii) categories of genes that experience 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 calculate (iii) the enrichment status of each category within each species, in order to identify lineage-specific shifts in the pattern of functional divergence. The expected values in each of these cases refers to the values obtained when averaging (i) over all species and (ii) over all functional categories and (iii) all species in each category.

2.3.6 Hierarchical clustering

We create 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 (R Development Core Team, 2010). 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 particular functional categories or convergent

functional divergence among phylogenetically unrelated sequences. We take this chance to remind the user that each cell of this heat map is the outcome of a statistical test declaring whether a category and species are enriched or impoverished for functional divergence.

2.3.7 Implementation

CAFS 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 by Williams *et al.*. (2010)) and the kind of multi-proteome investigation reported in chapters 3 and 5. The latest version of the code and documentation is available at:
http://bioinf.gen.tcd.ie/~faresm/software/software.html.

2.3.8 Assessment of susceptibility to false positives

Given the difficulty of finding definitive positive controls for an analysis of this nature we felt it important to demonstrate that this software is not susceptible to a large number of false positives. We simulated 20 alignments under a codon model with neutral evolution (non-synonymous , d_N, to synonymous, d_S, rates ratio $\omega = d_N/d_S = 1$) using the evolver package in PAML. These alignments were converted to amino acids and analysed with CAFS under the default alpha value of 0.05. With this value the expectation would be 5% of sites being reported as

functionally divergent. Our software reported an average of 2.3% of sites as functionally divergent. Given this low percentage we are confident that our methodology of significance testing and implementation of false discover rate is not susceptible to a large number of false positives.

2.4 Results

2.4.1 Analysis of the sensitivity of CAFS to false positives due to distance-based phylogenetic tree construction

We have developed this method with the purpose of performing large scale analyses upon whole proteomes and hence inferring shifts in functional categories. The larger the dataset the less feasible it is to create rigorous phylogenetic trees. After all, that would merely be the first step in this analysis. With the statements of the previous section in mind we have performed a comparison of our method using the calculated distance trees and trees built using RaxML (Stamatakis, 2006). We found that 92% of the sites identified using the distance-based trees were identical to those using the maximum likelihood method, these data are summarized in Table 2.1. Greater differences in the smaller alignments can be accounted for by the test statistic. This test statistic (equation 2.1) depends on two characteristics, firstly, the numerator will be large (either positive or negative) for a column in an alignment if there is a large amount of conservation in each clade whilst the amino acids in each clade are radically different. Secondly, the denominator depends on the variance in the aforementioned clades. These two factors mean that smaller alignments are subject to larger errors.

Comparison of sites when using Maximum likelihood or BioNJ trees				
# Sequences	#of ML FDsites	#BioNJ FD sites	Overlapping sites	%Overlap
10	10	3	3	100%
26	29	8	3	37.5%
34	72	40	23	57.5%
86	256	102	88	86%
204	134	87	82	94%
348	298	226	219	97%
692	293	230	225	98%
881	1092	696	643	92%

Table 2.1: ML vs BioNJ trees. The number of FD sites predicted at the 0.05 P-value level using BioNJ trees is consistently smaller. This indicates a stricter scoring scheme, which potentially reduces false positives. The final column shows the percentage of the FD sites detected through BioNJ trees that are also detected through ML trees increases. We note that with larger number of sequences in the alignment the percentage of overlapping sites approaches 100%. Only the smaller alignments show noticeable discrepancies but some of these can be explained by the effect of different tree topologies. Overall, the above results confirm the suitability of BioNJ for tree construction, particularly for alignments with a large number of sequences.

2.4.2 Time comparison: CAFS vs DIVERGE

Table 2.2 details the time taken to run eight sample alignments from the set of all alignments of all genes in the OMA database. All alignments were run upon the

same 4-core machine. It should be noted that the runtimes of the CAFS software includes the calculation of a BioNJ phylogenetic tree. The trees used in DIVERGE were pre-calculated trees and are not included in the time taken by DIVERGE. It is evident from table 2.2 that as the number of sequences in an alignment increases the greater the suitability of CAFS. It is also evident from this table that DIVERGE is not suitable at all for large alignments.

Comparison of CAFS and DIVERGE runtimes

#Species/Sequences	Alignment Length	Time(DIVERGE)	Time(CAFS)
10	334	NTN	NTN
26	510	11s	26s
34	247	10s	1m30s
86	316	3m16s	2m57s
204	492	Failed*	47m21s
348	550	Failed**	2h21m12s
692	587	Failed*	9h21m28s
881	1101	Failed*	20h20m54s

Table 2.2: Comparison of CAFS and DIVERGE runtimes. Shown here are the runtimes of CAFS and DIVERGE for a set of alignments of varying size. It is clear that CAFS performs better once any considerable number of sequences is used. It should be noted that in each case CAFS is calculating a BioNJ tree which accounts for a large portion of the times taken. NTN=no testable nodes. *Denotes the software failed constructing a tree or Reading a pre-made tree. **Denotes the software failed reading the alignment.

2.5 Discussion

2.5.1 Differences between our software and others

Given the differences between our software and that of DIVERGE it is very difficult to make comparisons in terms of the sites reported by each program. DIVERGE compares two clades after a duplication event and our software compares two clades and an outgroup. We perform our test in this manner in order to assess events of functional divergence at any testable node on a tree and also because our opinion is that using an outgroup provides stronger evidence of functional divergence. Another widely used piece of software is PAML (Yang 1997). PAML has the functionality to search for regions of positive selection. The test that is performed is d_N/d_S ratio in the yn00 and codeml packages of PAML. Whilst not directly comparable since PAML analyses nucleotides and CAFS analyses protein sequences it is important to discuss the differences between our program and that of PAML. As mentioned PAML searches for regions or genes which are under positive selection and as described in chapter 1 has been used in many analyses to great effect. A drawback of the d_N/d_S ratio is that there is no information pertaining to the amino acid properties. Whilst our software is similar to PAML in that they both assess some levels of "conservation" down a column of an alignment CAFS uses BLOSUMs to assess how radical a change is and therefore has an inherent estimation of the probability of a change in function. Additionally, our program can easily automate and carry out large analyses in addition to the ability to assess divergences on a grander scale (whole proteomes) which is not possible using PAML.

2.5.2 Comparison with DIVERGE

The most widely used program for detection of functional divergence is DIVERGE (Gu, 1999). Even though it is well-suited for individual analyses, it can not be used for a large-scale study, such as the those presented in chapters 3 and 5. This is because the size of alignments dealt with exceeds the limits of the data that DIVERGE can handle. It is also designed to be run interactively and can not be used in automated pipelines.

2.5.3 Justification for use of BioNJ

There are two main reasons why the use of BioNJ trees was chosen over maximum likelihood trees in this software. First of all, it runs much faster, which is critical for large-scale applications. Secondly, incorporating a maximum likelihood calculation (RaxML or PAML for example) would have a negative impact on both the less computer savy users as they would need an in depth knowledge of the ML package and also mean that future versions of this software would be constrained by the ML package. These elements would affect the automation of large scale analyses and also the "out-of-the-box" feel of the program. Additionally, the maximum likelihood methods employed in the Bio++ libraries which we use in our program are significantly slower than RaxML and would not have provided an adequate alternative.

2.5.4 A note about testable alignments

It should be noted that since DIVERGE could not run the four largest alignments in this subset of our dataset from chapter 3, we predict that at least half of the align-

ments in our full dataset can not be analyzed by the DIVERGE software. Indeed we were unable to run analysis on any alignment over 86 sequences long, however DIVERGE would read alignments up to 100 sequences long. Another problem about the calculations performed with DIVERGE2.0 is the impossibility to perform analyses collected in Gu2001 which pertains to the method in (Gu and Vander Velden 2002). This analysis did not work for any of the alignments above. All analyses failed with an error message (Please recheck input sequence data and tree information), for which we could not find documentation. We were also unable to obtain help directly from the author of the software.

2.6 Conclussions

In this chapter we present a novel model for the assessment of site-specific functional divergence. We believe the overall power of this software lies in the speed of runtime, flexibility of use and its user friendly nature. The implementation in $C++$ means that there is significant speed increases upon the simplified (both model and software) presented in Toft *et al.*.(2009). This speed increase allowed for much more complex calculations, such as simulation of alignments to assess significance along with a better estimation of the standard error (equation 2.2).

Further, we understand and acknowledge the value of phylogenetic trees built using both maximum likelihood and bayesian methods and we condone their use in small scale analyses. One of the biggest assets of our program is speed and automation. With this in mind, we wish to demonstrate to the user that whilst BioNJ trees are calculated within the program the results recieved are 92% comparable to those returned after the maximum likelihood analysis. Given the high level of similarity between the maximum likelihood built trees and those built with BioNJ any the

conclusions drawn in the results of chapter 3 would hold in either circumstance. It is our opinion that analyses on the scale of those detailed in the chapters 3 and 5 are not possible for many who have limited computational resources.

In reference to the DIVERGE comparisons we justified our calculations by demonstrating that DIVERGE is not ideal for large scale analyses and can be troublesome to the user for even relatively small alignments. In addition to the tests run above we would like to say that on large scale analyses all of the information about the sites tested are collected and automatically analysed according to any tagging system applied to the dataset and statistical tests performed "on the fly" to return the most information possible to the user. Large scale analyses are performed in chapters 3 and 5 upon 750 complete bacterial genomes and the clients and non-clients of the chaperonin GroEL.

Chapter 3

Proteome-wide analysis of functional divergence in bacteria: Exploring a host of ecological adaptations

3.1 Related manuscript

Caffrey B. E.*, Williams T. A.*, Jiang X., Toft C., Hokamp K., Fares, M. (2012) Proteome-Wide Analysis of Functional Divergence in Bacteria: Exploring a Host of Ecological Adaptations. PLoS ONE 7(4): e35659. doi:10.1371/journal.pone.0035659

*denotes equal contributions.

3.2 Introduction

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 & Conery, 2003); second, their ability to obtain genes by horizontal gene transfer (HGT), 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.

To investigate the evolution of prokaryotic diversity, we infer patterns of radical change for each protein individually, and then cluster species according to the functional categories (derived from COG: Clusters of Orthologous Groups; (Tatusov *et al.*, 2003)) in which they exhibit significant functional divergence. We perform an analysis of functional divergence on 750 bacterial proteomes using the software CAFS (Clustering Analysis of Functional Shifts). 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 ad-

ditional 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

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 user's 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 (Roth *et al.*, 2008; Schneider *et al.*, 2007), 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 ontology tags to each sequence (Tatusov *et al.*, 2003; Tatusov *et al.*, 2001). We then calculated the frequency of the modal COG tag in each group. To avoid ambiguity in the clustering of functional categories, we only analyzed those alignments which only had one COG category assigned. 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.

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 were 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.

3.4 Results and Discusion

3.4.1 A conserved functional core and variable crust in the evolution of bacterial proteomes

After assigning all alignments with a tag according to the COG tagging system and placing these into modal groups we found that the largest group (4,788 alignments) contained only one COG tag each, validating our approach to grouping homologs. Removing those groups with greater than one functional category and those with unknown function resulted in a final dataset of 3,813 groups of sequences.

An obvious sign of functional divergence would be a set of homologs that spans multiple COG categories. In this study we focus only on those alignments where all sequences have the same COG annotation. This represents the majority of homologs 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

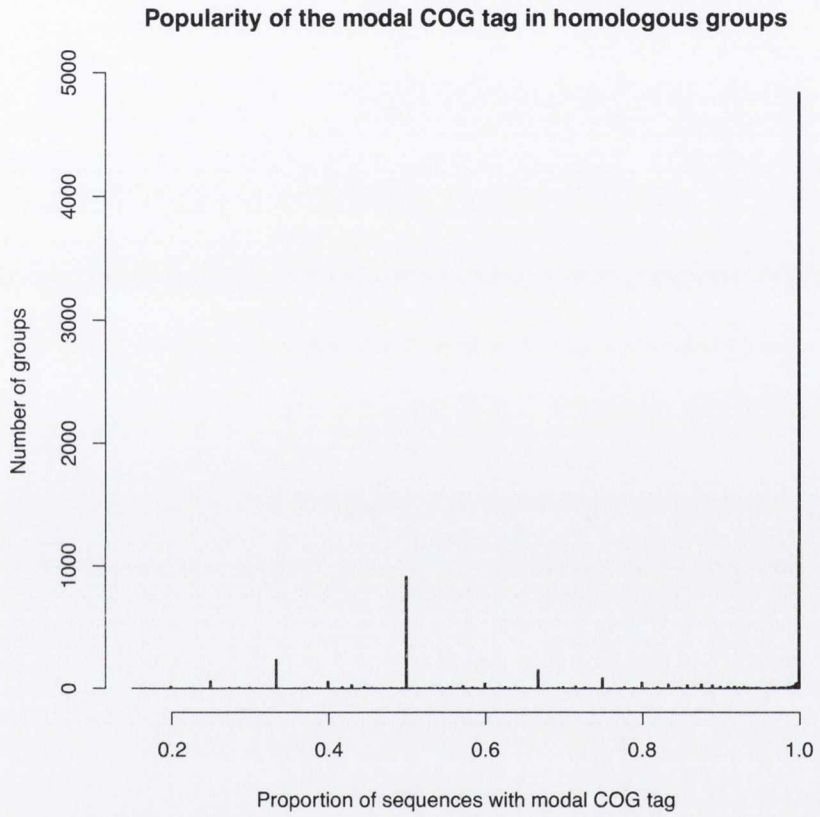

Figure 3.1: Frequency of modal COG tags. This plot shows a plot of the frequency of each modal group for the assignment of COG tags. That is, the x axis represents 1/(number of COG tags assigned to an alignment) and the y-axis represents how many groups had each value. It is evident that the majority of groups received only one COG assignment, thus justifying our methodology of groupings of homologous genes

tests to evaluate the differences in functional divergence between COG gene categories in our dataset (see Figure 3.2). 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, seventeen of the nineteen categories were either enriched or impoverished for functional divergence, while only two categories failed to 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 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. These findings are summarized in Table 3.1.

This supports the theory that bacterial proteomes comprise a relatively unchanging core (that is, genes in impoverished categories) coupled with a set of more variable functions (enriched categories) (Lake *et al.*, 1999; Makarova *et al.*, 1999; Mushegian & Koonin, 1996). The impoverished categories are almost exclusively those involved with information storage and processing, including DNA replication,

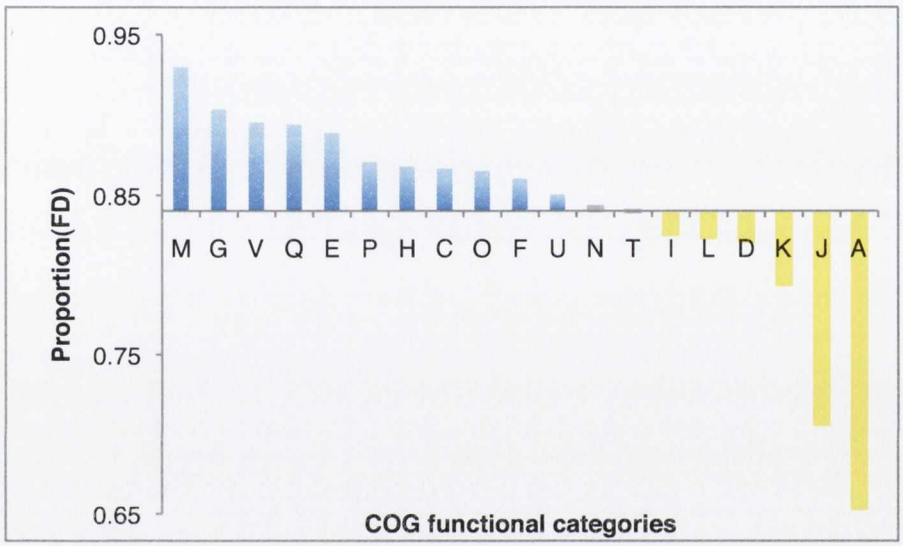

Figure 3.2: Enrichment of functional categories. This Figure shows the relative proportions of each COG category. Bars highlighted in yellow represent impoverishment in terms of functional divergence and bars in blue represent enrichment in functional divergence. The x-axis is centered at the mean proportion of $FDsites/Totalsites$. The colors, however, represent a chi-squared test that has been imposed on each of the categories with the aforementioned mean as the expected value.

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 among 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; Lake *et al.*, 1999; Makarova *et al.*, 1999; Mushegian & Koonin, 1996). 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 Cricks concept of the frozen accident (Crick, 1968), is that too many other genes depend on the basic functions of trans-

lation, transcription, and repair: functional changes in these genes would disrupt many other systems in the cell.

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 for each species individually using chi-squared tests. 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 3.2.

While there was no statistically 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 x 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

Enrichment Table		
COG Tag	Functional Category	Enrichment status
L	Replication, recombination and repair	Impoverished
D	Cell cycle control, cell division and chromosome partitioning	Impoverished
A	RNA processing and modification	Impoverished
J	Translation, ribosomal structure and biogenesis	Impoverished
K	Transcription	Impoverished
I	Lipid transport and metabolism	Impoverished
T	Signal transduction mechanisms	Not Enriched
N	Cell motility	Not Enriched
O	Posttranslational modification, protein turnover and chaperones	Enriched
U	Intracellular trafficking, secretion, and vesicular transport	Enriched
H	Coenzyme transport and metabolism	Enriched
Q	Secondary metabolites biosynthesis, transport and catabolism	Enriched
G	Carbohydrate transport and metabolism	Enriched
E	Amino acid transport and metabolism	Enriched
F	Nucleotide transport and metabolism	Enriched
C	Energy production and conversion	Enriched
P	Inorganic ion transport and metabolism	Enriched
M	Cell wall/membrane/envelope biogenesis	Enriched
V	Defense mechanisms	Enriched

Table 3.1: Description of functional categories. Presented here are the enrichment levels of the COG functional categories tested using CAFS. There is a stark separation between informational genes and energy production genes. The results tabulated here show that the analysis performed by CAFS not only returns useful information on an amino acid site resolution but also returns information according to the tagging system applied.

Lifestyle	Comparison	Enriched	Neither	Impoverished	Significance
Psychrophile	Mesophile	2/61	6/433	1/66	N.S.
Thermophile	Mesophile	7/61	22/433	1/66	N.S.
Pathogen	Non-pathogen	22/77	272/294	47/38	*** (-)
Intracellular pathogen	Other pathogen	0/22	26/246	4/43	N.S.
Symbiont	Non-symbiont	4/95	36/530	13/72	*
*Intracellular endosymbiont	All others	4/224	14/360	15/133	*** (-)
All interactors	Free-living	44/55	410/156	70/15	*** (-)

Table 3.2: Enrichment analysis of host-associated bacteria. 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, + or - denote the direction of the shift associated with the lifestyle being tested. For instance, interactors are significantly impoverished (-) compared to free-living bacteria

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 3.2). 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 3.2), their loss from host-associated bacteria may explain the relative impoverishment of functional divergence in these proteomes. Also, the remaining genes may be under strong constraints imposed by the specialized environment they live in, limiting therefore any opportunity for functional divergence (Toft and Fares., 2009). These enrichments are summarized in Table 3.2

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 144 kb and 792 kb depending on the host (in comparison, *E. coli* K12 has a genome size of 4.639Mb) (See for example

Enrichment of species according to biological properties.		
Term	Coefficient	P-value
Intercept	2.265	$8.23 * 10^{-7}$
Lifestyle (Host-assoc.)	-2.536	$1.28 * 10^{-13}$
Genome size (bp.)	$-1.799 * 10^{-7}$	0.0212

Table 3.3: Enrichment of species according to biological properties. We used a generalized linear model with binomial errors to assess the impact of lifestyle and genome size on the enrichment and 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.

(Degnan *et al.*, 2005; Gil *et al.*, 2002; Nakabachi *et al.*, 2006; Perez-Brocal *et al.*, 2006; Shigenobu *et al.*, 2000; Tamas *et al.*, 2002; van Ham *et al.*, 2003). 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 (Table 3.3). 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 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 From proteome-wide to residue-level 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 3.3a. 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 highlight-

Comparison of host-associated and free-living bacteria with regard to COG categories		
COG category	Free-living	Host-associated
C	Enriched	Enriched
P	Enriched	Enriched
M	Enriched	Enriched
V	Enriched	Enriched
L	Impoverished	Impoverished
D	Impoverished	Impoverished
Q	Enriched	Enriched
G	Enriched	Enriched
E	Enriched	Enriched
J	Impoverished	Impoverished
U	Not Enriched	**Enriched**
T	**Enriched**	Impoverished
K	Impoverished	Impoverished
O	Enriched	Enriched
N	Not Enriched	Not Enriched
I	Impoverished	Impoverished
H	Enriched	Enriched
F	Enriched	Enriched
A	Impoverished	Impoverished

Table 3.4: Comparison of functional categories for free-living and host-associated bacteria. Categories U and T show different levels of enrichment for functional divergence when the analysis is run on these groups of bacteria independently

ing 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 3.3), 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 (Figure 3.3b), 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 bacilliformus*, is enriched for functional divergence in cell motility genes (N), with the others being 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 bacilliformus* 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 (Dehio, 2001; Ihler, 1996). 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

(Figure 3.3c). As expected, closely related *E. coli* and *Shigella* strains cluster together at the bottom of the dendrogram (Figure 3.3a), 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 basis of an overall relaxed functional constraint, particularly in the categories of cell motility (N) and protein modification, turnover and chaperones (O): most others are 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 (Fricke *et al.*, 2008; Ren *et al.*, 2005). 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 O17: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 not enriched in this category.

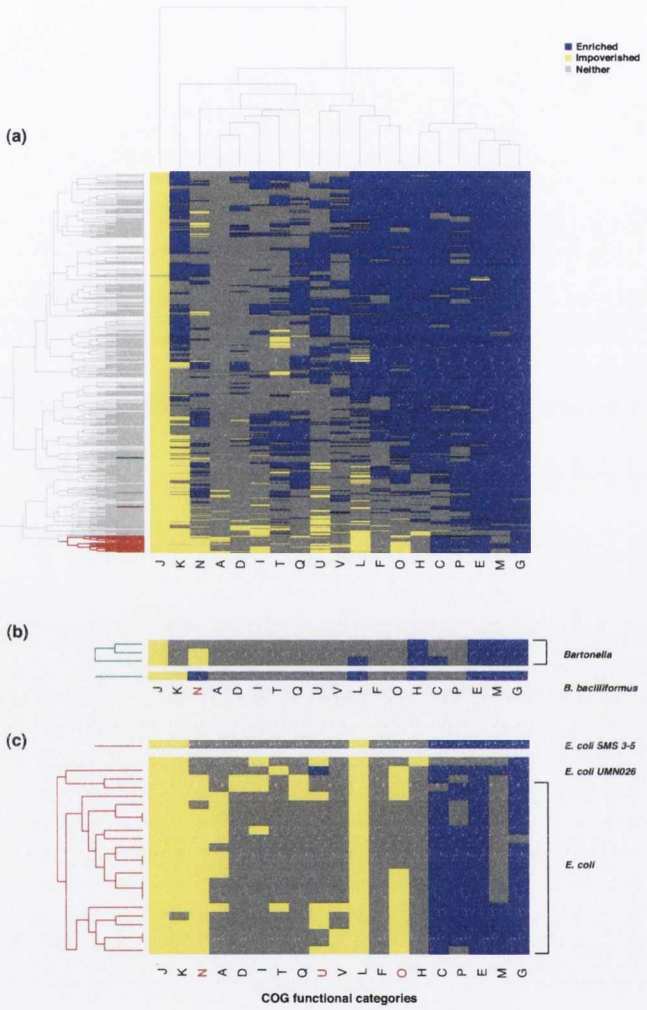

Figure 3.3: Visualizing high-level patterns of functional divergence. Section (a) shows a clustered heat map of all 750 species. Yellow color represents impoverishment for functional divergence. Blue represents enrichment for functional divergence. Obvious trends include impoverishment for ribosome structure and biogenesis (J) and transcription (K) genes. The energy production and metabolism genes are enriched (G, E, P, C, H) along with cell wall (M) and much of defense mechanisms (V). In sections (b) and (c) show examples of strains with atypical spectra of functional divergence.

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 3.4). To gain further insight into the possible implications of the UMN026-specific changes in these proteins, we mapped the specific residues under functional divergence in VirB8 (also output by CAFS) onto the Agrobacterium tumefaciens crystal structure (Bailey *et al.*, 2006).

Of the 15 sites under functional divergence (Table 3.5), six could be mapped onto the crystallized region of the protein. Of these six, five 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 homologs, is directly adjacent to two highly conserved sites, Ala-195 and Thr-196. Thr-196, which CAFS 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 CAFS 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

Summary table of sites under functional divergence in Vir-B9	
E. coli UMN026	*Agrobacterium tumefaciens*
Y116	F127
L176	V183
D189	T196
E206	L211
T213	V218

Table 3.5: Sites of functional divergence in Vir-B9. The left column shows the sites found in E.coli UMN026, the column on the right shows the homologous sites in Agrobacterium tumefaciens. Sites without a value for Agrobacterium tumefaciens represent sites, which have not been crystalised.

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.

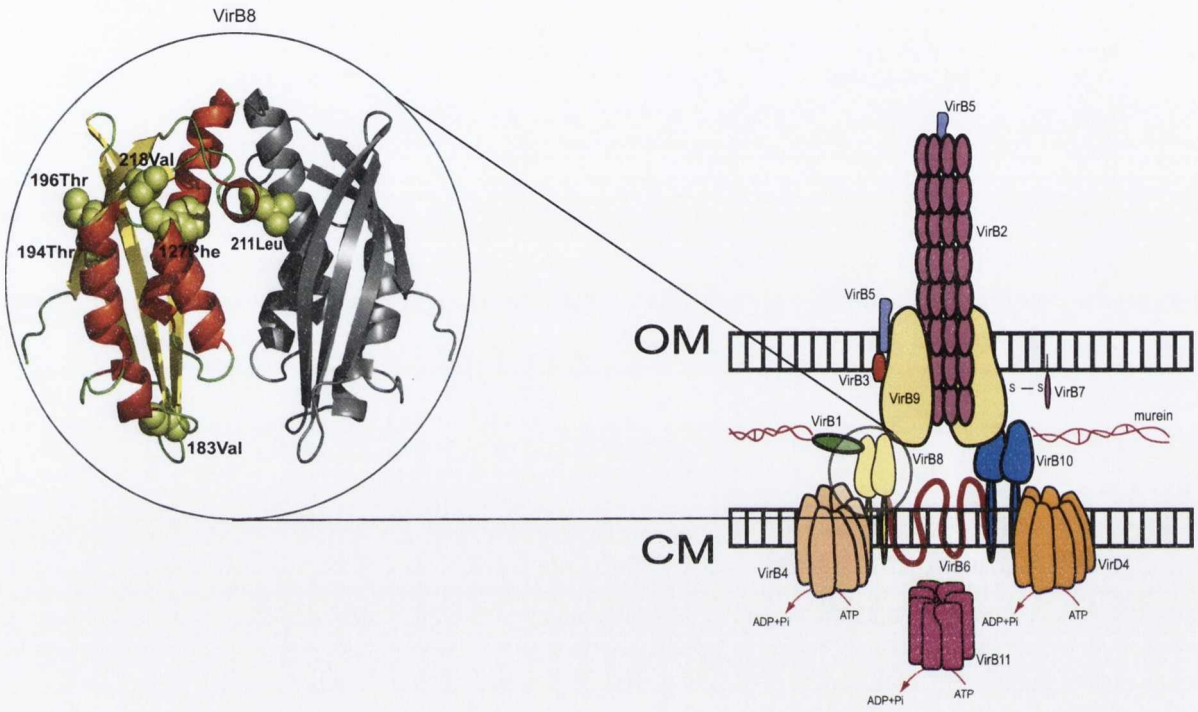

Figure 3.4: Amino acid residues under functional divergence in E. coli UMN026 VirB8. *Right*: the structure of a Type IV secretion system found only in two strains of E. coli. CM = cytoplasmic membrane, OM = outer membrane. The complex structure is based on that of Baron (2006). In UMN026, the central complex proteins VirB8 and VirB9 are under functional divergence. *Left*: Of the five sites detected by CAFS that could be mapped to the VirB8 crystal structure (Bailey *et al.*, 2006), there is evidence that four are involved in forming the VirB8 homodimer, suggesting that functional divergence at these positions is involved in altering the quaternary structure of the complex.

Chapter 4

CAPS 2.0: Disentangling phylogenetic and functional coevolution

4.1 Related manuscript

Caffrey, B.E., Hokamp, K, Williams, T.A., M. A. Fares. CAPS 2.0: Disentangling phylogenetic and functional coevolution. In preparation.

4.2 Introduction

In his remarkable book *On the Origin of Species* (Darwin 1859), Charles Darwin emphasized his observation on the variation of natural forms and attributed this to an underlying force called "natural selection". By this consideration, he challenged views at the time following the argument that inherited adaptations and coadaptations of species could not be attributed to external environmental conditions. Neither could the merit of coevolution between different forms be adduced to forces of creation, claims that he qualified as preposterous. Ever since his first

observations, Charles Darwin realized the need for uncovering the mechanism and means of modification and coadaptation.

The term coadaptation was first used in an ecological context to refer to the interaction and the "Evolution together" of interacting ecological entities. Anton De Bary broadly classified these interactions within the category called "symbiosis", including mutualists, pathogens and saprophytes (Anton De Bary, 1879). In the relationship between two organisms, especially between hosts and parasites, an arms race is established, *i.e.* that the host develops new systems to combat the pathogen while the pathogen develop better invading systems and the ability to escape the defense system of the host. The co-dependence or co-variation concept was also applied to account for co-variation of morphological characters (Pagel *et al.*. 1994).

In molecules, the concept of coevolution was first used in the understanding of covariation using DNA/RNA molecules (Schoniger and von Haeseler 1994; Rzhetsky 1995). However, we can identify different levels of coevolution that result from the fact that proteins do not act in isolation but they form part of complex networks of molecular interactions. Determining these interactions and their coevolutionary dynamics has been and remains to be a formidable challenge in the fields of proteomics and systems biology (Fares, Ruiz-Gonzalez, and Labrador 2011).

Molecular coevolution relies on the idea of covariation proposed by Fitch and Markowitz (Fitch and Markowitz 1970). According to their proposal, at any evolutionary time, proteins comprise two types of regions: those that are variable and those that are subjected to strong functional and structural constraints and are hence conserved. The fixation of changes in variable regions of the molecule can modify the fitness landscape of the protein in such a way that the strong constraints in the conserved protein regions may change so that they become variable. This idea was

later completed by Fitch (Fitch 1971) to account for co-dependence between amino acid regions.

Proteins rarely function in isolation but they interact in a fine-tuned way with other proteins in the cell or between cells to perform a particular task (Albert, Jeong, and Barabasi 2000). Those interactions that are important in the cell are usually conserved throughout evolution, implying that changes in one of the proteins (protein A) of the interaction pair will induce changes in their interacting partner (protein B) in a way that will ensure the conservation of the interaction. Reciprocally, the evolutionary changes in B will impose constraints on its interaction partner A. These reciprocal selective constraints will lead to coevolutionary dynamics of ensuring the coadaptation of both interacting molecules to become successfully fixed in the population.

This principle, which is the main focus in this thesis chapter is an area of research in evolutionary biology which has yet to be fully understood. Identifying pairs of coevolving proteins could in principle lead to the identification of protein-protein interactions. Describing the interactome has been the focus of high throughput experimental approaches, all of which involve high economic and effort expenses. Developing computational tools for the *in silico* identification of protein-protein interactions is of justifiable importance. There are many more advantages that computational methods can offer against high-throughput approaches: (a) computational methods are faster to perform; (b) computational methods are cheaper; (c) they can provide useful information not contained in experimental approaches, such as the amino acid sites involved in the interaction between proteins. Also identifying coevolution in molecules is useful from different perspectives: (a) Functional annotation of proteins encoded by unknown genes; (b) revealing interactions between amino acids within a protein, hence enabling predictions of protein three-

dimensional structures and folding dynamics; and (c) understanding how complex machineries undergo adaptive changes despite not having meaningful effects on the organism's fitness (Codoner and Fares 2008).

Despite the importance of identifying molecular coevolution for the prediction of protein-protein interactions, only one study has been developed and carried out so far (Pazos *et al.*. 2008). While this approach has shown promising results, methods remain to be far from performing reliably owing to the large amount of false positive signals present in the results of coevolution analyses: (a) stochastic coevolution and (b) phylogenetic historical coevolution. The remainder of this chapter will therefore be devoted to understanding the sources of false positive signals when carrying out both intra and inter-molecular coevolution.

When performing coevolution analysis, one needs to bear in mind that there is a significant information gap between the number of available protein sequences and the number of protein crystal structures. This gap is an important limitation in identifying functional coevolutionary relationships because structural amino acids proximity can explain their evolutionary dependence through their reciprocal functional and structural constraints. Coevolution methods designed to identify atomic interactions between amino acids in a protein are powerful tools that counterbalance the lack of three-dimensional structure information (Gobel *et al.*. 1994; Pazos, Olmea, and Valencia 1997). Traditionally, experimental biology has attempted to unravel these interactions by directed mutagenesis of single amino acids, which has proven frustratingly useless owing to the astronomically large number of combinations needed to characterize such interactions in a small molecule. Alternatively, covariation analyses considering a large evolutionary scale (for example, comparing the amino acid sequence variation for a protein taking distantly related organisms) has provided insightful information that could also be used in more directed muta-

genesis experiments (Fares 2006; Fares and Travers 2006; Travers and Fares 2007).

A more approachable problem is the identification of coevolutionary relationships between amino acid sites within a protein. This coevolution is expected if we consider that proteins fold according to the complex atomic interactions of the constituent amino acids. Since protein folding is essential for protein function, selection will favor the appropriate interaction between the amino acids to optimize protein folding. Within-protein amino acid coevolution is therefore the result of the complex interactions between amino acids (Fares, 2006; Codoner *et al.*, BMC Evol. Biol. 2007).

There have been many methods to identify the coevolution of amino acid sites, all of which can be broadly classified into two main categories: (a) Methods based on the measure of mutual information of pairs of amino acids; and (b) methods based on the measure of the correlated variation between amino acids (distance methods). These methods rely on the assumption that some functionally or structurally linked sites or those amino acid sites surrounding important functional or structural regions should covary to maintain the main protein features (Taylor and Hatrick 1994; Atwell *et al.*. 1997; Chelvanayagam *et al.*. 1997; Pazos, Olmea, and Valencia 1997; Martin *et al.*. 2005; Codoner, Fares, and Elena 2006; Kim *et al.*. 2006). Other methods, also based on coevolution, have focused on the identification of interactions between motifs or proteins (Goh *et al.*. 2000; Pazos and Valencia 2001; Goh and Cohen 2002; Goh, Milburn, and Gerstein 2004; Pazos *et al.*. 2008).

Covariaton between amino acid sites is due to several factors that can be explicitely determined. For example, two amino acids, i and j, can covary due to multiple dependencies, including:

$$C_{ij} = C_{phylogeny} + C_{structure} + C_{function} + C_{interactions} + C_{stochastic} \quad (4.1)$$

These terms are the main factors to disintangle by coevolutionary analyses and remain to be a challenge for most methods developed for this purpose (Atchley *et al.*. 2000). Phylogenetic covariation was first proposed by Felsenstein to highlight the historical dependence between species (Felsenstein 1985). With regard to the other coevolution terms, these usually produce confounding results and are very difficult to distinguish from one another. All methods developed to identify coevolution have as their main objective an aim to identify each of the terms of C_{ij}.

Methods to identify coevolution can be broadly classified into parametric and non-parametric. Of these, the two most widely used methods are those based on the calculation of the mutual information content (MIC) of two amino acids, based on information theory (Kullback 1959). These methods measure of the amount of information (variability) contained in an amino acid site within a multiple sequence alignment. This information can be measured using the Shanon entropy (H(i)), which determines the probability of one of the 20 amino acid states being present at any particular time at a position of the alignment:

$$H(i) = -\sum_{i} P(x_i) log P(x_i) \quad (4.2)$$

Where i is the number of sequences and x is are the transition probabilities. The

joint probability distribution for amino acids i and j can be written as:

$$H(i,j) = -\sum_{x_i,y_i} P(x_i,y_i) log P(x_i,y_i) \tag{4.3}$$

Using these two terms, the mutual information can be calculated as:

$$MI = H(i) + H(j) - H(i,j) \tag{4.4}$$

MI values range between 0 (no covariation between amino acid sites) and a positive value whose magnitude is proportional to the strength of covariation between the two amino acid sites considered. Unfortunately, the identification of covariation using these methods is highly dependent on the amount of conservation of the multiple sequence alignment (Fodor and Aldrich 2004). These methods have been intensively used to detect sites of coevolution due to functional constraints and those due to protein-protein interactions (Korber *et al.*. 1993; Clarke 1995; Atchley *et al.*. 2000; Wollenberg and Atchley 2000; Hoffman, Schiffer, and Swanstrom 2003; Tillier and Lui 2003; Gloor *et al.*. 2005; Hummel *et al.*. 2005; Tillier *et al.*. 2006; Tillier and Charlebois 2009; Gloor *et al.*. 2010).

At the opposite end of methods to detect coevolution, many other methods have been based on the identification of correlations between matrices of amino acid distances for particular sites of a multiple sequence alignment. Within these methods, several authors relied on the matrices being directly extracted from phylogenetic contexts, so that the length of branches from a tree built for one amino acid site were compared to that of another amino acid site (Neher 1994; Pazos *et al.*. 1997; Goh *et al.*. 2000; Goh and Cohen 2002; Kim, Bolser, and Park 2004; Pazos *et al.*. 2005; Kim *et al.*. 2006; Pazos *et al.*. 2008; Pazos and Valencia 2008). These meth-

ods however present certain limitations owing to the lack of correcting parameters that should account for how radical amino acid transitions are (Martin *et al.*. 2005; Fares and Travers 2006). These methods rely on the Pearson's correlation coefficient discussed in the Materials and Methods section.

Finally, alternatively to these non-parametric approaches, other authors have developed parametric methods based on probabilistic approaches to detect coevolution. Some of these methods are based on maximum-likelihood approaches (Pollock and Taylor 1997; Pollock, Taylor, and Goldman 1999; Choi, Li, and Lahn 2005), on Bayesian probabilities (Dimmic *et al.*. 2005), or sequence divergence based approximations (Fares 2006; Fares and McNally 2006; Fares and Travers 2006). Unfortunately, the performance of most these methods remains underwhelming and suffer from a large number of false coevolutionary signals between amino acid sites.

In this chapter we present the improvement of a previous method to detect coevolution (Fares and Travers 2006) that includes additional biological information, largely ignored by the hitherto developed approaches.

4.3 Materials and Methods

4.3.1 Workflow

Figure 4.1 illustrates the workflow of the CAPS 2.0 software. The input for this software/analysis is one alignment for intra-molecular analysis and two alignments for inter-molecular analysis. Optionally, a crystal structure and phylogenetic tree can be submitted for each alignment otherwise a tree will be calculated automatically by the software.

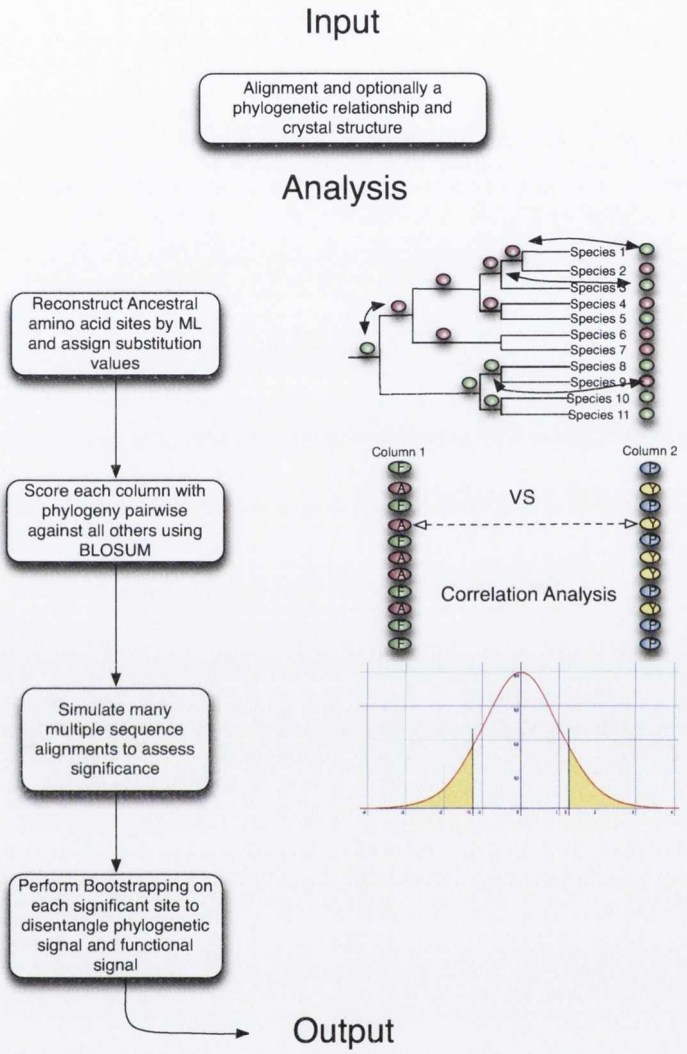

Figure 4.1: CAPS workflow. The workflow of CAPS 2.0 is shown above. The input to this analysis consists of an alignment or group of alignments. Optional input data can be given in the form of a crystal structure and phylogenetic relationship. CAPS reconstructs ancestral sequences with a maximum likelihood approach and using the JTT model. Next correlations are assessed between columns in the input data (either intra molecularly or inter-molecularly). Assessment of significance is performed by simulating multiple sequence alignments and drawing a distribution. Finally, a bootstrapping analysis is performed upon individual sites to separate purely phylogenetic signal from functional signal.

4.3.2 Evolutionary blindspot

One major limitation of CAPS version 1 (and many other programs) is the lack of use of evolutionary history. The current methodology is to take an alignment and analyze it without any assumption of evolutionary history or phylogenetic relationship. Consider Figure 4.2, which demonstrates the striking differences in substitutions when phylogeny is considered. Firgure 4.2a shows that without use of phylogenetic signal there is no option other than to compare all sequences against all others. Figure 4.2b shows an example where it is possible that one pair of amino acids had a correlated change in the last common ancestor of species 1-5. Figure 4.2c shows an example with multiple correlated changes along the collective evolutionary history of these proteins. This example is much more likely to suggest a pair of sites that are functionally coevolving. It should be noted that there is no reason to assume that the example in Figure 2abis not coevolving but rather that the example in 2c contains much stronger evidence of coevolution.

4.3.3 Scoring coevolution

Following the parsing/calculation of a phylogenetic tree and ancestral reconstruction of sequences at all inner nodes, we first calculate the mean transition parameter θ for each column, given by:

$$\bar{\theta} = \sum_{S=1}^{n} (\theta_{ij})_S \tag{4.5}$$

Where θ_{ij} is the transition score given by the weight matrix for an amino acid substitution from state i to state j and S sums over transitions on the phylogenetic

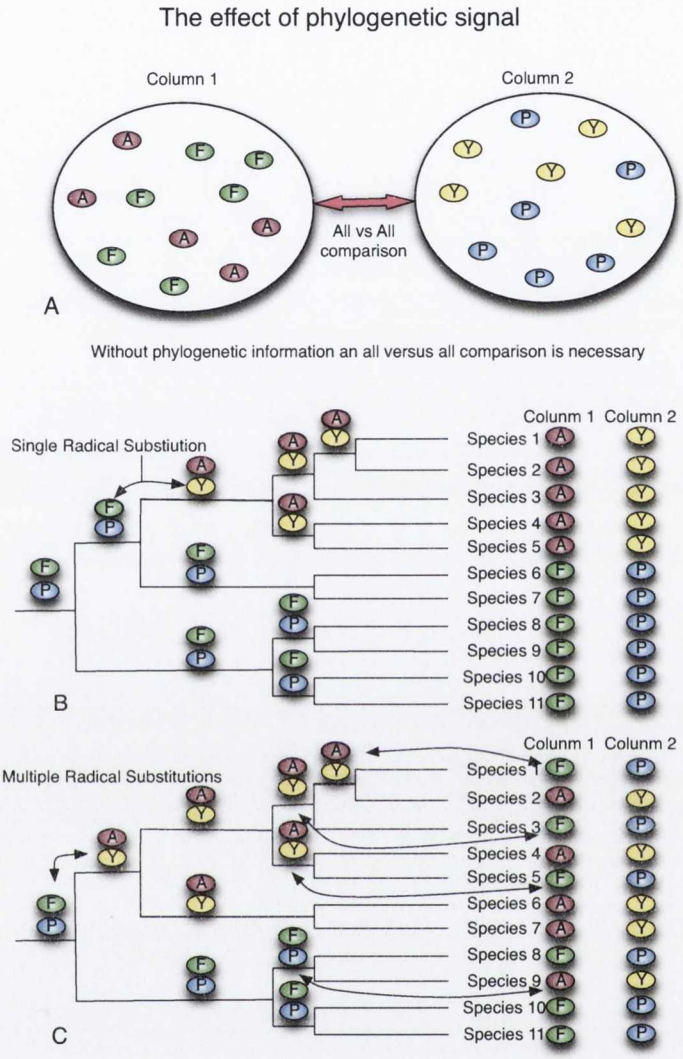

Figure 4.2: The necessity of phylogeny and ancestral reconstruction. This Figure illustrates the shortcomings of many current softwares for assessing coevolution. In part *A* we see the arrangement of amino acids and when an alignment alone is analyzed. Without knowing the phylogenetic relationships involved there is no other course of action than to compare all transitions against all. This leads to multiple testing of various sites; additionally, it is impossible to know of the significance of each transition. The two images in part **B** and **C** illustrate two alternate scenarios where phylogenetic relationships can lead to two very different outcomes. The first shows the case where there are two paired radical substitutions leading to all of the signal in the test whereas the second shows the case of multiple paired radical substitutions. The second gives much stronger evidence of coevolution. It is clear from these analogies that phylogenetic relationships are necessary in the assessment of coevolution.

tree in each column. Therefore each i-j state is the transition score from a child node to its parent. The next step is to calculate the variability of each amino acid transition compared to the mean transition parameter. This is given by:

$$D_{ij} = [(\theta_{ij} - \bar{\theta}]^2 \tag{4.6}$$

and the mean of D given by:

$$\bar{D} = \sum_{S=1}^{n} D_S \tag{4.7}$$

Correlation coefficients are then calculated according to Pearson's correlation coefficient ρ_{AB} :

$$\rho_{AB} = \frac{\sum_{S=1}^{n} [(D_S)_A - \bar{D_A}][(D_S)_B - \bar{D_B}]}{\sqrt{\sum_{S=1}^{n} [(D_S)_A - \bar{D_A}]^2} \sqrt{\sum_{S=1}^{n} [(D_S)_B - \bar{D_B}]^2}} \tag{4.8}$$

Where A and B are two amino acid site columns (from the same or different multiple sequence alignments) and the summing over the variable S is over the transitions from child to parent within each column. These values (ρ_{AB}) are then tested for significance against a null distribution drawn from neutrally evolving multiple sequence alignments (see below).

4.3.4 Simulations and significance test

The existing functionality of CAPS depends on the user providong a threshold cut-off for the correlation coefficient or for the program to decide the cut-off by re-sampling. Since re-sampling is carried out by sampling the original scores repeatedly, the identification of a cut-off value in this manner is flawed as any biases or trends which are inherent in the input data will also be inherent in the cut-off value. Additionally, the existing software substitution scores are normalized according to the relative distances between each pair of sequences. BLOSUM matrices correct for this except in cases where the distance between sequences is much greater than that of the mean pairwise distance. Therefore in cases where the distances are of an acceptable order we have a double correction. The proposed alternative should circumvent both this problem and that of the threshold cut-off.

To obtain a more intuitive cut-off value we propose simulating multiple alignments repeatedly according to a neutral theory of evolution under the JTT model (JTT, 1992) and scoring these by the same procedure as the tested data, thus building a distribution from which we can obtain a cut-off value. Since we wish to take phylogenetic distance into account we simulate the aforementioned alignments with the same pairwise distances between species, meaning that the corresponding cut-off value is scaled according to the distances involved. Neutral evolution here is understood as the fixation of amino acid substitutions following a neutral model in which coevolution is not parameterized.

4.3.5 Disentangling phylogenetic signal and functional signal

Bearing in mind the prior discussion of an evolutionary blind spot, we proposed an alternative model, which incorporates a phylogenetic tree and ancestral sequences. This new model allows us to search through regions of the tree, which may have lead to false positive pairs of sites through phylogenetic relationships alone. We perform a bootstrapping step on all sites that are reported as being statistically significant according to the simulations and significance testing. To bootstrap a pair of sites in an alignment (or pair of alignments) we randomly sample the amino acid substitution scores at those columns and then perform the correlation coefficient(given above) on these sampled sites. This procedure is performed 10000 times; the bootstrap value is calculated by dividing the number of times each of the bootstrap correlations still pass the threshold (obtained from the simulations described above) by 10000. The assigned bootstrap value can be used to remove results that fall below a user-specified bootstrap threshold.

4.3.6 Implementation

CAPS 2.0 was implemented in $C++$ and is available under the GNU General Public License v.3 for Linux and Mac. The code was written using the GNU Scientific Library and the Bio++ libraries(citation). The program is accompanied by full documentation and enables the user to perform intra-molecular or inter-molecular coevolution analysis under the new model and bootstrapping described above. The latest version of the code and documentation is available at http://bioinf.gen.tcd.ie/~faresm/software/software.html. Additionally, we have developed a web interface for CAPS available at http://bioinf.gen.tcd.ie/caps/ which

includes a Jmol and Cytoscape plugin for advanced visualizations of the results. Some screenshots are provided in Appendix B.

4.3.7 Data

All available crystalized protein structures for gamma-proteobacteria were downloaded from the Protein Data Bank (PDB, www.rcsb.org/pdb). This comprised 1090 entries, representing 20-25% of the *E. coli.* proteome. Protein sequences homologous to the structure-associated sequence were obtained by reciprocal Blast searching of 85 complete gamma-proteobacterial genomes. Hits with an E-value smaller than 10^{-1} were retained. Sets of homologs were aligned with ClustalW (Thompson *et al.*. 1994) using default parameters. The quality of these alignments was inspected manually. Cluster of Orthologous Genes (COG) tags were obtained from NCBI and assigned to genes for analysis of under- and over-representation of functional categories.

In the case of intra-molecular analysis, which included structural analysis, we only used columns of the alignments which could be aligned to the structural sequence and columns which had fewer than 20% gaps. Models were predicted for each alignment using ProtTest (citation) and maximum likelihood trees were constructed using RAxML with 100 bootstraps and the thorough maximum likelihood search.

4.4 Results

4.4.1 Removal of phylogenetic signal

Given the low numbers of well-known and experimentally validated coevolving sites it is hard to find a definitive positive control or benchmarking system. To make any estimates of coevolutionary signal which is functional, structural or involved in interactions, it is critically important to make use of structural information in any benchmarking, as it is well known that, while some coevolving sites can be proximally distant from each other (Plotnikova *et al.*, 2004), it is accepted that interacting residues generally lie close together in three-dimensional space purely because of the physiochemical properties of the various bonds underpinning protein structure. Under this reasoning we have chosen the dataset of all gamma-proteobacterial genes for which there are many known crystal structures.

We performed intra-molecular analysis of 1090 alignments with their corresponding structures and phylogenetic trees. A range of bootstrap values were used in order to demonstrate the relative effect of bootstrapping. The results are shown in Figure 4.3. We placed the pairs of coevolving sites into bins according to the distance between the amino acids in three-dimensional space and plotted this against the mean bootstrap value for each bin. The three-dimensional distances were calculated from the centroid of each amino acid followed by the Euclidian distance between each pair of amino acids. Figure 4.3 shows that the higher the bootstrap value the shorter the distance between the amino acids. This demonstrates that sites which coevolve throughout the phylogeny are more functionally or structurally related than those within sub-sections of the phylogeny. We also present the effect of bootstrapping in Figure 4.4, showing that there may be an upper bound to the positive affects

of bootstrapping, *i.e.* once the bootstrap threshold is set above 0.99 the software begins to remove functional signal also.

In order to test the susceptibility of CAPS version 2 to false positives we simulated 100 alignments using codeml from the PAML package without the effect of natural selection (ω =1). CAPS was run in intra-molecular mode with type I error (alpha) of 0.05 and with a modest bootstrap threshold of 0.7. We found that only 0.004% of all possible pairs of sites were predicted to be coevolving. The previous version of CAPS uses a resampling technique which returns 2% of sites coevolving with alpha=0.05. We also found setting the threshold bootstrap value higher does reduce the number of coevolving pairs, owing to a reduction of false positives (Figure 4.4). This Figure also shows that when the bootstrap value is raised sufficiently high there is a loss of signal.

4.4.2 Analysis of inter-molecular coevolution

To analyze the accuracy of inter molecular coevolution analysis performed by CAPS 2.0 we investigate the coevolutionary patterns of the co-complex structures available for gamma proteobacteria, *i.e.* definitive examples of interactions. We downloaded all of the co-complex structures from the Protein Data Bank (PDB, www.rcsb.org/pdb). We limited our analysis to those complexes which comprised only two different associated sequences. We eliminating those with multiple sequences because results obtained from the analysis of complexes of varying numbers of proteins make interpretation of any results obtained extremely difficult.

As a comparison we used the Mirrortree software (Pazos and Valencia, 2001), which is available from the developer in binary format. We ran each alignment against all others and then ranked them according to the highest signal. When

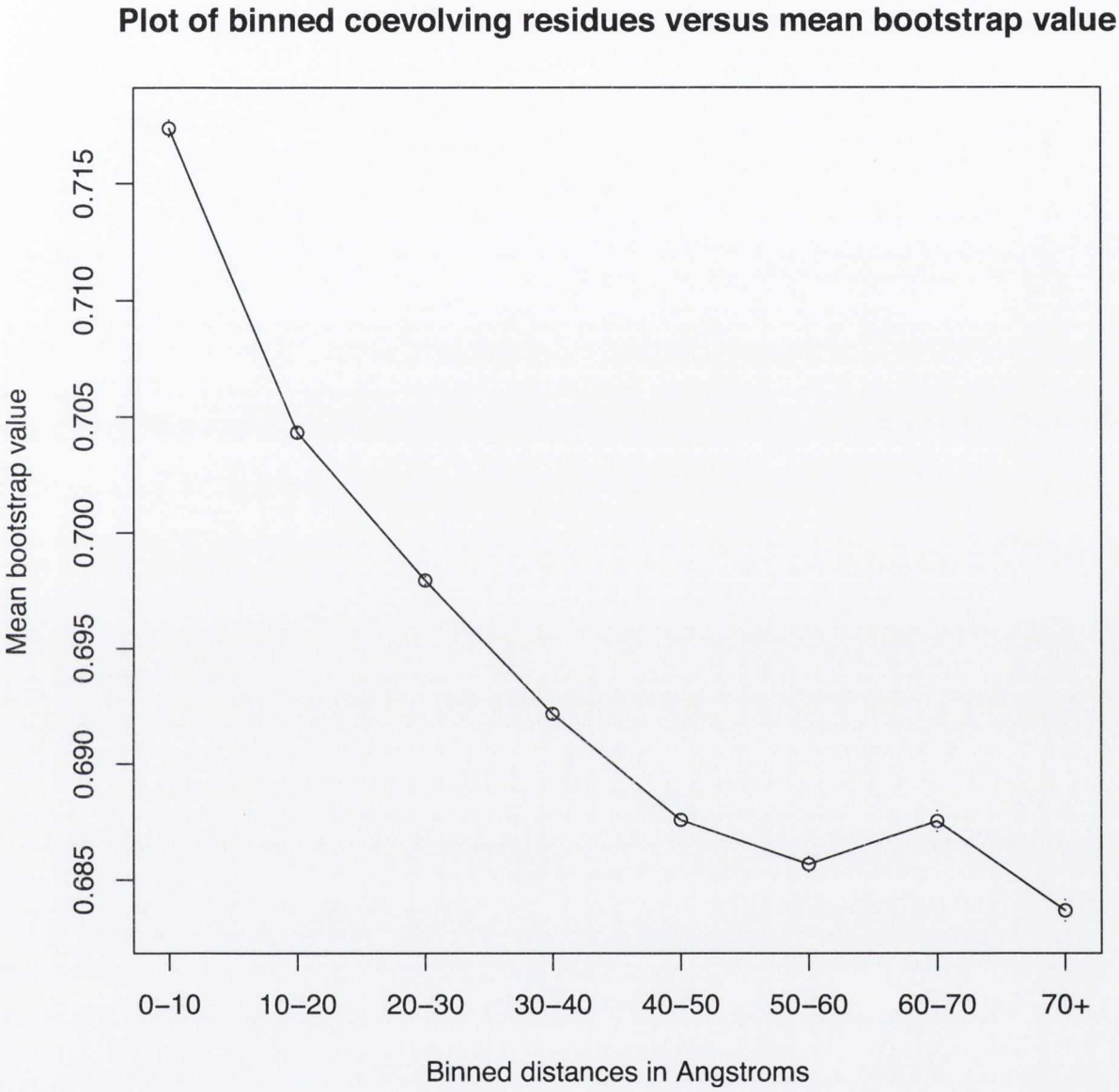

Figure 4.3: Plot of binned distances versus mean bootstrap value. This graph represents all pairs of intra-molecularly coevolving sites in 1090 proteins from the gamma proteobacteria. Pairs of sites were binned according to their structural pairwise distance calculated from their corresponding pdb structures. There is a clear trend showing that pairs of coevolving sites which are closer together have higher bootstrap values, thus justifying our bootstrapping method. **Note:** This graph would have error bars but the standard error is so small that it is not possible for the graphing software to show them

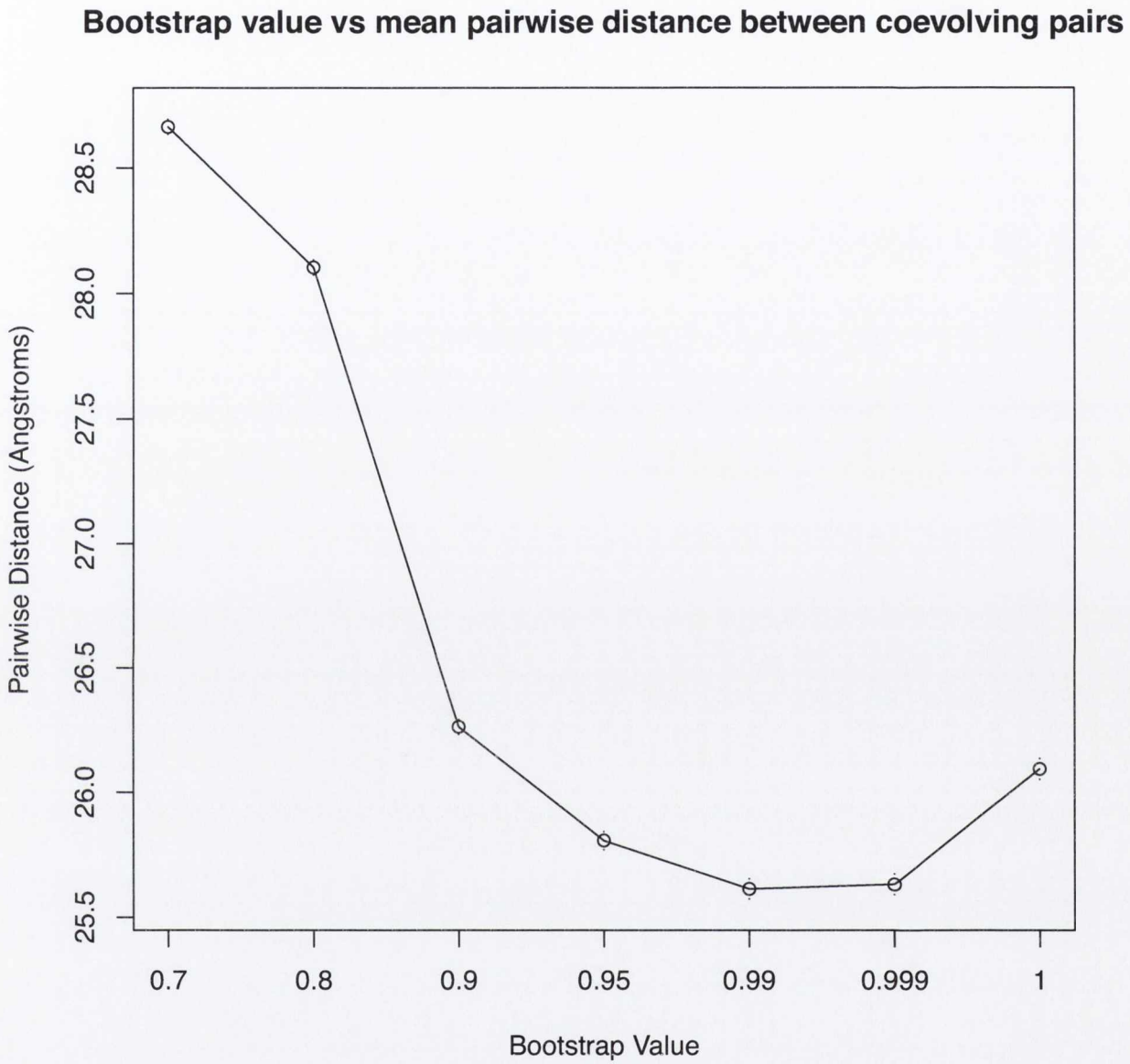

Figure 4.4: Plot of bootstrap value versus mean distance. This graph shows the average pairwise distance between coevolving pairs. It can be seen here that the average pairwise distance decreases as the bootstrapping threshold is increased. This suggests that stronger coevolutionary signal under this model is more likely to yield functionally important pairs of coevolving sites. **Note:** This graph would have error bars but the standard error is too small for the graphing software to show them.

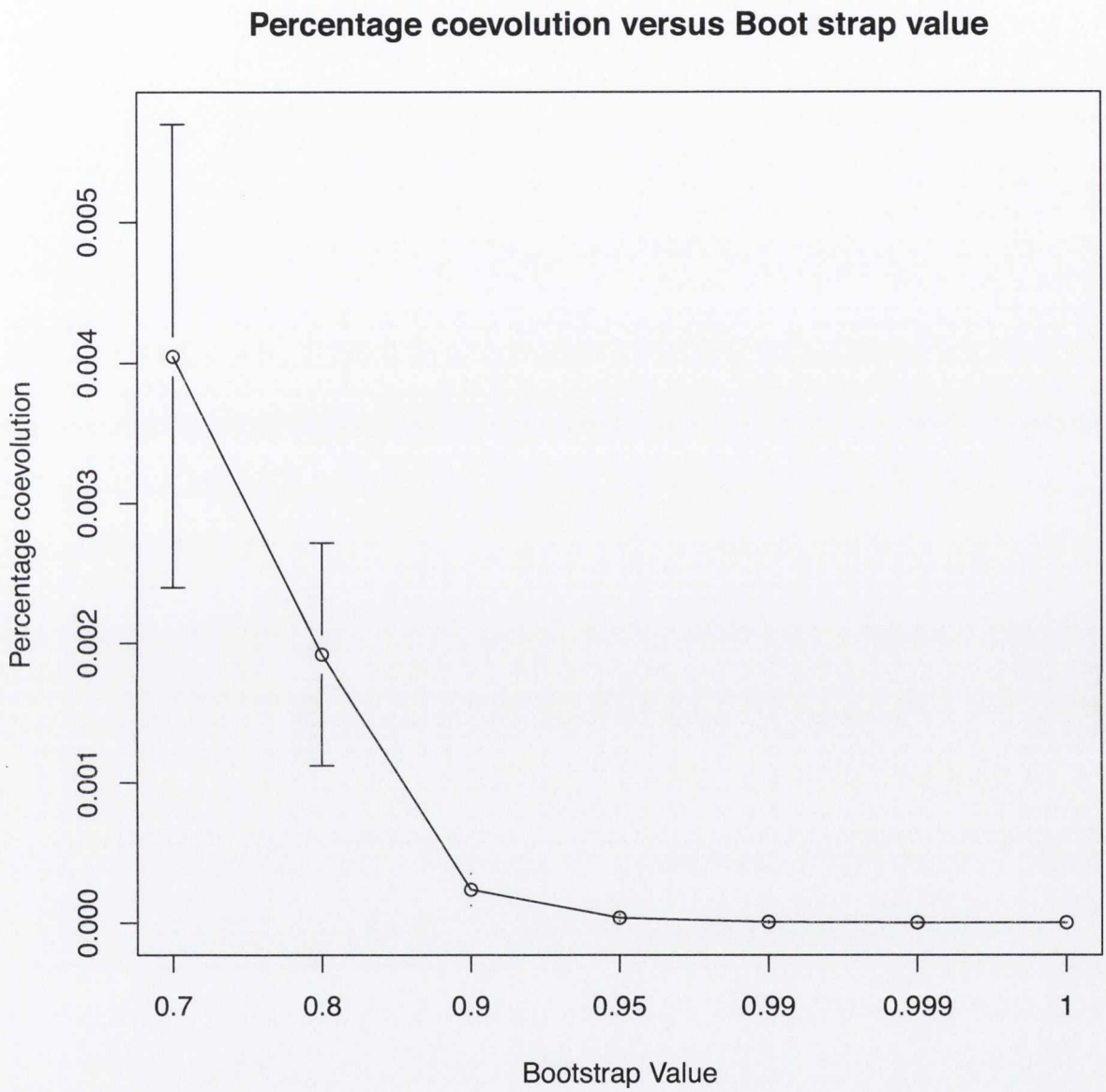

Figure 4.5: Susceptibility to false positives. This Figure represents the percentage of false coevolving pairs of sites detected by CAPS versus increasing bootstrap value. CAPS was run on 100 alignments simulated without selection pressures. The percentages of false positives here are extremely low and are shown to be even lower as the bootstrap threshold is increased. **Note:** The points with bootstrap values greater than or equal to 0.9 would have error bars but the standard error is too small for the graphing software to show them

ranking the pairs of alignments we used the percentage of sites coevolving between each pair of alignments for the CAPS analysis. In the case of Mirrortree we used the correlation coefficient returned by the software. The results are presented in Table 4.1, showing that each of the programs performed roughly the same. Again, we tested to see if bootstrapping improves the predictions of our software. Figure 4.7 shows a plot of the average value ranking of each pair of co-crystalized structures versus the bootstrap value. It is evident that bootstrapping improves the predictions, but the optimum bootstrap is smaller than that of the intra-molecular coevolution value(0.95).

We performed numerous correlation tests to see if there was a propensity for our software to perform better in one evolutionary circumstances over another. We tested the correlation of the rankings described in Table 4.1 for both programs versus the total distance on the phylogenetic trees, the absolute value of the distance on the trees, the number of sequences common to both alignments and the average identity of the alignments. The Pearson correlation coefficient which was most significant was the absolute value of the differences between the total distances on each tree. More clearly this is given by:

$$\epsilon = (T1 - T2) \tag{4.9}$$

where $T1$ is the total distance on tree one and $T2$ is the total distance on tree two.

This test returns +0.35 for CAPS and conversely -0.35 for Mirrortree. This suggests that CAPS performs better than mirrortree when the distances on the trees are very different and Mirrortree performs better than CAPS when the distances are similar. We show a graph in Figure 4.5 that displays bins of the ranking of each

Locus ID	Locus ID pair	CAPS ranking	Mirrortree ranking	Rule ranking
B0014	B2614	13	19.5	13
B0032	B0033	2	2	2
B0185	B2316	10.5	3.5	3.5
B0407	B0462	14	15	14
B0728	B0729	6	9	6
B0912	B1712	4.5	1	1
B1602	B1603	20.5	4.5	4.5
B1709	B1711	7.5	1.5	1.5
B1891	B1892	19.5	9	9
B2572	B2573	5.5	18	5.5
B3731	B3733	7.5	3.5	3.5
B3734	B3735	6	15.5	6
B4259	B4372	11	14.5	11
	Average	9.8	8.9	6.2

Table 4.1: A comparison of the CAPS and Mirrortree programs. The columns represent the average pairwise ranking of each gene with the definitive interactors. For example CAPS ranked B2614 as the 15th most highly coevolving protein for B0014; CAPS ranked B0014 as the 11th most highly coevolving protein with B2614, therefore the average of these values is 13. Of these definitive pairs of interacting proteins CAPS ranks 6 interacting pairs higher than Mirrortree and Mirrortree ranks 6 interacting pairs higher than CAPS. Both programs rank 1 pair equally. This Table illustrates that both programs are not very accurateunder certain conditions. We have proposed a method of drastically improving the predictions of protein-protein interactions by preferentially using one over program over the other depending on the rule proposed in section 4.4.3. The final column shows the scores obtained when implementing this rule.

pair of co-complexed proteins versus the experimentally verified partner. The graph shows results from CAPS, Mirrortree and a set of columns which is constructed by using the following rule: take the CAPS score if $\epsilon > 5$ the and Mirrortree if $\epsilon < 5$, where is the average value of . This rule drastically affects the prediction of interactions. It eliminates the worst predictions (the fifth column) and is an improvement on the prediction from either program. The assignment of an exact numerical value for which software should be used is not ideal. It would be much more pleasing to declare a rule which is grounded within an evolutionary context, however, we are not, as yet able to do so. This may be possible with the emergence of greater numbers of co-crystalized structures. In lieu of this we have presented this rule as a rough estimation of when to use CAPS or Mirrortree.

4.4.3 The effect of bootstrapping on inter-molecular coevolution

Figure 4.7 shows the effect of bootstrapping upon inter-molecular coevolution. The graph shows the average ranking of pair of co-complexed proteins in the data set under different runs of CAPS with altered bootstrap values (See average value given in Table 4.1). It shows that there is a steep decrease in the accuracy of predictions followed by a steep incline. This suggests that an ideal bootstrap cut-off exists (at least for this data set) which will optimize the results. This is not surprising since there is only so much phylogenetic signal that can be removed before all functional signal is removed also.

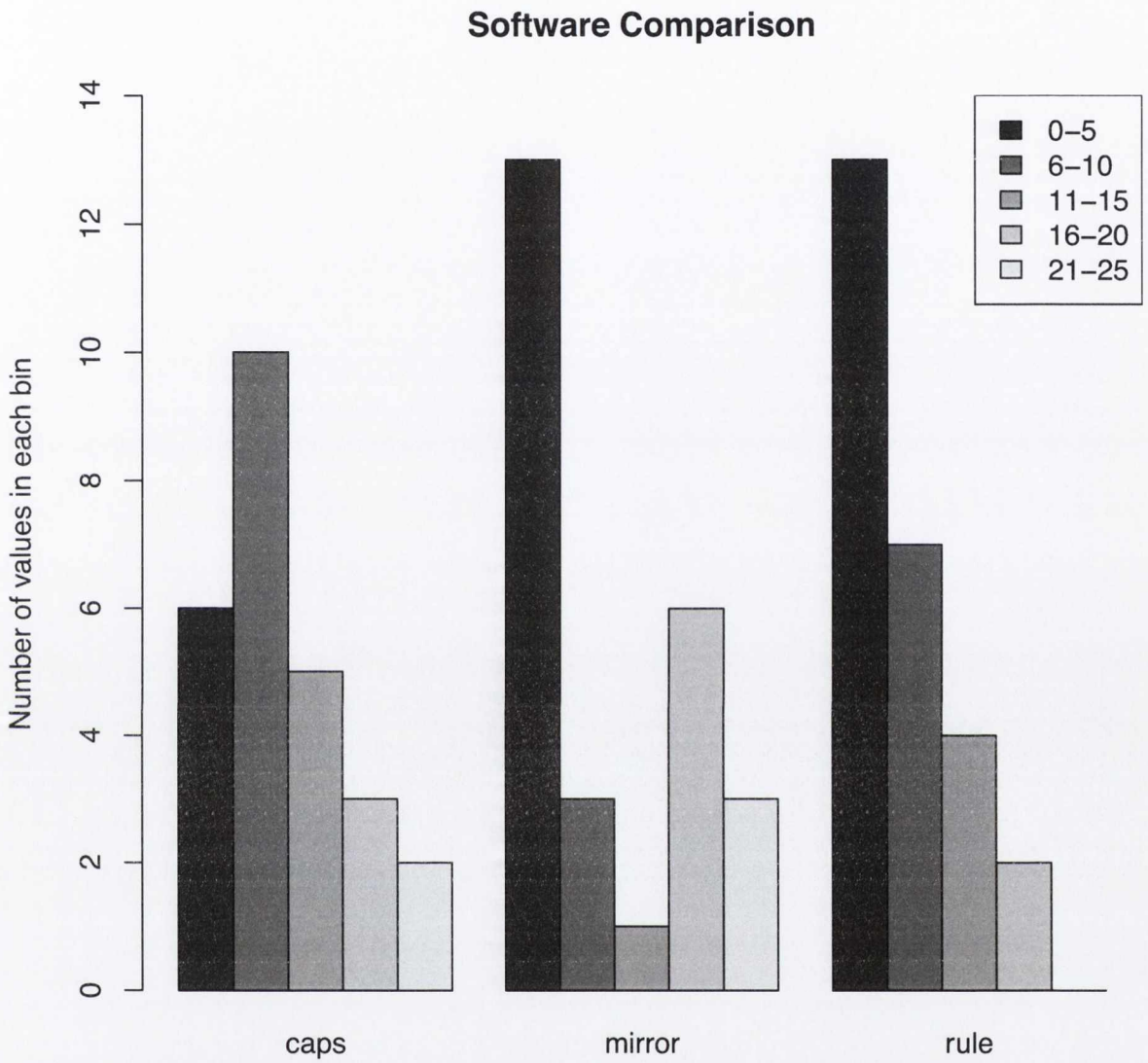

Figure 4.6: Graphical representation of PPI predictions in bins. Shown are the profiles of predictions of interactions through coevolution analysis. The ranking of each co-complex relative to its corresponding pair is binned into five categories. The CAPS software shows fewer predictions in the worst bins relative to Mirrortree but also fewer predictions than mirror tree in the best bins. According to the rule proposed above the third distribution eliminates those predictions in the very worst bins and pushes the distribution towards better predictions.

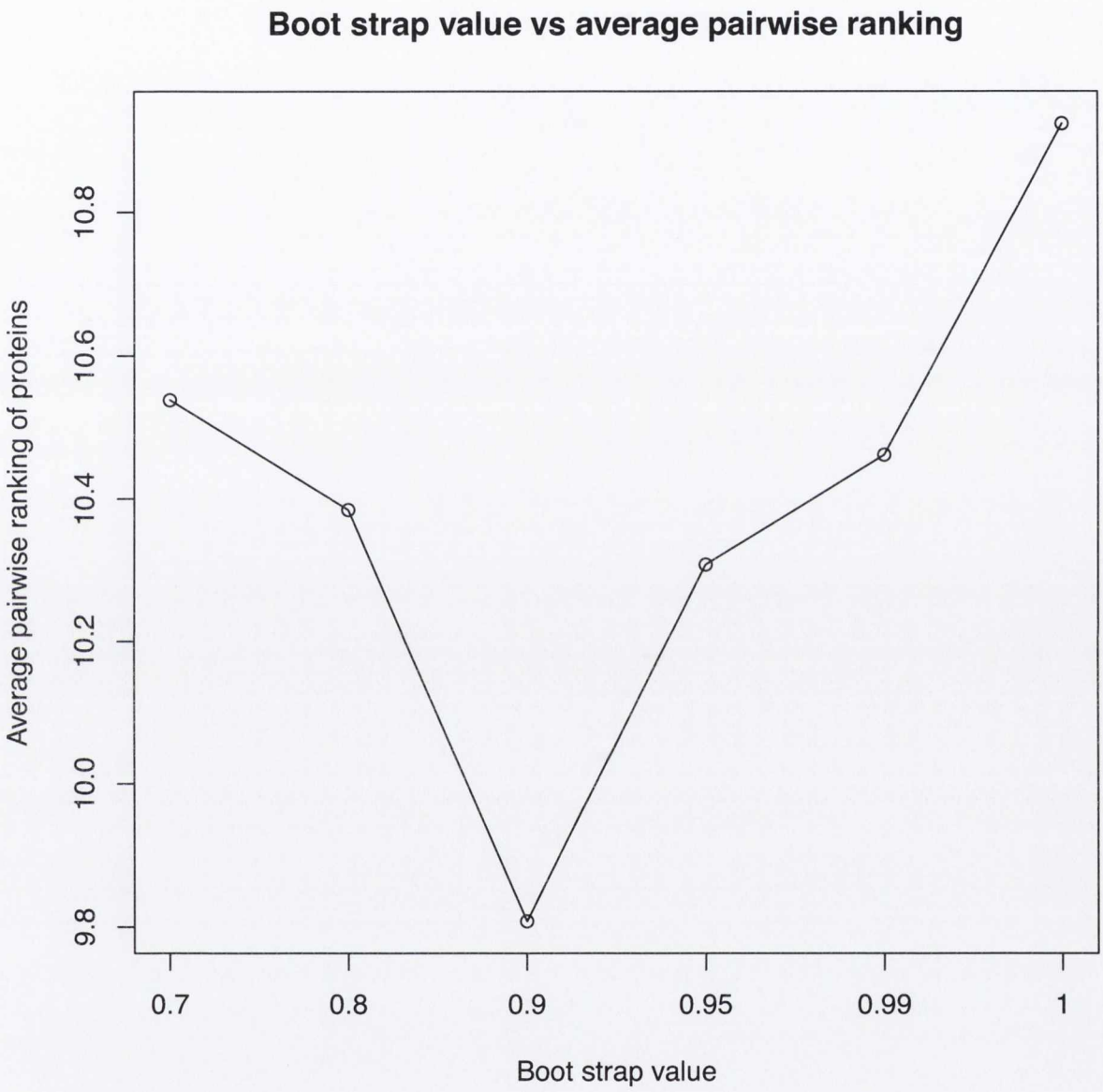

Figure 4.7: The effect of bootstrapping on PPI predictions. Shown is a graph that illustrates a possible ideal bootstrap value for inter-molecular coevolution. Six separate runs of CAPS were performed upon the data altering the bootstrap threshold. A bootstrap value of 0.9 is shown to be best.

4.5 Discussion

We have presented a novel model of analysis of molecular coevolution which can be applied either intra-molecularly or inter-molecularly. This method along with all other models which have been presented previously have aimed at disentangling the different parameters contained in Equation 4.1. Although two sites in a multiple sequence alignment have high numbers of correlated sites, it has been shown here that this can largely consist of stochastic and phylogenetic signal as opposed to functional, structural or interaction signal.

Our aim has been to remove the effects of $C_{phylogeny}$ and $C_{stochastic}$ from our model so as to address the remaining terms. In Section 4.3.3 we present an alteration to current methodologies which calculates ancestral sequences and uses these to circumvent any evolutionary blind spots which may be present without considering phylogeny. Our opinion is that the major contribution this step makes is removing $C_{stochastic}$ signal as is evident in Figure 4.4. This Figure shows that **prior** to bootstrapping there are extremely small levels of false positives. For obvious reasons the removal of stochastic signal is of utmost importance.

Removal of signal from phylogenetic sources is a somewhat more controversial problem. Mirrortree analyses are, after all based purely on phylogenetic signal and it has been shown both here and in other studies (Pazos and Valencia, 2008; Pazos et al.., 2005) that it can be extremely useful. The basis of these models is grounded in a theory that if two genes have extremely similar evolutionary history then they are most likely related. This assumption does indeed work in many cases, however, we propose two cases that illustrate this view is oversimplified. Firstly, many interacting proteins do so through binding domains which are small relative to the entire protein; this suggests that other regions of these proteins are free to

evolve under a separate set of selection constraints. Secondly, if a duplicated gene undergoes subfunctionalization in one region and now has the ability to interact with a new partner; this pair of interacting proteins are now constrained (theoretically) under similar conditions. In this case the newly evolved section of the gene may have a similar pattern of coevolution but the remaining section of this gene may have an extremely different evolutionary history.

In Section 4.4.3 we compared the results of analyses carried out by both CAPS and Mirrortree on a set of 13 experimentally confirmed interactors. We found that neither software can reproduce acceptable results all of the time. We propose a method to choose which software should be used dependent on the difference between the total distances on the phylogenetic trees considered; this rule drastically improves predictions. We have shown that CAPS performs better than Mirrortree in cases when the total length of the two phylogenetic trees involved are very different; correspondingly Mirrortree performs better when the lengths of the phylogenetic trees involved are very similar. This is compelling evidence that our method is indeed removing the effects of phylogenetic coevolution.

The identification of both site specific intra-molecular and inter-molecular coevolution have lead to discoveries in diverse areas including, myoglobin (Pollock *et al.*, 1999), ribosomal RNA (Dutheil *et al.*, 2005) and HIV (Travers *et al.*, 2007). These studies represent excellent examples of the insights that coevolution analysis can have into important proteins in the biological sciences. Our new model will provide further insight into the effects molecular coevolution has upon organisms and their diversity. Our CAPS web interface also offers an excellent resource to the research community.

Chapter 5

Investigation of the effect of chaperone buffering upon its clients

5.1 Introduction

Evolvability is the ability of an organism to undergo adaptive evolution. Wagner defines evolvability thus: a system is evolvable if mutations in it can produce heritable phenotypic variation (Wagner, 2008). From this definition we can extrapolate that certain systems can be more evolvable than others depending on the selection constraints acting on the system. In cases of relaxed selection pressure we can postulate that there is greater evolvability. Evolvabillity is not restricted, however, to point mutation accumulation. Indeed a bacterium can be said to be highly evolvable because of its ability to acquire genes according to lateral gene transfer, upregulated mutation rates and sexual recombination. These can all lead to the evolution of heritable phenotypic variation.

Robustness is the ability of an organism to accumulate change without fitness effects. It is therefore intrinsically linked with evolvability. Wagner defines robustness thus: a biological system is mutationally robust if its function or structure persist

after mutations in its parts. This definition implies that some genes or genomes may be more or less susceptible to the deleterious effects of accumulated mutations. Why are some genes more susceptible or less susceptible to mutations? The answer has been proposed to lie in molecular "capacitors" which buffer the effects of deleterious mutations. The effect of buffering of deleterious mutations was first demonstrated by Rutherford and Lindquist (1998). While working on *Drosophila* they reported that reduced Hsp90 activity causes developmental abnormalities. They concluded that Hsp90 was "buffering" the effects of deleterious mutations in the genome. A similar effect was noted by Queitsch *et al.*, (2002) when studying *Arabidopsis thaliana* and additionally, in *Saccharomyces cerevisiae* by (Cowen and Lindquist, 2005).

In addition to the seminal work presented by Lindquist and others on Hsp90 there have been many studies carried out on the chaperonin GroEL. This heat shock protein assists in the folding of proteins. GroEL has been shown to allow endosymbiotic bacteria cope with the high levels of genetic drift through over-expression (Moran, 1996), to recover the fitness of *E. coli* subjected to evolutionary bottlenecks by over-expression (Fares *et al.*, 2002) and also to buffer its clients (Williams and Fares, 2010). Even though there have been many excellent studies carried out in the area of buffering of mutations it remains unclear whether these buffering effects can lead to functional innovations.

Here we present an analysis of 85 complete genomes from gamma-proteobacteria which are systematically separated into groups according to client class (Kerner *et al.*, 2005) and essentiality ((Hashimoto *et al.*, 2005; Kato and Hashimoto, 2007). These genomes were selected, firstly, because they represent a class of bacteria which is well represented both in numbers of sequenced genomes but also in terms of the variety of phenotypic traits. We perform functional divergence and coevolution analysis upon this data set making use of CAFS (Clustering Analysis of Functional

Shifts) and CAPS 2.0 (Coevolution Analysis of Protein Sequences), described in Chapter2 and 4, respectively. Here we aim to develop an understanding of the dynamics by which GroEL clients and non-clients evolve. We contrast the different selection pressures and highlight possible lineage specific adaptations which may have arisen through chaperone buffering.

5.2 Materials and methods

5.2.1 Data

All available crystalized protein structures for gamma-proteobacteria were downloaded from the Protein Data Bank (PDB, www.rcsb.org/pdb). This comprised 1090 entries, representing 20-25% of the *E. coli* proteome. We chose this dataset with the explicit intention of presenting results on a structural level though the results from these data have not yet proven useful. Protein sequences homologous to the structure-associated sequence were obtained by reciprocal Blast searching of 85 complete gamma-proteobacterial genomes. Hits with an E-value smaller than 10^{-4} were retained. Sets of homologs were aligned with ClustalW (Thompson *et al.*. 1994) using default parameters. The quality of these alignments was inspected manually. We assigned categories to each of the alignments according to essentiality. Essentiality data was obtained from the SHI-GEN profiling of *E. coli* chromosome database (Hashimoto *et al.*, 2005; Kato and Hashimoto, 2007) where essentiality is assigned if strains carrying a null mutation cannot grow under any conditions. We also assigned tags of GroeEL client classes, which have been identified by Kerner *et al.*, (2005). There were 252 client proteins, 85 of which were found to be obligate clients for which GroEL was absolutely necessary to fold. These clients are split into

three classes (I, II, III). Class one and two clients are only partially dependent on GroEL, class three proteins are described as obligate clients.

5.2.2 Chi-squared analyses

In this study we perform multiple chi-squared tests. The chi-squared statistic is given as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5.1}$$

Where O_i is the observed frequency of the quantity being tested, E_i is the expected frequency and n is the number of possible outcomes of each event (in our case it was always 2). This test is a "goodness of fit test", it analyses whether an observed distribution differs from that of an expected distribution. The result from either of these methods is compared against a chi-squared distribution for the assignment of a P-value.

5.2.3 Coevolution and functional divergence analyses

Analysis of intra-molecular coevolution was performed with an alpha threshold (type I error) of 0.05 and a site-based bootstrap threshold of 0.99 as we discovered to be associated with the best functional results in Chapter 4. Functional divergence analysis was performed with the use of the CAFS software presented in Chapter 2 with an alpha (type I error) of 0.05. Phylogenetic trees for both of these analyses were constructed using RAxML (Statamakis, 2006) with 100 bootstraps and models chosen according to those predicted by Prottest (Abascal *et al.*, 2005).

5.3 Results

Functional divergence and intra-molecular coevolution analysis were performed upon 1090 multiple sequence alignments of gamma-proteobacteria. Below we present the results of these analyses contrasting the relative effects of GroEL buffering on clients versus non clients, the affect of essentiality upon protein evolution and special cases of lineage specific functional divergence.

5.3.1 GroEL buffering of mutations in clients versus non-clients

GroEL buffers mutations in its clients as described in the introduction above. Figure 5.1 presents the difference between site-specific functional divergence events in GroEL clients and non-clients. These analyses were carried out with the CAFS software presented in Chapter 2. The length of GroEL clients is significantly greater than that of non-clients and we therefore normalized the number of functional divergence sites by the length of each protein in all of the subsequent tests. Figure 5.2 presents the differences in levels of intra-molecular coevolution analysis performed using the CAPS software presented in Chapter 4. The profile is similar but perhaps more striking than that of the functional divergence graph. We performed a t-test upon these data to assess significance. We found that there is evidence to suggest that GroEL clients have both significantly more sites under functional divergence (P-value = 0.02) and a modest significance that there are also greater levels of intra-molecular coevolution (P-value = 0.0516). Taken together this implies that the buffering effects of GroEL are allowing for functional adaptation. We analyzed whether there were differences in the profiles of obligate clients versus clients with

only partial dependence. We found that there was a trend observed but statistical tests performed using a series of chi-squared tests did not return significance. We have included some of these graphs in supplementary materials (Appendix B).

5.3.2 Comparison of functional divergence and coevolution between essential and non-essential genes

Essential genes are expected to evolve slower and that any mutations therein will be highly deleterious thus they will undergo strong levels of purifying selection. Figure 5.3 and 5.4 compare levels of functional divergence and intra molecular coevolution between essential and non-essential genes. We found that after a t-test there was a significantly greater number of functionally divergent (P-value = 0.0149) sites in essential genes. There were greater levels of coevolution but the significance was marginal (P-value = 0.12). At first this may seem slightly contrary to the expectation, one expects very few mutations to accumulate in essential genes. When essential genes pick up a mutation which is beneficial, however, it is likely to have a large effect on fitness. This fitness effect may be fixed very quickly in a lineage through strong positive selection. The lower significance of the coevolution analysis may be explained by the fact that a slightly deleterious mutation is sometimes needed prior to a second mutation which restores (or even improves) the fitness of a gene.

5.3.3 Clustering analysis of functional divergence events

We performed a full functional divergence analysis upon all sequences in our dataset using the CAFS software described in Chapter 2. We tagged each of the alignments

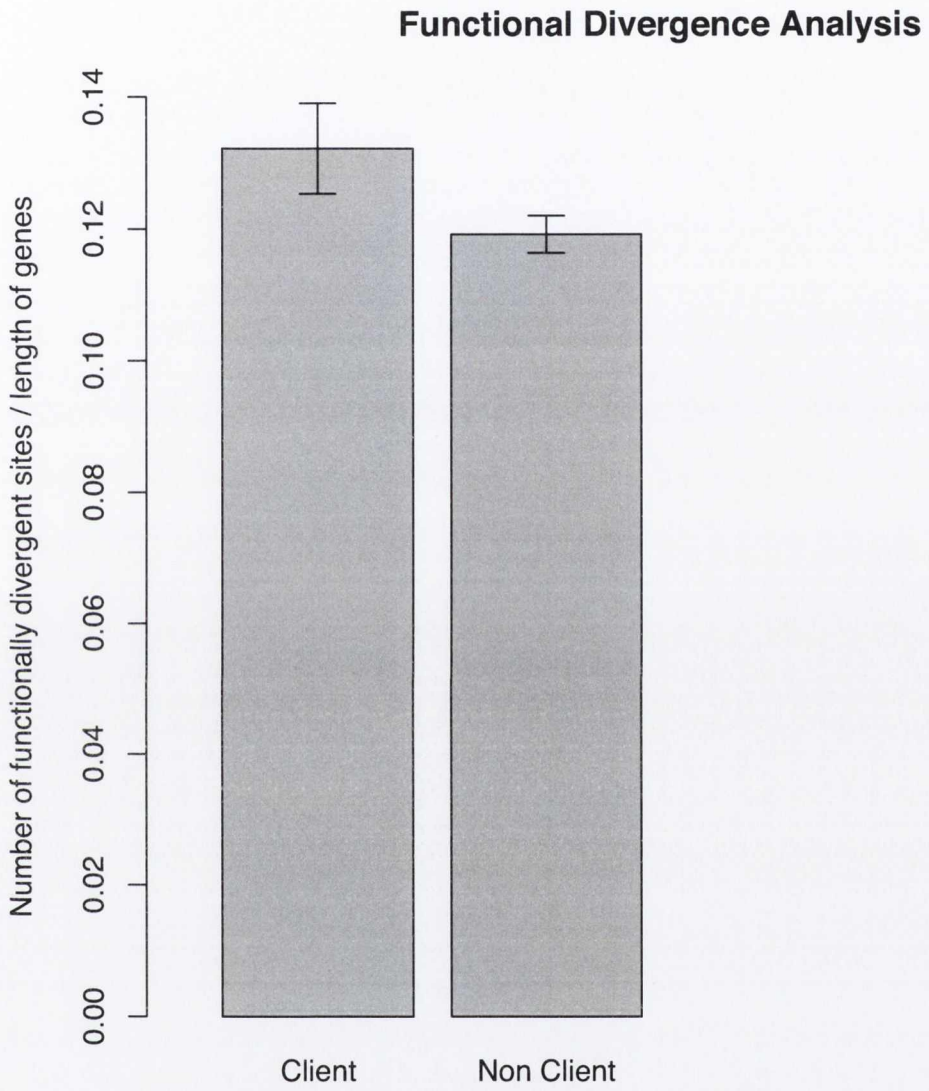

Figure 5.1: Buffering of functional divergence sites by GroEL. This Figure shows the profile of site specific functional divergence in GroEL clients versus non clients. This profile and a corresponding t-test (P-value = 0.018761) suggests that GroEL is buffering site specific functional divergence in clients.

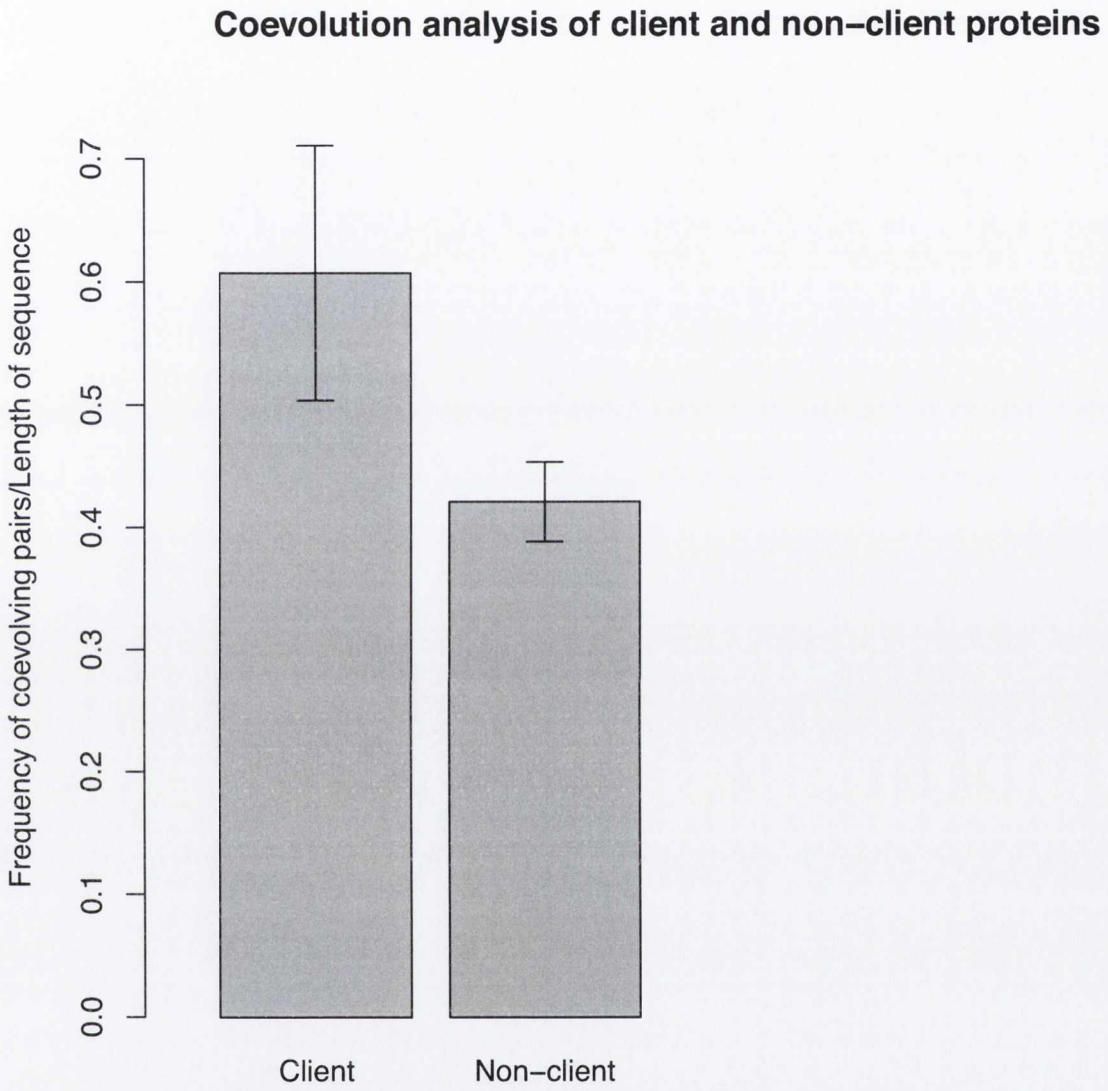

Figure 5.2: Buffering of coevolving sites by GroEL. This Figure shows the profile of site specific intra-molecular coevolution analysis in GroEL clients versus non clients. This profile and a corresponding t-test (P-value = 0.0516) suggests that GroEL is buffering intra-molecular coevolution in clients

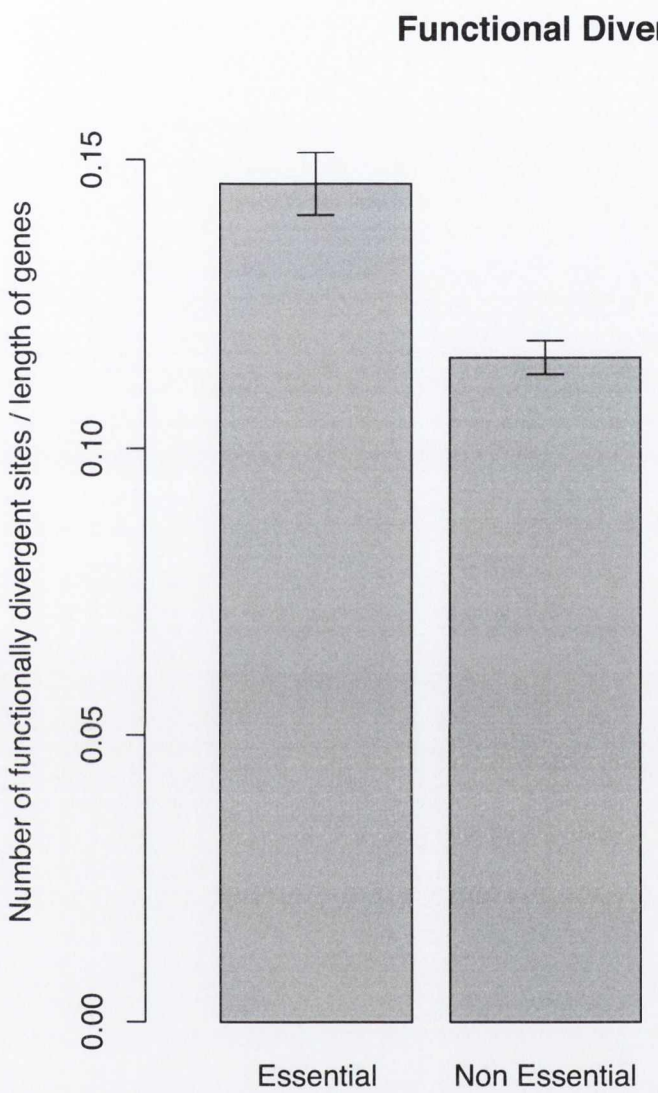

Figure 5.3: Profile of functional divergence sites in essential and non-essential genes. This table shows the profile of site specific functional divergence in GroEL essential genes versus non essential genes. This profile and a corresponding t-test (P-value = 0.0149) suggests that there are significantly more sites of functional divergence in essential genes. This is somewhat surprising but may be explained by the fact that in the very rare occasions when a mutation in an essential gene is advantageous it is fixed in the population very quickly by high levels of positive selection.

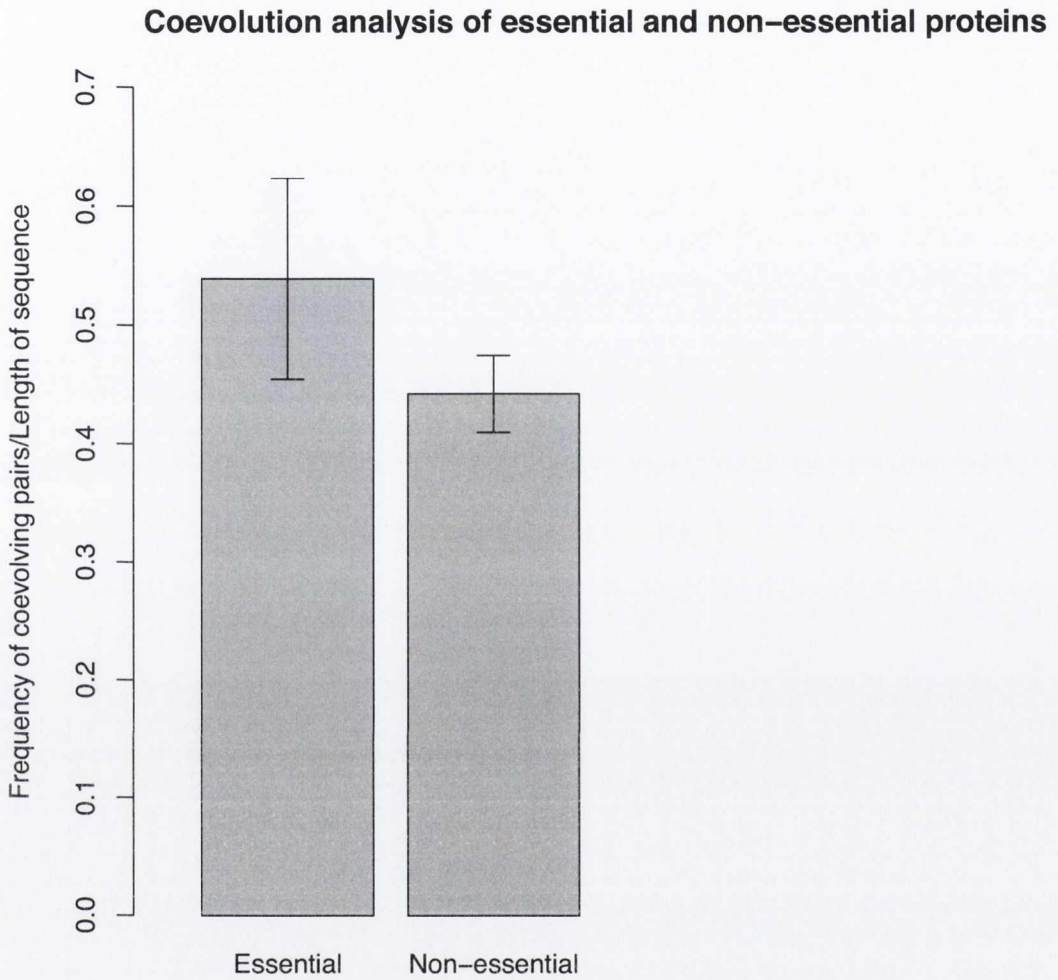

Figure 5.4: Profile of coevolving sites in essential and non-essential genes. This table shows the profile of intra-molecular coevolution analysis in GroEL essential genes versus non essential genes. This profile is similar to that of the functional divergence analysis, however, it is less pronounced and the corresponding t-test (P-value = 0.12) reflects this. This suggests that similarly to the functional divergence conjecture this may be explained by rare occasions of coevolving sites which are beneficial. In the case of coevolution, however, a slightly deleterious mutation is sometimes needed prior to a second mutation which restores (or even improves) the fitness. In this case only deleterious mutations which are extremely slight in fitness effects would be tolerated, hence the less significant profile.

with a level of dependency (I-III) of client proteins and whether the protein was essential or not. Figure 5.5 shows the heat map created using the R software package using the output from CAFS. In the heat map blue represents impoverishment for functional change and red represents enrichment for functional divergence. Each cell of the heat map represents a "goodness-of-fit" chi-squared test where the expected values are obtained by getting the average value of functionally divergent sites over all cells.

A clear division into two groups is visible: the top half of the Figure is dominated by functional divergence enrichment and the lower half is impoverished. The lower half represents *E. coli*, *Shigella* and *Salmonella* strains. These strains are mostly host-associatiated and were discussed in Chapter 3 as being impoverished for functional divergence. We performed a chi-squared test upon the numbers of host-associated bacteria in the lower half versus the upper half to test if there is a propensity for change in host-associated bacteria. As in Chapter 3. this was significant again (P-value$<$ 0.0001) with an even higher P-value. Whilst this is unsurprising it is confirmation of previously known data and hence reassurance that our tagging system has not been erroneous.

Firstly, the area marked in green is comprised of five *Acinetobacter* and two *Psychrobacter* and one *Cellvibrio* strains. The striking profile here is that these strains are the only strains enriched for functional divergence in the essential and client class I group. *Acinetobacter* strains are pathogenic and cause pneumonia and urinary tract infections. These bacterial species are known to be multi drug resistant but the full extent of resistance is unknown (Adams *et al.*, 2008). Many of these species gained drug resistance from laterally transferred genes but it has been shown that this is not their only mechanism for drug resistance (Adams *et al.*, 2008). Investigation into the *Psychrobacter* and *Cellvibrio* strains within this

section reveals that all but one are drug resistant. We found eight genes which are responsible for the enrichment of functional divergence in this category. One of these genes (rpoA) is part of an RNA polymerase which is the target of *Rifampicin.*

Mutations in *rpoB* (binding partner of *rpoA*) have been shown to cause resistance to rifampicin (Jin & Gross 1988). Of the other seven genes which caused this enrichment, two pairs of interacting proteins were found. Little is known of specific functions of these genes within *Acinetobacter*, indeed the only citation of significance to be found for one strain was the initial sequencing project. Since this group is composed of all (bar one) drug resistant strains and shows the greatest levels of functional divergence in this category, we propose that GroEL may be buffering mutations which lead to drug resistance in the genes involved, especially considering the information describing *rhoB.* Interestingly the *Cellvibrio japonicus* strain was alone in this group in having functional divergent sites in *grpE*, another heat-shock protein but also a client of GroEL.

5.3.4 Zooming in on functional divergence profiles

Since the majority of signal in the heat map in Figure 5.5 is drowned out by phylogeny we chose to create two further heat maps, one from the top portion of the graph representing the highly functionally divergent species and one from the bottom portion representing mostly host-associated bacteria. This approach allows us to describe some of the finer details of these strains in terms of functional divergence.

Figure 5.6 represents the top half of Figure 5.5. Here we have recalculated all chi-squared tests involved in Figure 5.5 upon the strains in the top portion of the heat map only. This allows us to see a gradient of differences between the strains in this subset. Firstly, in Figure 5.5 it can be observed that some *Psuedomonas*

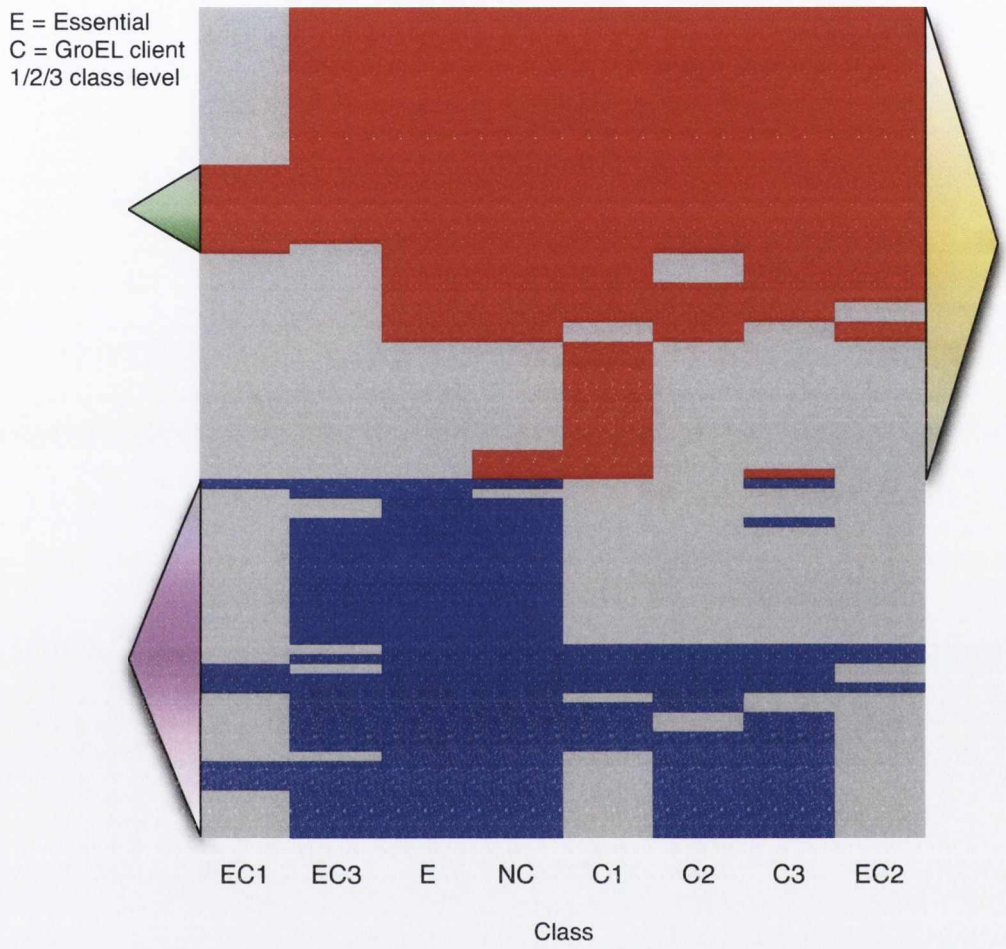

Figure 5.5: Heat map of functional divergence profiles tagged with client/essential categories. Presented here is a heat map of all species tested in the study. Each cell represents a chi-squared test which was either significant and over expectation (red), significant and under expectation (blue) or not significant (grey). It can be seen that there is a stark division between the top and bottom sections. This division mainly represents phylogeny with the *E. coli*, *Shigella* and *Salmonella* strains on the bottom half (magenta) of the heat map and *Pseudomonas*, *Acinetobacter*, and *Yersinia* species on the top half (yellow). The section marked in green represents a group of *Acinetobacter*, *Psychrobacter* and *Cellvibrio* strains, seven of these eight species are drug resistant suggesting a role for GroEL buffering of class 1 clients.

(marked with a purple triangle) strains are enriched for functional divergence whilst others are not. Investigation of the underlying genes reveals that genes *tpx* and *rbsB* are responsible. Somprasong *et al.*.(2012) showed that *Psuedomonas aeruginosa* is protected against hydrogen peroxide toxicity by *tpx*. The gene *rbsB*, was found to be differentially expressed in lung tissue of cystic fibrosis patients infected with the strain *Pseudomonas aeruginosa* (Sriramulu *et al.*, 2005). This explains differences between the functionally divergent patterns of *Pseudomonas aeruginosa* but strains, such as *Pseudomonas fluorescens* also have this profile and do have differing genes but we are yet to assign biological relevance to these.

Figure 5.7 represents the bottom portion of Figure 5.5. Again we have recalculated the chi-squared tests associated with Figure 5.5; this time removing species from the top half of the heat map. This heat map shows many more subtle differences. Firstly, the strain *Shigella sonnei Ss046* (marked with a pink triangle) is the only *Shigella* strain which is enriched for essential clients (class II). This strain is the cause of the enteric infectious disease *shigellosis* and is drug resistant bacteria. Genes involved in the enrichment of functional divergence in this strain were *gyrA* and *acrE*. Mutations in *gyrA* have been implicated in the resistance to quinolones (Dimitrov *et al.*, 2010). Sites discovered here did not correspond with the sites identified by Dimitrov *et al.*, however theirs was not a fully exhaustive study. The protein *acrE* has been shown to inhibit lipophilic inhibitors and was shown to affect drug susceptibilities (Ma *et al.*, 1993).

Shigella flexneri 5 str. 8401 (marked with a yellow triangle) is the only strain of *Shigella* which has essential class I and non-essential class II genes enriched for functional divergence. *Shigella flexneri* cause disentry and pose a significant threat to human health. The enrichment in essential class I genes is caused by sites detected as functionally divergent in the gapA gene. This gene codes for the protein gapdh,

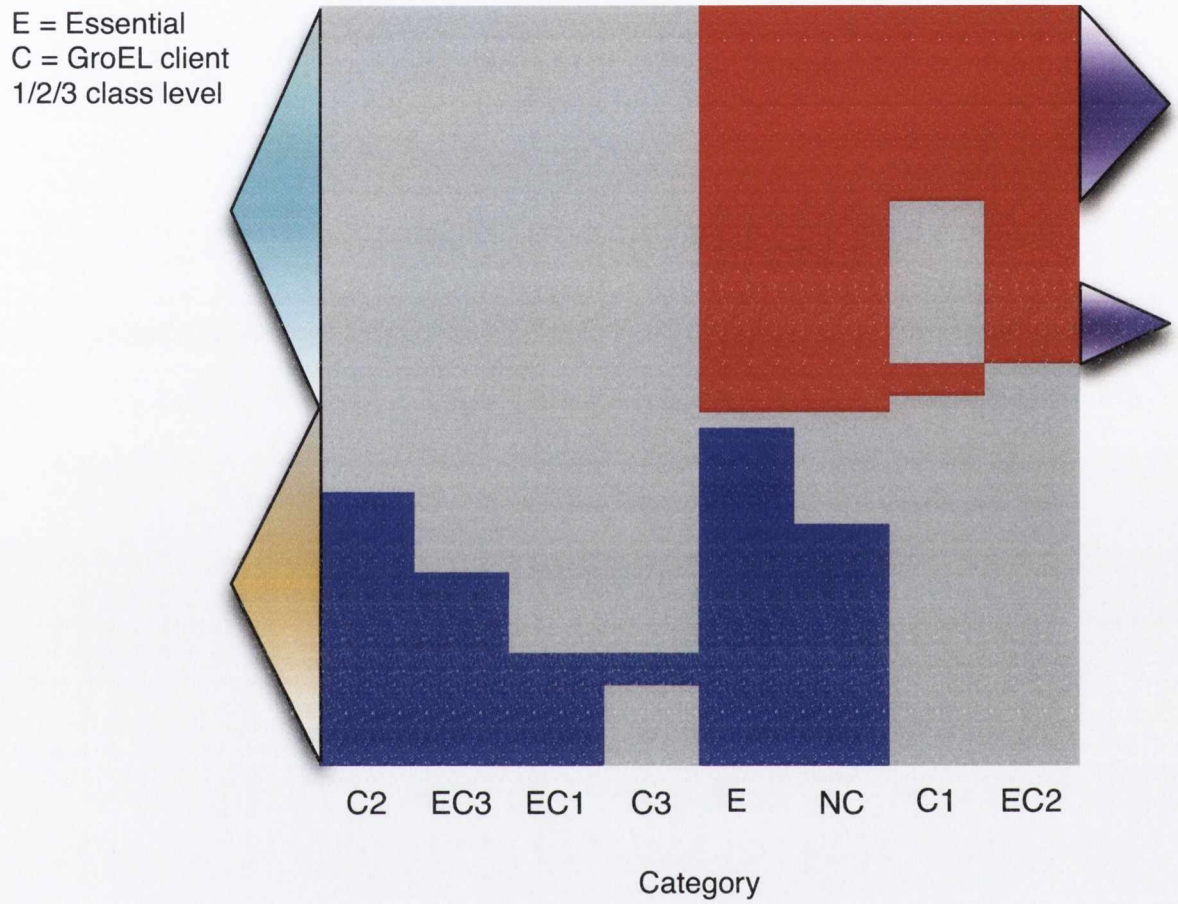

Figure 5.6: Sub heat map of top portion of Figure 5.5. This heat map represents the top half of Figure 5.5 reanalyzed with the bottom half removed. This shows an asymmetric profile between *Pseudomonas* and *Acinetobacter* (cyan) species on the top and mostly *Yersinia* (orange) species on the bottom. Interestingly there are some species of Pseudomonas which are not enriched for client class I genes while the remaining are, these are identified with purple triangles. We attribute this to two genes, *tax* and *rbsB*.

a protein which exhibits some heat-shock characteristics (Charpentier *et al.*, 1994), additionally, under conditions causing low growth rates cells may need gapdh activity to metabolize glucose from an alternative carbon source, such as lactose or sucrose (Toyoda *et al.*, 2008).

5.4 Discussion

The buffering of deleterious mutations by GroEL has been the focus of many studies (as discussed in the introduction), both experimentally (Moran, 1996; Fares *et al.*, 2002) and computationally (Williams and Fares, 2010; Warnecke and Hurst, 2010). Williams and Fares have shown that GroEL does indeed buffer mutations in its clients; and in the case of Warnecke and Hurst they have shown that GroEL buffering and codon usage may be two ways in which organisms may limit misfolding errors. On undertaking this analysis we wanted to build on this work by identifying lineage specific adaptations and the types of mutations which GroEL buffers.

In section 5.3.1 we showed differential profiles of site-specific functional divergence and intra-molecular coevolution analysis in GroEL clients versus non-clients. These profiles suggest that GroEL is buffering this type of mutations. In the case of functional divergence there is strong evidence that the buffering of clients by GroEL is a driver of functional innovation and therefore lends to the evolvability of the organisms involved. The buffering of coevolving sites suggests that sites within a protein which are functionally constrained have the possibility of exploring sequence space mediated by GroEL.

The essentiality of proteins makes them an extremely interesting group to study; when a null mutation is performed upon one of these genes it is known to have

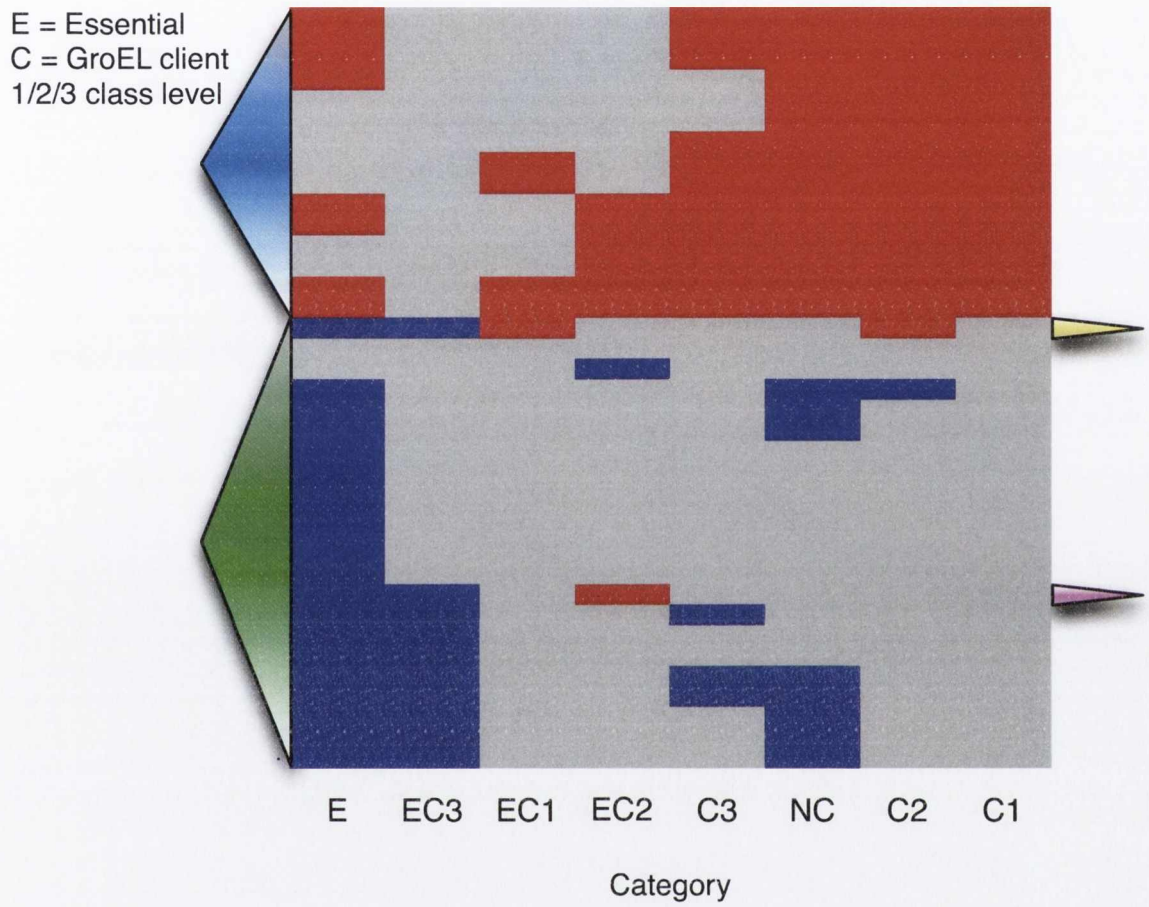

Figure 5.7: Sub heat map of bottom portion of Figure 5.5. Here we present the bottom half of Figure 5.5 in a heat map which has been reanalyzed without the strains from the top half of the original. The green triangle represents *E. coli* and *Shigella* strains. The blue triangle represents *Salmonella* strains. The yellow triangle represents the *Shigella flexneri 5 str. 8401* strain a virulent strain. The pink triangle represents the *Shigella sonnei Ss046* a drug resistant strain. Genes involved in the enrichment of *Shigella sonnei* are known to confer drug resistance to the strain.

fatal effects upon the organism. The initial expectation is that mutations within essential genes could result in fatality. We found that these genes were significantly more likely to exhibit site specific functional divergence. This somewhat surprising result can be explained as follows: if a gene is extremely important and it gains a deleterious mutation then it will most likely be removed from the population by purifying selection. On the other hand, in the extremely rare occasions that a mutation is advantageous this will impart a large fitness advantage upon the organism which may be fixed in the population through high positive selection. The profile of coevolution analysis upon essential and non-essential genes is similar but somewhat "dampened". This could be due to the lower statistical probability of a pair of coevolving sites in essential genes. Molecular coevolution involves a pair of correlated mutations. The first of these mutations may be neutral but in most cases it is slightly deleterious. A second mutation may restore or improve on the initial fitness. In the case of essential genes, however, even slightly radical mutations could have a significant fitness decrease resulting in purifying selection. Therefore any increases in fitness that coevolution may be able to facilitate through coadaptive mutations are more rare events.

Profiles of functional divergence events in tagged proteins proved a rich source of evidence for functional innovation through chaperone buffering. Our clustering analysis which makes use of multiple chi-squared tests, visualized as a heat map, returns evidence of site specific mutations in *Acinetobacter* strains which are essential class I clients of GroEL. One of the genes (*rhoA*) which was identified with sites under functional divergence was responsible for imparting *Rifampicin* drug resistance upon the strains. In *Cellvibrio japonicas* we found sites under functional divergence in *grpE* another heat-shock protein but also a client of GroEL. Another heat shock related protein (*gapA*) was found to cause enrichment of functional divergence en-

richment in *Shigella flexneri*. This protein is a class I essential client which has been shown to facilitate glucose metabolism through alternative food sources. This raises the question: is GroEL buffering mutations within another heat shock mechanism which allows that system adapt to add new clients? This theory, while an interesting thought is, however, at this moment merely conjecture. An analysis of functional divergence could possibly be performed with two functional classifications. The first classification being GroEL client tags and another which assigns tags according to clients of other heat shock proteins, such as Hsp70 and Hsp90.

Fine tuning our functional divergence analysis in Section 5.3.4 revealed more interesting profiles with peculiar biological properties. We discovered genes in *Pseudomonas* under strong levels of functional divergence. These genes have been shown to protect these species from toxicity (*tax*) and have been implicated in the adaptation to inhabiting lung tissue through differential gene expression in cystic fibrosis patients. These unusual profiles of functional divergence enrichment have uncovered biological adaptations in important genes through the application of statistical tests and protein classification. We feel that these examples are excellent evidence of buffering of client genes leading to functional innovation.

Chapter 6

Investigating the evolution of slightly deleterious mutations in a mutation-accumulation experiment: The role of molecular chaperones in evolution

6.1 Related manuscript

Investigating the evolution of slightly deleterious mutations in a mutation-accumulation experiment: The role of molecular chaperones in evolution. Mario X. Ruiz-Gonzalez*, Christina Toft*, **Brian E. Caffrey*** and Mario A. Fares. *In preparation.*

*denotes equal contributions.

6.2 Introduction

Understanding the forces which dictate microbe evolution is a complex problem rooted within the comprehension of the mutation rate, gene duplication, robustness, evolvability and all other attributes which affect evolution. Advances in our understanding of these areas has lead to much sought-after knowledge of how microbes adapt to new ecological niches and the relative effects these properties have on population genetics. Pioneering long-term experimental evolution carried out by the Lenski lab over the last twenty years have shed much light upon the evolution of new traits. Among their observations they have discovered the gain in ability of an *E. coli* strain to metabolize citrate (Blount *et al.*, 2008), measured the mutation rate (Wielgoss *et al.*, 2011) of *E. coli* and have estimated that fewer than 100 point mutations reached fixation in a population after 20,000 generations (Lenski, 2004). These advances sparked the interest of many scientists and opened the door to new questions and further exciting research that can be carried out on the ever-present *E. coli* bacteria.

Much evolutionary research has focus upon molecular chaperones. Chaperones are ubiquitous proteins which assist in the assembly of oligomeric complexes, transportation, and mediate the proper folding of proteins. Many proteins follow Anfinsen's dogma (Anfinsen, 1973) and fold independently of chaperones but some aggregate after leaving the ribosome and can be toxic to the cell. Many chaperones buffer the effects of environmental stress (e.g. heat-shock proteins) but have also been shown to buffer the effects of deleterious mutations (Fares *et al.*, 2004; Rutherford, 2003; Williams & Fares, 2010). GroEL has also been shown to rescue heat-sensitive mutants (Van Dyk *et al.*, 1989) and genes subjected to evolutionary bottlenecks (Fares *et al.*, 2002). Furthermore, the buffering capacity of molecular chaperones has been

proposed as facilitator of adaptive evolution (Rutherford and Lindquist, 1998; Fares et at 2002; Tokuriki and Tawfik; Lindquist, 2010).

Rutherford and Lindquist (1998) demonstrated that deleterious mutations in signaling proteins could be buffered by the effect of Hsp90. When Hsp90 activity was lowered *Drosophila* species developed abnormalities. The cornerstone of this work was to demonstrate that when artificial selection of abnormal phenotypes were placed together in populations to create a genome with multiple abnormalities the restoration of Hsp90 function could no longer suppress all variants.

The chaperonin GroEL/GroES is an extensively studied complex and was proposed to be buffering the effects of high levels of genetic drift (Moran, 1996) in endosymbionts. Moran proposed that the over-expression of GroEL/GroES in endosymbionts enabled the cell to maintain function despite the increase in deleterious mutations. This theory garnered support in the above mentioned study (Fares *et al.*, 2002) where *E. coli* was subjected to strong genetic drift which resulted in loss of fitness. This fitness was restored by over-expression of GroEL. In addition to these experimental approaches there have been computational analyses which have shown that GroEL clients experience weaker selection than non-clients for translationally optimal codons suggesting less need to finely regulate the mistranslation of clients (Warnecke and Hurst 2010). The expected corollary to the pioneering work by Rutherford and Lindquist (1998), Moran (1996) and others suggests that the evolutionary rate of GroEL clients should be higher than that of non-clients. This was initially found to not be the case, however, when accounting for the relative importance of the GroEL client proteins agreement was found with these early studies by Williams and Fares, (2010) and in this thesis (Chapter 5).

Here we present a computational analysis and comparison of two *E. coli* strains

which have been subjected to increased mutation rate through the deletion of the mutS gene and also to increased genetic drift due to evolutionary bottlenecks. Additionally, one strain was grown at 37 degrees and the second was constrained to grow at constant heat-shock of 48 degrees. We analyze the asymmetric profiles of mutations accumulated in both strains with respect to disparities between transitions and transversions, preferential codon substitutions, GC bias, functional categories of affected genes and GroEL clients versus non-clients.

6.3 Materials and methods

6.3.1 Experimental design

We evolved an ancestral *E. coli* K-12 strain under genetic drift and two conditions (37 and 48 degrees of temperature) to understand the fixation dynamics of slightly deleterious mutations in the bacterium genome. A strain lacking the repair gene mutS was used to increment the fixation rate of mutations. Strains lacking the mutS gene are predicted to have an increased mutation rate of between 100 and 1000 fold compared to wild-type. Two mutation accumulation (MA) lineages of the strain were serially passaged onto YPD by repeated streaking, each passage resulting from re-streaking a single colony. Re-streaking was carried out every 24 h for 100 days. Each lineage was passaged 100 times, which resulted in an estimated 2200 generations in total. ($\sim$ 22 generations per passage for 100 passages). A glycerol stock of each lineage was prepared every 10 passages ($\sim$ 220 generations) and stored at -80°C. Growth assays were performed for the two different evolved strains and their ancestral (parental strain) at 37 and 48 degrees of temperature.

6.3.2 Sequencing

Shotgun sequencing and 8kb PE 454 data was used for the parental strain. *De novo* assembly of the two sets of 454 data, using GS Data Analysis Software v2.6 (newbler)(http://454.com/products/analysis-software/index.asp) with default settings, were conducted. Mauve v2.3.1(Darling *et al.*, 2010) was used to compare this *de novo* assembly to the *E. coli* K-12, this showed that the parental strain had two plasmid genes inserted into its genome (size of insertion 3559 bp). Correcting for indels and SNPs were done by mapping the 454 data onto Ecoli k12 (NC_000913) using ssaha2 (http://www.sanger.ac.uk/resources/software/ssaha2/) - due to the error rate of 454 in homopolymer reagents, any indels or SNPs occurring here were discarded. Construction of the parental genome was done from the consensus of ssaha2, with manual addition of the 3.5kb insertion. A second check of the fasta sequence was done using breseq v0.16 available at:
http://barricklab.org/twiki/pub/Lab/ToolsBacterialGenomeResequencing. Annotation of NC_000913 and AJ277653 were added to the nucleotide sequence, using RATT (Otto *et al.*, 2011)

Sequences were quality trimmed using Sickle available at:
https://github.com/najoshi/sickle. The minimum length used was 85 and minimum quality was 30 which removed around 30% of all reads. This left 10.5 and 11.8 million read pairs in the 37 degrees and 48 degrees sample, respectively. Only half of the data was use for the mapping (~250 X coverage) against the parental genome.

Indels, SNPs, larger deletion and junctions were predicted using breseq v0.16 from the Barrick Lab. Breseq v0.16 uses ssaha2 to map reads against the reference genome. Breseq carries out the following steps:

1. Searches for mosaic read alignments, which could indicate new junctions in

the new genome compared to the reference genome.

2. Base substitution mutations and small *indels* are called by examining each position in the pileup of mapped reads to the reference. genome.

3. When reference regions has missing and low coverage, breseq would predicts deletions.

6.3.3 Chi-squared analyses

In this study we perform multiple chi-squared tests. These tests rely upon genetic information obtained from the whole genome of the *E. coli.* We obtained the full genome and associated data/annotations from NCBI (http://www.ncbi.nlm.nih.gov/genomes The chi-squared statistic is given as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{6.1}$$

Where O_i is the observed frequency of the quantity being tested, E_i is the expected frequency and n is the number of possible outcomes of each event (in our case it was always 2). This test is a "goodness of fit test", it analyses whether an observed distribution differs from that of an expected distribution. In section 6.3 we use a "independence" test using the chi-squared test statistic, consider the Table below (6.1). In this Table we ask whether category 1 and category 2 differ significantly, in our analyses. Category 1 and category 2 were represented by the 37 degree and 48 degree strains of *E. coli.* In this case the chi-squared statistic is given by:

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{6.2}$$

where a, b, c, and d are elements of a contingency Table as in Table 6.1. The result from either of these methods is compared against a chi-squared distribution for the assignment of a P-value.

Variable	Data 1	Data 2
Category 1	a	b
Category 2	c	d

Table 6.1: Example of contingency table. Example of a 2-way contingency Table analyzed in section 6.3

6.4 Results

After the 2200 generations growth tests showed an improved adaptation of the strain evolved under heat-shock conditions to 48 degrees, indicating possible fixation of temperature-advantageous mutations in the genome of this evolved strain. The strain grown at 37 degrees had lower fitness at 48 degrees than the adapted strain and, surprisingly, the adapted strain had lower fitness when returned to the 37 degree environment.

6.4.1 Transitions and transversions

There were 136 single nucleotide polymorphisms (SNPs) in the strain which evolved at 37 degrees; of these 128 (94.1%) were transitions and 8 were transversions (5.9%). There were 193 mutations in the strain which evolved at 48 degrees; of these 190 (98.4%) were transitions and only 3 (1.5%) were transversions. This not only demonstrates the high level of purifying selection on transversions but when these values were analyzed using a chi-squared test to assess a goodness of fit to the Tamura

& Nei(citation) model of evolution both were found to be hugely significant (both P-values$<$ 0.0001). Both strains were subjected to high levels of genetic drift by evolutionary bottlenecking. The expectation in this scenario is that there would be *greater* than normal levels of transitions as bottlenecking relieves some of the power of purifying selection.

In addition to the comparisons to expected levels of transitions we performed a contingency Table analysis *between* the profiles of transitions/transversions of the two strains to test whether there is an **even** greater level of purifying selection on the strain evolved under heat-shock conditions. Using a chi-squared test we found a significant asymmetric profile (P-value = 0.0293) between the categories of regular evolution and heat-shock evolution.

Table 6.2 shows that of the 128 and 190 transitions in the 37 degree and 48 degree *E. coli* respectively we found an asymmetric profile which favored mutations which lead to the fixing of a change from adenine (A) or thymine (T) to guanine (G) or cytosine (C). It is known that the nucleotides guanine and cytosine lend greater stability to DNA because of the greater number of hydrogen bonds between the two nucleotides. When chi-squared tests were performed upon these data we found no significant difference between the frequency of each type of transition when comparing the two strains.

Initial nucleotide	Final substitution	37 degree	48 degree
A	G	32.35%	34.19%
T	C	36.02%	29.53%
G	A	9.56%	19.19%
C	T	16.17%	15.54%
transitions	Any	5.88%	1.55%

Table 6.2: Purifying selection on transversions and GC bias. These data show that there is a large level of purifying selection upon mutations which result in transversions. It also shows that there is an asymmetric proportion of mutations resulting in GC content.

6.4.2 Pressures upon initial and resultant nucleotides

6.4.2.1 Initial nucleotides

To determine whether there are greater stresses upon each nucleotide being maintained or eliminated from the population we performed chi-squared tests upon the frequencies that each nucleotide is mutated. To perform a "goodness of fit" chi-squared test an expected value is needed, we generated the frequency of each nucleotide in the original strain (*E. coli* K-12). The GC content of this strain is 50.8%, therefore the expectation value for guanine and cytosine was 25.4% of the mutations and the expectation value for adenine and thymine was 24.6% of the mutations. We found that adenine was statistically more likely to be mutated in both strains, thymine was more likely to be mutated in the 37 degree strain and that guanine and cytosine were statistically less likely to be mutated in both strains. These data are summarized in Table 6.2.

6.4.2.2 Resultant nucleotides

To determine whether there are greater probabilities of each base pair being the resultant nucleotide in a mutation we again performed chi-squared tests, in this

Initial nucleotide	37 degree	48 degree
A	33.8% * (P-value = 0.0125)	34.7% ** (P-value=0.0011)
T	39.7% *** (P-value = 0.0001)	30.05% Not significant
G	9.55% *** (P-value = 0.0001)	19.17% *(P-value = 0.0468)
C	16.91% * (P-value = 0.023)	16.06% ** (P-value = 0.0029)

Table 6.3: Preferential accumulation of mutations in AT sites. These data show the percentage of each nucleotide prior to mutation in both strains. We performed a chi-squared test for each nucleotide. We calculated the expected frequency by calculating the percentage of each nucleotide in the whole genome. All but one of these tests were significant. There was a higher than expected number of adenine and thymines and a lower than expected number of guanine and cytosines.

case upon the frequencies of each nucleotide after a mutation. Again we required an expectation value, in this case we used the assumption that if the selection pressures were equal upon each nucleotide then each has an equal likelihood of being the resultant mutation. We make this assumption in order to test if the converse is true. Table 6.3 summarizes these analyses and is represented graphically in Figure 6.1. We found that in both the strain evolved at 37 degrees and the strain evolved under heat-shock there were significantly *more* guanines as the resultant nucleotide and significantly *fewer* thymines as the resultant nucleotide. In the strain evolved at 37 degrees there were significantly *more* cytosines and significantly fewer adenines as the resultant nucleotide. There was a higher percentage of cytosines and a lower percentage of adenines in the strain evolved in heat-shock but these were not significant. We also performed chi-squared tests to compare the categories of the heat-shock strain versus the strain evolved at 37 degrees using an "independence" chi-squared analysis. This test returned no significance.

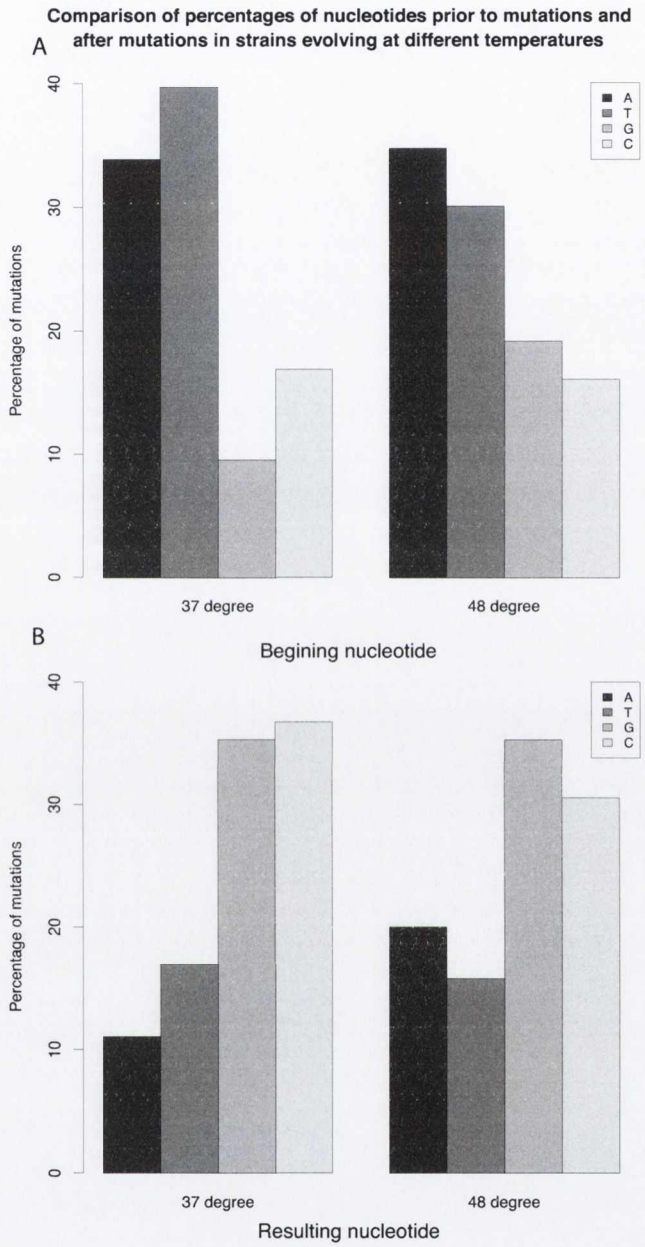

Figure 6.1: Preferential accumulation of mutations in AT sites in heat-shock strain resulting in higher GC content. Part (A) shows the proportion of sites comprised of each nucleotide **prior** to each mutation. The group of columns on the left shows the proportions of nucleotides from the strain evolving at 37 degrees and the group of columns on the right represents the proportion of nucleotides from strain evolving at 48 degrees. Part (B) shows the proportion of sites comprised of each nucleotide **after** each mutation. There is a clear preference in both strains for a loss of adenine and thymine and a retention of guanine and cytosine in both strains.

Resultant nucleotide	37 degree	48 degree
A	11% *** (P-value = 0.0002)	20% Not significant
T	16.9% * (P-value = 0.0295)	15.78% ** (P-value = 0.0021)
G	35.29% ** (P-value = 0.0056)	35.26% ** (P-value = 0.002)
C	36.76% ** (P-value = 0.0015)	30.53% Not significant

Table 6.4: Preferential gain of GC content. These data show the percentage of each nucleotide after mutation in both strains. We performed a chi-squared test for each nucleotide. We calculated the expected frequency by assuming there is an equal probability of a mutation to each other nucleotide. All but two of these tests were significant. There was a higher than expected number of adenine and thymines and a lower than expected number of guanine and cytosines.

6.4.3 Analysis of the location of mutations

If a mutation lies within an important region it is likely to be deleterious. Genes are extremely important regions as they code for the proteins and RNA which carry out the functions of the cell. Intergenic regions are less important; they do have function in regulation through transcription factor binding sites but much of intergenic regions is assumed to be non-functional. We investigated whether there was an asymmetry in the accumulation of mutations in intergenic regions by performing a chi-squared test. We tested those mutations within intergenic regions versus those within genes. To obtain an expected value for the calculation we calculated the proportion of the genome which is intergenic. We found that there were almost twice as many mutations in intergenic regions than is expected by random mutations. This number was significant according to the chi-squared test in both strains (P-value $<$ 0.0001). When tested for differences between the distribution of the two strains it was found that there was a modest (P-value = 0.08) difference between the strains; the strain which evolved at high temperature was more likely to pick up mutations in coding regions. This modest tendency for increased levels of mutations in the heat-shock strain is possible evidence of the buffering of mutations by chaperones

since chaperones are upregulated during heat shock.

6.4.4 Assymetric amino acid changes

We sought to identify whether codons which code for any amino acids are more or less susceptible to being lost and gained in bacterial protein coding regions under prolonged heat-shock exposure relative to those under normal conditions. We found that there were 97 non-synonymous changes in the strain evolved at 37 degrees and 117 non-synonymous changes in the strain evolved at 48 degrees.

6.4.4.1 Amino acid loss

Firstly, we investigated whether some amino acids are lost more readily than others in each strain. We performed chi-squared tests upon all amino acids, with expectation values calculated by the total amino acid composition in the whole genome of the strain. The results are summarized in Table 6.4. We found that both strains were significantly *more* likely to accumulate a mutation in a codon coding for valine or threonine. Both strains were also *less* likely to accumulated a mutation within a codon coding for leucine. The strain which evolved under heat-shock conditions was found to be more likely to accumulate mutations within codons coding for tyrosine but this was not the case in the strain evolved under normal conditions.

6.4.4.2 Amino acid gain

We investigated whether there is any asymmetry in the rates of introduction of amino acids into the genome by performing chi-squared tests upon the frequency of

Initial amino acid	37 degree	48 degree	greater or lower than expected
T	18.5% *** (P-value < 0.0001)	14.5% *** (P-value < 0.0001)	greater
V	11.3% ** (P-value = 0.0095)	12.8% ** (P-value = 0.0021)	greater
L	2.1% ** (P-value = 0.0061)	1.7% ** (P-value = 0.0029)	lower
Y		5.9% * (P-value = 0.0458)	greater

Table 6.5: Preferential accumulation of mutations in threonine and valine codons. These data summarize the chi-squared tests performed upon the accumulation of mutations in codons. We performed a chi-squared test for each set of codons. We calculated the expected frequency by finding the percentage of occurrence of codons which code for each amino acid in the genome. It was found that a greater number of mutations accumulate in the codons which code for threonine and valine in both strains. There were fewer than expected mutations in leucine codons in both strains. There were greater than expected numbers of substitutions accumulated in tyrosine codons in the heat-shock species only.

occurrence of codons for each resultant amino acid. Table 6.5 shows that there was greater than expected level of accumulations of alanine in both strains. In the strain subjected to high temperature conditions there was greater than expected accumulations of cysteine, serene and histidine at significant levels. For the strain evolved at 37 degrees there were greater accumulations of proline and glycine. There were lower than expected levels of leucine codons accumulated. Cysteine is the only amino acid which forms disulphide bonds; these bonds lend stability to proteins making them less susceptible to denaturation. Since denaturation can occur more easily at high temperatures it is unsurprising that the strain which evolved at high temperature has an enrichment in mutations leading to cysteine codons. The argument could be made that the propensity to accumulate mutations in codons coding for tyrosines may merely underlie the fact that tyrosine codons are AT rich and the exchange to cysteine is merely a reflection of GC bias. We investigated, this, however and discovered that half of the polymorphisms which resulted in a cysteine were from argenine and involved transitions from A/T to G/C. This provides further evidence

Resultant amino acid	37 degree	48 degree	greater or lower than expected
A	26.8% *** (P-value < 0.0001)	26.48% *** (P-value < 0.0001)	greater
C		5.1% ** (P-value = 0.0001)	greater
S		14.52% ** (P-value = 0.0001)	greater
H		5.1% * (P-value = 0.039)	greater
P	26.8% (P-value < 0.0001)		greater
G	14.43% (P-value < 0.0109)		greater
L	3%(P-value < 0.0214)		lower

Table 6.6: Preferential gain of alanine codons. These data summarize the chi-squared tests performed upon the resultant codons after a mutation. We performed a chi-squared test for each set of codons. We calculated the expected frequency finding the background percentage of occurrence of each codon in the genome. There are greater than expected numbers of mutations leading to alanine codons in both species. There are greater than expected mutations leading to cysteine, serene and histidine in the heat-shock strain. There are greater than expected mutations leading to proline and glycine codons in the 37 degree strain and fewer than expected mutations leading to leucine codons.

that this is a heat-shock dependent shift as this trend was not evident in the strain evolving at 37 degrees. Interestingly none of the genes which gained cysteine codons are clients of GroEL and therefore cannot make use of the benefits of chaperone buffering. These data along with the data in section 6.4.4.1 represent two sides of the same coin. The loss of tyrosine codons in the previous section are somewhat a precursor for the gain of cysteine codons in the 48 degree strain.

6.4.5 Analysis of GroEL clients

We sought to identify whether genes which are clients of the chaperonin GroEL accumulated a greater number of mutations than non-clients of GroEL. In the strain of *E. coli* evolving in normal conditions we observed that there were 5 (3%) mutations

	37 degree	48 degree
Mutations/indels in clients of GroEL	5 (3.1%)	155
Mutations/indels in non clients of GroEL	11 (8%)	126
Chi-squared versus expectation	P-value = 0.08 (below expected)	P-value = 0.3 (above expected)
Chi-squared test against each other	P-value = 0.062	P-value = 0.062

Table 6.7: GroEL buffering of mutations in client proteins at 48 degrees. These data show that there are fewer than expected mutations, insertions or deletions in genes which are clients of GroEL at 37 degrees suggesting that these genes are under high purifying selection. In the strain at 48 degrees this high level of purifying selection is absent. When a chi-squared test is performed between the two strains there is modest significance.

(including insertions and deletions) within GroEL client genes and 155 mutations in non-client genes. In the strain of *E. coli* which was evolved at 48 degrees we found 11 (8%) mutations within GroEL client genes and 126 mutations in non-client genes. To determine whether these numbers are greater or less than expected we calculated the percentage length of the *E. coli* genome which comprises clients. We then performed a chi-squared test on each strain. We found, when testing the 37 degree strain that there was modest significance (P-value = 0.08) of below expected numbers of mutations in GroEL clients. There was no significance in the 48 degree strain suggesting that during heat-shock (when the expression levels of GroEL are greater), there is a buffering effect upon mutations accumulated in GroEL genes. We also performed a chi-squared test between the strains which supported the conjecture that the profiles of mutations accumulated in client genes is different between strains evolving at high temperatures and those evolving at mesophilic temperatures.

6.4.6 Analysis of functional categories

Function underpins phenotype and as it is necessary to analyze whether there was a mutational bias towards one function over others. We took the COG functional categories (obtained from NCBI) and assigned them to each of the genes. We calculated the frequency of nonsynonymous mutations in each functional category. We tested these frequencies against the expected number of mutations in the genome. Due to the random nature of mutations there is an expectation that mutations will be evenly distributed throughout the genome. We performed chi-squared tests upon each functional category for both the 37 degree and 48 degree strains. Table 6.8 shows a summary of the results obtained. We found that both strains were enriched for mutations in coenzyme transport and metabolism (H) genes.

The 37 degree strain was enriched in cell wall/membrane (M) genes and transcriptional (K) genes. The advent of high numbers of mutations in cell wall/membrane genes is not surprising, it is known that these genes experience high levels of functional divergence relative to other functional categories (Caffrey *et al.*, 2012). What is somewhat surprising is the enrichment of mutations in transcriptional genes. These genes are known to be less likely to experience evolutionary change (Caffrey *et al.*, 2012), possibly because they represent a relatively small number of genes but are responsible for the regulation of a very large number of genes.

The 48 degree strain was enriched for mutations accumulating in Energy production and conversion (C) and interestingly under represented in Replication, recombination and repair genes (L). This strain was under imposed heat-shock and hence experienced high levels of stress. Chaperones can buffer mutations (as mentioned in introduction) this result suggests that mutations which hinder the DNA replication/repair mechanisms cause too much instability in the genome when combined

COG functional category	37 degree	48 degree	greater or lower than expected
H (Coenzyme transport and metabolism)	26.8% *** (P-value < 0.0001)	26.48% *** (P-value < 0.0001)	greater
C (energy production and conversion)		** (P-value = 0.0001)	greater
L (Replication recombination and repair)		** (P-value = 0.0001)	lower
M (cell wall/membrane)	(P-value < 0.0001)		greater
K (transcription)	(P-value < 0.0109)		greater

Table 6.8: Accumulation of mutations in functional categories. These data compare functional categories of genes in two strains of *E. coli*, one evolved at 37 degrees and one at 48 degrees. Both species were subjected to evolutionary bottlenecking and therefore experience greater genetic drift. Both species are enriched for mutations in coenzyme transport and metabolism, this suggests that the increase in genetic drift experienced by both strains causes a selection pressure that favors coenzyme transport and metabolism. The strain evolved at 37 degrees experiences enriched mutations in cell wall/membrane genes, this is not surprising as these are known to adapt more than others. The strain evolved at 48 degrees is enriched for mutations in energy production genes and impoverished for mutations in replication, recombination and repair genes. This suggests that mutations in replication/repair genes are extremely deleterious during the added stress of heat-shock and experience strong purifying selection.

with heat-shock.

6.4.7 Analysis of the relative importance of genes involved in mutations

6.4.7.1 Analysis of interacting proteins

Of the non-synonymous mutations there were 92 and 106 genes in the 37 degree and 48 degree strain respectively for which there was interaction data on the bacteri-

ome.org database. Of these genes there were 2 interacting genes in the 37 degree strain and 5 interacting genes in the 48 degree strain. To see if this was significant we needed an expected value of interacting genes from a distribution of this size in order to carry out a "goodness of fit" test. To retrieve this we randomly selected 100 genes from the whole genome of *E. coli* and assessed how many were interacting. We repeated this 1000 times to get an unbiased distribution. The expected percentage was 2%. We performed the fisher exact-test and found that the 48 degree strain had modest significance (P-value = 0.027) accumulate more mutations in interacting genes. One of the genes *oppA* was actually involved in two interactions. This gene is a client of GroEL. These data suggest that the buffering effect of GroEL (and possibly other Hsps) are buffering mutations within clients which is leading to coadaptation.

6.4.7.2 Analysis of expression levels

We analyzed the expression intensities of genes involved in mutations in both strains of *E. coli* using microarray data from the study by Covert *et al.*. (2004). This expression study reported dChip-normalised mean mRNA expression values across three replicates of wild-type *E. coli* cells growing in aerobic conditions. We only included genes which had expression intensities in all three replicates in the experiment. In an attempt to remove genes that were not expressed we plotted a histogram (Figure 6.2) and manually chose a cutoff to impose on the dataset. We removed all expression values below an intensity of 250. Figure 6.3 shows a graph of the comparison of the average expression level of genes which have accumulated mutations in the two strains studied. We performed a t-test on these and found that there was marginal significance (P-value = 0.19).

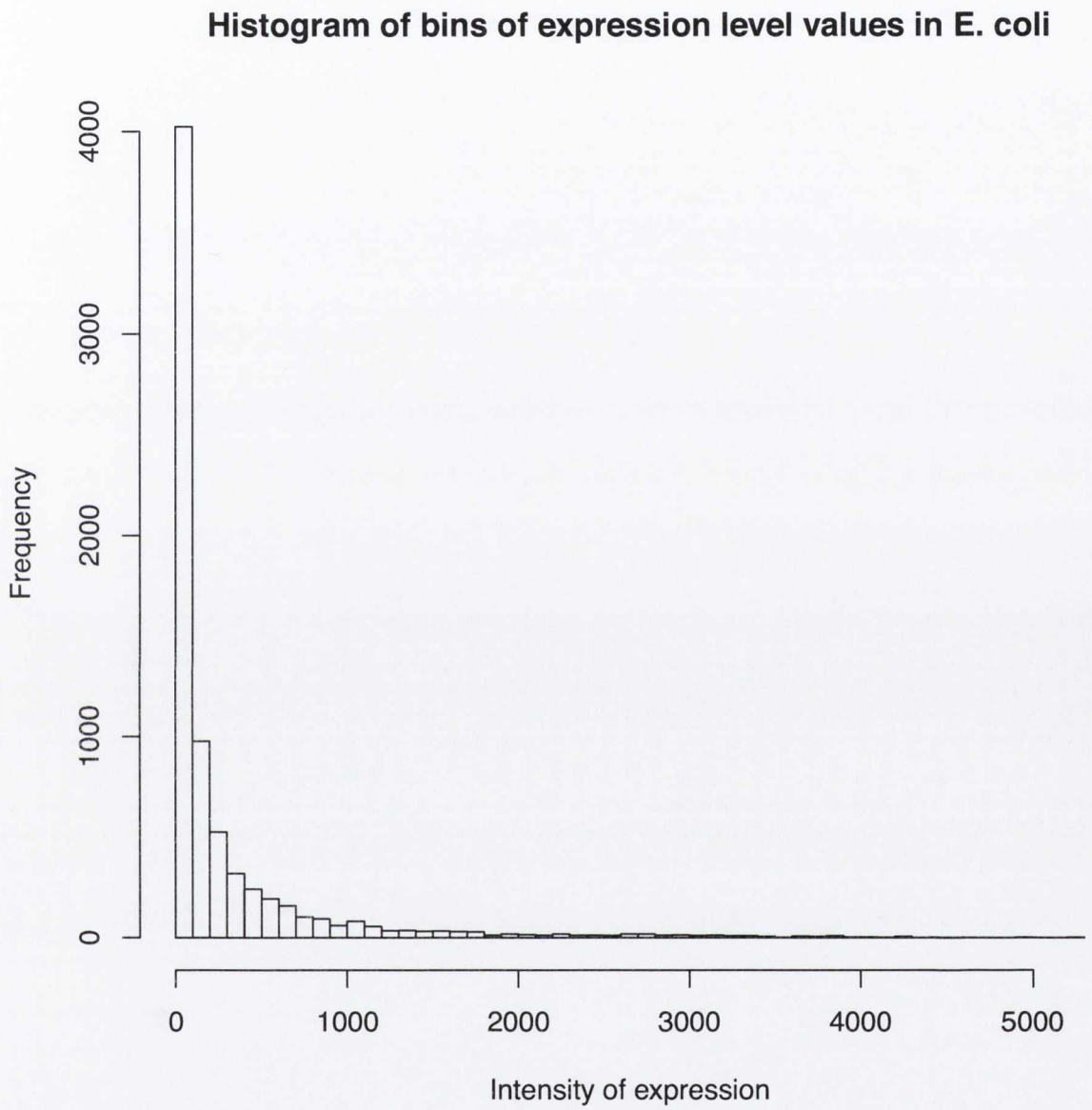

Figure 6.2: Histogram of gene expression. Displayed here is a histogram of expression level of all genes in the *E. coli* data set. We used these data to obtain a threshold of background genes for our analysis of differences in expression levels of genes involved in mutations in the strains grown at 37 degrees and those grown at 48 degrees.

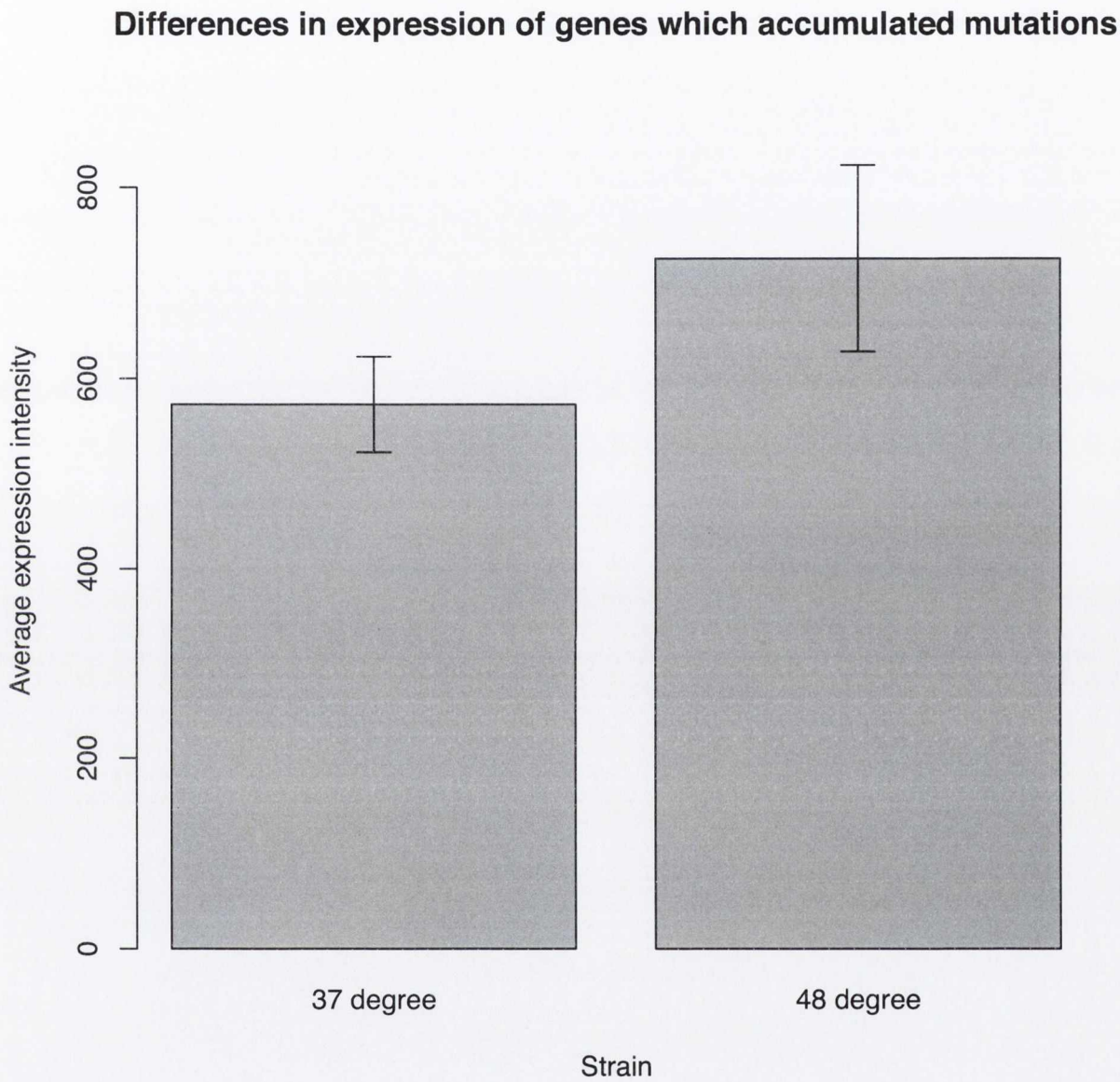

Figure 6.3: Profile of gene expressions in mutated genes. This graph illustrates the difference in mean expression values of genes involved in mutations in the 37 degree and 48 degree *E. coli* strains. This trend is small but it should be noted amongst the other analyses documented in this chapter.

6.5 Discussion

The multiple analyses documented above highlight two separate areas of discussion: Firstly, the effect which heightened genetic drift has upon *E. coli* strains irrespective of other evolutionary conditions and, secondly, the relative effects of mutations between strains of *E. coli* evolving under heat-shock versus those evolving at mesophilic temperatures.

6.5.1 Heightened genetic drift effects

We observed that *E. coli* undergoing strong effects of genetic drift have high purifying selection on transversions when compared to values expected by the Tamura Nei evolutionary model. We observed a significant preference for substitutions from adenines and thymines to guanines and cytosines. We found, contrary to work by Hildebrand *et al.*, that there was no evidence to suggest that there is any selective pressure upon the accumulation of synonymous substitutions which resulted in GC content. Hildebrand *et al.*. investigated the ratio of AT to GC versus GC to AT polymorphisms and concluded that there was a signifiant shift. When we compared these transitions in both synonymous and non synonymous sites we found there was no difference. This suggests that GC content bias could be due to periods of heightened genetic drift as opposed to a property of synonymous sites. We also found, as expected, that intergenic regions had significantly greater accumulations of mutations owing to their relative lack of function.

Additional similarities of asymmetric profiles were found in the relative loss and gain of codons. Both strains had a propensity to accumulate mutations in codons which coded for threonine and valine in place of alanine codons. The loss of these

codons can possibly be attributed to the greater GC content of alanine codons. Both strains were also statistically less likely to accumulate mutations in leucine codons which is somewhat surprising considering these codons are AT rich (56%).

6.5.2 Heat-shock specific variation

When contrasting the differences between the strain evolving under mesophilic conditions to that of the strain evolving at heat-shock we found a number of differences. We found significant evidence that the levels of purifying selection upon transversions in the 48 degree strain was greater than that of the 37 degree strain, despite the fact that the 37 degree strain had significance itself. This suggests that the added genome instability due to transitions coupled with heat shock is very highly deleterious to the cell.

Assymetric codon accumulations were observed in the 48 degree strain. Cysteine coding codons were enriched after heat-shock conditions. This could have resulted from the added stability of disulfide bonds and occurred in genes which cannot benefit from the buffering effects of GroEL. Other asymmetries in amino acid gain were a higher probability of gain of serine and histidine by the 48 degree strain and glycine by the 37 degree strain. Attempts at understanding a functional basis for these results have not as yet been fruitful. GroEL buffering of client proteins has been heavily studied as discussed in the introduction to this chapter. We found evidence that there was modest significance between the two strains in support of previous evidence that GroEL buffers mutations in its clients. This evidence lends significant weight to the theory that the buffering of mutations by GroEL can drive functional innovation and ecological adaptation. Furthermore it demonstrates the robustness of the *E. coli* genome to accumulated mutations and that robustness is

a facilitator of evolvability during environmental change. The implications of these results along with those in chapter 5 show that we have shown that GroEL actually buffers mutations in its clients during stress conditions as opposed to GroEL being a chaperone for proteins which accumulate greater numbers of mutations.

On a whole these results demonstrate the slow moving nature of evolution. Many effects of mutations are hard to detect even in cases where the rate of mutations is accelerated as in these data. We have presented a clear picture of the effects of heightened genetic drift upon *E. coli* along with a thorough contrast of the effects of heat-shock versus mesophilic environments. These results represent a clear example of ecological adaptation emphasized by the shift in environmental suitability of the strain grown under heat-shock conditions. In addition to the growth rate experiments performed the sequence data which accompanies it represents an excellent data source where a set of mutations are definitively known to have caused ecological adaptation.

We conclude that though the effects of heat-shock are small when taken separately the accumulation of these effects can cause phenotypic responses which are reflected in ecological adaptations.

Chapter 7

Conclusions

7.1 Functional divergence analysis

The inspirational breakthroughs made by Kimura (1968) and Ohno (1972a and 1972b) in developing the neutral and nearly neutral theories of evolution have laid the way for other excellent additions to evolutionary theory over the past forty years. The emergence of new sequencing technologies along with the computational advancements has seen the development of evolutionary theory to include topics, such as subfunctionalization, neofunctionalization (Force *et al.*, 1999), evolvability and robustness (Wright, 1932; Poole *et al.*, 2003; Wagner 2008). The emergence of new functions through these differing sources of variability is a topic of interest not limited to the field of molecular evolution but rather to the field of molecular biology as a whole.

The software presented in Chapter 2 aims at discovering amino acid sites which are evidence of functional divergence. This software can analyze single proteins but can be used to perform whole proteome analysis. In Chapter 2 we demonstrate the merit of whole proteome analyses by analyzing all proteins from 750 bacterial genomes. We identify specific amino acid sites of significant importance, such as those found in the VirB8 and VirB9 secretion system proteins in *E. coli*. In addition, we illustrated that with an ever increasing number of genomes available, whole proteome analyses

can shed significant light upon the diversification of genomes by demonstrating that host-associated strains are less functionally divergent than free-living strains. This suggests that the evolvability of these genomes has decreased.

7.2 Coevolution analysis

Much of molecular evolutionary research has been focused upon the importance of conserved sites in multiple sequence alignments. Sites which are highly conserved are known to be of importance. The theories of evolvability and robustness however allow for small changes to be acceptable in a genome. These small changes can lead to compensatory mutations which allow organisms to explore sequence space. This exploration can lead to increased fitness and niche specialists through functional innovation. Atchley *et al.*. (2000) described coevolution as composed of phylogenetic, structural, functional, interaction and stochastic signal. The separation of these variables is necessary to explore the role provided by coevolving sites in functional innovation.

The CAPS 2.0 software described in Chapter 4 demonstrates that ancestral sequence reconstruction removes almost all false positives from neutrally evolved alignments, implying that stochastic signal is removed. We have also demonstrated that site specific bootstrapping leads to the report sites that are closer in three dimensional structure for intra-molecular coevolutionary analysis. In the case of inter-molecular coevolutionary analysis we have shown that bootstrapping increases the reliability of predictions of interactions. In Section 7.3 I discuss the use of CAPS in identifying functional innovation in bacteria.

7.3 Phenotypic variation through chaperone buffering

Chaperons are ancient proteins found across all domains of life. Their role in the evolution of organisms is still poorly understood. It has been shown in many studies that chaperones buffer mutations in their clients (Rutherford and Lindquist, 1998; Queitsch *et al.*, 2002; Fares *et al.*, 2002; Williams and Fares; 2010) and therefore they increase the robustness of the organism. Though these studies have demonstrated chaperone buffering occurs there has not been much evidence of the increase in evolvability and specific cases of functional innovation facilitated by chaperones.

In Chapter 5 we presented an analysis of both functional divergence and coevolution upon a set of 85 gamma-proteobacteria. This analysis made use of both programs presented in Chapter 2 and Chapter 4. We assigned categories of GroEL client class and essentiality to the alignments. Results from this chapter included that GroEL buffers both mutations leading to sites of functional divergence and pairs of coevolving sites. We also found that essential genes are more likely to report functionally divergent sites along with higher levels of coevolution. We concluded, however, that levels of coevolution may be less significant because a lowered statistical likelihood of occurrence in essential genes. Most importantly, the clustering analysis of functional divergence performed uncovers evidence of phenotypic variability due to GroEL buffering. Sites in functionally important genes are identified in species with atypical lifestyles. Moreover, the genes predicted to be functionally divergent are those that have been experimentally verified as functionally relevant to their niche speciality.

7.4 Analysis of experimental evolution

The emergence of new function and phenotype is well accepted to arise from positive selection according to Darwin's theory. On a molecular basis the source of most (not all, e.g. LGT, polyploidy) functional innovations are point mutations within the genome. The relative effects of differing selection pressures, such as heat-shock, increased mutation rate, increased genetic drift and alternative food sources is of great interest to develop a better understanding of genome evolution and the diversity of life.

In Chapter 6 we analyzed the various asymmetries of accumulated mutations in two strains of *E. coli*. Both strains were subjected to increased mutation rate and genetic drift but differed in the fact that one was grown under continual heat-shock and the other under normal conditions. We found that genetic drift caused heightened purifying selection upon mutations resulting in transversions and caused a bias towards the accumulation of GC content. We found in the strain evolved at heat-shock that there was significant pressure upon the accumulation of mutations in cysteine coding codons. We also observed that GroEL clients were more likely to accumulate mutations when evolving under heat-shock. In addition to this we found that in the 48 degree strain interacting genes were more likely to accumulate mutations.

Taken together, these results build a picture of the mechanisms by which bacterial genomes can evolve to be niche specialists. After all, the strain of *E. coli* grown at 48 degrees had lower fitness at 37 degrees after 2200 generations. The preference for stabilizing cysteine residues along with buffered mutations in GroEL clients and mutations in interacting proteins gives us a greater understanding of the mechanisms of innovation. We conclude that these are subtle changes but they are measurable.

7.5 Perspective

In this thesis I have developed two novel methods of analysis of evolutionary events; functional divergence and molecular coevolution. These methods have been designed to form a better understanding of these two phenomena but also to aid in the understanding of the molecular mechanisms by which proteins, such as chaperones, affect the robustness and evolvability of organisms. There remain many open questions in both molecular evolution and biological research as a whole. Some of those which could be investigated through analyses of functional divergence and coevolution are (i) what are the effects of coevolution and functional divergence in molecular signaling pathways, (ii) what is relative importance of coevolution and functional divergence in the phylogenetic "tree of life", (iii) how do coevolution and functional divergence shape effect the genome following whole genome duplication with a focus upon gene loss and genome instability due to the deleterious effects of copy number variation and gene expression.

Bibliography

[1] Abascal, F., Zardoya, R., and Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21 (9): 2104-2105, 2005.

[2] Adams, M. D., Goglin, K., Molyneaux, N. *et al.* Comparative Genome Sequence Analysis of Multidrug-Resistant Acinetobacter baumannii. *J. Bacteriol* vol. 190 no. 24 8053-8064, 2008.

[3] Albert, R., H. Jeong, and Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* 406:378-382, 2000

[4] Altenhoff A. M. , Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology* 5: e1000262-e1000262, 2009

[5] Anderson B.E., Neuman M.A. Bartonella spp. as emerging human pathogens. *Clin Microbiol* Rev 10: 203-219, 1997

[6] Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* 2000;17:16478.

[7] Atwell, S., M. Ultsch, A. M. De Vos, and Wells, J. A. Structural plasticity in a remodeled protein-protein interface. Science 278:1125-1128, 1997

[8] Azuma Y, Ota M An evaluation of minimal cellular functions to sustain a bacterial cell. *BMC Syst Biol* 3: 111, 2009

[9] Bailey S, Ward D, Middleton R, Grossmann JG, Zambryski PC Agrobacterium tumefaciens VirB8 structure reveals potential protein-protein interaction sites. *Proc Natl Acad Sci* U S A 103: 2582-2587, 2006

[10] Baron C. VirB8: a conserved type IV secretion system assembly factor and drug target. *Biochem Cell Biol* 84: 890-899, 2006

[11] Baron C. Antivirulence drugs to target bacterial secretion systems. *Curr Opin Microbiol* 13: 100-105, 2010

[12] Benjamini Y., and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* Series B (Methodological) 57: 289-300, 1995

[13] Berglund A.C., Wallner B., Elofsson A., Liberles D.A. Tertiary windowing to detect positive diversifying selection. *J Mol Evol* 60: 499-504, 2005

[14] Blount, Z D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* 489, 513518, 2008.

[15] Brenner D.J., O'Connor S.P., Hollis D.G., Weaver R.E., and Steigerwalt A.G. Molecular characterization and proposal of a neotype strain for Bartonella bacilliformis. *J Clin Microbiol* 29: 1299-1302, 1991

[16] Cascales E., Christie P.J. The versatile bacterial type IV secretion systems. *Nat Rev Microbiol* 1: 137-149, 2003

[17] Charpentier, B. and Branlant, C. The Escherichia coli gapA gene is transcribed by the vegetative RNA polymerase holoenzyme E sigma 70 and by the heat shock RNA polymerase E sigma 32. *J. Bacteriol.* vol. 176 no. 3 830-839, 1994.

[18] Chelvanayagam, G., A. Eggenschwiler, L. Knecht, G. H. Gonnet, and Benner, S. A. An analysis of simultaneous variation in protein structures. *Protein Eng.* 10:307-316, 1997.

[19] Choi, S. S., W. Li, and Lahn, B.T. Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. *Nat Genet* 37:1367-1371, 2005

[20] Clarke, N. D. Covariation of residues in the homeodomain sequence family. *Protein Sci* 4:2269-2278, 1995.

[21] Codoner FM, Dars J-A, Sol RV and Elena SF The Fittest versus the Flattest: Experimental Confirmation of the Quasispecies Effect with Subviral Pathogens. *PLoS Pathog* 2(12): e136. doi:10.1371/journal.ppat.0020136

[22] Codoner, F. M., and Fares, M.A. Why should we care about molecular coevolution? Evolutionary Bioinformatics Onlines 4:9, 2008

[23] Codoner, F. M., M. A. Fares, and S. F. Elena. Adaptive covariation between the coat and movement proteins of prunus necrotic ringspot virus. *J Virol* 80:5833-5840, 2006

[24] Conant GC. , and Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9: 938-950, 2008

[25] Covert M. W., Knight E. M., Reed J. L., Herrgard M.J ., Palsson B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004; 429(6987):92-96.

[26] Cowen, L. E., Lindquist, L.Hsp90 Potentiates the Rapid Evolution of New Traits: Drug Resistance in Diverse Fungi.*Science*, Vol. 309 no. 5744 pp. 2185-2189, 2005.

[27] Crick F.H. The origin of the genetic code. *J Mol Biol* 38: 367-379, 1968

[28] Cuvier G. Sur le squelette d'une trs-grande espce de quadrupde inconnue jusqu' prsent, trouv au Paraguay, et dpos au cabinet d'histoire naturelle de Madrid, 1809

[29] Dagan T., Martin W. The tree of one percent. *Genome Biology* 7: 118-118, 2006

[30] Darwin, C *On the Origin of Species.*

[31] Dayhoff, M. O., Erk, R. V., and Park, C. M. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, volume 5, pages 89-99. National Biomedical Reasearch Foundation, Washington D.C.

[32] De Bary, A. and Trubner, K. J. The phenomenon of symbiosis. Strasbourg, Germany, 1879

[33] Degnan PH, Lazarus AB, Wernegreen JJ Genome sequence of Blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res* 15: 1023-1033, 2005

[34] Dehio C Bartonella interactions with endothelial cells and erythrocytes. *Trends Microbiol* 9: 279-285, 2001

[35] Dimitrov, T., Dashti, A. A., Albaksami, O., M M Jadaon, M. M. Detection of mutations in the gyrA gene in fluoroquinolone resistance Salmonella enterica serotypes typhi and paratyphi A isolated from the Infectious Diseases Hospital, Kuwait. *J Clin Pathol* 2010;63:83-87

[36] Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and Nielsen, R. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21 Suppl 1:i126-135, 2005.

[37] Douglas A.E. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. *Annu Rev Entomol* 43: 17-37, 1998

[38] Dramsi S., Cossart P. Intracellular pathogens and the actin cytoskeleton. *Annu Rev Cell Dev Biol* 14: 137-166, 1998

[39] Dutheil J, Gaillard S, Bazin E, Glemin S, Ranwez V, Galtier N, Belkhir K Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7: 188-188, 2006

[40] Dyall SD, Brown MT, Johnson PJ Ancient invasions: from endosymbionts to organelles. *Science* 304: 253-257, 2004

[41] Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek 73: 25-33, 1998

[42] Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*

[43] Fares M.A., Elena S.F., Ortiz J., Moya A., Barrio E. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol 55: 509-521, 2002

[44] Fares, M. A. Computational and Statistical methods to explore the various dimensions of protein evolution. *Current Bioinformatics* 1:207-217, 2006.

[45] Fares M.A. and Travers S. A. A. A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses. *Genetics* 173, no. 1 (May 1, 2006): 9-23.

[46] Fares, M. A., and McNally, D. CAPS: coevolution analysis using protein sequences. *Bioinformatics* 22:2821-2822, 2006

[47] Fares, M. A., M. X. Ruiz-Gonzalez, and Labrador, J. P. Protein coadaptation and the design of novel approaches to identify protein-protein interactions. IUBMB *Life* 63:264-271, 2011

[48] Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, 17(6):368-376, 1981

[49] Felsenstein, J. Phylogenies and the comparative method. *American Naturalist* 15:15, 1985

[50] Fisher RA The genetical theory of natural selection. Oxford Clarendon Press; 1930

[51] Fitch, W. M. Rate of change of concomitantly variable codons. J Mol Evol 1:84-96, 1971

[52] Fitch, W. M., and Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-593, 1970

[53] Fodor, A. A., and Aldrich, R.W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56:211-221, 2004

[54] Force A., Lynch M., Pickett, F.B., Yan Y.L., & Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4), 1531-1545. PMC1460548

[55] Fricke WF, Wright MS, Lindell AH, Harkins DM, Baker-Austin C, Ravel J, Stepanauskas R Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate Escherichia coli SMS-3-5. J Bacteriol 190: 6779-6794, 2008

[56] Fryxell, KJ The coevolution of gene family trees. *Trends Genet.*, 12:364-369.

[57] Galhardo, R. S., Hastings, P. J., Rosenberg, S. M. Mutation as a Stress Response and the Regulation of Evolvability. Vol. 42, No. 5 , Pages 399-435, 2007.

[58] Gans J, Wolinsky M, Dunbar J Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309: 1387-1390, 2005

[59] Gascuel O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 14: 685-695, 1997

[60] Gavin A. C., Bsche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M. *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141-147, 2002.

[61] Gil R., Sabater-Munoz B., Latorre A., Silva F.J., Moya A. Extreme genome reduction in Buchnera spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci U S A 99: 4454-4458, 2002

[62] Gillespie JH Population genetics: a concise guide: Johns Hopkins University Press.

[63] Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44:71567165.

[64] Gloor, G. B., G. Tyagi, D. M. Abrassart, A. J. Kingston, A. D. Fernandes, S. D. Dunn, and Brandl, C. J. Functionally compensating coevolving positions are neither homoplasic nor conserved in clades. Mol Biol Evol 27:1181-1191, 2010

[65] Gobel, U., C. Sander, R. Schneider, and Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* 18:309-317, 1994

[66] Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co- evolution of proteins with their interaction partners. *J Mol Biol.*, 299:283293.

[67] Goh CS, Cohen FE. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol.*, 324:177192.

[68] Goh, C. S., D. Milburn, and Gerstein, M. Conformational changes associated with protein-protein interactions. Curr Opin Struct Biol 14:104-109, 2004

[69] Goldman N, Yang Z A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736, 1994

[70] Graur, D., Gouy, M., & Wool, D. In retrospect: Lamarck's treatise at 200. *Nature*, 460(7256), 688-689

[71] Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16: 1664-1674

[72] Gu X (2001) Mathematical modeling for functional divergence after gene duplication. J Comput Biol 8: 221-234

[73] Gu X (2001) Mathematical modeling for functional divergence after gene duplication. J Comput Biol 8: 221234.

[74] Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23: 1937-1945

[75] Hakes L, Lovell SC, Oliver SG, Robertson DL. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci* USA. 104:79998004.

[76] Haldane JBS New paths in genetics. London: George Allen & Unwin; 1941

[77] Hartl, D. L., Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295(5561):1852-1858.

[78] Hasegawa, M., Kishino, H. and YanoDating, T.Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160-174.

[79] Hashimoto M, et al. Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol Microbiol.* 2005;55:137-149.

[80] Kato J, Hashimoto M. Construction of consecutive deletions of the Escherichia coli chromosome. *Mol Syst Biol.* 2007;3:132.

[81] Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*

[82] He, X. , Zhang, J., Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157-1164.

[83] Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genet* 6(9): e1001107. doi:10.1371/journal.pgen.1001107

[84] Hoffman, N. G., C. A. Schiffer, and Swanstrom, R. Covariation of amino acid positions in HIV-1 protease. *Virology* 314:536-548, 2003

[85] Hummel, J., N. Keshvari, W. Weckwerth, and Selbig, J. Species-specific analysis of protein sequence motifs using mutual information. BMC Bioinformatics 6:164, 2005.

[86] Hughes, A. L. and Nei, M. (1988). Patterm of nucleotide substitution at major histocompatibility complex class i loci reveals over dominant selection. *Nature*, 335(6186), 167-170.

[87] Ihler GM (1996) Bartonella bacilliformis: dangerous pathogen slowly emerging from deep background. FEMS Microbiol Lett 144: 1-11

[88] Innan H. , and Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* 11: 97-108, 2010

[89] Ito T., Chiba T., Ozawa R., Yoshida M., Hattori M., Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome *Proc Nat Acad Sci* 98(8):4569-74, 2001.

[90] Jin, D. J., Walter, W. A., Gross, C. A. Characterization of the termination phenotypes of rifampicin-resistant mutants. *Journal of Molecular Biology* Volume 202, Issue 2, 20 July 1988, Pages 245253.

[91] Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*: CABIOS 8: 275-282

[92] Jukes T, Cantor C (1969) Evolution of protein molecules. In Mammalian protein metabolism, Munro HN (ed), pp 21-123. New York: Academic Press

[93] Kerner MJ, et al. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli. Cell* 2005;122:209-220.

[94] Kim, W. K., D. M. Bolser, and Park, J. H. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics* 20:1138-1150, 2004.

[95] Kim Y, Subramaniam S. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins.*, 2006;62:111524.

[96] Kim, Y., M. Koyuturk, U. Topkara, A. Grama, and Subramaniam, S. Inferring functional information from domain co-evolution. Bioinformatics 22:40-49, 2006.

[97] Evolutionary rate at the molecular level. *Nature*, 217(5129), 624-626, 1968.

[98] Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111-120, 1980.

[99] Kimura M (1983) The neutral theory of molecular evolution: Cambridge University Press.

[100] Korber, B. T., R. M. Farber, D. H. Wolpert, and Lapedes, A. S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A 90:7176-7180, 1993.

[101] Kullback, S. Information theory and statistics. Wiley, New York, 1959.

[102] Kumar RB, Das A (2001) Functional analysis of the Agrobacterium tumefaciens T-DNA transport pore protein VirB8. J Bacteriol 183: 3636-3641

[103] Lamarck, J.B.

[104] Philisophie Zoologique, 1809

[105] Lake JA, Jain R, Rivera MC (1999) Mix and match in the tree of life. Science 283: 2027-2028

[106] Lenski, R. E. 2004. Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. *Plant Breeding Reviews* 24:225-265.

[107] Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res 37: D471-478

[108] Lovell, S. C., Robertson, D. L. An Integrated View of Molecular Coevolution in ProteinProtein Interactions, *Mol Biol Evol* (2010) 27 (11): 2567-2575.

[109] Lund PA (2009) Multiple chaperonins in bacteria–why so many? FEMS Microbiol Rev 33: 785-800

[110] Lynch M. , and Conery JS. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155, 2000

[111] Lynch M. , and Conery JS. The origins of genome complexity. *Science* 302: 1401-1404, (2003)

[112] Lynch M. , and Katju V. The altered evolutionary trajectories of gene duplicates. *Trends Genet* 20: 544-549, (2004)

[113] Madaoui H, Guerois R. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc Natl Acad Sci* USA. 105:77087713.

[114] Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9: 608-628

[115] Manges AR, Johnson JR, Foxman B, O'Bryan TT, Fullerton KE, Riley LW (2001) Widespread distribution of urinary tract infections caused by a multidrug-resistant Escherichia coli clonal group. N Engl J Med 345: 1007-1013

[116] Martin, L. C., G. B. Gloor, S. D. Dunn, and L. M. Wahl, L.M. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21:4116-4124, 2005.

[117] Mayer, M. P., Rdiger, S., Bukau, B. Molecular Basis for Interactions of the DnaK Chaperone with Substrates. *Biological Chemistry* Volume 381, Issue 9-10, Pages 877885,1431-6730, DOI: 10.1515/BC.2000.109, July 2005.

[118] McCutcheon, J. P. and Moran, N. A. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis, *Proc Natl Acad Sci U S A*, 104 (49) 19392-19397, 2007.

[119] McKenzie GJ, Harris RS, Lee PL, Rosenberg SM (2000) The SOS response regulates adaptive mutation. *Proc Natl Acad Sci U S A* 97: 6646-6651

[120] Miyata, T. and Yasunaga, T. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*, 16(1):23-36.

[121] Moran N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Nac Acad Sci*, 93(7), 2873-2878.

[122] Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108: 583-586

[123] Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y, Wang X Co-evolution of ligand-receptor pairs. *Nature* 368:251-255.

[124] Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A 93: 10268-10273

[125] Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314: 267

[126] Nei, M. Gene duplication and nucleotide substitution in evolution. *Nature*, 221(5175):40-42.

[127] Nei, M. and Gojobori, T. Simple methods for estimating the numbers of synonymous and non synonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418-426.

[128] Neher, E. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci* U S A 91:98-102, 1994

[129] Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936

[130] Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299-304

[131] Ohno S. Evolution by gene duplication: Springer Verlag.(1970)

[132] Ohta T. Population size and rate of evolution. Journal of Molecular Evolution, 1(4), 305-314.

[133] Ohta T. Evolutionary rate of cistrons, and DNA divergence. Journal of Molecular Evolution, 1(2), 150-157.

[134] Ohta T. Slightly deleterious mutatant substitution in evolution, *Nature*, 246(5428), 96-98.

[135] Otto, T. D., Dillon, G. P., Degrave W. S., and Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucl. Acids Res.* (2011) 39 (9)

[136] Pagel, M. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proc. R. Soc. Lond.* B 22, 1994

[137] Pazos, F., M. Helmer-Citterich, G. Ausiello, and Valencia, A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 271:511-523, 1997.

[138] Pazos, F., O. Olmea, and Valencia, A. A graphical interface for correlated mutations and other protein structure prediction methods. Comput Appl Biosci 13:319-321, 1997.

[139] Pazos, F., and A. Valencia. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng 14:609-614.

[140] Pazos F, Ranea JA, Juan D, Sternberg MJ. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol.*, 352:10021015, 2005

[141] Pazos, F., D. Juan, J. M. Izarzugaza, E. Leon, and Valencia, A. Prediction of protein interaction based on similarity of phylogenetic trees. Methods Mol Biol 484:523-535, 2008.

[142] Pazos, F., and A. Valencia. 2008. Protein co-evolution, co-adaptation and interactions. *EMBO J* 27:2648-2655.

[143] Pace N.R. A molecular view of microbial diversity and the biosphere. Science 276: 734-740

[144] Perez-Brocal V., Gil R., Ramos S., Lamelas A., Postigo M., Michelena J.M., Silva F.J., Moya A., Latorre A. A small microbial genome: the end of a long symbiotic relationship? Science 314: 312-313, 2006

[145] Perler F., Efstratiadis A., Lomedico P., Gilbert W., Kolodner R., Dodgson J. The evolution of genes: the chicken preproinsulin gene. *Cell*, 20(2):555-566.

[146] Plotnikova*, O. V., Kondrashov, F. A., Vlasovb, P. K., Grigorenkoa, A.P., Ginterc, E. K., Rogaeva, E. I. Conversion and Compensatory Evolution of the ?-Crystallin Genes and Identification of a Cataractogenic Mutation That Reverses the Sequence of the Human CRYGD Gene to an Ancestral State. *American Journal of Human Genetics* Volume 81, Issue 1, July 2007, Pages 3243

[147] Pikuta E.V., Hoover R.B., Tang J. Microbial extremophiles at the limits of life. Crit Rev Microbiol 33: 183-209 2007

[148] Pinto G., Mahler D.L., Harmon L.J., Losos J.B. Testing the island effect in adaptive radiation: rates and patterns of morphological diversification in Caribbean and mainland Anolis lizards. Proc Biol Sci 275: 2749-2757, 2008

[149] Pollock, D. D., and W. R. Taylor. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng 10:647-657.

[150] Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol* 287:187-198.

[151] Poole, A. M., Philips, .M. J., Penny, D. Prokaryote and eukaryoteevolvability*Biosystems*, 69(2-3), 163-185, 2003.

[152] Queitsch, C., Sangster, T., and Lindquist, S., Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889), 618-24, 2002.

[153] R Development Core Team R: A language and environment for statistical computing. http://www.r-project.org, 2010

[154] Rastogi, L., Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol*, 5(1):28.

[155] Ren CP, Beatson SA, Parkhill J, Pallen MJ The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from Escherichia coli. J Bacteriol 187: 1430-1440, 2005

[156] Rose, M. R. Oakley, T. H. The new biology: beyond the Modern Synthesis. *Biology Direct*, 2(1), 30, 2007.

[157] Rudwick M.J.S George Cuvier, Fossil Bones, and Geological Catastrophies: New translations and Interpretations of the Primary Text, (1ed). University of Chicago Press.

[158] Roth A, Gonnet G, Dessimoz C Algorithm of OMA for large-scale orthology inference. BMC bioinformatics 9: 518-518, 2008

[159] Rutherford, S. L. and Lindquist. Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709), 336-42.

[160] Rutherford, S. L. Between genotype and phenotype: protein chaperones and evolvability. *Nature Reviews Genetics* 4, 263-274, 2003

[161] Rzhetsky, A. Estimating substitution rates in ribosomal RNA genes. Genetics 141:771-783, 1995.

[162] Sandstrom J, Telang A, Moran NA Nutritional enhancement of host plants by aphids - a comparison of three aphid species on grasses. J Insect Physiol 46: 33-40, 2000

[163] Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. and Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341-345.

[164] Scherer DC, DeBuron-Connors I, Minnick MF Characterization of Bartonella bacilliformis flagella and effect of antiflagellin antibodies on invasion of human erythrocytes. Infect Immun 61: 4962-4971, 1993

[165] Schoniger, M., and von Haeseler, A. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol Phylogenet Evol* 3:240-247, 1994.

[166] Schluter D The ecology of adaptive radiation, New York: Oxford University Press, 2000

[167] Schuster-Bckler, B. Conrad, D. and Bateman, A. Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions. *PLoS ONE* 5, e9474.

[168] Schneider A, Dessimoz C, Gonnet G OMA Browser Exploring orthologous relations across 352 complete genomes. Bioinformatics 23: 2180-2182, 2007

[169] Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature 407: 81-86, 2000

[170] Singer GA, Hickey DA Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 317: 39-47, 2003

[171] Sriramulu, D. D., Nimtz, M., Romling, U. Proteome analysis reveals adaptation of Pseudomonas aeruginosa to the cystic fibrosis lung environment. *Proteomics* Volume 5, Issue 14, pages 37123721, 2005.

[172] Somprasonga, N., Jittawuttipokab, T., Duang-nkerna, J., Romsangb, A., Chaiyenc, P., Schweizere, H. P., Vattanaviboona, P., and Mongkolsuka, S. Pseudomonas aeruginosa Thiol Peroxidase Protects against Hydrogen Peroxide Toxicity and Displays Atypical Patterns of Gene Regulation. *J. Bacteriol.* vol. 194 no. 15 3904-3912, 2012.

[173] Sowa, M. E.,Bennett, E. J.,Gygi, S. P., andHarper, J. W. (2009) Defining the human deubiquitinating enzyme interaction landscape. Cell 138, 389403.

[174] Stephens, S. G., Possible Significance of Duplication in Evolution *Adv Genet.*, 4:247-265.

[175] Stryer L., Berg J., and Tymoczko J., Biochemistry, (5ed.). W.H. Freeman, 2002

[176] Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolution-arily conserved networks of residues mediate allosteric commu- nication in proteins. *Nat Struct Biol.* 10:5969.

[177] Suzuki Y New methods for detecting positive selection at single amino acid sites. J Mol Evol 59: 11-19, 2004

[178] Suzuki Y Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol Biol Evol 21: 2352-2359, 2004

[179] Suzuki Y, Gojobori T A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16: 1315-1328, 1999

[180] Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG 50 million years of genomic stasis in endosymbiotic bacteria. Science 296: 2376-2379, 2002

[181] Tamura, K., and M Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* (1993) 10 (3).

[182] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41, 2003

[183] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-28, 2001

[184] Tavere, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on Mathematics in the Life Sciences*, ed. Miura, R. H., volume 17, pages 57-86.

[185] Taylor, W. R., and Hatrick, K.Compensating changes in protein multiple sequence alignments. *Protein Eng* 7:341-348, 1994.

[186]

[187] Tillier, E. R., and Lui, T. W. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19:750-755, 2003

[188] Tillier, E. R., L. Biro, G. Li, and D. Tillo. Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins* 63:822-831, 2006.

[189] Tillier, E. R., and Charlebois, R. L. The human protein coevolution network. Genome Res 19:1861-1871, 2009.

[190] Toft C, Williams TA, Fares MA .Selection for Translational Robustness in Buchnera aphidicola, Endosymbiotic Bacteria of Aphids. *Mol Biol Evol* (2009) 26 (4): 743-751.

[191] Toft C, Williams TA, Fares MA Genome-wide functional divergence after the symbiosis of proteobacteria with insects unraveled through a novel computational approach. PLoS Comput Biol 5: e1000344, 2009

[192] Tokuriki, N. and Tawfik, D. S. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* 459, 668-673, 2009

[193] Toyoda, Koichi and Teramoto, Haruhiko and Inui, Masayuki and Yukawa, Hideaki Expression of the gapA gene encoding glyceraldehyde-3-phosphate dehydrogenase of Corynebacterium glutamicum is regulated by the global regulator SugR, *Applied Microbiology and Biotechnology*:81, 0175-7598, 291-301,

[194] Travers, S. A., and Fares, M. A. Functional coevolutionary networks of the Hsp70-Hop-Hsp90 system revealed through computational analyses. *Mol Biol Evol* 24:1032-1044, 2007.

[195] Travers, S A. A., Tully, D.C., McCormack, G. P. and Fares, M. A. A Study of the Coevolutionary Patterns Operating within the env Gene of the HIV-1 Group M Subtypes. *Mol Biol Evol* (2007) 24 (12): 2787-2801

[196] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., *et al.* A comprehensive analysis of proteinprotein interactions in *Saccharomyces cerevisiae*, *Nature* 403, 623-627 (2000).

[197] Van Dyk, T. K., Gatenby, A. A. & Larossa, R. A. Demonstration by genetic suppression of interaction of GroE products with many proteins. *Nature* 342, 451 - 453,1989.

[198] van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernandez JM, Jimenez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Moran F, Moya A Reductive genome evolution in Buchnera aphidicola. Proc Natl Acad Sci U S A 100: 581-586, 2003

[199] Van Kesteren, R.E., Tensen CP, Smit, AB, van Minnen, J., Kolakowsko, LF, Meyerhof, W, Richter, D., van Heerikhuizen, H., Vreugdunhil, E. and Geraerts, WP Coevolution of ligand-receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. *J. Biol. Chem*, 271:3619-26

[200] L. Van Valen A New Evolutionary law, Evolutionary Theory, Vol. 1 (1973)

[201] Waddell PJ, Kishino H, Ota R. Phylogenetic methodology for detecting protein interactions. Mol Biol Evol. 24:650659, 2007.

[202] Wagner A. Robustness and evolvability: a paradox resolved. *Proceedings. Biological Sciences / The Royal Society*, 275(1630), 91-100, 2008

[203] Wang ZO, Pollock DD. Context dependence and coevolution among amino acid residues in proteins. *Methods Enzymol.*, 395:779790.

[204] Warnecke, T. and Hurst, L. D. GroEL dependency affects codon usagesupport for a critical role of misfolding in gene evolution. *Molecular Systems Biology* Article number: 340, 2010.

[205] Whelan, S. and Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol Biol Evol*, 5(a):90-96.

[206] Williams TA, Codoner FM, Toft C, Fares MA Two chaperonin systems in bacterial genomes with distinct ecological roles. *Trends Genet* 26: 47-51, 2010

[207] Wielgoss, S., Barrick, J. E., Tenaillon, O., Cruveiller, S., Chane-Woon-Ming, B., Mdigue, C., Lenski, R. E., Schneider, D. Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. vol. 1 no. 3 183-186.

[208] Wollenberg, K. R., and Atchley, W. R. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci* U S A 97:3288-3291, 2000.

[209] Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the sixth international congress on Genetics*

[210] Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13(5):555-6. 1997.

[211] Yang Z, Bielawski JP Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503, 2000

[212] Yang Z, Nielsen R Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-917, 2002

[213] Young, C. J., , Barral, J.M., Hartl, F. U. More than folding: localized functions of cytosolic chaperones. *Trends in Biochemical Sciences.* Volume 28, Issue 10, October 2003, Pages 541547

[214] Zeldovich KB, Berezovsky IN, Shakhnovich EI Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3: e5, 2007

[215] Zhang J Evolution by gene duplication: an update. Trends Ecol Evol 18: 292-298, 2003

[216] Zhang J Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21: 1332-1339, 2004

[217] Zhang J, Nielsen R, Yang Z Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-2479, 2005

[218] Zhang, F., Gu, W., Hurles, M. E., Lupski, J. R. Copy Number Variation in Human Health, Disease, and Evolution. *Annu. Rev Genomics Hum. Genet.* 10, 451-481.

[219] Zhou H, Roy S, Schulman H, Natan MJ. Solution and chip arrays in protein profiling. *Trends Biotechnol.* S34-9. Review, 2001

Chapter 8

Appendix A

Some images are provided here which pertain to the CAPS server in Chapter 4.

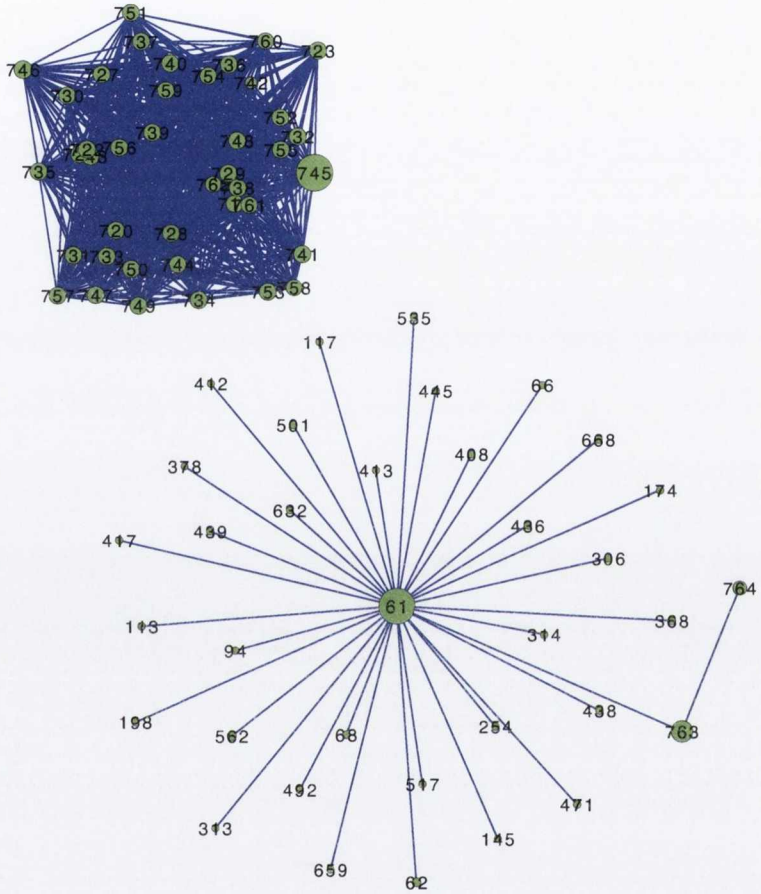

Figure 8.1: Example of CAPS web interface cityscape output. Depicted above is a cityscape layout of the output from the CAPS server. Pairs of sites are joined together with a line. Sites which are coevolving with more pairs are given a larger node. It can be seen in this example that there were two distinct groups. One group where there was a lot of coevolution within the group and a second group which had one central hub connecting the others. On the server these graphs are interactive.

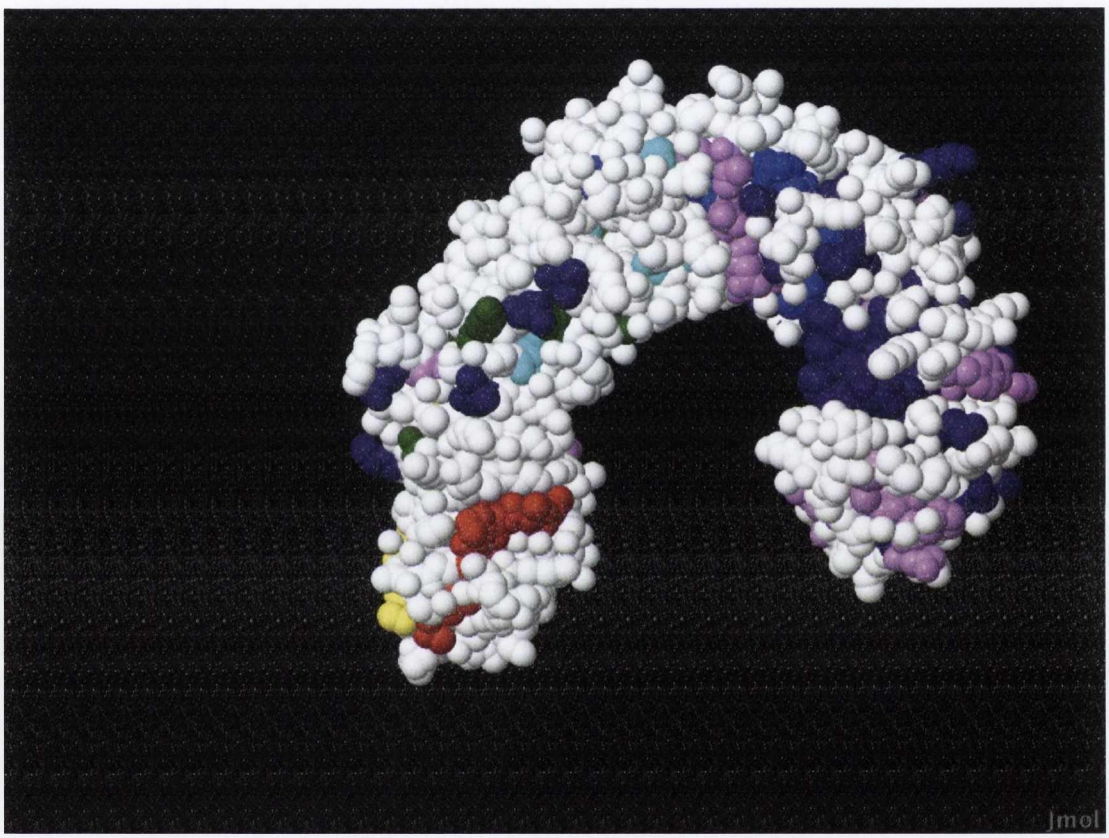

Figure 8.2: Structure of TLR1 with groups of coevolving amino acids mapped. Intra-molecularly coevolving groups are mapped to the three-dimensional structure of TLR1. It is shown that these residues are close geometrically.

Chapter 9

Appendix B

Some graphs are provided here which pertain to chapter 5.

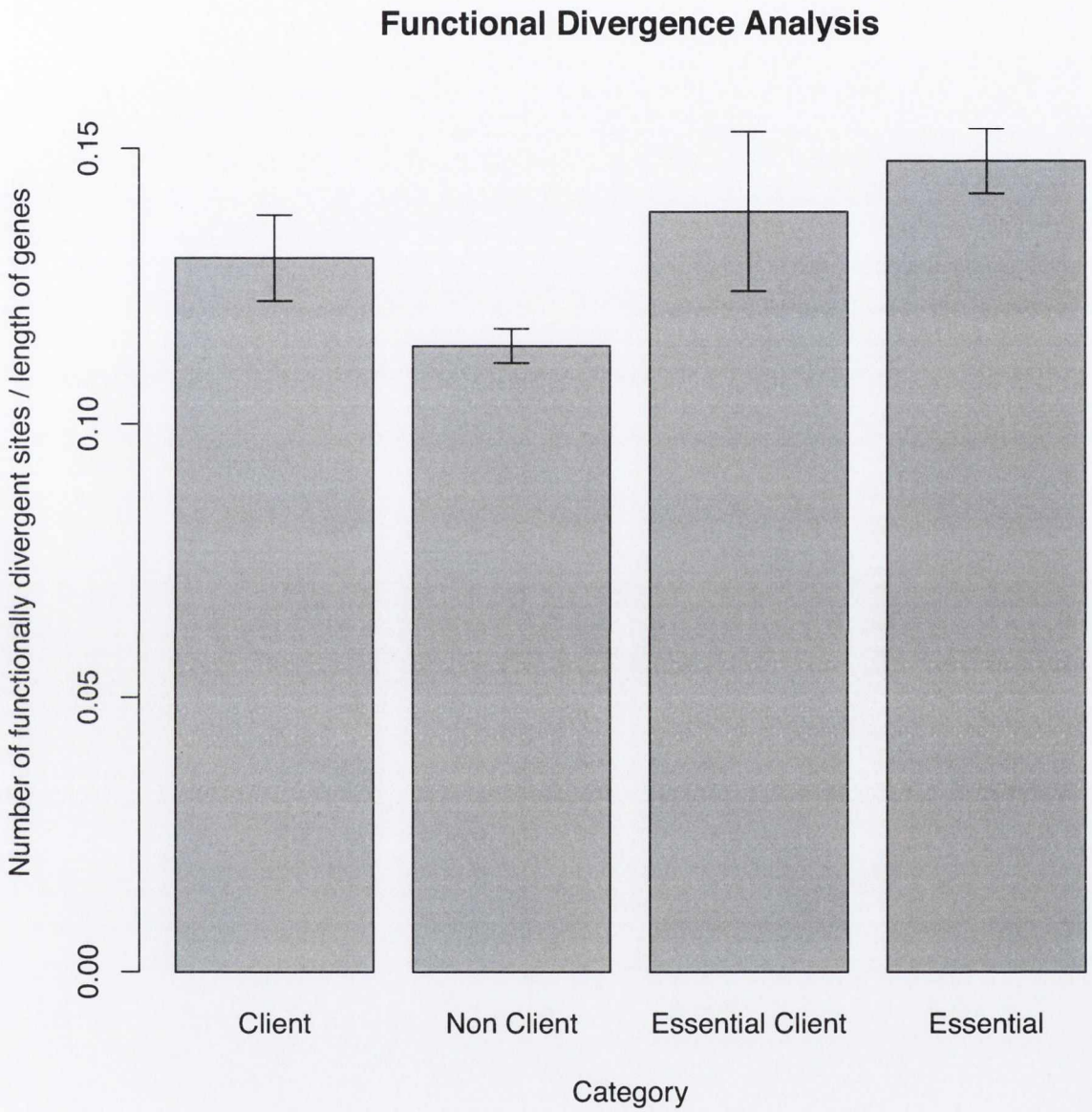

Figure 9.1: Graph of client versus essential genes. Here are contrasted the levels of functional divergent sites in clients, non-clients, essential and non-essential genes.

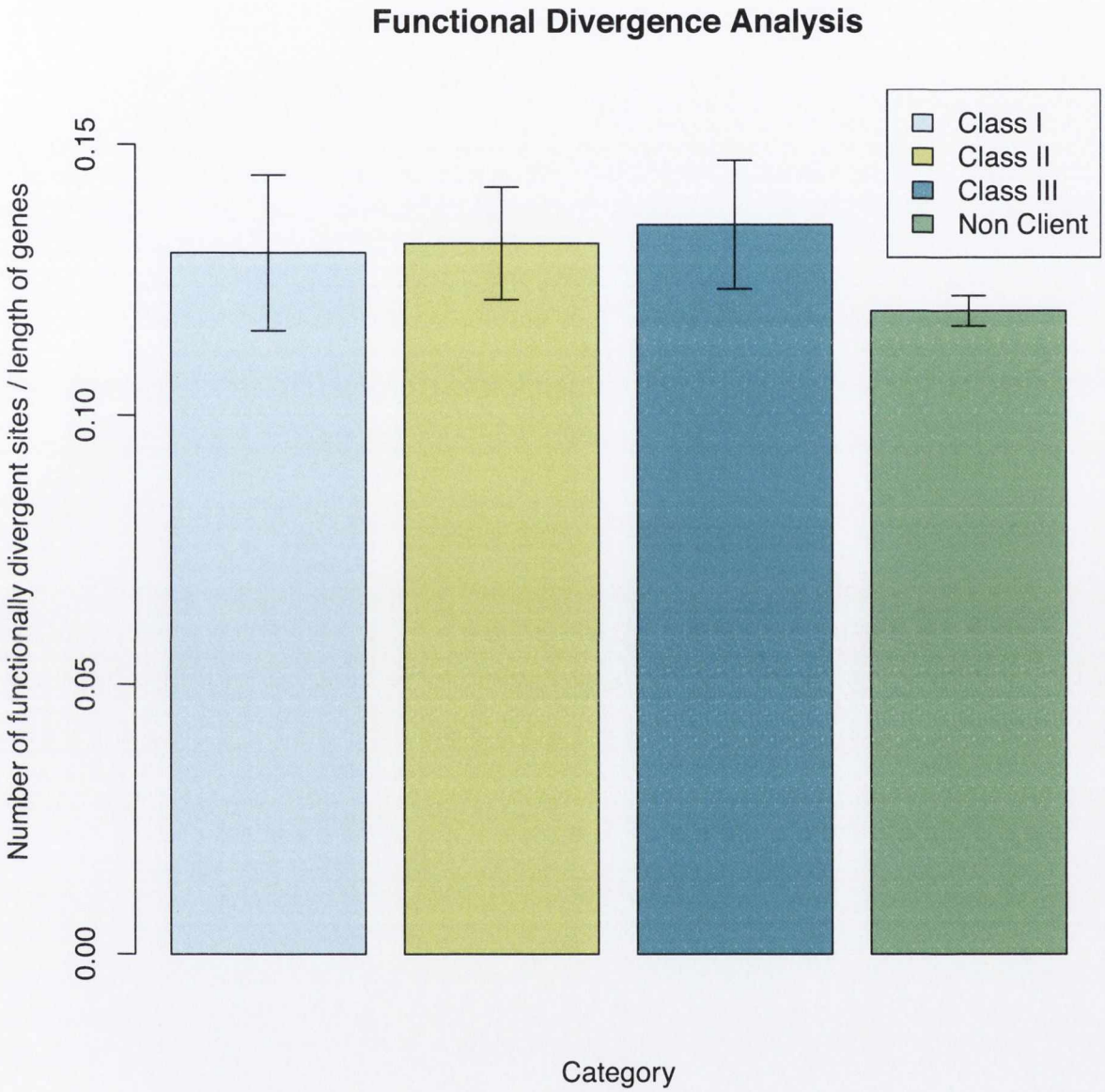

Figure 9.2: Graph of client classes. Presented here is a graph of the average number of functional divergence sites per position in alignment. There is clearly a trend; increasing from class I to class III, statistical tests do not find this trend significant suggesting chaperone buffering is a subtle effect.

Chapter 10

Appendix C

Data which pertains to this thesis can be found on an optical disc with this thesis:

File	Data
cafs-src.tgz	All source code pertaining to Chapter 2.
outputs.zip	All output files from functional divergence analysis in Chapter 3
caps-src.tgz	All source code pertaining to Chapter 4.
c_nc_data.tgz	Data set for Chapter 5
SNPs.tgz	All SNP data for Chapter 6

Table 10.1: Table listing all files in the supplementary DVD.