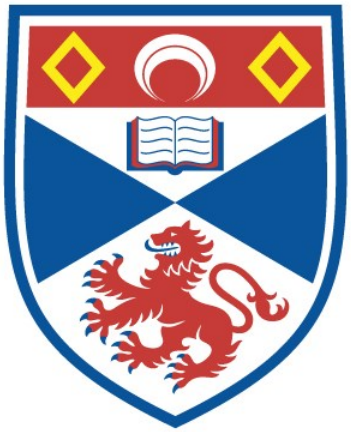
TRACKING THE EVOLUTION OF FUNCTION IN DIVERSE ENZYME SUPERFAMILIES

Rosanna G. Alderson

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

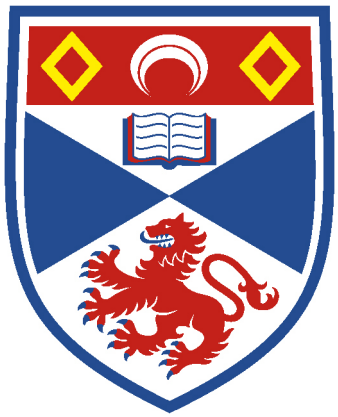


2016

Tracking the evolution of function in diverse enzyme superfamilies

Rosanna G. Alderson

March 2016



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of Doctor of Philosophy at the University of St Andrews

1. Candidate's declarations:

I, Rosanna Alderson, hereby certify that this thesis, which is approximately 40 000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September, 2011 and as a candidate for the degree of PhD. in August, 2012; the higher study for which this is a record was carried out in the University of St Andrews between 2011 and 2015.

I, Rosanna Alderson, received assistance in the writing of this thesis in respect of [language, grammar, spelling or syntax], which was provided by Dr Daniel Barker, Dr James McDonagh, Dr Leo Holroyd, Rachael Skyner, Kathryn Jean Swanson & Sih-Yu Chen.

Date: Signature of candidate:

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of …… in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: Signature of supervisor:

3. Permission for publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY
Embargo on all of print copy for a period of one year on the following ground:

- Publication would preclude future publication

ELECTRONIC COPY
Embargo on all of electronic copy for a period of one year on the following ground:

- Publication would preclude future publication

Date: Signature of candidate: Signature of supervisor:

“Most species do their own evolving, making it up as they go along, which is the way Nature intended. And this is all very natural and organic and in tune with mysterious cycles of the cosmos, which believes that there’s nothing like millions of years of really frustrating trial and error to give a species moral fiber and, in some cases, backbone.” — Terry Pratchett, Reaper Man

For Grace Alderson

Acknowledgments

Many thanks go to my supervisor, Dr John Mitchell and Dr Daniel Barker for all their supervision, guidance and encouragement in the last four years. I also thank Dr Tanja van Mourik, Dr V Anne Smith, Professor David O'Hagan and Dr Rebecca Goss for their input and interest in my research and progress of my PhD. Interesting discussions at conferences have proved vital in the critical evaluation of my own work, in particular, I would like to thank Professor Dannie Durand, Professor Christine Orengo and Dr Nicholas Furnham for their contributions to this. My immense gratitude goes to my fantastic colleagues Dr Neetika Nath, Dr Lazaros Maveridis, Dr Luna de Ferrari and Dr James McDonagh who have been fantastic (and patient!) tutors on the topics of statistics and scripting.

Engaging with the public on the topic of my research is an activity I take great pleasure in. Working with Dr Daniel Barker, Dr Heleen Plasier, Dr James McDonagh and Dr Mhairi Stewart on the 4273Pi project has been an absolute pleasure - thank you. Thanks to Dr Tanja van Mourik for inviting me to join the School of Chemistry Athena SWAN committee and helping me integrate this with my work as Student Champion of the Interconnect Network.

My gratitude also extends to my other colleagues - Luke, Ludo, Laz, Jose, Ava, Jan, Simon, Gregor, Ava, James, Rachael, Leo and Luna for the friendly and fun atmosphere in the lab and at our Friday night meet-ups at the Whey Pat tavern. Thank you Cindy, Kj, Ellen, Francois, Naomi, John, Martin, Gina, Elsbeth, David, Becky, Steve, Hauna and all the staff and friends at the St Andrews Sports Centre for much needed respite and fun.

Finally, I thank my family, James and his lovely family for their unfaltering support over the past years. Mum - thank you for all the phone calls, texts and nagging - thanks to you I have not become an insomniac or developed scurvy. Most importantly though, you have provided the motivation which has made this work possible - thank you.

Contents

List of Figures

List of Tables

List of abbreviations

AIC	Akaike Information Criterion
ASR	Ancestral Sequence Reconstruction
BIC	Bayesian Information Criterion
CAFASP	Critical Assessment of Fully Automated Structure Prediction
CASP	Critical Assessment of Structure Prediction
CSA	Catalytic Site Atlas
EC	Enzyme Classification
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
Indel	insertion or deletion
JTT	Jones, Taylor and Thornton
LBA	Long Branch Attraction
LRT	Likelihood Ratio Test
MAFFT	Multiple sequence Alignment based on Fast Fourier Transform
MCMC	Markov Chain Monte Carlo
MDA	Multiple Domain Architecture
ML	Maximum Likelihood

MRCA	Most Recent Common Ancestor
NJ	Neighbour Joining
NMR	Nuclear Magnetic Resonance
NNI	Nearest Neighbour Interchange
PDB	Protein Data Bank
PSSM	Position Specific Scoring Matrix
RMSD	Root Mean Square Deviation
SPR	Subtree Pruning and Regrafting
SSG	Structurally Similar Group
TBR	Tree Bisection and Reconnection
UPGMA	Unweighted Pair-Group Method using Arithmetic averages
WAG	Whelan and Goldman

Publications arising from work in this thesis

- Alderson, R. G., Ferrari, L. De, Mavridis, L., McDonagh, J. L., Mitchell, J. B. O., & Nath, N. (2012). **Enzyme informatics**. *Current Topics in Medicinal Chemistry,* 12(17), 1911–1923.
- Alderson, R. G., Barker, D., & Mitchell, J. B. O. (2014). **One origin for metallo-β-lactamase activity, or two? An investigation assessing a diverse set of reconstructed ancestral sequences based on a sample of phylogenetic trees.** *Journal of Molecular Evolution*, 79(3-4), 117–29.
- Alderson, R. G., Barker, D., & Mitchell, J. B. O. **Investigating the evolution of functions in the DUF-62 superfamily.** *Manuscript in preparation.*

Abstract

Tracking the evolution of function in enzyme superfamilies is key in understanding how important biological functions and mechanisms have evolved. New genes are being sequenced at a rate that far surpasses the ability of characterization by wet-lab techniques. Moreover, bioinformatics allows for the use of methods not amenable to wet lab experimentation. We now face a situation in which we are aware of the existence of many gene families but are ignorant of what they do and how they function. Even for families with many structurally and functionally characterized members, the prediction of function of ancestral sequences can be used to elucidate past patterns of evolution and highlight likely future trajectories. In this thesis, we apply *in silico* structure and function methods to predict the functions of protein sequences from two diverse superfamily case studies.

In the first, the metallo-β-lactamase superfamily, many members have been structurally and functionally characterised. In this work, we asked how many times the same function has independently evolved in the same superfamily using ancestral sequence reconstruction, homology modelling and alignment to catalytic templates. We found that in only 5% of evolutionary scenarios assessed, was there evidence of a lactam hydrolysing ancestor. This could be taken as strong evidence that metallo-β-lactamase function has evolved independently on multiple occasions. This finding has important implications for predicting the evolution of antibiotic resistance in this protein fold. However, as discussed, the interpretation of this statistic is not clear-cut.

In the second case study, we analysed protein sequences of the DUF-62 superfamily. In contrast to the metallo-β-lactmase superfamily, very few members of this superfamily have been structurally and functionally characterised. We used the analysis of alignment, gene context, species tree reconciliation and comparison of the rates of evolution to ask if other functions or cellular roles might exist in this family other than the ones already established. We find that multiple lines of evidence present a compelling case for the evolution of different functions within the Archaea, and propose possible cellular interactions and roles for members of this enzyme family.

1. Introduction & Motivation

"Pattern, like beauty, is to some extent in the eye of the beholder"

Peter Grant, 1979

1.1. What is enzyme evolution?

The majority of the population are familiar with Darwin's 'Tree of Life' - the idea that all organisms share common ancestry [1]. Indeed, the idea of our closest evolutionary relatives being primates is well publicised and accepted (e.g. [2]). Behind this organismal evolution lies the molecular evolution of genes and proteins which ultimately affect the fitness of an organism and its chance of survival. These levels of evolution are ultimately interlinked.

This work focuses at the level of enzyme evolution. Enzyme evolution comprises the evolution of enzyme coding genes, their products and the after effects on the metabolism of a given organism. Mutations affecting the metabolic machinery of an organism can change the ability to produce products and degrade substrates enabling the adaptation to external changes.

1.2. Why is studying enzyme evolution important?

We have theories as to how evolution works based on patterns and observations in nature. We discuss some of these as well as methods that have been developed to plot and predict evolution. Every day, many more protein sequences are discovered than can be manually structurally and functionally characterised [3]. This leaves us with a backlog of many sequences for which we have no clue of structure, function or cellular role. For example, according to Mudag et al., a quarter of known protein families are devoid of members with either structural or functional annotation [4]. A large part of the problem is that structurally and functionally characterising a gene product in the laboratory is a lengthy and expensive process, and takes far longer than genome sequencing.

Increasingly, *in silico* methods are being used to ascertain function from sequence data (e.g. by authors such as Radivojac et al.[5]). The benefit of these *in silico* techniques is that they are much cheaper in terms of time and resources, allowing for many more predictions of protein function to be made. The area of bioinformatics and chemoinformatics is rapidly expanding, with an ever increasing plethora of tools to choose from. Examples of studies which demonstrate the use of currently available *in silico* prediction tools include those by Gerlt et al., Alderson et al., Jacobson et al. and Steffen-Munsberg et al. [6, 7, 8, 9].

One of the main criticisms of such *in silico* methods is that although often quicker to perform, they can be less accurate in their designation of function [7, 10, 11]. This is because bioinformatic strategies tend to work from empirical knowledge of other related (homologous) or similar enzymes. Rather than experimentally *demonstrating* function, bioinformatics has the power to make informed *predictions* of enzyme function.

The prediction of past, present and future enzyme functions has a particularly relevant application in understanding disease and in the design of drugs to combat it. Kumar et al. describe the use of phylogenetics as a 'telescope' that allows us to review the results and consequences of billions of years of experimentation through natural selection [12]. Phylogenetics enables us to review evolutionary inventions that have tended to persist and be independently invented in evolution. Correlating this information with environmental changes can give us important information regarding folds, reactions and mechanisms that are likely to be selected for in a wide range of circumstances. A particularly significant example of this is in the evolution of antibiotic resistance, where the past patterns of resistance can inform our predictions of likely future trajectories [13, 14].

The more data we have on enzyme structures and functions the more accurate our predictions are likely to be. The real strength of bioinformatics lies in the number of proteins for which *in silico* predictions can be made. In addition to this, although one *in silico* strategy may be relatively inaccurate, its conjunction with the results of other orthogonal bioinformatics strategies can reveal patterns in a dataset with good reliability in a way that individual wet-lab experiments would not be able to achieve.

1.3. What does this thesis contribute to the field?

In this work, we explore and evaluate the use of such *in silico* techniques to ask specific questions about two very different enzyme superfamilies. Part II of this thesis examines the metallo-β-lactamase superfamily. This constitutes a case-study where we predict enzyme evolution in a superfamily with a wealth of available structural and functional information. In contrast, our study of the DUF-62

superfamily (Part III) constitutes a case study of functional prediction in a superfamily where relatively few members have been characterised, and for those that have, their cellular role is under question. Each of these superfamilies is populated with enzymes of medical interest and provides interesting examples of the evolution of catalytic machinery for a diverse array of substrates.

For each case study we started with the construction of alignment and phylogeny for each superfamily. In each case, we used differing methods as a reflection of the level of sequence, structural and functional annotation available; although a common theme throughout this thesis is the use of methods that are informed by structure. The level of structural and functional data available for each superfamily was key in the defining questions concerning the evolution of function that we could ask. For the DUF-62 family, where structural and functional data was sparse, we used *in silico* methods to explore how many other functions/cellular roles might exist in the family, other than the few characterised. For the metallo-β-lactamase superfamily, the wealth of structural and functional information allowed for a question to be asked at a finer granularity of detail. For this superfamily, we asked whether the same function, achieved by different catalytic machinery in different enzymes, was the result of one divergent or multiple independent events of evolution.

For both cases, we aimed to answer our question by predicting function and/or cellular interactions of sequences that were not characterised. In this way, we demonstrate methodological procedures of diagnosing function when: 1) Many functions are characterised throughout the superfamily but differentiating between different scenarios of evolution is not simple and 2) When hardly any functions are known and multiple sources of prediction must be combined. Making predictions for things other than protein function, such as the reconstruction of ancestral sequences, can only be achieved by bioinformatic methods. In this thesis, we exemplify the use of methods only achievable by computational analysis to decipher important biological questions. Moreover, constructing alignments and phylogenies for evolutionarily diverse superfamilies such as these is by no means a trivial task [7]. In this thesis we demonstrate the use and application of methods to construct robust alignments and phylogenies for enzyme superfamilies, whilst also considering how to deal with phylogenetic uncertainty.

The metallo-β-lactamase superfamily includes enzymes that confer one mechanism of antibiotic resistance - an ever-emerging threat. We aimed to assess the possibility that the same antibiotic-hydrolysing function has evolved more than once in the same superfamily. The results of our work are important in future drug-design efforts, since evolutionary studies such as these have the power to reveal patterns in evolution. The DUF-62 superfamily is populated by the SAM hydrolase enzyme, which although characterised in terms of chemical mechanism, is a mystery in terms of cellular role. We asked whether this function and cellular role is likely to be conserved across the superfamily, which includes organisms that live in diverse and extreme environments. Our findings indicate that

multiple cellular roles, and/or functions are likely to have evolved in this family, providing important first steps in the structural and functional characterisation of more members of this diverse superfamily.

Part I.

Enzyme evolution and the methods used to study it

2. Enzyme evolution from a theoretical perspective

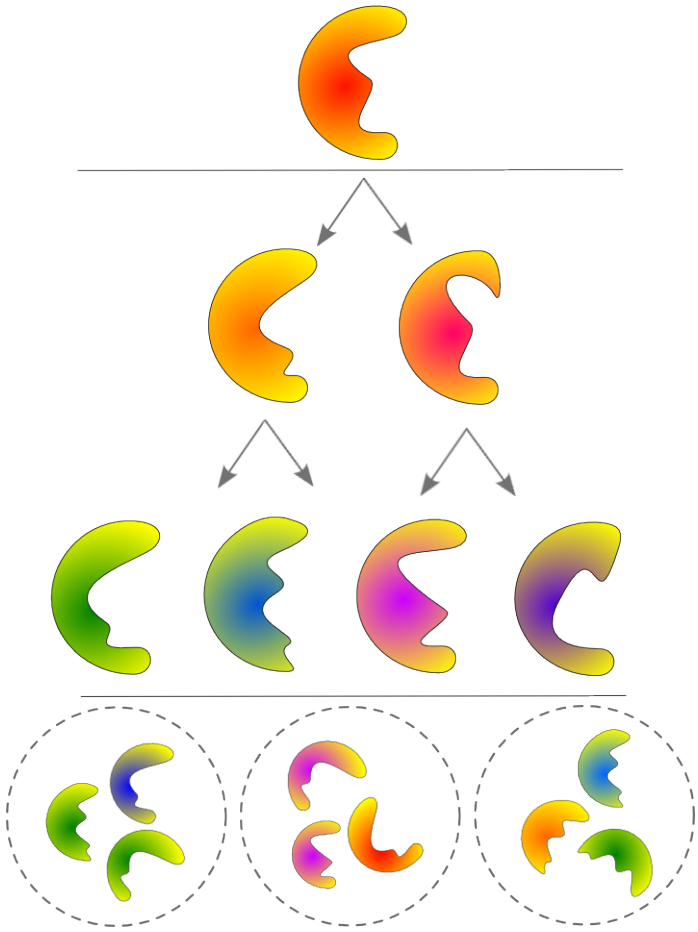


Figure 2.1.: *New enzymes are thought to to evolve by a process of duplication from the original, ancestral enzyme (top) followed by further divergence and specialisation, shown by changes in shape and colour (middle). Although divergent, we can see patterns of similarity in extant enzymes, represented here by colour and shape in related enzyme families (encompassed by dashed circles) that helps us determine their ancestral relationships.*

2.1. Why do enzymes evolve?

Natural selection acts on all heritable traits, from the morphological, right down to the biochemical level. As the process of transcription and translation is costly, in terms of resources, energy and time for the cell, useless biological molecules tend to get 'lost' in evolution. For example, pseudogenes, homologous to their functional relative, have lost their protein coding ability. This has come about *via* a process in which genes, not positively contributing to fitness, acquire mutations that eventually render them untranslatable. These pseudogenes become obsolete in the genome - a stark reminder that if something is not used it will eventually be lost in the process of evolution. Environments change over time, and a gene may find itself in a different organism or in the same lineage under different conditions. The enzyme must therefore evolve to meet the needs of the cell and ultimately the organism in order to be propagated in future generations.

2.1.1. What drives the evolution of metabolism?

Enzymes are the catalytic components of an organism's metabolism. Each component of the metabolic engine may undergo evolutionary changes that constitute positive, neutral and maladaptive traits. Much like Darwin's tree of life, enzymes also share common ancestry. Enzymes have diverged in evolution in order to carry out specific and required functions that differ between types of organisms and their environments. In this way, much as we do with organisms, we can group enzymes by ancestry - we say these related groups of enzymes are homologous and we call the whole set of homologous enzymes a superfamily. If these homologous enzymes have diverged after a speciation event we call these enzymes 'orthologous' and if they differ due to a duplication event, 'paralogous' [15, 16, 17, 18]. Within superfamilies, although related, exist diverse functions and chemical mechanisms of catalysis.

At this point, it is worth highlighting the fact that discerning genetic function has been viewed in different ways in the literature. For example, ENCODE the Encyclopedia of DNA Elements [19, 20] characterises gene function by virtue of demonstrable biochemical activity. However, as discussed by Graur et al. these biochemical activities used for functional classification by ENCODE, including gene transcription, transcription factor binding, histone modification, chromatin conformation and DNA methylation, are not always indicative of gene functionality [21]. It is argued that other factors such as effect on fitness and phenotype need to be taken into consideration [21, 22, 23]. In this work, we define function from information gained from experimentally characterised protein structures. For example, we might diagnose the function of a translated gene sequence from the presence of amino acids shown to contribute to catalysis in related, experimentally characterised structures.

Despite the variety of enzyme chemistries observed in nature there are fewer types of enzyme structure than there are genes in the human genome [24, 25, 26, 27] Viewed from this perspective, enzyme fold space is limited (as is protein sequence space [28]) and evolution has re-invented multiple protein structures to achieve a diverse variety of functions [29, 24]. What drives the evolution of new folds and structures? In general, current opinion can be summarised to fall into three main categories - substrate specificity, chemical mechanism and active site, as described by Gerlt et al [24].

Substrate specificity

In 1945 Horowitz suggested that the driver of the evolution of metabolic networks is the depletion of substrates in the environment - resulting in a 'backwards' evolution model [30]. Horowitz asked us to imagine an organism that was only capable of catalysing the last step of a modern metabolic pathway (*A)* [30]. Initially, this organism is capable of survival, since it finds the substrate (A) needed for the last step in a substrate-rich hypothetical environment [30]. However, as the population increases substrate A decreases [30]. In these conditions, organisms capable (by a chance repertoire of a combination of genes for example) of synthesising substrate A by say, enzyme *B* are at a clear selective advantage [30]. In this way, a new enzyme in the metabolic pathway (*B)* has evolved retrospectively from the point of view of the final product in the metabolic pathway [30]. Horowitz hypothesised that this process can happen repeatedly, resulting in long 'chains' of metabolic pathways in which the product of one enzyme reaction is the specific substrate for another [30].

Chemical Mechanism

A later theory, as discussed by Babbitt and Gerlt, and, Herschlag, D and O'Brien is that there existed a primordial pool of enzymes. Not optimised nor particularly specialised, but nevertheless, by chance, able to catalyse a range of partial reactions [31, 32, 33]. For example, take an enzyme from this primordial pool - it has the ability to catalyse reaction A quite well, but, by chance, it also on occasion successfully stabilises the transition state needed for reaction B. If there are multiple copies of the enzyme available, for example, after a duplication event, then whilst one copy of the enzyme is free to carry out its 'main' function, the duplicated copy is free of selection pressures and evolution can fine tune its sideline activity - reaction B. After years of divergent evolution we end up with two homologous enzymes *A* and *B,* both derived from one common ancestor but performing diverse chemistries, driven by an enzyme's inherent ability to catalyse a range of reactions [31, 32, 33].

This theory, of enzymes being 'recruited' to metabolic networks, has been explored by Caeteno-Anoles et al. in 2007 and is an extension of the theories of Nei, Ohno and Force et al. [34, 35, 36]. The constructed MANET database reveals that enzymes from the same homologous family have been recruited to perform a wide range of enzyme reactions [37]. Further, Caeteno-Anoles et al. propose that in a prebiotic world many enzymes had already been 'invented', adding support to the idea of a preexisting 'pool' of ancient (albeit inefficient) enzymes.

This theory goes along with the observation by others such as Babbitt, Gerlt and Todd et al. , that superfamilies tend to share some conservation in overall chemical mechanism but are able to catalyse a wide diversity of substrates [31, 33, 38]. Although Todd et al. do concede that in some cases substrate may be the driving force, as seen in some superfamilies such as the crotonase-like superfamily [38]. Evolution driven by either 'chemical mechanism' or 'substrate' is difficult to tease apart as they are often interlinked [38].

Active site

A third point of view, is that the very shape or architecture of the enzyme is under constraint [24]. From this point of view, we can imagine that enzyme *A* is able to support the reaction A, via positioning of various important functional groups involved with catalysis. However, in a different biological context (for example, in a different metabolic pathway) the same enzyme architecture may perform a different enzyme reaction by virtue of its optimally placed functional groups in a different location than needed for enzyme reaction A. This differs from the above proposal that chemical mechanism dominates, since in the above scenario the enzyme may adopt different conformations, therefore allowing different substrates to bind. Here, the enzyme is more of a 'fixed' platform in which fortunately placed chemical groups are able to catalyse a variety of different reactions [24]. The idea of evolution using a structural 'scaffold' has been discussed in relation to specific enzyme families, for example, by Aravind in reference to the diverse range of chemical reactions that have evolved in the metallo-β-lactamase superfamily [39].

A spectrum of hypotheses

We can therefore imagine the different hypotheses of the 'drivers' of evolution as lying on a spectrum (Fig. 2.2), with Horowitz 'backwards evolution' theory requiring that enzymes be extremely substrate specific, since both the original enzyme *A* and the next evolved enzyme *B* share specificity for either substrate or product respectively [24, 30]. In contrast, the hypotheses that either 'chemical mechanism' [31, 32, 33] or 'active site' [24] are the drivers for evolution require that promiscuity plays a role. For the 'chemical mechanism' driver hypothesis, it was noted that enzymes in homologous superfamilies tend to share at least a

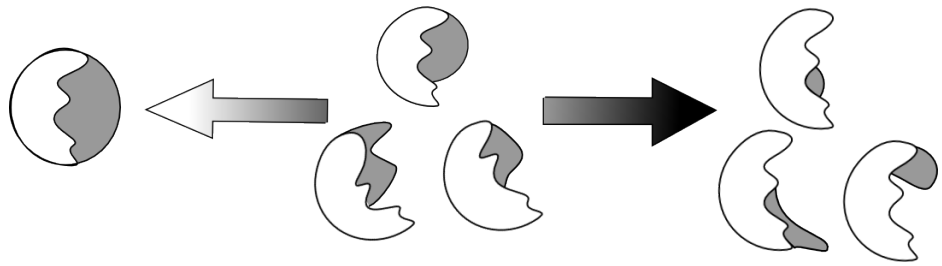


Figure 2.2.: *In this illustration, the enzyme active site is white and the substrate grey. In substrate driven evolution the enzyme is monogamous in terms of substrate (left). In chemical mechanism evolution the enzyme is promiscuous in terms of substrate but partially conserved in the chemical mechanism utilised, shown by differing substrates binding to the same area of the active site (centre). This is in contrast to active site driven evolution where the enzyme is substrate promiscuous and is also not necessarily conserved in terms of chemical mechanism, depicted here by different substrates binding to different areas of the active site (right).*

common partial chemical mechanism even when members catalyse a wide array of substrates. For this to be the case, the common structural scaffold of this family has had to allow for different substrates to bind – by virtue of a flexible enzyme active site but has used conserved steps in the chemical reaction to create diverse products. Looking to the other extreme, if the 'active site' acts as ultimate driver for evolution then then the protein scaffold remains fixed – the platform of catalytic amino acids are able to catalyse many different reactions and the enzyme is, by definition, highly substrate-promiscuous. A classic example is the TIM barrel fold, which hosts a dazzling array of molecular functions [40] and, as typified by the description of active site dominance by Gerlt and Babbitt, occurs in independent superfamilies [24].

2.1.2. 'Driven' evolution in context - arms races and evolutionary warfare

In effect, we are in competition for resources with other organisms. This driving force, in which organisms use their repertoire of enzymes to fight for limited resources, is a key example of 'why' enzymes evolve. The world is in flux - populations that do not adapt die, and species become extinct as a result. Understanding 'why' enzymes evolve is part of the process for developing and finding cost effective and rational novel drugs and targets. Enzyme evolution is a current and relevant process to all of human society - it has consequences far beyond abstract theories in textbooks. For example, antibiotic resistance, used as a case study in this thesis, is partially conferred by the ongoing ability of β-lactamase enzymes to render ineffective antibiotic drug molecules that would otherwise kill the pathogenic microbe. It is important to understand the driving forces for such

evolutionary changes if we are to advance in fields such as antibiotics and antimicrobials.

There have been many studies that have focused on evolutionary arms races, some with relevance to human health. Juarez *et al.* in 2008 pinpointed the structural sites in snake venom disintegrins that were under both positive and negative selection in evolution and discussed their rationalisation in terms of biological function [41]. Russell et al. in 2011 compared the evolution of insects and bacteria to xenobiotics introduced into the environment in the form of pesticides [42]. Insects and bacteria have evolved very different mechanisms to cope with these xenobiotics due to the different selection pressures of xenobiotics on insect and bacterial metabolism that is confounded by environmental, metabolic and genetic constraints [42]. For example, bacterial enzymes seem to have reached a higher level of optimisation compared to those of insects [42]. This result seems antiintuitive given the greater pressure of xenobiotics on insect life processes [42]. However, the authors reflect that bacteria have bigger populations, a greater frequency of Horizontal Gene Transfer (HGT) events, shorter generation times and are able to metabolise wider variety of compounds [42]. There are other examples of apparently rapid evolution of resistance to environmental by bacteria, such as [43], but, at the time of writing, a direct comparison of the ability to degrade xenobiotics in different organisms appears unique to [42]. The authors argue that as well as learning about insecticide resistance the bacterial response to insecticides may actually be helpful in exploring new remediation strategies and 'provide[s] unique opportunities to characterize this evolutionary process as it unfolds in real time' [42].

2.2. How do enzymes evolve?

How do the changes, driven by forces discussed above, come about? How do diverse superfamilies of enzymes, originally derived form a single ancestral enzyme, become populated with diverse and divergent enzymes of different functions?

2.2.1. The materials for enzyme evolution

In order for neofunctionalisation (the process of an enzyme acquiring a new function) to occur and become fixed in a population, there must be a way in which a given gene is able to acquire new mutations and not hamper the preexisting fitness of an organism. This is especially true when evolution occurs in housekeeping genes for which successful function is imperative for survival.

In 1970, Ohno proposed that duplication of genes, created in a stochastic fashion by unequal crossing over during recombination, retro-transposition or whole

chromosome duplication [44], provides the necessary material for neofunctionalisation [35]. By having an extra copy of a gene the organism is then able to maintain fitness whilst also providing evolution the opportunity to mutate the copy of the gene. Much like organismal evolution, the products of the two genes may compete for the same substrates and cellular niche or the gene may acquire an unrelated new function that provides an advantage to the organism and therefore may become fixed in the population. The majority of duplications lead to the generation of pseudogenes rather than new genes with new functions. However, in organisms with short generation times, such as pathogenic bacteria, these short generation times allow for rapid exploration of the fitness landscape [45, 46].

There is some debate as to exactly when the duplication event happens in the process of gene neofunctionalisation (for examples see [7]). It is known that many enzymes display promiscuity and/or substrate ambiguity. The question lies as to whether duplication allows a gene encoding an already promiscuous enzyme to acquire mutations that will allow the gene to fully sub-functionalize or whether duplication represents an obligatory event before an gene can start acquiring new mutations and undergo neofunctionalisation (Fig. 2.3) [47, 48].

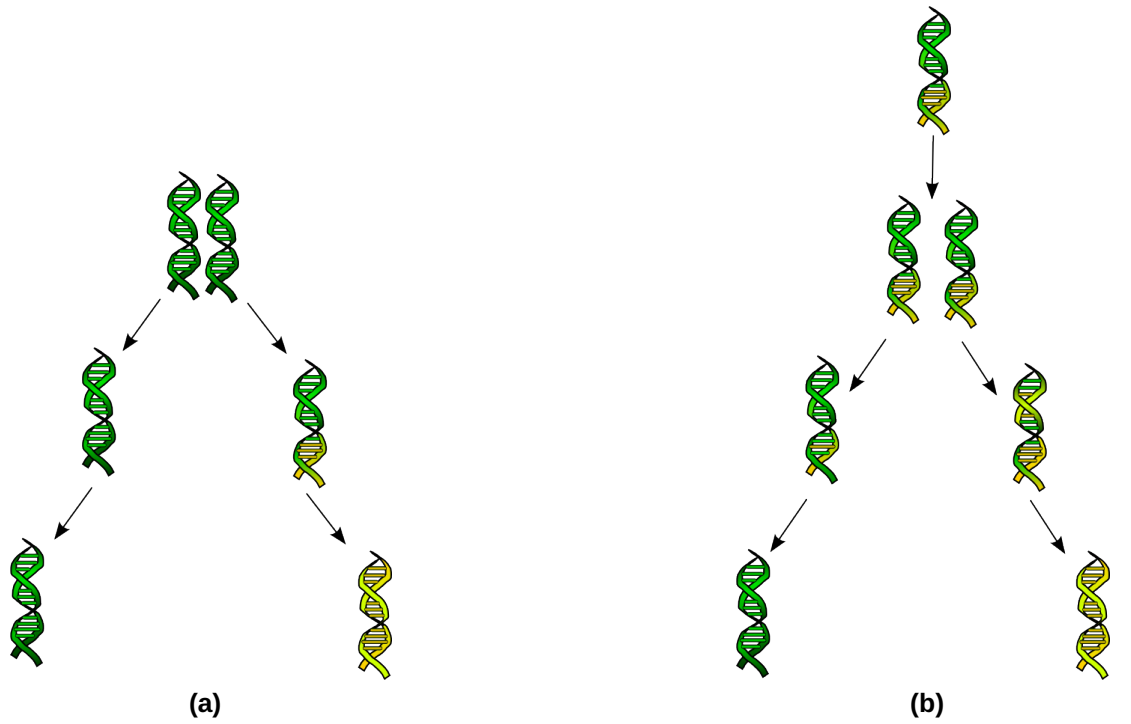


Figure 2.3.: *Two possible scenarios involving the process of duplication and neofunctionalisation. In Sub-Figure (a) the gene is duplicated before process of neofunctionalisation begins, represented by yellow colouring. This is in contrast to Sub-Figure (b) where the gene has already started the process of neofunctionalisation, coloured in yellow, before the process of duplication.*

2.2.2. The process of specialisation from promiscuity

The first enzymes

A recent paper by Carter discusses the possibility that the very first enzymes were in fact ‘urzymes’ (the ‘ur’ prefix from the Germanic root meaning ‘primitive’) [49]. It is thought that these urzymes might have composed a conserved structural core that is observed in modern day enzyme superfamilies and hence would have been much smaller in size and less complex than enzymes we see today. The authors state that these urzymes may have existed in a ‘molten globule’ state thereby allowing the same urzyme to have many different conformations [49].

What did the first enzymes look like? When considering this we must take into account their energetic surroundings. As Wolfenden points out, it was Darwin who mused that life may have began in a ‘warm little pond’ (Letter to J. D. Hooker, 1 Feb [1871]) [50]. Assuming, as Wolfenden has, that Earth was warmer than today then, in catalytic terms, this may have had big implications as to how we understand the ability of modern enzymes to overcome enthalpic barriers. Initially warm surroundings of the early Earth allowed what would be otherwise insurmountably high enthalpic barriers at current temperatures to be crossed much more easily [50](Fig. 2.4). In fact, higher temperatures are likely to increase the rate of spontaneous mutations [50]. This idea by Wolfenden [50] could be seen as analogous to a ‘greenhouse’ of enzymes, with conditions enabling a wide range of reactions to be sampled and polypeptide variants to be explored, before temperatures cooled and those achieving rate enhancements were preserved.

There is good evidence for ancient enzymes being - 1) more thermostable (e.g.[51, 52, 53, 54]) and 2) promiscuous (e.g.[55, 47, 56]) Studies have used ancestral sequence reconstruction to model the properties of ancient enzymes. Some of these studies have looked at enzyme families relevant to medicine, such as the class A β- lactamases and hypothesised that their increased promiscuity was probably due to dynamic properties rather than major differences in structure [57], tallying with the ‘molten globule’ ancient enzyme discussed by [50].

Promiscuity as an advantage for the modern-day enzyme

However, promiscuous activity is not just reserved for ’primitive’ enzymes existing on an archaic Earth - promiscuity appears to be a trait present in modern enzymes as well. As Copley so succinctly puts it - ‘Moonlighting is mainstream: Paradigm adjustment required’ [58]. There are subtle differences between the definition of ‘catalytic promiscuity’, ‘substrate ambiguity’ and ‘moonlighting’ that are briefly summarized in (Tab. 2.1). However, for the present purposes, these terms are all under the umbrella term ‘promiscuous’ meaning an enzyme able to perform

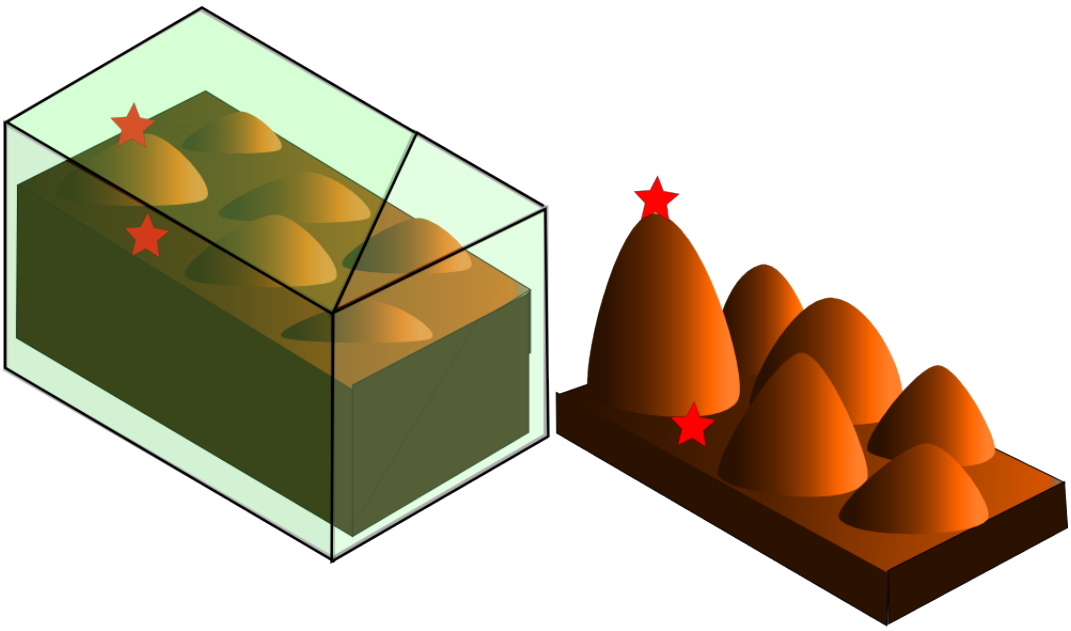


Figure 2.4.: *In this figure, a Greenhouse is used as an analogy for warmer temperatures, with the brown soil level representing the energy level landscape. On the left, in the Greenhouse, warmer temperatures allowed for a polypeptide (red star) to explore topological fold space more extensively, since the enthalpic peak of enzyme conformational changes were more accessible from a higher overall baseline energy level. This is in contrast to the figure depicted on the right, where colder temperatures make 'climbing mountains' more difficult since the baseline energetic start point is lower [50].*

more than one function - whether that be different substrates or reactions, within or outwith the principal active site.

Pandya et al. point out in their review that promiscuity is linked to evolvability - in this case, an enzyme's ability to neo- or subfunctionalise [59]. Khersonsky and Tawfik also review the mechanistic aspects that underlie the process of neofunctionalisation *via* promiscuous activities [47]. Evolvability is a term that can be used to describe different processes, as discussed by Brown [60].

Clearly, the classic 'textbook' definition of enzymes being fixed in structure and monogamous with their specific substrate ignores the wealth of evidence that promiscuity is not a trait reserved for 'primative' enzymes only but is a phenomenon that is apparently being selected for. Flexibility was found to be selected for in an ongoing process in the metallo-β-lactamases at the expense of stability [61] and may play a key role in maintaining substrate binding adaptability - key in the evolution of drug resistance. Baier and Tokuriki found that despite the distinct functions seen in this metallo-β-lactamase superfamily modern day enzymes still display promiscuity and is a sign of 'connectivity between catalytic landscapes' that may allow for novel functions to arise [62].

Catalytic promiscuity	Enzymes that catalyse substrates that are not necessarily similar to the native substrate, by chance
Substrate ambiguity	Enzymes that catalyse a range of chemically related substrates, similar to the native substrate
Moonlighting	Enzymes that have evolved other functions by use of regions other than the active site

Table 2.1.: *Working definitions of catalytic promiscuity, substrate ambiguity and moonlighting. Definitions derived from those described by Khersonsky and Tawfik [47].*

2.2.3. Evolutionary pressures shaping enzyme function

The process of duplication and enzyme promiscuity is key in allowing polypeptide chains to sample protein function space. Imagine an enzyme, recently duplicated, employing a promiscuous activity that confers fitness to the host. How does this variant gene become fixed (reaching a frequency of 100%) in a population, and go on to 'specialise' further in this promiscuous, variant activity? In the following paragraphs the theorised action of selective forces on protein structure and function will be discussed - mainly in reference to the excellent and wide ranging review by Pal et al. [63].

Positive selection

Positive selection, is a form of directional selection described by Darwin [1]. An example of a recent definition is provided by Pal et al. as '[...] the accelerated spread of a beneficial genetic variant in the population owing to the increased reproductive success of its carriers' [63]. In general, positive selection happens when functional constraint is lifted and the fitness gained by exploring sequence space outweighs the variants sampled that negatively impact fitness. Examples include arms races between host and pathogen or when compensatory mutations occur after some deleterious mutation affecting a protein's stability has occurred (an example of epistasis) [63]. For example Juárez et al. found that residues exposed on the surface of disintegrin, found in snake venom and responsible for stopping blood clotting, were under positive selection in evolution [41]. Strangely enough, these surface regions do not serve a specific functional role, but, as the authors speculate, these regions of positive selection may indicate past recruitment as a toxin - by a process of neofunctionalisation after being integrated into genes of the genome that encode venom proteins [41].

Purifying selection

Purifying selection or ‘negative selection’ is defined by Pal et al. as ‘The removal of a deleterious genetic variant from the population [arising from, for example, genetic drift or linkage to adaptive traits] owing to the reduced reproductive success of its carriers’ by [63]. In contrast to positive selection, negative selection, therefore, can be seen acting when functional constraints are conservative. Any possible fitness gain by sequence variation is outweighed by the chance that such a mutation could negatively impact protein function. In terms of protein structure and function, the less exposed to solvent a residue is, the more constrained it is by purifying selection [64]. This seems intuitive since those residues important for the stability of a protein (often found nearer the core) are more likely to be conserved. Pal et al. also go on to discuss how ‘multi-tasking’ proteins - or those with many different interacting proteins - are likely to be under higher levels of purifying selection than those with fewer interactions [63]. In this case, a mutation is much more likely to hamper at least one function if a protein has many of them, leading to less robustness in the tolerance of new mutations. Sometimes, the environment can impose these conditions - for example, Das and Misra found that functionally uncharacterised, ‘hypothetical’ proteins were under negative selection in *Deinococcus radiodurans,* known for its ability to recover from bouts of high levels of radiation [65]. The authors argue that this is indicative of these hypothetical proteins serving a fundamental and conserved role in recovery after DNA damage [65].

The neutral and nearly neutral hypothesis

As accounted by Nei, there has been much debate as to the role of selective evolution (positive and negative) versus neutral drift in maintaining the diversity of genotypes between different organisms [66]. Such a debate goes beyond this thesis, but assuming that both ‘selective’ forces of positive and negative evolution play a role in determining genotype - to what extent does neutral evolution shape the evolution of protein structure and function?

How we define a ‘neutral’ mutation is under debate [66]. For the purposes of this discussion, we define a ‘neutral’ mutation as one that does not have a substantial impact on the function of a protein in a specific biological context and on the fitness of an organism as whole. This definition takes into account Wagner’s proposal that we consider neutrality as a trait that allows for ‘robustness’ in evolution and hence evolvable innovation [67]. Wagner proposes that the innovations we see in enzyme evolution, are key examples of the role of neutral evolution in action. If some promiscuous function does not affect the ability of the native function to proceed then the mutation is neutral in *that specific cellular context.* A classic example, Wagner gives, is the product of the δ-crystallin gene which has differ-

ent functions in different tissues, and is often used as the ‘classic’ moonlighting example [7, 68].

2.2.4. Other processes in enzyme evolution

Until this point we have considered broad brush influences on enzyme evolution. Gene mutation results in observed changes to hereditary information (substitution) if not eliminated from the gene pool by natural selection. There are various processes that can happen pre-translationally and post-translationally that are not a direct result of specific nucleotide mutations but still affect how enzymes function and ultimately evolve (Fig. 2.5).

Regional genomic influences

The genomic environment of a gene means that neighboring genes do not evolve independently. It has been observed that not all parts of the genome evolve at the same rate, the most convincing evidence for such, as highlighted by Pal et al. [63], comes from studies in mammals [69]. It is those areas of the genome that are prone to recombination that tend to experience high mutation rate, as the double-strand break machinery tends to be error prone [70]. If an advantageous mutation goes into fixation it may carry with it a proximal mutation, which may be negative or neutral. This proximal mutation therefore goes into fixation not by the direct action of selective pressure but by what Smith and Haigh term ‘hitch hiking’ [71, 72]. An interesting consequence of the process of recombination is that it can break up areas in which disadvantageous mutations have ‘hitch-hiked’ thereby increasing the ‘efficiency of selection’ [63].

Gene fusion

Gene fusion can happen by exchange of genetic information between chromosomes resulting in two formally separated genes that now are translated into one protein product. Functionally, gene fusion is thought to confer benefits for metabolic processes. For example, in large cells, the compartmentalisation of different enzymes circumvents the reliance on diffusion [73]. Genes found to be fused together in the genome are often found to share similar functions [73]. In fact, in the biotechnology industry, gene fusion is sought after - with possibilities for ‘substrate channeling’ and increases in overall stabilities leading to a more efficient conversion of metabolic products [74].

Domain enlargements, recruitment & rearrangement

Insertions can provide an increase in functional specialisation and complexity [38]. Todd et al. provide a good example of the evolution of an allosteric regulatory site in phosphoenolpyruvate carboxylase - suspected to be by peptide addition [38]. As Todd and colleagues point out though, there are always exceptions to the rule and evolution does not always promote domain enlargement, it sometimes promotes domain shrinkage depending on the circumstances [38]. The combination of domains can make a difference to the overall function of a protein. For example, in the TIM barrel glycosyl hydrolase superfamily, the TIM barrel domain can combine with different domains to carry out a diverse array of functions [38].

Subunit assembly & oligomerisation

In many proteins, the assembly of identical protein chains does not change the function of a protein [38]. However, when the chains are not all identical, i.e. in a 'heterooligmer', overall changes in protein function can come about by their assembly [38]. The benefits of such an assembly can be similar to those discussed for 'gene fusion'. For example, the tryptophan synthase alpha-chain protein and the large subunit of carbonyl phosphate-synthase can associate and allow for the efficient channeling of intermediates [75, 76]. A similar channeling strategy has been shown to be used by proline catabolic enzymes PRODH and P5CDH [77].

2.3. What patterns do we see in enzyme evolution and do they match theory?

Superfamily members tend to be diverse in function - probably due to optimisation of promiscuous functionality. We see evidence of this in the lactamase and the DUF 62 superfamilies, in which a diverse range of reactions (although sharing commonalty in chemical mechanism) are present. In the metallo-β-lactamase superfamily, we have evidence that the ancestor of the metallo-β-lactamase enzyme may not have resembled either class of extant metallo-β-lactamase and may have served another function.

The diversity of functions seen in homologous superfamilies may be indicative of the promiscuous, generalist enzyme ancestors as hypothesised by Jensen 1976 [79]. It appears that in evolution, a pattern emerges where optimisation of structure is balanced with conformational flexibility. This need for balance was discussed by Tokuriki et al. in 2009 and is described as an 'evolvability and activity trade off' [80]. The scale of this balance is determined by the specific metabolic process that an enzyme is involved with, for example, Tokuriki et al.

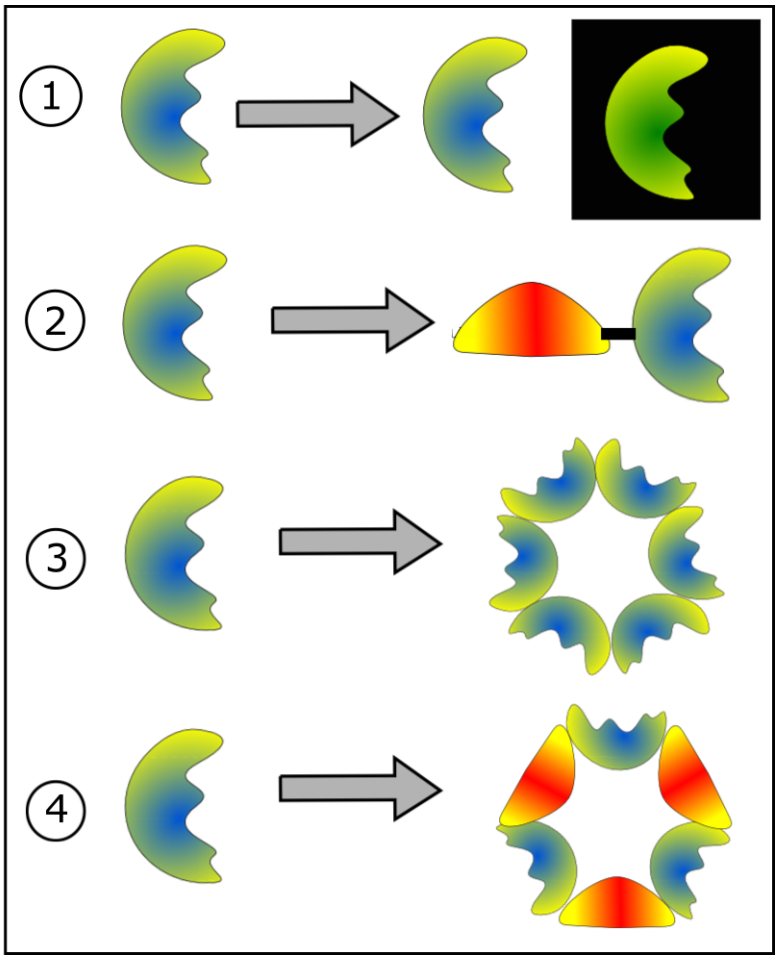


Figure 2.5.: *(1) In the process of gene recruitment a gene is recruited to another biological process where it performs a different function. For example, the δ-crystallin gene has been found to perform different metabolic and structural roles, depending on the biological context that the gene is expressed in [68]. (2) In gene fusion, the protein products of two genes are translated together in the same polypeptide chain. The fused protein product may now perform a different function. For example, in the cyanobacterium* Synechocystis, *gene fusion has resulted in a the A-type flavoproten being translated with a NAD(P)H:flavin oxidoreductase domain [78]. (3 & 4) Protein domains can assemble it oligomeric structures. These can either be with the same type of domain, homooligomers (3) or with different types of domains, heterooligomers (4). Thiamine pyrophosphate dependent enzymes use the combination of different domains to achieve different functions, although all members of the superfamily share two domains in common which bind thiamine pyrophosphate, as reviewed by Todd et al. [38].*

ask if enzymes involved in secondary versus primary metabolism might be more 'evolvable' [80]. We can imagine this reflects how secondary metabolism adapts to changing environments throughout evolution, whilst keeping 'core' primary metabolism conserved and functioning.

Structural 'scaffolds' allow for novel functions to evolve within a conserved active site. That is, within a diverse superfamily, there is often a clear, conserved structural scaffold in which catalytic residues necessary for a reaction occur [81]. But even within this conserved site occur areas of flexibility - both structurally and evolutionarily. The flexibility is necessary for the enzyme to balance evolvability and activity [80]. A key example of this is in loops that are often found to bind substrate and enable conserved structural scaffolds to evolve novel functions [80, 82].

Much work, surveying a large number of enzyme superfamilies, has found distinct trends in structural scaffold evolution. For example, Anantharaman et al. in 2003 noted that in general, catalytic residues stay conserved within superfamilies - it is the substrate binding and co-factor binding that differ [81]. The authors go on to give examples of structural scaffolds that have evolved multiple catalytic functions. Some ancient folds such as the P-Loop hydrolases are fairly narrow in terms of chemical reaction whereas others, such as the TIM barrel, are relatively broad [81]. The authors attribute this to two different 'modes' of evolution - wherein the P-loop hydrolase evolution was driven by the need for diverse substrate targets, whereas the evolution of TIM barrels by wide ranging exploration of chemical reaction space. Much like Anantharaman et al. in 2003, Todd 2001 et al. note that substrate specificity, rather than reaction chemistry, changes across superfamily members with different functions, although there is often a commonality in the chemistry of the substrates a superfamily may catalyse - even if diverse [38]. In some cases, such as the metallo-β-lactamase family, the same function appears to have evolved more than once independently. It seems that some folds are particularly amenable to certain chemical reactions. This also appears in the SAM hydrolase family in which the ability to modify SAM, by substitution or hydrolysis via the utilization of different nucleophiles, is conserved and is used in a variety of functions in this superfamily. Todd et al. in 2002 focuses on cases where the same way of catalysing a substrate has evolved twice in one homologous family [83]. Sometimes the residues between these two 'inventions' are identical, sometimes not, sometimes they lie in equivalent parts of structure, but often not [83]. In other cases, such as for the ferritins the analogous, functionally equivalent residues, are in a different active site altogether [83]. Todd et al., writing in 2001, also cite examples including the Zn peptidase superfamily [84] and the FAD/NADP(H)-dependent disulphide oxidoreductase superfamily [38].

3. How can we reconstruct the history of evolutionary events?

We can see evidence of enzyme evolution when comparing sequences, structures and enzymes across different superfamilies. How do we detect these patterns even across highly divergent families where the phylogenetic signal may be weak?

In this chapter, we discuss the methods used in the process of mapping the evolutionary history of enzymes. In the next, we look at methods that allow us to ask interesting biological questions based on these alignments and phylogenetic trees. In this thesis, we focus on phylogenetic tree construction. These trees are in fact, a subset of a wider phylogenetic network [85], in which non-vertical modes of inheritance are mapped along with the vertical ones. Such networks can be useful in the visualization of relationships between taxa, including: recombination, Horizontal Gene Transfer (HGT) and hybridization [85]. However, although Ancestral Sequence Reconstruction (ASR), as is utilised in this work, can be performed on networks, the procedure actually integrates the results of the multiple tree topologies (for example, as discussed by Arenas and Posada [86]).

Mapping enzyme evolution requires an informative data set, where the differences in specific character traits can be compared (in this case an alignment of amino acids or nucleotides). The difference, called the 'evolutionary distance' is quantified as the number of amino acid substitutions at a given site between two homologous sequences. Details of the relative substitution events at each position of the alignment can be used in the construction of a phylogeny. The tree can then be assessed by means of some statistical support measure.

Since this thesis focuses on the measurement of evolutionary distances within highly divergent enzyme superfamilies, often with a wide range of functions, sequence and structural diversity, only certain methods are appropriate. It is only these that will be discussed, particularly in their application to the enzyme superfamilies covered in this thesis, along with the benefits and challenges they bring forth.

When attempting to detect the phylogenetic signal in diverse enzyme superfamilies particular themes emerge. As discussed in Chapter 2, enzyme superfamilies tend to share a conserved structural scaffold in which different functions have evolved. This observation is congruent with the findings of Chothia and Lesk, and Illergård et al. that structure is more conserved in evolution than primary sequence

[87, 88]. That is, when analysing divergent superfamilies, a structurally informed approach will be most useful in identifying evolutionary changes.

This structurally informed approach is an important ally when dealing with highly divergent enzymes, as sometimes, the phylogenetic signal can be weak and conflicting, leading to poorly supported phylogenetic trees.

3.1. Alignments

3.1.1. Sequences & Structures

3.1.1.1. Sequence selection for phylogenetics

The data we use to infer the number of substitution, insertion and deletion (indel) events between a set of protein sequences is usually an alignment of the coding nucleotide or amino acid sequences. How do we choose these nucleotide and amino acid sequences? When we construct a phylogeny there are a number of key assumptions we make that are reflected in the corresponding choice of data set.

The fundamental assumption we make in phylogenetic construction is that our data set members are homologous (i.e. share a common ancestor). Enzyme superfamilies are a good example of this, although divergent in sequence and function, members share conserved sequence motifs and structural elements which indicate they are probably derived from a common ancestor. Of course, sometimes, similar sequences and structures can arise by convergent evolution outwith superfamilies which may lead to a false assumption of homology. However, the process of alignment, in which each nucleotide or amino acid is 'matched' to another similar amino acid or nucleotide in another sequence, iterated over all sequence positions reveals the overall similarity of a set of sequences and hence the extent of assumed homology. If two sequences had converged to have similar properties, we would not expect to see this homology over most positions in the sequence. The assertion of homology of two sequences is a subjective process, and involves the examination of homology (determined *via* alignment) throughout the entire sequence - but especially in suspected functional regions where we might expect high levels of conservation.

The next assumption we make is that the sequences, although similar enough to assume homology, have enough differences that they are phylogenetically informative. To be phylogenetically informative, a column in the alignment needs to have a distribution of characters that can be used to discriminate between different tree topologies. Examples of a phylogenetically uninformative column of an alignment include character states that are the same for all taxa, or where no

two character states are the same. Although for Maximum Likelihood (ML) methods, even columns of an alignment that share the same state are still informative - since they imply a low rate of evolution.

When looking at sequence based alignments we have two options, nucleotide or amino acid alignments. It could be argued that nucleotide alignments provide a greater depth of information about evolutionary changes in a gene. This is due to the fact that each amino acid is coded by three nucleotides forming a codon (which is translated by the ribosome). The code is degenerate, and some amino acids can be coded by more than one codon. This is mainly due to the 'wobble' in the third codon position. If a mutation happens in a codon but does not change the amino acid coded this is called a synonymous mutation. If the change results in a change in amino acid we call this a non-synonymous mutation. The strength of nucleotide analyses is that the data set reveals both synonymous and non-synonymous changes at each amino acid position, in contrast to amino acid alignments, which can only give us information on non-synonymous changes.

The problem arises in the fact that the possible state space of nucleotides is so much smaller than for amino acids. For nucleotides there are only 4 states possible whereas for amino acids there are 20 (only counting typical amino acids). So when analysing enzymes that have evolved over long evolutionary distances, the chances that two positions will have evolved to the same state by chance, rather than true homology, is much higher when looking at nucleotide alignments. This is not such a problem over shorter evolutionary distances where fewer mutations are likely to have occurred. However, since we are looking at families of highly divergent enzymes (e.g. in CATH, the cut-off for a sequence family is as low as 35% sequence similarity [89]) changes at the protein level are usually most representative of the underlying evolutionary processes.

Phylogenetic reconstruction assumes that the sequence set is an adequate representation of biological reality. A sequence data set is by its very nature limited and there tends to be a level of subjectivity in the choice of sequences, although the experimenter can make sure they use a consistent and repeatable approach, such as only picking sequences above a given significance threshold. Sequences should be from a wide range of organisms - to reflect their taxonomic distribution. Not only this, but sequences should reflect the relative proportions of organisms they are found in ensuring one group of organisms is not heavily overrepresented [90]. Finally, the best quality sequences should be used - ideally, full-length, with a solved structure in a curated database. Although, depending on data availability satisfying all these conditions is not always possible.

3.1.1.2. Sequence databases

Different sequence databases group sequences together by different criteria. Of those that are curated, UniProt is an example of a closed database covering 'all'

sequences [91] whereas RefSeq, in contrast, aims to provide a non-redundant, curated, database of sequences [92].

3.1.1.3. Uniting sequence, structure & function

We focus on sequence databases that utilise structural and functional information here - since this structural information is necessary for phylogenetic reconstruction of divergent superfamilies.

The CATH database uses structural similarity to define similar folds and structural classes [89]. In addition to making structural comparisons, CATH clusters sequences with more than 35% sequence similarity into homologous superfamilies (Fig. 3.1) [89]. The CATH definition of a homologous superfamily is used by many other programs discussed in this thesis. SCOP [93] and Dali Domain Dictionary [94] are examples of other databases of homologous enzyme superfamilies but differ in their assignment of enzymes to superfamilies. This is due to a difference in the definition of a domain and fold class [95, 96]. SCOP is the result of manual assignment into classes whereas CATH and DALI utilise automatic methods, although CATH uses manual inspection as part of the process for more difficult cases [89, 96, 97].

Databases such as MACiE [98], EzCatDB [99] and SFLD [100] focus on enzyme catalytic mechanisms. MACiE, focused on in this work, comprises a database of non-homologous enzyme chemical mechanisms. Rather than being defined by overall reaction, as is typified by the Enzyme Classification (EC) system, reaction steps are taken into account. Therefore two enzymes carrying out the overall same reaction, with the same protein fold (as defined by CATH), but by different reaction steps will be classified as two separate entries. This is useful for comparing the evolution of enzyme mechanisms [101] and for comparing cases of evolutionary convergence to the same function or mechanistic step [102].

The Catalytic Site Atlas (CSA) [103] integrates information from the Protein Data Bank (PDB) [104], EC, MACiE and the literature to annotate catalytic sites in proteins. The CSA can be used to annotate functional sites and compare a protein with other structurally solved enzymes. The CSA can also be used to construct ‘active site templates’ to help compare or discern function between different enzymes, examples of studies using active site templates include [102, 105, 106].

CATH homologous superfamilies have been used to create maps of functional evolution in FunTree using structurally informed sequence alignments to create phylogenetic trees of homologous superfamilies [107, 108]. By annotating the trees and alignments with information from the CSA and MACiE, FunTree enables the visualization of the evolution of function within a superfamily.

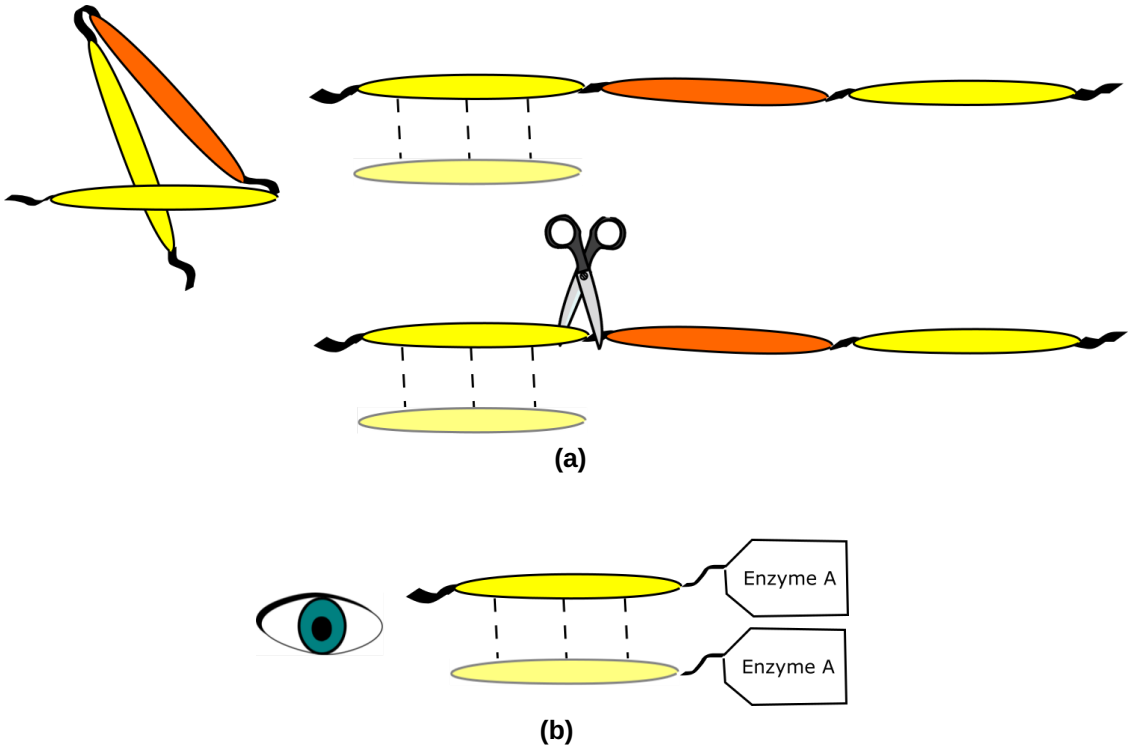


Figure 3.1.: *A schematic diagram showing the process of CATH domain searching and chopping. If chopped domain has >35% sequence similarity to enzyme in database then labeled as homologous, either automatically (a), or in trickier cases manually (b) (adapted from Greene et al. [97], Figure 3).*

3.1.1.4. Databases dealing with enzyme properties

Some databases group not by sequence similarity or structure, but offer information and specific properties for comparison . These include BRENDA [109] which annotates enzyme chemical reactions and pathways, KEGG [110] which offers information of an enzyme at the cellular level, including pathways and interactions, and STRING [111] in which genome context is used to infer functional association.

3.1.1.5. Searching for additional protein sequences

Expressions, position-specific scoring matrices (PSSMs) (Fig. 3.2) and hidden Markov models (HMMs) (Fig. 3.3) form the basis of many sequence alignment strategies, for examples of their early use see Srormo et al. [112] and reviews such as that by Yoon [113]. These strategies are used to find similar sequences from a database as compared to a query sequence. A walk-through of these techniques with basic examples is given in Figures 3.2 & 3.3 .

Examples of programs that use expressions as part of their search strategy include ProSite [114, 115] and InterPro [116] in which expressions from a range of different databases are used in an automatic process and amalgamated with manual editing. In contrast, PFAM [117] uses HMMs to populate its database with protein families.

There are a vast array of sequence searching tools - suitable for different methods. Here, we contrast BLAST [121] one of the most common search tools with HMMR3 [122] a more specialist and complex tool.

A basic BLAST search is best for finding near relatives. Variants, such as PSI-BLAST [121], have been developed to look for more distant relatives by using a constructed PSSM based on a set of aligned homologous sequences [120]. The BLAST search looks for similarity of blocks of sequence called 'words'. The sequences are then ranked by similarity score [120].

In contrast, HMMR3 is better for searching for more distant sequences. HMMR3 uses a profile as a starting point that encapsulates the properties of a whole set of sequences into one profile using a HMM (Fig. 3.4) [123]. The profile is derived by calculating the probabilities of reaching each state (amino acid) in the sequence.

3.1.2. Multiple alignment strategies

3.1.2.1. Progressive

Multiple alignment strategies aim to take a set of sequences, assumed to be homologous, and align them in such a way that the positions of amino acids are lined up to maximise homology.

The matching and weighting of alignment positions for any set of sequences depends on parameters of the PSSMs or Hidden Markov model. However, when there are many sequences in a data set, the problem becomes more computationally complicated. A balance between a strategy that is exhaustive enough to find an optimal (or near optimal) multiple alignment and one that is feasible in terms of computational resources must be found.

In short, most multiple alignment strategies are heuristic - that is, they are not guaranteed to find the most optimal solution [120]. It is therefore important to find a strategy that reduces the complexity of the problem whilst getting as close to the optimal solution as possible. Heuristic multiple alignment strategies tend to use one of two main strategies - iterative or progressive [120].

Progressive alignments are highly efficient in the way that they find a multiple alignment solution. As a first step, the data set of multiple alignments is aligned sequence by sequence in a pairwise manner to find sequences that are most similar to each other (shown as tips of the guide tree in Fig. 3.5). The nodes in

Sequence expressions

Here is a short sequence alignment:

```
MCA---YLT
GCAGGGYLW
MCAA--YVW
MMA---YLW
MCGV--YLW
```

The regular expression for this alignment:

[MG][CM][AG][Any amino acid] Y [LV][WT]

This does not differentiate between:

a) GMGR - - YVT
b) MCAG - - YLW

Where sequence a) is less likely to be homologous than sequence b) (the consensus sequence).

Position based sequence weight scheme

How do we create a scheme that differentiates between these two scenarios?

Let us take the first three columns of the alignment :

The process on the right calculates a number that is the weight of each residue, given its position in the alignment.

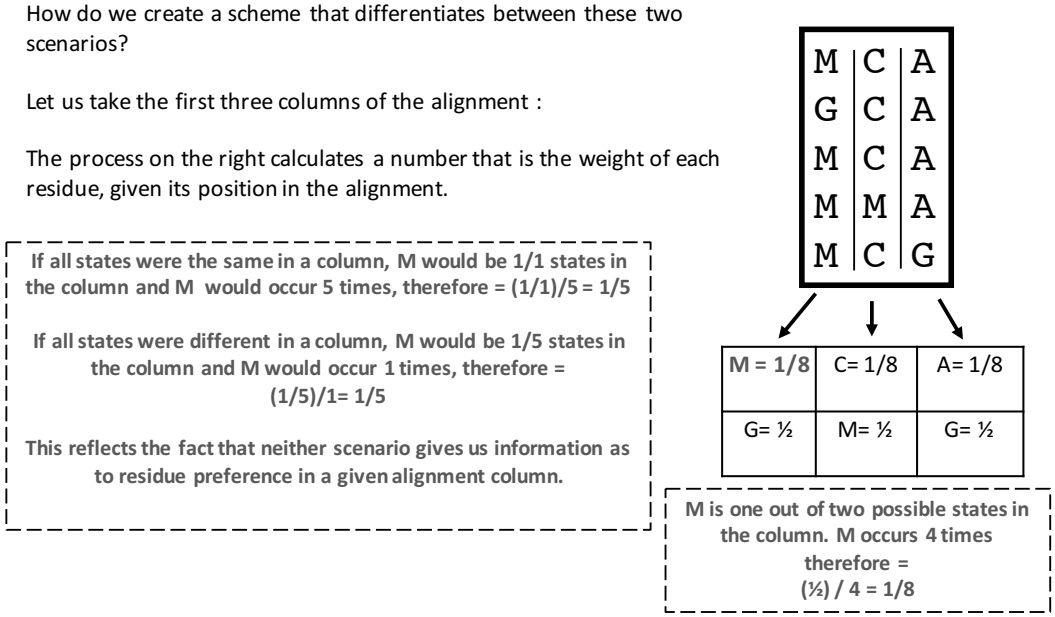


Figure 3.2.: *A walkthrough deriving a simple sequence expression and a position based sequence weight scheme. This example position based sequence weight scheme uses the Henikoff and Henikoff position-based sequence scheme [118]. The scheme assigns weights over alignment columns, rather than over whole sequences, as other methods do [118]. Figure adapted from and inspired by Henikoff and Henikoff, Krogh, Zvelebil and Baum [118, 119, 120].*

We can now use these position specific weights for each residue to define weights for each sequence:

Sequence	Column weights/3	Total weight
MCA	$\frac{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}{3}$	0.125
GCA	$\frac{\frac{1}{2}+\frac{1}{8}+\frac{1}{8}}{3}$	0.25
MCA	$\frac{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}{3}$	0.125
MMA	$\frac{\frac{1}{8}+\frac{1}{2}+\frac{1}{8}}{3}$	0.25
MCG	$\frac{\frac{1}{8}+\frac{1}{8}+\frac{1}{2}}{3}$	0.25

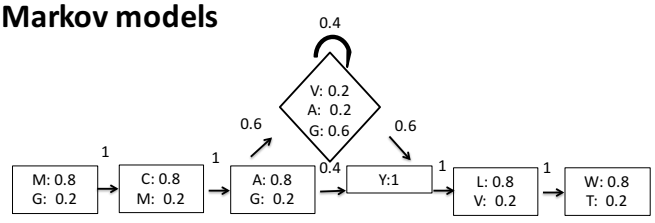
Total	1

Now we can evaluate the first three columns of our query sequence based on the position specific weights defined from our original alignment:

a) GMG: 0.5
b) MCA: 0.125

Even just working out the sequence weights for the first three columns gives us an indication that sequence a) is weighted much more heavily and therefore is less likely to be homologous than sequence b).

Hidden Markov models



Using this Hidden Markov model we can evaluate the probability of our consensus sequence belonging to our homologous sequence family–

Consensus: MCAG - -YLW

$$p(MCAGYLW) = 0.8 \times 1 \times 0.8 \times 1 \times 0.8 \times 0.6 \times 0.6 \times 0.6 \times 1 \times 1 \times 0.8 \times 1 \times 0.8 = 0.07$$

This is a very simplified example. In reality, the probabilities of each residue in each state would reflect the proportion of 20 amino acids. The assessment of a whole alignment would use a profile HMM where there would be multiple possible routes between states.

Figure 3.3.: *A walkthrough deriving a simple position based sequence weight scheme (continued from Fig. 3.2) and a simple HMM from the alignment in* (Fig. 3.2)*. Figure adapted from Krogh, Zvelebil and Baum [119, 120].*

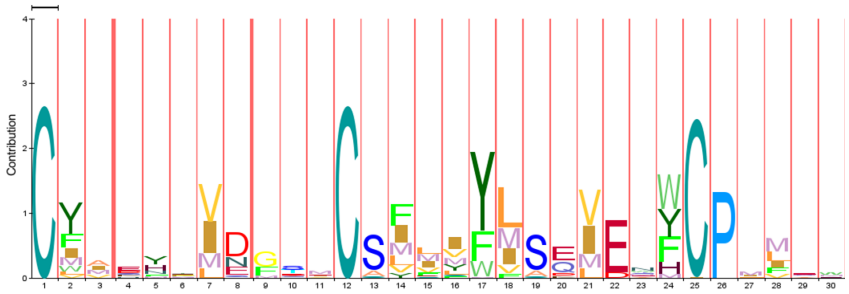


Figure 3.4.: *Pfam profile logo for glycosyltransferase family 18 (PF15024). Produced at: http://pfam.xfam.org/family/PF15024. The x-axis denotes the position in the alignment, the y-axis is the contribution measured in bits.*

(Fig. 3.5) represent intermediate alignments in the pairwise strategy. Intermediate alignments are built by aligning progressively similar sequences and intermediate alignments in a hierarchical manner forming a 'guide tree'. The problem with using this guide tree to build an alignment is that it very much depends on the the initial alignments and clustering. If these initial alignments are sub-optimal, then this is propagated through the whole process [120].

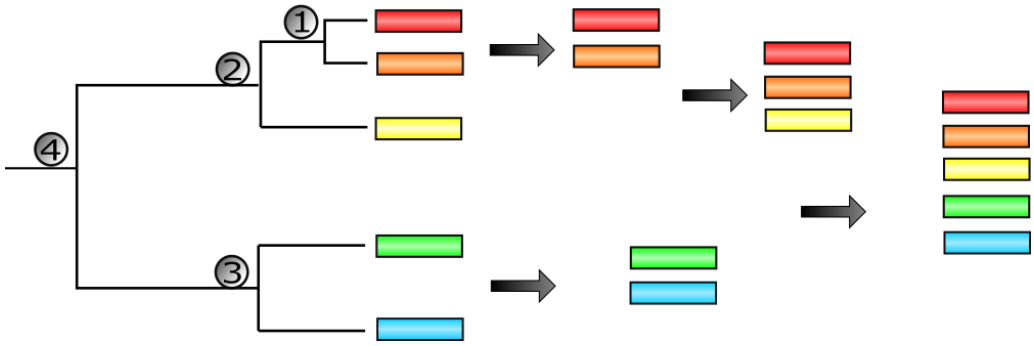


Figure 3.5.: *In a progressive alignment strategy, sequences are aligned in order that corresponds to the guide tree. In the cartoon above, sequences thought to be most similar according to the guide tree are aligned first (steps 1, 2 and 3) resulting in two groups of aligned sequences. These two groups are then aligned in the last step (4). Adapted from Figure 6.19 of Zvelebil and Baum, [120].*

Clustal [124] is an example of an alignment program that uses a progressive alignment strategy. Intermediate alignments can be weighted to correct for guide tree bias, for example, programs such as T-Coffee [125] do this [120].

3.1.2.2. Iterative

An iterative alignment strategy is one in which alternative alignments, other than those guided by an initial guide tree, are considered. Iterative alignments are

slower but often more exhaustive in finding the closest approximation of the global optimum [120].

Some iterative alignments consider all possible combinations of aligned sequences others, however, use the splits (topology) of the guide tree to inform the choice of combinations [120], thereby finding a reasonable compromise between computational time and accuracy.

The MUSCLE algorithm has been shown to be more accurate than ClustalW, which uses a progressive alignment strategy [126, 127]. The strategy employed by the MUSCLE program uses an initial guide tree and progressive alignment, from which a distance measure is calculated and used to make another guide tree and then a further progressive alignment. Finally this guide tree is then chopped up to make smaller sub-tree alignments which are then realigned to look for an increase in score [127]. The authors call this refinement and improvement of progressive alignment - an iterative strategy informed by distance measurements and subtree splits [127].

MAFFT, [128, 129, 130, 131, 132] (Multiple sequence Alignment based on Fast Fourier Transform) finds regions of homology using an efficient and novel method whereby each amino acid of an alignment is considered as a vector of its volume and polarity value [128, 129, 130, 131, 132]. This takes into account that evolution 'sees' the physico-chemical properties of amino acids, not the individual amino acids themselves. Therefore, by taking into account physico-chemical properties, the method accounts for structurally neutral substitutions - such as leucine to iso-leucine. Although somewhat indirectly, MAFFT therefore accounts for the structure-function relationships of amino acids.

MAFFT allows users to choose a strategy in aligning their multiple sequences [128, 129, 130, 131, 132]. Options range from a progressive alignment using a guide tree only to further options using iterative refinement and tree dependent restricted partitioning to inform further iterations.

3.1.2.3. Adjusting parameters for particular scenarios

As well as the strategy for multiple sequence alignment, other parameters can be changed in order to tailor the alignment strategy to individual data sets. For example, the substitution matrix used determines the weights, costs and overall costs of transitions from one amino acid to another. How are these matrices defined? We can use BLOSSUM matrices [133] or PAM matrices [134] which are derived from the empirical data of the conservation of amino acids between sequences at different levels. Depending on how divergent the sequences in an alignment are, a different substitution matrix may be more or less appropriate to model and infer evolutionary events [120].

Gap penalties can be incorporated into a HMM and are assigned as penalties (Fig. 3.3). Gaps in an alignment indicate insertions or deletions in a sequence.

These are relatively rare events, since indels occurring in functional parts of a polypeptide may have detrimental consequences. However, sometimes, when aligning highly divergent protein sequences too high a gap penalty can hinder the correct alignment of homologous blocks of sequence. Over large evolutionary distances, the chance of deletions and insertions increases, especially where a process of subfunctionalisation or neofunctionalisation has occurred. It is therefore necessary, at high evolutionary divergence to allow for the occurrence of gaps and not penalise too harshly. Alignment methods such as DIALIGN [135] work on this premise. Alternatively, incorporation of structural information as employed by Expresso [136] is the preferred choice in this thesis. For example, it is helpful if gap penalties in an alignment are weighted as a function of their position in the 3D structure of a protein.

3.2. Reconstructing phylogenies

3.2.1. Models of enzyme evolution

3.2.1.1. Transforming alignments into phylogenetic trees

The data about evolutionary events for each amino acid position in an alignment is transformed into a map of the evolution of each sequence. How do we map these evolutionary relationships between sequences based on a sequence alignment?

A multiple alignment represents the data we observe from the evolution of a group of homologous sequences. The model of evolution refers to the individual events of substitution, insertion or deletion of amino acids at each position in the polypeptide chain. There are many models available, based on empirical data on the evolution of protein sequences. We use statistical methods to find which of these already available, empirical models best fits our data. Once chosen, this model tells us about about the substitution events and rates of mutation between in different sites of the alignment, and this can be used to infer the relative evolutionary relationships between sequences in our alignment.

The information on the evolutionary relationships between our sequences can be used to construct a single tree topology, for example, in parsimony, by constructing the shortest tree that accounts for all the mutational events we have inferred from our alignment. Or, the information we have gained from our alignment can be used to distinguish between multiple candidate topologies, as is employed by Bayesian and Maximum Likelihood strategies.

3.2.1.2. Mechanistic approximations of protein sequence evolution

The *p*-distance A simple approximation of the evolutionary distance between two rows of an alignment, i.e. two polypeptide sequences, is to work out the fractional alignment distance (*p*-distance), where *L* equals the number of sites shared by two aligned sequences (excluding positions where one of the sequences has a gap) and *D* equals the number of sites that differ [120] :

$$p = \frac{D}{L}$$

This simple approximation does not take into the biological context in which sequences evolve. For example, sites may have undergone substitutions more than once, especially as evolutionary times increase. In addition, genes, chromosomes and organisms do not all evolve at the same rate (for examples see work by Lynch [137]).

The Poisson distribution can be used to correct for this. Using the Poisson distribution we can derive a probability of the number of mutations at any given site assuming a given rate. We can use this distribution to calculate the probability of a mutation event for any given site and call this the 'evolutionary distance'.

The Poisson distribution, where *n* represents the number of mutations and *rt* represents rate per site per time unit. The following ten equations are from those stated and discussed by Zvelebil and Baum [120]:

$$P(n; rt) = e^{-rt}(rt)^n/n!$$

Therefore, to calculate the probability of no mutation occurring in a sequence for a given time and rate [120]:

$$e^{-rt}(rt)^0/0!$$

$$e^{-rt}1/1$$

$$e^{-rt}$$

If we take two sequences, the probability of no mutations having occurred in each of given site for a given time and rate would be [120]:

$$e^{-2rt}$$

In terms of p-distance (which is a parameter we know) this is equal to (assuming no independent mutations have led to identical residues in both sequences by a convergent process) [120] :

$$e^{-2rt} = 1 - p$$

Since evolutionary distance (d) is approximated by the average number of mutations per site, this is equivalent to our measure of rt and can now be re-written as [120] :

$$1 - p = e^{-d}$$

But this is the probability of no mutations at either site in two sequences. To calculate the inverse, i.e. the probability of mutation for a site between two sequences, we can derive the Poisson corrected distance (d_p), [120] :

$$ln(1 - p) = -d$$

$$d_p = -ln(1 - p)$$

The gamma parameter

An assumption made in the calculation of fractional alignment distance (*p*-distance) is that rates between sites are the same [120]. A gamma distribution can be used to approximate the size of variation between sites, via one parameter, a [120]. At short evolutionary distances (p<0.2) the gamma approximation, p-distance and Poisson corrected p-distance all give similar estimates for evolutionary distance. It is at greater evolutionary distances where the estimation of evolutionary distances between the three approximations varies significantly [120].

3.2.1.3. Empirical approximations of protein sequence evolution

Mechanistic approximations, which explicitly model the biological process of sequence evolution, have been developed and become much more complicated than a p-distance approximation. It is therefore common to see complex mechanistic models being employed for mapping the evolution of nucleotide sequences. Examples of more complex mechanistic models employed for nucleotide evolution include: Jukes-Cantor [138], K80 model [139] and Tamura-Nei [140] all of which can be used with the gamma distribution to account for variability in mutation rate between sites.

When looking at protein sequence evolution, deriving a mechanistic model becomes unreliable. This is because parameter space is much smaller for four possible bases versus 20 amino acids. As such, empirical models are often used instead. These models use existing protein data sets to derive estimates about probabilities for protein mutations. The data sets sample from a wider protein population and are therefore more or less appropriate to model particular sets of proteins.

The JTT matrix [141] is derived from determination of substitution rates of similar sequences from a large protein database. This is an update on the method by Dayhoff 1978 [134], in which closely related known proteins (to try and alleviate the effect of multiple substitutions at the same site) were aligned and phylogenetic trees built. From these trees and alignments, Dayhoff and colleagues estimated ancestral sequences which were used to count the number of substitutions necessary to evolve today's extant proteins [142].

An update to this database is based on nuclear proteins is the Whelan and Goldman (WAG) [143] database based on 182 alignments of proteins. The more recent LG matrix uses 3,912 alignments and accounts for variability between sites in its matrix calculation [144].

3.2.1.4. Picking models - different criteria

The (LRT), Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) It might seem most intuitive to use the latest updated version of the Dayhoff matrix to model the evolution of a given protein alignment. This is certainly one strategy, but it must be remembered that a sequence set is a sample of a wider protein population and may have different characteristics that are better described by one model of evolution than another.

It is therefore wise to use a statistic to measure how closely a given data set matches the available substitution models. This is the process of model selection and generally, three statistics can be used for model selection - each with their own strengths and weaknesses.

The LRT The LRT can be used to discriminate between two nested models. It can be used to discriminate between use of the same model but with a different number of parameters [142]. For example, by distinguishing between two substitution models that only differ by their number of rate parameters

The likelihood ratio test statistic, with $2\triangle l$ being twice the log likelihood difference, L_0 being the likelihood of the null model, L_1 being the likelihood of the alternative model, l_0 and l_1 being the log values of each respective likelihood [142]:

$$2\triangle l = 2log(L_1/L_0) = 2(l_1 - l_0)$$

The log likelihood difference between two models can be used to determine if the more complicated model, which always fits the data better than the less complicated model, does so to significantly justify the use of extra parameters. The statistic approximates a χ^2 distribution which can be used to determine the significance of on the outcome depending on the number of degrees of freedom [142].

The LRT is somewhat limited in its use, since it can only feasibly compare between two models that are nested [142]. This is not usually the case when considering different substitution models.

The AIC and BIC In contrast, the AIC [145] and BIC [146] can be used when comparing multiple, non-nested models.

For example the AIC1 (Akaike information criterion), where p is the number of parameters, and l is the optimum log likelihood [142] :

$$AIC = -2l + 2p$$

and the BIC (Bayesian information criterion), where n is the sample size (whether that be sequence length, or sequence length multiplied by the number of sequences) [142]:

$$BIC = -2l + plog(n)$$

both determine the worth of extra parameters by assessing if they improve the optimum log likelihood of a model given the data suffieciently. The difference is that the BIC penalises models with more parameters to a greater extent as a function of sequence length than the AIC. Although, a ‘corrected’ version of the AIC exists, called AIC2 that penalises extra parameters more heavily [147].

Accounting for invariant sites Some sites in an alignment are phylogenetically uninformative. Their total conservation between sequences does not allow us to discern alternative evolutionary scenarios or topologies. These sites can be accounted for separately in model estimation with a separate rate parameter which is thought to help model the sites under variation more accurately [148, 149, 150]. However, as discussed by Yang [142], some believe that adding an invariant site proportion to a gamma distribution is rather futile, since the gamma distribution takes account of low rate at sites with $a \leq 1$. In addition, the proportion of invariable sites p_0 is not robust to the addition of sequences and their level of divergence, since more sequences with divergence leads to a lower proportion of constant sites, which in turn lowers the estimate of p_0 [142].

3.2.2. Sequential generation of tree topologies

Distances between sequences can be used for generation of a tree topology by a strategy of stepwise clustering [120]. One variant of these clustering methods is Neighbour Joining (NJ) which adopts the principle of parsimony - that is, the truest approximation of the tree will be that of the shortest length [151].

No molecular clock is assumed in NJ, unlike Unweighted Pair-Group Method using Arithmetic averages (UPGMA) [152] which produces an ultrametric tree where a constant rate of evolution is assumed - leading to all tips being equal in their distance back to the root of the tree. The root of the tree is assumed to be ancestral to the other taxa (ingroup). Whether this method is appropriate depends on the likelihood that all taxa have evolved with substitutions at an equal rate in the data set.

The Fitch-Margoliash method [153] generates a tree using a similar strategy to UPGMA. However, although additive like UPGMA, it is not ultrametric and not rooted. Its additivity means that any two branches connected to two nodes can be added to get the distance between nodes [120].

These methods are relatively quick and efficient, and so often used for large sets of sequences. However their clustering strategy means only one complete tree topology is ever explored and does not allow for the comparison of alternative tree topologies. This is problematic when the inferred topology is not representative of the true tree. For example, when assuming a parsimonious model of evolution.

3.2.3. Evaluating tree topologies using a global optimality criterion

The number of trees possible increases factorially with the number of taxa [154]. An efficient tree topology exploration will assess each topology variant encountered by a criterion ‘s’. The goal, is either to find a topology with the greatest

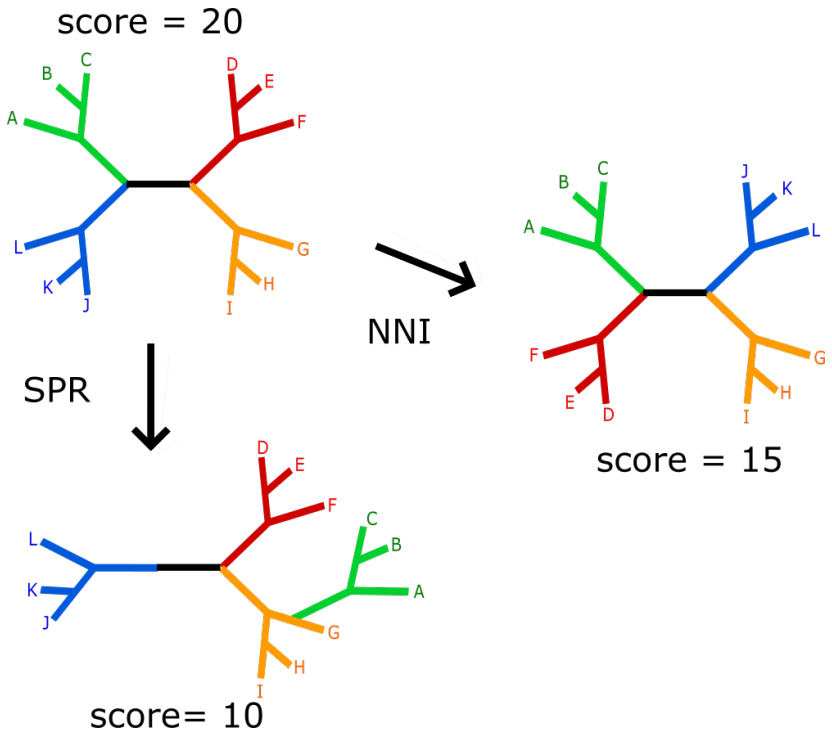


Figure 3.6.: *In exploring tree topology space, step size is important. Using NNI moves may actually result being 'trapped' in a local minimum. NNI moves simply swap neighboring subtrees by changing their relative connectivity to their shared branch. In contrast, SPR moves, which use a larger step size by 'pruning' subtrees and 'regrafting' at another point in the tree may be more successful in more widely sampling the tree topology landscape and so finding the global minimum value of 's'.*

or smallest value of 's', depending on what measure is used to evaluate trees. For example, this might be increasing the likelihood or decreasing the overall length of a tree as is aimed for in a parsimony strategy.

3.2.3.1. Search strategies

A branch bound strategy to search for optimal tree topology or topologies is an efficient way to navigate through tree topology space to a maximum or minimum 's' [120]. The algorithm starts at a baseline topology, before adding an additional sequence in all possible positions [120]. Those topologies with a higher value of 's' than the baseline tree will be rejected and only a subset of topologies will be kept for further sequence addition [120]. The branch and bound algorithm, although exhaustive in its results, is also computationally expensive [120].

3.2.3.2. Optimising tree topology

For bigger data sets, even the branch and bound method is too intensive since the number of tree topologies to evaluate increases dramatically as a function of the number of sequences in the underlying alignment.

For data sets containing more sequences and therefore more possible trees it is necessary to take smaller steps from an initial starting tree. The steps are forms of branch swapping and include, at their smallest NNI (nearest neighbour interchange), SPR (Subtree Pruning and Regrafting) to TBR (Tree Bisection and Reconnection) which is the largest of these optimisation strategies [120]. Smaller steps are more prone to getting stuck in local minima (Fig. 3.6) than methods that take larger steps [120].

3.2.3.3. An example in action - The PhyML process

PhyML is feasible for use with small to medium sized data sets. As such, PhyML relies on smaller moves to optimise a reasonable tree [155]. In its original version, PhyML used NNI moves to explore the tree topology landscape, however, this was susceptible to being caught in local minima [155]. As such, the latest version of PhyML makes use of SPR moves also. To reduce the computational burden of introducing these SPR moves a filtering strategy similar in methodology as the branch and bound method is used [155].

3.2.4. Evaluation of multiple tree topologies - character based methods

'Character based' methods including parsimony, Bayesian and maximum likelihood (ML) can be used to evaluate multiple, alternative tree topologies. How do we evaluate these tree topologies? The most simple, parsimony, will choose the tree topology which involves the fewest number of mutations to reconstruct the evolutionary history of the alignment [156]. However, this assumes a situation of minimum evolution that is not always reflective of the true tree, or data, at hand.

For a given topology, we wish to have a measure of the probability that the tree hypothesis (H) is correct given the alignment (D) that is P(H|D). Being a conditional probability, we would need to know the probability of our tree to estimate the probability of our alignment (assumed unknown). In reality, we do not know the probability of our tree but do know the probability of our alignment. So how do we infer our confidence of an unknown hypothesis (our tree) given an alignment (our data)?

3.2.5. Bayesian methods to construct phylogenetic trees

There are two ways to approach this problem. One is to use the law of joint probability to derive a way of finding P(H|D), assuming a set of prior probabilities as part of the process - this is a Bayesian approach [120].

Bayes' theorem allows for a derivation of a direct measure of confidence for a tree topology given the data. Bayes' theorem uses the law of joint probability to derive P(H|D). The following four equations are taken from [120]:

The probability of two joint events [120]:

$$P(D, H) = P(D)P(H|D)$$

This can be also written in reverse [120]: :

$$P(D, H) = P(H)P(D|H)$$

since both are equal to $P(D, H)$ then [120]:

$$P(D)P(H|D) = P(H)P(D|H)$$

If we divide both sides by $P(D)$ to get $P(H|D)$ [120]:

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

So now we have a way to calculate the posterior probability of the hypothesis given the data P(H|D) - which involves dividing the likelihood of the data given the hypothesis P(D|H) (multiplied by a prior probability of the hypothesis) divided by the probability of the data (e.g. the residue composition in the alignment).

The issue of contention with this approach is the use of a 'prior'. The prior introduces knowledge about the 'true' tree topology before the evaluation of tree topologies has begun - effectively biasing the sample to be sampled.

The problem is, how to select these priors objectively? In some cases, practitioners of Baysian phylogeny use 'flat' priors - in an attempt to circumnavigate the subjectivity in prior assumptions. However, setting a 'flat' prior for one parameter can have unexpected consequences for another parameter - for example, topology on clade size [142, 157, 158].

3.2.5.1. Maximum Likelihood (ML)

The ML method [159] assumes that picking the topology that maximises the likelihood of the data is the most likely tree, L(H|D) [120, 157]. The process of ML is summarised for a simple tree in (Fig. 3.7).

Therefore, there is no measure of objective confidence in the tree. The maximum likelihood strategy gives the probability of the data given the tree topology.

3.3. Assessing phylogenetic support and reliability

Despite the subjectivity of prior assumptions used in Bayesian phylogeny reconstruction the interpretation of phylogenetic support is much more straightforward using a Bayesian strategy. The output is a posterior probability of the tree topology and other parameters being true given the underlying alignment and prior assumptions made. In contrast, determining what a 'likelihood' means is by no means as straightforward [142]. A likelihood does not say anything about the probability of the tree topology being true within the entire population of possible trees. Rather, a likelihood gives a measure of the best-fit of the topology to the data, it does not give any indication as to how significant this is in terms of a wider population of all possible trees.

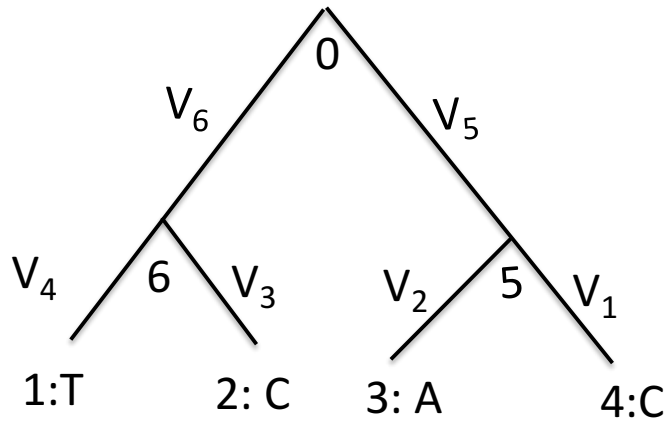
One alternative measure of support for trees is a measure of reliability of each bifurcation - this bootstrap procedure is commonly used to assess the support of trees. However, the interpretation of a bootstrap value is not straightforward - as is evidenced by the wide range of viewpoints on the matter [142].

3.3.1. What does the Bootstrap actually mean?

A bootstrap sample repeats the tree building process (with all conditions the same as the original) with a perturbed alignment which is sampled with replacement. The perturbed alignment is formed by sampling the original alignment columns to form a pseudo alignment - where columns from the original alignment may be missing or duplicated. The overall number of columns in this perturbed alignment is the same as the original alignment.

- Repeatability - It has been argued that bootstrapping is a way of further sampling the tree topology space for that alignment and can be approximated to sampling the full distribution of trees [142, 160, 161, 162].
- Frequentist - type I error rate/false positive rate. This views the probability of the tree being true as opposed to some null hypothesis that it isn't true [163].

Maximum Likelihood



Illustrated above is a tree of four species (1,2,3 & 4). At a particular site the data observed is T,C,A,C for each of the species respectively. Nodes 6, 5 are ancestral and 0 is the root. Branch lengths are measured as the number of substitutions per site and labelled as 'V'.

The probability for a nucleotide (x_hth column in the alignment) at the root is given by:

$$p(x_h|\,\theta) = \sum_{x_0}\sum_{x_6}\sum_{x_5}[gx_0 Px_0\,x_6(V_6)Px_6T(V_4)\;Px_6C(V_3)\;Px_0x_5(V_5)\;Px_5A(V_2)\;Px_5C(V_1)]$$

Where θ constitutes the branch length (V) and transition/transversion ratio parameters included in the model (g and P).

gx_0= Prior probability that node 0 has nucleotide x_0, $\left(\frac{1}{4}\right)$ according to K80 model.

Px_0= Probability of transition of unknown sequence at $x_{n\,to}\;x_m$.

We do not know the sequences for ancestral nodes. Therefore all possibilities for nodes 5 and 6 must be taken into account.

We assume each site to be independent, so the product of the probabilities at each site gives us the probability over the whole tree. An alternative measure is the log likelihood which takes the sum of all sites:

$$l = \log(L) = \sum_{h=1}^{n} \log\{p(X_h \mid \theta)\}$$

Figure 3.7.: *A walkthrough of the maximum likelihood method for a simple tree. Adapted from Yang[142], Figure 4.1.*

- Accuracy - this is the most ambitious of interpretations [164]. This sees the result as an equivalent of the probability that the given clade is present on the true tree.

In this work, the closest approximation we make is that the bootstrap is a measure of repeatability. We approximate that perturbing the data set by bootstrapping is an approximation of building trees from another data set derived from the same genes with the same parameters. We look for common features in this foray into tree topology space, and use it as a guide to test the robustness of our ML tree.

3.3.1.1. Alternatives to the bootstrap

The bootstrap is computationally costly and can be difficult to interpret. There are variations and alternatives to this procedure. For example the bootstrap interior branch test [120]. The data is permutated and sampled with replacement - much like bootstrapping. However, in this scenario, the tree topology is constrained to remain the same and the branch lengths differ. The test aims to look at the proportion of zero length branches inferred when re-sampling the data set as a measure of confidence in the branch [120].

Despite challenges in interpreting the bootstrap we found the bootstrap allows us to explore alternative scenarios of evolution. The lack of clear consensus within our bootstrap set may have been indicative of a particularly rugged, or particularly smooth likelihood landscape [102]. Other methods, such as the interior branch strategy do not allow for the direct exploration of other topologies by the user. A Bayesian strategy would have constrained the distribution of topologies to explore - resulting in a less exhaustive search.

3.3.1.2. Long branch attraction

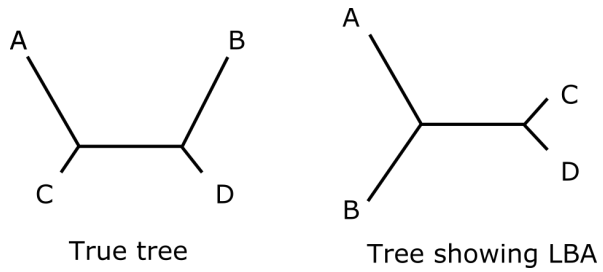


Figure 3.8.: *In Long Branch Attraction, the wrong tree can be inferred due to the 'attraction' of long branches. This is shown in this figure as taxa A and B being inferred as sharing a common ancestor on an unrooted tree when in fact, taxon A is a sister group to taxon C and taxon B is a sister group to taxon D, as depicted in the 'true tree'. Adapted from of Zvelebil and Baum [120], Figure 8.21.*

Even if trees are assessed as optimal and well supported by the data artifacts can occur that lead to misleading topologies. One of these is Long Branch Attraction (LBA) (Fig. 3.8) and is a particular problem when using a model of parsimony when building phylogenies e.g.[165, 166].

The phenomenon is seen when longer branches are wrongly grouped as sister taxa on a phylogenetic tree [142]. This happens because statistically, those branches with more substitutions are likely to share similar amino acid compositions at any one site through sheer chance [142]. This can be misinterpreted during phylogeny building as close homology [142].

This problem can be overcome by using a more sophisticated model of evolution, which incorporates different rates of evolution for different lineages although the model has to be realistic, even if more complex [142]. It is also important that sequences are selected carefully for phylogenetic analyses. For example, the use of adequate taxon sampling breaks up the long branches into a larger number of short branches [167, 168].

Being aware of the possibility of LBA in our work, we have carefully selected models of evolution based on the data at hand, stayed away from maximum parsimony strategies and carefully selected sequences that are from a wide range of taxa and roughly representative of taxonomic group size [102].

4. Investigating the evolution of function *in silico*

Assessment of enzyme function as restricted to 'wet' laboratory assays is an excellent way to discern function at a case-by-case level. However, bioinformatics has the power of empirical knowledge and pattern detection - therefore enabling bioinformaticians to make informed, intelligent hypotheses about function that can be ratified in the lab. An approach such as this, using bioinformatics in conjunction with more 'traditional' lab techniques, allows for a greater area of enzyme function space to be explored whilst being much more efficient in terms of time and money.

We can use phylogenies to explore evolutionary relationships between biological sequences. Sometimes, having a 'bird's eye view' of enzyme data (as bioinformatics does) can reveal patterns that would not be obvious by lab work alone. Having an idea as to the order in which sequences have evolved, we can ask interesting questions. These questions might revolve around sequence divergence or ancestral sequence inference and can provide the data for further more specific investigations on enzyme function in the lab.

The study of bioinformatics is, by definition, restricted by the data available. Therefore, there exists an inherent bias in families with more data which then tend to be more amenable to bioinformatic methods and yield more statistically and biologically robust results. Not all possible methods are discussed in this chapter, but rather those used in this thesis are evaluated.

4.1. Ancestral sequence/state reconstruction (ASR)

4.1.1. Theory

Originally proposed by Pauling and Zuckerkandl in 1963 [169], ancestral sequence reconstruction involves the inference of character states for each column of the alignment based on extant character states (Fig. 4.1). As discussed for different tree building methods, a parsimony based strategy cannot quantify the certainty of any ancestral reconstructions. Maximum likelihood and Bayesian approaches consider a range of possibilities, due to their ability to incorporate more complex models of protein evolution. Fully Bayesian methods give a probability

but are based on subjective prior assumptions. Although, some level of subjectivity in prior assumptions is in common with any method.

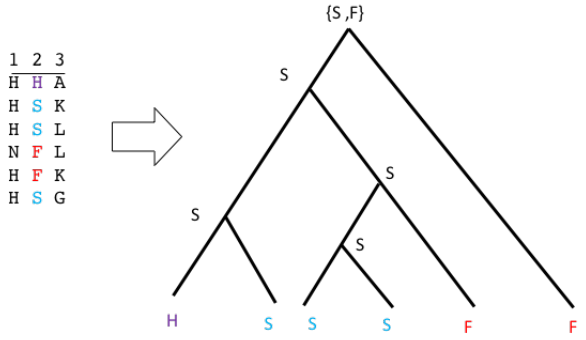


Figure 4.1.: *Illustrated is a simple parsimonious approach for carrying out ASR on a phylogeny for the column 2 of the above section of an alignment. Ancestral nodes are labelled with possible states going from leaves to tips, before resolving state unions by passing back down from the root to tips. More sophisticated models can be incorporated in ML and Bayesian methods.*

4.1.2. Applications

Ancestral sequence reconstruction is important for understanding how protein functions within homologous protein families have evolved. By looking at the sequences of ancestors we can infer possible functions (either currently extant or not) and use the information to model the trajectory of past evolution [56, 102]. In some cases, methods that reveal patterns in the evolution of function within a superfamily can be used to understand the evolution of modern day proteins [170].

4.1.3. Different methods and software available

In some ASR studies, a Bayesian method of ancestral sequence reconstruction (e.g. by Huelsenbeck and Ronquist [171]) is preferred as its results are more easily interpretable - with the posterior probability value being read as the degree of confidence in a given ancestral reconstruction. It has been demonstrated that a fully Bayesian strategy can be more robust in the modelling of thermostability of protein sequences [172] and so has been the method of choice for some [52, 57]. However, a study comparing ASR Bayesian and ML analysis of LeuB enzymes showed that an ML strategy actually generates kinetically more feasible estimates in contrast to those generated by a Bayesian strategy [51].

Along with other authors, including Edwards and Shields, and Ashkenazy et al. [173, 174], we have chosen to use an ML ASR strategy [102]. ML ASR predictions

use an empirical Bayes method [175, 176], where prior weightings are derived from the data at hand. Our justification for this lay in the need to set subjective priors and the limited sampling of trees as is necessary in a fully Bayesian ASR analysis [102].

4.2. Homology modelling

4.2.1. Theory

Ancestral sequences can be used to model the possible functions of ancestral proteins. Clues as to the function of such a protein might be found in its amino acid sequence. However, the linear sequence of a protein is not representative of its native, folded 3D structure. Although we code these complex structures as strings of amino acid units for the purposes of alignment, this is not representative of function *in vivo*.

The process of homology modelling can be used to infer the likely three dimensional structure for a given protein sequence. This is possible due to the fact that structure is more conserved than sequence [87]. Even though two sequences may have little in common at the sequence level, they still may fold into a similar structure. Comparing the structure of two proteins can help amplify any low level signal of homology that may not be obvious at the sequence level - given the plethora of evolutionary events that can change the exact sequence of a protein. Using this observation, we can extrapolate that certain sequences tend to fold in certain ways and use this empirical information to build homology models.

A detailed explanation and discussion on the topic of homology modelling is beyond the scope of this thesis. However the main principles and steps are discussed here:

Structural homologs are found and aligned to template

The query sequence, from now on referred to as the ‘target’, must be queried against other proteins in a structural database (e.g. the PDB) to look for proteins with a similar sequence, or a similar section of sequence [120]. At this point one or more of these sequence matches (now referred to as a ‘template’) is aligned to the target sequence [120].

The accuracy of this alignment is incredibly important. Since a template is being used to infer the 3D structure of a protein, the template and the target must be correctly matched up. Mismatches can have big consequences for the integrity of the downstream structure. For example, when charged residues are modelled to

lie within a hydrophobic core, this goes against our prior empirical knowledge of protein structure [177].

Modelling the structurally conserved core

Conserved parts of the protein, that align to the template without major insertions or deletions, are modelled first [120]. These tend to correspond to secondary structure elements such as beta sheets and helices. Amino acid side chains are also modelled at this stage. In more than 90% of cases, the amino acid conformations can be directly transferred between query and template [120]. In cases where this direct transfer is not applicable, empirical information as to the most common conformations of amino acid chains in the form of a rotamer library is used [120].

Modelling less conserved parts of the sequence

Gaps and deletions in the alignment are modelled as highly flexible loop structures, which are most likely to vary in evolution. They are often the site at which diverse substrate specificity evolves within superfamilies [178, 179]. It is therefore necessary to model these structures in order to elucidate function - although their variability makes this challenging. A template approach surveys a database of known protein structures to find a match with sufficient sequence similarity to the loop sequence so that its structure can be inferred [120]. Alternatively, *ab initio* methods can be used to predict a loop structure using empirical knowledge of amino acid geometry [180].

4.2.2. Different methods and software available

Determining the possible structure of a polypeptide sequence is highly in demand for researchers in many different fields. As such, there are a variety of online tools that automate the process of homology modelling, e.g. the I-TASSER and Phyre servers [181, 182]. These are assessed by the Critical Assessment of Structure Prediction (CASP) [183] and the Fully Automated Structure Prediction (CAFASP) [184]. The assessment consists of researchers using their software to predict the structure of proteins whose 3D structure has been solved - for example by X-Ray crystallography or NMR - but whose structure has not been released into the public domain. The attempt at predicting this known 3D structure is then assessed by independent reviewers or, for the case of CAFASP, by an automated evaluation technique.

4.3. Reconciling gene and species trees - inferring evolutionary events

Gene and species trees often differ. These differences can be used to diagnose gene duplication, transfer and loss events. It is important to know the history of a gene in this genomic context because duplication events can be the seeds for neofunctionalisation and subfunctionalisation.

4.3.1. Theory

Reconciling a binary species tree and a binary gene tree is conceptually quite straightforward . When mapping a gene tree and species tree inferring transfers, losses and duplications are kept to a minimum - a program such as Notung uses a parsimonious strategy [185]. As such, the duplication, loss and transfer score (π) is minimised [186]:

$\pi = c_D D + c_L L + c_T T$

However, the inclusion of transfers makes the mapping more complex. Inclusion of the possibility of transfers means that the gene tree is no longer constrained by the species tree topology and multiple solutions to the mapping problem can be found - sometimes these are equally optimal [185] [186].

The weighting of these events can be changed by the user. This is important when considering different biological contexts, for example, the cost of transfers could be down-weighted to reflect horizontal gene transfer prevalence in a prokaryote phylogeny [187].

4.3.2. Different methods and software available

As discussed, the inference of evolutionary events is important for many applications. As a result, there are many algorithms and programs available that can be used to infer HGT events, some of the more well known being RIATA-HGT [188], EEEP[189] and Prunier [190].

It is necessary to take into account transfers, duplications and losses in order to accurately reconcile species trees and gene trees. Notung is unique in the way that it considers a broad range of evolutionary events, including duplication, transfer, loss and incomplete lineage sorting [185].

4.4. Comparing evolutionary rates between different phylogenetic groups

4.4.1. Theory

How do we ask questions about the evolution of sequence, structure and function after a transfer or duplication event?

As discussed in Chapter 2, duplication events provide an unconstrained copy of a gene which is then free to accumulate mutations and possibly, change in function [35]. The difference in evolutionary constraints between these two genes can be detected as a difference in evolutionary conservation [191].

How do we detect these differences? One way might be to analyse the proportion of non-synonymous rates d_N of substitution as compared against synonymous rates d_S of substitution, as expressed by the normalised ratio ω used in software such as PAML [192]:

$(\omega = d_N/d_S)$

A greater proportion of non-synonymous than synonymous mutations indicates positive selection (ω= >1), a greater proportion of synonymous mutations than non-synonymous indicates negative selection (ω= <1) and the same frequency of non-synonymous and synonymous mutations (ω =1) indicates neutral evolution. One of the problems with this approach is that the codon level can become swamped at large evolutionary distances (as discussed in Chapter 3).

Algorithms exist to compare substitution rates at the amino acid level. Two examples we will consider are the Evolutionary Trace [193] method and DIVERGE [194, 195].

Both methods employ the fact that at the amino acid level, lack of conservation of particular regions of a protein can imply a lifting of evolutionary pressure and therefore potential for neofunctionalisation. Residues conserved within their functional subgroup are called ‘class specific residues’ [196]. Both Evolutionary Trace [193] and DIVERGE [194, 195] assume that functional subgroups can be inferred by sequence conservation.

In the Evolutionary Trace method, these class-specific residues are ranked based on the number of divisions a tree must undergo for them to be class specific [196]. As such, those with a rank of a low number (1,2,3...) represent class specific residues that occur and are conserved near the root of the tree - their conservation early in evolution may indicate a fundamental role in protein function [196]. Class residues with higher numbered ranks (4,5,6...) represent those in which only small subgroups are the residues conserved [196]. This may be indicative of subfunctionalisation or neofunctionalisation. As the ranks get higher in number

the power to detect signal from noise decreases with decreasing group/sample size [196].

These ranked class specific residues can then be mapped onto structure. Class specific residues that cluster spatially can be indicative of functional sites [196]. If these differ between two diverged groups this may be indicative of a change in function.

The problem with evaluating conservation of residues this way is that there is no statistical basis on which to differentiate functional divergence and neutral drift. A statistical framework needs to be implemented to assess the significance of differences detected. Gu uses the measure of evolutionary rate between two subgroups as a proxy for a measure of conservation [197]. DIVERGE tests whether the rates of evolution between two subgroups are correlated after a duplication event [197]. If the subgroup rates are correlated, a low coefficient of divergence is assumed. If, however, the subgroup rates are found to be independent a high coefficient of divergence is inferred.

Gu et al. measure the extent of type I functional divergence between cluster 1 and cluster 2 by the coefficient, θ_{12} [198]. It is equivalent to $P(S_1)$, the probability that at least one of two clusters being compared is under functional constraint [198, 197]. In this state, the rates between equivalent sites in two clusters are statistically independent [198, 197]:

$\theta_{12} = P(S_1)$

Therefore, 1-θ_{12} is the probability that neither of the two clusters are under functional constraint. In this state, the rates between equivalent sites in two clusters are not statistically independent [198, 197]:

$1 - \theta_{12} = P(S_0)$

The significance of the value of this coefficient can be calculated using a Likelihood Ratio Test [197] :

$H_0 : \theta_{12} = 0$

versus [197]

$H_A : \theta_{12} > 0$

The likelihood ratio statistic ($2\Delta l$) can then be tested for significance against a χ^2 distribution with one degree of freedom [197].

If there is enough evidence to reject H_0 with significance then we can ascertain that there is sufficient evidence that functional constraints differ enough between the two genes that they can be seen as two functionally divergent groups of enzymes.

Part II.

Exploring the evolution of antibiotic resistance in the metallo-β-lactamase superfamily

5. Reconstructing the evolution of the metallo-β-lactamase superfamily

One mechanism of survival for antibiotic resistant bacteria is their ability to hydrolyse β-lactam molecules. The ability of bacteria to do this is one of the key mechanisms of antibiotic resistance [199] and predates modern medicine [200, 201]. In fact, a study in [202] dates the emergence of metallo-β-lactamase function to more than two billion years ago. Despite this early emergence of metallo-β-lactamase function, the intensive use of antibiotics by humankind has encouraged the appearance of resistant strains, which are a growing threat [203]. Some bacterial enzymes have evolved the ability to hydrolyse a wide range of substrates - making the design of a specific inhibitor more challenging [204]. These broad spectrum enzymes include the metallo-β-lactamases which use activated water in nucleophilic attack of the lactam ring via a zinc ion(s) bound in the active site [205]. This is in contrast to the serine-β-lactamases, which although similar in function (and E.C. class) to the metallo-β-lactamases, use a serine residue to mitigate nucleophilic attack on the lactam ring rather than a bound zinc ion [206].

The metallo-β-lactamase gene is found in many pathogenic gram negative species of bacteria including: *Bacteroides fragilis*, *Pseudomonas aeruginosa, Aeromonas hydrophila, Serratia marcescens* and *Elizabethkingia meningoseptica* [205]. It appears that the gene (carried on a mobile DNA element) is becoming widely distributed [209]. The gene's existence in the genomes of 'environmental species' constitutes an additional threat [210, 211, 212, 213].

It is therefore important that we understand the distribution, transmission and evolution of this gene. By the use of phylogenetics we can pinpoint organisms and trends that might be most fruitful to study in the design of new antibiotics [214]. Studying past patterns of evolution using phylogenetics cannot be used as a 'crystal ball' to predict future evolutionary events, but it can help pinpoint more likely ones [215, 216, 217, 218].

Building a well-supported phylogeny for the metallo-β-lactamase family has proven a challenge. Past studies have focused on the B1 and B3 lactamases only, building separate phylogenies for these two groups [202]. Others have used a small dataset of structural representatives to build a phylogeny of the whole

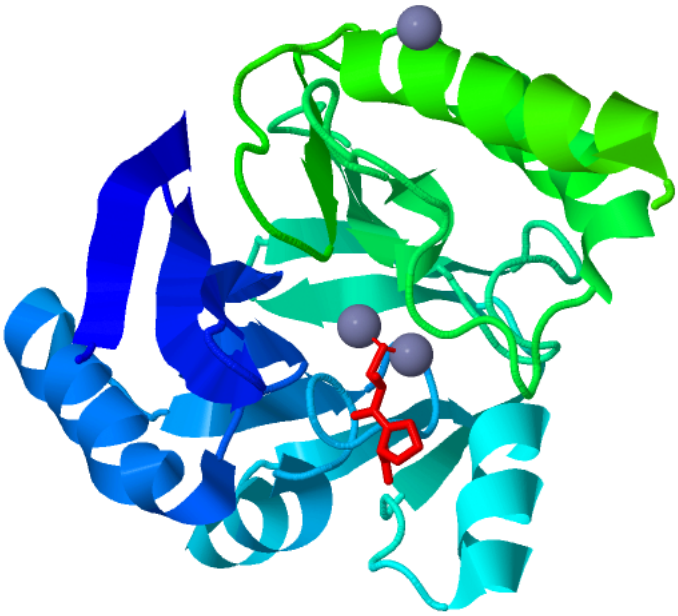


Figure 5.1.: *Crystal structure of 1M2X, a BlaB metallo-beta-lactamase from* Chryseobacterium meningosepticum *bound with* D*-captopril inhibitor (red) [207]. Zinc ions are shown as grey spheres, those bound in the active site are important for the lactamase reaction. Figure generated using Jmol [208].*

family [219]. However, despite these efforts, statistical support for divergences for relationships between more divergent superfamily members remains low.

Part of challenge in attempting to build a phylogeny for this superfamily lies in the fact that enzyme functions in this superfamily are so diverse. Annotated in the CATH database (CATH 3.60.15.10) [220] this family includes the metallo-β-lactamases, A-type flavoproteins, the glyoxalase IIs and the RNase Z enzymes. The structural and functional diversity of these enzymes is such that they are separated into two clusters of Structurally Similar Groups (SSGs) by the protein structure–function phylogeny suite FunTree [108, 107] for the purposes of alignment and phylogenetic tree building. SSGs are the result of clustering portions of protein sequences that contain only the domain of interest based on their sequence and structural similarity [108, 107].

At the level of the metallo-β-lactamases, there exists another level of separation which is much contended in the literature. The metallo-β lactamases have traditionally been classified into three subclasses B1, B2 and B3 [221], by virtue of their sequence identity or substrate specificity. However, the degree of evolution-

ary separation between these subclasses appears not to be equal - with the B1 and B2 subclasses sharing more more sequence identity to each other than to group B3 [222]. Despite this difference, common catalytic and mechanistic features are shared by the B1/B2 and B3 subclasses. Namely, the amide bond of a lactam ring is hydrolysed by a zinc-activated water initiating nucleophilic attack at the carbonyl carbon. The area of contention lies in the fact that although the overall mechanism is the same in these subclasses, the transition state is stabilised by different residues [223, 224, 225, 226, 227].

This observation is congruent with the general substrate profile of this family. That is, excluding the flavoproteins, although substrate identity varies, the overall mechanism of hydrolysis is conserved. As discussed by Aravind et al. [39] this superfamily seems to exemplify the case in which different substrates have been accommodated by a conserved scaffold through the course of evolution [83, 81]. As such, the subject of classification of the B1/B2 and B3 metallo-β-lactamases remains contentious. In fact, some argue that the B1/B2 and B3 may represent independent evolutionary inventions and this should be reflected in their classification, no matter how similar they might be in function [228]. The difficulty in discerning the relationship between the B1/B2 and B3 subgroups is compounded by the fact that although these subgroups are similar in overall structure, they have distinct and individual structural features [223].

Here, we utilise the FunTree resource as a base to generate an improved phylogeny for the whole superfamily - by widening the distribution of taxa included in the alignment, using an ML strategy with a model of evolution selected to fit the data.

5.1. Past work by others and available data

A structurally informed approach can be an aid in diagnosing signal in evolutionarily diverse superfamilies. It therefore comes as no surprise that previous studies have used structural information as a basis for alignment and tree building.

Here, we contrast the alignment generated by Garau et al. [219] with that produced by the FunTree resource [108, 107] which uses different strategies to generate a structural alignment.

5.1.1. Alignments

Garau et al. [219] used a small set of structural representatives to create an alignment of the metallo-β-lactamase superfamily. This sequence alignment was then edited to only include those parts of the structural alignment that were conserved (Fig. 5.2). In doing this the authors ensured that positions in the alignment were

aligned to maximize identity. However in doing so, there is concern that key evolutionary information may have been disregarded.

FunTree, in contrast, uses a different strategy to maximize identity in alignment positions between structural representatives [108, 107]. Using the CATH classification of homologous superfamily domains (in this case 3.60.10.15), it is recongnised that superfamily members can be so diverse it is not always easy to align them at the structural core. Therefore, the FunTree strategy clusters superfamily members into Structurally Similar Groups. These SSGs can then be more confidently aligned (than the whole family).

The alignment by Garau et al. includes many of the same protein representatives as those found in FunTree, but does not share any of the same representatives of the flavoproteins as the FunTree phylogeny. Interestingly, 1VJN & 1WRA share the metallo-β lactamase CATH code (predicted by CATHEDRAL [229]) but are put into separate SSGs in FunTree [108, 107].

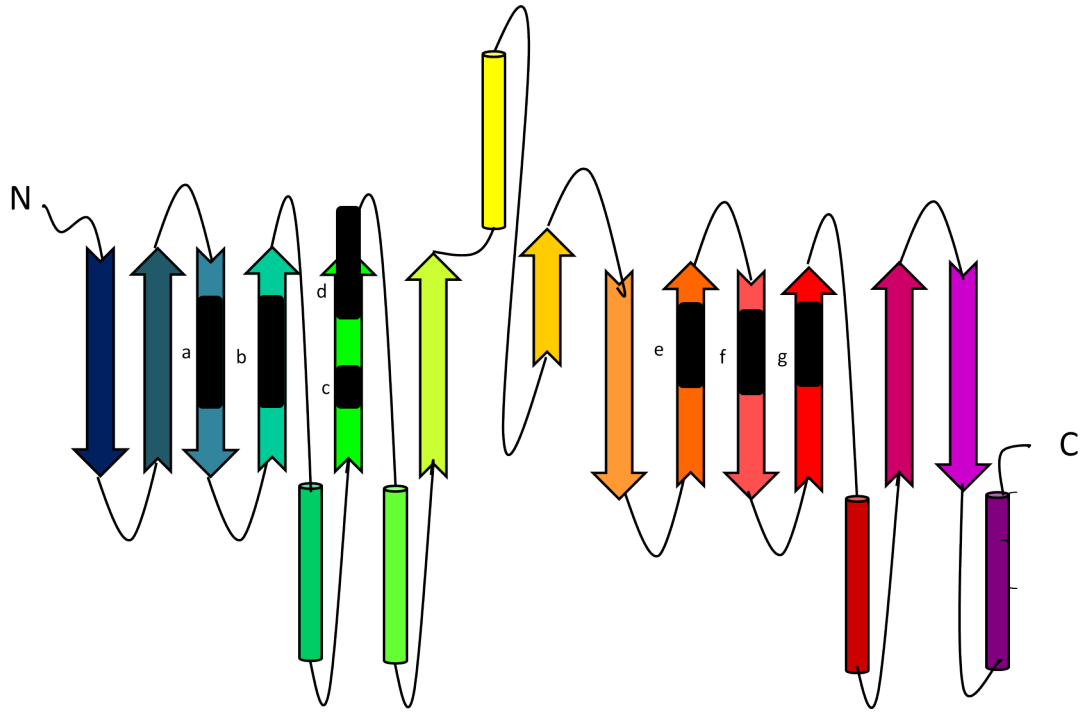


Figure 5.2.: *Schematic illustration of the secondary structure of the metallo-β-lactamase topology as presented by [230]. Sections a - g of the structurally conserved regions used in the alignment by Garau et al. [219] are highlighted. Adapted from Garau et al. [219].*

In the alignment by Garau et al. residues such as: Tyr 191 in 1SML & 228 1K07 and Cys 141 in 1QH5 (numbering as documented in the Catalytic Site Atlas [103]) thought to be involved in catalysis [103] were not included in the alignment. However, even though these residues lie in areas not so well structurally conserved,

their inclusion is important in the building of phylogenetic trees that faithfully represent the history of evolution in this superfamily.

The above comparisons demonstrate that there are fundamental differences between the alignments of Garau et al. and FunTree. One of the main reasons for this may lie in the fact that Garau and colleagues concatenated all B1 sequences as a first step and then used this resulting sequence as a template for the rest of the alignment. By doing so, the authors actually biased the alignment to include features found in the B1 class of metallo-β-lactamases.

5.1.2. Phylogeny

As discussed, the FunTree strategy gives results that are more accurate in the alignment of residues of functional importance than that of Garau et al. [219]. We then went on to examine how differences in the underlying alignment strategies used by Garau et al. and FunTree impacted the resulting phylogeny.

Determining the evolutionary ordering of the major functional groups in this superfamily is by no means a trivial task. In part, this is evidenced by the work of Garau et al. [219] in which phylogenies based on structural diversity scores versus those based on a structurally informed sequence alignment differ extensively in their topology.

The FunTree phylogeny agrees with the phylogeny based on structural alignment as generated by [219] in that the B3 and B1/B2 lactamases form two distinct groups that diverged back in evolutionary history (Fig. 5.3). The clear evolutionary separation of these groups evidenced by both FunTree and [219] adds weight to the argument by Hall and Barlow [228] that despite functional similarities of chemical mechanism, the B1 and B3 classifications of these enzymes should remain distinct.

However, the relation of these two major groups of lactamases is a source of contention. The phylogeny by Garau et al. implies that the B3 subclass is more closely related to the glyoxalases and rubredoxin than the B1/B2 group which is most closely related to rubredioxin [219]. This is in contrast to the FunTree phylogeny, where the B1/B2 group is more closely related to the glyoxalases, and the B3 group most closely related to the flavoproteins.

5.1.3. The contribution we make to this field

Constructing phylogenies of divergent superfamilies represents a formidable challenge. Despite this, we felt that the existing attempts for this superfamily could be improved by use of strategies shown to be useful in the literature.

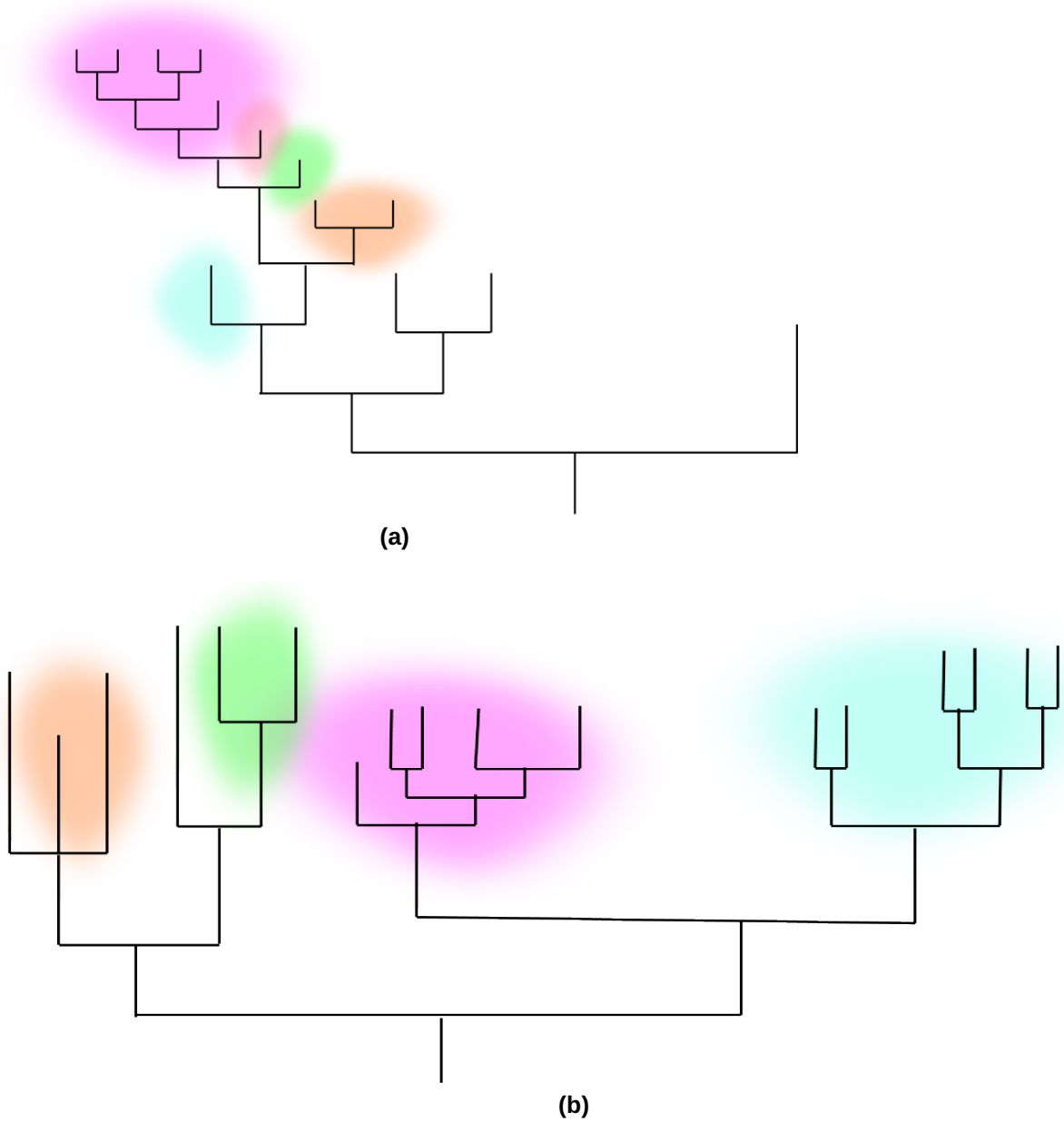


Figure 5.3.: Top*: A phylogeny based on structures of the metallo-β-lactamases by Garau et al.[219]. Figure adapted from Garau et al. [219]* Bottom*: Phylogeny adapted from FunTree.[107, 108] Enzyme functions are colour coded as follows:* cyan *glyoxalase IIs,* green *A-type flavoproteins,* pale pink *subclass B2 metallo-β-lactamases,* magenta *B1 metallo-β-lactamases,* orange *B3 metallo-β-lactamases,* no colour *function not included in FunTree SSG1 phylogeny. Branch lengths are not biologically meaningful.*

The first strategy was to increase the size of the dataset by adding additional homologous sequences. By doing this, we hoped to increase signal to noise ratio, break up long branches [168, 167, 231] and provide a more realistic and broad ranging view of evolution. We felt that after inspection (discussed above), the FunTree alignment did an excellent job of aligning homologous and important catalytic residues. Therefore, we used the FunTree alignment of CATH domain 3.60.15.10 as a profile to search for similar additional sequences and as a base during alignment (discussed in more detail in sec. 5.2).

The ML method used by FunTree involves fewer unrealistic assumptions than the neighbour-joining used by [219]. FunTree uses a JTT model of evolution [141] by default. We hoped to improve on this by using a statistical framework to assess the best model of evolution for our alignment (see sec. 5.2).

The FunTree software suite generates its phylogenies by use of a species guide tree. For some datasets this may improve phylogenetic accuracy but only when the evolution of a gene correlates with that of the overall evolution of its species. For prokaryotes, this is not always the case, given the high prevalence of HGT [232, 187, 233]. We therefore felt that for this particular dataset, with its high proportion of prokaryote members, that constraining the gene tree to the species tree topology was not appropriate.

FunTree phylogenies are not rooted [107, 108]. Including an outgroup as a root would enable the unfolding of evolutionary events throughout this superfamily to be seen more clearly. An outgroup should be clearly more evolutionarily distinct than those members of the ingroup, whilst remaining detectably homologous. We used the FunTree definition of SSGs within a superfamily to define an ingroup and outgroup (for more detail see sec. 5.2).

5.2. Methods

5.2.1. Selection of additional sequences

Using the FunTree multiple alignment (FunTree 3.60.15.10 SSG1) [107, 108] a profile hidden Markov model was created in HMMR [122] to search for additional sequences in the UniProtKB database using default parameters. The profile effectively represented SSG1 of CATH H-level superfamily 3.60.15.10.

In its phylogenies, FunTree uses a filtering strategy so that trees include groups of taxa that proportionally represent their occurrence in nature [108, 107, 90]. Whilst adding sequences, we were careful to maintain the approximate proportions of these functional groups. A diverse and significant group of sequences were picked by using keywords for different metallo-β-lactamase members (as listed by Bebrone [205]) and other functional groups: ‘flavoprotein’, ‘nitric oxide

reductase' (NOR) and 'Hydroxyglutathionehydrolase'/'glyoxalase II' which were ordered by score. We excluded draft sequences but not all sequences had been reviewed.

We extracted the 3.60.15.10 domain from the sequences by submitting to Gene3D [234, 235] and CATH (for those sequences that had solved structures) PDBsum [236] and Gene3D were also used to trim off signal peptide sequences.

Choice of outgroup sequences Composed of two structurally similar outgroups, the CATH superfamily 3.60.15.10 includes SSG1 - including the metallo-β- lactamases, and SSG 2 which includes the ribonucleases (tRNase Z). We used structurally solved members of SSG2, with their experimentally designated functional residues as our outgroup. These members of SSG2 satisfied the conditions for a good outgroup - showing homology but suitably distinct from all members of the ingroup (SSG1).

Alignment of additional sequences We used the FunTree structurally informed multiple sequence alignment (FunTree 3.60.15.10 SSG1) [108, 107]) of the superfamily as a basis for which to align additional sequences. Trimming of the alignment by BMGE [237] was used in preliminary analyses (data not shown) but the results of which were not used in further analyses. By visual inspection, we found that metal coordinating residues, thought to be conserved in the alignment, were well aligned.

The bias of long branch attraction [165, 238] can be reduced by adding additional sequences [168, 167, 231]. We aligned our additional sequences, found using the above search strategy, using he profile aligning facility in MAFFT [129, 131, 130], with the L-INS-I algorithm, JTT 100 matrix with gap opening penalty of 1.0 and an extension penalty of 0.0. The gap penalties were lowered as compared to the default to account for the high level of sequence and structural divergence expected given the wide range of functions within this family. The JTT 100 matrix achieved the lowest number of gapped sites and therefore best alignment of catalytic residues according to visual inspection.

Model testing and inference of phylogenetic trees Using MODELGENERATOR [239], specified with four gamma catagories, we found that the WAG-I-G [143] model was chosen as a best fit to our alignment by the BIC, AIC and AIC2 criteria.

PhyML 3.0 [155] was used to infer the Maximum Likelihood phylogeny. We allowed PhyML to optimise the I and G parameters and use a strategy of the best of 'Nearest Neighbour Interchange' and 'Subtree Pruning and Regrafting' rearrangements with 100 bootstrap replicates.

Since we had good structural evidence that SSG2 was more divergent than any members of the ingroup SSG1, we only included trees in which the ingroup was monophyletic for further analysis. We used the R [240] package Ape [241] and 'Root' function to manually test for the monophyly of the ingroup across the whole bootstrap set. Using this criterion, 98 of 100 bootstrap trees were used for further analysis [90, 157].

5.3. Results & Discussion

5.3.1. Alignment of sequences and structures

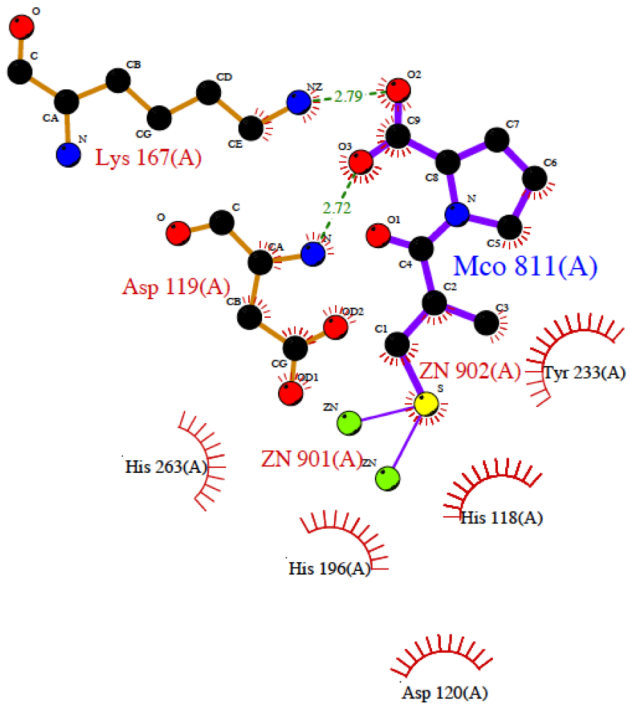


Figure 5.4.: *LIGPLOT [242] schematic depiction of the protein ligand interactions between 1M2X, BlaB metallo-beta-lactamase and* D*-captopril inhibitor (labelled above as Mco 811A) [207]. Hydrogen bonds are shown as green dashed lines. Curved red combs show other types of interactions, such as ligand and hydrophobic interactions.*

Our alignment for this superfamily, viewed with Taylor [243] colouring at 30% conservation shows that conservation of residues only lies within a specific region that surrounds Asp 170 (residue numbering from (Fig. 5.5)) and the motif surrounding it. This motif is not so well aligned for sequence Q3M8K9, where the position at 170 has been inferred as a gap '-' character and the nearest aspartate residue has been aligned with the histidine and glutamic acid residues of

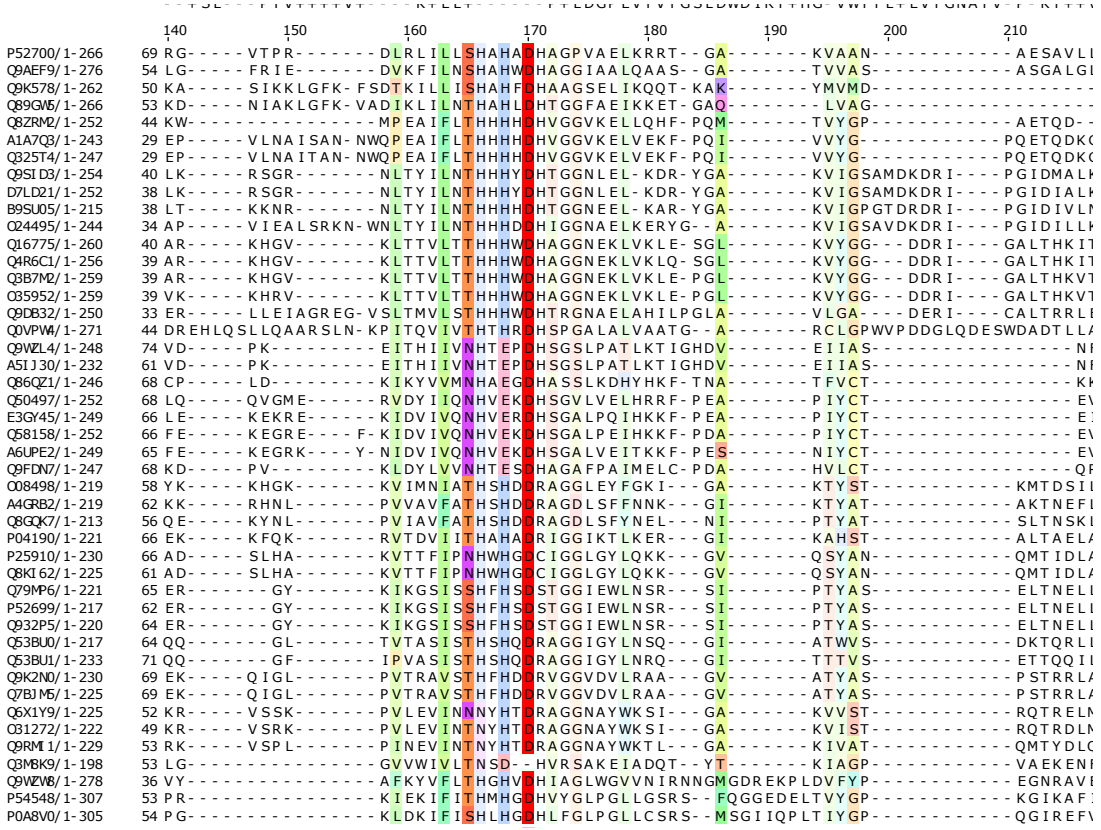


Figure 5.5.: *Section of input alignment [102] viewed with Taylor colouring [243] at 30% conservation in Jalview 2.8.2 [244].*

column 168. This may represent misalignment by the software or be possibly indicative of a mutation in this region for this particular sequence. This motif has clear roles in metal ion binding [245]. Interestingly, the motif changes consistently dependent on the function, with the motif H-x-H-x-D corresponding to metal ion binding in the lactamases, ribonucleases and glyoxalases, and H-x-E-x-D corresponding to metal binding in the flavoproteins. These observations are in line with [245]. Interestingly, the ribinucleases bind one or two zinc ions, the flavoproteins bind iron, and the glyoxalses are capable of binding iron, zinc and manganese [246, 247, 248, 249].

In fact, the glyoxalases have been so well studied in their ability to bind different metal ions as compared to their superfamily relatives that researchers have managed to change the specificity, and resulting function, of a glyoxalse II scaffold [250]. Different functions within this superfamily have been annotated in terms of their metal coordination in both sites, which differs dependent on function [245].

We expect that residues performing similar functions should cluster closely in 3D

space between different enzymes. We used the CSA's definition of a catalytic residue *via* the FunTree resource to superimpose structural members and compare clustering of active site residues for the ingroup. At the time of writing, no CSA annotations existed for members of the outgroup.

Within clade (interclade) distances show how well residues thought to be doing an equivalent function in the same enzyme class cluster together in three-dimensional space (Tab. 5.1). The between clade (intraclade) distances give an idea of how these clusters of functional residues relate in terms of distance between each other (Tab. 5.1).

In general, within clade distances of catalytic residues with the same identity in the same functional group tended to cluster closely.

Unsurprisingly, given its role as a hydrogen bond acceptor, electrostatic stabiliser, activator and in increasing nucleophilicity, Asp 120 (MACiE [98] 1SML numbering) is extremely well conserved both in sequence alignment and structural superposition. The conservation of this residue throughout the metallo-β-lactamase superfamily was discussed some time ago by Aravind [39]. Focusing on its role in the metallo-β-lactamases, the role of this aspartate residue in the binding of the second of the zinc ions [61] is key, as the positioning of this zinc ion has been shown to be involved in the optimal positioning of the substrate [251, 252, 224]. The only protein for which the Asp 120 is not aligned is an uncharacterised protein, discussed previously (UniProt acc: Q3M8K9).

Other residues labeled as catalytic by the CSA tend to cluster according to function/substrate type, as can be seen by the interclade distances of representatives. These other catalytic residues tended to share roles in transition state and substrate stabilization.

Strikingly, the common catalytic Cys residue found in the glyoxalases did not cluster so 'tightly'. More specifically, the position of this cysteine residue in 2QED differed substantially in comparison to the two other glyoxalase structures. There are a number of reasons why this may be the case, and it must be remembered that crystal structures only provide a static view of an enzyme in action [253]. It is possible that the 2QED structure represents a different stage in reaction and enzyme conformation than the other two glyoxalase structures. Although the difference could emanate from the fact that 2QED glyoxalase comes from a prokaryotic source, whereas 1QH5 & 2Q42 are from eukaryotic sources. It seems likely that these residues that cluster together with high proximity are performing equivalent functions with enzyme members of the same function.

We then took representative enzymes from each function and measured the distance between catalytic residues (other than the highly conserved catalytic Asp residue, which seemed invariant in position), substantial distances were recorded. As such, it seems that in this family, catalysis of different substrates has evolved by an active site that has evolved and specialised to bind different substrates using

(a)

Proximity of catalytic residues within Funtree clades		
PDB Codes	**Catalytic Residues**	**Distance in Angstroms**
B3 metallo lactamases		
1sml/1k07	D120/D120	2.91
1sml/1k07	Y191/Y228	2.52
Flavoproteins		
1ycg/2q9u	D85/D89	0.28
1ycg/2ohh	D85/D87	0.12
2ohh/2q9u	D87/D89	0.34
1ycg/2q9u	Y195/Y199	0.26
1ycg/2ohh	N198/N201	0.54
1ycg/2q9u	H25/H31	1.05
B1 metallo lactamases		
1znb/1m2x	D103/D120	0.95
1znb/1mqo	D103/D120	2.08
1znb/1dd6	D103/D81	0.68
1znb/1ko3	D103/D120	1.15
1znb/1mqo	N193/N233	0.86
1znb/1dd6	N193/N167	0.78
1znb/1ko3	N193/N233	0.18
1ko3/1m2x	S227/S228	4.12
Glyoxylases		
2qed/2q42	D57/D58	4.61
2qed/1qh5	D57/D58	5.50
2q42/1qh5	D58/D58	1.65
2qed/2q42	C134/C138	9.69
2qed/1qh5	C134/C141	9.93
2q42/1qh5	C138/C141	0.66

(b)

Proximity of catalytic residues in-between Funtree clades				
First Enzyme (PDB code)	**Second Enzyme(PDB code)**	**Second Enzyme Function**	**Catalytic Residues**	**Distance in Angstroms**
B3 metallo lactamase				
1sml	1znb	B1 metallo lactamase	Y191/N193	6.60
1sml	2qed	Glyoxalase	Y191/C134	12.90
1sml	1qh5	Glyoxalase	Y191/C141	11.70
1sml	1ycg	Flavoprotein	Y191/Y195	11.01
1sml	2ohh	Flavoprotein	Y191/N201	6.96
1sml	2q9u	Flavoprotein	Y191/H31	10.94
Flavoprotein rep (PDB code)				
1ycg	1sml	B3 metallo lactamase	H25/Y191	10.94
1ycg	1znb	B1 metallo lactamase	H25/N193	7.20
1ycg	2qed	Glyoxalase	H25/C134	17.40
1ycg	1qh5	Glyoxalase	H25/C141	16.70
B1 metallo lactamase				
1znb	2qed	Glyoxylase	N193/R136	5.38
1znb	2q42	Glyoxylase	N193/K140	9.28
1znb	2qed	Glyoxylase	N193/C134	10.03
1znb	1qh5	Glyoxylase	N193/C141	12.75

Table 5.1.: *Interclade and intraclade distances of FunTree annotated residues measured in Angstroms between amino acid alpha carbons. Residue numbers as documented in the FunTree resource.*

discrete and divergent locations of residues within its active site (Fig. 5.6), whilst those residues involved in metal ion coordination remain conserved.

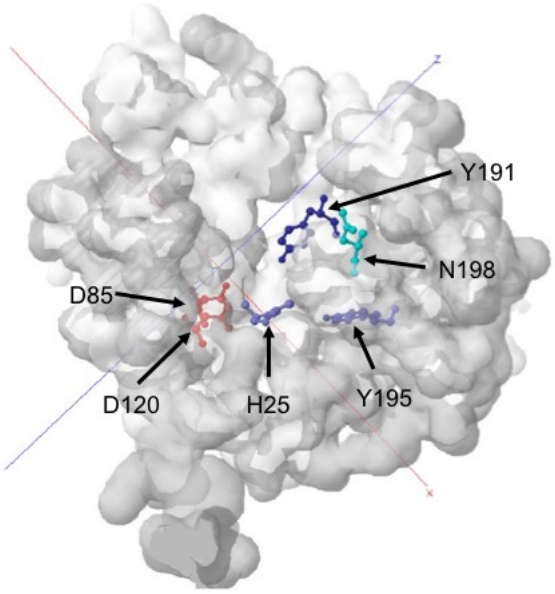
Aravind in 1999 described the structure of this active site as an exaptation [254] of an ancient structural scaffold that has allowed for specialisation of different substrates and therefore functions, whilst employing a similar chemical mechanism [39]. In the flavoproteins, the situation is a little more complicated, with a two domain fusion event having modified overall function. This superfamily exemplifies a common theme seen in enzyme evolution, that often a superfamily contains members that catalyse similar reaction mechanisms but on different substrates. Exaptation seems a plausible explanation to explain the pattern of binding and catalysis for a wide range of substrates in this superfamily. If so, this may help us to understand antibiotic resistance in this superfamily - since such an exaptation may lead to the same functionality evolving twice by independent means, as speculated by phylogenetic analysis by Aravind 1999 [39] and covered in the following chapter.

5.3.2. Comparison with Baier *et al.* alignment

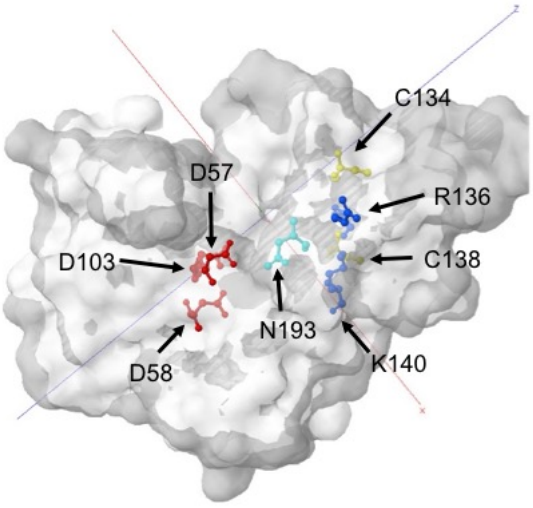
The idea that the structural scaffold of this superfamily has demonstrable capability of catalysing hydrolysis of a wide range of substrates could imply that specialisation to different substrates evolved via a process from promiscuous activity. This notion was explored by Baier and colleagues, who mapped the connections between diverse reactions in this superfamily and found them to be evolutionarily traversable [62].

The sequence set used in the study by Baier et al. included members with less than 5% sequence similarity, but homology was confirmed by means of a conserved motif (Fig. 5.7) and by the overall structural fold [62]. Since such divergent sequences were being used, sequence similarity networks were used to explore broad relationships between members as the authors deemed phylogenetic analysis and multiple alignments were not appropriate for such a low level of sequence homology [62].

The results of Baier et al's analysis can be viewed as complementary to our own. Our own alignment of the metallo-β-lactamase family with a selection of Baier's sequences [alignment using MAFFT with default parameters] shows conservation of the metal ion binding motif discussed earlier. Since Baier's sequence set contains more divergent members than our own, its perspective on evolution extends further, although the low level of sequence identity means the details are more broad brush. By taking this broader, less detailed view, Baier et al. hypothesise that ancestrally, this enzyme evolved to hydrolyse nucleotide derivative substrates, before diverging into catalysis of non-nucleotide substrates such as β-lactams [62]. This correlates with the FunTree divisions of SSG1 and SSG2 and our choice to use SSG2 as the outgroup.



(a)



(b)

Figure 5.6.: *a) 1SML (B3 metallo-lactamase) and 1YCG (flavoprotein) superimposed in Jmol [208] as provided by the FunTree resource [107, 108]. Catalytic residues (1SML): Tyr191, Asp120, catalytic residues (1YCG): Asp85, Asn198, His25, Tyr195. b) 2Q42 (glyoxalase), 2QED (glyoxalase) and 1ZNB (B1 metallo-lactamase) superimposed in Jmol [208]. Catalytic residues (2Q42): Asp58, Cys138, Lys140. Catalytic residues (2QED): Asp57, Cys134, Arg136. Catalytic residues (1ZNB): Asp103, Asn193.*

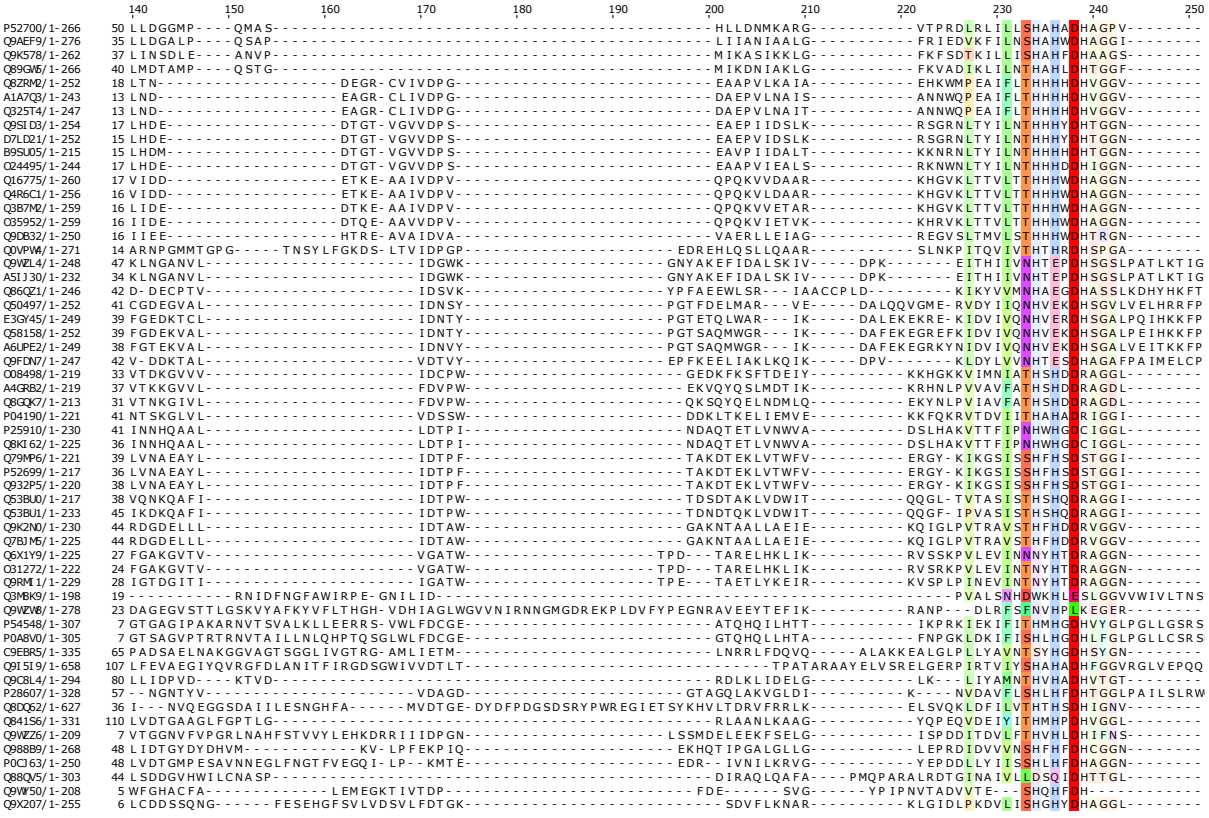


Figure 5.7.: *Input alignment [102] with additional sequences from Baier and Tokuriki [62], realigned with default parameters in MAFFT [128, 132] viewed with Taylor coloring [243] at 30% conservation in Jalview 2.8.2 [244]. The conserved sequence motif (positions 230-242 above) are similar to our input alignment* (Fig. 5.5)*.*

Baier and colleagues found that some of the promiscuous activities between different members of this superfamily corresponded with evolutionary divergences and some did not [62]. For example, enzymes not closely related to the metallo-β-lactamases were able to catalyse a metallo-β-lactamase reaction [62]. It was noted that metallo-β-lactamase functionality can happen within active sites that vary substantially in volume and hydrophobicity, adding credence to the idea that this fold is particularly amenable to lactamase substrates [62]. It was found that with enzymes with sub-optimal promiscuous activities the substrate was not well aligned for attack by the activated water molecule [62] corroborating the role of the highly conserved catalytic aspartate residue within this site [251].

Many of the differences between the dataset used by [62] and by ourselves are due to Baier et al.'s inclusion of more divergent sequences that are beyond the homology cutoff by FunTree. However, according to Baier et al.'s Fig 3., there are enzyme functions within our scope (i.e. between the B3, B1 and ribonucleases) that have not been included in our dataset. The difference emanates from the source of our sequences - Baier et al. used Pfam followed by a BLAST search

[62]. We used the FunTree alignment to search for additional sequences using HMMR. Most importantly, we used the FunTree SSGs as a template for which functions to include - aiming only to add sequences to already included functions. Additional functions, such as lactonases & phosphodiesterases can be found in the FunTree MDA 2 (Multiple Domain Architecture) for this family. However, in this study, we focused on SSG alignments since their homology is easier to determine. Later in this thesis we experiment with building an alignment and phylogeny for a superfamily in which multiple domains are included in the alignment.

5.3.3. Phylogenetic tree

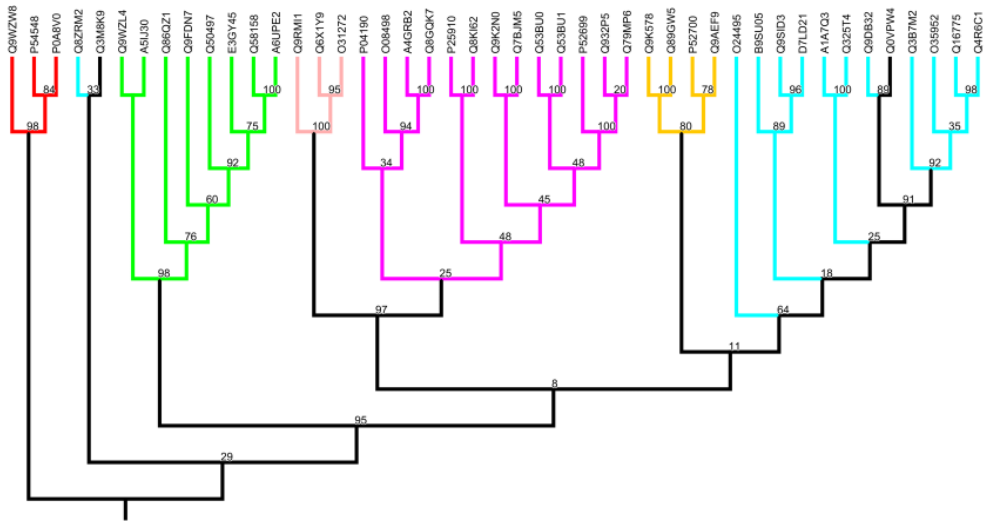


Figure 5.8.: *ML phylogenetic tree of the metallo-β-lactamase superfamily labelled with percentage bootstrap support values. Taxa are labelled by UniProtKB accession numbers. groups are colour coded by function:* red *ribonucleases,* cyan *glyoxalase IIs,* green *A-type flavoproteins,* pale pink *subclass B2 metallo-β-lactamases,* magenta *B1 metallo-β-lactamases,* orange *B3 metallo-β-lactamases,* black *function not assigned. The phylogeny was edited in Mesquite [255]. Figure taken from Alderson et al. [102].*

The lowest support for divergences in our phylogenetic tree (Fig. 5.8) is at the deeper splits, between the glyoxalase clade and the B3 lactamases, and between this glyoxalase /B3 clade and the B1/B2 clade. This correlates with the findings of Garau et al., in which the more ancient divergences were harder to resolve, and with bootstrap support values seen for the FunTree tree [219, 108, 107]

Although not strongly supported, the hypothetical protein (UniProt acc: Q3M8K9) forms a monophyletic group with (UniProt: Q8ZRM2) from *Salmonella typh-*

imurium that is a sister group to the remainder of the ingroup. Other than this, the main functional groups fall in well supported monophyletic groups (<50% cut off). We assume that our ingroup, being structurally distinct, should be monophyletic - phylogenetic support for this is strong, at 98%. Our phylogenetic tree demonstrates the difficulties in resolving the exact evolutionary order of functional subgroups within this superfamily. However, subclades of members sharing the same enzyme function are well supported - giving credence to our structurally informed alignment method and careful choice of members.

5.4. Conclusion

In studying divergent enzyme superfamilies, a structural approach is necessary. An approach such as the one outlined in this chapter ensures that related sequences from a wide evolutionary scope are selected, which may not be obvious from looking at sequence features alone. As such, the CATH and FunTree resources can be used as a starting point for exploration of enzyme superfamilies, at the single and multi domain level. We widened the taxonomic representation of the dataset to gain a more detailed picture of the evolution of a functions within a particular superfamily, using the FunTree seed alignment as a structural profile. To optimise the ML tree building strategy, we aimed to make sure that the model of evolution matched the data at hand [239]. The use of a species tree to guide gene tree topology, as is used in the FunTree pipeline, can be helpful, but for the prokaryotes, where HGT is common, constraining the gene tree by species tree would not be a judicious choice. By taking these measures, we have developed and improved the phylogeny for this superfamily, giving a more detailed picture of evolutionary relationships in this whole superfamily than others had previously published [219, 107, 108].

Using structure to inform sequence alignment is essential for building high quality trees of such divergent sequences. However, a balance must be struck. For example, an approach such as that found in Garau et al. [219], where only structurally aligned residues to specified members of the dataset were used, leads to a short and biased alignment, and therefore a questionable tree. The quest to find the true signal in noisy datasets such as these is not an easy one.

Our work, and the work of others, have shown that aligned members of this superfamily share a common scaffold in which metal coordination in order to activate water is conserved. The wide range of functions in this superfamily, particularly in the ability to hydrolyse a broad range of lactam substrates, is conferred by an active site that may be exapted to bind a range of substrates. Understanding this trend is key in understanding why this enzyme is so readily adaptable to overcoming diverse antibiotic challenges.

6. One origin for metallo-β-lactamase activity, or two?

The diversity seen in this superfamily includes three classes of metallo-β-lactamases - the B1,B2 and B3 subclasses. These three groups all hydrolyse lactam substrates by means of nucleophilic attack by a zinc activated water. The B1 and B2 classes share more sequence similarity to each other than to the B3 class and therefore are often grouped together [222]. In fact, analyses based on sequence, structure and phylogenetic reconstruction postulate that the B1/B2 and B3 subgroups arose by independent evolutionary events [202, 39, 222]. Hall et al. date emergence of B1/B2 lactamase functionality at one billion years ago and the B3 class as arising 2 billion years ago [222, 202]. The notion that these groups are products of independent evolutionary events tallies with their low sequence similarity. However, at the structural level, lactamases from these two groups share some structural similarities, although both groups also display unique functional features [223]. This evidence, therefore, does not constitute a clear basis on which to discriminate between scenarios of single or multiple origins of lactamase functionality.

In surveying the literature, the evolution of the same function *via* independent events within the same superfamily seems rare [256, 257]. However, the convergent evolution of the same function by independent means between different superfamilies seems more common, for example - aldehyde reductase and glycolate oxidase which have representatives from more than one enzyme superfamily [258]. In converging to evolve a similar function, these enzymes can employ different mechanisms of catalysis or more rarely, employ the same chemical mechanism, using different catalytic machineries [258]. The B1/B2 and B3 subgroups may constitute an example of this rarer case, in which the mechanism of nucleophilic attack of the lactam ring by zinc activated water is the same, but the transition-state stabilising residue differs (Fig. 6.1).

The observation that convergent evolution of the same function is rarer within superfamilies than between them seems intuitive, since the evolutionary landscape explored by the evolution within one superfamily is necessarily smaller than that explored by multiple superfamilies. However, given sufficiently strong selective pressure, the occurrence of evolutionary events leading to the same

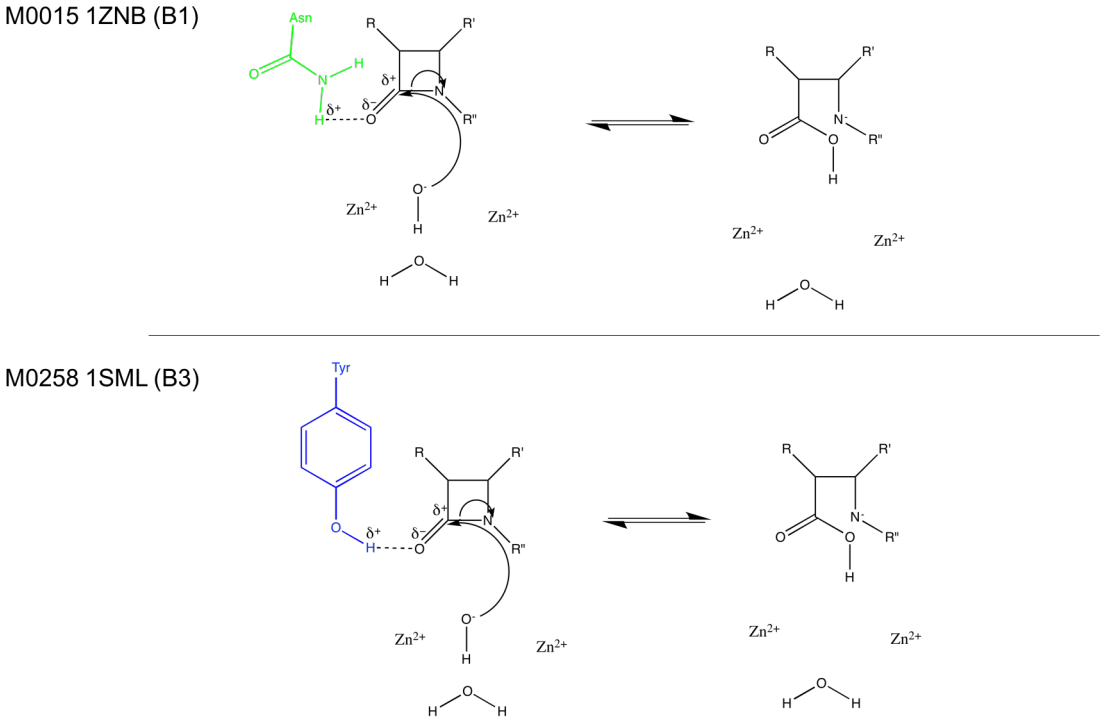


Figure 6.1.: *The chemical mechanism for B1 and B3 lactamase catalysis as displayed in MACiE entries M0015 and M0258 [259, 98]. Figure adapted from MACiE and created using ChemDraw [260]. The transition state stabilising residue, Asn for the B1 lactamase and Tyr for the B3 lactamase donates a hydrogen bond to the carbonyl oxygen, electrostatically stabilising the carbonyl group thus encouraging nucleophilic attack by the hydroxide ion.*

function becomes more likely, as exemplified in enzymes with a role in host-pathogen relationships - such as in iron-transporter ferric ion-binding protein found in *Haemophilus influenzae* [256], in plant resistance genes [261] and in the phosphatidylinositol-phosphodiesterase superfamily, which includes a member that catalyses the production of sicariid spider venom [108].

In addition to the influence of selective pressures, certain protein folds can be more amenable to the binding of certain substrates than others. In this family, Aravind noted that this fold is particularly amenable to substrates with similar chemistry [39]. This may be indicative of a fold exapted (preadapted) [254] to bind lactam substrates and therefore has important implications in understanding the evolution of antibiotic resistance within this superfamily. For example, understanding the past phylogenetic history of lactam hydrolysing enzymes can be used to predict more likely trajectories of future evolution [215, 216, 217, 218]. If indeed the same lactamase function has evolved twice in this family by independent events, then recognising the possibility that this fold is exapted to do so has implications for the future of antimicrobial drug design and clinical practice.

6.1. Our contribution to the field

Although there are indications, based on sequence and structure comparisons, that the evolution of lactam hydrolysis has occurred twice on two separate occasions, no study has been designed to directly assess this possibility. In our work, we look to assess the function of the common ancestor of the B1/B2 and B3 subgroups. We assume that if the evolution of these two groups occurred by independent means then the common ancestor should be devoid of lactamase activity. Conversely, if it is found that the ancestor is likely to have possessed lactamase activity, we interpret this as evidence that the B1/B2 and B3 functions evolved in a divergent process from ancestral lactamase activity.

In order to assess the function of the MRCA of the metallo-β-lactamase subgroups, we need a robust phylogeny on which to base our predictions. The challenges in reconstructing an unambiguous phylogeny for this superfamily are discussed in the previous chapter. We use a Maximum Likelihood strategy to infer the phylogeny and base our analysis on a set of topologies from the bootstrapped alignments, broadly similar to the approach used by Latysheva et al. [262].

Discerning function *in silico* is not easily accomplished by analysing sequence features alone. Nor is function necessarily well represented by the overall structure of a protein. In this study, we assess function based on the use of 3D catalytic templates - in which we can compare the positioning of catalytic residues in 3D space between our homology models and extant lactamase enzymes. This approach has been used by others, such as Meng et al. and Torrance et al. [105, 106].

6.2. The challenge of low bootstrap support

As discussed in the previous chapter, unambiguously reconstructing a phylogeny for this superfamily is a challenge [202, 219]. The low bootstrap support, especially between clades of different functions makes ascertaining the exact evolutionary ordering of events difficult.

However, since our question is broad based - in which we wish to know the general ordering of the evolution of functions in this family, we can use the bootstrap sample as a set of possible evolutionary scenarios. It seems likely that, although bootstrap support for some nodes is low, there are broader commonalities across the bootstrap sample which are not necessarily reflected in the bootstrap scores on the ML tree.

By predicting MRCA sequences for each topology in the bootstrap set, we can identify commonalities at the sequence level and, through homology modelling, at the structural level too. An alternative method, such as Bayesian Markov Chain Monte Carlo (MCMC) used by authors such as Lutzoni et al. [263] could have

been utilised but was deemed to be not as appropriate for our particular dataset. Using a Bayesian strategy is beneficial in terms of its statistical interpretability, since each predicted MRCA sequence can be assigned a posterior probability of confidence, although, this is conditional on the subjective priors applied and the underlying alignment and tree topology. One can attempt to create a more objective Bayesian strategy by the use of uniform priors, but this is not reasonable for all variables - such as branch length. In addition, even if we could reasonably estimate a prior for one variable, such as topology, this can have unintended effects on the priors for other variables, such as clade size [158, 264, 157]. Although a bootstrap value is not a direct indicator of confidence in a result, it does give an indication of the robustness of the topology given the underlying alignment [265]. There exists inherently more variability within a bootstrap sample than within a Bayesian sample, and so the bootstrap is less in danger of reporting inflated statistical support due to an underlying sample with little variability. Given the challenges in unambiguously reconstructing the evolution of this family, by ourselves and by others, constraining our sample by subjective priors would not be as effective in exploring the different past evolutionary trajectories of this family. If the evolutionary signal in this family were strong, then effects of the prior may be overcome, but this is not the case for this dataset [158].

We can think of these MRCA predictions, each based on one topology from the bootstrap set, as a sample from evolutionary probability space. Obviously, our sample is not exhaustive, and there will be other possible evolutionary scenarios not included. However, our careful use of sequence selection, structure based alignment and ML building strategy are constructive measures to ensure our sample is likely to reflect the evolutionary signal in the data.

6.3. Methods

We only included trees in which the ingroup was monophyletic for further analysis - resulting in 98 different tree topologies.

Prediction of ancestral sequences using GASP

We set branch lengths of trees to a minimum of 0.0001 and submitted these along with the original, un-bootstrapped alignment to GASP [174], using default settings and specifying a WAG substitution matrix [143] and outgroup sequences.

Selection of MRCA node

The Ape package [241] in R [240] was used to view output GASP trees. Using R, the relevant node number for each MRCA for each tree was used to extract

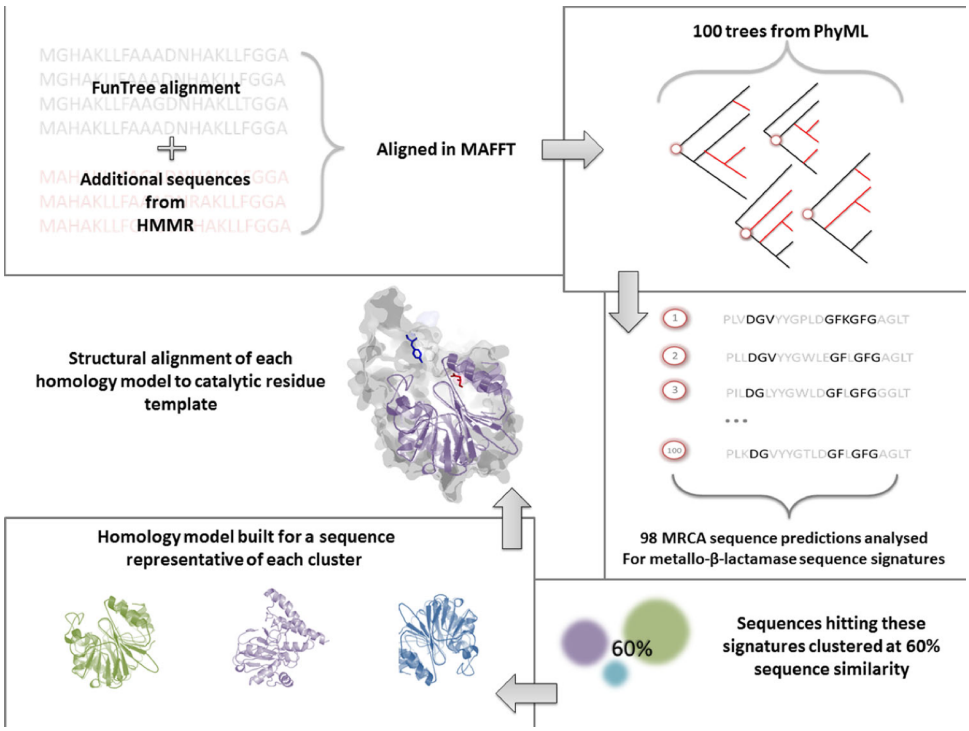


Figure 6.2.: *A schematic overview of the study. Alignment and phylogeny construction as described in Chapter 5 are shown, before the process of predicting the MRCA sequence for each tree in the bootstrap set, clustering, homology modelling and alignment to catalytic templates as described in this chapter. Figure taken from [102].*

the relevant GASP predicted sequence for each tree using the 'SeqinR' package [266].

Submission of sequences to InterProScan

We submitted the resulting 98 GASP MRCA sequence predictions to InterProScan [267]. 44 of these sequences were hits for metallo-β-lactamase signatures [116].

Clustering in CD-HIT

The 44 MRCA sequences that had positive hits for the metallo-β-lactamase signature were clustered at 60% in CD-HIT [268]. This step was to filter the dataset to a smaller number of representatives that represented the variability of the sequences. The resulting 11 representatives from the 11 clusters were then tractable for homology modelling and alignment to catalytic templates.

Homology modelling

We used PHYRE2 to model the 11 representative MRCA sequences [182, 269]. The coordinates for the highest scoring model for each sequence submission were used for the next stage.

Construction of catalytic templates

We found that high quality, publicly available templates, such as the ProFunc server [270], did not discriminate between B1/B2 and B3 structures. We therefore created our own templates, by using PDB structures of extant metallo-β-lactamase enzymes and their respective catalytic residues as annotated in MACiE [259, 98] or the CSA [271].

Alignment of homology models to catalytic templates

We used a structure-based strategy, rather than a sequence based strategy, to align our homology models to the catalytic templates. CEAlign [272] as utilised by ourselves in PyMOL Version 1.6.0.0 [273] is a structural based alignment strategy which we deemed more appropriate given the low level of sequence similarity between members of this superfamily.

We then used two criteria to filter our 11 MRCA homology models. Both these filters were applied to distinguish the models that possessed the minimal machinery for lactamase activity according to our catalytic templates. The first of these was to assess if our MRCA models had an equivalent residue to that defined by our catalytic template within a five Angstrom radius. The MRCA models that passed this test were then assessed to see if the distance between the predicted 'catalytic' residues was within $\pm$ two Angstroms of the distance between catalytic residues in the template. These distance filters of five and two Angstroms are purposefully generous, allowing for imprecision in the homology models, particularly in loop regions where it is known that some of these catalytic residues reside.

6.4. Results & Discussion

Ancestral sequences results from clustering analysis

We chose to use 'GASP' (Gapped Ancestral Sequence Prediction for proteins) [174] for prediction of MRCA sequences. GASP is not able to differentiate between multiple, near optimal predictions, as is implemented by programs such as FASTML [173]; despite this, its unique handling of gaps is more biologically

realistic. From a practical point of view, GASP is available as source code allowing for the prediction of sequences from our 98 trees in a tractable manner. The FASTML server is more sophisticated in its prediction of sequences and of gaps, but the use of this server was impractical for the size of the dataset. No source code was available with this indel prediction functionality included and we found the use of FASTML source code produced sequences that were too long - as a result of gaps being inappropriately treated.

Using the alignment of MRCA cluster representatives (Fig. 6.3) we noticed that a metal binding motif, as described by Gomes *et al.,* is well conserved as discussed in the previous chapter [245]. We can compare the first part of Gomes et al.'s metal binding motif (since the second part of this motif is not as well conserved and more scattered) to look at identity. Most sequences have 'H-X-D-H' which corresponds to *S.maltophilia* or with the glyoxalase enzyme signature as described by Gomes et al. [245]. One sequence, (46) differs from this 'H-X-D-R' which corresponds to *B.cereus* [245]. Based on this comparison, it seems that most MRCA sequences correspond most closely with the B3 lactamases, or to the glyoxalase. It is striking that the metal binding motif of the these predicted MRCA sequences does not tend to correspond with the B1 lactamases. If the MRCA of the metallo-β-lactamases did have lactamase activity, it is more likely to have had B3 characteristics, with the B2/B1 groups being derived from these. This corresponds with the prediction that the B3 lactamases are older than the B1 class. If the MRCA of the lactamases did not possess lactamase activity then it is likely to have been similar to a glyoxalase according to this comparison. We notice that regions 25-40, 120-135, and 185- 200 of our alignment appear to be most variable - with a large number of gaps. These regions may well correspond to features, such as loops, which are less constrained in evolution. When comparing our alignment to the annotated structure 1SML, we find beta hairpin secondary structures in these approximate regions.

It could be argued that the noted variability across our MRCA sequence set demonstrates our success in the generation of a wide sample of tree topology space. Yet, it appears our approach has been sufficient to highlight common motifs that would be harder to identify solely from the comparison of tree topologies. A comparative analysis of MRCA sequences across our bootstrap sample has highlighted convincing functional properties of the MRCA, despite ambiguities in the ML tree for this superfamily.

Homology modelling of structural representatives and alignment with catalytic templates

After the two filtering criteria, the only representative (sequence 51) that passed both our criteria for having metallo-β-lactamase activity had the closest structure to 1SML - a B3 lactamase (Fig. 6.4). This representative was from a cluster of five

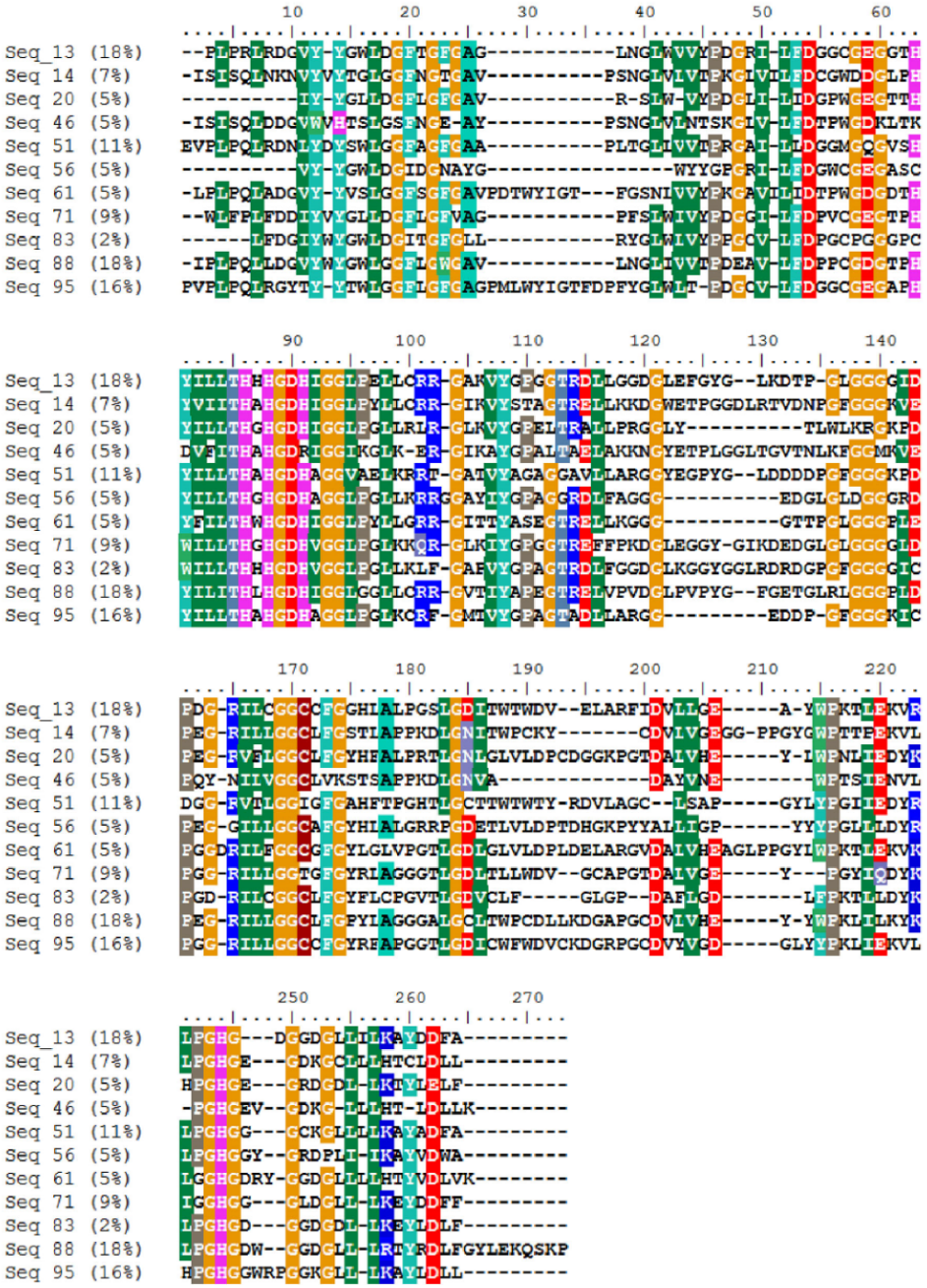


Figure 6.3.: *Sequence alignment of the 11 cluster representatives, with the percentage weight of the cluster in the dataset of 44 sequences possessing IPR001018 signatures. The sequences were aligned with default settings in MAFFT and columns are coloured according to a 70% similarity threshold in BioEdit [274]. Figure taken from [102].*

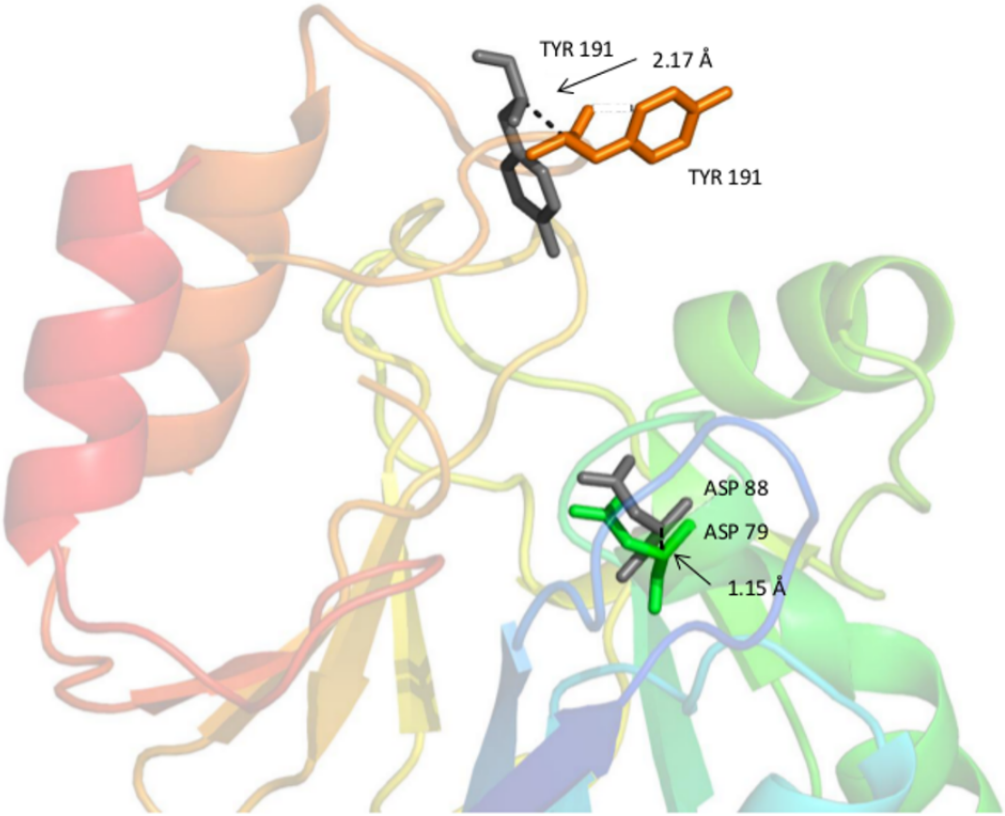


Figure 6.4.: *PHYRE2 homology model of sequence 51 aligned with 1SML with 3.5 angstroms RMSD, which passed both our criteria for being most like a B3 metallo-β-lactamase. 1SML catalytic residues Asp and Tyr are shown in grey, their distance in angstroms to the homology model's predicted catalytic residues, Asp in green and Tyr in orange, are shown. Image was generated using Pymol [273]. Figure taken from [102].*

MRCA sequences, and so represents 5/98 MRCA sequence candidates. A close runner up was the homology model from MRCA sequence 46, which was most similar to our B1 lactamase template (PDB 1M2X). Interestingly, this was the only sequence in our MRCA set that had the metal binding signature of a B1 lactamase (see above). This result correlates with the sequence based observations above. The variation seen in the sequence set was also reflected in the results of homology modelling for each MRCA sequence, with top used templates including more similar A-type flavoproteins and even an alkylsulfatase, which was deemed a homologous member of this superfamily according to Baier and Tokuriki [62].

On the face of it, and statistically, the fact that only five of the 98 MRCA sequence predictions passed our criteria for lactamase activity adds evidence against a lactamase hydrolysing MRCA and the hypothesis that lactamase activity has evolved twice, on two separate occasions in the B1/B2 and the B3 lineages. However, as Weinreich et al. demonstrated, some trajectories are more probable than others in evolution of fitter proteins [275]. It may not, therefore, be biologically ap-

propriate to assess our result on the assumption that all of our phylogenetic trees are equally probable in the evolutionary history of this family.

Even when looking at the MRCA sequence which passed both our criteria for lactamase activity, we should be aware that our criteria only prescribe the machinery (i.e. catalytic residues) that is necessary for catalysis. The presence of these residues does not imply sufficiency - this would need to be assayed *in vivo.* Moreover, our criteria for lactamase activity are based on the alignment with static crystal structures. Although useful for structural characterisation of larger proteins, crystal structures do not reflect protein dynamics, which play a key role in evolution [80, 276]. The crystal structures on which we base our analysis are of extant enzymes, which we know are successful in catalysing lactamase substrate. However our definition does not include all possible ways this function may have been accomplished in the past, for example, by early, promiscuous activities of other enzymes.

In general, phylogeny only has the power to model the process of evolution at a residue-by-residue granularity [173, 175, 277]. Even those that do take correlated evolution of sites into account [278, 279, 280, 281] do not evaluate each mutational step by its fitness, given the existing selection pressures and the size of population.

6.5. Conclusion

Our aligned MRCA cluster representatives, which were positive hits for InterPro metallo-$\beta-$lactamase signatures, showed high variability in regions - possibly loops involved in binding substrate and product. Despite this variability, all cluster representatives had the easily identifiable portion of the conserved metal binding motif highlighted in Gomes et al. [245]. Interestingly, the majority of these motifs indicated most similarity to a B3 lactamase or a glyoxalase. This observation was consistent at the structural level, in which we found the only model to pass both our criteria for being a metallo-β-lactamase most closely resembled our B3 lactamase catalytic template.

The finding that 54 of our MRCA sequence predictions did not match the metallo-β-lactamase InterPro signature has some bearing on the interpretation of our results. Are these sequences simply ‘scrambled’ predictions - for which all possible enzyme resemblance had been lost as a result of methodological inaccuracy? If so, then the proportion of ‘successful’ lactamase MRCA predictions amongst only those that matched the metallo-β-lactamase Interpro signature scales to 5/44 - over 10% of MRCA candidates. Looking at it this way, one might ascertain that 10% of the bootstrap sample actually supports the scenario that the B1/B2 and B3 lactamases arose as a result of a single, divergent event in evolution. Or, conversely, could it be that these 54 sequences may be the result of accurate

phylogenetic reconstruction and ASR, and are simply proteins for which the function is unknown? This would not be unusual, since the literature is littered with numerous examples of discovered proteins with no functional assignment or sequence signature [282, 283, 284]. In this case, possibilities open for an MRCA of unknown function, possibly exapted to catalyse lactam substrates. Despite the constraints on our study, including sample size, the inability to model all evolutionary parameters and catalytic templates based on static 'snapshots', phylogenetic methods have been shown to adequately model the process of protein evolution, at least from the viewpoint of structural viability [285].

What does our study mean for future efforts to combat antibiotic resistance? Firstly, if we assume that the methods we have employed are capable of determining the general properties of MRCA sequences in this family then statistically speaking, there appears to be few paths which evolution could have taken to yield a lactamase hydrolysing ancestor. This adds to the evidence for an exapted fold particularly amenable to evolving machinery capable of hydrolysing metallo-β-lactamase substrates.

If our 5% of MRCA sequences that passed both of our criteria for lactamase activity actually represents a more probable evolutionary trajectory, we can glean that this protein is likely to have resembled a B3 lactamase in terms of metal ion coordination and in terms of catalytic amino acids. One common theme is that whether able to hydrolyse a lactam molecule or not, the ancestor was likely to have been metal coordinating (according to our alignment of cluster representatives). The evolutionary flexibility in this family appears to lie in regions other than this metal coordinating motif, presumably in more flexible loop regions. It is therefore likely that whether via multiple independent evolutionary innovations, or by rapid divergence from a lactam hydrolysing ancestor, these enzymes have the flexibility to adapt their exapted structure to face new, but related structures of drug molecules.

It seems that targeting metal coordination in these enzymes may be an effective way to hinder the evolution of resistant variants, and has been exploited in the development of antimicrobial compounds which chelate the bound zinc ions, including biphenyl tetrazoles, mercaptocarboxylate, *D*-captopril and thiomandelic acid [206]. However, an efficient inhibitor for all subgroups of metallo-β-lactmases has not been discovered [206]. This is in part due to the differences in structure between the B1/B2 and B3 lactamases, and the fact that no covalent intermediate is involved to specifically target. Apart from biological uptake and *in vivo* efficiency, this superfamily includes important enzymes in human metabolism such as glyoxalase II. This makes the design of a broad ranging zinc chelator without dangerous off-target effects a challenge for drug design.

Part III.

Investigating the evolution of a Domain of Unkown Function - The DUF-62 gene

7. Tracking the evolution of the DUF-62 gene

The DUF-62 gene has evolved a diverse array of functions, including chlorinase, fluorinase and *S*-adenosyl methionine (SAM) hydrolase activity. The chlorinase, fluorinase and SAM hydrolase all use SAM as a substrate, initiating nucleophilic attack *via* halide ions or water [286].

Known structures and functions within this family tally with the general trend of a conserved scaffold being able to accommodate a range of substrates - thereby allowing for diversity of function with enzyme superfamilies [83, 81, 102, 39]. The overall structure of enzymes within this family is well conserved (Fig. 7.1) although there are key differences within the active site [286], for example, the fluorinase has an 'insert loop' that can be seen in aligned sequences and structures (Fig. 7.1) [286].

Figure 7.1.: *2WR8, DUf-62 enzyme from* Pyrococcus horikoshii, *2Q6I, chlorinase from* Salinispora tropica *and 1RQP, fluorinase from* Streptomyces cattleya *chain A aligned with CEAlign [272] in PyMOL [273], the fluorinase insert loop is encapsulated by a dashed grey circle in the above diagram.*

The chlorinase and fluorinase enzymes are exciting avenues of study for medicine with applications to oncological chemotherapy. The chlorinase catalyses the production of a metabolite involved in production of salinosporamide A (proteasome inhibitor and possible anti cancer agent) [287]. Halogenated metabolites

can provide important starting points for the generation of new antibiotics [288] since halogenation of a drug can improve its bioavailability, stability and activity [289]. In the last thirty years, 70 % of antimicrobials and 60% of anticancer drugs that have entered clinical trials have been based on natural products [290, 291].

From an evolutionary perspective, the generation of a wide range of natural products by an organism is thought to confer a selective advantage - by allowing a range of 'tools' that can be used in different and changing environments [292]. The mechanisms of generating these diverse metabolites can be attributed to metabolic enzyme promiscuity or incomplete processing of substrate leading to the release of a different product [292]. It seems likely that both the fluorinase and chlorinase produce metabolites that are involved in the generation of products useful to the organism, as is the case for other prokaryotes [293].

Nucleophilic halogenation is relatively rare, with the only other examples being halomethane production [288]. However, enzymes that catalyse halomethane production substitute the halide at the S-methionine position rather than the C5' position of adenosyl methionine as seen in this family [288]. Despite this difference, it has been reported that the halomethane synthase from *Arabidopsis thaliana* [294] has a similar structure to members of this family [288]. We find that these enzymes are unlikely to be homologous - given their different CATH [89] codes (2WR8 - 3.40.50.10790 & 2.40.30.90, 3LCC - 3.40.50.150). A structural alignment of PDB (Protein Data Bank) structures 3LCC and 2WR8 using CEAlign [272] in PyMol [273] revealed that although structural similarity is present for parts of the sequence, this is not conserved over the whole domain (Fig. 7.2). In addition, the *S*-adenosylhomocysteine (SAH) substrate binds in a very different location between the two enzymes - these differences are reflected in an RMSD (Root Mean Square Deviation) of 6.30 angstroms over 88 residues (Fig. 7.2).

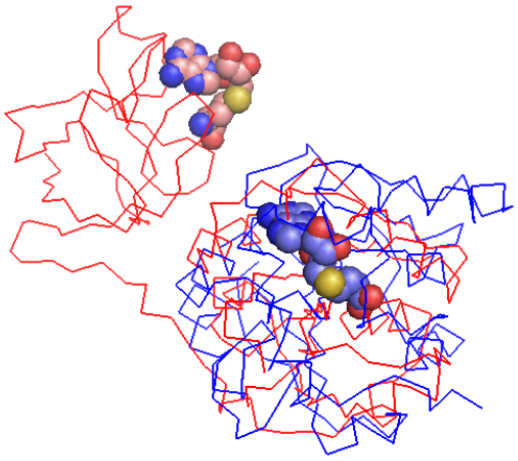


Figure 7.2.: *2WR8 chain A in red with bound SAH molecule in pink (inhibitor), 3LCC in blue with associated SAH molecule in lavender (inhibitor).*

On the face of it, the cellular role of the SAM hydrolase seems to be counter-intuitive, unlike the halogenases, which produce an 'expensive' product. By modifying SAM the SAM hydrolase appears to break down a 'high-currency' molecule in the cell to relatively 'cheap' constituent parts [286]. Available experimental evidence indicates that the SAM hydrolysis reaction is not reversible [295]. In the literature, two hypotheses have been posed as to the role of this enzyme. One is that the SAM hydrolase acts as a regulator of SAM levels by breaking SAM down when its concentration gets too high [295]. Another, more striking hypothesis is that since every catalytic cycle of the SAM hydrolase generates one proton this enzyme could have a role in maintaining cellular pH, particularly since around 20% of DUF-62 members are found in the archaea, many of which are extremophiles [286].

Unlike the metallo-β-lactamase family, where we had a wealth of knowledge of annotated structures, functions and cellular roles, this family provides an example of exploring function within a family that has much less annotation. We broke our study into smaller questions and hypotheses, whilst keeping the broad question of the role of SAM hydrolase in our mind when interpreting results.

Firstly, we do not know that all DUF-62 genes function as SAM hydrolases (or as chlorinases or fluorinases for that matter). Secondly, even if all DUF-62s do function as SAM hydrolases we do not know if their cellular role is the same.

As a first step, we aimed to gain a better understanding as to how well distributed the DUF-62 gene is across all organisms. Previous attempts, such as those by Eustáquio et al. and by Deng and O'Hagan [295, 286] have been by BLAST search - which, although sensitive, is not necessarily that accurate for identifying distantly related homologs. Instead, we make use of the fact that homologous proteins tend to share similar motifs rather than overall sequence similarity. We use the Pfam family [117] to look at the distribution of these genes throughout the kingdom of life, with its sensitive HMM search method.

Using the Pfam family as our core dataset, we build an ML phylogenetic tree to map the history of evolution in this family. We then go on to infer the root position and ancestral habitat of this gene as well as possible transfers assuming a parsimonious model of evolution. From this, we infer major transfer events in this family and determine well supported clades for further analysis.

7.1. Available data in the literature

7.1.1. Distribution of the DUF-62 family members across the tree of life

The DUF-62 gene is well represented across the prokaryotic archaeal and bacterial kingdoms of life but less so than in eukaryotes with only three representatives - two species of parasitic protozoan - *Entamoeba histolytica* and *Entamoeba dispar* as well as one from the castor oil plant - *Ricinus communis.* Within the prokaryotes, according to the Pfam, species presence of the gene seems well conserved. In the main, the gene occurs as a single copy, although some species have more than one, with *Gloeobacter violaceus* (strain PCC 7421)*, Frankia* sp. EUN1f and *Spirochaeta smaragdinae* having three copies of the DUF-62 sequence per species (source: Pfam). The highest copy numbers of this gene appear in the bacteria rather than the archaea.

In the bacteria, the gene is found in extremophiles, pathogenic species and in species that are neither extremophiles or pathogens. Chlorinase, fluorinase and SAM hydrolase functions have been annotated in the bacteria. In the archaea the gene is found in halophiles, some of which are also alkaliphiles, thermophiles, acidophiles and methanogens.

7.1.2. Alignment

The Clustal [124] progressive alignment strategy employed by such as those by Eustáquio et al.[295] is quick and efficient but does not have the ability to incorporate additional evolutionary information, such as from known structures, to inform the alignment process. We found that many more residues are highly conserved in this family than highlighted in such as those by Eustáquio et al. [295], in which 11 homologs were found by BLAST search. These additional residues are highlighted in (Fig. 7.3). Moreover, those positions highlighted by Eustáquio et al., thought to be involved in catalysis, are more variable in residue type than is portrayed by these authors. The variation seen at these alignment positions may be indicative of other functions within this family and is further supported by a range of functional annotations with IPR002747 sequence motif Fig. 7.4.

7.1.3. Phylogeny

Eustaquio et al.[295] used this alignment to build a neighbour joining tree using ClustalX [297]. Their phylogeny, although limited in taxa and choice of tree building method, shows well supported branchings of the members between the ar-

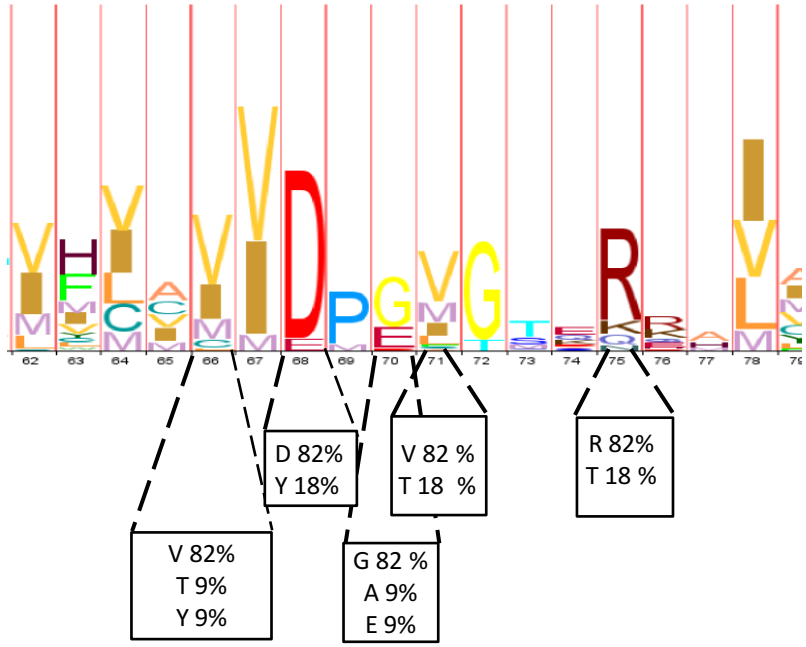


Figure 7.3.: *Comparison of Pfam sequence conservation in a section of the alignment containing key residues as compared to Eustáquio et al. [295]. Drop down boxes correspond to the percentage of each residue type in the corresponding alignment column of Eustáquio et al. [295] as compared to the proportion of residues presented for each sequence position by the Pfam sequence logo. Sequence logo downloaded from the Pfam resource [296].*

chaea and the bacteria, and subdivisions of these - the Crenarchaeota, the Eurarchaeota, the Proteobacteria and the Actinobacteria are well supported.

Such strong support may seem surprising having used such a small dataset and using a neighbour joining strategy. The neighbour joining strategy can be seen as a heuristic to find the tree topology that represents minimum evolution. This method is time-efficient, but is susceptible to LBA and only ever finds one heuristic solution. Although the bootstrap values could be taken as an indicator of the quality of this tree, the bootstrap value indicates the degree of repeatability not accuracy [298]. Phylogenetic inconsistencies including compositional bias, LBA and heterotachy are systematic biases that mean the wrong tree topology can be inferred with high statistical support [299]. Increased taxon sampling has been shown to improve the accuracy and minimise these inconsistencies in phylogenetic trees [298] as has using a probabilistic model that can more realistically model the evolutionary process such as ML or Bayesian methods [299, 300, 301, 302].

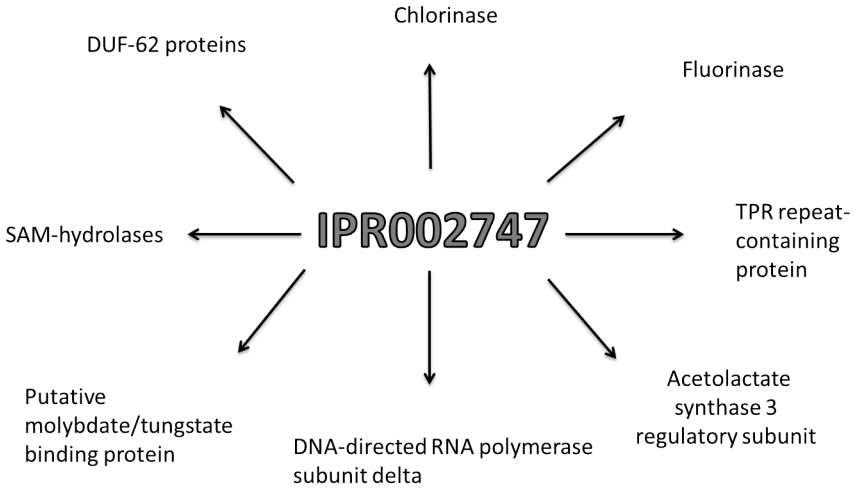


Figure 7.4.: *The diversity of functions assigned in the IPR002747 sequence family [116].*

7.1.4. What contribution do we make to the field?

In our work, we aim to expand and improve upon this phylogeny. First, we aim to to improve the taxon sampling by utilising the Pfam database. We use Pfam as our base as its use of profiles and HMMs enables proteins that are more divergent to be found - giving us a greater indication as to the extent of the conservation of this gene across the tree of life. As discussed, this gene is well distributed throughout the prokaryotes - with only a few representatives in eukaryotes. We found that this gene is therefore present in a much wider range of species than was demonstrated by such as those by Eustáquio et al. [295]. Using our more broadly representative taxon sample, we use a more sophisticated alignment strategy that incorporates information from solved structures to inform sequence alignments [136]. Rather than assuming that DUF62 superfamily evolution can be modelled by an assumption of 'minimum evolution', as underlies the NJ strategy [142], we pick a model that fits our data using the AIC and BIC criteria and use a maximum likelihood phylogenetic strategy that can incorporate this model to find the most likely tree topology given our alignment.

Using this tree topology, we root our gene tree by reconciliation with a species tree. Eustáquio et al. assumed an evolutionary ordering of functional groups where the halogenases are derived from the SAM hydrolases - but have insufficient evidence to propose this - since their phylogeny was not rooted using a species tree informed approach [295].

Although Eustáquio et al. in 2008 noted that no conserved operon structure exists for the DUF-62 gene no attempts have been made to survey the gene context across a phylogenetic tree. Surveying any changes in gene cluster structure in conjunction with phylogenetic groupings and root has the power to reveal changes

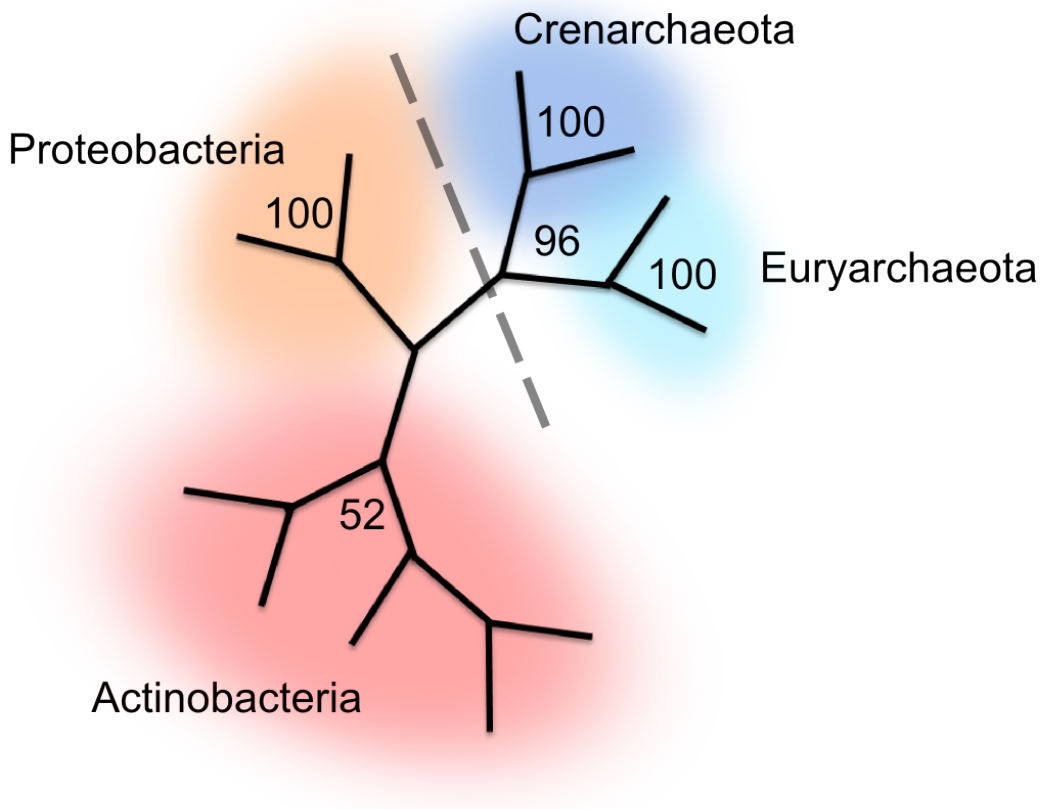


Figure 7.5.: *Neighbour Joining Tree of archaeal and bacterial DUF-62 sequences by Eustáquio et al. adapted from Figure 1 [295]. Original bootstrap values (2000 times in study) have been rounded to whole percentages for figure. Grey dashed line indicates division between archaea and bacteria.*

in function or cellular role for this gene [293].

7.2. Methods

7.2.1. Construction of phylogeny from whole Pfam seed

The seed alignment for family PF01887 was downloaded from Pfam on 12.08.14. Since this dataset consisted of a high number of sequences RaxML [303] was used to build phylogeny specifying LG [144] as the substitution model as selected by MODELGENERATOR AIC & BIC criteria [239] with four discrete gamma categories, without the use of empirical base frequencies and letting RaxML estimate the proportion of invariant sites.

7.2.2. Construction of phylogeny from only archaeal members of the Pfam seed

Seed sequence accessions were taken from archaeal Pfam family 01887 and used to download sequences from UniprotKB in batch-mode [91]. It was found that initial runs with DIVERGE [197, 304] for this sequence set generated errors. Investigations determined that this was probably due to the inclusion of two sequences that were too close in sequence identity to one another. Therefore, the Pfam sequence accession from *Pyrococcus furiosus* was replaced with *Pyrococcus yayanosii* since this was more divergent from *Pyrococcus horikoshii* but did not alter broad phylogenetic position, this accession was then used to download the sequence from UniprotKB.

The sequences were then aligned using Expresso from T-Coffee [136] at http://tcoffee.crg.cat/apps/tcoffee/do:expresso. All relevant methods (excluding RNA) were chosen to construct the library. PDB structures were automatically associated with each input sequence.

MODELGENERATOR [239] was used with four discrete gamma categories to estimate the model of evolution that best fitted the data. LG+I+G was chosen by AIC2 and by BIC, LG+I+G+F by AIC1. The PhyML 3.0 [155] server at http://www.atgc-montpellier.fr/phyml/ was used to build the phylogeny using an LG model of evolution, with model equilibrium frequencies, allowing PhyML to estimate the gamma parameter and proportion of invariant sites, with four substitution rate categories. BioNJ was used as the starting tree, SPR & NNI moves were used to search the topology. Both topology and lengths were allowed to be optimised. The dataset was bootstrapped 100 times.

Shortened taxa names were required for PhyML input (Phylip format). These were converted back to ‘accession_species’ name format in R [240]. In preparation for the following reconciliation analysis, it was made sure that names matched between species tree and gene tree using the NCBI taxonomy browser http://www.ncbi.nlm.nih.gov/taxonomy [305] [306] as a guide for alternative names.

7.2.3. Extraction and editing of species tree

In order to reconcile the gene tree we needed an appropriate species tree. We used the SILVA Tree of Life http://www.arb-silva.de/projects/living-tree/ since it is a consensus of various studies, uses an rRNA dataset, is manually curated and its output does not contain polytomies [307, 308]. The Tree of Life was accessed and downloaded in October 2014 and the archaea clade extracted in Dendroscope [309]. The archaea subtree was then edited to only include species names (deleting family names and accession numbers) using a simple text-string find and replace method and was checked in Dendroscope. It was noted that there were

some duplicate taxa, with the same accession numbers in different positions on the tree. This is due to possession of more than one copy of the ribosomal (rrn) operon in the genomes of some organisms [307, 308]. For most species, sequence similarity is great enough that only one of these paralogs needs to be considered, since there is insufficient difference between them to affect phylogenetic position. However, one species included in our study*, Haloarcula marismortui*- ATCC 43049, has two possible positionings [308]. We therefore created two species tree versions - each with one copy of *Haloarcula marismortui* in each of the two alternative positions, and both versions were used for further analyses.

7.2.4. Reconciliation of core gene tree with species tree using Notung

We reconciled both variants of the species tree with the gene tree for all following analyses. Parameters were varied systematically in Notung [185] and documented. These included: 'rearrange' on and off, with default weightings, 'duplication cost' & 'loss' up-weighted and 'transfer costs' down-weighted. We rationalised the down-weighting of transfers as a result of the known contribution of horizontal gene transfer in prokaryote evolution [233, 187]. We also experimented with the above parameters at a different branch rearrange threshold - at 50%.

Notung does not accept unrooted gene trees for analysis. As such, it is necessary to root the tree arbitrarily before rooting analysis. However, the 'arbitrary root' position can have an impact on subsequent rearrangement analyses, causing a bias between procedures in which we rearranged before rooting and cases in which we did not. This would mean that comparisons between results of rearranged and non-rearranged rooting analyses would not be valid. Hence, we used a strategy in which an arbitrary root was chosen, before a rooting analysis was carried out. We then used this suggested root for further analyses where the tree was rearranged and rooting analysis was carried out again, on this rearranged tree. As such, the tree was rooted in the same position for rearranged and non-rearranged analysis (Procedure suggested by Notung developers, personal communication).

7.3. Results & Discussion

7.3.1. Pfam full seed set

7.3.1.1. Alignment

We analysed the conservation of residues thought to be important for function, as annotated in the literature across the seed alignment downloaded from Pfam.

The alignment was analysed in Jalview [244] using Taylor colouring [243] at 30% identity threshold (Fig. 7.6).

Both Deng et al. in 2008 and Eustáquio et al. highlighted the conservation of a triad of residues postulated to be essential for SAM hydrolysis [286, 295, 310, 311]. In fact, this triad is thought to be involved in the activation of water for the enzyme reaction to proceed by an S_N2 mechanism as proposed by [311, 286] and corroborated by [312].

Using the Pfam seed alignment as a diverse representative for the DUF 62 family we find that Asp 68 is totally conserved and Arg 75 is well conserved (residue numbering as seen in [310]). These residues are indicative of SAM hydrolase functionality rather than halogenase functionality [286, 295, 311]. We find that the third member of this catalytic triad, His 127, is much lower in its level of conservation ([310] numbering). We found this His residue to be substituted for Tyr, Pro, Arg and Thr amino acids in many cases. A wide ranging BLAST search (~300 members) by Deng et al. also revealed this trend [310]. The authors rationalise this lack of conservation of the His residue as a sign that different nucleophiles might be able to be accommodated [310]. We also found that the 'GV' (70 & 71 fluorinase numbering as found in Deng et al. [286]) of the DPGVG motif is surprisingly unconserved, given that these residues were particularly well conserved in alignments by Eustáquio et al., Deng and O'Hagan and Deng et al. [295, 286, 311]. We also found the Arg in the motif TFHGRD (129 fluorinase numbering as found in Deng et al. [286]) not well conserved, this is in contrast to Eustáquio et al., Deng and O'Hagan and Deng et al. [295, 286, 311]. On the other hand, we find the PDNG (120-124 [286] fluorinase numbering) motif well conserved, especially 'N' & 'G' which corroborates the alignments generated by Eustáquio et al., Deng and O'Hagan and Deng et al. [295, 286, 311].

The general similarity in conservation of motifs such as that of 'PDNG' (120-124 fluorinase numbering as found in Deng et al. ([286]) indicates that Pfam members of this superfamily share homology with sequences found in Eustáquio et al., Deng and O'Hagan and Deng et al. [295, 286, 311] . Overall, the high conservation of two members of the SAM hydrolase triad indicates a superfamily in which many members perform a function more similar to the SAM hydrolase versus the chlorinase or the fluorinase. However, the lack of conservation of the His 127, numbering as found in Deng et al. [286], is intriguing, as this may point to flexibility/promiscuity in the binding of substrate and nucleophile in this family, allowing for a diverse range of functions to evolve.

7.3.1.2. Phylogeny

The groupings on this tree reflected the separation of the archaea from the bacteria, and the separation between the Crenarchaeota and the Euryarchaeota which was well supported. However, the bacterial group is not monophyletic and

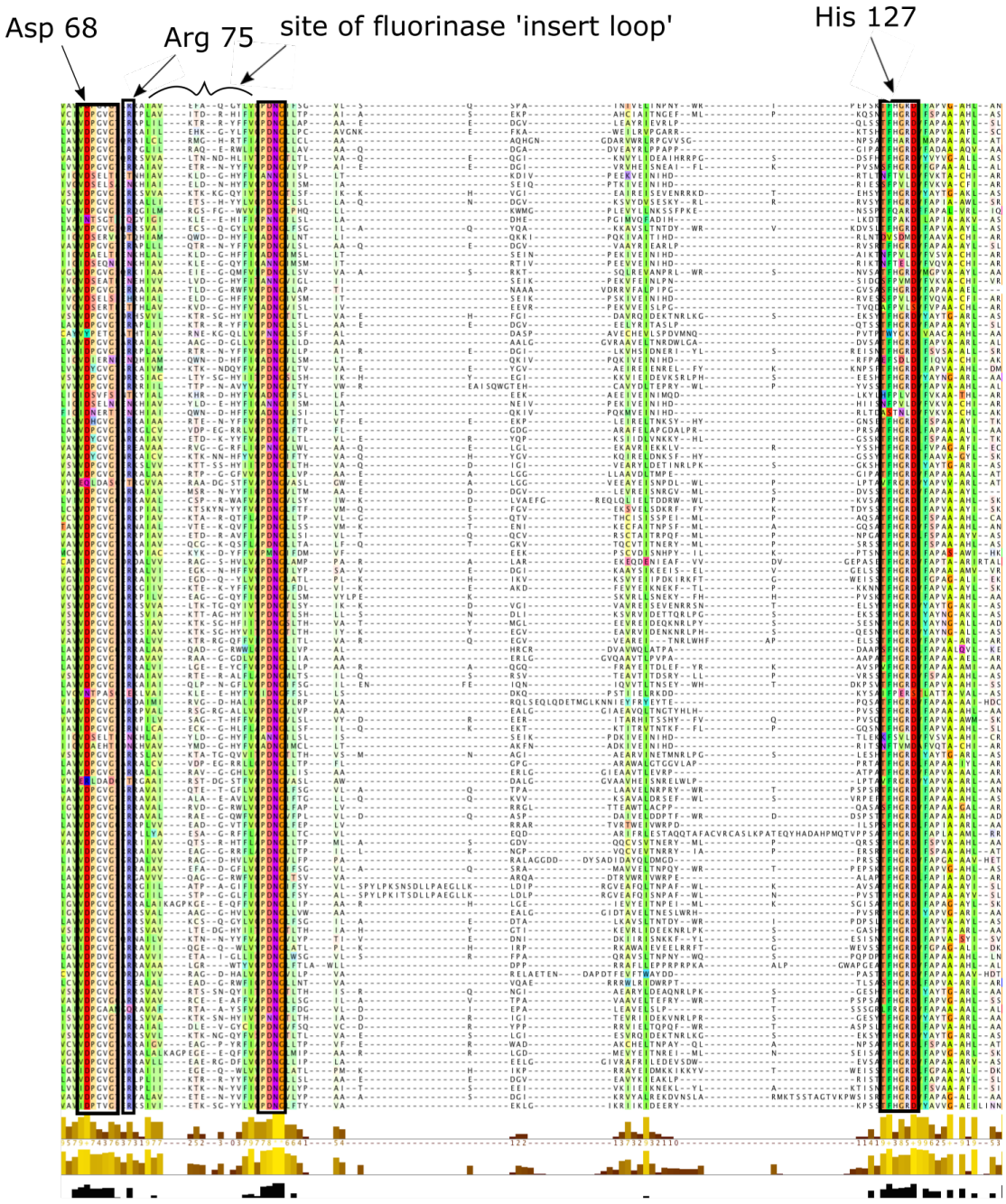
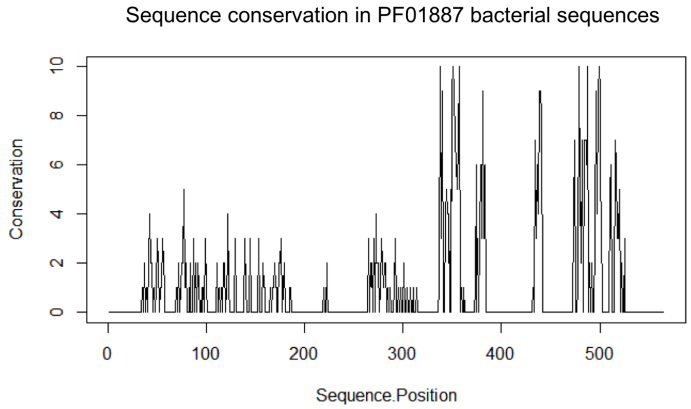


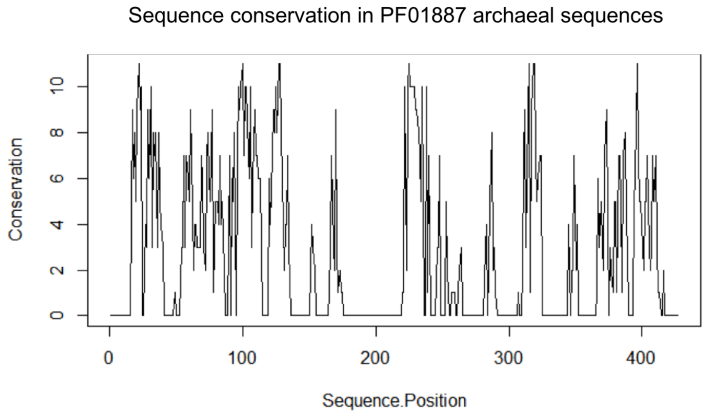
Figure 7.6.: *Section of the alignment from the full Pfam seed as retrieved from PF01887 [296]. Visualised using Taylor colouring [243] at a 30% conservation threshold in Jalview [244]. Highly conserved motifs have been highlighted, including those of the SAM hydrolase catalytic triad.*

although appearing to emerge twice is also not well supported. Overall, the tree had many nodes with extremely low measures of bootstrap support, especially within the bacteria. We therefore chose not to use this dataset for further analysis. The phylogenetic tree and associated alignment can be found in the supplementary information.

7.3.1.3. Domain conservation



(a)



(b)

Figure 7.7.: *Index of conservation (y-axis) plotted against sequence position for bacterial and archaeal sequence alignments downloaded from Pfam [296].*

Low bootstrap support can be indicative of evolutionary signal that is not consistent across the length of an alignment. We examined the level of conservation of residues across the length of the Pfam alignment (Fig. 7.7). These alignments represent all archaeal and all bacterial members of this Pfam family.

Sequence and structural analysis (Fig. 7.1) reveal that members of this superfamily have proteins composed of two domains. This is evidenced at a structural level as CATH classifies members as being a product of two domain structures. Pfam, however, classifies the two domains by one Pfam signature. It therefore appears that at least for many members of this superfamily, the domains are translated together.

It was found that in the archaea a good level of conservation was seen across both domains of the whole alignment, although less conservation was seen around sequence position 200. This is to be expected, since these positions correspond to the linker region between the two domains.

In bacteria high conservation of residues across both domains is not found. In the first domain, many sequences show a level of conservation that is approximately less than half that seen in the archaea. In the second domain, conservation levels are more similar in magnitude to the archaea, but areas of high conservation occur at a lower frequency across this second domain sequence than seen in the archaea.

The difference in conservation between the domains across the archaea and bacteria calls into question the relative functionalities of these two domains. Could it be that the lack of conservation of the first domain in the bacteria indicates a difference in function or cellular role?

Using solved structural representatives can help indicate the role of these two domains. For the SAM hydrolase (*Pyrococcus horikoshii*, PDB: 2WR8 chain A) the two domains have distinct roles in catalysis. The first domain (CATH - 3.40.50.10790) is where the catalytic triad, which activates the nucleophile, is found [310, 311].

The fluorinase functions as a ‘dimer of trimers’ [313] with three monomers in which each N and C terminal of each domain are in contact. SAM is bound between these domains along with a fluoride ion [314]. The SAM hydrolase is also assembled as a trimer with monomers assembled in a similar manner. The chlorinase is similar in trimer structure [287].

The fluorinase binds the fluoride ion with residues Thr 80 & Ser 158 which are found in the first domain (3.40.50.10790) [286]. However the substrate SAM also plays a role in binding the fluoride ion - effectively trapping it and desolvating it [314]. The unique insert loop found in the N-terminal domain of the fluorinase moulds the active site to bring the substrate into contact with Ser 158 and Thr 80, numbering as found in Senn et al. [314, 287]. It is thought that the difference in halide specificity between the fluorinase and chlorinase is in part due to the modification of active site structure created by this unique insert loop [287].

The chlorinase nucleophile bonding is similar to the fluorinase - except Gly 131 replaces Ser 158 to bind with the chloride ion in the first domain [287]. Both the chlorinase and fluorinase appear to have evolved to desolvate their respective

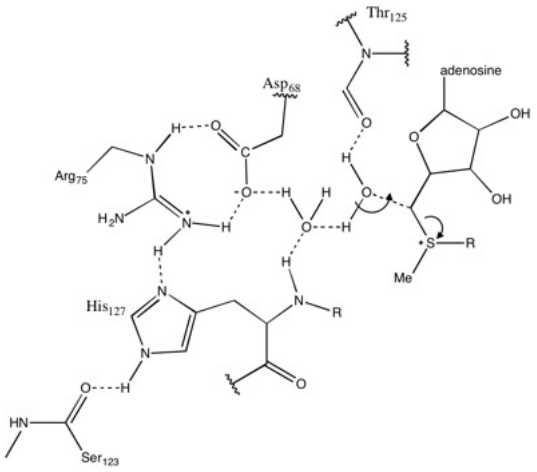


Figure 7.8.: *A mechanism proposed for nucleophilic substitution of SAM by activated water [311]. In the active site of the* Pyrococcus horikoshii *DUF-62 enzyme, the His127, Arg75 and Asp68 triad are thought to constitute a hydrogen bond network that participates in a 'electronic proton relay' as shown in the diagram above [311]. In the proposed mechanism, the hydrogen bond between His127 and Arg75 weakens the hydrogen bond interactions between Arg75 and Asp68 [311]. This increases the basicity of the Asp68 carbonyl group, where it can now abstract a proton from a nearby bound water molecule [311]. It is thought that this now activated water molecule can directly initiate nucleophilic attack on SAM, or, as shown in the above diagram, can abstract a proton from another proximal water molecule, that then initiates nucleophilic attack on the C5' carbon of SAM [311]. Diagram created in ChemDraw [260] and based on scheme 4, from Deng et al. [311].*

halide ions, thereby activating them for nucleophilic attack [287]. The importance of this halide ion desolvation and close juxtaposition of halide ion and SAM molecule contributes to the spectacular rate enhancements these two halogenase enzymes achieve, 2 x 10^{15} and 1 x10^{17} fold increase in the fluorinase and chlorinase reaction rates respectively as compared to the uncatalysed reaction [312].

Comparing the roles of these domains in different structurally solved members of this superfamily, it becomes clear that the N-terminal domain, CATH - 3.40.50.10790, has a role in activating a range of nucleophiles. The SAM substrate is bound *via* different combinations of N and C terminal domains for different functions. All three structural representatives carry out their reactions in an assembly where the substrate is catalysed between the interface of an N and C terminal domain of different monomer units.

For other annotated functions in this superfamily, including acetolactate synthase 3 regulatory subunit, DNA-directed RNA polymerase subunit delta and putative molybdate/tungstate binding protein, a common theme exists of binding nucleotide or amino acid substrates. For example, the structurally solved acetolactate synthase 3 regulatory subunit binds valine between the N-termini of two monomers, positioned back to back (in contrast to the above) [315].

The difference in conservation patterns of sequences between archaeal and bacterial members may indicate a large shift in function between these groups, possibly after a HGT event. The low bootstrap values of our Pfam seed tree incorporating bacterial and archaeal members may be indicative of the difference in conservation in the two groups. The archaeal sequences were, in general, better conserved across the whole length of the alignment so we chose to use archaeal seed members only for further analysis.

7.3.2. Archaeal seed sequences

Although we decided to restrict our analyses to archaeal sequences only, the sequence set still exhibits much diversity when looking across archaea. In fact, members of the Crenarchaeota and the Euryarchaeota are very distinct and even can be thought of as sub-domains rather than phyla [316]. We aim to capture the diversity of sequence features across the wide occurrence of the gene across the archaea.

7.3.2.1. Alignment

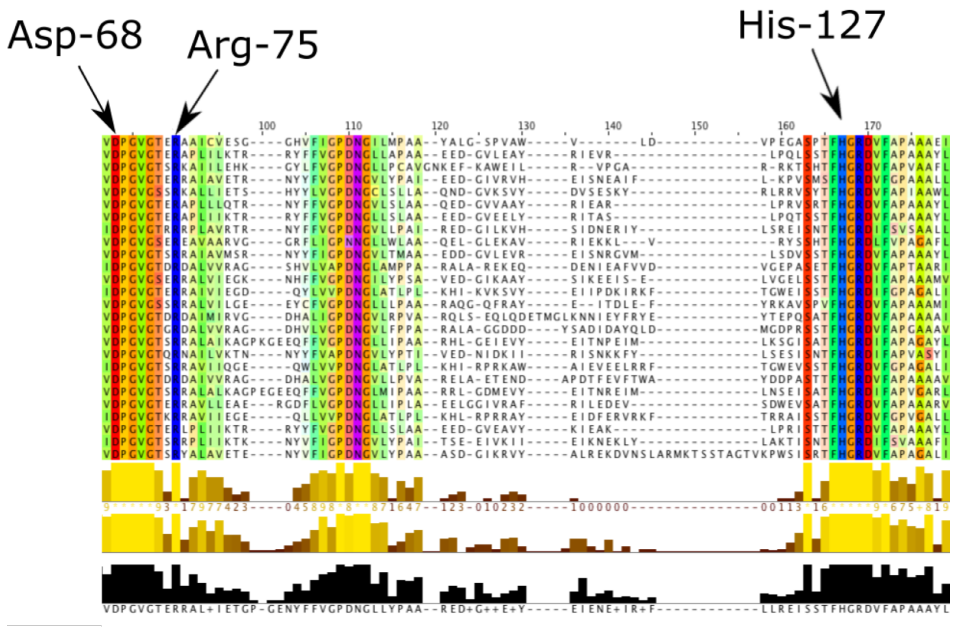


Figure 7.9.: *Alignment of archaeal seed members of PF01887 [296]. Visualised using Taylor colouring [243] at a 30% conservation threshold in Jalview [244]. The alignment positions corresponding to the catalytic triad are highlighted.*

Our alignment of the archaeal seed members (Fig. 7.9) shows that the catalytic triad (as discussed previously) is totally conserved. It is interesting to note that

His 127, numbering as found in Deng et al. [286], is completely conserved in alignment, despite its apparent flexibility in conservation as noted by ourselves and by others [310]. The pattern of conservation in this family indicates that this gene is likely to function by activating water as a nucleophile in the archaea, as opposed to a chloride or fluoride ion in which the high conservation of this triad is not found.

The motif DPGVG {68 - 72 numbering as found in Deng et al. [286]) although not well conserved across the whole Pfam seed alignment is well conserved for archaeal members. This pattern is in common with the Arg in the motif TFHGRD (125 -130 numbering as found in Deng et al.[286]) which, totally conserved in our alignment, is not well conserved when looking at the alignment of the whole Pfam seed set.

When looking across the whole alignment, it is clear that despite this absolute conservation in residues involved in SAM hydrolysis and binding, there lie intervening stretches of amino acids that are not as well conserved. This pattern is in common with our study of the metallo-β-lactamases. This phenomenon seems common for divergent superfamilies, in which the large evolutionary time scale means only residues under strong selection pressure to retain function are conserved.

7.3.2.2. Phylogeny

At the broad level, the phylogeny built from the Pfam seed alignment (archaeal members only) reflects species groupings, with divergence between the Crenarchaeota and Euryarchaeota having 81% support (Fig. 7.10). This is in line with the high levels of divergence between these groups [316]. More recent divergences within these groups are less well supported, in particular, in the grouping of the haloarchaea with the uncultured methanogenic archeaon (*Methanocella aravozae*), the position of *Haracula marismortui*, the placing of *Staphylothermus marinus* and the placing of *Thermofilum pendens*, *Ignicoccus hospitalis*, *Aeropyrum pernix* and *Hyperthermus butylicus*. It comes as some surprise that some of the least well supported divergences are not those found deepest in the tree, as compared to general observations from other superfamilies.

Low bootstrap support can be seen as the level to which the alignment supports the given tree topology and is an indication of how well support for a given topology is distributed over the alignment. It therefore seems likely that the lack of conservation for regions of the alignment, where gaps are frequent, may mean that certain taxa ‘jump around’ in the bootstrap set.

Alignments from divergent superfamilies tend to contain regions of low conservation and gaps. One way of attending to this problem is to ‘mask’ alignments with programs such as BMGE [237] and TrimAL [317]. However, for divergent

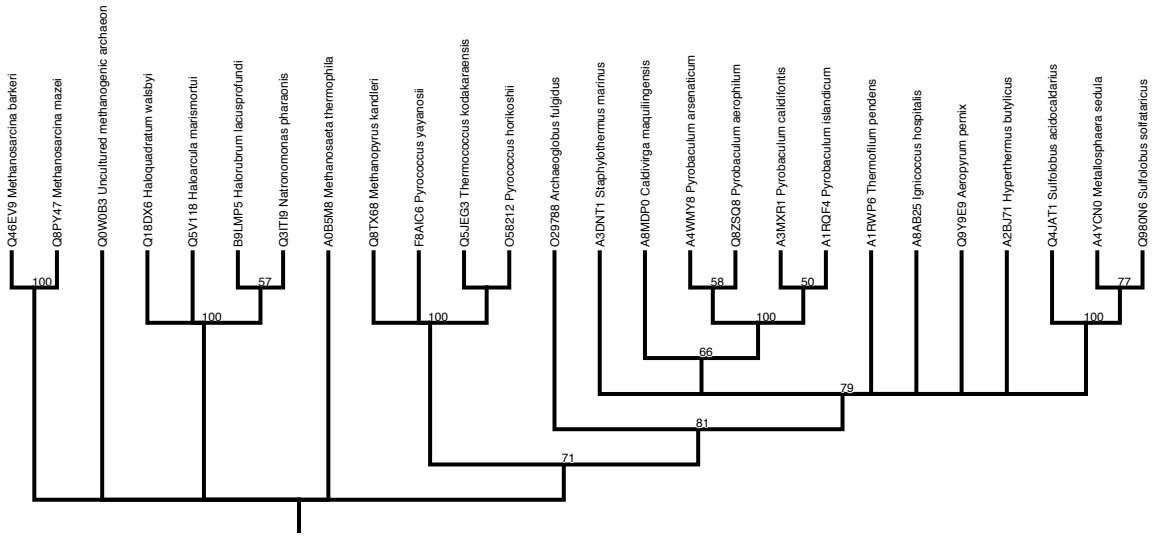


Figure 7.10.: *ML phylogeny of archaeal seed sequences from PF01887, nodes with less than 50% bootstrap support have been collapsed. Phylogenies visualised using Mesquite [255].*

sequences such as these, the risk that all signal may be lost is high. We found that BMGE trimming of a superfamily alignment worsened tree support. This corroborates the results of the BMGE paper which found that for a divergent set of sequences, alignment trimming does not always lead to a better supported tree and is therefore not always appropriate [237].

7.3.3. Rooting analysis in Notung

In order to find the origin of this gene in the archaea, we reconciled the gene and species trees according to the protocol outlined in (sec. 7.2). Tab. 7.1 details parameters that were changed and the effect on inferred root position and root score.

It was found that adjustment of event cost weights (duplications, transfers and losses) had an effect on the inference of root position, as did allowing Notung to rearrange branches with support at 50% and 90%. Notung provides the user a way to explore alternative evolutionary scenarios dependent on evolutionary assumptions made by the user. Starting with default parameters, we explored different assumptions using an iterative process (Tab. 7.1). Rearrangement thresholds were varied to explore confidence in our tree topology. The costs of transfers and losses as opposed to duplications were varied to explore the likely dominant process in evolution.

Costs	Rearrange?	Rearrange event score	Root position inferred between all sequences and ...	Root score
default (transfer- 3, duplications- 1.5)	TRUE	0	uninformative	0
default (transfer- 3, duplications- 1.5)	FALSE	NA	Tied	37
transfer- 1.5, duplications- 3	TRUE	9	Methanogen & Pyrococcus clade	2.5
transfer- 1.5, duplications- 3	FALSE	NA	Haloarcula marismortui	19
edge weight threshold- 50%, transfer- 1.5, duplications- 3	TRUE	36	Methanopyrus kandleri	9
edge weight threshold- 50%, transfer- 1.5, duplications- 3	FALSE	NA	Haloarcula marismortui	19
edge weight threshold- 50%, transfer- 1.5, duplications- 3, loses- 3	TRUE	78	Tied	13.5
edge weight threshold- 50%, transfer- 1.5, duplications- 3, loses- 3	FALSE	NA	Haloarcula marismortui	19.5

Table 7.1.: *Results of rooting analysis in Notung with different event costs and thresholds for rearrangement.*

We found that with default parameters and no rearrangements the root was inferred as a tie between multiple clades and *Methanopyrus kandleri*, and with rearrangement at 90% the root was inferred as lying between the most basal bifurcating branches.

When the cost for duplication was increased, and the cost for transfer was decreased, it was found that the root was inferred to lie between *Haloarcula marismortui* and the other species when the tree was not rearranged. This is in contrast to the result when the same event weightings were used, but Notung was allowed to rearrange the tree at 90% and 50% threshold levels. At 90%, the root is inferred as positioned between a large clade consisting of the methanogens and Pyrococcus, and a clade consisting of all other species in the study. At 50%, the root is inferred between *Methanopyrus kandleri* and all other species in the study.

How do we choose between these different evolutionary scenarios? The simplest way is to use Notung's strategy of minimising costs in evolution - thereby assuming a parsimonious model of evolution. In general, root cost scores are lowered when we allow Notung to rearrange the tree and when transfers are given a lower cost than duplications.

If we work purely off the 'lowest cost' logic, then we find that the lowest root lies between the clade of methanogens and *Pyrococcus* species and the rest of the taxa in this study, given a root score of 2.5. However, this is when we allow Notung to rearrange all bifurcations with support of less than 90% to a more parsimonious solution (i.e. closer to the topology of the species tree). The problem with this strategy is that evolution does not always take the simplest route and gene trees often do not match species trees.

In order to balance the logic of parsimony and yet allow room for gene tree deviation from the path of minimum evolution, we lowered the Notung threshold for rearrangement to 50%. In doing so, we put into place a prior assumption that divergences supported by equal to or more than 50% bootstrap support are not to be rearranged even if they do not follow the most parsimonious path of evolution. Lowering this rearrange threshold led to root scores with a higher cost than when we allowed Notung to rearrange a wider selection of branches (equal to or

above 90%). This strategy, which determined a root lying between *Methanopyrus kandleri* and all other species in this study, with a root score of 9 provided the best balance between accounting for HGT in evolution whilst ensuring that those divergences we were confident about according to alignment data were maintained.

There was a concern that although we pruned our tree (to ensure species not present on our species tree were not penalised as 'losses'), inferring transfers also, by definition assumes a loss event. Since our gene dataset is filtered, it is possible that loss costs are being inferred as a result of transfer events. In an effort to buffer these loss costs, we increased the 'loss' event cost. It was found that doing this led to many optimal roots all with high costs (*Methanopyrus kandleri* was one of these) but the same number of transfers, duplications and losses were inferred during rearrangement.

Be that as it may, our representative but sparse dataset means we do not get a picture of copy number - so we cannot infer the full picture of duplication and transfer events at this stage. Despite this limitation, we can use this reconciliation method to infer the likely root and ancestral habitat, under varying parameters.

Analyses with species trees 1 and 2, in which *H. marismortui* differed in its phylogenetic position, led to exactly the same results for root position and costs.

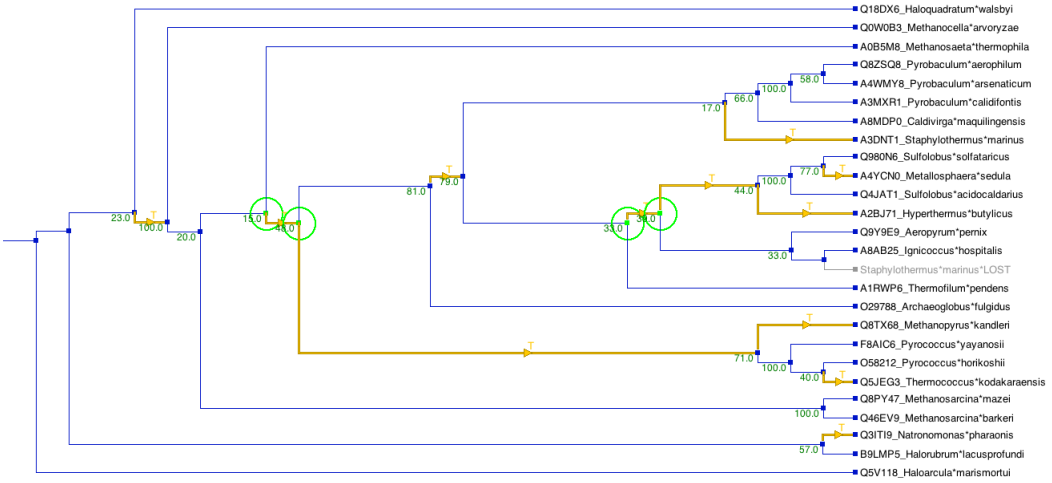
7.3.4. Rationalising different root hypotheses in Notung

We can analyse evolutionary histories under different parameters in order to rationalise the root inferred under a biological context.

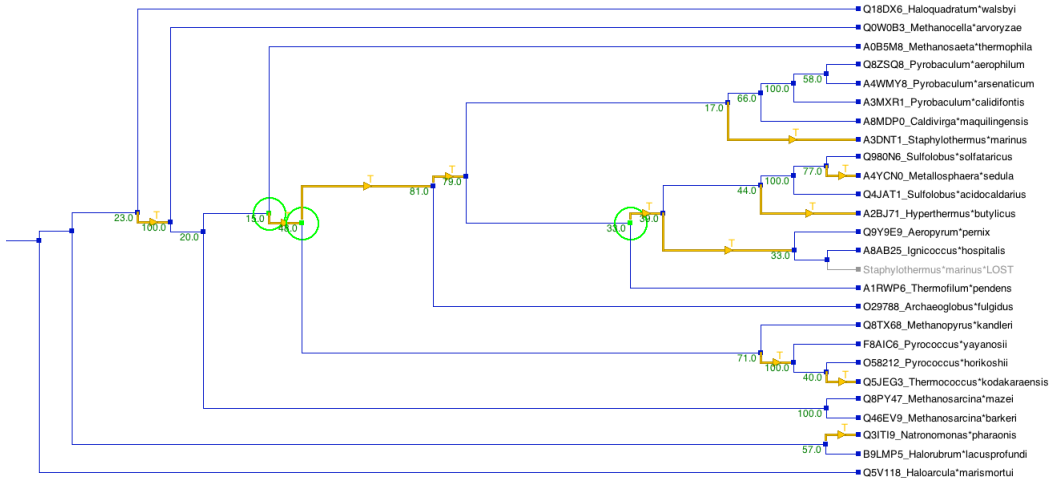
7.3.4.1. Rearrange off - *Haloarcula marismortui* as the sister group of all other species on the unrearranged tree

Under this scenario, by a parsimonious reconstruction of the ancestral environmental state, the root of this gene in the archaea resides in a hypersaline environment. Due to the number of transfers inferred (12) multiple optimal solutions can be found, most of these transfers occur within the hyperthermophilic clade. Looking across all possible transfer histories in Notung, a transfer between hypersaline members to the methanogens and another transfer event between the methanogens and those of the hydrothermal vent clade is in common for all (Fig. 7.11). Methanogen and hyperthermophilic environments are derived traits, with methanogen evolving first and giving rise to hyperthermophilic.

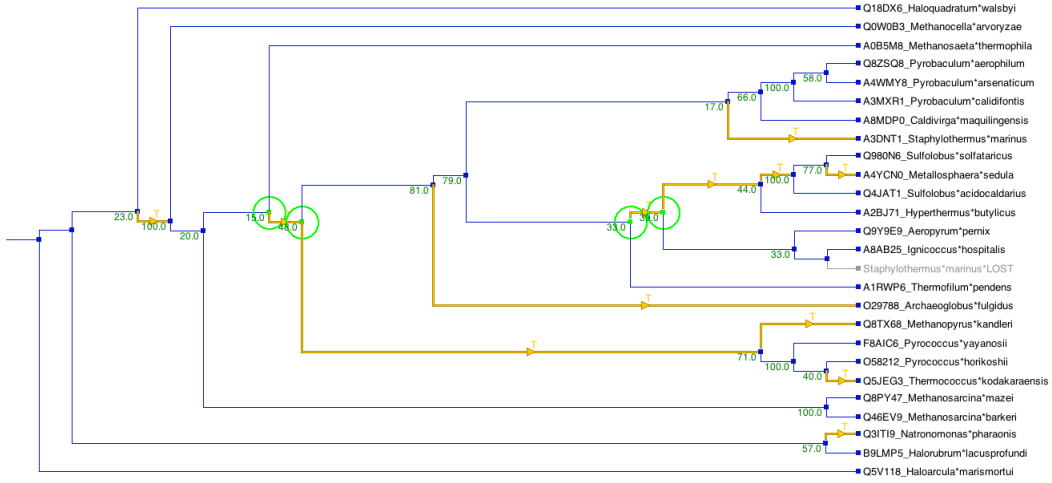
Evidence for a hypersaline ancestor, as is proffered by the unrearranged tree, is thin [316] as it is generally thought that a hyperthermophillic archaeal ancestor is a more likely scenario[316, 318, 54]. In addition, HGT is a well known phenomenon within the archaea but across vastly differing environments seems rare.



(a)



(b)



(c)

Figure 7.11.: *ML phylogeny of archaeal seed sequences from PF01887 drawn with the root lying between* Haloarcula marismortui *and all other species on the unrearranged tree. Predicted transfer events under this scenario are shown in yellow. Sub-Figures depict a range of optimal solutions as calculated by Notung. These optimal solutions were selected by clicking on the green circles. Figure generated in Notung 2.8.1.2 [185, 186].*

7.3.4.2. Rearrange on - *Methanopyrus kandleri* as the sister group of all other species on the rearranged tree at 50% threshold

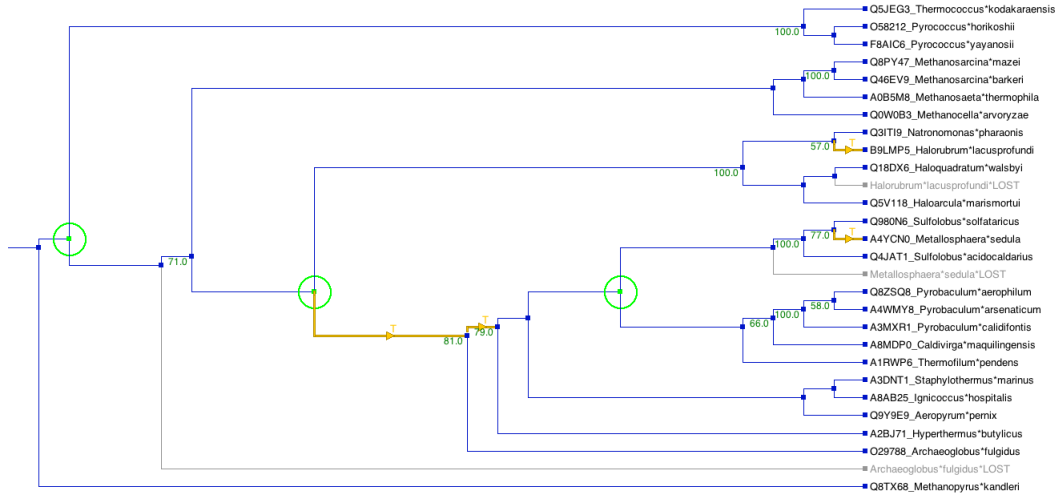
Under this scenario, the root of this gene in the archaea resides in a hydrothermal vent environment. Due to the number of transfers inferred (four) multiple optimal solutions can be found. Note that this, as compared to the un-rearranged tree above, is a smaller number and indicates a more parsimonious result. Although in this scenario it is clear that the ancestor of the archaea had a hydrothermal vent environment, it is not clear how many times a hydrothermal vent habitat emerges on the tree, although it would be parsimonious to assume that it is the ancestral state. If the ancestral trait of the hydrothermal vent only appears once, then methanogen and hypersaline would be the derived traits, with hypersalinity evolving either after HGT or divergence. HGT to hypersaline members in this scenario would mean transfer between organisms of different environments, which would seem unlikely. However, this is one of a range of possible optimal scenarios for this reconciliation, some of which do not involve transfer between different environments (Fig. 7.12). This is in contrast to the unrearranged tree reconciliation with a hypersaline root, in which two transfers between differing environments are always needed for an optimal solution. In contrast to the lack of a hypersaline ancestor, much evidence exists for a hyperthermophillic Last Common Ancestor (LCA) (e.g. [316, 318, 54]).

7.3.4.3. Ancestral habitat - the argument of parsimony

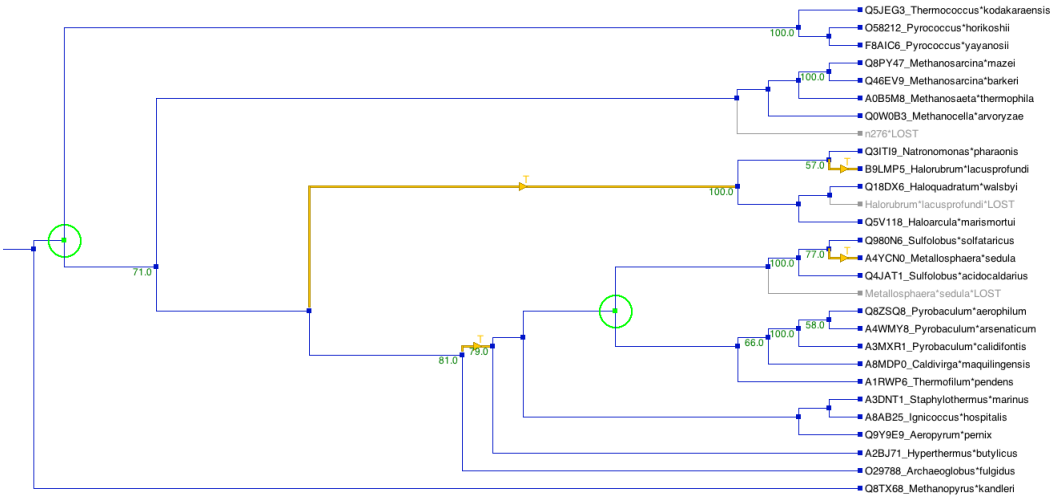
Taking the above analyses and arguments of parsimony into account, it seems likely that the root lies between *Methanopyrus kandleri* and all other species in the study. In this proposal, we assume that transfers are more likely to have occurred than losses, and that clades of gene tree with more than 50% support accurately represent evolution even if not parsimonious in reference to the species tree.

In reference to other studies, it is interesting that the high rate of evolution of *Methanopyrus kandleri* makes it hard to place [316]. There is some debate as to whether to place *Methanopyrus kandleri* with other methanogens (e.g.[319]) or nearer the root (e.g. [318, 320, 321]). For our gene tree, the placement of *Methanopyrus kandleri* is away from the methanogens and monophyletic with *Pyrococcus* and *Thermococcus* with moderately high support (71%).

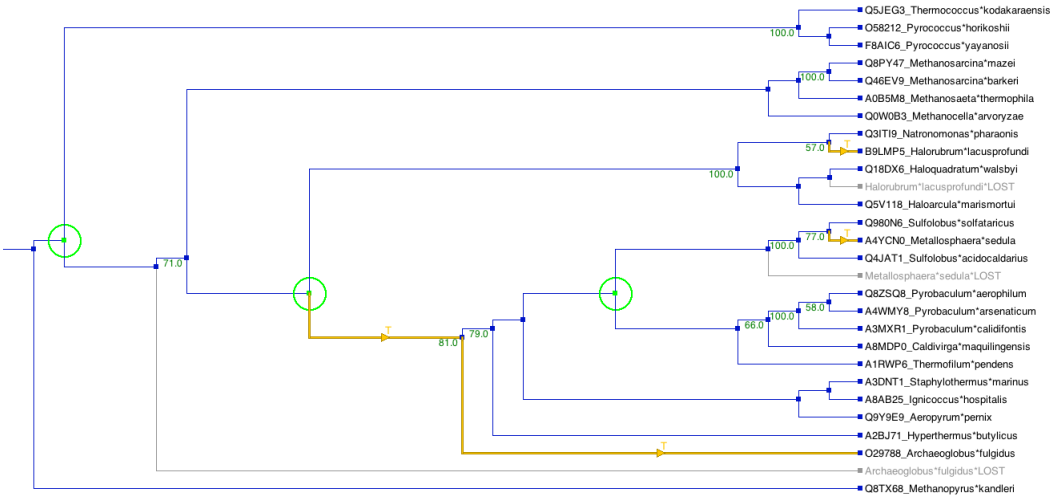
Methanopyrus kandleri, although present in a hydrothermal vent environment (black smoker), also exhibits features of other extremophiles - including the ability to carry out methanogenesis and a high salinity tolerance. *Methanopyrus kandleri*'s survival in hypersaline environments appears to be made possible by a higher than average level of negatively charged amino acids in its proteome [319]. It seems that many other archaeal members in this superfamily may thrive in high salinity environments - such as other 'hydrothermal vent' members and those from



(a)



(b)



(c)

Figure 7.12.: *ML phylogeny of archaeal seed sequences from PF01887 drawn with the root lying between* Methanopyrus kandleri *and all other species on the rearranged tree at 50% threshold. Predicted transfer events under this scenario are shown in yellow. Sub-Figures depict a range of optimal solutions as calculated by Notung. These optimal solutions were selected by clicking on the green circles. Figure generated in Notung 2.8.1.2 [185, 186].*

hypersaline lakes. This may constitute a counter argument to the hypothesis that the generation of protons by enzymes in this family confers a fitness advantage in extreme environments commonly found in this superfamily. In high salinity environments, commonly encountered by archaeal members of this superfamily, the generation of protons is unlikely to aid in the generation of negatively charged amino acids.

The prevailing opinion appears to be that methanogenesis evolved only once in the Euryarchaeota but was then lost on multiple occasions [321]. In fact, Gribaldo and Brochier-Armanet propose a controversial theory that the LCA of the archaea was a methanogen, but this activity was then lost in Crenarchaeota and some linages of Euryarchaeota [316].

7.3.5. Modelling the structure and inferring the possible function of the DUF-62 gene from *Methanopyrus kandleri*

Since we postulate that *Methanopyrus kandleri* has had little possibility of habit change since the root in the archaea, we submitted the protein sequence of this gene for homology modelling. We used I-TASSER [322] to do this, thanks to its high rankings in recent CASP experiments [323] and its integrated function prediction - including predicted ligand binding and EC number [322].

TM-align, as utilised by I-TASSER, predicted our protein to structurally closest to 2ZBU chain A, an uncharacterised conserved protein from *Thermotoga maritima*. Whereas the prediction of ligand and EC number both predicted our *Methanopyrus kandleri* homology model closest to the fluorinase 2V7X.

Both these results come as a surprise. Given its phylogenetic position, one might expect the sequence to be closest to the archaeal structurally solved structure from *Pyrococcus horikoshii* (2WR8, UniProtKB O58212 included in our study). However, its close similarity to bacterial members may be indicative of its basal position in this phylogeny.

7.4. Conclusion

In conclusion, sequence analysis of this superfamily indicates that the pattern of conservation of residues is more diverse than demonstrated by sequences found in Eustáquio et al., Deng and O'Hagan and Deng et al. [286, 295, 311] and may be indicative of more, as yet, unexplored functions within this superfamily. Building a phylogeny for this whole family remains a challenging task - leading to trees with low statistical support. This may be due to a difference in the conservation of

sequence between the archaeal and bacterial domains. Sequence conservation distribution appeared higher for archaeal members of this Pfam family, as such, we only included archaeal members for future analysis, in which we built a new, better supported phylogeny.

Reconciliation of this archaeal gene tree with the species tree generates different evolutionary histories and estimation of the root depending on the parameters used. One of the key differences is the difference in root inferred when allowing Notung to rearrange or not. We determined that allowing Notung to rearrange branches with less than 50% support struck the balance of using a parsimonious strategy, whilst allowing evidence from our gene tree and alignment to overule if well supported. This approach, in conjunction with decreasing transfers costs and increasing duplication costs, as expected in this family, gave us a root position between *Methanopyrus kandleri* and all other species included in this study, suggesting a hydrothermal vent dwelling ancestor of this superfamily. This correlates with studies that propose a hyperthermophilic last common ancestor.

Homology modelling and function prediction of the *Methanopyrus kandleri* DUF-62 gene yielded surprising results. Given its phylogenetic position, one might expect this model to most closely resemble its structurally solved DUF-62 gene from its relative *Pyrococcus horikoshii*. In fact, the homology model indicates that structurally, and functionally, *Methanopyrus kandleri* resembles homologous members in the bacteria and actually postulates fluorinase ligand binding and function. This surprising result may be indicative of its basal position in the evolution of this superfamily.

8. How many roles does the DUF-62 gene perform in the archaea?

The DUF-62 gene is well conserved across the prokaryotes, and the archaea. The archaeal kingdom comprises of organisms that live in a wide range of habitats, including hydrothermal vents, hypersaline lakes, and sewage. These organisms have evolved particular adaptations in their metabolism for such environments, for example, *Methanopyrus kandleri* has an unusually high proportion of negatively charged amino acids in its proteome, thought to be of benefit in high salinity conditions [324].

As pointed out by Eustáquio et al. [295], the wide distribution of this gene across the prokaryotes makes it likely that its role lies in primary, rather than secondary metabolism. For example, the hypothesis by /todoDeng and O'Hagan Deng and O'Hagan [286] that this gene may play a role in cellular homoeostasis in extreme conditions. However, we note there are many different 'extreme' conditions encountered by this family and yet the presence of the gene is conserved across many organisms. Taking this in conjunction with the fact that much diversity exists in the archaeal sequences away from the catalytic triad [311, 295] it seems possible that multiple functions or cellular roles have evolved in the superfamily, that may, or may not, have a link to the specific 'extreme' environment of the organism.

Using our rooted ML phylogeny, derived by methods described in the last chapter, we hope to discern the possibility of divergent functions or cellular roles within this family. Unlike the metallo-β-lactamases, as described in Part II, and other well studied superfamilies, only a handful of members of the DUF-62 superfamily have been structurally and functionally characterised. This poses a challenge; for the metallo-β-lactamases we were able to discern the function of unknown members by comparison and reference to a wealth of well characterised examples. In particular, we used knowledge of structure and spatial location of amino acids to predict function.

Despite the challenges involved with studying enzyme families with few structurally characterised members, the literature has many examples of studies which have attempted to overcome this. Hicks et al, in their study of the strictosidine synthase-like proteins, used sequence similarity networks in conjunction with phylogenetics and gene context analysis, despite the existence of few structural

representatives in this family [325]. Gene context, especially for prokaryote genomes, is a useful measure of a gene's interaction partners and cellular interactions. Such a strategy has been utilised in studies by authors such as Gerlt et al. [6] for assigning functions in the enolase family and by Makarova et al. [326] to infer metabolic pathways in the archaea. Most recently, an investigation by Mudgal et al. compared the functional prediction accuracy and sensitivity of a range of methods thought to be suited to the identification of remote similarity between proteins [4]. The authors found that using a strategy in which the results of multiple methods combined could be used as a measure of the robustness of the prediction [4]. A study such as this highlights the benefits of using multiple methods to predict function, and is a strategy we implement in this work.

8.1. What we contribute to the field

In this study, we ask how many functions and/or cellular roles have evolved in this superfamily. This goes beyond the broad functions already assigned to these proteins (e.g. by UniProtKB).

Although not functionally characterised, many members of this gene family are well documented in terms of their sequence and genomic context. We can use *in silico* techniques to track possible changes in function or cellular role across this phylogeny in conjunction with our ML tree, in which we hypothesise the evolutionary ordering of taxa and their environments. In doing this we choose to remain, as in the rest of this thesis, at the protein level, reflecting the high levels of divergence found within enzyme superfamilies.

Since the gene is found almost exclusively in prokaryotes we can make use of a given gene's location in the genome to infer function. This is due to the fact that in prokaryotes, enzymes that function in the same metabolic pathway are often translated from neighbouring genes in the genome; these clusters of genes operating together are called 'operons'. The existence of operons is prevalent in the prokaryotic kingdom [327] (although some examples from the eukaryotes exist [328]).

We used the STITCH resource [329] to query each DUF-62 gene in our dataset for possible protein and chemical functional interactions. The STITCH resource is a database which integrates information from a range of sources including those from experiments, databases and text mining of the existing literature [329]. STITCH gives a measure of confidence for the resulting predicted interactions [329]. Such an approach in integrating data from different sources increases the robustness of predictions as compared to using protein-protein interaction datasets alone [330]. Other methods which build functional linkage networks from integrated data sources have been developed including VIRGO (VIRtual Gene Ontology) [331] and VisANT. Unfortunately, the VIRGO server seems

no longer maintained http://whipple.cs.vt.edu:8080/virgo and VisANT is aimed at establishing disease-protein function relationships. Since no known members of the archaea have been established as pathogenic [332] we established that the STITCH databases were most appropriate for our study.

We then looked to quantify the results of the STITCH analysis. We assume that change in predicted protein interactions may be evident in a protein's amino acid sequence. We use DIVERGE [198, 197] (discussed in Chapter 4) to assess this.

Although structural information is sparse for this particular family, we can still make best use of the structures we do have by use of a program such as DIVERGE. By quantitatively comparing sequence conservation between phylogenetic groups, DIVERGE is able to identify common residues conserved within a family and those that differ between phylogenetic groups.

The residues that are restricted to particular phylogenetic clades can be important indicators of neofunctionalisation events. The method used by DIVERGE in which the phylogenetic conservation of residues is used to indicate the importance of residues has much in common with the Evolutionary Trace method (ET) [193, 196, 333]. However, the evolutionary trace method does not give an indication of how statistically significant a particular residue of 'importance' is, making distinguishing between residues that have diverged by neofunctionalisation versus neutral drift (for example) particularly difficult. DIVERGE uses a framework in which it tests the assumption that residues are correlated across different phylogenetic lineages. The extent of this correlation is given a score, allowing the user to determine the 'functionally divergent' residues in a more objective manner. For our study, we only know the residues that are important (and therefore conserved) for SAM hydrolysis functionality. We hope that our study may reveal other important functional residues in this superfamily, in an objective manner using DIVERGE.

Once the evolutionarily important residues have been identified (either by ET or in our case, DIVERGE) they can be then mapped onto an available structure, where the spatial clustering of these evolutionarily 'important' residues might be seen. This mapping onto structure is key, since identifying the conservation of residues at the sequence level does not reveal as much about catalytic function as diagnosing their relative position in 3D space - for example, around a binding, catalytic or allosteric site. Even though we only have a limited number of structural representatives for this family, the mapping of these evolutionarily 'important' residues onto any homologous structure is useful - since approximate clustering and proximity to the known catalytic site can be observed.

We then annotate the predictions of protein associations and change in function onto our phylogenetic tree. Using the reconciliation software Notung [185, 186], we infer the history of duplication and transfer events. As was discussed in the previous chapter, our current reconciled gene-species tree topology does not include all possible information on duplication events - due to our 'sparse' dataset.

We noted that the use of *all* sequences, even from the archaeal family only, resulted in a badly supported topology hence our decision to use the PFAM 'seed' set. In order to introduce extra information on duplications without introducing additional ML tree ambiguity, we chose to only add sequences for those clades which had the highest levels of predicted divergence as inferred by DIVERGE.

8.2. Methods

STITCH prediction of protein interactions Proteins were queried as a batch job at http://stitch.embl.de/cgi/show_input_page.pl (accessed 05.05.15) with each entry represented by a UniProtKB accession code. All parameters were left as default including the use of a medium confidence level score threshold (0.4).

Protein-protein interaction data and chemical interaction was downloaded as a plain text file for each entry. It should be noted that at the time of retrieval, STITCH did not have an entry for *Pyrococcus yayanossi*, so information of interactions of the closely related *Pyrococcus furiosus* was downloaded instead.

The text data on predicted protein interactions for each entry was then used to annotate the ML topology, using a different colour to distinguish each protein interaction partner.

Analysis of seed gene tree in DIVERGE 3.0 The seed gene tree (from previous chapter) was used to test clades with different habitats. Sequences had to be more than 4 per clade. The methanogens did not form an isolated clade, so were excluded from this analysis. The methanogen clade was then used to root the tree, leaving 3 clades to test: 'Hypersaline lake' (Hsl), 'Hydrothermal vent 1' (Hv_1) and 'Hydrothermal vent 2' (Hv_2). We only picked clades which were well supported. Nodes were collapsed at below 50% and groups were chosen after this.

It should be noted that initial tests with the Pfam seed set yielded 'NaN' numbers with DIVERGE 3.0 using the 1999 and 2001 algorithms. After testing, it was found that this seemed to be due to very similar sequences *Pyrococcus horikoshii* and *Pyrococcus furiosus*. Since *Pyrococcus horikoshii* is the only taxon with a structurally solved DUF-62 protein, we resolved the problem by replacing *Pyrococcus furiosus* with a slightly less similar sequence, *Pyrococcus yayanossi,* that still lay within the *Pyrococcus* clade. This resolved the problem for the 1999 type I algorithm, but not the type II or 2001 algorithm.

Results from DIVERGE were analysed at the whole sequence level but also at the residue level. The posterior probability of each residue contributing to functional divergence can be analysed in isolation. These can be mapped onto structure as described previously. Those specific residues that had a posterior probability

above a given threshold (at default = 0.5 and 0.7) were then analysed in terms of their mapping for a structural representative in this sequence set (PDB: 2WR8, from *Pyrococcus horikoshii*) with annotated catalytic amino acids.

Subtree Analysis Using the taxonomy browser in the Pfam family the *Thermoproteaceae* and the *Halobacteria* were selected and downloaded in 'FASTA' format and sequence accessions, since these were the two family levels that contained seed sequences from the clades with the highest level of divergence. These were combined with the original seed set from Pfam. The sequences were sorted to remove duplicate accessions - since original 'core' sequences and additional sequences had been added - leading to duplicate accessions in selected clades. The sequences were then submitted to 'Expresso' with parameters used in seed gene tree analysis (previous chapter). The aligned combined sequences were submitted to 'Modelgenerator' using 4 gamma categories. The AIC, AIC2 and the BIC all chose LG+G+F as a model. The same alignment was submitted to PhyML 3.0 at http://www.atgc-montpellier.fr/phyml/ specifying the LG model with empirical equilibrium frequencies, '0' proportion of invariant sites, 4 substitution rate categories, SPR & NNI moves and 100 bootstrap replicates.

Tips were converted using 'R' by the method as described in the previous chapter for the Pfam seed gene tree. Like the gene tree and species tree it was found that there were alternatives for some taxa names, these were reconciled by the same methods used in the previous chapter.

The subtree-gene tree was reconciled with the species tree, in Notung version 2.8.1.2.beta, using the same protocol as for the seed gene tree and same phylogeny for the species tree as the previous chapter, whilst keeping the outgroup taxon *Methanopyrus kandleri* consistent and rearranging at a 50% threshold level each time.

8.3. Results & Discussion

8.3.1. Prediction of DUF-62 interactions using STITCH

After colour coding each enzyme function as a different colour, we annotated the whole, unrearranged tree with the results only for those clades with 50% support or more.

From manual inspection, and using a qualitative approach, it appears that there is a non-random pattern of gene distribution, with particular clades appearing to have different patterns according to our colour coding of the different gene predictions.

Strikingly, and despite different patterns seen in different clades, the majority (excluding *Methanopyrus kandleri* and *Pyrobaculum aerophilum*) share predicted interactions with one copy or more of nicotinimide-nucleotide adenyltransferase (*nadR*) genes. This gene encodes an enzyme that synthesises nicotinamide adenine dinucleotide (NAD+) from nicotinamide *D*-ribonucleotide, adenosine triphosphate (ATP) and a proton [334]. NAD+ is an important coenzyme, and oxidising agent. The conserved pattern of predicted interactions between the DUF-62 gene and the *nadR* gene adds evidence to its possible role in primary metabolism in the archaea.

If we focus on clades that were used for the DIVERGE analysis (described in 'Methods') and inspect the colour distribution of STITCH predicted interactions and their gene names we find particular sets of gene interactions tend to correspond with the taxonomic grouping within each clade.

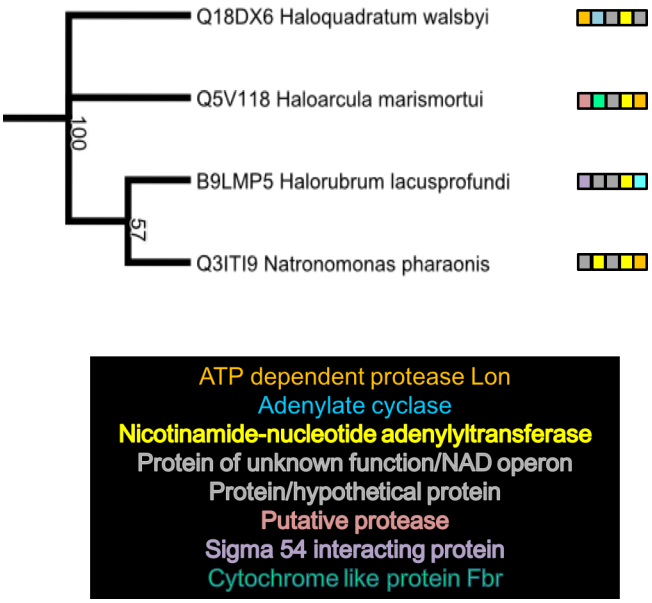


Figure 8.1.: *STITCH predicted interactions of the DUF-62 genes in the Hsl clade. Coloured boxes to the right of the phylogeny correspond to the coloured names of proteins in the black coloured box. Phylogeny visualised using Mesquite [255].*

Hypersaline lake In this clade, we not only see the nadR gene, but also, frequent occurrences of ATP dependent protease Lon and also, in two out of the four members, adenylate cyclase (Fig. 8.1).

ATP dependent protease Lon has roles in the degradation of misfolded and abnormal proteins, as well as possible DNA binding activity [335]. Focusing on its physiological role in the Archaea, experiments show that in a fellow halophile

Haloferax volacanii the Lon protease is essential for viability due to its purported role in the regulation of membrane lipid composition [336]. A later proteomic study by the same authors revealed that Lon may be involved in many more cellular roles, including co-enzyme metabolism, amino acid biosynthesis, genetic processes, transcriptional regulation and ABC transportation [337].

ATP is converted to 3'-5' cyclic AMP by the adenylate cyclase enzyme. Cyclic AMP is an important secondary cell messenger for a range of functions, including signal transduction.

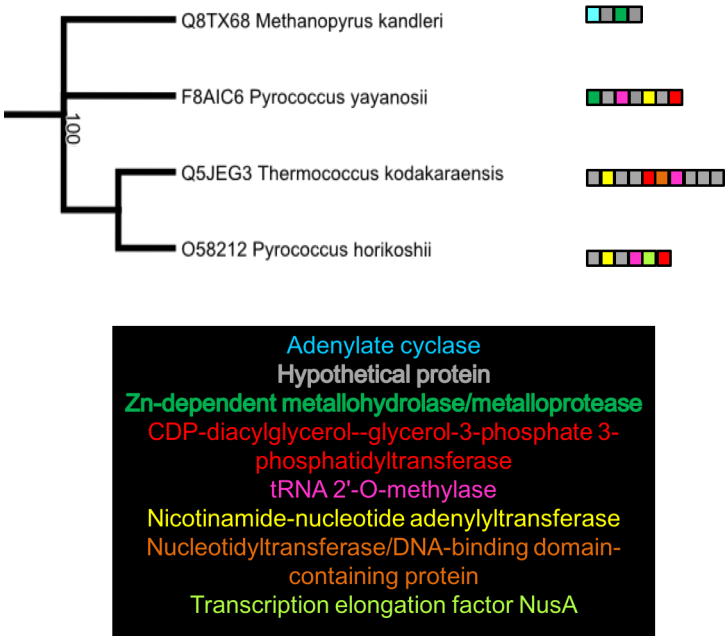


Figure 8.2.: *STITCH predicted interactions of the DUF-62 genes in the Hv_1 clade. Coloured boxes to the right of the phylogeny correspond to the coloured names of proteins in the black coloured box Phylogeny visualised using Mesquite [255].*

Hydrothermal vent 1 Three out of four members of this clade have predicted interactions with CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase (PGP synthase) (Fig. 8.2). This enzyme, also a transferase (like NADR), catalyses the production of CMP from CDP-diacylglycerol and glycerol 3-phosphate. This enzyme plays a role in the metabolism of glycerophospholipid, an important component of biological membranes [334].

The tRNA guanosine-2'-*O*-methyltransferase, another transferase, is predicted to interact with the DUF-62 gene for three out of four members (Fig. 8.2). The tRNA-2'-*O*-methylase catalyses the formation of *S*-adenosyl-*L*-homocysteine from *S*-adenosyl-*L*-methionine and tRNA substrates [334]. *S*-adenosyl-*L*-homocysteine can be then converted to homocysteine which has many different roles in the cell.

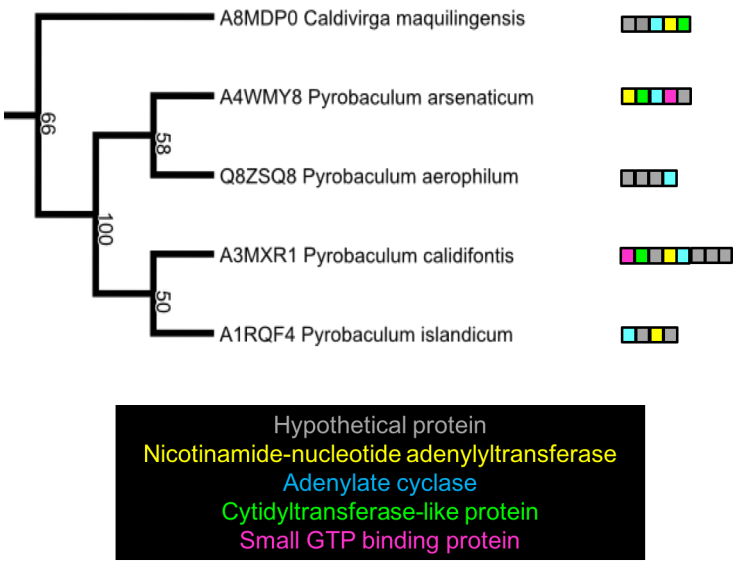


Figure 8.3.: *STITCH predicted interactions of the DUF-62 genes in the Hv_2 clade. Coloured boxes to the right of the phylogeny correspond to the coloured names of proteins in the black coloured box Phylogeny visualised using Mesquite [255].*

Hydrothermal vent 2 A cytidyltransferase-like protein is a predicted interacting partner for three of five members of this clade (Fig. 8.3). Cytidylyltransferase uses CTP as substrate before releasing pyrophosphate after catalysis [334]. A web search yields information on multiple types of these enzymes, including - choline-phosphate cytidylyltransferase, phosphatidate cytidylyltransferase and ethanolamine-phosphate cytidylyltransferase. All of these enzymes are capable of partaking in glycerophospholipid metabolism. All members of this clade are predicted to interact with adenylate cyclase (Fig. 8.3).

A striking feature of this qualitative study is that each DUF-62 gene in each clade has predicted interactions with enzymes that contribute in some way to membrane biosynthesis. Notably though, these are not through the same routes in metabolism for each clade.

There are a number of mechanisms by which archaeal membranes differ from those of bacteria and eukaryotes. It is thought that some of these differences may constitute adaptations. Firstly, archaeal lipids are constructed *via* an ether bond joining the alcohol and lipid components [339]. This is in contrast to eukaryotes in which an ester bond joins the glycerol and lipid moieties [339]. Secondly, the structure of the lipid chains often differs in the archaea - featuring branched, or less commonly, ring structures in side chains. It is thought that both these differences may make for 'tougher' membrane structures, since ether bonds are chemically more stable than ester bonds and branched lipid chains may help create a membrane that is less permeable to leakage [339]. The degree of cyclisation has been shown to be adaptable by organisms such as *Sulfolobus solfataricus [340].*

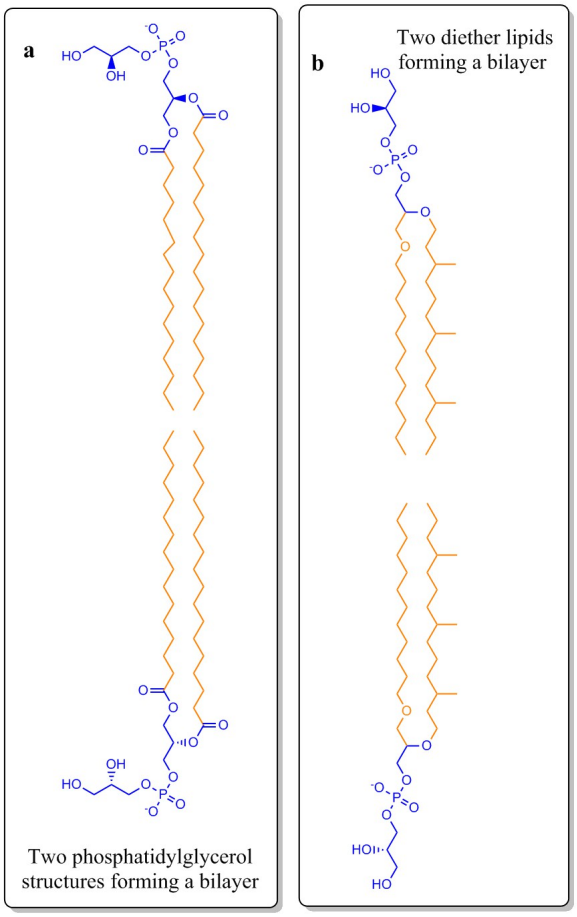


Figure 8.4.: *Comparison of bacterial (left) and archaeal (right) lipid structures. Adapted from Albers and Meyer [338], Box 1 and created using ChemDraw [260].*

Using the STITCH resource has allowed us to review predicted interactions of the DUF-62 gene across our archaeal phylogeny and allowed us to determine for which clades these predicted interactions differ. However, these predictions are not quantitative, and a change in interaction, even if accurately predicted does not necessarily imply a change in function.

If the DUF-62 gene does indeed have different interactions in different clades of our tree, we assume this will be evident at the sequence and structural level. At such a point, we might be in a better position to evaluate whether any noted changes in sequence conservation between these phylogenetic groupings may have a bearing on the function of these proteins, using our limited knowledge about structure and function for this family.

Statistic	hsl/hv_1	hsl/hv_2	hv_1/hv_2
ThetaML	0.06	0.32	0.25
AlphaML	1.58	1.03	1.18
SE Theta	0.14	0.11	0.10
LRT Theta	0.17	8.35	6.26

Table 8.1.: *Results of the DIVERGE analysis using the 1999 algorithm. The analysis was performed in a pairwise manner for each of the selected clades. ThetaML is the ML estimate of the coefficient of divergence, AlphaML is the ML estimate of the among site variation, SE Theta is the standard error of ThetaML, LRT Theta is the log likelihood ratio score as compared to the null hypothesis (ThetaML = 0).*

8.3.2. Testing for differences in sequence conservation across different clades

DIVERGE allows the user to compare the rates of evolution between sequences of different clades in a phylogenetic tree, measured by the conservation of sequences. This difference is then expressed as a coefficient from 0-1, with 1 being the greatest degree of inferred divergence. Although in general, no one comparison has a particularly high measure for ThetaML, our results show that the highest level of divergence lies between the Hsl and Hv_2 clades, the next between Hv_1 and Hv_2 clades and the least between Hsl and Hv_1 clades. However, it should be noted that the result for the comparison between Hsl and Hv_1 clades comes with a particularly large value of standard error.

Since DIVERGE tests the level of correlation between the rates of evolution of two clades it is possible for the user to assess the significance of ThetaML using the calculated LRT against a χ^2 distribution with 1 degree of freedom (df). The 5% cut-off value for 1 degree of freedom is 3.841, in which the Hsl/Hv2 and Hv_1/Hv_2 LRT statistics exceed whilst the Hsl/Hv_1 LRT statistic does not exceed this value. In fact, the Hsl/Hv_2 value also exceeds the 1% cut-off value with 1df of 6.635. On the basis of this, it is tempting to argue that there is evidence for the difference in the rates between Hsl and Hv_2 being statistically significant.

However, we must be aware that our experimental design involves multiple independent comparisons. Such a strategy increases the chance of false positives (Type I errors). To do this, we can convert our likelihood ratio statistics to p-values using the χ^2 distribution before using a correction, such as the Bonferroni. However, this approach is thought to be very conservative (an example of a discussion on this topic can be found by authors such as Perneger [341] along with its response) possibly leading to the generation of false negatives. The False Discovery Rate (FDR) is a less stringent approach, and has been utilised by ourselves

to adjust the p-values calculated from the Likelihood ratio statistics (found in Supplementary Information). We found that when adjusted by FDR, our DIVERGE results for Hsl/Hv_2 and Hv_1/Hv_2 were both significant at the 5% level but the Hsl/Hv_2 result was no longer significant at the 1% level.

However, using a χ^2 distribution to convert between a likelihood ratio statistic and a p-value is only an approximation. Corrections for multiple testing, such as the Bonferroni, violate the likelihood principle (e.g. by authors such as Barnard et al.[342]) since they assume that the context in which the data were obtained should be considered whereas a likelihood approach assumes the likelihood of an event can be estimated only by data that has already been collected and the model alone.

Given the philosophical issues with accounting for multiple tests when using LRT statistics, it is important to be careful in assessing the 'significance' of our results. With this in mind, we will discuss the results in terms of the 'strongest' and 'weakest' evidence for functional divergence of clades. Looking from this perspective, the comparison of clades in which we have the strongest evidence of divergence is between Hsl and Hv_2, although as noted previously the magnitude of this result is low on the 0-1 scale of ThetaML values of divergence.

In relating this back to the prediction of interactions in STITCH we remind ourselves that the Hsl clade members have common predicted interactions including ATP dependent protease Lon and adenylate cyclase, whereas in the Hv_2 clade frequently predicted interactions include cytidylyltransferase-like protein and also adenylate cyclase. The predictions of these interactions appear to represent different metabolic pathways, and so the finding that members from these two clades have some evidence for divergence is not surprising. We might have expected the two clades from the same class of habitat (Hv_1 & Hv_2) to exhibit very little evidence for divergence - this was not the case, with thetaML being the 2nd ranked in the set. Although we cannot evaluate our result in terms of overall statistical significance, this result could be indicative that the role of the DUF-62 gene is not correlated with any particular environment. Although, we should be aware that our classification of habitats, including 'hydrothermal vent', is particularly broad and may not capture distinct environments within this niche.

Taking a closer look at protein interaction data for clades of particular interest STITCH generates a large amount of information on predicted interactions, not all of which was discussed previously. We therefore decided to take a closer look at the clades with the highest level of divergence as calculated in DIVERGE and take a deeper look into predicted interactions and the source of these.

To do this, we took a member of each of the Hsl and Hv_2 clades with the most diversity and range of protein interactions, as examined in our previous STITCH analysis. In looking at the predicted chemical interactions two chemicals in particular appear surprising. The first is a chloride ion predicted to interact by our

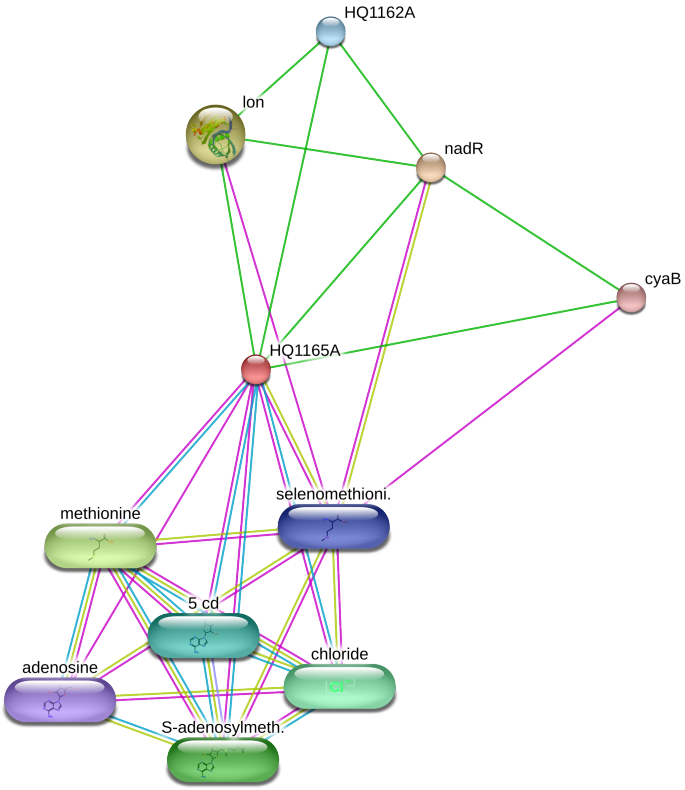


Figure 8.5.: *Predicted protein and chemical interactions as predicted by STITCH 4 for Hsl clade representative* - Haloquadratum walsbyi *DUF-62 protein (red). Line colours are representative of the source of evidence for each predicted interaction, green: gene neighbourhood, fuchsia: experiments, yellow: text mining, cyan: databases. Figure generated using STITCH 4 [329].*

Hsl representative. However, further investigation into the source of this 'experimental' evidence reveals that this is probably in reference to homology with the chlorinase and therefore may be a false positive.

Another curious prediction is that of selenomethionine for both the Hsl and Hv_2 representatives. Not only is selenomethionine predicted to interact with the DUF-62 gene but also other chemicals and proteins too. We suspect that the prediction of an interaction with selenomethionine is actually an artefact from homologous structures in which selenomethionine was incorporated to help elucidate the crystal structure [343]. It therefore should be noted that the combined score of some of these interacting proteins, when they have predicted links to our DUF-62 gene and selenomethionine, is over-inflated by this false positive result. There is still evidence, however, of interactions between our DUF-62 gene representatives and interacting proteins by gene context (green lines (Fig. 8.5) & (Fig. 8.6)).

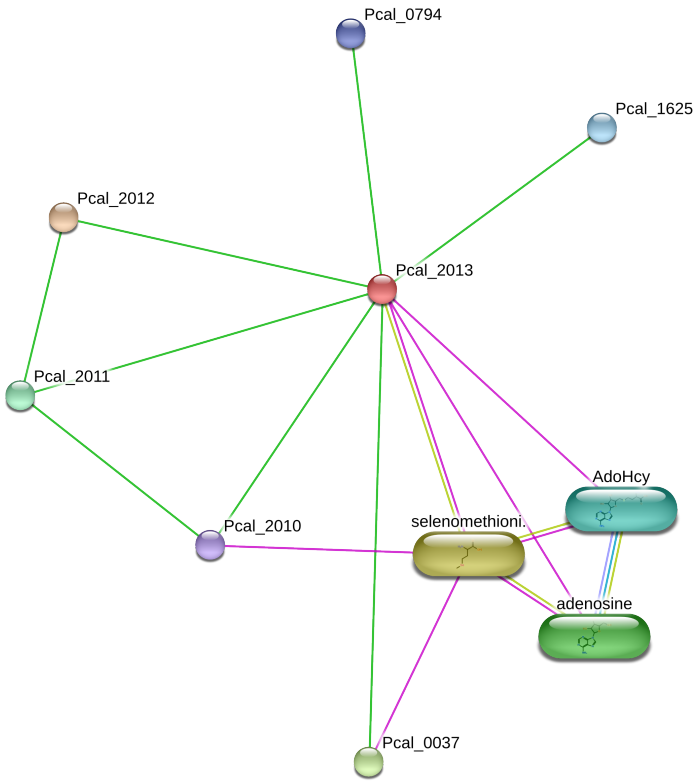


Figure 8.6.: *Predicted protein and chemical interactions as predicted by STITCH 4 for Hv_2 clade representative -* Pyrobaculum calidifontis *DUF-62 protein (red). Line colours are representative of the source of evidence for each predicted interaction, green: gene neighbourhood, fuchsia: experiments, yellow: text mining, cyan: databases, lilac: homology. Figure generated using STITCH 4 [329].*

Adding sequences to clades of interest and reconciliation in Notung If indeed there exists divergence in function between the Hsl and Hv_2 clades, we might expect to see evidence of duplication or transfer events between these clades. After adding additional sequences from PFAM to the Hsl and Hv_2 clades we reconciled them in Notung, using a 50% statistical threshold for rearrangement and placing *Methanopyrus kandleri* as the root.

Under these conditions, no temporally feasible solutions are found with a transfer weight of less than 2.0. With a transfer threshold of over 2.0 and duplications at various weights a transfer event appears to have happened between an ancestor of the *Pyrobaculum* clade (including Hv_2) to the ancestor of the Hsl clade. When rooting with *Methanopyrus kandleri*, the *Pyrococcus* species appear to have diverged before this event. In fact, this major transfer seems to split the Haloarchaea and the methanogens from *Pyrobaculum* (Hv_2). This is in line with our DIVERGE results, in which the greatest evidence of divergence was found

Transfer weight	Duplication weight	Result - event cost	Transfers	Losses	Duplications
1.5	3	No temporally feasible solutions found	NA	NA	NA
3	1.5	27	4	12	2
2	2	22	6	10	0

Table 8.2.: *Results of reconciliation of the species and subtree gene tree in Notung. Transfers and duplication weights were varied and the resulting event cost recorded. A threshold of 50% was maintained for rearrangement each time and rooted with* Methanopyrus kandleri*.*

between Hsl and Hv_2 clades. A transfer event early on in their history seems a convincing scenario in this case. This major transfer event between these groups holds for all parameter weightings tested with feasible solutions (Tab. 8.2). Differences in conservation between members of this family may be less to do with adaptation to different environments and more the result of a transfer event far back in evolutionary history. An interesting implication of this result is that members of the *Halobacteria* clade may not have possessed a DUF-62 gene before transfer and therefore its activity was not required in the *Halobacteria* clade native state, however, after transferral, it appears that the gene diverged and possibly neofunctionalised, taking its place in a new metabolic pathway.

Mapping the most divergent positions onto structure Using DIVERGE we can pinpoint the posterior profiles of the columns of the alignment that contribute most to the MLtheta score. We identified alignment positions with a posterior probability of equal to or more than 0.5 (coloured yellow on Fig. 8.8), and also those positions with a probability equal to or more than 0.7, coloured orange on (Fig. 8.8). We reasoned that those residues surpassing the 0.7 posterior probability were likely to be the strongest contributors to divergence.

We used Jalview [244] to map these alignment annotations onto the SAM hydrolysing enzyme from *Pyrococcus horikoshii*, (PDB:2WR8). Jalview also annotates known information on binding sites (pink) and areas of interest (green). Using Jalview we are able to see the correspondence at both the alignment and structural level of our highly scoring (assumed to be most divergent) columns and previously known information on function.

At the 0.5 posterior probability level, there are 29 residues contributing to the divergence score. In both the alignment and structure, these are spread out, and do not appear to cluster in any specific area. However, some positions coincide with those involved in binding and areas of interest. When we take it to the 0.7 level, all three are close to the binding site (distributed over the two structural domains). One of these, position 313, corresponds to an asparagine residue that is shown to bind substrate in the crystal structure of 2WR8.

None of the ‘divergent’ residues, at either 0.5 or 0.7 thresholds, coincide with the highly conserved motif of the catalytic triad (see previous chapter). It therefore

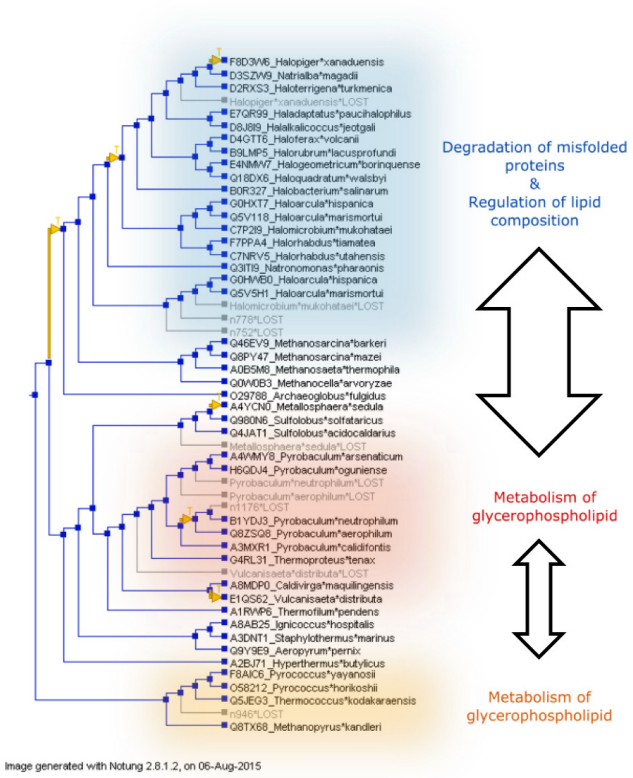


Figure 8.7.: *Reconciled and rearranged subtree gene tree in Notung (50% threshold) with duplications and transfers both weighted at 2 (result 1.3 in supplementary data). Greyed out taxa are those predicted to be lost under this reconciliation scenario. The tree is rooted at* Methanopyrus kandleri. *Colours approximately correspond to core gene tree groupings: Hsl: blue, Hv_1: orange, Hv_2: red. Possible cellular roles for each coloured node are annotated. The arrows (not to scale) correspond to the degree of functional divergence between clades as calculated by DIVERGE. Figure generated in Notung 2.8.1.2.*

seems likely that any possible divergence of function does not lie in the overall chemical mechanism - i.e. hydrolysis catalysed by the conserved catalytic triad rather, divergence of function seems more feasible in those residues involved in substrate/product binding. DUF-62 members of these two phylogenetic groups may hydrolyse different substrates and take part in different metabolic processes possibly interacting with different protein partners.

There has been another protein structure solved for this family - the product of the DUF-62 gene from *Methanocaldococcus jannaschii* (PDB: 2F4N). This organism was not included in our analysis but is presumably most closely related to the other methanogens. No function could be found for this protein and although compared to the fluorinase, the authors conclude that the enzyme is devoid of halogenase activity [345]. This similarity to the fluorinase is in common with predicted EC number and ligand binding of the homology model of the DUF-62 gene product

Figure 8.8.: *Positions with a posterior probability of 0.5 (yellow) and 0.7 (orange) mapped onto PDB:2WR8 in Jalview [244]. Green areas denote regions of interest and pink areas denote binding sites from the annotations from the PDB [344].*

from *Methanopyrus kandleri*, as presented in the last chapter.

8.4. Conclusion

In this work, we have used the topology of our ML tree and our prediction of root taxon as a basis on which to predict the possible existence of multiple functions or cellular interactions for the DUF-62 gene in the archaea. Utilising both a qualitative and quantitative approach and comparing our results to known information on the structure and function for members of this superfamily, we infer that there is evidence from our different lines of enquiry that multiple functions/interactions of the DUF-62 gene may have evolved in the archaea.

Discrete phylogenetic groupings of predicted interactions are observed. Strikingly, the differences in predicted interactions of the DUF-62 gene between the Hsl and Hv_2 clade, and to a lesser extent, the Hv_1 and Hv_2 clade have support from DIVERGE results, and coincide with a major transfer event far back in the history of this superfamily which divided these two groups. Despite the difference in these metabolic pathways, all have connections with membrane composition and therefore may constitute a pathway to regulate lipid composition in extreme environments. Based on this evidence we do not predict that the function of the DUF-62 gene specifically changes in a correlative manner according to organismal environment habitat. However, our definition of ‘environment’ is inherently broad, particularly in our grouping of all ‘hydrothermal vent’ habitats into one category, when in fact, this genre is likely to contain many more diverse sub-environments.

Our study only sheds light on a subsection of the evolution of the DUF-62 gene -

that is, in the archaeal domain of life. In actual fact, and as discussed in the previous chapter this gene is well distributed throughout the organisms, and is found in many species of bacteria, including pathogenic ones. As more structures are solved for proteins in the superfamily, our power to resolve evolutionary signal between distant family members will increase by improvement of the robustness and accuracy of alignments and phylogenetic trees. This will be of particular interest in determining the relationships between the archaea and other members of the DUF-62 family. In this study, we make use of available structural information in our alignment strategy to build ML trees but note that many bifurcations do not have strong statistical support. Therefore, by necessity, our inference of root position is at least in part constrained by the topology of the underlying species tree and our prediction of divergence between clades is limited to those with 50% bootstrap support or more. Although we have made steps to constructively add information from additional sequences in this family to determine duplication, loss and transfer events, we are restricted in our addition of sequences as our previous experiments (previous chapter) show that phylogenies built from larger sets of sequences are not well statistically supported.

As discussed, assessing the significance of multiple independent tests on the same data is a challenge. In this study we have explored the idea of attempting to correct our results using both the Bonferroni and FDR methods. We have mapped residues which were the most divergent between the Hsl and and Hv_2 clades onto the only archaeal structure with functional annotations (PDB:2WR8). Although useful in determining the spatial position of these residues in relation to the SAM hydrolase substrate (S-homocysteine used as an inhibitor in this case) we cannot be sure that all of our DUF-62 representatives have the exact same overall and active site structure as *Pyrococcus horikoshii*. Despite this, a common observation in the comparison of our DUF-62 alignment at the sequence level and at the structural level indicates that divergence of function in this family is likely to reside near the substrate binding site, rather than within sites involved in catalysis. From a BLAST search in Deng et al. the conservation of the catalytic triad presumed necessary for nucleophilic attack is well conserved across many of the DUF-62 sequences, spanning the domains of life [310].

Part IV.

Concluding remarks and future work

9. Conclusions

9.1. The metallo-β-lactamase superfamily

In this work, we have explored the prediction of function of sequences from two enzyme superfamilies. In the first case, our study of the well characterised metallo-β-lactamase superfamily attempted to improve published phylogenies before going on to explore whether the MRCA of the extant metallo-β-lactamases was capable of hydrolysing lactam substrates.

To do this, we needed a good quality phylogeny on which to base ancestral sequence reconstruction. Although phylogenies for this enzyme superfamily had been published in the literature, we noted that at deeper levels their statistical support was low. In order to improve upon this, we used a structurally informed approach by adding sequences to the pre-existing structurally informed FunTree alignment. Using structure to inform sequence alignment is essential for building high quality trees of such divergent sequences. However, a balance must be struck. As reviewed in the literature, a strategy in which only the structurally alignable residues are included generates a short and strongly biased alignment. We extended the number of taxa represented and used a statistical approach to choose the best model of evolution that fitted the data. This improved upon existing phylogenies for the superfamily which either assumed a parsimonious model of evolution, or had evaluated multiple tree topologies by ML analysis but had not chosen the model of evolution to fit the data. FunTree phylogenies use a species guide tree in their construction, essentially setting a prior belief that the evolution of the gene is congruent with that of the species. For our phylogeny construction attempt, such a constraint was not appropriate in discerning the gene tree of prokaryotes well known to have undergone horizontal gene transfer events.

By taking these measures, we have developed and improved the phylogeny for this superfamily, giving a more detailed picture of evolutionary events than others have previously published. Our work, and the work of others, have shown that aligned members of this superfamily share a common scaffold in which metal coordination in order to activate water is conserved. The wide range of functions in this superfamily, particularly in the ability to hydrolyse a broad range of lactam substrates, is conferred by an active site that seems exapted to bind a range of substrates. Understanding this trend is key in deciphering why this enzyme is so readily adaptable to overcoming diverse antibiotic challenges.

Despite taking measures to improve the phylogeny for this superfamily, we found that some of the deepest bifurcations were still not well supported. In general, the major functional groups formed discrete clades with high support, although the low bootstrap support on some deeper nodes meant we were still unsure as to which order these groups diverged. Given the lack of confidence in the deeper branchings of our ML tree, we decided to perform ASR for the MRCA of the metallo-β-lactamases for each tree in the bootstrap sample of 100 trees. By doing this, we were able to observe any commonalities in the properties of the MRCA not apparent when looking at phylogeny topologies only. The majority of MRCA motifs indicated most similarity to a B3 lactamase or a glyoxalase. This observation was consistent at the structural level, in which we found the only homology model to pass both our criteria for being a metallo-β-lactamase most closely resembled our B3 lactamase catalytic template. We found that only 5% of representative homology models for these sequences passed our criteria for having metallo-β-lactamase activity, based on alignment with a metallo-β-lactamase template.

What does our study mean for future efforts to combat antibiotic resistance? Although our results cannot be used to definitely support the case of one or two origins of lactamase evolution in this family, our foray into possible MRCA sequences and structures has yielded useful insights into common trends. Firstly, if we assume that the methods we have employed are capable of determining the general properties of MRCA sequences in this family and we take the results purely at face value, there appears to be few paths in which evolution could have taken to yield a lactamase hydrolysing ancestor. Adding evidence for an exapted fold particularly amenable to evolving machinery capable of hydrolysing metallo-β-lactamase substrates. One common theme is that whether able to hydrolyse a lactam molecule or not, the ancestor was likely to have been metal coordinating (according to our alignment of cluster representatives). The evolutionary flexibility in this family appears to lie in regions other than this metal coordinating motif, presumably in more flexible loop regions. It is therefore likely that whether via multiple independent evolutionary inventions, or by rapid divergence from a lactam hydrolysing ancestor, these enzymes have the flexibility to adapt their exapted structure to face new, but related structures of drug molecules. It seems that targeting metal coordination in these enzymes may be an effective way to hinder the evolution of resistant variants, although the design of a broad ranging zinc chelator without dangerous off-target effects is a challenge for drug design.

One of the main limitations in our study is that it is not possible to know how exhaustively we have sampled tree topology space and if those sampled represent biologically viable candidates. For example, that 54 of our MRCA sequence predictions did not match the metallo-β-lactamase InterPro signature has some bearing on the interpretation of our results. Are these sequences simply 'scrambled' predictions - for which all possible enzyme resemblance has been lost as a result of methodological inaccuracy? If so, then the proportion of 'successful' lactamase MRCA predictions scales to 5/44 - over 10% of MRCA candidates. Looking at

it this way, one might ascertain that over 10% of the bootstrapped phylogenies support the hypothesis that the B1/B2 and B3 lactamases arose as a result of a single, divergent event in evolution. Or, conversely, could it be that these 54 sequences may be the result of accurate phylogenetic reconstruction and ASR, and are simply proteins for which the function is unknown? As discussed, our experimental design means we have no way of rationally assigning probabilities to different tree topologies. Therefore, we have no way of determining if some trajectories are more likely in evolution. Specifically, if the 5% of our MRCA sequences that passed both of our criteria for lactamase activity actually represents a more likely evolutionary trajectory, this creates additional complexity in the interpretation of our 5% statistic.

Even if we were able to identify the more evolutionarily feasible tree topologies in our bootstrap set, our assessment of function is inherently limited by our catalytic templates. Our catalytic templates are based on extant crystal structures - static snap shots of dynamic, flexible structures. Although the catalytic templates give a good indication as to the relative location of catalytic amino acids in 3D space, they give no information as to the flexibility, compensatory and epistatic mechanisms the structure may utilise. As such, our work demonstrates the likely minimal machinery needed by an MRCA to hydrolyse metallo-β-lactamase substrates and quantifies the degree of support by topologies from an intelligent sampling strategy.

9.2. The DUF-62 superfamily

In this work, we have used the topology of our ML tree and our prediction of root position as a basis on which to predict the existence of multiple functions or cellular interactions for the DUF-62 gene in the archaea. We used this case study to contrast with the well studied metallo-β-lactamase superfamily, by using in silico methods to produce a phylogenetic tree and predict function, despite the lack of structurally and functionally characterised members of this family. By utilising both a qualitative and quantitive approach and comparing our results to known information on the structure and function for members of this superfamily, we infer that there is evidence from our different lines of enquiry that multiple functions and/or cellular roles of the DUF-62 gene have evolved in the archaea.

We have demonstrated a protocol in which we use a number of alternative tools to those used for the metallo-β-lactamase family. We extended our search to look for phylogenetic support for neofunctionalisation events, for example after a duplication or transfer event, by reconciling the species tree with the gene tree. Sequence analysis of this superfamily indicates that the pattern of conservation of residues is more diverse than demonstrated by previous studies and may be indicative, of more, as yet, unexplored functions within this superfamily. Building

a phylogeny for this whole family remains a challenging task - leading to trees with low statistical support. In fact, we found that sequence conservation for this gene in the archaea was much higher than for the bacteria, leading to our decision to restrict our analysis to archaeal members only. Reconciliation and rooting analysis generated an evolutionary history in which the root position lay between *M. kandleri* and all other species in the study, with the habitat of the root likely to be a hydrothermal vent. This correlates with studies that propose a hyperthermophilic last common ancestor (LUCA). We chose to model the *M.kandleri* taxon, since it was one of four taxa that were unlikely to have changed habitat (hydrothermal vent) since diverging from the root. Given the phylogenetic position, one might expect this model to most closely resemble its structurally solved DUF-62 gene from its relative *P.horikoshii*. In fact, the homology model indicates that structurally and functionally *M.kandleri* resembles homologous members in the bacteria with fluorinase ligand binding and function. By the use of gene context analysis, we observed a pattern of discrete phylogenetic groupings of predicted interactions. Despite the difference in these metabolic pathways, all have connections with membrane composition and therefore may constitute a pathway to regulate lipid composition in extreme environments. We used the program DIVERGE to test for differences in the rate of evolution between these phylogenetic groupings, and noted the biggest difference between the Hsl and Hv_2 clades. After adding extra sequences to these clades of interest, we reconciled the gene tree with the species tree and noted that this difference in the rate of evolution coincides with a major transfer event far back in the history of this superfamily which appeared to divide the Hsl and Hv_2 groups.

The main challenge in studying the DUF-62 superfamily is the lack of data available. Our work pushes new frontiers in this way, since many studies focus on families that have more members structurally and functionally characterised. We found that reconciliation of our archaeal gene tree with the species tree generated different evolutionary histories and estimation of the root, depending on the parameters used. This may indicate the low support of some bifurcations, especially since the root inferred when we allowed Notung to rearrange or not changed consistently. We determined that allowing Notung to rearrange branches with less than 50% support only struck the balance of using a parsimonious strategy, whilst allowing evidence from our gene tree and alignment to overrule if well supported. Our inference of root position is therefore constrained by the topology of the underlying species tree and our prediction of divergence between clades is limited to those with 50% bootstrap support or more. As more structures and functions are solved for this superfamily the resolution of phylogenies will improve, for example, by the use of structurally informed alignments.

Our study sheds light on the evolution of the DUF-62 gene only in the archaeal domain of life. In fact, this gene is well distributed throughout the prokaryotes, and is found in many species of bacteria, including pathogenic ones. Although we have made steps to constructively add information from additional sequences

in this family to determine duplication, loss and transfer events, we were restricted by adding only sequences of interest. As more structures are solved for proteins in this superfamily, our power to resolve the evolutionary signal between distant family members will increase. This will be of particular interest in determining the relationships between the archaea and other members of the DUF-62 family.

Assessing the significance of multiple independent tests on the same data is a challenge. In this study, we have explored the idea of attempting to correct our results using both the Bonferroni and FDR methods. We concluded that attempting to correct for multiple tests when using likelihood ratio statistics was difficult since they do not map directly to p-values. We then mapped residues which had the greatest degree of divergence between the Hsl and and Hv_2 clades onto the only archaeal structure with functional annotations (PDB:2WR8). We should bear in mind that the definition of 'environment' was inherently broad, particularly in our grouping of all 'hydrothermal vent' habitats into one category, when in fact this genre is likely to contain many more diverse sub-environments. Future studies might examine any differences in rates of evolution between well supported clades of the same environment. Although useful in determining the spatial position of these residues in relation to the SAM hydrolase substrate (*S*-homocysteine used as an inhibitor in this case) we cannot be sure that all of our DUF-62 representatives have the exact same overall and active site structure as the *Pyrococcus horikoshii* structure (PDB:2WR8). Despite this, a common observation in the comparison of our DUF-62 alignment at the sequence level and at the structural level indicates that divergence of function in this family is likely to reside near the substrate binding site, rather than within sites involved in catalysis. This is in common with other studies in which it was found that the evolution of new function often proceeds *via* the divergence in the residues involved in substrate or product binding, not by changes in those that are involved in the catalytic mechanism.

10. Methodological questions and suggestions for future directions

"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is"

Douglas Adams, 1979

10.1. Can we build an accurate phylogeny of the evolution of function in enzyme superfamilies?

Despite our structurally informed approach when building phylogenies, bifurcations are not always supported well statistically. This may be symptomatic of the fact that only small sections of the alignment are conserved (for example near the active site) whereas other parts of the alignment have undergone much change, possibly as a result of high mutation rates in evolutionary less constrained areas.

One way to circumvent this issue might be to only use parts of the alignment which are well conserved. However, this narrows down to an extremely small dataset and may miss important patterns of variation diagnostic of the evolution of new functions. Using an ML strategy and accounting for rate heterogeneity and invariant sites is one way in which we attempted to model the different categories of rate evolution in our alignments. Further partitioning of the alignment may be an interesting and appropriate path to follow in the future modelling of highly divergent superfamilies.

We have used the bootstrap procedure because it is conceptually tangible and has proved tractable for use in further investigations. However, statistically, the meaning of the bootstrap value is disputed. In this thesis, we have used a working

definition of the bootstrap value as a measure of how well the given alignment supports the topology, or, how robust a given topology is to the underlying data.

A high degree of robustness, or precision, is not always indicative of accuracy. For the case of enzyme superfamilies, we do not know the 'true' history of evolution in a given superfamily, so a true measure of accuracy is difficult to obtain. For example, using alignment trimming and masking, it might be possible to filter a 'messy' superfamily alignment to a set of columns with high conservation, resulting in a tree with higher bootstrap values but not necessarily accurate in its depiction of the true topology. Although alternative measures of statistical support are available, these can only ever be a measure of precision, rather than accuracy for any given topology.

What does this mean for the future construction of phylogenetic trees of the genes of highly divergent superfamilies? With an absence of knowledge of a 'true' tree topology at hand, and no fossil records to limit our hypotheses, we must reframe our question - rather than asking how 'accurate' our phylogeny is - we can ask how 'rational' our phylogeny is (inspired by online discussion, see [346]).

In some ways, this relates to our discussion of whether to use a Bayesian approach in the construction of phylogenies. The source of rational evaluation of a given phylogeny should come from a structural and functional perspective. Structure is more conserved in evolution than a protein's primary amino acid sequence, and only those proteins that are functional persist long-term in evolution. At present, no method constrains every node in a phylogenetic tree to model a functional protein, or even for the modelled protein to be structurally viable and foldable. One way forward might be to conduct ML analyses to allow for a large and diverse sample of possible tree space to be explored before evaluating each node to see if the ancestral protein structure and sequence is viable. Only those topologies that represent trajectories in evolution that satisfy these criteria should be considered as rational, if not accurate, hypotheses of the evolution of an enzyme superfamily.

10.2. How can we accurately diagnose function?

The above suggestion is at the moment, intractable, at least in part due to the computational expense needed to calculate the 'structure and function' of all given nodes across a sample of phylogenies. Moreover, such an assertion yields another question - 'How to diagnose function from a enzyme's primary sequence?', or in this particular scenario, from a *prediction* of the state of some ancestral node?

How do we define enzyme function? Throughout this thesis, the working definition is that function is an enzyme's capability to successfully catalyse a given reaction and therefore fulfil its cellular role. When one has the means to synthesise

ancestral sequences in the laboratory, many questions under the umbrella term 'functional' can be answered: Does the protein fold correctly? Can it turn over the expected substrate at a reasonable rate? Are there any promiscuous activities? What other proteins does the enzyme interact with? However, for experiments in which many alternative scenarios of evolution are sampled, wet-laboratory essays become impractical.

There are many tools for which the experimenter can query protein function *in silico,* these tend to be based on empirical data, rather than knowledge from first principles. For *in silico* queries of a mystery protein's function, the more you know about the homologs of the protein the more informed you are to make a prediction of its function. In this thesis, we have compared two case studies, the metallo-β-lactamases, for which many structural and functional representatives were available, against the DUF-62 superfamily, for which very few structurally and functionally solved representatives exist. The difference in the information available for the two superfamilies was reflected in the granularity of the questions we were able to ask. For the metallo-β-lactamases, we were able to ask a very precise question about function, since, due to the wealth of structural and functional information available, we were able to distinguish between B1 and B3 types of lactamase function. For the DUF-62 superfamily however, the lack of available functional information meant that we took a much broader stance, ascertaining that there was some evidence towards the existence of multiple functions and/or cellular roles in the superfamily, beyond those already characterised.

Even when many structural and functionally characterised members are available for a family, our ability to diagnose function is still limited. In this work, we make use of 'catalytic templates' as the minimal machinery to define function. We work from the rationale, that rather than the presence of particular residues being indicative of an ability to perform a particular catalytic reaction, these residues form the *minimal* set of machinery a given protein must have to perform a particular reaction. Much like the discussion about phylogenetic accuracy, assessing enzyme function accurately *in silico* is not possible, because without a wet-laboratory experiment we have no idea if all of the many variables needed for successful catalysis are present in a given amino acid sequence.

Perhaps the closest we can get to accurate estimate of protein function is by something such as docking or simulation, where the process of binding and of catalysis can be modelled explicitly, whilst incorporating information on dynamics and kinetics. However, increased accuracy comes at a price in terms of time and cost. It is therefore important to frame the question being asked carefully in the study of enzyme superfamilies. The results presented in this thesis use the information we have at hand to highlight, and to some extent quantify, the likelihood of more rational trajectories in the evolution of the metallo-β-lactamase and the DUF-62 genes.

A. Supporting data for analyses of the metallo-β-lactamase superfamily

Supplementary data can be found on the enclosed CD, under ‘MBL’ and at http://link.springer.com/article/10.1007%2Fs00239-014-9639-7 as described.

Notes taken verbatim from those included with the Supplementary Information of [102].

‘1_GASP’

This folder includes the full bootstrap sample of trees in NEWICK format, all phylogenetic trees and sequence files from running the GASP program in NEWICK and FASTA format respectively and the WAG matrix used in the running of the GASP program.

Input Protein sequence alignment of the superfamily can be found in ‘Bootstrap_sample_trees_and_alignment’.

Bootstrap_sample_trees are unrooted and include two trees where the ingroup isn’t monophyletic (tree #62 & #93).

Tiny/zero length branches were rounded up in ‘r’-

for(i in 1:length(trees)) + {trees[[i]]$edge.length[which(trees[[i]]$edge.length< 0.0001)] =0.0001}

Only monophyletic trees were used in analysis. Resulting trees (GASP_trees_out) should be rooted correctly by GASP.

Corresponding sequence files including sequence predictions can be found in ‘GASP_sequences_out’.

Ancestral sequence predictions are labelled as follows- ‘X Node X (Y,Z,A)’

Where X is the number of the node, Y and Z are numbers of descendant nodes and A is the ancestor node to node X.

‘2_MRCA’

This folder includes the node numbers corresponding to the MRCA of the metallo-β-lactamases in each of the 98 trees output by GASP and the corresponding ancestral sequence predictions.

Node numbers of the MRCA for all metallo-β-lactamases in each of the 98 trees can be found in -’Ancestral_node_numbers’.

The corresponding sequences pulled from GASP output can be found in ‘MRCA_sequences’.

‘3. INTERPROSCAN’

This folder contains the results of the InterPro search of the 98 MRCA sequence predictions.

Gap (-) characters were removed from MRCA sequences before being submitted to InterPro. Results of the analysis are shown in ‘InterProScan_results’, only results which hit signature PR001018 are shown.

‘4. CD-HIT’

This folder contains the results of clustering of sequences at 60% identity in CD-HIT.

The 44 sequences with sequence signature IPR001018 were then clustered at 60% identity using CD-HIT. The results of the clustering are in ‘Cluster_results’. Sequences with annotated with a star are cluster representatives and were submitted to PHYRE2 for homology modelling.

‘5. PHYRE2’

This folder includes all of the 11 MRCA models built in PHYRE2. Also included are the distances of between catalytic residues of aligned MRCA models and templates.

The best model for each sequence submitted is included here as a PDB file and annotated with its sequence identifier. Distances measured between catalytic residues and described in the manuscript can be found in ‘Distances_summary’.

B. Supporting data for analyses of the DUF-62 superfamily

Supplementary data can be found on the enclosed CD, under ‘DUF-62’ as described:

‘1_PFAM_seed_gene_tree_analysis’

‘1_Sequences’

This folder contains the archaeal sequences downloaded from UniProt *via* the accession codes for the PFAM seed for ‘SAM_adeno_trans (PF01887)’.

1) Seed sequences were downloaded from Pfam.

2) The file was delimited in excel and accession numbers extracted.

3) This list of accession numbers was used to query UniProt.

4) The results were filtered to include archaeal sequences only and downloaded.

‘2_Sequence_Alignment’

This folder contains the output alignment by Expresso in DND, FASTA and PHYLIP formats. The folder also contains log, scoring and template information.

‘3_Modelgenerator’

This folder contains the output file for Modelgenerator on the alignment file generated by Expresso. Four gamma categories were specified and the output details the results for AIC, AIC2, and the BIC criterion.

‘4_PhyML’

This folder contains all PhyML output files including log, statistics and likelihood files for ML and bootstrap trees.

‘5_Renaming_tips_in_R’

This folder contains the conversion table to convert accession numbers to accession_gene name’ format, the function used in ’R’, the resulting tree and a log of alternative names used in order to reconcile species & gene trees successfully.

‘6_Notung_analysis’

This folder contains the input gene tree, and both the input species trees, with the *H.marismortui* taxon in its alternative position in each. Also included is a spreadsheet detailing the results with different parameter changes, it makes references to the original Notung files which can be found in ‘...3_Analysis/Analysis_files’.

‘7_DIVERGE_analysis’

This folder contains the raw output from the DIVERGE analysis, using the 1999 Type I algorithm on clades with $\geqslant$50% support. Also contained are the mapping of predicted divergent residues to PDB structure in a Jalview file and False Discovery Rate (FDR) adjustment calculations.

‘8_STITCH_analysis’

This folder contains raw output from STITCH 4.0, queried with default medium confidence level (0.4). The output data includes predictions for chemical and protein interactions for each taxon and is organised into ‘Hsl’, ‘Hv_1’ and ‘Hv_2’ clades respectively.

‘2_Gene_tree_addional_sequences_analysis’

‘1_Additional_sequences’

This folder contains additional sequences selected using the taxanomy browser in the PFAM. We selected PF01887 sequences from the *Thermoproteaceae* and the *Halobacteria,* the accessions for which can be found in ‘selected_sequence_accessions.txt’. The folder also includes the additional sequences combined with the core seed gene sequence set, with duplicates removed and *P.yayanossi* replacing *P.furiosus*.

‘2_Sequence_alignment’

This folder contains the output alignment by Expresso in DND, FASTA and PHYLIP formats. The folder also contains log, scoring and template information.

‘3_Modelgenerator’

This folder contains the output file for Modelgenerator on the alignment file generated by Expresso. Four gamma categories were specified and the output details the results for AIC, AIC2, and the BIC criterion.

‘4_PhyML’

This folder contains all PhyML output files including log, statistics and likelihood files for ML and bootstrap trees.

‘5_Renaming_tips_in_R’

This folder contains the conversion table to convert accession numbers to ‘accession_gene name’ format, the function used in ’R’, the resulting tree and a log of alternative names used in order to reconcile species & gene trees successfully.

‘6_Notung_analysis’

This folder contains the input gene tree, and both the input species trees, with the *H.marismortui* taxon in its alternative position in each. Also included is a spreadsheet detailing the results with different parameter changes, it makes references to the original Notung files which can be found in ‘...3_Analysis/Analysis_files’.

‘3_Whole_PFAM_seed_tree’

This folder contains the sequence alignment and resulting ML tree from the whole PFAM seed set.

The seed alignment for family 01887 was downloaded from PFAM on 12.08.14. Since this dataset consisted of a high number of sequences RaxML was used to build phylogeny specifying LG as the substitution model as selected by MODELGENERATOR AIC & BIC criteria with 4 discrete gamma categories, without

the use empirical base frequencies and letting RaxML estimate the proportion of invariant sites.

It should be noted that some tips in the tree are labelled ‘NA’, as the seed accessions from Pfam could not be found in UniprotKB to associate with full species names at the the time of analysis.

Bibliography

[1] Darwin, C. (1859) On the Origin of Species by Means of Natural Selection. *Murray. London*

[2] Sarich, V. M., and Wilson, A. C. (1967) Immunological time scale for hominid evolution. *Science (New York, N.Y.) 158*, 1200–1203.

[3] Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2008) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics 23*, 1–25.

[4] Mudgal, R., Sandhya, S., Chandra, N., and Srinivasan, N. (2015) De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. *Biology Direct 10*, 38.

[5] Radivojac, P. et al. (2013) A large-scale evaluation of computational protein function prediction. *Nature Methods 10*, 221–7.

[6] Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. (2012) Divergent evolution in enolase superfamily: Strategies for assigning functions. *Journal of Biological Chemistry 287*, 29–34.

[7] Alderson, R. G., Ferrari, L. D., Mavridis, L., McDonagh, J. L., Mitchell, J. B. O., and Nath, N. (2012) Enzyme informatics. *Current Topics in Medicinal Chemistry 12*, 1911–1923.

[8] Jacobson, M. P., Kalyanaraman, C., Zhao, S., and Tian, B. (2014) Leveraging structure for enzyme function prediction: Methods, opportunities, and challenges. *Trends in Biochemical Sciences 39*, 363–371.

[9] Steffen-Munsberg, F., Vickers, C., Kohls, H., Land, H., Mallin, H., Nobili, A., Skalden, L., van den Bergh, T., Joosten, H.-J., Berglund, P., Höhne, M., and Bornscheuer, U. T. (2015) Bioinformatic analysis of a PLP-dependent enzyme superfamily suitable for biocatalytic applications. *Biotechnology Advances 33*, 566–604.

[10] Furnham, N., Garavelli, J. S., Apweiler, R., and Thornton, J. M. (2009) Missing in action: enzyme functional annotations in biological databases. *Nature Chemical Biology 5*, 521–525.

[11] Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology 5*, e1000605.

[12] Kumar, S., Dudley, J. T., Filipski, A., and Liu, L. (2011) Phylomedicine: An evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends in Genetics 27*, 377–386.

[13] Baquero, F., Coque, T. M., and de la Cruz, F. (2011) Eco-Evo Drugs and Strategies: The Need for Novel Tools to Fight Antibiotic Resistance. *Antimicrobial Agents and Chemotherapy 55*, 3649–3660.

[14] Barlow, M., and Hall, B. G. (2003) Experimental prediction of the natural evolution of antibiotic resistance. *Genetics 163*, 1237–41.

[15] Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Systematic zoology 19*, 99–113.

[16] Petsko, G. A. (2001) Homologuephobia. *Genome Biology 2*, comment1002.1.

[17] Koonin, E. V. (2001) An apology for orthologs - or brave new memes. *Genome Biology 2*, Comment1005.1.

[18] Jensen, R. A. (2001) Orthologs and paralogs - we need to get it right. *Genome Biology 2*, interactions1002.1–interactions1002.3.

[19] ENCODE Project Consortium, (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science 306*, 636–640.

[20] Kellis, M. et al. (2014) Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America 111*, 6131–8.

[21] Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013) On the immortality of television sets: ”Function” in the human genome according to the evolution-free gospel of encode. *Genome Biology and Evolution 5*, 578–590.

[22] Doolittle, W. F. (2013) Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences of the United States of America 110*, 5294–300.

[23] Eddy, S. R. (2013) The ENCODE project: Missteps overshadowing a success. *Current Biology 23*, R259–R261.

[24] Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual review of biochemistry 70*, 209–46.

[25] Zhang, C., and DeLisi, C. (1998) Estimating the number of protein folds. *Journal of Molecular Biology 284*, 1301–1305.

[26] Govindarajan, S., Recabarren, R., and Goldstein, R. A. (1999) Estimating the total number of protein folds. *Proteins 35*, 408–414.

[27] Wolf, Y. I., Grishin, N. V., and Koonin, E. V. (2000) Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology 299*, 897–905.

[28] Dryden, D. T. F., Thomson, A. R., and White, J. H. (2008) How much of protein sequence space has been explored by life on Earth? *Journal of the Royal Society, Interface / the Royal Society 5*, 953–6.

[29] Orengo, C., and Jones, D. (1994) Protein superfamilies and domain superfolds. *Nature 372*.

[30] Horowitz, N. (1945) On the evolution of biochemical syntheses. *Proceedings of the National Academy of Sciences 31*.

[31] Babbitt, P. C., and Gerlt, J. A. (1997) Understanding Enzyme Superfamilies. *The Journal of Biological Chemistry 272*, 30591–30594.

[32] Herschlag, D., and O'Brien, P. J. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* R91–R105.

[33] Gerlt, J. A., and Babbitt, P. C. (1998) Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Current Opinion in Chemical Biology 2*, 607–12.

[34] Nei, M. (1969) Gene Duplication and Nucleotide Substitution in Evolution. *Nature 221*, 40–42.

[35] Ohno, S. *Evolution by gene duplication*; Springer-Verlag: New York, 1970; pp pp. xv + 160 pp.

[36] Force, A., Lynch, M., Pickett, B., Amores, A., Yan, Y.-L., and Postlethwait, J. (1999) Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics 151*, 1531–45.

[37] Caetano-Anollés, G., Kim, H. S., and Mittenthal, J. E. (2007) The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences of the United States of America 104*, 9358–63.

[38] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology 307*, 1113–43.

[39] Aravind, L. (1999) An Evolutionary Classification of the Metallo-β-Lactamase Fold Proteins. *In Silico Biology 1*, 69–91.

[40] Wierenga, R. (2001) The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters 492*, 193–198.

[41] Juárez, P., Comas, I., González-Candelas, F., and Calvete, J. J. (2008) Evolution of snake venom disintegrins by positive Darwinian selection. *Molecular Biology and Evolution 25*, 2391–407.

[42] Russell, R. J., Scott, C., Jackson, C. J., Pandey, R., Pandey, G., Taylor, M. C., Coppin, C. W., Liu, J.-W., and Oakeshott, J. G. (2011) The evolution of new enzyme function: lessons from xenobiotic metabolizing bacteria versus insecticide-resistant insects. *Evolutionary Applications 4*, 225–248.

[43] Raushel, F. M. (2002) Bacterial detoxification of organophosphate nerve agents. *Current Opinion in Microbiology 5*, 288–295.

[44] Zhang, J. (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution 18*, 292–298.

[45] Elena, S. F., and Lenski, R. E. (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews. Genetics 4*, 457–469.

[46] Lenski, R. E., and Travisano, M. (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America 91*, 6808–6814.

[47] Khersonsky, O., and Tawfik, D. S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual Review of Biochemistry 79*, 471–505.

[48] Copley, S. D. (2012) Toward a Systems Biology Perspective on Enzyme Evolution. *The Journal of Biological Chemistry 287*, 3–10.

[49] Carter, C. W. (2014) Urzymology: Experimental Access to a Key Transition in the Appearance of Enzymes. *The Journal of Biological Chemistry* 30213–30220.

[50] Wolfenden, R. (2014) Massive Thermal Acceleration of the Emergence of Primordial Chemistry, the Evolution of Enzymes, and the Tempo of Spontaneous Mutation. *The Journal of Biological Chemistry*

[51] Hobbs, J. K., Shepherd, C., Saul, D. J., Demetras, N. J., Haaning, S., Monk, C. R., Daniel, R. M., and Arcus, V. L. (2012) On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus. *Molecular Biology and Evolution 29*, 825–35.

[52] Butzin, N. C., Lapierre, P., Green, A. G., Swithers, K. S., Gogarten, J. P., and Noll, K. M. (2013) Reconstructed ancestral Myo-inositol-3-phosphate synthases indicate that ancestors of the Thermococcales and Thermotoga species were more thermophilic than their descendants. *PloS One 8*, e84300.

[53] Akanuma, S., Nakajima, Y., Yokobori, S.-i., Kimura, M., Nemoto, N., Mase, T., Miyazono, K.-i., Tanokura, M., and Yamagishi, A. (2013) Experimental evidence for the thermophilicity of ancestral life. *Proceedings of the*

National Academy of Sciences of the United States of America 110, 11067–72.

[54] Nasir, A., Kim, K., and Caetano-Anollés, G. (2014) A Phylogenomic Census of Molecular Functions Identifies Modern Thermophilic Archaea as the Most Ancient Form of Cellular Life. *Archaea 2014*, 1–26.

[55] Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. *Current Opinion in Chemical Biology 10*, 498–508.

[56] Voordeckers, K., Brown, C. A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., and Verstrepen, K. J. (2012) Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biology 10*, e1001446.

[57] Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *Journal of the American Chemical Society 135*, 2899–902.

[58] Copley, S. D. (2012) Moonlighting is mainstream: Paradigm adjustment required. *BioEssays : news and reviews in molecular, cellular and developmental biology 34*, 578–88.

[59] Pandya, C., Farelli, J. D., Dunaway-Mariano, D., and Allen, K. N. (2014) Enzyme Promiscuity: Engine of Evolutionary Innovation. *The Journal of Biological Chemistry*

[60] Brown, R. L. (2014) What evolvability really is. *British Journal for the Philosophy of Science 65*, 549–572.

[61] Tomatis, P. E., Fabiane, S. M., Simona, F., Carloni, P., Sutton, B. J., and Vila, A. J. (2008) Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *Proceedings of the National Academy of Sciences of the United States of America 105*, 20605–10.

[62] Baier, F., and Tokuriki, N. (2014) Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily. *Journal of Molecular Biology 426*, 2442–56.

[63] Pál, C., Papp, B., and Lercher, M. J. (2006) An integrated view of protein evolution. *Nature Reviews. Genetics 7*, 337–48.

[64] Bustamante, C. D., Townsend, J. P., and Hartl, D. L. (2000) Solvent Accessibility and Purifying Selection Within Proteins of Escherichia coli and Salmonella enterica. *Molecular Biology and Evolution 17*, 301–308.

[65] Das, A. D., and Misra, H. S. (2013) Hypothetical proteins present during recovery phase of radiation resistant bacterium Deinococcus radiodurans are under purifying selection. *Journal of Molecular Evolution 77*, 31–42.

[66] Nei, M. (2005) Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution 22*, 2318–42.

[67] Wagner, A. (2005) Robustness, evolvability, and neutrality. *FEBS Letters 579*, 1772–8.

[68] Piatigorsky, J., O'Brien, W. E., Norman, B. L., Kalumuck, K., Wistow, G. J., Borras, T., Nickerson, J. M., and Wawrousek, E. F. (1988) Gene sharing by delta-crystallin and argininosuccinate lyase. *Proceedings of the National Academy of Sciences of the United States of America 85*, 3479–3483.

[69] Ellegren, H., Smith, N. G., and Webster, M. T. (2003) Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development 13*, 562–568.

[70] Rattray, A. J., and Strathern, J. N. (2003) Error-prone DNA polymerases: when making a mistake is the only way to get ahead. *Annual review of genetics 37*, 31–66.

[71] Smith, J. M., and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genetical Research 23*, 23–35.

[72] Birky, C. W., and Walsh, J. B. (1988) Effects of linkage on rates of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America 85*, 6414–6418.

[73] Yanai, I., Derti, A., and DeLisi, C. (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America 98*, 7940–5.

[74] Elleuche, S. (2014) Bringing functions together with fusion enzymes—from nature's inventions to biotechnological applications. *Applied Microbiology and Biotechnology*

[75] Hyde, C., Ahmed, S., and Padlan, E. (1988) Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from Salmonella typhimurium. *Journal of Biological Chemistry 263*, 17857–71.

[76] Thoden, J., Holden, H., and Wesenberg, G. (1997) Structure of carbamoyl phosphate synthetase: a journey of 96 Åfrom substrate to product. *Biochemistry 2960*, 6305–6316.

[77] Sanyal, N., Arentson, B. W., Luo, M., Tanner, J. J., and Becker, D. F. (2014) First Evidence for Substrate Channeling Between Proline Catabolic Enzymes: A Validation of Domain Fusion Analysis for Predicting Protein-Protein Interactions. *The Journal of Biological Chemistry 290*, 2225–2234.

[78] Vicente, J. A. B., Gomes, C. M., Wasserfallen, A., and Teixeira, M. (2002) Module fusion in an A-type flavoprotein from the cyanobacterium Synechocystis condenses a multiple-component pathway in a single polypeptide chain. *Biochemical and biophysical research communications 294*, 82–7.

[79] Jensen, R. A. (1976) Enzyme recruitment in evolution of new function. *Annual Review of Microbiology 30*, 409–25.

[80] Tokuriki, N., and Tawfik, D. S. (2009) Protein Dynamism and Evolvability. *Science 324*, 203–207.

[81] Anantharaman, V., Aravind, L., and Koonin, E. V. (2003) Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Current Opinion in Chemical Biology 7*, 12–20.

[82] Perona, J. J., and Craik, C. S. (1997) Evolutionary Divergence of Substrate Specificity within the Chymotrypsin-like Serine Protease Fold. *Journal of Biological Chemistry 272*, 29987–29990.

[83] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002) Plasticity of enzyme active sites. *Trends in Biochemical Sciences 27*, 419–26.

[84] Makarova, K. S., and Grishin, N. V. (1999) The Zn-peptidase superfamily: functional convergence after evolutionary divergence. *Journal of Molecular Biology 292*, 11–7.

[85] Huson, D. H., and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution 23*, 254–267.

[86] M, A., and D, P. (2010) The Effect of Recombination on the Reconstruction of Ancestral Sequences. *Genetics 184*, 1133–1139.

[87] Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *The EMBO Journal 5*, 823–826.

[88] Illergå rd, K., Ardell, D. H., and Elofsson, A. (2009) Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics 77*, 499–508.

[89] Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research 43*, D376–D381.

[90] Buschbom, J., and Barker, D. (2006) Evolutionary history of vegetative reproduction in Porpidia s.L. (Lichen-forming ascomycota). *Systematic Biology 55*, 471–84.

[91] The Uniprot Consortium, (2014) UniProt: a hub for protein information. *Nucleic Acids Research 43*, 204–212.

[92] Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., and Tolstoy, I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Research 42*, D553–9.

[93] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology 247*, 536–40.

[94] Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., and Holm, L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Research 29*, 55–57.

[95] Hadley, C., and Jones, D. T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure 7*, 1099–1112.

[96] Day, R., Beck, D. A. C., Armen, R. S., and Daggett, V. (2003) A consensus view of fold space : Combining SCOP , CATH , and the Dali Domain Dictionary. *Protein Science 12*, 2150–2160.

[97] Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research 35*, D291–7.

[98] Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., and Pearson, W. R. (2012) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Research 40*, D783–D789.

[99] Nagano, N. (2005) EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Research 33*, D407–12.

[100] Akiva, E. et al. (2014) The Structure-Function Linkage Database. *Nucleic Acids Research 42*, D521–30.

[101] Nath, N., Mitchell, J. B. O., and Caetano-Anollés, G. (2014) The natural history of biocatalytic mechanisms. *PLoS Computational Biology 10*, e1003642.

[102] Alderson, R. G., Barker, D., and Mitchell, J. B. O. (2014) One origin for metallo-β-lactamase activity, or two? An investigation assessing a diverse set of reconstructed ancestral sequences based on a sample of phylogenetic trees. *Journal of Molecular Evolution 79*, 117–29.

[103] Furnham, N., Holliday, G. L., de Beer, T. a. P., Jacobsen, J. O. B., Pearson, W. R., and Thornton, J. M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research 42*, D485–9.

[104] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology 112*, 535–542.

[105] Meng, E. C., Polacco, B. J., and Babbitt, P. C. (2004) Superfamily active site templates. *Proteins 55*, 962–976.

[106] Torrance, J. W., Bartlett, G. J., Porter, C. T., and Thornton, J. M. (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *Journal of Molecular Biology 347*, 565–581.

[107] Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Rahman, S. A., Laskowski, R. A., Orengo, C. A., and Thornton, J. M. (2012) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Research 40*, D776–82.

[108] Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Laskowski, R. A., Orengo, C. A., and Thornton, J. M. (2012) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Computational Biology 8*, e1002403.

[109] Chang, A., Schomburg, I., Placzek, S., Jeske, L., Ulbrich, M., Xiao, M., Sensen, C. W., and Schomburg, D. (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Research 43*, D439–46.

[110] Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research 28*, 27–30.

[111] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research 41*, D808–15.

[112] Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E.coli. *Nucleic Acids Research 10*, 2997–3912.

[113] Yoon, B.-J. (2009) Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current genomics 10*, 402–415.

[114] Sigrist, C., Cerutti, L., and Hulo, N. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics 3*, 265–274.

[115] Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Research 41*, D344–D347.

[116] Mitchell, A. et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research 43*, D213–D221.

[117] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014) Pfam: the protein families database. *Nucleic Acids Research 42*, D222–30.

[118] Henikoff, S., and Henikoff, J. G. (1994) Position-based sequence weights. *Journal of Molecular Biology 243*, 574–578.

[119] Krogh, A. *Computational Methods in Molecular Biology*; Elsevier, 1998; Chapter Chapter 4, pp 45–63.

[120] Zvelebil, M., and Baum, J. O. *Understanding Bioinformatics*; Garland Science, Taylor & Francis Group, LLC, 2008.

[121] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research 25*, 3389–402.

[122] Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research 39*, W29–W37.

[123] Schuster-Böckler, B., Schultz, J., and Rahmann, S. (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics 8*, 1–8.

[124] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology 7*, 539.

[125] Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology 302*, 205–217.

[126] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research 22*, 4673–4680.

[127] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*, 1792–7.

[128] Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research 30*, 3059–66.

[129] Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research 33*, 511– 518.

[130] Katoh, K., and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics 9*, 286–298.

[131] Katoh, K., and Frith, M. C. (2012) Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics (Oxford, England) 28*, 3144–3146.

[132] Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution 30*, 772–80.

[133] Henikoff, S., and Henikoff, J. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences 89*, 10915–10919.

[134] Dayhoff, M. O., and Schwartz, R. M. *Atlas of Protein Sequence and Structure*; 1978; Chapter 22.

[135] Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M., and Morgenstern, B. (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics 6*, 66.

[136] Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Research 34*, W604–8.

[137] Lynch, M. (2010) Evolution of the mutation rate. *Trends in Genetics : TIG 26*, 345–52.

[138] Jukes, T., and Cantor, C. *Evolution of Protein Molecules*; Academic Press: New York., 1969; Chapter Chapter 2, pp 21–132.

[139] Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution 16*, 111–120.

[140] Tamura, K., and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution 10*, 512 –526.

[141] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics 8*, 275–282.

[142] Yang, Z. *Computational Molecular Evolution*; Oxford University Press, 2006.

[143] Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution 18*, 691–699.

[144] Le, S. Q., and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Molecular Biology and Evolution 25*, 1307–20.

[145] Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactionson Automatic Control 19*, 716–723.

[146] Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

[147] Tsai, C.-l. (2014) Regression and time series model selection in small samples. *76*, 297–307.

[148] Hirt, R., and Logsdon, J. (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences 96*, 580–585.

[149] Lockhart, P., and Larkum, A. (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proceedings of the National Academy of Sciences 93*, 1930–1934.

[150] Gu, X., Fu, Y. X., and Li, W. H. (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution 12*, 546–557.

[151] Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4*, 406–425.

[152] Sokal, R. R., and Michener, C. D. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin 28*, 1409–1438.

[153] Fitch, W., and Margoliash, E. (1967) Construction of Phylogenetic Trees. *Science 155*, 279–284.

[154] Felsenstein, J. (1978) The Number of Evolutionary Trees. *Systematic Zoology 27*, 27–33.

[155] Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology 59*, 307–21.

[156] Fitch, W. M. (1971) Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology 20*, 406–416.

[157] Barker, D. (2015) Seeing the wood for the trees: philosophical aspects of classical, Bayesian and likelihood approaches in statistical inference and some implications for phylogenetic analysis. *Biology & Philosophy 30*, 505–525.

[158] Yang, Z., and Rannala, B. (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology 54*, 455–70.

[159] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*, 368–376.

[160] Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution 39*, 783–791.

[161] Efron, B., Halloran, E., and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences 93*, 7085–7090.

[162] Efron, B. (1979) Bootstrap methods: another look at the jackknife. *The Annals of Statistics 7*, 1–26.

[163] Felsenstein, J., and Kishino, H. (1993) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology 42*, 193–200.

[164] Berry, V., and Gascuel, O. (1996) On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Molecular Biology and Evolution* 999–1011.

[165] Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology 27*, 401–410.

[166] Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005) Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology 5*, 50.

[167] Bergsten, J. (2005) A review of long-branch attraction. *Cladistics 21*, 163–193.

[168] Hendy, M., and Penny, D. (1989) A framework for the quantitative study of evolutionary trees. *Systematic Zoology 38*, 297–309.

[169] Pauling, L., Zuckerkandl, E., Henriksen, T., and Lövstad, R. (1963) Chemical Paleogenetics. Molecular ”Restoration Studies” of Extinct Forms of Life. *Acta Chemica Scandinavica 17 supl.*, 9–16.

[170] Thornton, J. W. (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews. Genetics 5*, 366–375.

[171] Huelsenbeck, J. P., and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics 17*, 754–755.

[172] Williams, P. D., Pollock, D. D., Blackburne, B. P., and Goldstein, R. a. (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Computational biology 2*, e69.

[173] Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., and Pupko, T. (2012) FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research 40*, W580–W584.

[174] Edwards, R. J., and Shields, D. C. (2004) GASP: Gapped Ancestral Sequence Prediction for proteins. *BMC Bioinformatics 5*.

[175] Yang, Z., Kumar, S., and Nei, M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics 141*, 1641–1650.

[176] Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. W. (2010) Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution 27*, 1988–99.

[177] Kauzmann, W. *Advances in Protein Chemistry*; Advances in Protein Chemistry; Elsevier, 1959; Vol. 14; pp 1–63.

[178] Todd, A. E., Orengo, C. A., and Thornton, J. M. (1999) Evolution of protein function, from a structural perspective. *Current Opinion in Chemical Biology 3*, 548–56.

[179] Park, H., Brothers, E. N., and Merz, K. M. (2005) Hybrid QM/MM and DFT investigations of the catalytic mechanism and inhibition of the dinuclear zinc metallo-beta-lactamase CcrA from Bacteroides fragilis. *Journal of the American Chemical Society 127*, 4232–41.

[180] Kelm, S., Vangone, A., Choi, Y., Ebejer, J. P., Shi, J., and Deane, C. M. (2014) Fragment-based modeling of membrane protein loops: Successes, failures, and prospects for the future. *Proteins: Structure, Function and Bioinformatics 82*, 175–186.

[181] Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics 9*.

[182] Kelley, L. A., and Sternberg, M. J. E. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols 4*, 363–371.

[183] Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics 23*, ii–iv.

[184] Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K. J., Kelley, L. a., MacCallum, R. M., Pawowski, K., Rost, B., Rychlewski, L., and Sternberg, M. (1999) CAFASP-1: Critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function, and Genetics 37*, 209–217.

[185] Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics 28*, i409–i415.

[186] Danicic, D., Durand, D., Goldman, A., Lai, H., Stolzer, M., Vernot, B., and Xu, M. Notung 2.8 : A Manual. 2014.

[187] Gogarten, J. P., and Townsend, J. P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nature Reviews. Microbiology 3*, 679–87.

[188] Nakhleh, L., Ruths, D., and Wang, L.-S. In *Computing and Combinatorics SE - 11*; Wang, L., Ed.; Lecture Notes in Computer Science; Springer Berlin Heidelberg, 2005; Vol. 3595; pp 84–93.

[189] Beiko, R., and Hamilton, N. (2006) Phylogenetic identification of lateral genetic transfer events. *BMC Evolutionary Biology 17*, 1–17.

[190] Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics 11*, 324.

[191] Landgraf, R., Fischer, D., and Eisenberg, D. (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Engineering 12*, 943–51.

[192] Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution 24*, 1586–91.

[193] Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology 257*, 342–58.

[194] Gu, X., and Vander Velden, K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics (Oxford, England) 18*, 500–1.

[195] Gu, X., Zou, Y., Su, Z., Huang, W., Zhou, Z., Arendsee, Z., and Zeng, Y. (2013) An update of DIVERGE software for functional divergence analysis of protein family. *Molecular Biology and Evolution 30*, 1713–9.

[196] Lichtarge, O., and Sowa, M. E. (2002) Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology 12*, 21–7.

[197] Gu, X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular Biology and Evolution 18*, 453–64.

[198] Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution 16*, 1664–74.

[199] Smith, R. A., M'ikanatha, N. M., and Read, A. F. (2014) Antibiotic Resistance: A Primer and Call to Action. *Health communication* 1–6.

[200] D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., and Wright, G. D. (2011) Antibiotic resistance is ancient. *Nature 477*, 457–61.

[201] Coulson, A. (1985) ß-Lactamases: Molecular Studies. *Biotechnology and Genetic Engineering Reviews 3*, 219–254.

[202] Hall, B. G., Salipante, S. J., and Barlow, M. (2004) Independent origins of subgroup Bl + B2 and subgroup B3 metallo-beta-lactamases. *Journal of molecular evolution 59*, 133–41.

[203] Oelschlaeger, P. (2008) Outsmarting metallo-β-lactamases by mimicking their natural evolution. *Journal of Inorganic Biochemistry 102*, 2043–51.

[204] Liénard, B. M. R., Garau, G., Horsfall, L., Karsisiotis, A. I., Damblon, C., Lassaux, P., Papamicael, C., Roberts, G. C. K., Galleni, M., Dideberg, O., Frère, J.-M., and Schofield, C. J. (2008) Structural basis for the broad-spectrum inhibition of metallo-beta-lactamases by thiols. *Organic & Biomolecular Chemistry 6*, 2282–2294.

[205] Bebrone, C. (2007) Metallo-beta-lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily. *Biochemical pharmacology 74*, 1686–701.

[206] Bebrone, C., Lassaux, P., Vercheval, L., Sohier, J.-S., Jehaes, A., Sauvage, E., and Galleni, M. (2010) Current challenges in antimicrobial chemotherapy: focus on ß-lactamase inhibition. *Drugs 70*, 651–79.

[207] García-Saez, I., Hopkins, J., Papamicael, C., Franceschini, N., Amicosante, G., Rossolini, G. M., Galleni, M., Frère, J.-M., and Dideberg, O. (2003) The 1.5-A structure of Chryseobacterium meningosepticum zinc beta-lactamase in complex with the inhibitor, D-captopril. *The Journal of Biological Chemistry 278*, 23868–73.

[208] Jmol: an open-source Java viewer for chemical structures in 3D. Http://www.jmol.org/.

[209] Walsh, T. R., Toleman, M. A., Poirel, L., and Nordmann, P. (2005) Metallo-β-Lactamases: the Quiet before the Storm? *Clinical Microbiology Reviews 18*, 306–325.

[210] Saavedra, M. J., Peixe, L., Sousa, J. a. C., Henriques, I., Alves, A., and Correia, A. (2003) Sfh-I, a subclass B2 metallo-β-lactamase from a Serratia fonticola environmental isolate. *Antimicrobial Agents and Chemotherapy 47*, 2330–2333.

[211] Rossolini, G. M., Condemi, M. A., Pantanella, F., Docquier, J. D., Amicosante, G., and Thaller, M. C. (2001) Metallo-β-Lactamase Producers in Environmental Microbiota: New Molecular Class B Enzyme in Janthinobacterium lividum. *Antimicrobial Agents and Chemotherapy 45*, 837–844.

[212] Simm, A. M., Higgins, C. S., Pullan, S. T., Avison, M. B., Niumsup, P., Erdozain, O., Bennett, P. M., and Walsh, T. R. (2001) A novel metallo-beta-lactamase, Mbl1b, produced by the environmental bacterium Caulobacter crescentus. *FEBS Letters 509*, 350–354.

[213] Stoczko, M., Frère, J. M., Rossolini, G. M., and Docquier, J. D. (2006) Postgenomic scan of metallo-β-lactamase homologues in rhizobacteria: Identification and characterization of BJP-1, a subclass B3 ortholog from Bradyrhizobium japonicum. *Antimicrobial Agents and Chemotherapy 50*, 1973–1981.

[214] Hall, B. G. (2004) Predicting the evolution of antibiotic resistance genes. *Nature Reviews Microbiology 2*, 430–435.

[215] Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., and Fitch, W. M. (1999) Predicting the Evolution of Human Influenza A. *Science 286*, 1921–1925.

[216] Plotkin, J. B., Dushoff, J., and Levin, S. A. (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proceedings of the National Academy of Sciences of the United States of America 99*, 6263–6268.

[217] Lemey, P., Kosakovsky Pond, S. L., Drummond, A. J., Pybus, O. G., Shapiro, B., Barroso, H., Taveira, N., and Rambaut, A. (2007) Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Computational biology 3*, e29.

[218] Palmer, A. C., and Kishony, R. (2013) Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nature Reviews. Genetics 14*, 243–8.

[219] Garau, G., Guilmi, A. D., and Hall, B. G. (2005) Structure-based phylogeny of the metallo-β-lactamases. *Antimicrobial Agents and Chemotherapy 49*, 2778–2784.

[220] Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M., and Orengo, C. A. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research 41*, D490–D498.

[221] Galleni, M., Lamotte-brasseur, J., Maria, G., Spencer, J., Dideberg, O., Frère, J.-m., Rossolini, G. M., and metallo-β-lactamase working group, T. (2001) Standard Numbering Scheme for Class B β -Lactamases. *Antimicrobial Agents and Chemotherapy 45*, 660–663.

[222] Hall, B. G., Salipante, S. J., and Barlow, M. (2003) The metallo-β-lactamases fall into two distinct phylogenetic groups. *Journal of Molecular Evolution 57*, 249–54.

[223] Wang, Z., Fast, W., Valentine, A., and Benkovic, S. (1999) Metallo-β-lactamase: structure and mechanism. *Current Opinion in Chemical Biology 3*, 614–622.

[224] Spencer, J., Read, J., Sessions, R. B., Howell, S., Blackburn, G. M., and Gamblin, S. J. (2005) Antibiotic recognition by binuclear metallo-β-lactamases revealed by X-ray crystallography. *Journal of the American Chemical Society 127*, 14439–44.

[225] Ullah, J. H., Walsh, T. R., Taylor, I. A., Emery, D. C., Verma, C. S., Gamblin, S. J., and Spencer, J. (1998) The crystal structure of the L1 metallo-β-lactamase from Stenotrophomonas maltophilia at 1.7 A resolution. *Journal of Molecular Biology 284*, 125–36.

[226] Wang, Z., Fast, W., and Benkovic, S. J. (1999) On the mechanism of the metallo-β-lactamase from Bacteroides fragilis. *Biochemistry 38*, 10013–23.

[227] Xu, D., Guo, H., and Cui, Q. (2007) Antibiotic deactivation by a dizinc β-lactamase: mechanistic insights from QM/MM and DFT studies. *Journal of the American Chemical Society 129*, 10814–22.

[228] Hall, B. G., and Barlow, M. (2005) Revised Ambler classification of β-lactamases. *The Journal of Antimicrobial Chemotherapy 55*, 1050–1.

[229] Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M. G., and Orengo, C. A. (2007) CATHEDRAL: A fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Computational Biology 3*, 2333–2347.

[230] Garau, G., Bebrone, C., Anne, C., Galleni, M., Frère, J.-M., and Dideberg, O. (2005) A metallo-beta-lactamase enzyme in action: crystal structures of the monozinc carbapenemase CphA and its complex with biapenem. *Journal of Molecular Biology 345*, 785–95.

[231] Holton, T. A., and Pisani, D. (2010) Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biology and Evolution 2*, 310–24.

[232] Eisen, J. A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development 10*, 606–611.

[233] Garcia-Vallve, S. (2000) Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. *Genome Research 10*, 1719–1725.

[234] Lees, J., Yeats, C., Redfern, O., Clegg, A., and Orengo, C. (2010) Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Research 38*, D296–D300.

[235] Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., and Orengo, C. (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research 40*, D465–D471.

[236] Laskowski, R. A. (2009) PDBsum new things. *Nucleic Acids Research 37*, D355–D359.

[237] Criscuolo, A., and Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology 10*, 210.

[238] Huelsenbeck, J. (1998) Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Systematic Biology 47*, 519–537.

[239] Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., and Mclnerney, J. O. (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology 6*.

[240] The R Development Core Team, R: A language and environment for statistical computing. 2015; http://www.r-project.org/.

[241] Paradis, E., Claude, J., and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics 20*, 289–290.

[242] Wallace Andrew C, T. J. M., Laskowski Roman A (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering 8*, 127–134.

[243] Taylor, W. R. (1997) Residual colours : a proposal for aminochromography. *Protein Engineering 10*, 743–746.

[244] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics 25*, 1189–91.

[245] Gomes, C. M., Frazão, C., Xavier, A. V., Legall, J., and Teixeira, M. (2002) Functional control of the binuclear metal site in the metallo-beta-lactamase-like fold by subtle amino acid replacements. *Protein science : a publication of the Protein Society 11*, 707–712.

[246] Zang, T. M., Hollman, D. A., Crawford, P. A., Crowder, M. W., and Makaroff, C. A. (2001) Arabidopsis glyoxalase II contains a zinc/iron binuclear metal center that is essential for substrate binding and catalysis. *The Journal of Biological Chemistry 276*, 4788–95.

[247] Cameron, A. D., Ridderström, M., Olin, B., and Mannervik, B. (1999) Crystal structure of human glyoxalase II and its complex with a glutathione thiolester substrate analogue. *Structure 7*, 1067–78.

[248] Campos-Bermudez, V., Leite, N., and Krog, R. (2007) Biochemical and structural characterization of Salmonella typhimurium glyoxalase II: new insights into metal ion selectivity. *Biochemistry* 11069–11079.

[249] Marasinghe, G., Sander, I., and Bennett, B. (2005) Structural studies on a mitochondrial glyoxalase II. *Journal of Biological Chemistry 280*, 40668–40675.

[250] Park, H.-S., Nam, S.-H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J., and Kim, H.-S. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. *Science 311*, 535–8.

[251] Llarrull, L. I., Fabiane, S. M., Kowalski, J. M., Bennett, B., Sutton, B. J., and Vila, A. J. (2007) Asp-120 locates Zn2 for optimal metallo-beta-lactamase activity. *The Journal of Biological Chemistry 282*, 18276–85.

[252] González, J. M., Medrano Martín, F. J., Costello, A. L., Tierney, D. L., and Vila, A. J. (2007) The Zn2 Position in Metallo-β-Lactamases is Critical for Activity: A Study on Chimeric Metal Sites on a Conserved Protein Scaffold. *Journal of Molecular Biology 373*, 1141–1156.

[253] Gutteridge, A., and Thornton, J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *Journal of Molecular Biology 346*, 21–28.

[254] Gould, S. J., and Vrba, E. S. (1982) Exaptation-A Missing Term in the Science of Form. *Paleobiology 8*, 4–15.

[255] Maddison, W. P., and Maddison., D. Mesquite: a modular system for evolutionary analysis. 2011; http://mesquiteproject.org.

[256] Bruns, C., Nowalk, A., and Arvai, A. (1997) Structure of Haemophilus influenzae Fe+3-binding protein reveals convergent evolution within a superfamily. *Nature Structural Biology 4*, 919–924.

[257] Burroughs, a. M., Allen, K. N., Dunaway-Mariano, D., and Aravind, L. (2006) Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *Journal of Molecular Biology 361*, 1003–34.

[258] Gherardini, P. F., Wass, M. N., Helmer-Citterich, M., and Sternberg, M. J. E. (2007) Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of Molecular Biology 372*, 817–45.

[259] Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., and Pearson, W. R. (2011) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Research 2005*, 1–7.

[260] Inc, P. I. ChemDraw Professional.

[261] Ashfield, T., Ong, L. E., Nobuta, K., Schneider, C. M., and Innes, R. W. (2004) Convergent evolution of disease resistance gene specificity in two flowering plant families. *The Plant Cell 16*, 309–318.

[262] Latysheva, N., Junker, V. L., Palmer, W. J., Codd, G. A., and Barker, D. (2012) The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics 28*, 603–6.

[263] Lutzoni, F., Pagel, M., and Reeb, V. (2001) Major fungal lineages are derived from lichen symbiotic ancestors. *Nature 411*, 937–40.

[264] Autzen, B. (2011) Constraining prior probabilities of phylogenetic trees. *Biology & Philosophy 26*, 567–581.

[265] Alfaro, M. E. (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution 20*, 255–266.

[266] Charif, D., and Lobry, J. (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *Structural Approaches to Sequence Evolution*

[267] Jones, P. et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics 30*, 1236–1240.

[268] Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics 26*, 680–682.

[269] Wass, M. N., Kelley, L. A., and Sternberg, M. J. E. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research 38*, W469–W473.

[270] Laskowski, R. A., Watson, J. D., and Thornton, J. M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research 33*, W89–W93.

[271] Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research 32*, D129–D133.

[272] Shindyalov, I. N., and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering 11*, 739–747.

[273] Schrodinger LLC, The PyMOL Molecular Graphics System, Version 1.3r1. 2010.

[274] Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic acids symposium series. 1999; pp 95–98.

[275] Weinreich, D., Delaney, N., DePristo, M., and Hartl, D. L. (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science 312*, 111–114.

[276] Klinman, J. P., and Kohen, A. (2014) Evolutionary Aspects of Enzyme Dynamics. *The Journal of Biological Chemistry* 30205–30212.

[277] Menzel, P., Stadler, P. F., and Gorodkin, J. (2011) maxAlike: maximum likelihood-based sequence reconstruction with application to improved primer design for unknown sequences. *Bioinformatics 27*, 317–325.

[278] Yang, Z. (1995) A space-time process model for the evolution of DNA sequences. *Genetics 139*, 993–1005.

[279] Felsenstein, J., and Churchill, G. A. (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*, 93–104.

[280] Pagel, M., and Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology 53*, 571–581.

[281] Fitch, W. M. (1971) Rate of change of concomitantly variable codons. *Journal of Molecular Evolution 1*, 84–96.

[282] Yamamura, A., Ohtsuka, J., Kubota, K., Agari, Y., Ebihara, A., Nakagawa, N., Nagata, K., and Tanokura, M. (2008) Crystal structure of TTHA1429, a novel metallo-beta-lactamase superfamily protein from Thermus thermophilus HB8. *Proteins 73*, 1053–7.

[283] Alfredson, D. A., and Korolik, V. (2007) Identification of putative zinc hydrolase genes of the metallo-β-lactamase superfamily from Campylobacter jejuni. *FEMS immunology and medical microbiology 49*, 159–64.

[284] Shimada, A., Ishikawa, H., Nakagawa, N., Kuramitsu, S., and Masui, R. (2010) The first crystal structure of an archaeal metallo-β-lactamase superfamily protein; ST1585 from Sulfolobus tokodaii. *Proteins 78*, 2399–2402.

[285] Lakner, C., Holder, M. T., Goldman, N., and Naylor, G. J. P. (2011) What's in a likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. *Systematic Biology 60*, 161–74.

[286] Deng, H., and O'Hagan, D. (2008) The fluorinase, the chlorinase and the duf-62 enzymes. *Current Opinion in Chemical Biology 12*, 582–92.

[287] Alessandra S„ Pojer, F., Noel, J. P., and Moore, B. S. (2008) Discovery and characterization of a marine bacterial SAM-dependent chlorinase. *Nature Chemical Biology 4*, 69–74.

[288] Smith, D. R. M., Grüschow, S., and Goss, R. J. M. (2013) Scope and potential of halogenases in biosynthetic applications. *Current Opinion in Chemical Biology 17*, 276–283.

[289] Goss, R. J. M., and Grüschow, S. (2014) Enzymology: A radical finding. *Nature Chemical Biology 10*, 878–879.

[290] Newman, D. J., and Cragg, G. M. (2007) Natural products as sources of new drugs over the last 25 years. *Journal of Natural Products 70*, 461.

[291] Roy, A. D., Grüschow, S., Cairns, N., and Goss, R. J. M. (2010) Gene expression enabling synthetic diversification of natural products: Chemogenetic generation of pacidamycin analogs. *Journal of the American Chemical Society 132*, 12243–12245.

[292] Grüschow, S., Rackham, E. J., and Goss, R. J. M. (2011) Diversity in natural product families is governed by more than enzyme promiscuity alone: establishing control of the pacidamycin portfolio. *Chemical Science 2*, 2182.

[293] Fischbach, M. A., Walsh, C. T., and Clardy, J. (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences of the United States of America 105*, 4601–4608.

[294] Schmidberger, J. W., James, A. B., Edwards, R., Naismith, J. H., and O'Hagan, D. (2010) Halomethane biosynthesis: Structure of a SAM-dependent halide methyltransferase from arabidopsis thaliana. *Angewandte Chemie - International Edition 49*, 3646–3648.

[295] Eustáquio, A. S., Härle, J., Noel, J. P., and Moore, B. S. (2008) S-Adenosyl-L-methionine hydrolase (adenosine-forming), a conserved bacterial and archaeal protein related to SAM-dependent halogenases. *Chembiochem : A European Journal of Chemical Biology 9*, 2215–9.

[296] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014) Pfam: The protein families database. *Nucleic Acids Research 42*, 222–230.

[297] Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics 23*, 2947–2948.

[298] Heath, T. A., Hedtke, S. M., and Hillis, D. M. (2008) Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution 46*, 239–257.

[299] Delsuc, F., Brinkmann, H., and Philippe, H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics 6*, 361–375.

[300] Gaut, B. S., and Lewis, P. O. (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution 12*, 152–162.

[301] Huelsenbeck, J. P. (1995) Performance of Phylogenetic Methods in Simulation. *Systematic Biology 44*, 17–48.

[302] Kuhner, M. K., and Felsenstein, J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution 11*, 459–468.

[303] Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics 30*, 1312–1313.

[304] Gu, X., Zou, Y., Su, Z., Huang, W., Zhou, Z., Arendsee, Z., and Zeng, Y. (2013) An update of DIVERGE software for functional divergence analysis of protein family. *Molecular Biology and Evolution 30*, 1713–9.

[305] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009) GenBank. *Nucleic Acids Research 37*, D26–D31.

[306] Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Research 32*, D35–D40.

[307] Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F. O. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research 42*, D643–8.

[308] Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F. O., and Rosselló-Móra, R. (2010) Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Systematic and Applied Microbiology 33*, 291–9.

[309] Huson, D. H., and Scornavacca, C. (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology 61*, 1061–1067.

[310] Deng, H., McMahon, S. A., Eustáquio, A. S., Moore, B. S., Naismith, J. H., and O'Hagan, D. (2009) Mechanistic insights into water activation in SAM hydroxide adenosyltransferase (duf-62). *Chembiochem : A European Journal of Chemical Biology 10*, 2455–9.

[311] Deng, H., Botting, C. H., Hamilton, J. T. G., Russell, R. J. M., and O'Hagan, D. (2008) S-Adenosyl-L-methionine:Hydroxide Adenosyltransferase: A SAM Enzyme. *Angewandte Chemie 120*, 5437–5441.

[312] Lohman, D. C., Edwards, D. R., and Wolfenden, R. (2013) Catalysis by desolvation: the catalytic prowess of SAM-dependent halide-alkylating enzymes. *Journal of the American Chemical Society 135*, 14473–5.

[313] Deng, H., Ma, L., Bandaranayaka, N., Qin, Z., Mann, G., Kyeremeh, K., Yu, Y., Shepherd, T., Naismith, J. H., and O'Hagan, D. (2014) Identification of fluorinases from Streptomyces sp MA37, Norcardia brasiliensis, and Actinoplanes sp N902-109 by genome mining. *Chembiochem : A European Journal of Chemical Biology 15*, 364–8.

[314] Senn, H. M., O'Hagan, D., and Thiel, W. (2005) Insight into enzymatic C-F bond formation from QM and QM/MM calculations. *Journal of the American Chemical Society 127*, 13643–55.

[315] Kaplun, A., Vyazmensky, M., Zherdev, Y., Belenky, I., Slutzker, A., Mendel, S., Barak, Z., Chipman, D. M., and Shaanan, B. (2006) Structure of the Regulatory Subunit of Acetohydroxyacid Synthase Isozyme III from Escherichia coli. *Journal of Molecular Biology 357*, 951–963.

[316] Gribaldo, S., and Brochier-Armanet, C. (2006) The origin and evolution of Archaea: a state of the art. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 361*, 1007–22.

[317] Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics 25*, 1972–1973.

[318] Burggraf, S., Stetter, K., Rouviere, P., and Woese, C. (1991) Methanopyrus kandleri: An Archaeal Methanogen Unrelated to all Other Known Methanogens. *Systematic and Applied Microbiology 14*, 346–351.

[319] Slesarev, A. I. et al. (2002) The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences of the United States of America 99*, 4644–4649.

[320] Brochier, C., Forterre, P., and Gribaldo, S. (2004) Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. *Genome Biology 5*, R17.

[321] Brochier, C., Forterre, P., and Gribaldo, S. (2005) An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evolutionary Biology 5*, 36.

[322] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015) The I-TASSER Suite: protein structure and function prediction. *Nature Methods 12*, 7–8.

[323] Moult, J., Fidelis, K., Rost, B., Hubbard, T., and Tramontano, A. (2005) Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins 61 Suppl 7*, 3–7.

[324] Slesarev, A. I. et al. (2002) The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences of the United States of America 99*, 4644–9.

[325] Hicks, M. A., Barber, A. E., Giddings, L.-A., Caldwell, J., O'Connor, S. E., and Babbitt, P. C. (2011) The evolution of function in strictosidine synthase-like proteins. *Proteins* 3082–3098.

[326] Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., and Koonin, E. V. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Research 30*, 482–496.

[327] Price, M. N., Arkin, A. P., and Alm, E. J. (2006) The life-cycle of operons. *PLoS Genetics 2*, e96.

[328] Blumenthal, T. (2004) Operons in eukaryotes. *Briefings in Functional Genomics & Proteomics 3*, 199–211.

[329] Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H., Von Mering, C., Jensen, L. J., and Bork, P. (2014) STITCH 4: Integration of protein-chemical interactions with user data. *Nucleic Acids Research 42*, 401–407.

[330] Rentzsch, R., and Orengo, C. A. (2009) Protein function prediction - the power of multiplicity. *Trends in Biotechnology 27*, 210–219.

[331] Massjouni, N., Rivera, C. G., and Murali, T. M. (2006) VIRGO: Computational prediction of gene functions. *Nucleic Acids Research 34*, 340–344.

[332] Cavicchioli, R., Curmi, P. M. G., Saunders, N., and Thomas, T. (2003) Pathogenic archaea: Do they exist? *BioEssays 25*, 1119–1128.

[333] Lichtarge, O., Yao, H., Kristensen, D. M., Madabushi, S., and Mihalek, I. (2003) Accurate and scalable identification of functional sites by evolutionary tracing. *Journal of Structural and Functional Genomics 4*, 159–66.

[334] Mitchell, A. et al. (2014) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research 43*, D213–D221.

[335] Lee, I., and Suzuki, C. K. (2008) Functional mechanics of the ATP-dependent Lon protease- lessons from endogenous protein and synthetic peptide substrates. *Biochimica et Biophysica Acta - Proteins and Proteomics 1784*, 727–735.

[336] Cerletti, M., Martínez, M. J., Giménez, M. I., Sastre, D. E., Paggi, R. a., and De Castro, R. E. (2014) The LonB protease controls membrane lipids composition and is essential for viability in the extremophilic haloarchaeon Haloferax volcanii. *Environmental Microbiology 16*, 1779–1792.

[337] Cerletti, M., Paggi, R. A., Guevara, C. R., Poetsch, A., and De Castro, R. E. (2015) Global role of the membrane protease LonB in Archaea: Potential protease targets revealed by quantitative proteome analysis of a lonB mutant in Haloferax volcanii. *Journal of Proteomics 121*, 1–14.

[338] Albers, S.-V., and Meyer, B. H. (2011) The archaeal cell envelope. *Nature Reviews. Microbiology 9*, 414–426.

[339] Albers, S. V., van de Vossenberg, J. L., Driessen, A. J., and Konings, W. N. (2000) Adaptations of the archaeal cell membrane to heat stress. *Frontiers in Bioscience : A Journal and Virtual Library 5*, D813–D820.

[340] De Rosa, M., and Gambacorta, A. (1988) The lipids of archaebacteria. *Progress in Lipid Research 27*, 153–175.

[341] Perneger, T. V. (1998) What's wrong with Bonferroni adjustments. *BMJ (Clinical research ed.) 316*, 1236–1238.

[342] Barnard G. A.; G. M. Jenkins„ and Winsten, C. B. (1962) Likelihood Inference and Time Series. *Journal of the Royal Statistical Society, Series A 125 (3): 321–372 125*, 35–9238.

[343] Hendrickson, W. A. (1999) Maturation of MAD phasing for the determination of macromolecular structures. *Journal of Synchrotron Radiation 6*, 845–851.

[344] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Research 28*, 235–42.

[345] Rao, K. N., Burley, S. K., and Swaminathan, S. (2008) Crystal structure of a conserved protein of unknown function (MJ1651) from Methanococcus jannaschii. *Proteins: Structure, Function, and Bioinformatics 70*, 572–577.

[346] Tanner, A. R. How can you tell if a phylogenetic tree is accurate? Message posted to:, 2014; http://www.researchgate.net/post/How_can_you_tell_if_a_phylogenetic_tree_is_accurate.