MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Exploring deep phylogenies using protein structure

Ashar J. Malik

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Biochemistry.

Institute of Natural and Mathematical Sciences
Massey University
Auckland, New Zealand
2018

Abstract

Recent times have seen an exponential growth in protein sequence and structure data. The most popular way of characterising newly determined protein sequences is to compare them to well characterised sequences and predict the function of novel sequences based on homology. This practice has been highly successful for a majority of proteins. However, these sequence-based methods struggle with certain deeply diverging proteins and hence cannot always recover evolutionary histories. Another feature of proteins, namely their structures, has been shown to retain evolutionary signals over longer time scales compared to the respective sequences that encode them. The structure therefore presents an opportunity to uncover the evolutionary signal that otherwise escapes conventional sequence-based methods.

Structural phylogenetics refers to the comparison of protein structures to extract evolutionary relationships. The area of structural phylogenetics has been around for a number of years and multiple approaches exist to delineate evolutionary relationships from protein structures. However, once the relationships have been recovered from protein structural data, no methods exist, at present, to verify the robustness of these relationships. Because of the nature of the structural data, conventional sequence-based methods, e.g. bootstrapping, cannot be applied. This work introduces the first ever use of a molecular dynamics (MD)-based bootstrap method, which can add a measure of significance to the relationships inferred from the structure-based analysis.

This work begins in Chapter 2 by thoroughly investigating the use of a protein structural comparison metric Q_{score}, which has previously been used to generate structural phylogenies, and highlights its strengths and weaknesses. The mechanistic exploration of the structural comparison metric reveals a size difference limit of no more than 5-10% in the sizes of protein structures being compared for accurate phylogenetic inference to be made. Chapter 2 also explores the MD-based bootstrap method to offer an interpretation of the significance values recovered. Two protein structural datasets, one relatively more conserved at the sequence level than the other and with different levels of structural conservation are used as controls to

simplify the interpretation of the statistics recovered from the MD-based bootstrap method.

Chapter 3 then sees the application of the Q_{score} metric to the aminoacyl-tRNA synthetases. The aminoacyl-tRNA synthetases are believed to have been present at the dawn of life, making them one of the most ancient protein families. Due to the important functional role they play, these proteins are conserved at both sequence and structural levels and well-characterised using both sequence and structure-based comparative methods. This family therefore offered inferences which could be informed with structural analysis using an automated method. Successful recovery of known relationships raised confidence in the ability of structural phylogenetic analysis based on Q_{score} to detect evolutionary signals.

In Chapter 4, a structural phylogeny was created for a protein structural dataset presenting either the histone fold or its ancestral precursor. This structural dataset comprised of proteins that were significantly diverged at a sequence level, however shared a common structural motif. The structural phylogeny recovered the split between bacterial and non-bacterial proteins. Furthermore, TATA protein associated factors were found to have multiple points of origin. Moreover, some mismatch was found between the classifications of these proteins between SCOP and PFam, which also did not agree with the results from this work. Using the structural phylogeny a model outlining the evolution of these proteins was proposed.

The structural phylogeny of the Ferritin-like superfamily has previously been generated using the Q_{score} metric and supported qualitatively. Chapter 5 recovers the structural phylogeny of the Ferritin-like superfamily and finds quantitative support for the inferred relationships from the first ever implementation of the MD-based bootstrap method. The use of the MD-based bootstrap method simultaneously allows for the resolution of polytomies in structural databases. Some limitations of the MD-based bootstrap method, highlighted in Chapter 2, are revisited in Chapter 5.

This work indicates that evolutionary signals can be successfully extracted from protein structures for deeply diverging proteins and that the MD-based bootstrap method can be used to gauge the robustness of relationships inferred.

In the loving memory of Malik M. Raza

Acknowledgements

The completion of this work would not have been possible without the help of my supervisors Drs. Jane Allison and Ant Poole, to whom I will always be indebted.

Additionally, I would like to thank Dr Thomas Collier, Ivan, William, Shamim, Apara and Jack who helped with the proof-reading of this thesis. Within the Allison group a special thanks to Ivan for tolerating my non-sense during the time spent sharing an office.

I would also like to add that this stage marks an important milestone in what has been a very long personal journey which has been influenced by numerous people and events. A sincere thanks to all of them. A special thanks to my parents, siblings and relatives for tolerating my insanity.

Finally, to Kausar, Tehreem and Amina, the three strongest and most influential people in my life, without whom I would truly be lost.

Contents

Abstract **i**

Acknowledgements **v**

List of Figures **xi**

List of Tables **xvi**

1 Introduction **1**

1.1 Introduction 3
1.2 Protein sequence 5
1.2.1 Twilight zone of sequence homology 6
1.3 Protein structure 8
1.4 Protein databases 11
1.4.1 PFam 11
1.4.2 RCSB 13
1.4.3 SCOP and CATH 14
1.5 Sequence-based phylogenetics 15
1.5.1 Sequence data 16
1.5.2 Comparative analysis 16
1.5.2.1 *Pairwise alignments using dynamic programming* 17
1.5.2.2 *Multiple sequence alignment* 21
1.5.3 Inferential method 22
1.5.3.1 *Distance methods* 23
1.5.3.2 *Character methods* 24
1.5.4 Phylogenetic tree 24
1.5.4.1 *Parametric and non-parametric bootstrap* . . 25

1.6 Structure-based phylogenetics 27
1.6.1 Hybrid sequence-structure methods 27
1.6.2 Molecular phylogentics: From sequence to structure . 28
1.7 Structural comparison 30
1.7.1 Structural representation 30
1.7.2 Structural alignment 31
1.7.3 Scoring function 32
1.8 Structure comparison metrics 32
1.8.1 RMSD 32
1.8.2 DALI 33
1.8.3 TM-Align 34
1.8.4 CE 35
1.8.5 VAST 37
1.8.6 MAMMOTH 37
1.8.7 Secondary structure matching-based Q_{score} 38
1.8.7.1 *Algorithm summary* 38
1.8.7.2 *Pairwise protein comparison: Sequence and structure* 41
1.9 Inferential method 44
1.9.1 Neighbour-joining: Algorithm summary 44
1.10 Phylogenetic tree 46
1.10.1 Conventional bootstrap 47
1.10.2 Molecular dynamics-based bootstrap method 47
1.11 Molecular dynamics for conformational sampling 49
1.11.1 Molecular dynamics summary 50
1.11.2 System representation 50
1.11.3 Force fields 50
1.11.4 System considerations 52
1.11.4.1 *Statistical ensemble* 53
1.11.4.2 *Simulation environment* 53
1.11.4.3 *Boundary conditions* 53
1.11.5 Molecular dynamics: Method breakdown 54
1.11.5.1 *Energy minimization* 54
1.11.5.2 *Molecular dynamics: Method breakdown* . . . 54
1.11.6 MD simulation: An example 56
1.12 Summary 57

Bibliography 61

2 Method Development **73**

2.1 Overview 75
2.2 Secondary structure matching-based Q_{score} 75
2.3 Method 76
2.3.1 *Part* 1: The size effect 76
2.3.2 *Part* 2: The shape effect 80
2.3.3 The MD-based bootstrap method 81
2.4 Results 83
2.4.1 *Part* 1: The size effect 83
2.4.2 *Part* 2: The shape effect 89
2.4.3 The MD-based bootstrap method 90
2.5 Discussion 96
2.6 Conclusion 97
2.7 Future work 98
Bibliography 101

3 Aminoacyl-tRNA synthetases **113**

3.1 Aminoacyl-tRNA synthetases 115
3.1.1 Evolutionary analysis of aaRSs: What is known so far? 116
3.1.2 Mitochondrial aaRSs 118
3.1.3 Structure-based phylogenetics: Recovering the known 119
3.2 Method 119
3.3 Results 121
3.3.1 Subclasses of aaRSs 123
3.3.2 Cytoplasmic, Mitochondrial and Bacterial aaRS . . . 126
3.3.3 Eocyte hypothesis 127
3.4 Discussion 128
3.5 Future Work 130
Bibliography 133

4 The histone fold **145**

4.1 Introduction 147
4.1.1 Histone fold and the core histone proteins 147
4.1.2 Nucleosome formation and properties of the histone fold 148
4.1.3 Prevalence of the histone fold 148

4.1.4 Histone-like proteins and the phylogenetic history of the histone fold . . . 150
4.2 Method . . . 154
4.3 Results . . . 157
4.3.1 Long branch attraction . . . 157
4.3.2 Presence of an evolutionary signal . . . 157
4.3.3 SCOP and Pfam organisation . . . 158
4.3.4 TATA binding protein associated factors and the histone fold . . . 163
4.3.5 Centromere-forming histones . . . 165
4.4 Discussion . . . 167
4.5 Conclusion . . . 168
4.6 Future Work . . . 168
4.6.1 Structure-based method for inferring phylogenies . . . 169
4.6.2 Histone fold phylogeny . . . 169
Bibliography . . . 171

5 The ferritin-like superfamily 181
5.1 Introduction . . . 183
5.1.1 PFam, SCOP and CATH . . . 183
5.1.2 Structural methods in phylogenetics . . . 184
5.1.2.1 *Structural alignments and scoring* . . . 184
5.1.2.2 *Robustness of phylogenetic relationships* . . . 185
5.2 Methods . . . 186
5.2.1 Structural data . . . 186
5.2.2 Structural phylogeny . . . 187
5.2.3 MD simulations and the bootstrap-like analysis . . . 187
5.3 Results . . . 192
5.3.1 PFam and SCOP classifications . . . 192
5.3.2 MD trajectory stability . . . 193
5.3.3 Interpretation of results from the MD-based bootstrap method . . . 195
5.3.4 Structural phylogeny of the ferritin-like superfamily . 195
5.4 Discussion . . . 199
5.5 Conclusion . . . 200
5.6 Future Work . . . 201

Bibliography 203

6 Summary **215**

6.1 Method development 217

6.2 Protein structural data 218

6.3 Protein databases 219

6.4 Future directions 219

Appendices **221**

Appendix-I 225

Appendix-II 229

Appendix-III 241

Appendix-IV 253

List of Figures

1.1 Protein structure 9
1.2 Solvent surface representation of a protein 10
1.3 SCOP and CATH organization 15
1.4 Sequence-based phylogenetic analysis 16
1.5 Dynamic programming matrix 18
1.6 Global sequence alignment using dynamic programming . . . 19
1.7 Local sequence alignment using dynamic programming 20
1.8 Protein pairwise sequence alignment 22
1.9 Protein multiple sequence alignment 23
1.10 A rooted phylogenetic tree 25
1.11 A rooted phylogenetic tree with support 26
1.12 Structure-based phylogenetic analysis 29
1.13 Properties of vertices and edges of the graphs assigned to calculate Q_{score} 39
1.14 Superposition of structure using secondary structure matching-based Q_{score} 41
1.15 Pairwise sequence alignment of α and β-haemoglobins 42
1.16 Superposed structures of histone H3 and H4 42
1.17 Pairwise sequence alignment of H3 and H4 histone proteins . 43
1.18 The neighbour-joining algorithm 46
1.19 The conventional non-parametric bootstrapping method . . . 48
1.20 Molecular dynamics trajectories 49
1.21 Force field terms 52
1.22 A conventional Molecular dynamics routine 56
1.23 Conformational energy landscape 58

2.1 Distribution of sizes of proteins in RCSB 77

2.2 Fractional structural analysis of proteins in the cytochrome family . . . 84
2.3 Fractional structural analysis of proteins in the ferritin family. 85
2.4 Fractional structural analysis of proteins in the globin family. 86
2.5 Distance between trees with fractional and complete structure: cytochrome . . . 87
2.6 Distance between trees with fractional and complete structure: Globins . . . 88
2.7 Distance between trees with fractional and complete structure: Ferritins . . . 88
2.8 RMSD trends for protein simulations . . . 89
2.9 The shape factor from Q_{score} calculated from protein simulations . . . 90
2.10 Limited MD-based bootstrap trials on structures from the globin family . . . 91
2.11 α and β-haemoglobin structures . . . 92
2.12 Limited MD-based bootstrap trials on structures from the cytochrome family . . . 93
2.13 Protein crystal structures from ribonucleotide reductase-like family . . . 94

3.1 Aminoacyl-tRNA synthetase conservation . . . 118
3.2 Structural phylogeny of class I aminoacyl-tRNA synthetases . 122
3.3 Structural phylogeny of class I aminoacyl-tRNA synthetases : Neighbour-net . . . 123
3.4 Structural phylogeny of class II aminoacyl-tRNA synthetases 124
3.5 Structural phylogeny of class II aminoacyl-tRNA synthetases : Neighbour-net . . . 125
3.6 The three domain and eocyte trees . . . 127
3.7 Support for the Woese three domain tree . . . 129

4.1 The histone fold . . . 147
4.2 The structure of the eukaryotic nucleosome . . . 148
4.3 Structural topology of the core histones . . . 152
4.4 Structural superimposition of the core histones . . . 152
4.5 Representative proteins having the histone fold . . . 153
4.6 Structural phylogeny of the histone fold . . . 159

4.7 Structural phylogeny of the histone fold: Neighbour-net . . . 160
4.8 Size distribution of histone fold proteins from eukaryotes and bacteria . . . 161
4.9 Structural phylogeny of the histone fold: SCOP classification 162
4.10 Structural phylogeny of the histone fold: PFam classification 163
4.11 Evolutionary model of the histone fold . . . 166

5.1 Polytomies in hierarchical databases . . . 184
5.2 Molecular dynamics trajectories of protein structures . . . 191
5.3 RMSD trends of molecular dynamics simulaitons . . . 194
5.4 Structural phylogeny of the ferritin-like superfamily : Neighbour-net . . . 196
5.5 Structural phylogeny of the ferritin-like superfamily with support . . . 197
5.6 Conserved structural core amongst members of the ferritin-like superfamily . . . 198

List of Tables

2.1 Ferritins, globins and cytochromes used to test contribution of the size factor in Q_{score} 78
2.2 Protein structures used to test the shape factor in Q_{score} . . . 82
2.3 Protein structures used to test the MD-based bootstrap method 82

3.1 Classification of aminoacyl-tRNA synthetases 116
3.2 Class I aminoacyl-tRNA synthetases 120
3.3 Class II aminoacyl-tRNA synthetases 121

4.1 Nucleosome core histone proteins 149
4.2 Histone-like protein structures 155
4.3 SCOP and PFam classification of histone-like proteins 156
4.4 TATA-binding protein associated factors 164

5.1 Members of the ferritin-like superfamily 188
5.2 PFam and SCOP classification of ferritin-like superfamily . . 189

Chapter 1

Introduction

Overview

The work in this thesis builds on the current practice in structural phylogenetics, through the addition of a novel method to quantitatively test the robustness of evolutionary inferences from structural data.

This chapter starts by introducing some gaps in the understanding of biological data that arise when sequences from deeply diverging proteins are analysed. Protein structures are introduced as possible substitutes to sequence data. The chapter continues to introduce the area of sequence-based phylogenetics and draws parallels with the established sequence-based method to present a structure-based phylogenetic approach. The structure-based approach is broken down and each part is explained in detail, starting with an explanation of pairwise comparisons of protein structures to generate distances which is followed by the neighbour-joining distance clustering method. The chapter ends with a detailed discussion on the molecular dynamics-based bootstrap method which is introduced to test tree topology and a detailed discussion of the molecular dynamics simulation method used to generate structural data for use with the molecular dynamics-based bootstrap method.

1.1 Introduction

The inheritance of genetic information between successive generations requires numerous biological processes. These processes are not error free which is why through mutations and insertion or deletion (indels) the inherited information may not be identical to that of the ancestral generation committing it [1]. The genetic information inherited is directly responsible for the phenotype presented by the progeny, and therefore the changes within may be deleterious, silent, i.e. produce no effect, or contribute positively, enhancing the fitness of the progeny.

While changes in the genetic information over successive generations might be minuscule, these tend to become significant on evolutionary time scales. The genetic information is composed of a string of nucleotides repeated along the length of deoxyribonucleic acid (DNA) [1]. Mutations are substitutions of one nucleotide for another whereas indel insert or delete nucleotides from a DNA sequence [1]. Even if there has been significant

accumulation of changes in the genetic content, molecular data may retain enough similarity to act as a guide in uncovering molecular phylogenies, which may in certain instances act as guides in recovering organismal phylogeny.

Woese et al. [2] established the three domain system of life, i.e. classification of species into one of three groups, archaea, bacteria and eukarya based on the sequences of the ubiquitously present ribosomal ribonucleic acid (rRNA). Presence of the rRNA in archaea, bacteria and eukarya provided a basis to compare otherwise incomparable organisms. The three domain system allows the evolutionary history of all living organisms to be traced back to a single organism, the last universal common ancestor (LUCA). Similar to the rRNA, sequences from other molecules, like the DNA itself, other types of RNA and proteins can also be used to recreate organismal phylogenies.

Unfortunately, the process of comparing the sequences of molecules for phylogenetic inferrence is non-trivial. Divergence increases on evolutionary time scales or alternatively it can be said that similarity between organisms emerging from a common ancestor reduces as they move further away in time [3]. This presents a challenge for phylogenetic analysis as sufficient similarity is necessary in separating an evolutionary signal, i.e. where organisms are related, from noise, i.e. where they are not. As is later discussed in this work, Section 1.2.1, sequence data from extant organisms is not always sufficiently conserved, making it difficult to extract an evolutionary signal.

This divergence in sequence-based data from DNA, RNA and proteins shifts the attention towards higher-level structural organization of RNA and proteins which can be probed for evolutionary signals. The area that explores the structure-based data to infer molecular phylogenies is referred to as structural phylogenetics. This thesis focuses on three-dimensional (3D) structures of proteins. Proteins are molecules that function based on their 3D structures. While the protein sequence to structure relation is not a straightforward one, it is noted in Section 1.3 that protein structure remains conserved over longer time scales as opposed to the underlying sequence [4–7]. The conservation in structure can therefore be used to recover phylogenies which escape traditional sequence-based evolutionary analysis.

The area of protein structure-based evolutionary analysis faces two challenges, namely in the use of an inference method for the comparison of protein structures and testing the evolutionary relationships determined from

structural analysis. The inference method presents a challenge as no established models of protein structure evolution exist. Structural distance has been used previously [8], with a special class of distance-based methods to infer evolutionary relationships. However, these relationships inferred from distance-based methods cannot be robustly tested using structural data alone as a method to accomplish this does not exist. A method to achieve this is presented in this thesis.

The following sections are intended to act as an introduction to the sequences and structures of proteins. Conventional methods of novel protein sequence characterization are introduced, followed by a challenge these methods face which create gaps in our understanding of biological data. Protein structures are introduced to show that while the sequences may lose evolutionary information, it may still be retained at a structural level. Following the introduction of protein sequences and structure the chapter introduces current practices in the area of sequence-based phylogenetics, to which parallels are drawn and a structure-based phylogenetics approach is presented. Current approaches in structure-based phylogenetics are briefly discussed and a novel addition in the form of a molecular dynamics (MD)-based bootstrap method is presented to augment the area of structure-based phylogenetics. As the primary focus of this thesis is the introduction of the MD-based bootstrap method, it builds on an existing approach of determining structure-based phylogenies [8] and couples it with this novel addition to associate a measure of significance to the relationships inferred.

1.2 Protein sequence

Protein synthesis is an intricate process where information from the DNA is used to generate an intermediate molecule, the messenger RNA (mRNA), which is subsequently used at the ribosome to produce protein molecules [1]. The ribosome uses 20 natural and two modified amino acids in the translation process to synthesize proteins. The function carried out by the protein molecule requires the protein to adopt a structure through a folding process which is poorly understood [9]. This section focuses on the sequence of the protein while the following section introduces the structure of these biomolecules.

Protein molecules have amino acids as their fundamental building blocks

which are linked together by polypeptide bonds. The sequence of amino acids is what is referred to as the protein sequence, which is essentially a combination of 20 amino acids repeated along the length of the sequence. These amino acids are usually represented by a single letter code in the sequence. An example of a protein sequence from human α-haemoglobin [10] is shown below.

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVK
GHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA
EFTPAVHASLDKFLASVSTVLTSKYR

Mostly, the first step of characterizing any novel protein sequence is by comparison with sequences of other fully characterized proteins. The comparison enumerates the number of identical and similar amino acids between the protein sequences. Numerous factors, like the length and position of regions of the sequences found similar between proteins and the association of these with the functional residues, dictate if the sequences would be similar. However it could empirically be stated that the closest protein in structure and function will be one that shares a significant sequence similarity, i.e. above a certain threshold. Although a consensus does not exist on the threshold, similarity higher than 30% is a popularly accepted value [11].

The reason behind sequence similarity correlating to structural and functional equivalence is that the protein amino acid sequence is directly responsible for the structure of the protein and the structure in turn is responsible for the function that protein molecules carry out. This relation between sequence, structure and function dictates that a similarity at the sequence level would consequently result in a structural and functional similarity, with some exceptions [12]. Protein sequences that share sequence-level similarity and consequently structure and function are referred to as being homologous [13]. This concept is widely used in protein structural modelling where a characterized protein, or a set thereof, is used to predict the structure of novel protein sequences [14].

1.2.1 Twilight zone of sequence homology

Homology is fundamental to establishing evolutionary relations however this concept is complicated by the observation of a "twilight zone". To un-

derstand this problem it is necessary to look at protein sequence comparison in more detail. In essence protein sequences are lists of characters denoting amino acids. When comparing two protein sequences one of the following two scenarios can emerge:

1. protein sequences share enough similarity to be safely classified as similar and thus homologous

2. protein sequences do not share a strong similarity and probably do not share homology

The interest, in the above cases, is to establish, from the range of similarity scores, a threshold value beyond which homology cannot be inferred. Work by Rost [11] illustrated this in a more quantitative way by looking at sequence comparison between proteins of known structure. The results of his analysis divided the sequence comparison space into three zones. These are:

- like (1) above, a region where the similarity in the aligned sequence is high enough to safely assume homology, which is termed as the "safe zone".

- the region immediately below the "safe zone" called the "twilight zone", which is marked by 10-25% sequence similarity where it becomes hard to discern an evolutionary signal from noise, resulting in the conclusion that sequences may or may not share homology

- the bottom zone of sequence similarity. This zone is occupied by sequence matches that occur purely by chance and, hence, are not significant.

The twilight zone imposes a constraint on the inferences from sequences having low similarity between them [15, 16]. PFam [17], an initiative to group protein sequences based on sequence similarity, uses notably advanced comparison methods, discussed in some detail in Section 1.4, in diverged protein sequence data in an attempt to extract weak evolutionary signals.

In cases where sequence-based analysis is complicated by the "twilight zone", sometimes protein structures act as suitable substitutes to sequences for recovering evolutionary relationships. The following section introduces protein structure and attempts to explain the reasons why structures remain conserved on longer time scales as their respective sequences diverge.

1.3 Protein structure

Protein structure refers to the specific 3D arrangement of the atoms, of the amino acids, that the biomolecule takes up once synthesized. The sequence, discussed in the previous section, is referred to as the primary structure, which is an intermediate step in the formation of a complex tertiary structure. The primary structure leads to secondary and tertiary levels of organization.

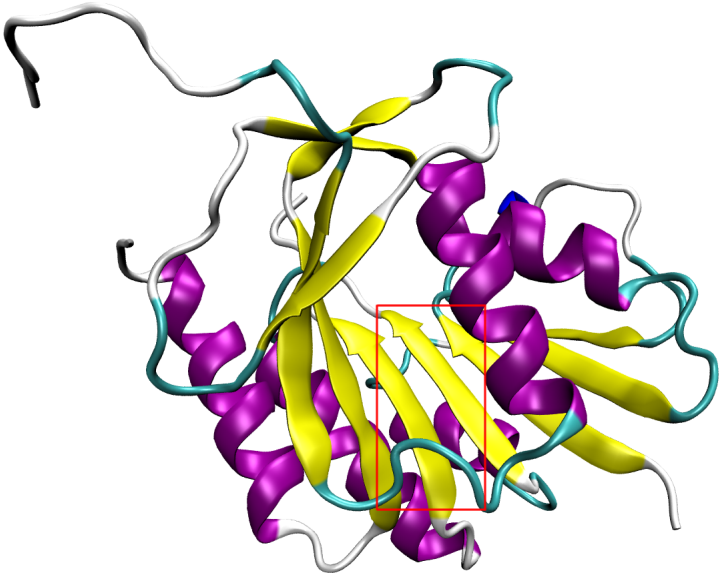
The contributions of the amino acids towards the final structure can be understood by decomposing each of them into two parts i.e. a backbone and a side chain, with the exception of glycine which presents a hydrogen instead of a sidechain. Both these parts are responsible for contributing stability towards the final 3D structure. At the secondary level backbone components of all amino acid residues interact forming localised hydrogen bonds. Successive hydrogen bonds formed between backbones of the ith and ith + 4 amino acids result in the formation of an α-helix, whereas those formed between backbones of distant amino acids result in the formation of β-sheets [1].

These secondary structure components lead to the tertiary level, where the interactions are stabilized by side chains through non-bonded interactions, e.g. hydrogen bonds, cation-π interactions, salt bridges, van der Waals interactions etc. [9, 18]. Apart from helices and sheets, sections of the sequence that fail to form an ordered element are categorized as loops [19] or disordered regions whereas short structural elements connecting two secondary structure elements (SSEs) are referred to as turns.

Complexity increases as one moves from sequence to structure as regions of the protein which are distant in the one-dimensional sequence may be adjacent in 3D space once the protein achieves its final folded state. Figure 1.1 shows a simplified representation of a protein structure. However, a more accurate representation can be approximated by wrapping a surface [20] around the structure, see Figure 1.2. This is how a protein appears to its surrounding.

The models used in protein sequence-based evolutionary analysis assume that each amino acid is capable of mutating independently [21]. While this assumption is necessary to make analysis using sequence-based methods tractable, it is somewhat removed from reality. Numerous factors like size,

chemical nature and position of the amino acid, can lead to certain amino acid pair interactions which may be crucial to the folding process, i.e required to fold the protein in the intricate way shown in Figure 1.1. To illustrate this further two cases are considered.



{...}GQDIVYANLTGEDLDIQANSVIAAMKACDVKRLIFVLSLGIYD{...}

Figure 1.1: Protein structure: This protein comprises 236 amino acids which fold in 3D space to achieve the structure shown. The structure is shown in cartoon representation. Purple regions are α helical, yellow are β sheets, cyan are turns and white regions are loops. The amino acids corresponding to the two β-strands, enclosed by the red box, are highlighted in red in the sequence underneath the structure. These are distant in the sequence yet adjacent in the final structure. The sequence is cropped on the terminals to show relevant regions only. Generated from PDB 3qvo [22].

- A single mutation may be sufficient to significantly impact the behaviour of a protein, e.g. the mutation of the sixth residue in β-haemoglobin from glutamic acid to valine is characteristic of sickle cell disease and causes a change in function of haemoglobin.
- A sequence similarity of just 68% is seen between α-haemoglobins and 69% between β-haemoglobins from *Homo sapiens* and *Anser indicus*

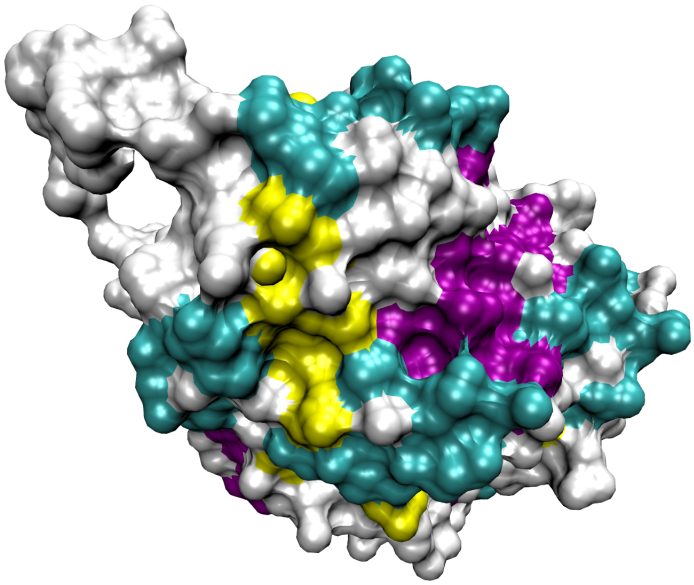


Figure 1.2: Protein surface: This representation of the protein shows cavities lining the surface which indicate the depth to which a water molecules can penetrate into the protein core, referred to as solvent accessible surface [20]. Packing of side chains of amino acids introduces steric effects stopping further penetration by external agents. The surface is coloured according to the underlying SSEs, same as Figure 1.1. Generated from PDB 3qvo.

(Goose). The heterotetrameric ($\alpha_2\beta_2$) haemoglobin carries out identical functions, of transporting oxygen, in both organisms.

Although the relationship between sequence and structure is not fully understood, these examples illustrate that equal likelihood of all amino acids to be substituted is also incorrect. These examples reveal sensitivity of protein structure to certain residues and an inattentiveness to others. From an evolutionary point of view this is suggestive that while all amino acids have a chance to be substituted, only those will undergo substitution that do not negatively impact the structure and hence the function of the protein. In the context of a structure, two groups of amino acids can therefore be formed, namely key and non-key amino acids.

Key amino acids are those whose locus in 3D space is critical for the formation of the correct structure capable of carrying out a function. A change in these residues therefore may result in a complete loss of function. The non-key residues are amino acids occupying sites that contribute towards secondary structure stability, through backbone interactions. The backbone

being common to all amino acids allows for a selection from a repertoire of amino acids.

The key and non-key amino acid grouping attaches a higher probability of mutations to non-key amino acids as opposed to key residues. The low frequency of changes in key residues is because this requires correlated evolution [23]. This is a process in which amino acids change in a way that the interactions that are lost due to mutations are simultaneously replaced by interactions from the new amino acids that are incorporated in their place. The transition from the previous state to the post-mutated state is possible because the new residues contribute either in the same way, i.e. neutral change, or enhance the protein's activity, also known as positive epistasis [24]. The key amino acids therefore change slowly, staying conserved for longer periods, whereas the non-key residues mutate relatively faster.

While all residues are undergoing change, albeit at different rates, the structure remains conserved to continue carrying out the specific function of the protein [25–27]. Hence, as the analysis moves deeper on the evolutionary time scale, an evolutionary signal lost to the "twilight zone" while comparing sequences, can be recovered from structure.

1.4 Protein databases

As protein sequence and structural data have become more abundant, databases have been set up in an effort to organise it in a systematic way. A number of databases therefore are available for both sequences and structures. This section briefly discusses some important ones relevant to the scope of this work, namely, PFam for sequences and PDB [28], SCOP and CATH [29] for structures.

1.4.1 PFam

PFam is a protein sequence database. At an organizational level the sequences in PFam are arranged into protein families and clans [30]. Each family comprises a number of protein sequences and each clan a number of protein families. Significant sequence similarity is the determining factor for grouping sequences into families. This similarity based grouping implies that members of a particular family share evolutionary history [17].

Sequence profiles and hidden Markov models (HMM) are constructed for each family. These act as a family signature against which databases are screened for new family members [31]. HMMs and profiles can, loosely, be thought of as probabilistic averages of all the sequences used to formulate them. In case of divergent families, a single average may not be sufficient to capture all novel members when screening databases and hence multiple may be constructed. These multiple entries then assist with capturing all members of a diverse family and are grouped into clans.

A simplified example is used to illustrate clans as used by PFam. Consider a divergent family of proteins, "F", for which two entries "A" and "B" are constructed i.e. HMMs and profiles. Both entries screen databases for new members to be added to the family "F". If entry "A" recovers a set of four sequences "S", comprising $seq1, seq2, seq3, seq4$ and entry "B" recovers a set of three sequences "T" comprising $seq5, seq6, seq7$, PFam attempts to construct a single profile to capture both sets of sequences "S" and "T". If this attempt fails to construct a single model, these sets of sequences, families, are grouped into a superfamily called a clan. This example gives a condensed explanation of clans [30], highlighting the determination of evolutionary relatedness when sequence comparisons are in the "twilight zone". See [30] for specific details regarding thresholds in sequence assignment to families and clans.

PFam uses four criteria to group sequences into clans. These are:

1. related structure
2. related function
3. significant matching between a novel sequence and HMMs of different families
4. profile-profile comparisons between protein families

The structural information in (1) is incorporated from protein structure classification database, SCOP [32]. This four step criterion collates protein families which are similar i.e. evolutionarily related. The clans therefore attempt to encompass protein families that share an evolutionary origin.

PFam attempts to infer homology in distantly related sequences using advanced comparison methods. However, these relationships should be taken

with a grain of salt, because as discussed earlier clans are formulated by sequence comparisons in the "twilight zone", in a mix of evolutionary signal and noise, separation of which is not a straightforward task.

1.4.2 RCSB

RCSB (Research Collaboratory for Structural Bioinformatics) provides a protein data bank (PDB, www.rcsb.org) which is a structure database. This database allows researchers to submit structural data from different sources such as X-ray crystallography, solution NMR (nuclear magnetic resonance), EM (electron microscopy).

The meta-data accompanying structures is also accessible. This usually includes several pieces of information of which some important ones are the resolution at which the structure was solved (in the case of X-ray structures), the name of the organism to which the structure belongs, the expression system in which the protein was expressed and purified, any amino acids that might be missing from the data submitted.

In terms of the structural details, Cartesian coordinates are submitted for each atom in all the amino acids of the protein with some exceptions, like hydrogens in case of X-ray crystallography-based structures. Other structural details include alternate conformers, symmetry matrices, B-factors, multiple conformations in case of solution NMR-based structures.

The RCSB database originally started out as a structural repository, but has progressively integrated useful tools that provide additional insight when analysing structure data. These tools include various structural comparison methods, e.g. rigid and flexible alignments by FATCAT [33], combinatorial extension [34] and others like MAMMOTH [35] and TM-Align [36] which are linked externally. Other tools include methods of sequence comparisons [37], structural symmetry analysis, assessment of structural quality, data imported from PFam, SCOP, CATH and UniProtKB [38].

All these combined resources provide exceptional insight into protein structures with just a few key strokes, however it is lacking in providing any evolutionary insight i.e. an organization of structures similar to SCOP and CATH.

1.4.3 SCOP and CATH

The lack of evolutionary insight from RCSB is partially filled by the SCOP and CATH databases. These databases provide a wealth of information by clustering protein structures based on structural similarity. Like PFam organizes sequences, SCOP and CATH organize protein structures hierarchically [39, 40]. However both have different approaches to organizing data. SCOP uses automated methods in some empirical steps e.g. determining similarity at a sequence level. Beyond this all assignments are manually curated. CATH uses a mix of manual and automated methods for its curation. Difficult cases are handled manually whereas the majority of proteins are automatically assigned based on sequence similarity determining methods, i.e. global alignments, Section 1.5, and protein structure comparisons program, SSAP [41]. Due to different organizational levels in SCOP and CATH, proteins can end up being classified differently. The key similarities between SCOP and CATH [42–44] are outlined in Figure 1.3.

Both SCOP and CATH attempt to infer evolutionary relationships. The lowest two levels of SCOP, i.e. Family and Superfamily and bottom level of CATH, i.e. Homology are occupied by proteins which have clear relationships between them, i.e. have significant sequence similarity or similar structures and functions, suggestive of a common origin. Superfamily in SCOP and Topology in CATH could be considered equivalent as they are both occupied by proteins which are significantly dissimilar in sequence, i.e. the sequence comparison would generate a score in the "twilight zone", yet share enough structural or functional similarity to be grouped together. Above these hierarchical levels, both databases discriminate based on structural content and organization which may not be a result of shared ancestry.

SCOP and CATH are a step up from PFam, in that the structural databases provide a deeper evolutionary insight as opposed to PFam. The structural databases, however, do not completely delineate relationships at relevant levels of their hierarchies, i.e. no relationships between protein structures are determined, which are grouped at superfamily and family levels in SCOP and Homology level in CATH. This results in a polytomous relationship at these levels in the hierarchies, which still leaves a gap in our understanding of how the proteins are evolutionarily related. While the manual curation tends to create groups of structures, lack of a structure-based phylogenetic method leaves unresolved the relationships

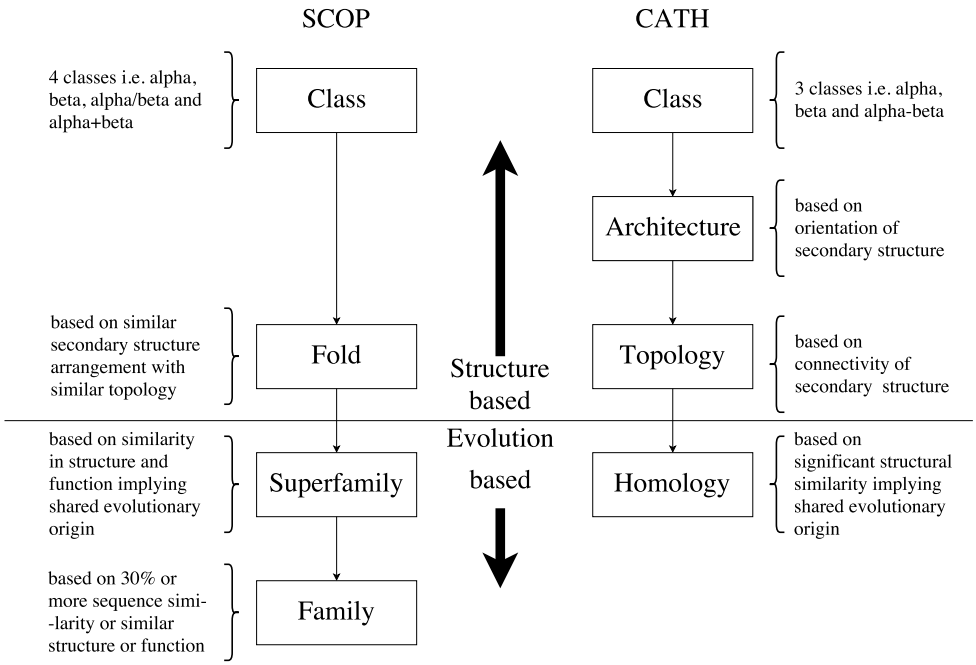


Figure 1.3: SCOP and CATH organization: SCOP arranges protein structures into classes, Folds, Superfamilies and Families. CATH uses Classes, Architectures, Topologies and Homologies to organize protein structures. The horizontal split marks a boundary which separates structure and evolution-based groupings. Structures grouped together in Homology (CATH) and Family and Superfamily (SCOP) share evolutionary origin.

between them. The work conducted in this thesis aims to present a phylogenetic approach which can address this gap in our understanding.

The preceding discussion highlights problems with conventional sequence-based methods in that they may sometime struggle to recover deep evolutionary relationships. Structure was discussed as an alternative to sequence-based methods in cases where the "twilight zone" may complicate recovery of evolutionary signals. The following section looks in some detail at the conventional method of inferring phylogenies using protein sequences.

1.5 Sequence-based phylogenetics

The starting point of any form of evolutionary analysis is collating a dataset that is a collection of comparable entities. In this section only sequences are considered, whereas structures are discussed in Section 1.6.

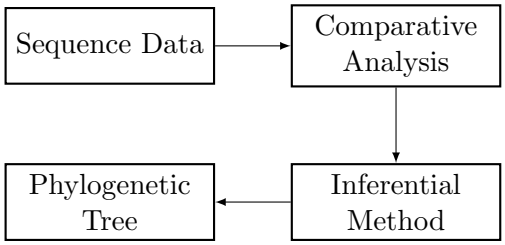


Figure 1.4: Sequence-based phylogenetic analysis. The sequence data undergoes comparative analysis to gauge similarity between all sequence pairs. Depending on the choice of inferential method, distance or character-based, a phylogenetic tree is constructed.

Figure 1.4 illustrates a scheme of steps undertaken to perform a sequence-based phylogenetic analysis. These steps are individually discussed in some detail below.

1.5.1 Sequence data

Sequence data is comparable when empirical evidence supports a shared evolutionary history between sequences populating the dataset. The empirical evidence can be significant sequence similarity, i.e. protein families in PFam, or structure or functional relatedness, i.e. proteins occupying the same family or superfamily in SCOP or that carry out similar functions. This step is crucial as comparing non-comparable sequences might render the results meaningless.

1.5.2 Comparative analysis

Comparative analysis of proteins refers to aligning protein sequences which is a non-trivial problem. Two types of alignments i.e. pairwise and multiple, can be created for sequence data and each in one of two ways i.e. local and global. The pairwise alignment with local and global alignment methods is explained first. Examples to illustrate local and global pairwise comparisons use two short stretches of sequences shown below.

Sequence 1: HEAGAWGHEE
Sequence 2: PAWHEAE

1.5.2.1 *Pairwise alignments using dynamic programming*

Pairwise alignments, as the name suggests are alignments between two sequences. The alignment problem can be separated into two parts, generating the alignment and scoring the alignment. Dynamic programming is employed to address the first part, i.e. aligning sequences. In case of pairwise comparisons, an $(N+1)*(M+1)$ matrix is used, where N and M are individual lengths of sequences i.e. number of amino acids in each of the sequences. An additional row and column in the matrix is used for initialization purposes, see Equation 1.3. Either a local or global alignment method is used. These methods are explained below.

Local and global alignments are two ways in which alignments are achieved. A local alignment [45] is one where emphasis is laid on short regions of the alignment. This type of alignment can reveal multiple short regions of similarities between the sequences compared. Alignments of this type are characterized by fewer gaps being opened to achieve the alignment. In contrast to this a global alignment [46] would have multiple gaps to ensure a high number of characters match between the compared sequences as the entire length of the sequences are aligned. Global alignments use the criteria in Equation 1.1 and local alignment uses Equation 1.2 to fill up the dynamic programming matrix.

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j), \\ F(i,j-1) - d, \\ F(i-1,j) - d. \end{cases} \tag{1.1}$$

where i and j are indices of the matrix and x_i and y_j are indices of characters in the alignment, d is the assigned gap penalty. The s-score is explained later in this section.

$$F(i,j) = \max \begin{cases} 0, \\ F(i-1,j-1) + s(x_i, y_j), \\ F(i,j-1) - d, \\ F(i-1,j) - d. \end{cases} \tag{1.2}$$

Solution of a dynamic programming problem is three tiered:

1. Initialization: an $(N+1)*(M+1)$ matrix is created and the initial row and column are filled either using:

$$Initialization: \begin{cases} F(0,0) = 0 \\ F(0,j) = -j*d \\ F(i,0) = -i*d \end{cases} \tag{1.3}$$

   in case of global alignments or completely filling the initial row and column with "0" in case of local alignment, see Figures 1.6 and 1.7.

2. The initialization is followed by applying Equations 1.1 or 1.2 to fill the cells in the matrix for global or local alignments, respectively. This is illustrated in Figure 1.5. For every *max* value that is obtained, a traceback is retained indicating which of the three neighbouring boxes contributed to the value.

3. Traceback: Once the entire matrix is filled a traceback is performed. In case of global alignment, a path is traced between the lower-right cell and the upper-left cell. For local alignments , a stretch is traced back until a "0" is reached.

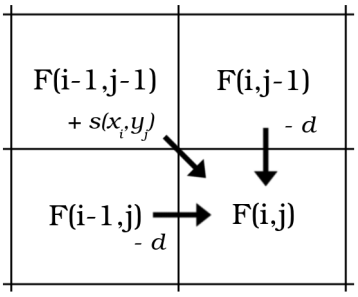


Figure 1.5: Dynamic programming matrix calculation: The criteria in Equations 1.1 or 1.2 are used to calculate value for *F(i,j)*. This step is iterated until the end of alignment. Once the *max* score is chosen, a point is left behind indicating the source of the score.

Both global and local alignments use a scoring function $s(x_i,y_j)$. This function allows for a score depending on the comparison between amino acid

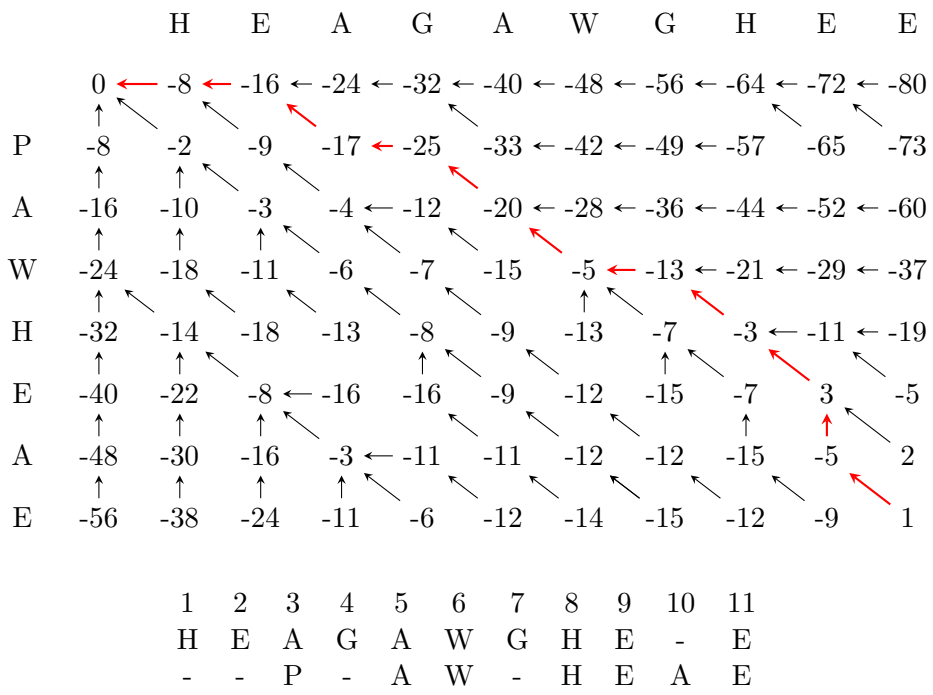


Figure 1.6: Global sequence alignment using dynamic programming. Equation 1.1 is used to completely fill the matrix (top) and tracebacks are retained. Once filled, a final traceback, shown here in red gives the global alignment. Gaps are indicated by horizontal and vertical arrows whereas aligned pairs are given by diagonal arrows. A vertical arrow introduces a break in Sequence 1 whereas a horizontal arrow does the same for Sequence 2. The final alignment is shown below the filled matrix.

x_i and y_j from the two sequences. In the case of nucleotides multiple models exist for the calculation of this score e.g. JC69 (Jukes and Cantor, 1969) [47], K80 (Kimura, 1980) [48], F81 (Felsenstein 1981) [49], HKY85 (Hasegawa, Kishino and Yano 1985) [50] and GTR (Generalised time-reversible) [51]. In the case of proteins, as is the case here, point accepted mutations (PAM) [52] and block substitution matrices (BLOSUM) [53] are used. These matrices are 20x20 matrices which give penalties for substituting one amino acid for the other. Some aspects of the PAM and BLOSUM matrices are listed below.

- Methods studying closely related protein sequences use PAM whereas those investigating evolutionarily divergent proteins use BLOSUM.
- Creation of a PAM matrix is based on global alignments whereas that

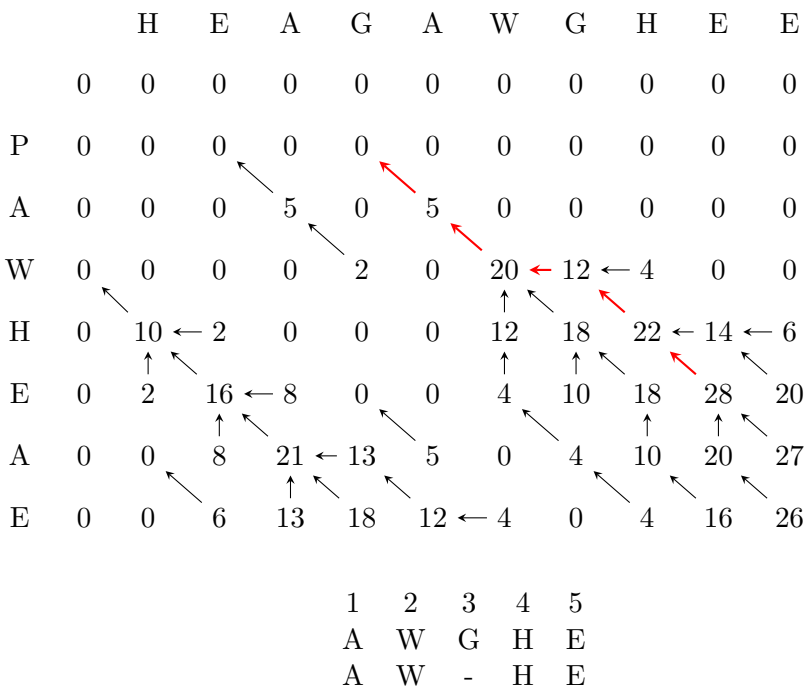


Figure 1.7: Local sequence alignment using dynamic programming. Equation 1.2 is used to completely fill the matrix (top). Instead of using gap penalties, zeros are used to initialize the matrix. In contrast to global alignment, a local alignment can end when a "0" is reached. Arrows in red indicate an alignment. Gaps are introduced in a manner similar to global alignments. The final alignment is shown below the filled matrix.

of the BLOSUM matrix is based on local alignments.

- The preferred usage of PAM matrix is for generating global alignments and that of BLOSUM is for local alignments.

PAM and BLOSUM were each designed through observation of amino acid substitutions in related sequences. Numbers are appended to the names of these matrices indicating the divergence in data used while designing these substitution matrices. For instance BLOSUM80 is used for less divergent data relative to BLOSUM45 which, in turn, will be used for sequences that are more diverged. The opposite is true for PAM, where PAM250 is for diverged data and PAM120 for related sequences. In case of BLOSUM the number indicates the similarity between the clustered sequences used to populate the matrix. PAM uses a different method where the PAM1

matrix defines the rate of observing mutations in 1% of the amino acids and all following matrices are iterations of PAM1, e.g. PAM250 would be 250 iterations of PAM1. Both PAM and BLOSUM, in essence, therefore, are empirical models for amino acid evolution which provide probabilistic measures of observing a substitution in closely related sequences. These models are frequently implemented when aligning sequences.

Once an alignment is achieved the number of matches, mismatches, instances of gap opening and extensions are counted and presented as a ratio of the alignment length. These ratios act as indicators of the quality of the alignment.

1.5.2.2 *Multiple sequence alignment*

Alignments can be between two sequences, i.e. pairwise as illustrated earlier, Figure 1.8, or more sequences, i.e. a multiple sequence alignment (MSA), Figure 1.9. Both these methods have their own advantages. While pairwise sequence alignment might recover regions of similarity between two sequences, an MSA would be successful at picking regions of similarity across multiple sequences which might be too subtle to be identified through pairwise comparisons [54].

The time complextity ($\mathcal{O}(L^{\mathrm{N}})$ for N sequences of length L) makes MSA, using the dynamic programming method, impractical when more than a few sequences are compared. As MSA is a popular choice, multiple methods have been developed to bypass this time complexity problem, namely progressive alignment methods like T-Coffee [55], iterative methods like MUSCLE [56], hidden Markov methods i.e. Probcons [57], genetic algorithms [58] and simulated annealing [59].

Although a detailed discussion of these methods is beyond the scope of this work, the motivation behind these methods is two fold, reducing errors in alignment when methods move away from the exact algorithm of dynamic programming and achieving reasonable results in practical time scales e.g. the time complexity of MUSCLE is $\mathcal{O}(L^2+N^2)$. Since no specific MSA methodology has been popularly accepted [60–62], different results are achieved depending on the choice of method [62, 63].

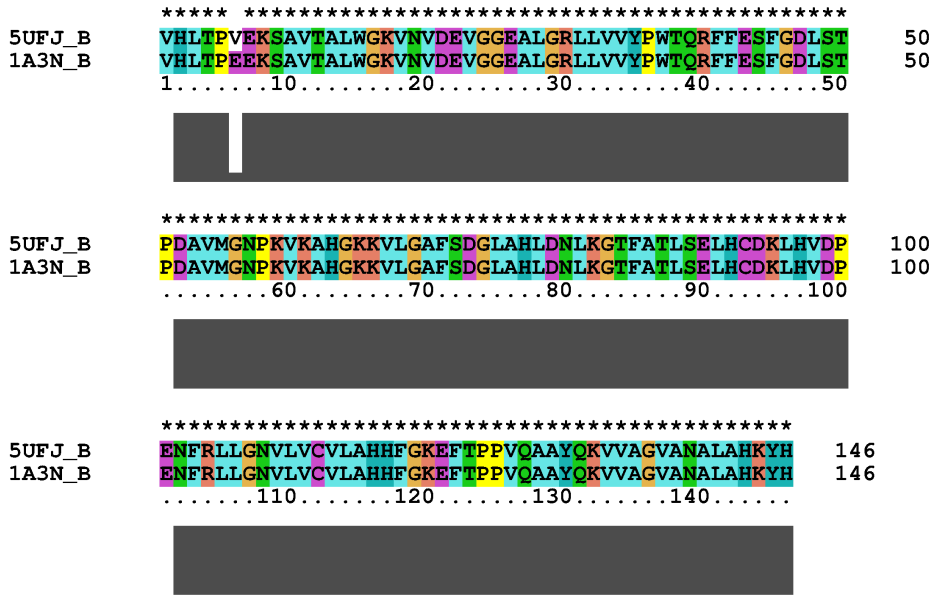


Figure 1.8: Protein pairwise sequence alignment. The wild-type β-haemoglobin is compared to one that has a mutation at the sixth position characteristic of sickle-cell disease. Protein sequences were obtained from PDB structures 5ufj [64], chain B and 1a3n [65], chain B. A "*" symbol on top of each column indicates consensus, at that position, across sequences. A single colour across the column indicates an identity or chemical similarity. Residue positions are indicated as numbers. The grey histogram shows consensus across all positions except position six.

1.5.3 Inferential method

Once the sequence data is aligned, a distance or character-based method is chosen. These methods are discussed in some detail below. The primary difference between the two is that distance methods compute a measure of divergence, i.e. a distance for each sequence pair in the alignment, and use the distance to create a phylogenetic tree. Character-based methods, on the other hand, find a tree which best fits the alignment [62]. Both methods have weaknesses; distance methods result in loss of information when alignments are converted to distances [69], whereas character-based methods are sensitive to homoplasy (i.e. convergence) and the search for an optimal tree makes this process slow [62, 70, 71].

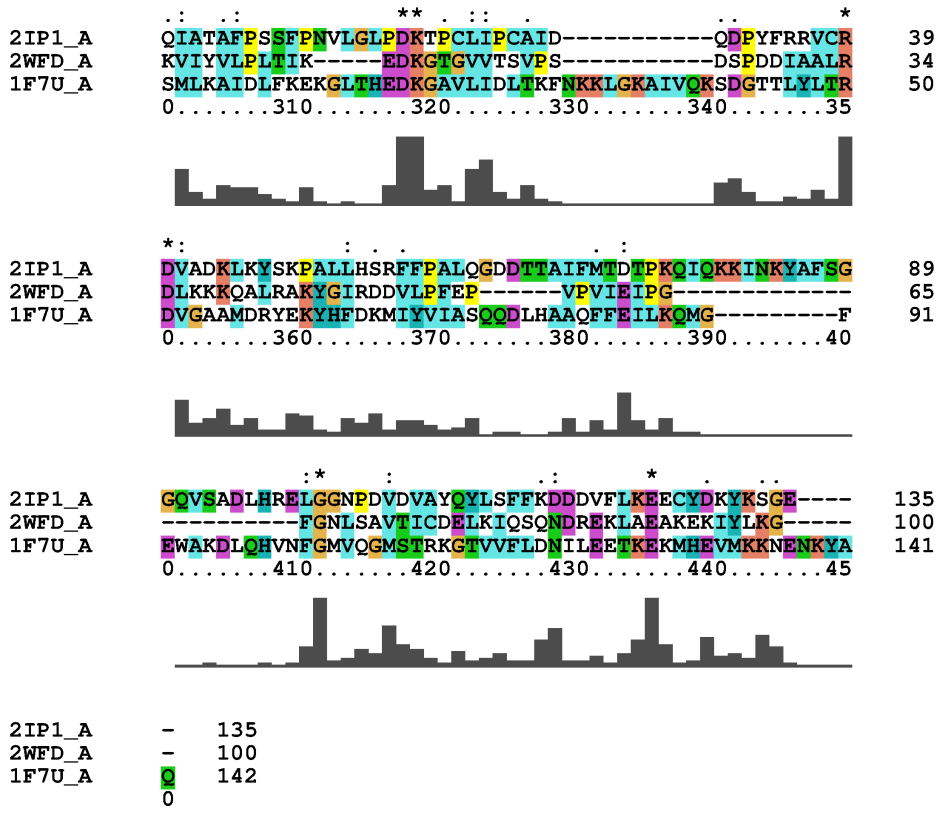


Figure 1.9: Protein multiple sequence alignment. A subsection of the full alignment of three aminoacyl tRNA synthetases. Sequences were extracted from their protein structures (PDB 2ip1 [66], chain A, 2wfd [67], chain A and 1f7u [68], chain A). "*" indicate consensus across positions, ":" and "." are indicative of falling levels of similarity."-" indicate gaps opened to optimize the alignment.

1.5.3.1 *Distance methods*

Distance-based methods include clustering algorithms and those utilizing objective functions. Clustering algorithms include unweighted pair group method with arithmetic mean (**UPGMA**) [72] and neighbour-joining (**NJ**) [73]. Minimum evolution (**ME**) [74] and least squares (**LS**) [75–77] are distance methods that use objective functions.

Amongst the distance methods UPGMA assumes a constant rate of evolution which does not hold for most biological data. The NJ method, in contrast to UPGMA does not assume a constant rate and hence is more popular. Both UPGMA and NJ are fast as they generate a single tree and

lack a tree searching and optimality criterion. This can also be considered their weakness as it means that the single tree generated may or may not be correct.

ME, in contrast to UPGMA and NJ, has an objective function to minimize evolutionary distance by selecting the shortest tree as the best tree. The ME method suffers from the same loss of information, when sequence data is converted to distances, as other distance-based inferential methods. Furthermore the ME method is slow, as it searches for the best tree. LS is another distance-based method which minimizes the sum of squared branch length differences between the given and predicted data. LS has the same weaknesses as ME i.e. loss of information and slow nature of the algorithm.

1.5.3.2 *Character methods*

Character-based methods include maximum parsimony (**MP**) [78, 79], maximum likelihood (**ML**) and Bayesian methods [49]. MP attempts to find the tree with the minimum number of changes, whereas ML finds a tree, based on likelihood, that best fits the observed data. Bayesian methods are a type of likelihood method and are discussed in some detail in Section 1.6.1.

Parsimony does not explicitly assign probabilities of amino acid substitutions and is known to struggle with long branch attraction. ML, on the other hand, utilizes quantitative models to describe the evolution of sequences in the alignment and makes use of probability to find the best tree that supports the observed data. Both these methods are sensitive to homoplasy (i.e. convergence) and are slow.

1.5.4 Phylogenetic tree

Phylogenetic trees demonstrate relationships between compared entities [76]. Phylogenetics describes evolution as an empirical process of successive furcations from the root (last common ancestor) to the entities being examined. A tree ignores events like horizontal gene transfers and hybridization events, suggesting only vertical transfer of information as entities evolve. Reticulation networks, on the other hand, are sometimes used, as alternatives to trees, to gauge the accuracy of data that is converted into a tree. Departures of a network from tree-likeness are suggestive of the aforementioned events. The ability of networks to inform the analysis of transfer and

hybridization events give them an edge over regular phylogenetic trees.

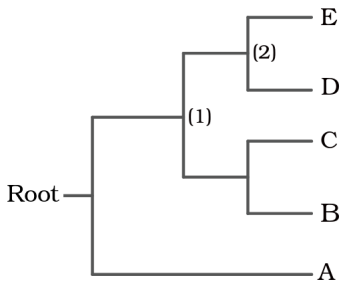


Figure 1.10: A rooted phylogenetic tree: A-E are entities (e.g. protein sequences) of interest. The tree reveals a bifurcation at (1) resulting in the ancestors of E,D and C,B. The E,D ancestor bifurcates to form the entities as currently observed E and D. (1) and (2) are nodes. Horizontal lines are clades which connect nodes. The horizontal lengths are indicative of distances or character changes depending on the method used. Vertical distance is for visual purposes only.

Phylogenetic trees may be rooted or unrooted. Rooted trees, Figure 1.10, can be generated when the root is known a priori, which is not always the case. Roots if not known, can be introduced in a few ways, e.g. midpoint rooting is a method which places the root midway between the two most distant taxa. Figure 1.10 is an example of a midpoint rooting with the root placed midway between taxa A and E. Another way in which rooting can be done is by using out groups, where the out group is known to be significantly different from the group of interest e.g. avian species can form an out group to mammals.

An inferred tree may not always be reflective of true evolutionary history, which is why statistical testing is used to measure the robustness of the signal represented by the tree. There exist a number of methods for statistical testing in phylogenetics, e.g. Jackknife [80] and bootstrap [81], of which only bootstrap is discussed in this work. Bootstrapping is a popular method, two forms of which, parametric and non-parametric [62], are explained below in the context of sequence-based analysis.

1.5.4.1 *Parametric and non-parametric bootstrap*

The evolutionary model presented by a tree is dependant on the sequence alignment. This alignment is used to construct a single tree. Bootstrapping

aims to create perturbation in the alignment and uses it to gauge how sensitive the relationships reflected on the tree are. Ideally if the tree correctly estimates the evolutionary history, small perturbations should not result in significantly different trees. The perturbations in the alignment, to test tree topology, are created either using a parametric or non-parametric approach.

Parametric bootstrapping generates new datasets by simulating protein sequence evolution according to some evolutionary model [82]. In non-parametric bootstrapping [82], the initial sequence alignment is "resampled with replacements" to construct new datasets, with each dataset constituting a trial. Resampling with replacements means that an alignment is divided into columns with each column being independent [83]. The samples (columns) are then drawn at random from the alignment and a new alignment of the same size is constructed. The new alignments can have one column repeated multiple times and some columns never sampled.

Regardless of the choice of parametric or non-parametric methods, once new alignments are generated, each alignment is used to construct a tree. Relationships observed in the new trees are compared to a reference tree. An example of this would be counting the number of times a relationship between two taxa is recovered and expressed as a fraction of the total number of trials or expressed as a percentage. This is illustrated in Figure 1.11.

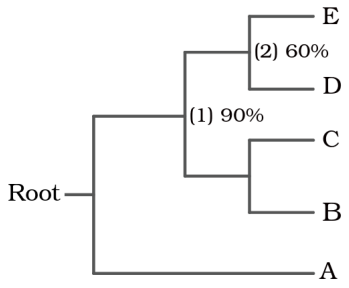


Figure 1.11: A rooted phylogenetic tree with support. Same as Figure 1.10 with statistical support only shown at nodes (1) and (2). The support values illustrated on the nodes were obtained in the following way. 10 bootstrap trials were conducted out of which six trials resulted in the same connectivity between taxa E and D and nine of 10 trials recovered E,D and C,B as sister groups.

There is some debate over the meaning of the bootstrap values with some supporting that they reflect the robustness [82] of the relationships i.e. re-

peatability while others argue that they are a measure of accuracy [62]. Despite these debates the non-parametric bootstrap remains a popular choice to test topology of phylogenetic trees.

1.6 Structure-based phylogenetics

The primary focus of this work is to attempt to fill the gap created by sequence-based methods while inferring evolutionary relations for distantly related protein sequences. It has been noted in earlier sections that protein structures may be used to recover evolutionary signals. Before a formal introduction to the structure-based method and the novel addition introduced in this work, some existing attempts to include structural information are presented.

1.6.1 Hybrid sequence-structure methods

Previously, sequence-based phylogenetic methods have been discussed at length. To understand the hybrid methods, the reader is reminded of two important steps in determining phylogenetic relationships using sequences, namely the step of aligning protein sequences, Section 1.5.2, and use of an inferential method, Section 1.5.3.

Protein sequence alignment is informative, as previously discussed, when there exists a significant proportion of similar sites between sequences compared, i.e. the comparison does not fall in the "twilight zone". Deep evolutionary analysis faces this exact problem. Hybrid methods use an amalgam of protein sequence and structural information to achieve a better sequence alignment, i.e a sequence alignment which is informed by conservation in structure. This hybrid approach is useful when comparing divergent sequences [84, 85]. One example of this hybrid approach to alignment can be seen implemented in 3D-Coffee [86], a plugin within the T-Coffee program, that utilizes structural information to better align sequences.

Bayesian methods are a class of character method that have recently gained attention [82, 87], and more recently been coupled with structure-informed sequence methods [84] or a joint sequence-structure method [85] to uncover deep evolutionary relationships. These methods use a likelihood function to obtain posteriors using an evolutionary model, consequently resulting in a tree which best fits the given data.

The first stochastic model using both sequence and structure was proposed by Challis et al. [88]. The evolutionary model included three components, namely mutations, indels and structural drift. The joint likelihood of the combined sequence-structure model, conditional on the indel process, is given by the product of the likelihoods of the individual models. This Bayesian method was further extended by Herman et al. [85] for a complete phylogenetic analysis of some members of the globin protein superfamily.

The importance of protein structure has long been known to the scientific community, and with more protein structures becoming available the methods including structural data have gained traction. While a structure-informed sequence alignment results in the recovery of a more robust evolutionary signal from deeply diverging proteins, the use of character methods continues to assume independent evolution of sites in the alignment, including the recent developments which include a structure evolution model [85, 88]. As discussed in detail previously, Section 1.4, this is not entirely accurate. Bayesian methods also attract controversy from assignment of priors [82], as these directly impact the outcome.

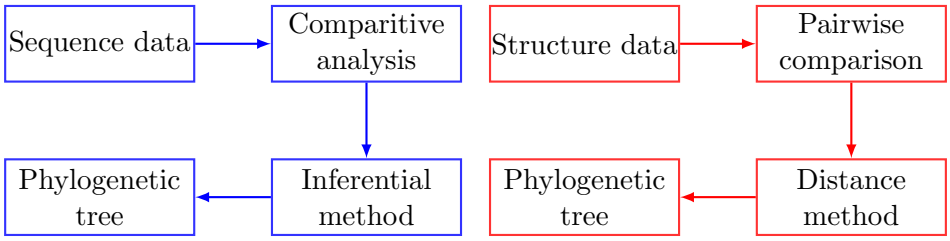
The character methods search from a tree-space, a tree which best fits the data. The tree-space is not exhaustive and strongly dependent on the choice of the model chosen to define the evolution of the data. The character-based methods, therefore, may not always present the "true tree". The distance-based methods avoid the use of models to define the evolution of structural data. While distance-based methods may perhaps suitably recover evolutionary signals from structural data, they suffer from a lack of a quantitative method to robustly test the inferred relationships. This work introduces a novel approach to enhance the ability of the distance-based method used with structural data to recover evolutionary relationships.

The following section draws parallels with sequence-based methods and introduces the structure-based method to infer phylogenetic relationships.

1.6.2 Molecular phylogentics: From sequence to structure

Section 1.5 introduced sequence-based phylogenetics which was followed by some details of each of the steps. Here, parallels are drawn with the sequence-based method to construct a structure-based approach. This structural approach, shown in Figure 1.12, extends on earlier work by Lundin et al. [8] with the addition of a novel phylogenetic tree testing mechanism, pre-

viously missing, making the approach presented here a complete method for structure-based phylogenetic analysis.



(a) Sequence-based (b) Structure-based

Figure 1.12: Structure-based phylogenetic analysis: (a) same as Figure 1.4. (b) Structure data from comparable proteins are collected and pairwise compared using a structure comparison metric Q_{score}. The similarity data are converted to distance data and the resulting distance matrix is used with the distance-based NJ clustering method to generate a phylogenetic tree. The tree is then tested using the MD-based bootstrap method. Colours are for visual separation only.

Similar to the sequence-based method, the structural approach starts with collecting data which in this case are comparable 3D protein structures. Examples of sources for data of this type would be the RCSB structural database. The structures are then pairwise compared using a structural comparison metric, Q_{score} [89], based on secondary structure matching (SSM). The Q_{score} metric is compared to other structural metrics and explained in detail in Section 1.8. The similarity scores from Q_{score} are converted to distance values and a pairwise-distance matrix is populated for the proteins compared. A distance-based clustering method, NJ, detailed in Section 1.9, is then used to infer a phylogenetic tree. To test the relationships inferred from the tree, a novel MD-based bootstrap method is used, using structural data only, to gauge the statistical significance of the inferred relationships.

The hierarchical classification of structures by SCOP and CATH, manually or using semi-automated methods, clusters structures based on structural similarity. The use of Q_{score} aims to quantitatively capture the similarity by measuring structural distance in a similar manner. Furthermore, this method of using protein structures alone, to determine evolutionary relationships, through structural similarity, moves away from the caveats of hybrid methods. Firstly, the use of a distance method, NJ, with a measure of structural distance as quantified by the SSM-based Q_{score}, considers structure as

a whole, instead of individual residues being able to evolve independently. Secondly, this method does not require assumptions regarding the evolution of protein structures and empirically tests for the presence of an evolutionary signal. This allows the distance-based structural comparison method to simultaneously move away from choosing an evolutionary model, in case of likelihood methods and additionally priors in case of Bayesian methods.

The following sections explains in more detail the pairwise protein structure comparison, the NJ algorithm and the MD-based bootstrap method, all of which come together in this approach to use protein structures for inferring deep evolutionary relationships.

1.7 Structural comparison

The protein structure comparison problem is perhaps as old as experimentally determined protein structures themselves. Comparison of a protein structure with another serves the same purpose as does the sequence comparison i.e. to identify similarity between proteins. The function of an uncharacterised protein can easily be predicted if it shares significant structural similarity with a functionally characterized protein as structural similarity implies functional similarity.

As is the case with sequence comparison, structure comparison is a non-trivial problem. For a complete understanding of this problem, in the context of 3D protein structures, it is decomposed into the following three parts:

1. structural representation,
2. structural alignment and
3. generating a similarity score for the alignment.

These are discussed in order below.

1.7.1 Structural representation

Protein structures are typically represented as a set of atomic Cartesian coordinates. However rotation and translation can make two otherwise-similar sets of Cartesian coordinates appear vastly different. Therefore the first step in solving the structural comparison problem is the conversion

of the structure to a form of representation which is positionally, i.e. rotationally and translationally, independent. A common example of this is representing the structure through inter-atomic distances listed in a two dimensional distance matrix rather than Cartesian coordinates. Another example of this is to represent the protein structure as vectors i.e. secondary structure elements (SSEs) are represented as a set of vectors. The comparison is then made between these matrices or vectors instead of the structures. In essence this conversion eliminates contributions from translational and rotational effects, therefore readily allowing the structures to be compared without the need to eliminate position-specific constraints.

1.7.2 Structural alignment

Once the structures have been internally represented, their alignment is carried out. In sequence alignment methods, as discussed in the previous chapter, substitution matrices provide transition probabilities to optimize the alignment. The effectiveness of substitution matrices for structures is low, but they have been generated [90, 91]. The methods discussed here do not employ these substitution matrices, hence they are not included in this discussion.

The structural alignment can be carried out in two ways i.e. in a rigid or flexible manner. Rigid alignments are only used when proteins with identical amino acid content or identical lengths are compared. These are unrealistic, in an evolutionary context, because rigid alignments do not allow for the introduction of gaps which are equivalent to insertion-deletion events on evolutionary time scales. Flexible alignments are ones where gaps are introduced to optimally align structural segments which are more similar e.g. a gap may be introduced in either structure if, as a consequence of it, a better alignment is achieved between two α-helices.

The focus of this work is only on flexible alignments and some popular metrics that perform flexible alignments are discussed in Section 1.8. The result of a structural alignment is a list of amino acids across the two structures which are considered equivalent and hence used for the calculation of a similarity score.

1.7.3 Scoring function

The last step is the use of a scoring function which quantifies the similarity or difference amongst the equivalent pairs of amino acids determined in the previous step. This can be done in a number of ways. One example could be that the structural alignment is converted to its sequence alignment and the empirical scoring function discussed in the previous sections is used to quantify similarity of the proteins. However this is not a common practice and a number of complex functions are used depending on the metric employed. Some of popular metrics and their scoring functions are discussed in the following section.

1.8 Structure comparison metrics

There are a number of structural comparison metrics available. Those that appear more frequently in the scientific literature are RMSD, DALI [92], TM-Align [36], CE [34], VAST [93], MAMMOTH [35] and SSM [89]. Each metric generates different results primarily because of the sub-optimal nature of these algorithms. Although for this work, SSM-based Q_{score} was employed, any other metric as long it meets a certain basic criterion, discussed towards the end of section 1.8.6, can be used. The following discussion introduces the popular metrics mentioned earlier, for comparison purposes, followed by an in depth discussion of the SSM-based Q_{score}.

1.8.1 RMSD

Root mean squared deviation (RMSD) is a method of quantifying dissimilarity between structures and hence can be regarded as a scoring function. RMSD, essentially, calculates distance between a pair of points or, in the case of protein structures, atoms. The mathematical relationship to quantify this difference is given by:

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)} \quad (1.4)$$

where v and w are two sets of n points and x, y, z are 3D Cartesian coordinates.

RMSD therefore generates a measure of difference in the coordinates of the two sets of points compared. In case of rigid superposition of structures before employing RMSD it is essential to remove any contributions from rotational and translational motions by protein structures in 3D space. This is because the empirical superposition approach using RMSD does not represent protein structure in a position-independent manner. To achieve an optimal superposition, a pairwise correspondence is considered between points in v, e.g. coordinates of atoms from structure A, as they are mapped onto points in w, e.g. coordinates of atoms from structure B. During superposition one structure is kept as a target onto which the other structure is mapped. A common method of mapping uses the Kabsch algorithm [94], to generate a transformation matrix which maps one structure onto the other such that the RMSD between the superposed structures is minimized.

Other comparison metrics differ from this RMSD-based superposition method in that they first represent the structures internally. This eliminates the need to remove position-dependant rotational and translational motions. Secondly, the metrics considered here allow for flexible alignments i.e. the structural comparison will not assume a one to one relation. This means that while rigid superposition would superpose atom with index *"m"* in structure "A" with atom *"m"* in structure "B", a flexible alignment could compare atom *"m"* from structure "A" to *"m - k"* or *"m + k"* in structure "B" where *"k"* can be a number that introduces an offset of a few amino acid residues. This alignment process decides which atoms are considered equivalent and hence should be compared, while retaining connectivity. Once an equivalence is established RMSD (as a scoring function only) or other complex functions could be used to quantify the difference between them.

Following from the above discussion, RMSD could not be employed for this work as in itself it does not employ a flexible method of aligning structures. For phylogenetic analysis, the metric needs to be able to perform flexible alignments to better capture evolutionary processes.

1.8.2 DALI

DALI (distance matrix alignment) is a structural comparison metric that utilizes a distance matrix to represent a protein structure internally. It calculates distances between all α-carbons (Cα) in a structure and uses the distances to populate a square distance matrix in a sequence dependant

manner. The distance matrix now represents the three dimensional structure. For pairwise comparison both structures are first represented by their respective distance matrices. This is followed by the comparison process which starts by reducing each distance matrix to a set of distance matrices of six amino acids each. Each of these matrices essentially represents a hexapeptide fragment from the original protein structure. The same process is carried out with the distance matrix of the other structure. An all by all comparison between the two sets of hexapeptide matrices is carried out. The final alignment is such that the scoring function in Equation 1.5 is maximized [95].

$$S = \sum_{i \in core} \sum_{j \in core} (\theta - \Delta(d_{ij}^A, d_{ij}^B))\omega(d_{ij}^A, d_{ij}^B) \tag{1.5}$$

Where for structures A and B, core is the set of structurally equivalent pairs (i^A, j^B), Δ is the deviation of intramolecular Cα-Cα distances between (i^A, j^A) and (i^B, j^B), relative to their arithmetic mean (d). θ is an empirical threshold of similarity, empirically set to 0.2, ω is an envelope function such that $\omega = \exp(-d^2/r^2)$ where $r = 20$ Å.

Once a final alignment is achieved a Z-score is returned which indicates how similar or different a particular structural alignment is against a distribution of random structural alignments.

Although effective in many cases, the DALI metric has two drawbacks a) loss of information when converting a 3D structure to a 2D distance matrix and b) the output score (Z-score) is an open ended probability score i.e. it is not normalised. Holm et al. [96] have suggested a way in which to normalize the score but this uses a background distribution of comparisons. This is problematic because a change in the limited structural data used to calculate the background distribution may impact the normalization process. Furthermore probability scores are not equivalent to distances and hence even in a normalized state the Z-score cannot be used as a measure of structural distance.

1.8.3 TM-Align

TM-Align uses α-carbons while reducing a structure to its respective SSEs. A helix, sheet or coil is assigned by looking at a five amino acid neighbour-

hood of each α-carbon. Dynamic programming is then used to align the SSEs across two structures, with "1" indicating a match, "0" a mismatch and penalty of "-1" for a gap. In the second step an alignment is performed, using dynamic programming, where no gaps are inserted and the best alignment, using TM-score in Equation 1.6, is selected.

$$TM_{score} = \frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \frac{d_i}{d_0}} \tag{1.6}$$

where L_N is the length of the native structure, L_T is the number of residues aligned to the template structure, d_i is the distance between the ith pair of aligned amino acid residues and d_0 is a scaling factor given by:

$$d_0 = 1.24 \sqrt[3]{L_N - 15} - 1.8 \tag{1.7}$$

A third alignment then reintroduces gaps but uses a mix of scores from the first two alignments. The three initial alignments are followed by a heuristic approach which uses the rotation matrix obtained in the earlier alignments to iteratively improve the alignments such that the TM-score is maximized.

TM-Align provides results with a score normalized between 0 and 1 however the normalization is performed with a scaling factor d_0, see Equation 1.7, in which the empirical constants are based on random structural matches. Furthermore the score formulation, see Equation 1.6, only considers the length of one structure onto which the other is being mapped i.e. L_N. Although the output generates two scores, each normalized on the length of the respective structure, significantly different scores emerge when the two structures are of different sizes, which is usually the case with significantly diverged proteins.

1.8.4 CE

Combinatorial extension (CE) is based on internal distance matrices similar to DALI, but instead of computing the distance matrix for the entire structure it considers octameric fragments. These fragments are then aligned across the two structures, with the best aligned pair acting as the seed for the remaining alignment. CE defines the best alignment as the longest con-

tinuous path P of aligned fragment pairs (AFPs) between two structures n_A and n_B amongst all AFPs. Three criterion determine the continuity of AFPs constituting an alignment. These are:

$$p_{i+1}^A = p_i^A + m \ \wedge \ p_{i+1}^B = p_i^B + m \tag{1.8}$$

or

$$p_{i+1}^A > p_i^A + m \ \wedge \ p_{i+1}^B = p_i^B + m \tag{1.9}$$

or

$$p_{i+1}^A = p_i^A + m \ \wedge \ p_{i+1}^B > p_i^B + m \tag{1.10}$$

where p_i^A and p_i^B are the starting residues in protein A and B at the ith position in the alignment, m is size of the similarity matrix set to eight for octameric fragments. Equation 1.8 describes consecutive AFPs whereas Equations 1.9 and 1.10 represent consecutive AFPs aligned with gaps in structure A and B, respectively.

The seed alignment is extended to maximize the alignment by incorporating other aligned fragments using Equations 1.8, 1.9 and 1.10. To minimize gaps Equations 1.9 and 1.10 are improved through the addition of gaps and given by:

$$p_{i+1}^A = p_i^A + m + G \tag{1.11}$$

$$p_{i+1}^B = p_i^B + m + G \tag{1.12}$$

where G is the maximum gap size of 30, chosen empirically. A heuristic approach is used coupled with distance measures for extending seed alignments. Following the alignment the aligned fragment pairs are listed along with a Z-score which uses a normal distribution with mean "0" and standard

deviation "1". The Z-score gives the quality of the alignment by providing the probability of attaining a better alignment from comparing random structures.

Similar to DALI, CE loses information in reducing the structure to a matrix of distances. Moreover CE does not provide a normalized score.

1.8.5 VAST

The vector alignment search tool (VAST) [93] method only allows for comparison between a query structure and a target database and therefore cannot be used for pairwise comparisons. Although this method is used in the literature, it is not suitable for generating pairwise structural distances, therefore it is not discussed in detail here.

1.8.6 MAMMOTH

Matching molecular models obtained from theory (MAMMOTH) represents a continuous fragment of seven α-carbons as a unit vector. The vector points from the ith to the ith + 1 α-carbon. The vectors are mapped to the origin as unit vectors. For comparison between two structures, unit vectors from one structure are mapped onto the unit vectors of the other structure and a rotation matrix is determined. The final rotation matrix is such that it minimizes the sum of squared distances between corresponding unit vectors and the square root of the minimized sum gives the URMS (unit vector root mean square) distance. In the second step the URMS is converted to a similarity score by comparing the observed URMS with a minimum URMS. The minimum URMS is given by:

$$URMS^R = \sqrt{2 - \frac{2.84}{\sqrt{n}}} \tag{1.13}$$

where R is a random set of n unit vectors. The similarity score is thus given by:

$$S^{AB} = \frac{URMS^R - URMS^{AB}}{URMS^R} \Delta(URMS^R, URMS^{AB}) \tag{1.14}$$

where $\Delta(URMS^R, URMS^{AB})$ is "10" if $URMS^R > URMS^{AB}$ and "0" otherwise. Comparison between all possible heptapeptides therefore populates a scoring matrix. In the following step dynamic programming is applied to the similarity matrix to build an alignment using the global alignment method, discussed in the previously, Section 1.5.2.1. Finally a maximum subset of local structures having α-carbons within 4 Å is identified and reported as PSI (percentage of structural identity). Extreme value fitting is used to compute a P-value to assess the score quality by comparing to random structural alignments.

All the algorithms outlined in this section either decompose structure to distance, which may result in a loss of information, or have non-normalized comparison scores, i.e. there is no upper bound for the score, which increases as the sizes of proteins compared increases. Additionally, instead of scores some metrics generate a probability value to assess the quality of the comparison, probability values that may change when the background distribution used to calculate them changes. For these reasons, none of these metrics were used here. Instead the SSM-based Q_{score} was used. This metric is outlined in the following section.

1.8.7 Secondary structure matching-based Q_{score}

1.8.7.1 *Algorithm summary*

SSM is an algorithm which uses structural elements to build graphs and match them to quantify a measure of similarity between compared protein structures. The SSEs are identified using PROMOTIF [97] and only amino acids lying in α-helices and β-strands are used for subsequent analysis. Vertices and edges of the graph representing a structure are labelled. The vertex carries the labels of length (number of amino acid residues in an SSE) and type of SSE (α-helix or β-strand). Edges connecting vertices carry labels defining geometrical details concerning the connected SSEs e.g. connectivity information, angles between the SSEs represented by the connected vertices etc. This is illustrated in Figure 1.13.

SSEs have different mathematical representations, depending on type i.e. α-helix or β-strand. A vector for an SSE (r) is defined as:

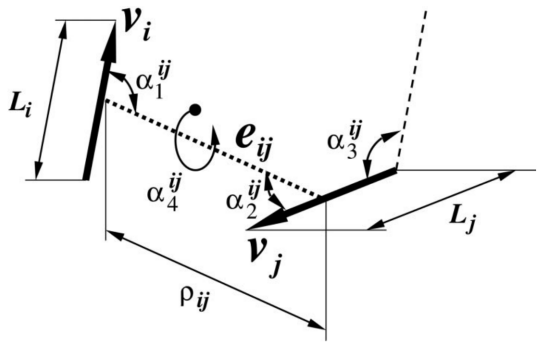


Figure 1.13: Properties of vertices and edges of the SSE graph. Vertices v_i and v_j are represented as vectors. Edge e_{ij} connects the centres of the vectors. ρ_{ij} and angles α_k^{ij} are parameters that define the relative orientations of element i and j.

$$r_{sse} = r_b - r_e \tag{1.15}$$

where for an α-helix:

$$\begin{cases} r_b = (0.74r_p + r_{p+1} + r_{p+2} + 0.74r_{p+3})/3.48 \\ r_e = (0.74r_{q-3} + r_{q-2} + r_{q-1} + 0.74r_q)/3.48 \end{cases} \tag{1.16}$$

and for a β-strand:

$$\begin{cases} r_b = (r_p + r_{p+1})/2 \\ r_e = (r_{q-1} + r_q)/2 \end{cases} \tag{1.17}$$

where indices p and q denote the serial numbers of the first and last amino acid residues in the SSE.

Between vertices, edges and labels a complete representation of the structure is possible. Connectivity of SSEs is introduced using a "connect" function for three scenarios of connectivity i.e. no, flexible and strict connectivity.

After the structures are internally represented the SSEs are matched. Following the initial matching, the best aligned α-carbons are used as seeds to expand the alignment. The following function, Equation 1.18, gives the final score of the comparison between the structures.

$$Q_{score} = \frac{N_{align}^2}{[1 + (\frac{RMSD}{R_0})^2]N_1 N_2} \tag{1.18}$$

where N_{align} is the number of residues aligned, R_0 is an empirical parameter set to 3 Å and N_1 and N_2 are the lengths of the proteins compared.

This expression returns a normalized score between 0 and 1, with 1 indicating identical proteins and the score approaching 0 as dissimilarity increases. In the expression for the quality of the alignment, Equation 1.18, the ratio of N_{align} to RMSD scales in favour of a longer alignment with minimal RMSD.

The SSM-based Q_{score} was selected for use in this work because during the comparison process it considers the structure, albeit as vectors, instead of reducing it to a distance matrix which in its granularity loses the structural context. Moreover it accounts for the alignment size and lengths of both the proteins compared, which is important because the metric needs to able to quantify the quality of alignment. By considering the number of residues aligned relative to the sizes of the compared proteins, it returns a normalized score. This is important because the structural comparison metrics gauge similarity and therefore to convert the similarity score to a distance an upper bounded value is required. In the case of SSM-based Q_{score}, the similarity was subtracted from "1" to generate a measure of distance.

$$(structural\ distance)\ d = 1 - Q_{score} \tag{1.19}$$

This structural distance is now analogous to the distance between sequences, and hence can be used with distance-based methods for phylogenetic analysis. Any metric that satisfies the basic criteria just listed should be able to replace the SSM-based Q_{score} satisfactorily.

1.8.7.2 *Pairwise protein comparison: Sequence and structure*

Pairwise protein comparison performed using SSM generates a quality score (Q_{score}) which is used as a measure of distance between the structures. This section demonstrates the use of the SSM metric for protein comparison in two cases, one of conserved and the other of divergent proteins.

First, the α and β-haemoglobins from *Anser indicus* are compared with one another as they are known to be the result of a relatively recent gene duplication and divergence event. The SSM metric generates a rotation matrix which is used to superpose the structures shown in Figure 1.14, illustrating the alignment. The sequence alignment, for α and β-haemoglobins in Figure 1.15, shows 34% identity and 54% similarity. This case shows strong sequence similarity equating to strong structural similarity which is captured by the SSM alignment.

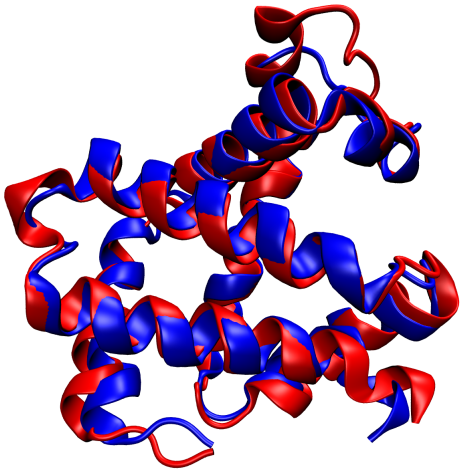


Figure 1.14: Superposition of structure using SSM: α (comprising 141 amino acid residues) and β (comprising 146 amino acid residues) haemoglobins from *Anser indicus* (PDB 1hv4) are superposed using the transformation matrix from SSM. α and β chains are in red and blue respectively. A Q_{score} of 0.63 was achieved, with an RMSD of 1.35 Å over 125 aligned residues.

Next, the two nucleosome-forming histone proteins H3 and H4, from *Homo sapiens*, are compared, see Figures 1.16 and 1.17. In this case, the proteins H3 and H4 are part of the histone family. This case is analogous to the first one, of α and β-haemoglobins, with a small difference i.e. the gene duplication and divergence in case of H3 and H4 is a result of a deeper evolu-

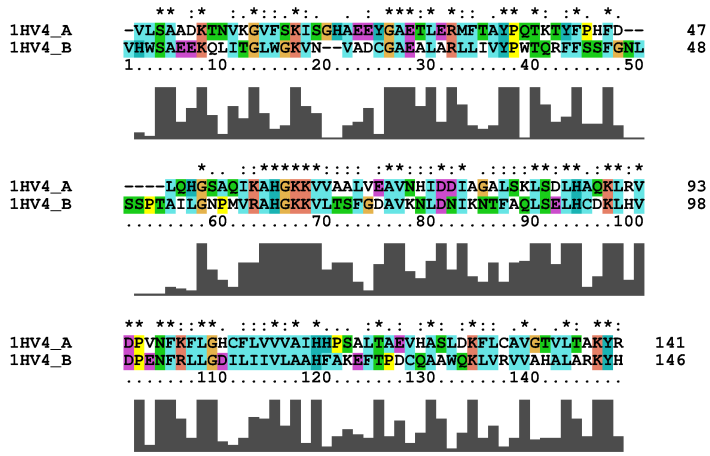


Figure 1.15: Pairwise sequence alignment of α and β-haemoglobins. The alignment shows 34% identity (labelled "*") and 54% similarity (labelled ":" and "."). The similar residues include those labelled identical.

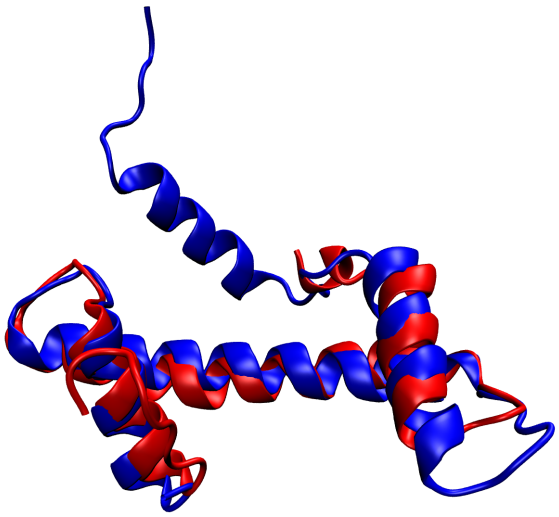


Figure 1.16: Superposed structures of Histone H3 (136 residues) and H4 (103 residues) proteins from *Homo sapiens* (PDB 2cv5). H3 and H4 are in red and blue respectively. A Q_{score} of 0.43 was achieved, with an RMSD of 1.92 Å over 68 aligned residues.

tionary event, relative to the haemoglobin case. Due to their biological role these proteins have been conserved at a structural level but to a lesser extent at a sequence level. This is reflected in the sequence alignment, between H3

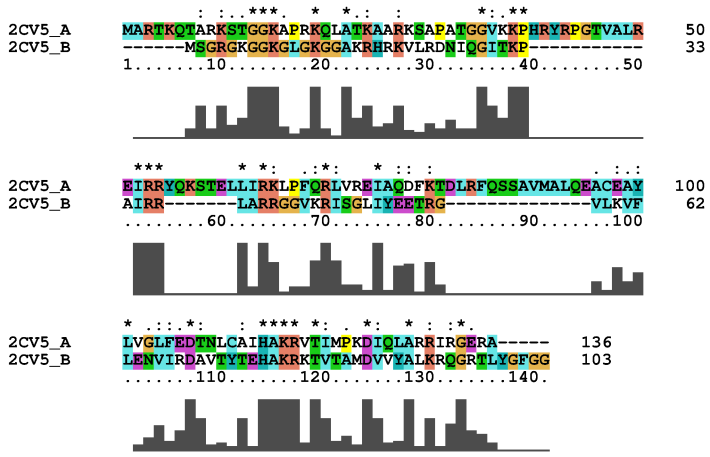


Figure 1.17: Pairwise sequence alignment of H3 and H4 histone proteins. The alignment indicates 23% identity (labelled "*") and 36% similarity (labelled ":" and "."). The similar residues include those labelled identical.

and H4, which illustrates a 23% identity and 36% similarity between these proteins. Due to the sequence alignment scores falling in the "twilight zone" a unified sequence-based phylogenetic analysis has not been attempted for these proteins. However, the conservation in structure can be detected by an SSM-based superposition, Figure 1.16, and reflected in a Q_{score} of 0.43.

SSM has been in use since 2004 and has been tested thoroughly by the authors [89]. It also satisfies the criteria of utilizing structural components, albeit as vectors, instead of reducing them to distances, considering the aligned sequence as well as the lengths of the individual proteins and, finally, generating a normalized score which can be seen equivalent to distance between structures compared. This coupled with the two cases that have been examined here, one where the sequences of proteins were slightly different (i.e. the α and β-haemoglobins, comparison of which is not in the "twilight zone") and the other where they were significantly different (i.e. the H3 and H4 histone proteins) indicates that this metric is a satisfactory choice for generating distances between structures for use in structure-based phylogeny determination, as done previously by Lundin et. al [8].

1.9 Inferential method

The method of inference used in this approach is neighbour-joining. The choice of this method amongst other distance and character-based methods, discussed earlier, is clear. For one, the metric used for structural comparison generates a score which can directly be interpreted as structural distance which justifies the use of a distance method. Secondly, the specific choice of NJ out of the distance methods is for the purpose of convenience as the other reliable method, minimal evolution, requires the use of an optimality criterion which in the case of structures cannot be determined in a straightforward way. Thirdly, character-based methods e.g. in the case of Bayesian methods require the use of an evolutionary model and prior probabilities. These approaches cannot be satisfactorily extended to utilizing structure only, and have been highlighted previously in, Section 1.6.2. Thus, NJ becomes a suitable choice.

Once all the proteins in the structural data set are pairwise compared, the similarity scores are converted to distances. The distances are used to construct a square distance matrix of size $n * n$, where n is the number of structures being analysed. Each value in the matrix, d_{xy}, corresponds to the distance between two structures listed on row x and column y. Using this nomenclature, the following section explains in detail the NJ algorithm [98] and lists all the steps for converting the distances in the matrix to an unrooted phylogenetic tree.

1.9.1 Neighbour-joining: Algorithm summary

For all pairwise comparisons in the distance matrix calculate a quantity L_{xy} such that:

$$L_{xy} = d_{xy} - \frac{r_x - r_y}{n - 2} \tag{1.20}$$

where r_x and r_y are the sum of distances in row x and column y and given by:

$$r_x = \sum_i d_{xi} \tag{1.21}$$

$$r_y = \sum_i d_{yi} \tag{1.22}$$

Once L_{xy} is calculated for all pairwise comparisons, the smallest L_{xy} is chosen (if multiple have same value one is chosen) and structures represented by row x and column y are clustered into an entry z. The node z represents a split grouping the structures clustered. The distance of each structure from the split (x, y) and (y, z) is given by:

$$d_{xz} = \frac{1}{2(n-2)}((n-2)d_{xy} + r_x - r_y) \tag{1.23}$$

$$d_{yz} = \frac{1}{2(n-2)}((n-2)d_{xy} + r_y - r_x) \tag{1.24}$$

z replaces the structures clustered in the distance matrix. The adjusted distances from z to all other members are calculated:

$$d_{iz} = \frac{1}{2}(d_{xi} + d_{yi} - d_{xy}),\ k \neq x, y \tag{1.25}$$

where i iterates over all entries for which an adjusted distance is to be calculated. This completes one cycle. As two structures have been grouped, n is reduced by one. The above steps are repeated until $n = 2$, at which point:

$$L_{\mathrm{xy}} = d_{\mathrm{xy}} \tag{1.26}$$

gives the distance from the last structure.

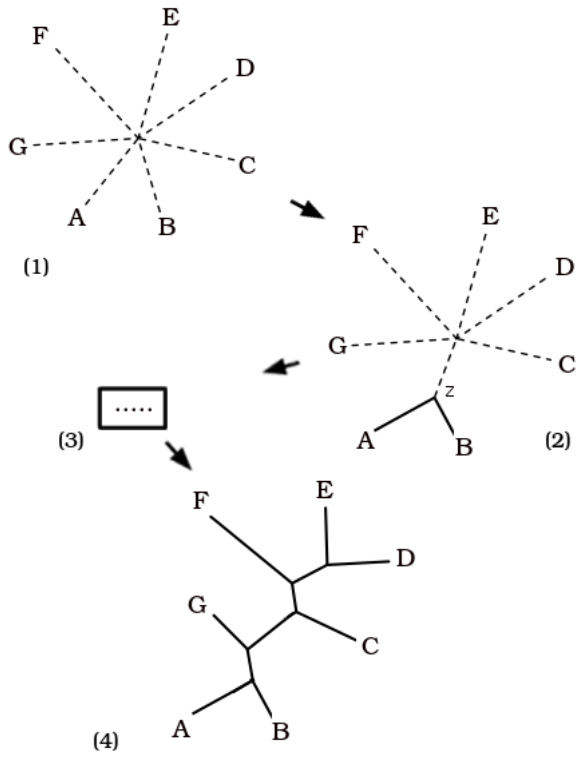


Figure 1.18: NJ tree: The NJ algorithm starts with a star-like tree topology (1) and successively clusters structures. The end of the first cycle is illustrated in (2) which groups A,B into a new node called z. The process is repeated a number of times (3) until the entire tree is resolved (4).

NJ starts with a star topology illustrated in Figure 1.18. In successive iterations of Equations 1.20 through 1.25 clusters are formed such as z in Figure 1.18. When $n = 2$, Equation 1.26 creates the final split resulting in a fully resolved tree, see Figure 1.18.

1.10 Phylogenetic tree

Once the tree is determined using the inferential method explained earlier, the topology of the tree is tested using the MD-based bootstrap method. This section first explains the conventional bootstrap method as used in sequence analysis and then introduces the reader to a MD-based bootstrap method used in this structure-based approach.

1.10.1 Conventional bootstrap

Parametric and non-parametric are two ways in which the bootstrap method is employed. The difference between the two, as discussed earlier, is that in the parametric method, the sequences in the original alignment are modified through simulation using a model of evolution whereas in the non-parametric method the sequence alignment is resampled with replacements. With either method "N" trials are conducted, where each trial starts with generating a data set equivalent in size to the original alignment. A reference tree is made using the original data and the topology of the tree is then tested against the data sets from the trials. The relationships recovered from the trials are counted and expressed as a fraction of the number of trials on the nodes of the reference tree. These fractions reveal the robustness of the relationships reflected by the node. A visual illustration of the non-parametric bootstrap process is presented in Figure 1.19.

1.10.2 Molecular dynamics-based bootstrap method

This structure-based approach of inferring phylogenies could not make use of the bootstrap method as is. The choice of moving away from sequence to structure was to probe the conservation in higher dimensionality of the structure and therefore using the sequence alignment (either informed from the aligned structures or conventional sequence alignment methods) to check the topology of the tree is counterintuitive.

The parametric method could not be employed as an established model for protein structure evolution is missing, at present. The non-parametric method resamples with replacements the columns of a multiple sequence alignment. This cannot be done for structure because, primarily, a method for multiple sequence alignment for structures is not established which is necessary for column-based re-sampling and, secondly, a re-sampled structure may not be real because the amino acids in the new order may not fold in the same way.

A MD-based bootstrap method was formulated, to serve the purpose of the bootstrap method without its impediments for the purpose of structure-based evolutionary analysis. The MD-based bootstrap method is outlined below assuming a reference dataset of M structures for which pairwise comparisons are carried out, and an unrooted NJ reference tree (R_{tree}) is created.

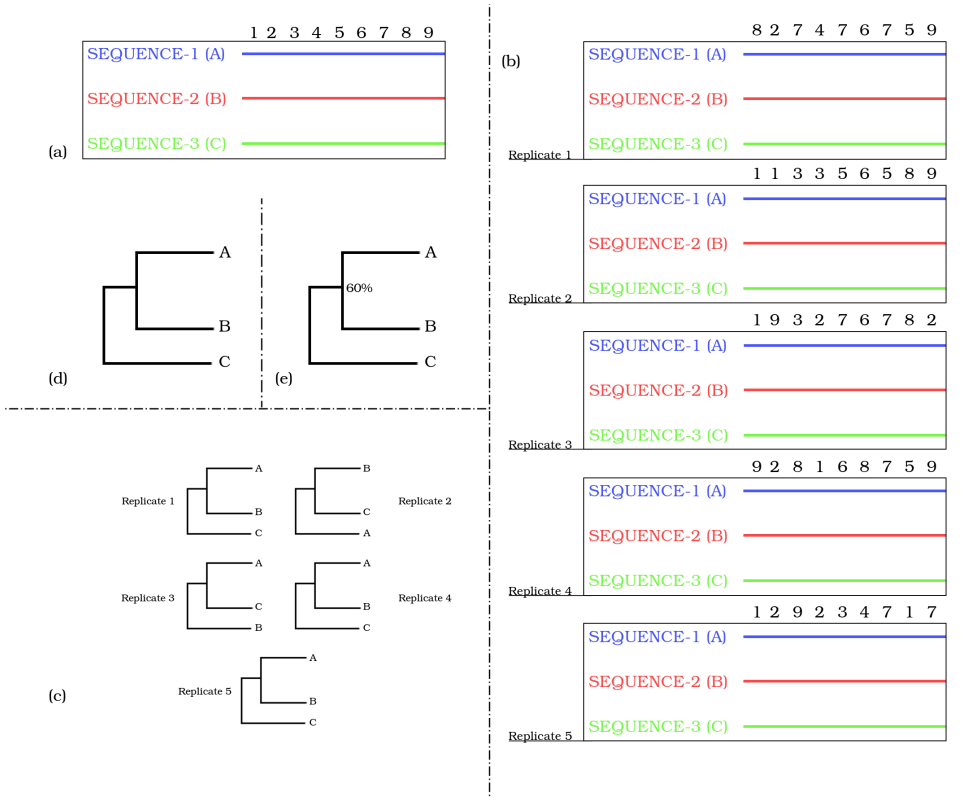


Figure 1.19: The bootstrap method. (a) The original alignment serves as a starting point. The columns comprising the alignment are resampled with replacements (b) to create N trials (only five shown here). For each replicate a tree is generated (c). (d) shows the reference tree generated from the original alignment (a). The relationships in the reference tree are sought in the replicate trees and expressed as a ratio of the total number of trials (e). The reference tree topology is recovered in three of the five replicates resulting in 60% support.

The bootstrap-like analysis is as follows.

- For each of the M structures, a MD simulation, described in section 1.11, is performed and conformations recorded in trajectories, see Figure 1.20.
- For one trial, a conformation is randomly selected from each of the M trajectories to populate a new replicate dataset.
- The previous step is repeated N times to create N replicates for N trials.
- In each replicate, pairwise comparison generates new distances against

which NJ trees are created, T_i, where i is the replicate number e.g. T_{130} would refer to the tree from the *130*th replicate trial.

- Each replicate tree, T_i, is compared to the reference tree R_{tree}. If the relationship between structures in the reference tree R_{tree} are recreated in the replicates, T_i, they are counted.
- After all replicate trees have been compared to the reference tree, counted relationships, in the previous step, are expressed as a ratio of the number of trials, N, on their respective nodes, in the same way as shown in Figure 1.19.

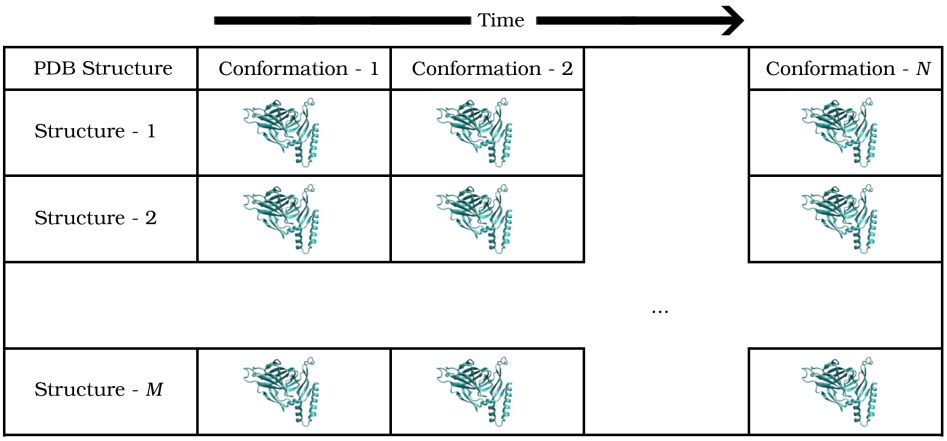


Figure 1.20: MD trajectories. MD simulations are carried out for each of the M structures in the dataset and N conformations are recorded.

1.11 Molecular dynamics for conformational sampling

MD simulations were used to generate conformations used in the MD-based bootstrap method to test the topology of the phylogenetic tree. This section briefly explains the MD simulation method in general, outlines certain specific system parameters and provides a canonical procedure in which these simulations can be conducted.

1.11.1 Molecular dynamics summary

MD can be defined as a simulation method that uses the classical Newtonian equations of motion to determine particle positions at discrete intervals of time. The particles themselves either represent individual atoms or groups of atoms depending on the resolution of the system. The particles are assigned velocities and their motions observed over a certain period of time. This motion is subsequently used to recover system properties that would other wise be difficult to recover using a non-computational experimental approach. In this work, these simulations allowed for alternative conformations to be captured for each of the protein structures for use with the MD-based bootstrap method.

1.11.2 System representation

Physical systems that are probed with MD simulations are usually represented at either atomic or lower, "coarse-grained" resolution, so that each particle represents either an individual atom or a group of atoms.

Coarse-graining compromises the level of detail and potentially also the accuracy of the results in favour of computational speed, as the grouping of atoms significantly reduces the number of pairwise calculations during a simulation. For this work atomic systems are considered, i.e. each particle models a single atom.

Particles in the MD simulations are arranged into bonded and non-bonded groups. In MD simulations the interactions between particles are governed by a force field. This is further expanded below.

1.11.3 Force fields

The chemical properties of the atoms during a simulation are governed by a set of potential energy functions that give rise to physical forces that the atoms experience during the simulation. These include short range repulsions, intermediate range dispersion interactions and electrostatic interactions between non-bonded atoms. There are also terms to describe bond lengths, bond angles and torsional rotation around bonds for bonded atoms.

These collections of potential energy functions are known as "force fields". Several version of force fields are widely available and used for biomolecular simulations including CHARMM [99], AMBER [100], GROMOS [101],

OPLS [102]. In the CHARMM36FF force field, used for this work, the total potential energy is the sum of the potential energies of both the bonded and non-bonded groups and is given by:

$$U_{CHARMM} = U_{bonded} + U_{non-bonded} \tag{1.27}$$

U_{bonded} includes:

$$\begin{aligned} \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{UB} K_{UB}(b^{1-3} - b_0^{1-3})^2 \\ + \sum_{dihedrals} K_\varphi(1 + \cos(\eta\varphi - \delta)) + \sum_{impropers} K_\omega(\omega - \omega_0)^2 \\ + \sum_{residues} U_{CMAP} \end{aligned} \tag{1.28}$$

where K are force constants, b_0, θ_0 and ω_0 are equilibrium values for bond lengths, angles and improper torsional angles, and b_0^{1-3} is the equilibrium bond length between any two atoms connected by two bonds, η and δ are dihedral multiplicity and phase shifts. The U_{CMAP} term returns the contributions from Φ and Ψ backbone angles. In each term the current values b, θ, b^{1-3}, ω, φ and for U_{CMAP} the Φ and Ψ terms are generated from the simulation whereas equilibrium values are determined from experiments and provided by the force field.

$U_{non-bonded}$ includes:

$$\sum_{nbLJ} \varepsilon_{ij}((\frac{r_{ij}^{min}}{r_{ij}})^{12} - 2(\frac{r_{ij}^{min}}{r_{ij}})^{6}) + \sum_{nbElec} \frac{q_i q_j}{\epsilon r_{ij}} \tag{1.29}$$

where r_{ij} refers the distance between two non-bonded atoms, r_{ij}^{min} is the most favourable distance between two atoms, ε_{ij} to the lowest interatomic

energy, q_i and q_j to the charges of the two atoms and ϵ to the dielectric permittivity. From amongst these terms r_{ij} is determined in runtime from the simulation, where the remaining parameters are provided by the force field. These terms are illustrated in Figure 1.21.

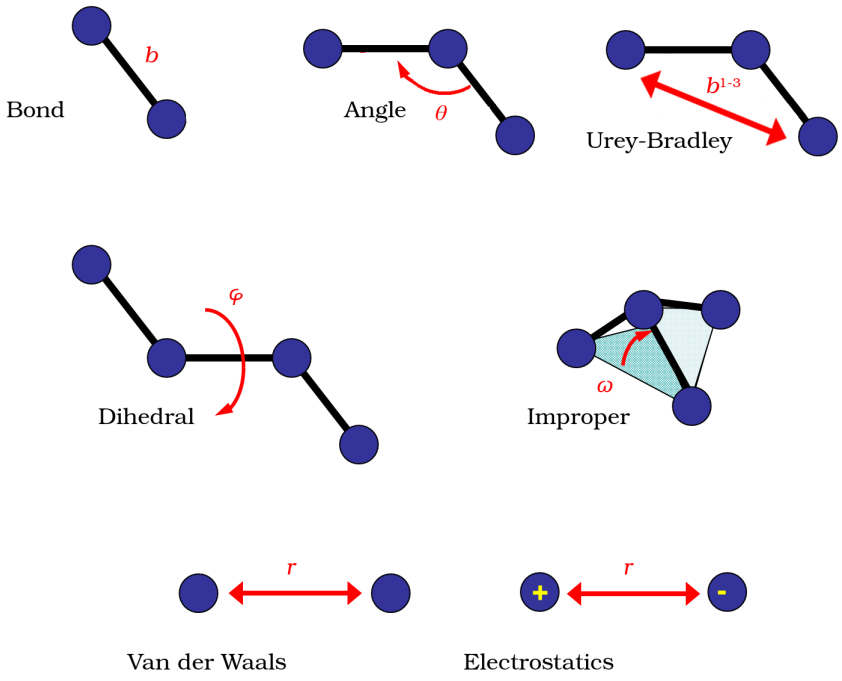


Figure 1.21: Force field terms. Illustration of the bonded and non-bonded atomic interactions in Equations 1.28 and 1.29.

1.11.4 System considerations

The purpose of using the MD simulation method is to reproduce the behaviour of a system of interest. For the subsequent analysis from the recorded observations to be meaningful, the system needs to be simulated in a manner which is close to reality. As an example, a simulation of a protein carried out at 373 K will not reproduce the behaviour at physiological temperature and therefore must be simulated at 310 K for the observations to be applicable. This is a simple example illustrating the need to control physical variables for the simulation to produce useful results. Similarly, the protein samples conformations from a statistical mechanical ensemble which

requires consideration of certain properties. This section briefly summarizes some important aspects which require consideration when attempting to realize results from MD simulations of a particular system.

1.11.4.1 *Statistical ensemble*

A number of statistical ensembles are available, each of which maintains certain properties of the system while allowing others to fluctuate. Two ensembles, the canonical (NVT) and the isothermal-isobaric (NPT) ensemble, are popularly used for biological systems, as they (especially NPT) mimic "real life". In these ensembles either the volume, V in NVT, or pressure, P in NPT, is kept constant along with the number of atoms (N) and temperature (T). Coupling the system with a barostat and thermostat allows for pressure and temperature to be kept constant.

1.11.4.2 *Simulation environment*

To reproduce results from simulation of biological systems which are close to reality, the system of interest is subject to conditions which approximate its native environment. For proteins this is usually the cellular environment, where the protein is, mostly, surrounded by water. It is therefore important to solvate the protein, using water as a solvent.

Apart from solvent, proteins and other cellular molecules experience an environment which has a certain pH and is ionized. For the simulation to produce accurate results, these native conditions must be replicated. In protein-based systems, the pH is introduced in the form of protonated amino acids. At a neutral pH (7.0), arginine, lysine and histidine carry positive charges whereas aspartate and glutamate are negatively charged. To change the ionic strength in the solvent around the protein sodium (Na^+) and chloride (Cl^-) ions are introduced. In addition, ions may be used to neutralise the unbalanced charge of a protein which is necessary when using particle mesh ewald (PME) summations for accurate long-range electrostatic measurements.

1.11.4.3 *Boundary conditions*

The computational cost associated with simulating a system is directly dependent on the number of particles it contains. An increase in (N) expo-

nentially increases the time required for the simulation to complete. However, as long-range interactions play crucial roles in certain systems, a large (N) is necessary to produce realistic results. Another problem that arises from simulating closed systems is edge effects, i.e. how particles behave at system boundaries. To bypass this periodic boundary conditions are used to replicate bulk-like environments with a limited number of particles (N) and mitigate the edge effects. Other boundary conditions are also available, however due to the ability of periodic boundaries to capture bulk-like effects, others are rarely employed.

1.11.5 Molecular dynamics: Method breakdown

1.11.5.1 *Energy minimization*

A number of experimental techniques, e.g. X-ray crystallography or solution NMR (nuclear magnetic resonance), are employed to determine the structure of biomolecules e.g. proteins. These techniques generate a structure which may not be optimized according to the force field to be used with the MD simulations. Energy minimization is, therefore, usually carried out prior to simulating a molecule, to optimize the geometry according to a certain force field.

The aim of this routine is to find a geometry which sits at a minimum on the potential energy surface. A number of minimization routines are available e.g. steepest descent [103], adopted basis Newton-Raphson method (ABNR) [103] and conjugate gradient [103, 104].

1.11.5.2 *Molecular dynamics: Method breakdown*

The MD method can be decomposed into the following steps.

1. A starting structure provides reference coordinates of all the atoms in the system. The starting structure is static i.e. has no thermal energy and hence no motion. The particles are initialized by assigning velocities from a distribution of choice such that the total linear momentum of the system is zero and the system attains a desired temperature e.g. 310 K.

2. Once the velocities are assigned, the atoms begin to move and therefore change coordinates with time. The equations of motion are used to obtain the new coordinates at discrete intervals of time. Although multiple ways of solving Newtonian equations exist, to obtain the new coordinates, velocity-verlet is a popular choice. This method has two steps.

   Step 1:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \tag{1.30}$$

$$v(t+\frac{\delta t}{2}) = v(t) + \frac{1}{2}a(t)\delta t \tag{1.31}$$

   Step 2:

$$a(t+\delta t) = -\frac{1}{m}\nabla_r U[r(t+\delta t)] \tag{1.32}$$

$$v(t+\delta t) = v(t+\frac{\delta t}{2}) + \frac{1}{2}a(t+\delta t)\delta t \tag{1.33}$$

   Position, r, and velocity, v, are updated first, Equations 1.30 and 1.31. Whereas positions are updated for a full time step δt, velocity is updated for a half time step $\delta t/2$. The initial acceleration, $a(t)$, is calculated in the same way as shown in step 2, Equation 1.32, using positions at time t. The new positions, $r(t+\delta t)$, are used to update acceleration, Equation 1.32. For this the potential energy, U, is calculated according to the force field, Section 1.11.3, and the derivative of the potential energy provides the force acting on an atom of mass m. This is followed by the final recovery of velocities, Equation 1.33, at $t+\delta t$. The terms *r(t)* and *v(t)* are position and velocity at time t. The choice of the time step δt is important so as to balance running a long simulation with energy conservation. A common choice for this variable is 2 fs, as it is the largest time step which can be used while still conserving energy.

3. Step two is repeated, in δt time step increments, to reproduce the behaviour of the system for a period of time, e.g. 10 ns. A consensus does not exist on the length of the simulation. A usual practice is to

observe the system for a certain length of time until a desired property equilibrates, also called the equilibration period. Once this is achieved, the simulation is continued, in a production period, from which the data is used for analysis. Under the premise of "more is better", the simulation can be extended in the production period to any length, subject to the computational resources available.

1.11.6 MD simulation: An example

MD simulations are carried out by highly developed large software programs like NAMD [105], CHARMM [103], AMBER [100] and GROMACS [106] etc. Each program can make use of either its own force field, Section 1.11.3, or in certain cases, namely NAMD and GROMACS, other force fields. To explain the simulation process, a conventional MD routine, as used in this work, is outlined, for the NAMD simulation program using the CHARMM force field, in Figure 1.22 and discussed below.

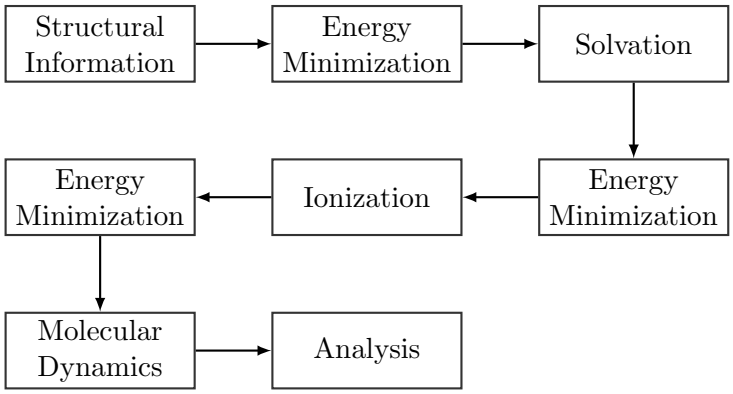


Figure 1.22: A conventional MD routine. See text for details on each stage.

The NAMD program requires two **structural information** input files, a PDB (or coordinate file) and a PSF file. The PDB file holds Cartesian coordinates for each of the atoms contained within the system. A PSF file holds structural information other than the coordinates e.g. atom types and charges and reference codes for extracting relevent force field parameters from separate force field files.

The starting structure is usually determined using X-ray crystallography or solution NMR. The first **energy minimization** step makes this

experimentally determined structure consistent with the force field. The subsequent minimization steps are carried out every time the number or nature of particles in the system is changed. This is to remove any steric clashes which might be introduced as a result of alterations to the system.

The **solvation** step adds a relevant solvent around the protein which is mostly water. Although multiple water models exist TIP3P [107] is used in this work, as it is compatible with the CHARMM force field.

This method employs the PME (particle mesh ewald) [108] method for calculation of long-range electrostatic interactions which requires the system to be neutral. **Ionization** counteracts the unbalanced charges from the amino acids of the protein by adding positive (Na^+) or negative ions (Cl^-).

The **MD** stage has been covered, in detail, at the start of this section. In the context of conducting the simulation some important parameters are set. These include the temperature at which the system is to be simulated e.g. 310 K, the type of thermostat and barostat used, periodic boundries set up to include the protein and the solvent around it and the number of integration steps (each of δt time step) for which the system is to be observed.

The **analysis** is usually specific to the type of system probed and problem investigated. A MD simulation allows for the protein or any other system of interest to sample conformations from an energy landscape, which are recorded in trajectories. This landscape is theoretically modelled in Figure 1.23. These conformations represent the range of structures accessible to the system as a result of thermal fluctuations. In the context of the MD-based bootstrap method, these conformations are used to build replicate trees to measure the robustness of the phylogenetic relationships given by the reference tree.

1.12 Summary

This chapter presented, in detail, the individual steps involved in generating phylogenies using protein structure. These are:

- Pairwise structural comparison of protein structures.
- Conversion of structural similarity to distance between structures.

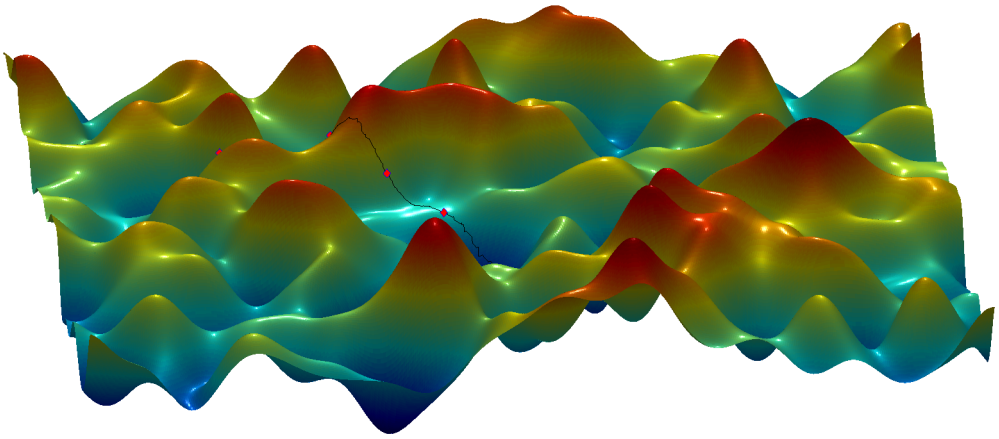


Figure 1.23: Illustration of a conformational energy landscape: MD simulations sample conformations from a landscape similar to this. The troughs and crests represent the local minimum and maximum energy states. Each point on the surface represents a conformation. The black line traces a possible path a MD simulation may take with the red diamonds indicating conformational snapshots recorded for subsequent analysis.

- Use of the NJ method to build and unrooted phylogenetic tree from the distances between structures.
- Use of MD simulations to generate alternate conformations of each structure in the dataset.
- Use of the MD-based bootstrap method to generate a measure of robustness for the relationships shown in the phylogenetic tree.

At each step, methods are compared and contrasted to justify the choices of algorithms used in this approach.

The work included in the following chapters tests weather a suitable metric can be used to extract a signal from empirical protein structure comparison and, moreover, weather that can be used to infer evolutionary relationships between compared proteins.

Chapter 2, breaks down the structural comparison metric and explores its effects on the comparison of protein structures and how the results can be interpreted.

Chapter 3, applies the metric to a model protein structural dataset to recover known relationships. The aim of this chapter is to raise confidence in the method and assess its capability outside test datasets.

Chapter 4, presents, using histones, an example of the types of protein structural families that can be explored using the method in this work. Although the complete method is not applied on this structural dataset, due to time constraints, it presents a coherant example highlighting data collation and interpretation of structure-based phylogenies.

Chapter 5, shows the complete use of the MD-based bootstrap-like method to assess the validity of the previously determined structural phylogeny of the ferritin-like superfamily. The aim of this chapter is to depict that the automated MD-based bootstrap-like method can act as a suitable replacement for careful manual phylogeny assessment.

Bibliography

[1] Alberts, B. *Molecular Biology of the Cell.* Garland Science, 2017.

[2] Woese, C. R. and Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[3] Estes, S. and Arnold, S. J. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. *The American Naturalist*, 169(2):227–244, 2007.

[4] Holm, L. and Sander, C. Mapping the protein universe. *Science*, 273 (5275):595–602, 1996.

[5] Rost, B. Protein structures sustain evolutionary drift. *Folding and Design*, 2:S19–S24, 1997.

[6] Panchenko, A. R., Wolf, Y. I., Panchenko, L. A., and Madej, T. Evolutionary plasticity of protein families: coupling between sequence and structure variation. *Proteins: Structure, Function, and Bioinformatics*, 61(3):535–544, 2005.

[7] Illergård, K., Ardell, D. H., and Elofsson, A. Structure is three to ten times more conserved than sequence — a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

[8] Lundin, D., Poole, A. M., Sjöberg, B.-M., and Högbom, M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *Journal of Biological Chemistry*, 287(24):20565–20575, 2012.

[9] Dill, K. A. and MacCallum, J. L. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[10] Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bryant, S. H., Canese, K., and Church, D. M. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 42(Database Issue):D7, 2014.

[11] Rost, B. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

[12] Fiser, A. Comparative protein structure modelling. In *Protein Structure to Function with Bioinformatics*, pages 91–134. Springer, 2017.

[13] Pearson, W. R. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, pages 1–3, 2013.

[14] Webb, B. and Sali, A. Protein structure modeling with MODELLER. *Protein Structure Prediction*, pages 1–15, 2014.

[15] Vogt, G., Etzold, T., and Argos, P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *Journal of Molecular Biology*, 249(4):816–831, 1995.

[16] Yona, G. and Levitt, M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology*, 315(5):1257–1275, 2002.

[17] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., and Sangrador-Vegas, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279 – D285, 2016.

[18] Dill, K. A. Dominant forces in protein folding. *Biochemistry*, 29(31): 7133–7155, 1990.

[19] Berg, J. M., Tymoczko, J. L., and Stryer, L. Biochemistry, 7th edn WH Freeman, New York; dt.(2007) Biochemie, 5, 2011.

[20] Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 1983.

[21] Balding, D. J., Bishop, M., and Cannings, C. *Handbook of Statistical Genetics*. John Wiley & Sons, 2008.

[22] Cuff, M. E., Xu, X., Cui, H., Edwards, A., Savchenko, A., and Joachimiak, A. PDB ID: 3qvo, Structure of a Rossmann-fold {NAD}(P)-binding family protein from Shigella flexneri. 2011.

[23] Pollock, D. D. and Taylor, W. R. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering*, 10(6):647–657, 1997.

[24] Bastolla, U., Dehouck, Y., and Echave, J. What evolution tells us about protein physics, and protein physics tells us about evolution. *Current Opinion in Structural Biology*, 42:59–66, 2017.

[25] Ponting, C. P. and Russell, R. R. The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, 31(1):45–71, 2002.

[26] Coulson, A. F. W. and Moult, J. A unifold, mesofold, and superfold model of protein fold use. *Proteins: Structure, Function, and Bioinformatics*, 46(1):61–71, 2002.

[27] Sousounis, K., Haney, C. E., Cao, J., Sunchu, B., and Tsonis, P. A. Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Human Genomics*, 6(1):10, 2012.

[28] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. In *International Tables for Crystallography Volume F: Crystallography of Biological Macromolecules*, pages 675–684. Springer, 2006.

[29] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[30] Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., and Durbin, R. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl_1):D247 – D251, 2006.

[31] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., and Mistry, J. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1): D222 – D230, 2013.

[32] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4): 536–540, 1995.

[33] Ye, Y. and Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, 32(suppl_2):W582 – W585, 2004.

[34] Shindyalov, I. N. and Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.

[35] Ortiz, A. R., Strauss, C. E. M., and Olmea, O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Science*, 11(11):2606–2621, 2002.

[36] Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

[37] Tatusova, T. A. and Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, 174(2):247–250, 1999.

[38] Consortium, U. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158 – D169, 2017.

[39] Chandonia, J.-M., Fox, N. K., and Brenner, S. E. SCOPe: Manual curation and artifact removal in the structural classification of proteins–extended Database. *Journal of Molecular Biology*, 429(3): 348–355, 2017.

[40] Dawson, N. L., Sillitoe, I., Lees, J. G., Lam, S. D., and Orengo, C. A. CATH-Gene3D: Generation of the resource and its use in obtaining structural and functional annotations for protein sequences.

Protein Bioinformatics: From Protein Modifications and Networks to Proteomics, pages 79–110, 2017.

[41] Taylor, W. R. and Orengo, C. A. Protein structure alignment. *Journal of Molecular Biology*, 208(1):1–22, 1989.

[42] Hadley, C. and Jones, D. T. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, 7(9): 1099–1112, 1999.

[43] Day, R., Beck, D. A. C., Armen, R. S., and Daggett, V. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Science*, 12(10):2150–2160, 2003.

[44] Csaba, G., Birzele, F., and Zimmer, R. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Structural Biology*, 9(1):23, 2009.

[45] Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[46] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[47] Jukes, T. H., Cantor, C. R., and Munro, H. N. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3(21):132, 1969.

[48] Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[49] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[50] Hasegawa, M., Kishino, H., and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[51] Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17 (2):57–86, 1986.

[52] Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. 22 A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. National Biomedical Research Foundation Silver Spring, MD, 1978.

[53] Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89 (22):10915–10919, 1992.

[54] Notredame, C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):131–144, 2002.

[55] Notredame, C., Higgins, D. G., and Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, 2000.

[56] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[57] Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2):330–340, 2005.

[58] Notredame, C. and Higgins, D. G. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Research*, 24(8):1515–1524, 1996.

[59] Kim, J., Pramanik, S., and Chung, M. J. Multiple sequence alignment using simulated annealing. *Bioinformatics*, 10(4):419–426, 1994.

[60] Wang, G. and Dunbrack, R. L. Scoring profile-to-profile sequence alignments. *Protein Science*, 13(6):1612–1626, 2004.

[61] McClure, M. A., Vasi, T. K., and Fitch, W. M. Comparative analysis of multiple protein-sequence alignment methods. *Molecular Biology and Evolution*, 11(4):571–592, 1994.

[62] Baxevanis, A. D. and Ouellette, B. F. F. *Bioinformatics: a practical guide to the analysis of genes and proteins*, volume 43. John Wiley & Sons, 2004.

[63] Edgar, R. C. and Batzoglou, S. Multiple sequence alignment. *Current Opinion in Structural Biology*, 16(3):368–373, 2006.

[64] Metcalf, B., Chuang, C., Dufu, K., Patel, M. P., Silva-Garcia, A., Johnson, C., Lu, Q., Partridge, J. R., Patskovska, L., Patskovsky, Y., Almo, S. C., Jacobson, M. P., Hua, L., Xu, Q., Gwaltney, S. L., Yee, C., Harris, J., Morgan, B. P., James, J., Xu, D., Hutchaleelaha, A., Paulvannan, K., Oksenberg, D., and Li, Z. Discovery of GBT440, an orally bioavailable R-state stabilizer of sickle cell hemoglobin. {*ACS*} *Medicinal Chemistry Letters*, 8(3):321–326, 2017.

[65] Tame, J. R. H. and Vallone, B. The structures of deoxy human haemoglobin and the mutant Hb Tyrα42His at 120 K. *Acta Crystallographica Section D: Biological Crystallography*, 56(7):805–811, 2000.

[66] Malkowski, M. G., Quartley, E., Friedman, A. E., Babulski, J., Kon, Y., Wolfley, J., Said, M., Luft, J. R., Phizicky, E. M., and DeTitta, G. T. Blocking S-adenosylmethionine synthesis in yeast allows selenomethionine incorporation and multiwavelength anomalous dispersion phasing. *Proceedings of the National Academy of Sciences*, 104 (16):6678–6683, 2007.

[67] Seiradake, E., Mao, W., Hernandez, V., Baker, S. J., Plattner, J. J., Alley, M., and Cusack, S. Crystal structures of the human and fungal cytosolic Leucyl-tRNA synthetase editing domains: a structural basis for the rational design of antifungal benzoxaboroles. *Journal of molecular biology*, 390(38):29502–29510, 2009.

[68] Delagoutte, B., Moras, D., and Cavarelli, J. tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. *The EMBO Journal*, 19(21):5599–5610, 2000.

[69] Steel, M. A., Hendy, M. D., and Penny, D. Loss of information in genetic distances. *Nature*, 336(6195):118, 1988.

[70] Huelsenbeck, J. P. Is the Felsenstein zone a fly trap? *Systematic Biology*, 46(1):69–74, 1997.

[71] Felsenstein, J. The number of evolutionary trees. *Systematic Zoology*, 27(1):27–33, 1978.

[72] Sneath, A. and Sokal, R. R. Principles of numerical taxonomy. *San Francisco and London I*, 963, 1963.

[73] Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[74] Rzhetsky, A. and Nei, M. A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution*, 9(5):945, 1992.

[75] Cavalli-Sforza, L. L. and Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570, 1967.

[76] Fitch, W. M. and Margoliash, E. Construction of phylogenetic trees. *Science*, 155(3760):279–284, 1967.

[77] Felsenstein, J. An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46(1):101–111, 1997.

[78] Farris, J. S. Methods for computing Wagner trees. *Systematic Biology*, 19(1):83–92, 1970.

[79] Fitch, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.

[80] Quenouille, M. H. Notes on bias in estimation. *Biometrika*, 43(3/4): 353–360, 1956.

[81] Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer, 1992.

[82] Felsenstein, J. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

[83] Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.

[84] Cvicek, V., Goddard III, W. A., and Abrol, R. Structure-based sequence alignment of the transmembrane domains of all human GPCRs: Phylogenetic, structural and functional implications. *PLoS Computational Biology*, 12(3):e1004805, 2016.

[85] Herman, J. L., Challis, C. J., Novák, Á., Hein, J., and Schmidler, S. C. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Molecular Biology and Evolution*, 31(9):2251–2266, 2014.

[86] O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2):385–395, 2004.

[87] Huelsenbeck, J. P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[88] Challis, C. J. and Schmidler, S. C. A stochastic evolutionary model for protein structure alignment and phylogeny. *Molecular Biology and Evolution*, 29(11):3575–3587, 2012.

[89] Krissinel, E. and Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12): 2256–2268, 2004.

[90] Rice, D. W. and Eisenberg, D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *Journal of Molecular Biology*, 267(4):1026–1038, 1997.

[91] Goonesekere, N. C. W. and Lee, B. Context-specific amino acid substitution matrices and their use in the detection of protein homologs. *Proteins: Structure, Function, and Bioinformatics*, 71(2):910–919, 2008.

[92] Holm, L. and Sander, C. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20(11):478–480, 1995.

[93] Gibrat, J. F., Madej, T., Spouge, J. L., and Bryant, S. H. The VAST protein structure comparison method. In *Biophysical Journal*, vol-

ume 72, pages MP298 – MP298. Biophysical Society 9650 Rockville Pike, Bethesda, MD 20814-3998, 1997.

[94] Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[95] Holm, L. and Park, J. DaliLite workbench for protein structure comparison. *Bioinformatics*, 16(6):566–567, 2000.

[96] Holm, L. and Sander, C. Dictionary of recurrent domains in protein structures. *Proteins: Structure, Function and Genetics*, 33(1):88–96, 1998.

[97] Hutchinson, E. G. and Thornton, J. M. PROMOTIF - a program to identify and analyze structural motifs in proteins. *Protein Science*, 5 (2):212 – 220, 1996.

[98] Haubold, B. and Wiehe, T. *Introduction to computational biology: an evolutionary approach*. Springer Science & Business Media, 2006.

[99] Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, 2017.

[100] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, 1995.

[101] van Gunsteren, W. F., Daura, X., and Mark, A. E. GROMOS force field. *Encyclopedia of Computational Chemistry*, 1998.

[102] Jorgensen, W. L. OPLS force fields. *Encyclopedia of Computational Chemistry*, 1998.

[103] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. a., and Karplus, M. CHARMM: a program for macro-

molecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[104] Fletcher, R. and Reeves, C. M. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 1964.

[105] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[106] Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., and van der Spoel, D. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[107] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[108] Darden, T., York, D., and Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

Chapter 2

Method Development

2.1 Overview

The structure-based phylogenetic method developed and applied in this thesis uses the secondary structure matching (SSM)-based Q_{score} to quantify the degree of similarity between structures, and molecular dynamics (MD) simulations to sample alternative conformations of the compared protein structures to add a measure of significance to the inferred phylogenies. This chapter mechanistically explores the effect of the Q_{score} and of MD simulations on the success and reliability of structure-based phylogenetics, and outlines their limitations. Firstly, the Q_{score} is decomposed into two parts, one that considers the sizes of the proteins compared and the other that considers the morphometric contributions. Both these parts are explored through the use of control datasets to see their respective behaviours. Secondly, again through the use of control protein structural datasets, the MD-based bootstrap method is explored. In summary, the chapter aims to offer insight into the effect of these individual aspects of the structural phylogeny method on the nature and reliability of the evolutionary relationships inferred.

2.2 Secondary structure matching-based Q_{score}

The comparison between protein structures, amongst which an evolutionary relationship is to be delineated, is carried out through the SSM-based Q_{score} metric. The algorithm underlying the calculation of the structural comparison score, Equation 2.1, has previously been discussed at length, Chapter 1, Section 1.8.7.

$$Q_{score} = \frac{N_{align}^2}{[1 + (\frac{RMSD}{R_0})^2]N_1N_2} \tag{2.1}$$

While some discussion comparing the Q_{score} metric to other structural comparison metrics has been outlined earlier; in summary, the choice of the Q_{score} metric is primarily because of the inclusion of both size and structural variations to generate a normalised structural comparison score. The Q_{score} metric can therefore be divided into parts which account for these aspects, i.e. size, Equation 2.2, and shape, Equation 2.3.

$$(Part\ 1 : Size\ factor)_{Q_{score}} = \frac{N_{align}^2}{N_1 N_2} \tag{2.2}$$

$$(Part\ 2 : Shape\ factor)_{Q_{score}} = \frac{1}{[1 + (\frac{RMSD}{R_0})^2]} \tag{2.3}$$

As can be seen from Equation 2.1, the length of the alignment N_{align} and individual lengths of the proteins N_1 and N_2 scale the morphometric contribution, i.e. from the use of root mean square deviation (RMSD) between protein residues aligned. To assess contributions from each aspect separately, the effect of the other is ignored, see Section 2.3 for details on how this is achieved.

The second aspect of the structure-based phylogenetic method developed in this thesis is to associate a measure of robustness to the inferred phylogenies. This is achieved through use of alternative conformations, sampled using MD simulations, for each of the structures used in the analysis. The exploration around this aspect uncovers the mechanism through which a measure of significance is associated to the inferred relationships and elucidates a possible limitation of this method when analysing structural data spread over different evolutionary time scales.

2.3 Method

2.3.1 *Part* 1: The size effect

For a comprehensive analysis of the size contributions towards the Q_{score}, a broad size survey of all the protein structures in the RCSB (www.rcsb.org) structural database was conducted. Protein size, in the context of this work, refers to the length of the protein sequence, i.e. the number of amino acids in the protein. The distribution of these sizes is shown in Figure 2.1.

K-means clustering was used in the most dense part of the distribution, comprising 150,000 protein structures, to recover three clusters. These were centred on protein sequence lengths of 125 amino acids, 184 amino acids and

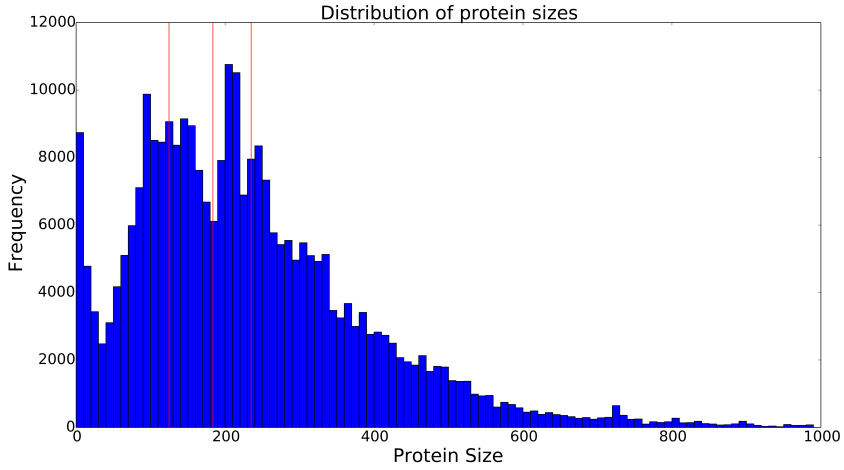


Figure 2.1: Protein size distribution for all proteins present in the RCSB database (structural data extracted on: 01-Nov, 2017). Three clusters were identified in the most dense part of the distribution using k-means clustering. The centroids of each cluster are shown with red vertical lines at protein sizes of 125, 184 and 234 amino acids.

229 amino acids, respectively. These centres then provided the size values on which size contributions could be robustly tested.

This analysis aimed at probing the contribution of the sizes towards recreating evolutionary histories. Therefore, to ensure only those protein structures which shared an evolutionary history were placed on a phylogenetic tree, the SCOP structural database was used. The SCOP structural database groups protein structures into families based on a shared evolutionary history. All the protein families were therefore probed for those which had 20 protein structures whose sizes were either close to or exactly the same as the cluster means. The quantity 20 was chosen for convenience purposes only.

Cytochrome C and globins were found to have structures distributed around 125 and 184 amino acids and hence were chosen to represent these cluster means. A protein family with at least 20 structures with sizes distributed around 234 amino acid residues was not found and hence the closest family at 225 amino acids, namely the ferritin family, was selected instead. 20 structures each were selected from these three families which, along with their respective sizes, are listed in Table 2.1.

Table 2.1: Protein structures, their chain identifiers and respective sizes of the three protein families, ferritins, globins and cytochromes used to test the contribution of the size factor to the overall Q_{score}.

Ferritins			**Globins**			**Cytochrome C**		
PDB	**Chain**	**Size**	**PDB**	**Chain**	**Size**	**PDB**	**Chain**	**Size**
3A9Q	I	193	3V57	B	177	1COT	A	121
4KVR	A	222	1XG0	C	174	1CXA	A	124
4QUW	A	223	4ILZ	B	137	1CXC	A	124
3PW8	A	247	3QM5	A	145	1DT1	A	129
4KVS	A	222	4FH6	A	184	1E8E	A	124
2VUX	B	255	3LB1	B	137	1F1C	A	129
3PWQ	K	234	3QM6	A	145	1FOC	A	128
4PGK	A	221	4FH7	A	137	1FOC	B	128
4PG0	A	221	3LB4	A	137	1GU2	A	124
4RC7	A	220	3V58	B	177	1GU2	B	123
4RC7	B	222	4O35	A	145	1L9B	C	124
4RC8	A	222	4LM6	B	175	1L9J	C	124
4RC8	B	222	4FH6	B	137	1L9J	D	124
4PG1	A	222	3O7N	A	137	1OAE	A	124
4KVQ	A	223	4DWT	A	137	1OAE	B	124
4TW3	A	232	4DWU	A	137	2BH4	X	121
4RC5	A	221	3V58	D	177	2CXB	A	123
4RC5	B	222	4LM6	D	177	2CXB	B	123
3PW8	B	247	3QM9	A	145	2FWL	A	129
4RC6	A	215	3V57	D	177	4YE1	A	122

To remove the influence of shape, i.e. by setting Equation 2.3 to 1, while only the size contributions, Equation 2.2, were explored, the following was done.

- Each protein structure in the analysis was decomposed into 10 fractions, see Appendix-I for an illustration.
- Each fractional part comprised the first $n\%$ * N amino acids of the protein, starting from the most N-terminal residue present in the structure, where $\{n \in 10, 20, ..., 100\%\}$ and N was the number of residues in the structure.
- The label of each protein fraction was preceded by n and was coloured red for easy visualization. The labels of the complete counterparts were coloured black, see Figures 2.2, 2.3 and 2.4. For example, for a multi-chain protein with a PDB ID 1ABC from which chain A, having 200 amino acids, was extracted and used, the 10 fractions would be **10_1ABC_A, 20_1ABC_A, 30_1ABC_A, ... , 90_1ABC_A, 100_1ABC_A**, where a fraction **20_1ABC_A** would have 40 amino

acids from the N-terminal region of chain A representing 20% of the content of the protein.

- For each protein family, 10 phylogenetic trees as outlined below were created, each using a different one of the 10 sets of fractional structures along with the complete protein structures. For instance, for the cytochrome family of 20 structures, 10 trees are generated $T_{10\%}$, $T_{20\%}$, $T_{30\%}$, ... , $T_{90\%}$, $T_{100\%}$ where a tree $T_{60\%}$ uses a structural dataset of 40 structures of which 20 are complete structures as provided by RCSB and the remaining their respective 60% fractions.

For the generation of the phylogenetic trees the following steps were followed.

- Pairwise comparisons were done for each protein structure using Superpose [1]. Due to the nature of the algorithm, comparisons are order specific i.e. A $\cong$ B $\neq$ B $\cong$ A. Both pairwise comparisons were, therefore, performed and the scores were averaged to attain a final score "q" of the comparison between structure A and B.
- The similarity score was subtracted from one (d = 1 - q) to generate a distance (d) value.
- A matrix was populated with the pairwise distances.
- The neighbour-joining (NJ) algorithm [2] as implemented by the Phylo [3] package in Biopython [4] was used to generate a NJ-tree from the matrix. The tree was then visualized using Figtree [5].

This section generated 10 trees for each of the protein families. The final tree $T_{100\%}$ used two structural datasets, namely the original crystal structures and the 100% fraction generated by the method previously listed. It should be noted that the 100% fraction was an exact replica of the complete structures. The final tree, referred to as the true tree in this work for each family represents the actual positions of the structures in an evolutionary context. As the two datasets used to generate the $T_{100\%}$ tree are identical, the red and black labelled structures share the same clade. This approach empirically tested the size effect in that if size of the structures compared could impact the evolutionary analysis it would result in the fractional structures moving away from their true positions as delineated by the true tree, i.e. $T_{100\%}$.

To quantitatively assess the difference between the fractional trees and the true tree the Euclidean distance between these was quantified. For this the phylogenetic tree comparison program treecompare as made available by DendroPy [6], a python library for phylogenetic computing, was used.

2.3.2 *Part* 2: The shape effect

No direct way was found for the analysis of the morphometric contribution to the Q_{score} metric. Ideally the shape effect can be explored by analysing a number of evolutionarily related protein structures with exactly or nearly identical sizes. This is realistically not possible as such structural data does not exist.

Another method could have been a controlled increase in the pairwise compared structural distance, i.e. RMSD. While a controlled increase in size was done to probe the size effect by generating fractional proteins in the previous section, it is not possible to do the same for shape as generating protein structures with continuously increasing structural pairwise comparison distance is not a trivial task. Hence an indirect approach was used.

The indirect approach used MD simulations to generate alternative conformations for protein structures and used the shape factor, Equation 2.3, to look at fluctuations within Q_{score} when only the shape of a protein changes and not its size. Comparing the starting protein structure with its alternative conformation sets Equation 2.2 to 1, as a protein structure is compared to itself.

To explore the shape effect, 50 protein structures, from the ferritin-like superfamily, listed in Table 2.2, were simulated using the NAMD [7] program along with the CHARMM36FF [8] force field. The following steps were followed to generate alternative conformations:

1. Each structure was energy minimized for 200 steps using the default minimiser in NAMD [7].

2. Each structure was solvated using the TIP3P [9] water model. In this step, a minimum cubic box was created with dimensions that fit the protein, i.e. between the minimum $(x_{min}, y_{min}, z_{min})$ and maximum $(x_{max}, y_{max}, z_{max})$ coordinates of the protein. The boundaries were

extended by 15 Å in each direction and the newly added volume was filled with the solvent.

3. After minimization, excess charge was neutralized (if present) through the addition of Na^+ and Cl^- as counter ions, for effective calculation of long-range electrostatics using PME [10] summations. The ions were added by randomly selecting a water molecule and substituting it with the ion.

4. After solvation and ionization, the system was minimized (for 300 steps using the default minimizer) to remove clashes and find a new potential energy minimum.

5. Following the previous step, a heating phase of the MD simulations was conducted. To achieve the simulation temperature of 310 K, the temperature was increased from 0 K in 5 K increments every 10000 integration steps, with each step being 2 fs apart.

6. After a temperature of 310 K was achieved, each structure was simulated for an additional 20 ns and conformations were recorded every 2 ps.

The starting structure for each of the simulations was compared to all the alternative conformations sampled within its particular trajectory and the all-atom positional RMSD value was calculated for each. The RMSD value was then used in conjunction with Equation 2.3 to gauge indirectly the contribution from fluctuations in shape while keeping the size contributions constant at one.

2.3.3 The MD-based bootstrap method

The MD-based bootstrap method has been discussed at length in the previous chapter, Chapter 1, Section 1.10.2. This section adds further insight into the limitations of this novel addition to structural phylogenetics by applying it to two control structural datasets from the globin and ferritin protein families. The structures used in this part are listed in Table 2.3 and illustrated in Figures 2.11 and 2.13, respectively.

The globins include α- and β-haemoglobins which are known from literature from sequence-based analysis to be a result of a relatively recent

Table 2.2: PDB and chain identifiers of protein structures used in testing the shape based contribution to the Q_{score}. [11–82]

PDB	Chain	PDB	Chain
2CHP	A	2FKZ	A
2JD7	A	2FZF	A
1BCF	A	2UX1	A
1BG7	A	2ZA7	A
1DPS	A	3E1Q	A
1EUM	A	3E6S	A
1JGC	A	1AFR	A
1JI4	A	1JK0	A
1JI5	A	1MTY	B
1JIG	A	1MXR	A
1JTS	A	1OQU	A
1KRQ	A	1OTK	A
1LB3	A	1R2F	A
1LKO	A	1UZR	A
1N1Q	A	1W68	A
1NFV	A	1ZA0	A
1O9R	A	2INC	B
1QGH	A	2UW1	A
1R03	A	2UW1	B
1TJO	A	2UW2	A
1TK6	A	3DHG	A
1UVH	A	2INP	C
1VLG	A	2VZB	A
1YUZ	A	3EE4	A
2FJC	A	3QHB	A

Table 2.3: The ribonucleotide reductase-like (RNR-like) and globin structures used as controls to test the MD-based bootstrap method. In the globins column chains with identifier "A" are α-haemoglobins and those with identifier "B" are β-haemoglobins.

RNR-like		Globins	
PDB	Chain	PDB	Chain
1AFR	A	1GCV	A
1MTY	B	1GCV	B
1R2F	A	1HV4	A
1UZR	A	1HV4	B
1ZA0	A	2DN2	A
2INP	C	2DN2	B
2UW1	A	3HRW	A
2UW2	A	3HRW	B

gene duplication and divergence event [83]. On the other hand, the other structural dataset includes ribonucleotide reductase-like proteins which are a subset of the ferritin-like protein superfamily and are known from literature to be more diverged [84]. Ribonucleotide reductase-like structures are

a part of the ferritin-like super family and were already simulated as a part of the previous section. The globin simulations were conducted in a similar manner as outlined previously, Section 2.3.2.

The starting structures of the simulations were used for the creation of a reference tree, T_0. For convenience and clarity purposes only four alternative coordinate sets were then created by randomly extracting four conformations from the trajectory of coordinates collected during the simulation of each structure. Again for convenience and clarity purposes only four trial trees were built from each of the alternative coordinate sets, T_1 .. T_4. The relationships inferred from the reference tree were then enumerated in the trial trees and summarised onto the nodes of the reference tree. The reference, trial and annotated trees, for each of the families, are shown in Figures 2.10 and 2.12.

2.4 Results

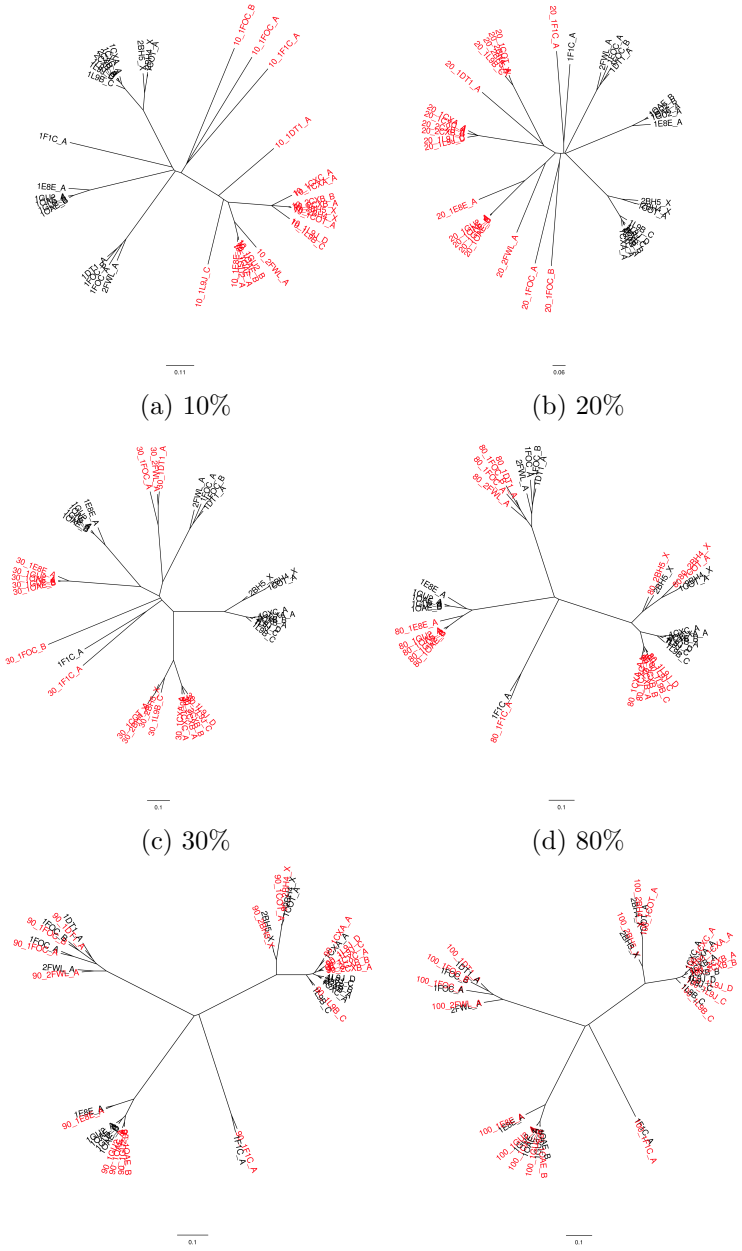
2.4.1 *Part* 1: The size effect

20 protein structures from each of three protein families, namely cytochromes, globins and ferritins, were used in this analysis. As discussed in the methods sections each protein was decomposed into fractions, with each fraction starting from the N-terminal region of the protein and increasing by 10% of the total number of residues in the protein from the previous one. The fractions were labelled to hold the name of the fraction of the protein they represented.

Each of the phylogenetic trees that are generated holds two structural datasets, with one comprising all 20 complete structures, and the other one of the 10 sets comprising different fractions of the structure, see methods, Section 2.3.1.

Figure 2.2, which for better visualization only shows six of the 10 cases, illustrates the outcome for the cytochrome family. The figure clearly illustrates that the fractions cluster separately, e.g. in Figure 2.2 (a) the 10% fractions of the complete structures occupy a separate split.

The fractions are identical to their respective structural fragments in the complete structures, i.e. for a structure 1ABC_A, its 10% fraction from the N-terminal region, 10_1ABC_A, is structurally identical to the first 10% of the amino acids. As the size of the fractions increase, they converge to their

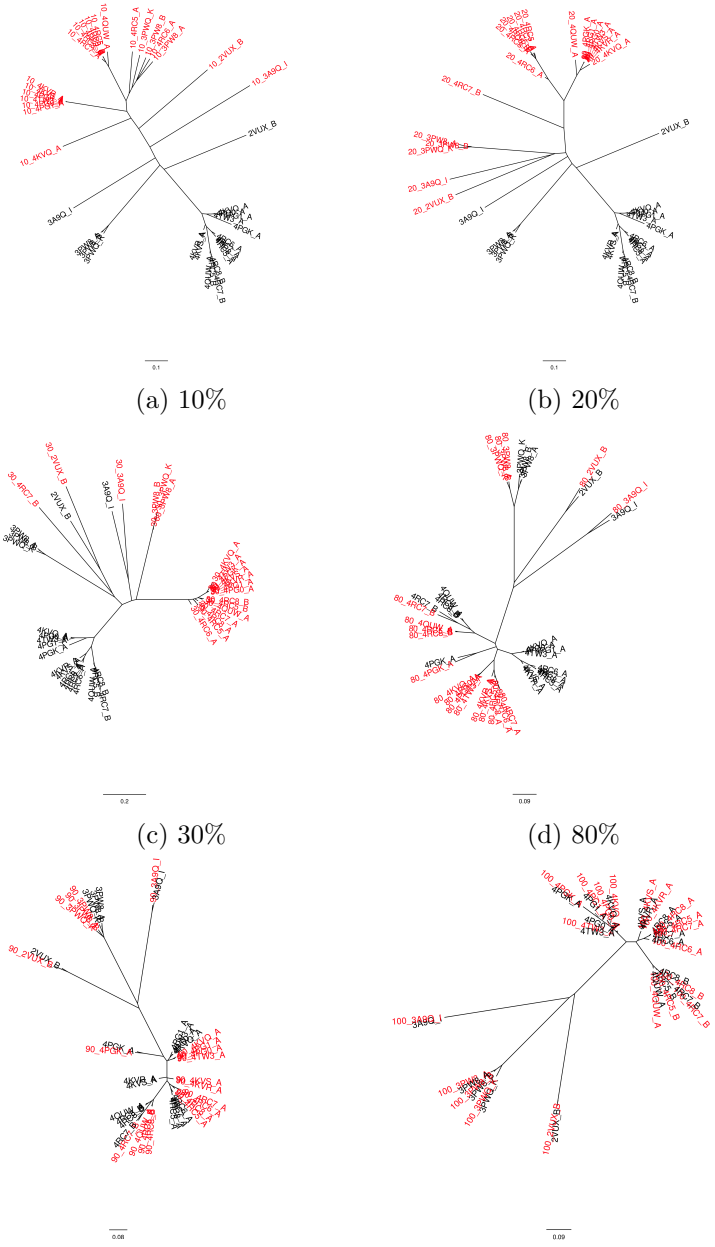


(a) 10% (b) 20%

(c) 30% (d) 80%

(e) 90% (f) 100%

Figure 2.2: Fractional structural analysis of proteins in the cytochrome family. Only six of the total 10 fractional structural datasets are visualized. The fractional structures are shown in red, and the complete structures in black. See the results section for details. See Appendix-II for enlarged figures.

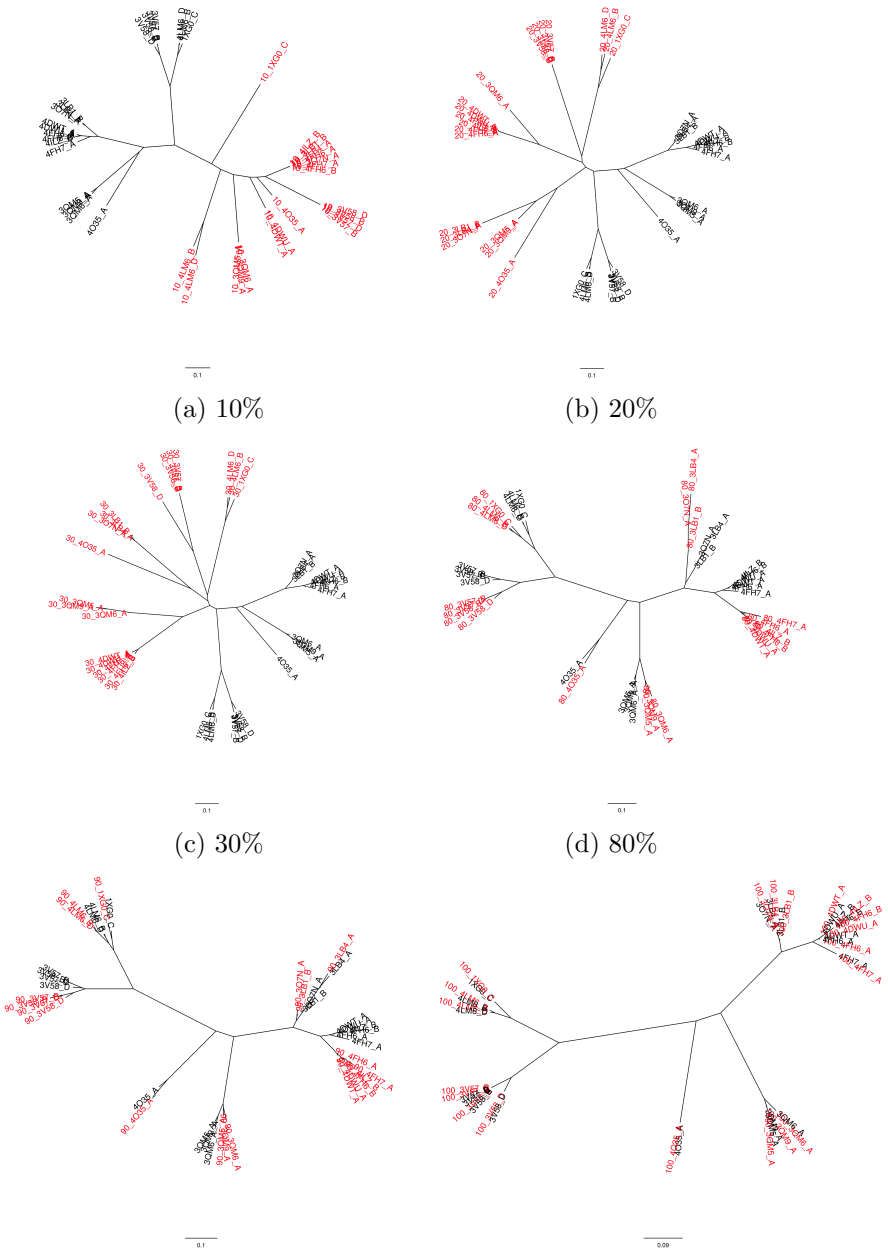


(a) 10% (b) 20%

(c) 30% (d) 80%

(e) 90% (f) 100%

Figure 2.3: Fractional structural analysis of proteins in the ferritin family. Only six of the total 10 fractional structural datasets are visualized. The fractional structures are shown in red, and the complete structures in black. See the results section for details. See Appendix-III for enlarged figures.



(a) 10% (b) 20%

(c) 30% (d) 80%

(e) 90% (f) 100%

Figure 2.4: Fractional structural analysis of proteins in the globin family. Only six of the total 10 fractional structural datasets are visualized. The fractional structures are shown in red, and the complete structures in black. See the results section for details. See Appendix-IV for enlarged figures.

true positions in the tree, as determined by $T_{100\%}$, with full recovery of positions as the sizes of the fractions converge towards 100% of the complete structures. In the context of this work recovery implies that the fractional structure occupies the same clade as its complete structural counterpart. The trend of fractional structures clustering separately and slowly converging to their respective positions in the true tree is also observed with ferritins and globins as illustrated in Figures 2.3 and 2.4.

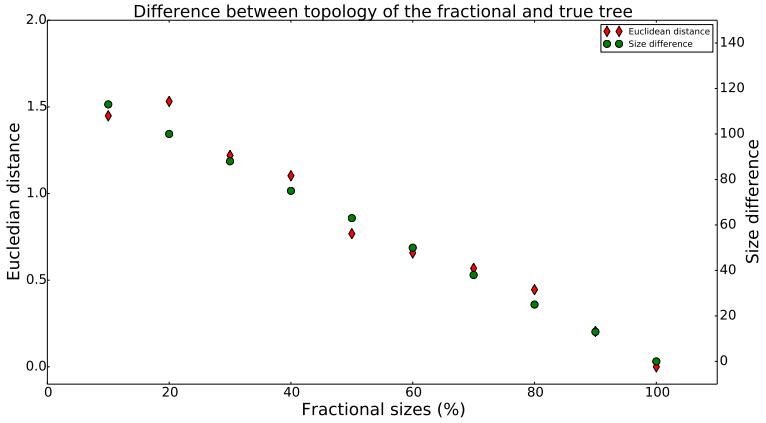


Figure 2.5: The Euclidean distance between fractional trees, i.e. $T_{10\%}$ through $T_{90\%}$, and the true tree, i.e. $T_{100\%}$ for the cytochrome family with a mean structural size of 125 amino acids. As the size difference between the complete and fractional structure reduces, the topology of the fractional trees approaches that of the true tree, i.e. $T_{100\%}$.

While the Figures 2.2 through 2.4 illustrate the effect qualitatively, the same can be illustrated quantitatively by measuring the Euclidean distance between each fractional tree, i.e. $T_{10\%}$ through $T_{90\%}$, and the true tree, i.e. $T_{100\%}$. This distance quantitatively describes the difference between the trees. The distance trends between intermediate and final tree for each of the three families are shown in Figures 2.5, 2.6 and 2.7. These figures reveal that as the size difference between the two structural datasets, i.e. between the fractional and complete structures, becomes smaller the Euclidean distance between the trees built from the fractional intermediate datasets, i.e. $T_{10\%}$ through $T_{90\%}$, and the true tree, i.e. $T_{100\%}$, also reduces.

This exercise aimed at uncovering a size threshold quantity beyond which protein structures should not be compared using the Q_{score} metric. The results for the cytochromes, globins and ferritins unanimously reveal that a

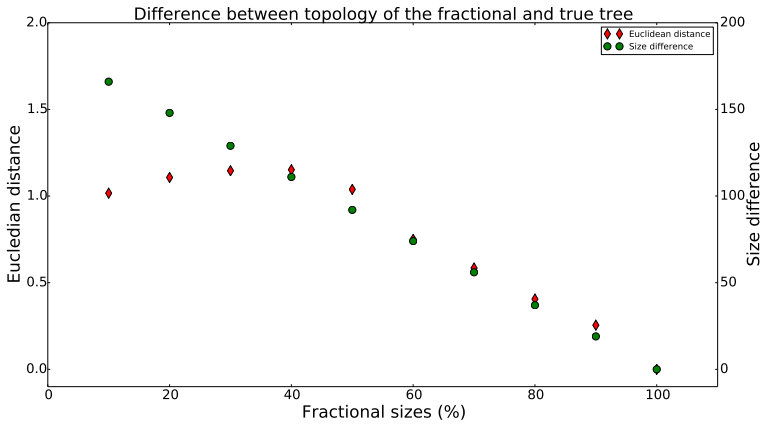


Figure 2.6: The Euclidean distance between fractional trees, i.e. $T_{10\%}$ through $T_{90\%}$, and the true tree, i.e. $T_{100\%}$ for the globin family with a mean structural size of 184 amino acids. As the size difference between the complete and fractional structure reduces, the topology of the fractional trees approaches that of the true tree, i.e. $T_{100\%}$.

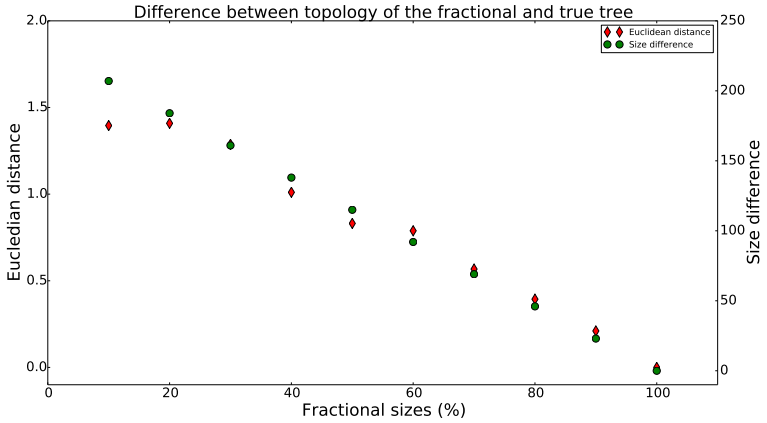


Figure 2.7: The Euclidean distance between fractional trees, i.e. $T_{10\%}$ through $T_{90\%}$, and the true tree, i.e. $T_{100\%}$ for the ferritin family with a mean structural size of 229 amino acids. As the size difference between the complete and fractional structure reduces, the topology of the fractional trees approaches that of the true tree, i.e. $T_{100\%}$.

smaller size difference results in better recovery. While some fractions, at just 80% of their complete counterparts, begin occupying the same clade, complete resolution is only achieved between 90% and 100%, i.e. with a size difference of no more than 12 amino acids for cytochromes, 18 amino acids

for globins and 22 amino acids for ferritins. This exercise therefore places a limit of 10% or smaller for variation in structural size of the proteins being compared using the Q_{score} metric for accurate phylogenetic recovery.

2.4.2 *Part* 2: The shape effect

The complexity of protein structure makes exploring the contribution of morphometric changes towards the Q_{score} a non-trivial problem. An indirect approach was adopted in this work. MD simulations were used to sample alternative conformations of 50 protein structures from the ferritin-like superfamily. Post-simulation, the starting conformation was compared with all subsequently sampled conformations by computing the atom-positional RMSD. Using Equation 2.3, the RMSD values were converted into shape based contributions towards the Q_{score}. The sampled conformations from simulations only change shape and hence the size effect is excluded.

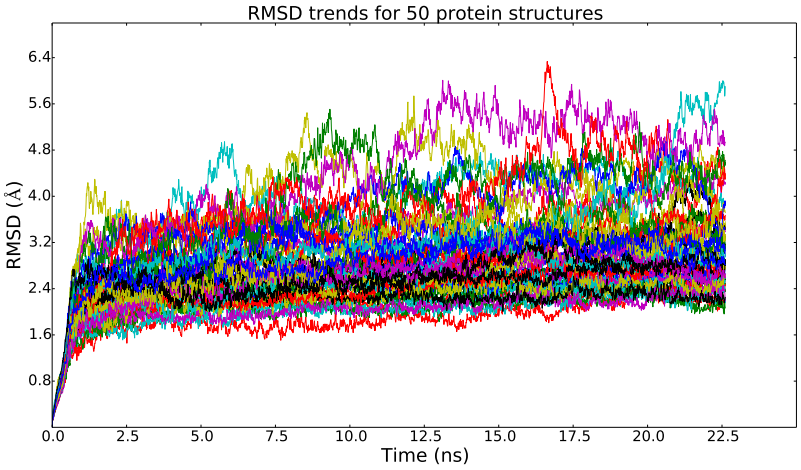


Figure 2.8: RMSD from the starting conformation of each of 50 protein structures from the ferritin-like superfamily during MD simulations.

The RMSD and corresponding shape effects, Equation 2.3, are shown in Figures 2.8 and 2.9. During the simulation process a structure samples conformations from the conformational landscape, Chapter 1, Figure 1.23, and hence moves away from its starting structure. This is illustrated by an initial increase in the RMSD value, which then gradually plateaus indicating equilibration. The shape contribution to the Q_{score} shows an inverted trend, also reflected in the mathematical formulation, Equation 2.3. It should be

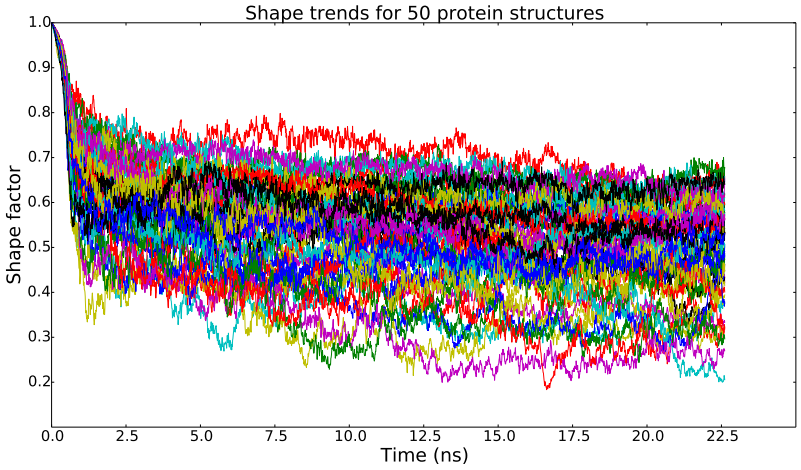


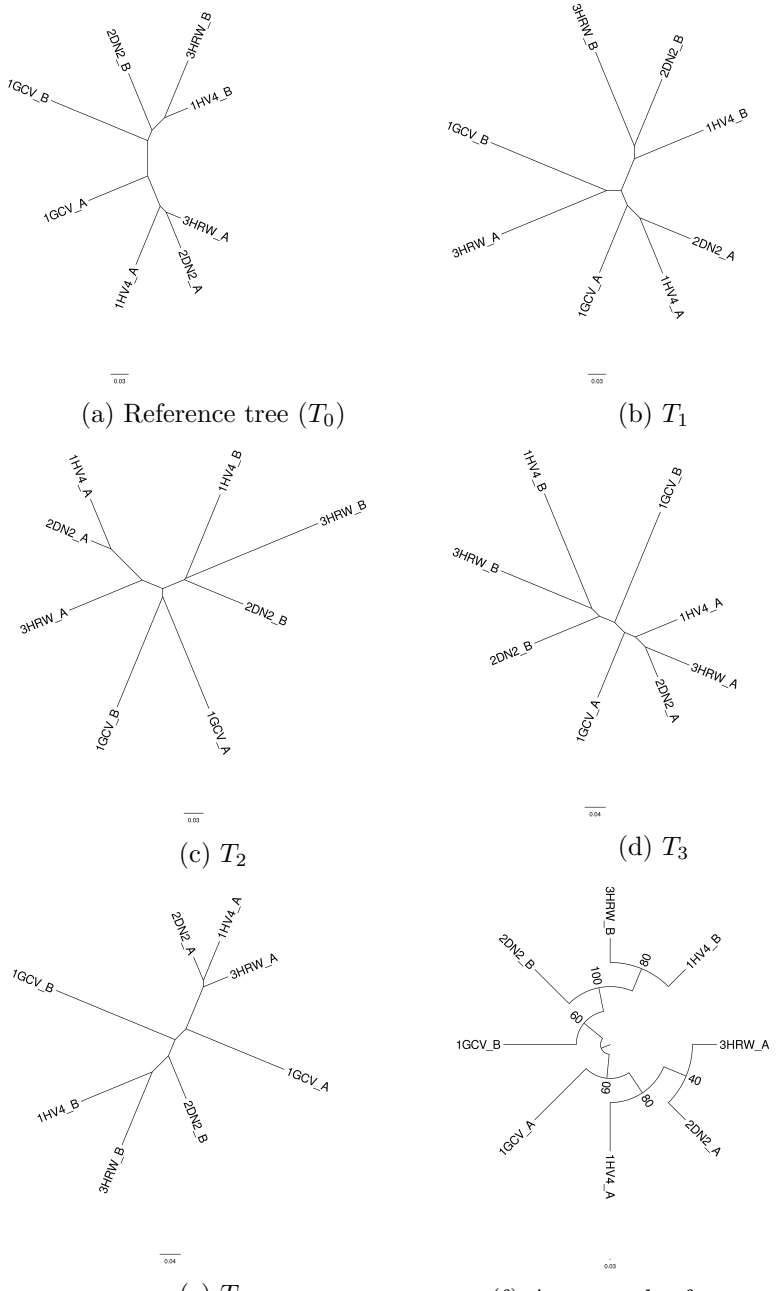
Figure 2.9: The shape factor calculated from Equation 2.3 and using the RMSD values as shown in Figure 2.8 for all the 50 protein structures from the ferritin-like superfamily. This shows the shape factor during the simulation for comparisons with the starting conformation.

noted that a Q_{score} of "1" indicates a perfect match, and decreases with increasing difference between compared structures, whereas the RMSD shows the opposite behaviour, increasing with the distance between structures.

While the morphometric contribution could not be tested in a manner similar to the size investigation, the fluctuation in RMSD and its corresponding shape contribution to the Q_{score} shows a high sensitivity to subtle changes in the structure sampled from MD simulations. This observation is helpful as it shows that the MD-based bootstrap can be used confidently to introduce fluctuations in the structural data to gauge the robustness of the inferred relationships. The same observation also serves as a warning. Protein molecules are highly dynamic. Therefore subtle differences in the structures used to infer evolutionary relationships may result in vastly different tree topologies and hence lead to an alternative evolutionary inference.

2.4.3 The MD-based bootstrap method

Eight protein structures from each of the two protein families, namely globins and ribonucleotide reductase-like, were simulated using MD simulations. The eight protein structures from the globin family comprise four α and four β chains. The α and β chains emerged from a gene duplication

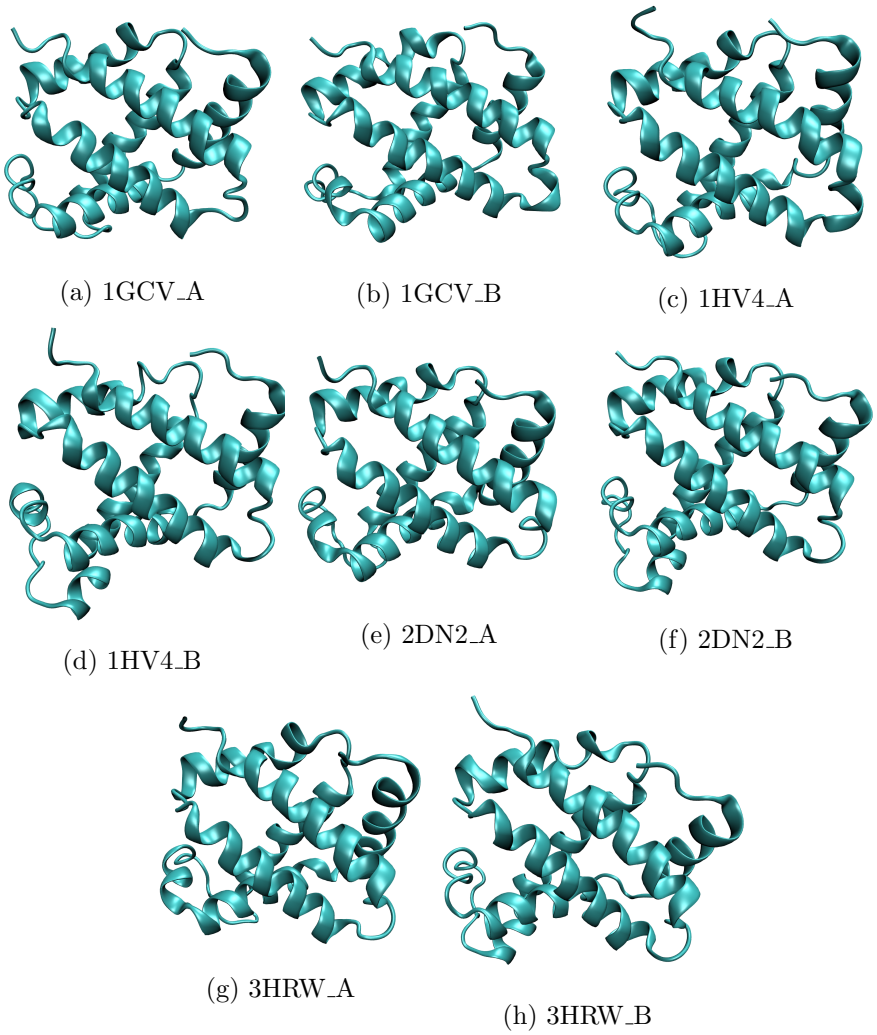


(a) Reference tree (T_0) (b) T_1

(c) T_2 (d) T_3

(e) T_4 (f) Annotated reference tree

Figure 2.10: MD-based bootstrap trials on structures from the globin family. The annotated tree (f) uses T_0, the reference tree, and shows the relationships recovered as a percentage of the trials conducted (in this case 5, T_0 ... T_4).

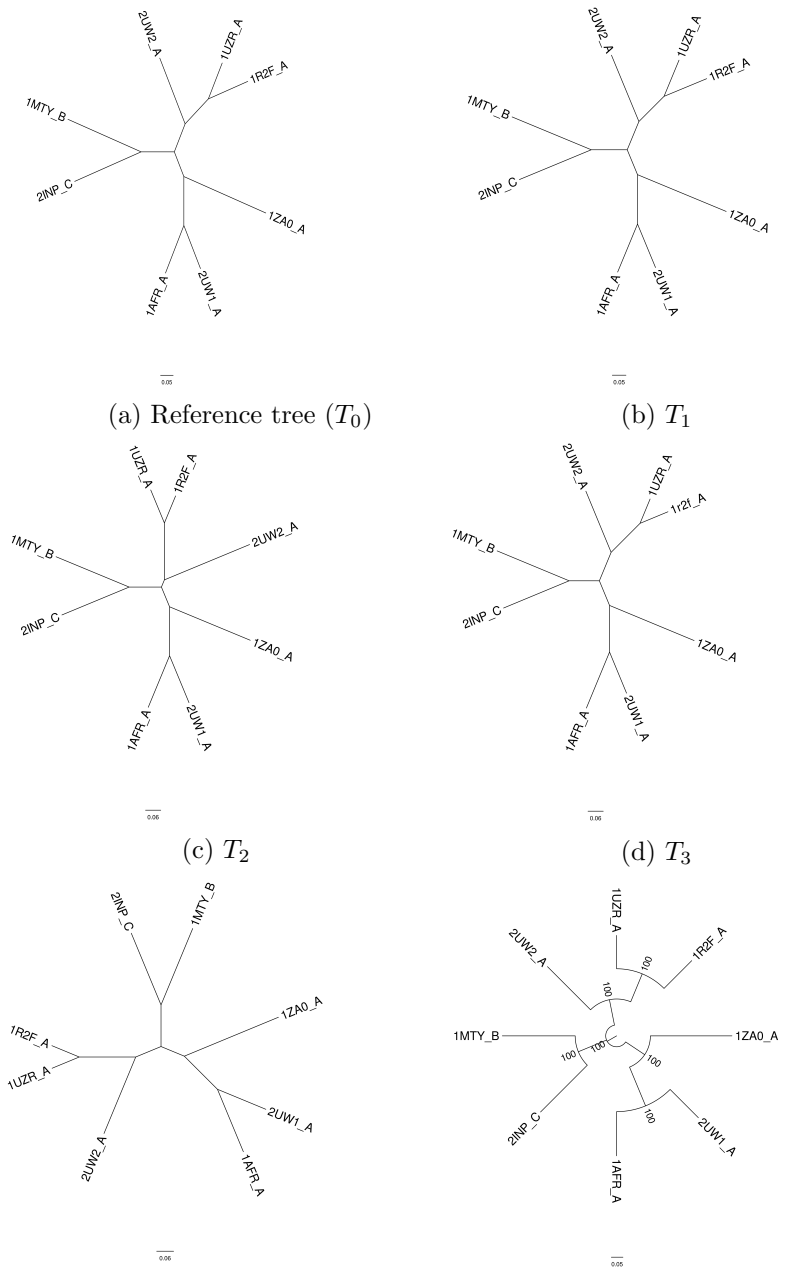


(a) 1GCV_A (b) 1GCV_B (c) 1HV4_A

(d) 1HV4_B (e) 2DN2_A (f) 2DN2_B

(g) 3HRW_A (h) 3HRW_B

Figure 2.11: The crystal structures of the α and β-haemoglobins which show a high degree of structural conservation.

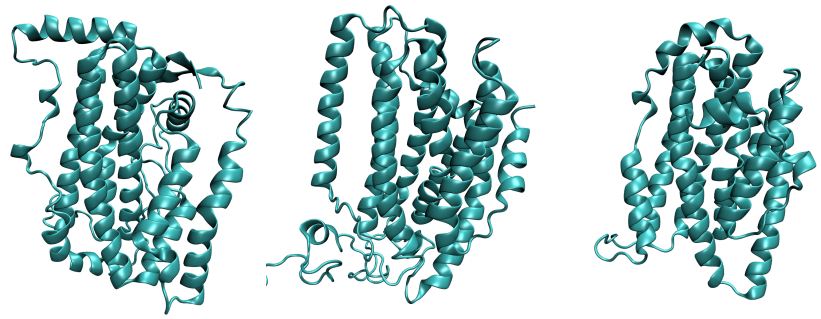


(a) Reference tree (T_0) (b) T_1

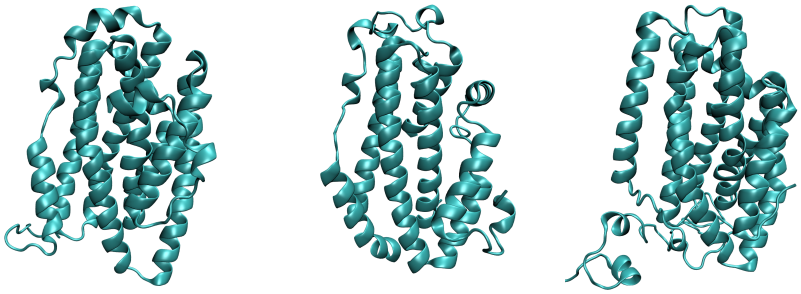
(c) T_2 (d) T_3

(e) T_4 (f) Annotated reference tree

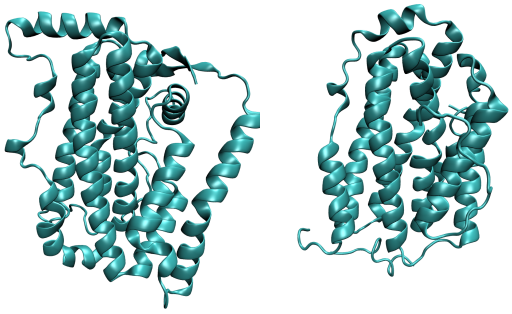
Figure 2.12: MD-based bootstrap trials on structures from the ribonucleotide reductase-like family. The tree in (f) is annotated as in Figure 2.10.



(a) 1AFR_A (b) 1MTY_B (c) 1R2F_A



(d) 1UZR_A (e) 1ZA0_A (f) 2INP_C



(g) 2UW1_A (h) 2UW2_A

Figure 2.13: The crystal structures of the ribonucleotide reductase-like proteins which show more structural divergence compared to the α- and β-haemoglobins.

and subsequent divergence, an event that is classified as occurring recently in the context of evolutionary history [83]. Meanwhile the ribonucleotide reductase-like family is more diverged [84]. The structural conservation in globins and divergence of the ribonucleotide reductase-like family can be seen reflected in their respective structures shown in Figures 2.11 and 2.13.

From the coordinate trajectories of each of the structures, post-simulation, four alternative conformations were randomly extracted. The starting structures were used to create two reference phylogenetic trees, T_0 in Figures 2.10 and 2.12, one for each protein family. Trees T_1 through T_4 were generated from the alternative conformations of each of the structures. Sub-figure "f" in Figures 2.10 and 2.12 summarise the relationships reflected by the reference trees which were also present in the replicates.

The statistics shown in the annotated reference tree highlights an important aspect of the MD-based bootstrap method. It has previously been explained, Chapter 1, Section 1.10.2, that higher numbers on the node reflect a more robust relationship. In this instance a contrast is seen between the values on the nodes of the annotated globin and ribonucleotide reductase-like trees.

The statistics shown on the tree tend to decrease when structures are similar as is the case with globins, Figure 2.11, and hence the alternative conformations sampled overlap, generating a variety of distance values, which obfuscates the evolutionary relationships. In contrast to this, Figure 2.13, shows that structures from the ribonucleotide reductase-like protein family are more structurally diverged and hence little or no overlap occurs in the sampled structures, leading to better statistics.

The weak statistics can also be attributed to under-sampling in the MD simulations, as these proteins structures were simulated for just 20 ns. While this could impact the statistics obtained, i.e. result in weak statistics, in this instance it can safely be said that the structural proximity is the cause of the poor statistics and not under-sampling. Although only illustrated for a reference with four trials for simplicity purposes in this case, the same holds when more conformations are sampled and more trials are conducted, see Chapter 5, Section 5.3.4.

2.5 Discussion

The purpose of this chapter was to look at the respective behaviour of the two key parts of the structure-based phylogenetic method developed in this work, namely the use of the Q_{score} metric to quantify a distance between structures compared and the use of the MD-based bootstrap method. The Q_{score} metric was further divided into two parts, Equations 2.2 and 2.3, each of which was evaluated in a controlled manner.

Equation 2.2, termed the size factor, was tested by generating fractions of structures and then comparing each fractional structural dataset to the complete structural dataset. This ensured that the RMSD contributions were 0, thus evaluating Equation 2.3 to 1.

Equation 2.3, termed the shape factor, was tested using MD simulations. Each simulation allowed a structure to sample alternative conformations. Thus when comparing between multiple conformations for the same structure the size factor, Equation 2.2 evaluated to 1, allowing for only the influence of shape to be probed.

While the incorporation of the RMSD, alignment length and individual lengths of the two proteins compared gives Q_{score} an advantage over other structural comparison metrics, as discussed in Chapter 1, Section 1.8, it also has caveats. These are illustrated through the analysis of the size and shape factors presented in this chapter.

The size factor reveals that, if the size difference between the compared protein structures varies by more than 5-10%, the structural comparison may generate incorrect results in an evolutionary context, as reflected by the Euclidean distance measurements between $T_{100\%}$ and the fractional trees, Figures 2.5-2.7.

The shape factor highlights another important aspect. Depending purely on the RMSD, the shape factor fluctuates considerably and strongly impacts the Q_{score}, which in some instances is reduced to as low as 0.2, see Figure 2.9. While this test was not done in an evolutionary context, i.e. it only observed fluctuations in the Q_{score}, it does strongly indicate that subtle changes in the structure may result in different tree topologies and hence different evolutionary interpretations.

Apart from the Q_{score} investigation, the MD-based bootstrap also provided useful insight. Using two different control structural datasets, the

meaning of significance beyond counting recovered relationships in trials was illustrated. In the recently diverged dataset, poor support values were seen, contrary to the structural dataset that was deeply diverged. This concept can be considered analogous to the way species and subspecies are classified. Subspecies are defined as those that have lesser differences or alternatively have more similarities, whereas if the differences increase beyond a threshold, the organisms are classified as separate species. Analogous to this, the MD-based bootstrap generates stronger support when there exists a certain degree of difference between the structures being compared.

Although not tested at present, it can be conjectured that if protein-based datasets cannot be characterised by sequence-based methods, due to high divergence, only then should they be processed using the structural method coupled with MD-based bootstrap for gauging robustness. This can be backed by the control datasets where α and β-globins are well characterised by sequence-based methods, implying homology detectable at the sequence level, and hence do not possess structural diversity, resulting in conformational overlap and hence weaker statistics. The ribonucleotide reductase-like family, contrary to globins, has weaker sequence level homology and hence a higher structural diversity, which results in better statistics.

2.6 Conclusion

The important aspects highlighted in this work signify that a certain amount of care needs to go into the interpretation of results from the use of this method. While there exists a tendency to compare any two structures, it should be noted that the Q_{score}, beyond a certain size difference of 5-10%, is dominated purely by the size difference and offers no meaningful evolutionary insight in terms of structural distance.

Similarly, proteins are thermodynamic molecules and exist as an ensemble of conformations due to thermal motions. Results in this work illustrate that the nature of the phylogeny that is generated depends on which conformation is selected from each protein's conformational ensemble, as was shown in the RMSD and shape factor plots, Figures 2.8 and 2.9.

As discussed in the previous section a certain degree of divergence is necessary to successfully enact the MD-based bootstrap method. As to what this degree is, and how to go about quantifying a threshold remains

an open problem. As stated earlier, based on conjecture, one could limit the use of this method, as is originally intended, to cases of deeply diverging protein datasets with which sequence-based methods struggle.

It is therefore necessary to avoid comparison of protein structures which are either too similar, like the α and β-globin example, or structures that vary significantly in size.

2.7 Future work

From the analysis in this work some of the questions that still need to be addressed are presented along with the future direction in which they will be analysed.

1. Can protein structural alignments be optimized through the introduction of gap penalties, to compare structures that vary considerably in size? In an evolutionary context, rarely would structural dataset obey the 5-10% size limit. Inclusion of gap penalties would make this method applicable to more evolutionarily diverse proteins where entire folds, i.e. multiple α-helices or β-strands or a mix of both, have been lost or gained.

2. What type of gap penalties can be incorporated in protein structural comparisons? While in sequence-based methods affine gap penalties are commonly used, i.e. a mix of gap opening and extension, it is not yet clear how the same approach can be extended to protein structures. At a structural level entire folds are gained or lost. Hence a better understanding of protein structural evolution is required before a complex model of gap penalties is applied, until which perhaps the affine model can be used with empirical parameters that can be optimised in cases where evolutionary histories of protein families are known.

3. How can the "breathing motion" of proteins be incorporated, if at all, into structural comparisons? This question perhaps needs a careful resolution as the fluctuations from alternative conformations are the foundation of the MD-based bootstrap. To satisfactorily address this it is perhaps needed that reference trees should not be used and instead

50% consensus or 95% strict consensus trees are used. The reference tree, generated from the distance-based neighbour-joining algorithm, may or may not be accurate. The use of consensus trees will generate results for the most robust relationships allowing the continued usage of this method without the need for significant alterations.

4. Can a distance-based cut-off be determined which informs when structures are similar and when they are not for use with the MD-based bootstrap method? It was addressed earlier in this chapter that the MD-based bootstrap method can be limited to only deeply diverging data, as is its intended purpose, hence only used with substantially diverged sequences, where the divergence gives rise to the requisite level of structural divergence, making the conformational ensembles non-overlapping. Apart from this any approach which summarises the distance between proteins as a scalar will not be sufficient due to substantial loss of information when reducing the 3D comparison between structures to a single number. Thus a new distance measure is required which maximally retains the information from the 3D structural comparison.

Bibliography

[1] Krissinel, E. and Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12): 2256–2268, 2004.

[2] Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[3] Talevich, E., Invergo, B. M., Cock, P. J. A., and Chapman, B. A. Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13(1):209, 2012.

[4] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., and Wilczynski, B. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[5] Rambaut, A. FigTree v1. 4. *Molecular Evolution, Phylogenetics and Epidemiology*, 2012.

[6] Sukumaran, J. and Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

[7] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[8] Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom

protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.

[9] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[10] Darden, T., York, D., and Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[11] Cobessi, D., Huang, L. S., Ban, M., Pon, N. G., Daldal, F., and Berry, E. A. The 2.6 Å resolution structure of *Rhodobacter capsulatus* bacterioferritin with metal-free dinuclear site and heme iron in a crystallographic "special position". *Acta crystallographica. Section D, Biological crystallography*, 58(Pt 1):29–38, 2002.

[12] Andersson, M. E., Högbom, M., Rinaldo-Matthis, A., Blodig, W., Liang, Y., Persson, B. O., Sjöberg, B. M., Su, X. D., and Nordlund, P. Structural and mutational studies of the carboxylate cluster in iron-free ribonucleotide reductase R2. *Biochemistry*, 43(24):7966–7972, 2004.

[13] Eriksson, M., Jordan, A., and Eklund, H. Structure of *Salmonella typhimurium* nrdF ribonucleotide reductase in its oxidized and reduced forms. *Biochemistry*, 37(38):13359–13369, 1998.

[14] Gauss, G. H., Benas, P., Wiedenheft, B., Young, M., Douglas, T., and Lawrence, C. M. Structure of the DPS-like protein from *Sulfolobus solfataricus* reveals a bacterioferritin-like dimetal binding site within a DPS-like dodecameric assembly. *Biochemistry*, 45(36):10815–10827, 2006.

[15] Kim, J., Malashkevich, V., Roday, S., Lisbin, M., Schramm, V. L., and Almo, S. C. Structural and kinetic characterization of *Escherichia coli* TadA, the wobble-Specific tRNA deaminase. *Biochemistry*, 45(20): 6407–6416, 2006.

[16] Lawson, T. L., Crow, A., Lewin, A., Yasmin, S., Moore, G. R., and Le Brun, N. E. Monitoring the iron status of the ferroxidase center of

Escherichia coli bacterioferritin using fluorescence spectroscopy. *Biochemistry*, 48(38):9031–9039, 2009.

[17] Sazinsky, M. H., Dunten, P. W., McCormick, M. S., DiDonato, A., and Lippard, S. J. X-ray structure of a hydroxylase-regulatory protein complex from a hydrocarbon-oxidizing multicomponent monooxygenase, *Pseudomonas sp.* OX1 phenol hydroxylase. *Biochemistry*, 45 (51):15392–15404, 2006.

[18] Swartz, L., Kuchinskas, M., Li, H., Poulos, T. L., and Lanzilotta, W. N. Redox-dependent structural changes in the *Azotobacter vinelandii* bacterioferritin: New insights into the ferroxidase and iron transport mechanism. *Biochemistry*, 45(14):4421–4428, 2006.

[19] Whittaker, M. M., Barynin, V. V., Igarashi, T., and Whittaker, J. W. Outer sphere mutagenesis of *Lactobacillus plantarum* manganese catalase disrupts the cluster core. Mechanistic implications. *European Journal of Biochemistry*, 270(6):1102–16, 2003.

[20] Högbom, M. and Nordlund, P. A protein carboxylate coordinated oxo-centered tri-nuclear iron complex with possible implications for ferritin mineralization. *FEBS Letters*, 567(2-3):179–182, 2004.

[21] Uppsten, M., Davis, J., Rubin, H., and Uhlin, U. Crystal structure of the biologically active form of class Ib ribonucleotide reductase small subunit from *Mycobacterium tuberculosis*. *FEBS Letters*, 569(1-3):117–122, 2004.

[22] Gauss, G. H., Reott, M. A., Roha, E. R., Young, M. J., Douglas, T., Smith, C. J., and Lawrence, C. M. Characterization of the *Bacteroides fragilis* bfr gene product identifies a bacterial DPS-like protein and suggests evolutionary links in the ferritin superfamily. *Journal of Bacteriology*, 194(1):15–27, 2012.

[23] Yoshizawa, K., Mishima, Y., Park, S. Y., Heddle, J. G., Tame, J. R. H., Iwahori, K., Kobayashi, M., and Yamashita, I. Effect of N-terminal residues on the structural stability of recombinant horse L-chain apoferritin in an acidic environment. *Journal of Biochemistry*, 142(6):707–713, 2007.

[24] Ceci, P., Ilari, A., Falvo, E., and Chiancone, E. The Dps protein of *Agrobacterium tumefaciens* does not bind to DNA but protects it toward oxidative cleavage. X-ray crystal structure, iron binding, and hydroxyl-radical scavenging properties. *Journal of Biological Chemistry*, 278(22):20319–20326, 2003.

[25] Ceci, P., Ilari, A., Falvo, E., Giangiacomo, L., and Chiancone, E. Reassessment of protein stability, DNA binding, and protection of *Mycobacterium smegmatis* Dps. *Journal of Biological Chemistry*, 280(41): 34776–34785, 2005.

[26] Guy, J. E., Whittle, E., Kumaran, D., Lindqvist, Y., and Shanklin, J. The crystal structure of the Ivy δ4-16:0-ACP desaturase reveals structural details of the oxidized active site and potential determinants of regioselectivity. *Journal of Biological Chemistry*, 282(27):19863–19871, 2007.

[27] Moche, M., Shanklin, J., Ghoshal, A., and Lindqvist, Y. Azide and acetate complexes plus two iron-depleted crystal structures of the diiron enzyme δ9 stearoyl-acyl carrier protein desaturase: Implications for oxygen activation and catalytic intermediates. *Journal of Biological Chemistry*, 278(27):25072–25080, 2003.

[28] Papinutto, E., Dundon, W. G., Pitulis, N., Battistutta, R., Montecucco, C., and Zanotti, G. Structure of two iron-binding proteins from *Bacillus anthracis*. *Journal of Biological Chemistry*, 277(17):15093–15098, 2002.

[29] Strand, K. R., Karlsen, S., Kolberg, M., Røhr, Å. K., Görbitz, C. H., and Andersson, K. K. Crystal structural studies of changes in the native dinuclear iron center of ribonucleotide reductase protein R2 from mouse. *Journal of Biological Chemistry*, 279(45):46794–46801, 2004.

[30] Granier, T., D'Estaintot, L. B., Gallois, B., Chevalier, J. M., Précigoux, G., Santambrogio, P., and Arosio, P. Structural description of the active sites of mouse L-chain ferritin at 1.2 Å resolution. *Journal of Biological Inorganic Chemistry*, 8(1-2):105–111, 2003.

[31] Iyer, R. B., Silaghi-Dumitrescu, R., Kurtz, D. M., and Lanzilotta, W. N. High-resolution crystal structures of *Desulfovibrio vulgaris* (Hildenbor-

ough) nigerythrin: Facile, redox-dependent iron movement, domain interface variability, and peroxidase activity in the rubrerythrins. *Journal of Biological Inorganic Chemistry*, 10(4):407–416, 2005.

[32] Tatur, J., Hagen, W. R., and Matias, P. M. Crystal structure of the ferritin from the hyperthermophilic archaeal anaerobe *Pyrococcus furiosus*. *Journal of Biological Inorganic Chemistry*, 12(5):615–630, 2007.

[33] Stillman, T., Hempstead, P., Artymiuk, P., Andrews, S., Hudson, A., Treffry, A., Guest, J., and Harrison, P. The high-resolution X-ray crystallographic structure of the ferritin (EcFtnA) of *Escherichia coli*; comparison with human H ferritin (HuHF) and the structures of the Fe3+ and Zn2+ derivatives. *Journal of Molecular Biology*, 307(2):587–603, 2001.

[34] D'Estaintot, B. L., Santambrogio, P., Granier, T., Gallois, B., Chevalier, J. M., Précigoux, G., Levi, S., and Arosio, P. Crystal structure and biochemical properties of the human mitochondrial ferritin and its mutant Ser144Ala. *Journal of Molecular Biology*, 340(2):277–293, 2004.

[35] Hamburger, A. E., West, A. P., Hamburger, Z. A., Hamburger, P., and Bjorkman, P. J. Crystal structure of a secreted insect ferritin reveals a symmetrical arrangement of heavy and light chains. *Journal of Molecular Biology*, 349(3):558–569, 2005.

[36] Ren, B., Tibbelin, G., Kajino, T., Asami, O., and Ladenstein, R. The multi-layered structure of Dps with a novel di-nuclear ferroxidase center. *Journal of Molecular Biology*, 329(3):467–77, 2003.

[37] Trikha, J., Theil, E. C., and Allewell, N. M. High resolution crystal structures of amphibian red-cell L ferritin: potential roles of structural plasticity and solvation in function. *Journal of Molecular Biology*, 248 (5):949–967, 1995.

[38] Zanotti, G., Papinutto, E., Dundon, W., Battistutta, R., Seveso, M., Giudice, G., Rappuoli, R., and Montecucco, C. Structure of the neutrophil-activating protein from *Helicobacter pylori*. *Journal of Molecular Biology*, 323(1):125–30, 2002.

[39] Jin, S., Kurtz, D. M., Liu, Z.-J., Rose, J., and Wang, B.-C. X-ray crystal structures of reduced rubrerythrin and its azide adduct: a structure-based mechanism for a non-heme diiron peroxidase. *Journal of the American Chemical Society*, 124(33):9845–55, 2002.

[40] McCormick, M. S., Sazinsky, M. H., Condon, K. L., and Lippard, S. J. X-ray crystal structures of manganese(II)-reconstituted and native toluene/o-xylene monooxygenase hydroxylase reveal rotamer shifts in conserved residues and an enhanced view of the protein interior. *Journal of the American Chemical Society*, 128(47):15108–15110, 2006.

[41] Sazinsky, M. H. and Lippard, S. J. Product bound structures of the soluble methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): Protein motion in the α-subunit. *Journal of the American Chemical Society*, 127(16):5814–5825, 2005.

[42] Schönafinger, A., Morbitzer, A., Kress, D., Essen, L. O., Noll, F., and Hampp, N. Morphology of dry solid-supported protein monolayers dependent on the substrate and protein surface properties. *Langmuir*, 22 (17):7185–7191, 2006.

[43] Marchetti, A., Parker, M. S., Moccia, L. P., Lin, E. O., Arrieta, A. L., Ribalet, F., Murphy, M. E. P., Maldonado, M. T., and Armbrust, E. V. Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature*, 457(7228):467–470, 2009.

[44] Frolow, F., Kalb, A. J., and Yariv, J. Structure of a unique twofold symmetric haem-binding site. *Nature Structural Biology*, 1(7):453–60, 1994.

[45] Grant, R. A., Filman, D. J., Finkel, S. E., Kolter, R., and Hogle, J. M. The crystal structure of Dps, a ferritin homolog that binds and protects DNA. *Nature Structural Biology*, 5(4):294–303, 1998.

[46] Ilari, A., Stefanini, S., Chiancone, E., and Tsernoglou, D. The dodecameric ferritin from *Listeria innocua* contains a novel intersubunit iron-binding site. *Nature Structural Biology*, 7(1):38–43, 2000.

[47] Macedo, S., Romão, C. V., Mitchell, E., Matias, P. M., Liu, M. Y., Xavier, A. V., LeGall, J., Teixeira, M., Lindley, P., and Carrondo, M. A.

The nature of the di-iron site in the bacterioferritin from *Desulfovibrio desulfuricans*. *Nature Structural Biology*, 10(4):285–90, 2003.

[48] Zeth, K., Offermann, S., Essen, L.-O., and Oesterhelt, D. Iron-oxo clusters biomineralizing on protein surfaces: Structural analysis of *Halobacterium salinarum* DpsA in its low- and high-iron states. *Proceedings of the National Academy of Sciences*, 101(38):13780–13785, 2004.

[49] Voegtli, W. C., Ge, J., Perlstein, D. L., Stubbe, J., and Rosenzweig, A. C. Structure of the yeast ribonucleotide reductase Y2Y4 heterodimer. *Proceedings of the National Academy of Sciences*, 98(18): 10073–8, 2001.

[50] Andersson, C. S. and Högbom, M. A *Mycobacterium tuberculosis* ligand-binding Mn/Fe protein reveals a new cofactor in a remodeled R2-protein scaffold. *Proceedings of the National Academy of Sciences*, 106(14):5633–5638, 2009.

[51] Bailey, L. J., McCoy, J. G., Phillips, G. N., and Fox, B. G. Structural consequences of effector protein complex formation in a diiron hydroxylase. *Proceedings of the National Academy of Sciences*, 105 (49):19194–19198, 2008.

[52] Högbom, M., Galander, M., Andersson, M., Kolberg, M., Hofbauer, W., Lassmann, G., Nordlund, P., and Lendzian, F. Displacement of the tyrosyl radical cofactor in ribonucleotide reductase obtained by single-crystal high-field EPR and 1.4 Å X-ray data. *Proceedings of the National Academy of Sciences*, 100(6):3209–3214, 2003.

[53] Zeth, K., Offermann, S., Essen, L.-O., and Oesterhelt, D. Iron-oxo clusters biomineralizing on protein surfaces: structural analysis of *Halobacterium salinarum* DpsA in its low- and high-iron states. *Proceedings of the National Academy of Sciences*, 101(38):13780–13785, 2004.

[54] Elango, N., Radhakrishnan, R., Froland, W. A., Wallar, B. J., Earhart, C. A., Lipscomb, J. D., and Ohlendorf, D. H. Crystal structure of the hydroxylase component of methane monooxygenase from *Methylosinus trichosporium* OB3b. *Protein Science*, 6(3):556–568, 1997.

[55] Hindupur, A., Liu, D., Zhao, Y., Bellamy, H. D., White, M. A., and Fox, R. O. The crystal structure of the *Escherichia coli* stress protein YciF. *Protein Science*, 15(11):2605 – 2611, 2006.

[56] Dyer, D. H., Lyle, K. S., Rayment, I., and Fox, B. G. X-ray structure of putative acyl-ACP desaturase DesA2 from *Mycobacterium tuberculosis* H37Rv. *Protein Science*, 14(6):1508–1517, 2005.

[57] Havukainen, H., Haataja, S., Kauko, A., Pulliainen, A. T., Salminen, A., Haikarainen, T., Finne, J., and Papageorgiou, A. C. Structural basis of the zinc- and terbium-mediated inhibition of ferroxidase activity in Dps ferritin-like proteins. *Protein Science*, 17(9):1513–21, 2008.

[58] Rosenzweig, A. C., Brandstetter, H., Whittington, D. A., Nordlund, P., Lippard, S. J., and Frederick, C. A. Crystal structures of the methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): implications for substrate gating and component interactions. *Proteins*, 29(2):141–52, 1997.

[59] Thumiger, A., Polenghi, A., Papinutto, E., Battistutta, R., Montecucco, C., and Zanotti, G. Crystal structure of antigen TpF1 from *Treponema pallidum*. *Proteins: Structure, Function and Genetics*, 62 (3):827–830, 2006.

[60] Hogbom, M. The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass. *Science*, 305(5681):245–248, 2004.

[61] Cooley, R. B., Rhoads, T. W., Arp, D. J., and Karplus, P. A. A diiron protein autogenerates a valine-phenylalanine cross-link. *Science*, 332 (6032):929–929, 2011.

[62] Johnson, E., Cascio, D., Sawaya, M. R., Gingery, M., and Schröder, I. Crystal structures of a tetrahedral open pore ferritin from the hyperthermophilic archaeon *Archaeoglobus fulgidus*. *Structure*, 13(4):637–648, 2005.

[63] Lindqvist, Y., Huang, W., Schneider, G., and Shanklin, J. Crystal structure of $\delta 9$ stearoyl-acyl carrier protein desaturase from castor seed and its relationship to other di-iron proteins. *The EMBO Journal*, 15 (16):4081–92, 1996.

[64] Takagi, H., Shi, D., Ha, Y., Allewell, N. M., and Theil, E. C. Localized unfolding at the junction of three ferritin subunits. A mechanism for iron release? *The Journal of Biological Chemistry*, 273(30):18685–8, 1998.

[65] Joint Center for Structural Genomics. PDB ID: 2oh3, Crystal structure of COG1633: Uncharacterized conserved protein (ZP_00055496.1) from *Magnetospirillum magnetotacticum* MS-1 at 2.00 Å resolution. 2007.

[66] Joint Center for Structural Genomics. PDB ID: 3ez0, Crystal structure of NTF2-like protein of unknown function (YP_270605.1) from *Colwellia psychrerythraea* 34H at 1.61 Å resolution. 2010.

[67] Joint Center for Structural Genomics. PDB ID: 2itb, Crystal structure of putative tRNA-(ms(2)io(6)a)-hydroxylase (NP_744337.1) from *Pseudomonas Putida* KT2440 at 2.05 Å resolution. 2006.

[68] Joint Center for Structural Genomics. PDB ID: 3fse, Crystal structure of two-domain protein containing DJ-1/ThiJ/PfpI-like and ferritin-like domains. (YP_324989.1) from *Anabaena variabilis* ATCC 29413 at 1.90 Å resolution. 2009.

[69] Joint Center for Structural Genomics. PDB ID: 2rcc, Crystal structure of putative class I ribonucleotide reductase (NP_241368.1) from *Bacillus halodurans* at 1.90 Å resolution, 2007.

[70] Luo, J., Liu, D., White, M., and Fox, R. PDB ID: 1jts, DNA protection and binding by *Escherichia coli* DPS protein. 2003.

[71] Ebihara, A., Yokoyama, S., and Kuramitsu, S. PDB ID: 2cwl, Structural and functional analysis of pseudocatalase from *Thermus thermophilus* HB8. 2005.

[72] Kim, Y., Joachimiak, G., Wu, R., Patterson, S., Gornicki, P., and Joachimiak, A. PDB ID: 2qqy, Crystal structure of ferritin-like, diiron-carboxylate proteins from *Bacillus anthracis* str. Ames. 2007.

[73] Osipiuk, J., Evdokimova, E., Kudritska, M., Savchenko, A., Edwards, A., and Joachimiak, A. PDB ID: 2gyq, X-ray crystal structure of YcfI protein, a putative structural protein from *Rhodopseudomonas palustris*. 2006.

[74] Ramagopal, U., Rutter, M., Adams, J., Toro, R., Groshong, C., Sauder, J., Burley, S., and Almo, S. PDB ID: 2qf9, Structure of putative secreted protein DUF305 from *Streptomyces coelicolor*. 2007.

[75] Yu, M., Bursey, E., Radhakannan, T., Kim, C., Kaviratne, T., Woodruff, T., Segelke, B., Lekin, T., Toppani, D., Terwilliger, T., and Hung, L. PDB ID: 2ib0, Crystal structure of a conserved hypothetical protein, rv2844, from *Mycobacterium tuberculosis*. 2006.

[76] Zhang, R., Joachimiak, A., Edwards, A., Savchenko, A., and Skarina, T. PDB ID: 1otk, The 2 A crystal structure of protein paaC from *Escherichia coli*. 2003.

[77] Seattle Structural Genomics Center for Infectious Disease. PDB ID: 3ix6, Crystal structure of thymidylate synthase thyA from *Brucella melitensis*. 2009.

[78] Fu, Z.-Q., Liu, Z.-J., Lee, D., Kelley, L., Chen, L., Tempel, W., Shah, N., Horanyi, P., Lee, H., Habel, J., Dillard, B., Nguyen, D., Chang, S.-H., Zhang, H., Chang, J., Sugar, F., Poole, F., Jr, J., F.E, Adams, M., Rose, J., and Wang, B.-C. PDB ID: 2fzf, Hypothetical protein Pfu-1136390-001 From *Pyrococcus furiosus*. 2006.

[79] Hortolan, L., Saintout, N., Granier, G., Langlois d'Estaintot, B., Manigand, C., Mizunoe, Y., Wai, S., Gallois, B., and Precigoux, G. PDB ID: 1krq, Crystal structure analysis of *Campylobacter jejuni* ferritin. 2002.

[80] Joint Center for Structural Genomics. PDB ID: 4h0a, Crystal structure of a hypothetical protein (SAV1118) from *Staphylococcus aureus* subsp. aureus Mu50 at 1.90 Å resolution. 2012.

[81] Joint Center for Structural Genomics. PDB ID: 1vlg, Crystal structure of Transaldolase (EC 2.2.1.2) (TM0295) from *Thermotoga maritima* at 2.40 Å resolution. 2004.

[82] Welin, M., Ogg, D., Arrowsmith, C., Berglund, H., Busam, R., Collins, R., Edwards, A., Ehn, M., Flodin, S., Flores, A., Graslund, S., Hammarstrom, M., Hallberg, B., Holmberg Schiavone, L., Hogbom, M., Kotenyova, T., Magnusdottir, A., Moche, M., Nilsson-Ehle, P., Nyman,

T., Persson, C., Sagemark, J., Sundstrom, M., Stenmark, P., Uppenberg, J., Thorsell, A., Van Den Berg, S., Wallden, K., Weigelt, J., and Norlund, P. PDB ID: 2uw2, Crystal structure of human ribonucleotide reductase subunit R2. 2007.

[83] Lewin, R. Evolutionary history written in globin genes. *Science*, 214 (4519):426–429, 1981.

[84] Reichard, P. Ribonucleotide reductases: the evolution of allosteric regulation. *Archives of Biochemistry and Biophysics*, 397(2):149–155, 2002.

Chapter 3

Aminoacyl-tRNA synthetases

Overview

The previous chapter introduced, in depth, an approach to recover evolutionary relationships from protein structures. This chapter presents a test of this method on the aminoacyl-tRNA synthetase family. The purpose of this work is to inform current opinions on the evolutionary relationships shared by the synthetase family using protein structures. Despite the ancient origin of this family, aminoacyl-tRNA synthetases have significant sequence similarity allowing for conventional sequence-based methods to infer evolutionary relationships. These known relationships are what make this synthetase family a good control to gauge the effectiveness of this structural method.

The biological role of the aminoacyl-tRNA synthetase family is introduced first, followed by the significance of understanding their evolution to uncover events at the dawn of cellular life. This is followed by a summary of the current state of knowledge regarding their evolution and the intended scope of this work. This is followed by a summary of the methods employed, results and a discussion around the effectiveness of this method and its future use in cases where relationships are not known.

3.1 Aminoacyl-tRNA synthetases

Aminoacyl-tRNA synthetases (aaRSs) constitute an important class of proteins playing a crucial role in the process of decoding the information contained within the DNA. These proteins charge tRNA (transfer RNA) molecules with their cognate amino acids [1] which are then used by the ribosome while translating an mRNA (messenger RNA) molecule into a protein. The following equation shows the charging of a tRNA with an appropriate amino acid.

$$\text{Amino acid} + \text{ATP} + \text{tRNA} \rightleftharpoons \text{aminoacyl}-\text{tRNA} + \text{AMP} + \text{PP}_\text{i}$$

There are 20 different amino acids that are used to build proteins, thus for specificity, 20 different tRNAs and 20 different aaRSs are also required. There are known cases where reduced specificity causes mischarging of tRNAs [2] or where organisms lacking an aaRS make use of a different syn-

Table 3.1: Classification of aminoacyl-tRNA synthetases. The structure column indicates if the structure is a monomer or oligomer i.e. number of respective protein chain and their types i.e. α or β [9] e.g. CysRS is monomeric with one α protein only whereas PheRS is a oligomeric structure with two α and two β chains. Most of the oligomeric structures have one protein only α unit with the exception of GlyRS and PheRS.

Class I	Structure	Class II	Structure
CysRS	α	AlaRS	α_4
MetRS	α_2	GlyRS	$\alpha_2\beta_2$
ValRS	α	SerRS	α_2
IleRS	α	ThrRS	α_2
LeuRS	α	ProRS	α_2
ArgRS	α	HisRS	α_2
GluRS	α	AspRS	α_2
GlnRS	α	AsnRS	α_2
TyrRS	α_2	LysRS	α_2
TrpRS	α_2	PheRS	$\alpha_2\beta_2$

thetase to carry out the necessary job [3, 4].

The aaRSs are grouped into two separate classes i.e. class I and II, see Table 3.1, [5] where class I proteins have a characteristic Rossmann fold [6] and class II synthetases are identified by a repertoire of three sequence motifs which result in an anti-parallel beta fold. Class II aaRSs bears no sequence or structural similarity with class I aaRSs apart from the reaction they catalyse, with the exception [7] of some structural similarity between alanyl (class II) and glutamyl-tRNA (class I) synthetases. This structural dissimilarity between class I and II causes a functional difference with the aminoacylation site being the 2'-OH in case of class I and 3'-OH in case of class II [5, 8] on the ribose of the 3' terminal adenosine.

3.1.1 Evolutionary analysis of aaRSs: What is known so far?

The association of aaRSs with the protein synthesis process and their universal presence across all domains makes them an interesting protein family with which to probe deep evolutionary relationships [10]. The aaRSs can assist with explaining unsolved biology questions like

- evolution of the genetic code [11, 12], i.e. the use of a limited number of amino acids, just 20, in protein synthesis and the redundancy of

codons and

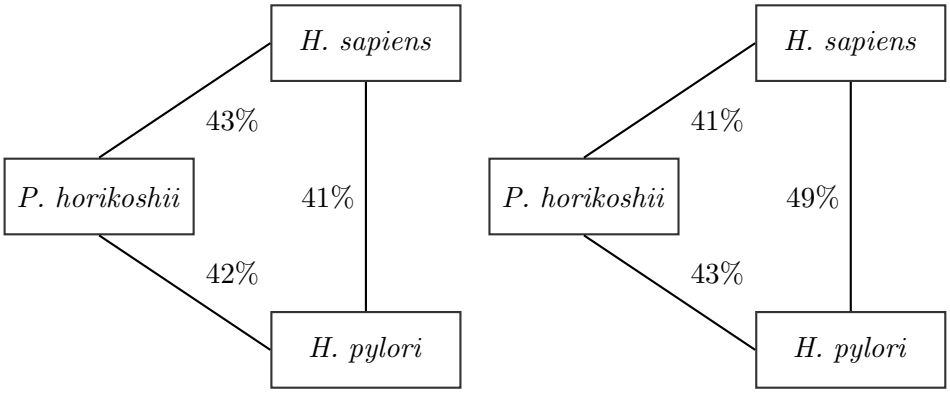
- evolutionary dynamics of the last common ancestor diverging into the three domains of life [11].

Furthermore, class I utilizes the Rossmann fold which is said to be an ancient fold [13–15] and is responsible for the synthesis of some of the hydrophobic residues, see Table 3.1. It could be conjectured here that relatively simplistic protein function, governed by a limited number of amino acids, emerged initially. The hydrophobic nature of these amino acids resulted in the emergence of the "protein folding process" followed by the incorporation of other amino acids, class II, that later resulted in the protein landscape expanding to incorporate more complex functions. These are some of the questions that can be answered by protein families, like the aaRSs, which have roots in the Precambrian era.

Evolutionary analysis of aaRSs [5] has, so far, concluded that the pronounced difference between the two classes i.e. in sequence, structure and specificities to 2' or 3' hydroxyl groups, points to independent origins for each. This is further supported by significant similarities within members of the same class. Two cases are illustrated in Figure 3.1 which show the significant similarity from pairwise sequence comparison of Arginyl-tRNA synthetase in class I and Alanyl-tRNA synthetase in class II. The comparison spans the three domains of life with *Homo sapiens* from eukarya, *Pyrococcus horikoshii* from domain archaea and *Helicobacter pylori* from bacteria. This sequence similarity is also reflected in their corresponding structures, in cases where structures have been experimentally determined.

Substantial conservation at a sequence level in aaRSs is affected by horizontal gene transfer (HGT) which makes recovery of organismal relationships, using aaRS phylogenies, difficult [11, 16–19]. Fortunately, these gene transfer events are easily isolated and removed from the analysis, albeit in a manual way, allowing aaRSs to recreate a phylogeny consistent with other translation related protein families [18].

Apart from recovering organismal evolutionary relationships, the sequence conservation of aaRSs across distant lineages has allowed for the formulation of another conclusion about their evolutionary history, namely the further categorization of class I and II into three canonical subclasses. Class I has Ia (comprising MetRS, ValRS, LeuRS, IleRS, CysRS, ArgRS),



(a) Arginyl-tRNA synthetase - Class I (b) Alanyl-tRNA synthetase - Class II

Figure 3.1: Conservation in class I and II across *H. sapiens*, *P. horikoshii* and *H. pylori*. Pairwise sequence comparison of Arginyl and Alanyl-tRNA synthetase across the three domains shows sequence similarity well above the "twilight zone".

Ib (comprising GluRS, GlnRS) and Ic (comprising TyrRS, TrpRS). Class II has IIa (comprising SerRS, ProRS, HisRS, ThrRS, AlaRS, GlyRS), IIb (comprising LysRS, AsnRS, AspRS) and IIc (comprising PheRS). Like the main classes, these subclasses are also characterised by sequence similarities, with each subclass having its own ancestor descending from the progenitor of the entire class [12].

3.1.2 Mitochondrial aaRSs

The presence of the mitochondria in eukaryotic cells is also of interest as mitochondria have their own set of aaRSs (mt-aaRSs). Genes encoding these mt-aaRSs reside within the nuclear genome [20]. Eukaryotes therefore carry two sets of aaRS genes, cytosolic and mitochondrial (chloroplast in case of plants). Studies have shown that mt-aaRSs are more similar to bacterial aaRSs then they are to eukaryotic ones [20], reflective of their shared origin as suggested by the endosymbiotic hypothesis. This generates interesting insight like the presence of cytosolic aaRSs in the eukaryotic cell prior to the endosymbiotic event. This protein family, aaRSs, therefore is well poised to help uncover a set of events leading to the incorporation of the pre-mitochondrial cell into the pre-eukaryotic cell and leading to a mitochondria as an organelle inside eukaryotic cells.

3.1.3 Structure-based phylogenetics: Recovering the known

The primary purpose of this work is to recreate well established evolutionary relationships using the structural approach as used previously [21] and introduced in the preceeding chapters. The case of aaRSs is of particular interest primarily because of their deep evolutionary history and secondly because of the availability of results from sequence-based analysis that can be used to gauge the effectiveness of the structural method [21]. The aim therefore is to successfully recover established signals like:

- Subcategories of aaRSs classes which are detectable at the sequence level and
- the shared evolutionary origin of bacterial and mitochondrial aaRSs.

will lend confidence to this structure-based phylogenetic method.

3.2 Method

To maximize the evolutionary depth, *H. sapiens* and *S. cerevisiae* in domain eukarya, *E. coli* in domain bacteria, *P. horikoshii* (in euryarchaeota) and *S. tokadii* (in crenarchaeota) in domain archaea were included in this work. Protein databank (PDB) identifiers were collected from UniProt [22] for aaRSs, both class I and II, against the aformentioned organisms using simple word searches (e.g. "tRNA-synthetase") and the corresponding three dimensional structures against those identifiers were obtained from the RCSB database (www.rcsb.org/) [23]. These PDB identifiers are listed in Tables 3.2 and 3.3. Both mitochondrial and cytoplasmic protein structures were collected, where available in case of eukaryotes. In case of multimeric proteins with identical chains only one chain was considered e.g. for MetRS a single α chain was considered. In case of non-identical units comprising a multimer e.g. PheRS a single α and a single β chain was used and treated as separate structures.

The gathered structural data was processed in a manner similar to Lundin et. al [21], using the following steps.

1. Pairwise comparisons were done for each protein structure using Superpose [68]. Due to the nature of the algorithm comparisons were order

Table 3.2: Class I aminoacyl-tRNA synthetases. Sub-cellular localization indicates region where protein is present i.e. cytoplasm, mitochondria or both. Column "AA" lists the amino acid that each aaRS loads onto the tRNA. "X" indicates non-organeller organisms. Colour coded species are for visual purposes only with eukarya in blue, bacteria in orange and archaea in red. [24–45].

Species	**PDB,Chain**	**AA**	**Sub. Localisation**
H. sapiens	4ZAJ,A	Arg	Cytoplasm
	4YE6,A	Gln	Cytoplasm
	2WFD,A	Leu	Cytoplasm
	1R6U,A	Trp	Cytoplasm
	5EKD,A	Trp	Mito. Matrix
	1N3L,A	Tyr	Cytoplasm
	2PID,A	Tyr	Mito. Matrix
S. cerevisiae	1F7U,A	Arg	Cytoplasm
	4H3S,A	Gln	Both
	2IP1,A	Trp	Cytoplasm
	2DLC,X	Tyr	Cytoplasm
E. coli	4OBY,A	Arg	X
	1U0B,B	Cys	X
	1QTQ,A	Gln	X
	4ARC,A	Leu	X
	3H99,A	Met	X
	5V0I,A	Trp	X
	2YXN,A	Tyr	X
P. horikoshii	2ZUE,A	Arg	X
	1WKB,A	Leu	X
	3JXE,A	Trp	X
	2CYC,A	Tyr	X

specific i.e. A $\cong$ B $\neq$ B $\cong$ A. The pairwise scores were, therefore, averaged to attain a final score "q" for the comparison between structures A and B.

2. The similarity score was subtracted from one (i.e. d = 1 - q) to generate distance (d).

3. A matrix was populated with the pairwise distances.

4. The neighbour-joining (NJ) algorithm [69] as implemented by the Phylo [70] package in Biopython [71] was used to generate an unrooted NJ-

Table 3.3: Class II aminoacyl-tRNA synthetases. The table is similar to table 3.2. [46–67]

Species	PDB,Chain	AA	Sub. Localisation
H. sapiens	4XEM,A	Ala	Cytoplasm
	4J15,A	Asp	Cytoplasm
	4AH6,A	Asp	Mito. Matrix
	2ZT5,A	Gly	Cytoplasm
	4G84,A	His	Cytoplasm
	4YCU,A	Lys	Cytoplasm
	3L4G,A	Phe	Cytoplasm
	3L4G,B	Phe	Cytoplasm
	3CMQ,A	Phe	Mito. Matrix
	4L87,A	Ser	Cytoplasm
	4HWT,A	Thr	Cytoplasm
S. cerevisiae	1EOV,A	Asp	Cytoplasm
	3UH0,A	Thr	Mito. Matrix
E. coli	3HY0,A	Asp	X
	1HTT,A	His	X
	1BBU,A	Lys	X
	3PCO,A	Phe	X
	3PCO,B	Phe	X
P. horikoshii	2ZZE,A	Ala	X
	1X54,A	Asn	X
	2CXI,A	Phe	X
	2DQ0,A	Ser	X
S. tokodii	1WYD,A	Asp	X

tree from the matrix. The tree was then visualized using Figtree [72].

5. Splitstree [73] was used to obtain a neighbour-net network from the matrix in 3.

3.3 Results

The pairwise distances, as discussed in the method section, between structures were used to construct an un-rooted phylogenetic tree using the neighbor-joining algorithm, see Figures 3.2 and 3.4. The same data was used to generate a neighbor-net splits network, see Figures 3.3 and 3.5. The purpose of generating the network is to observe the tree likeness of the data.

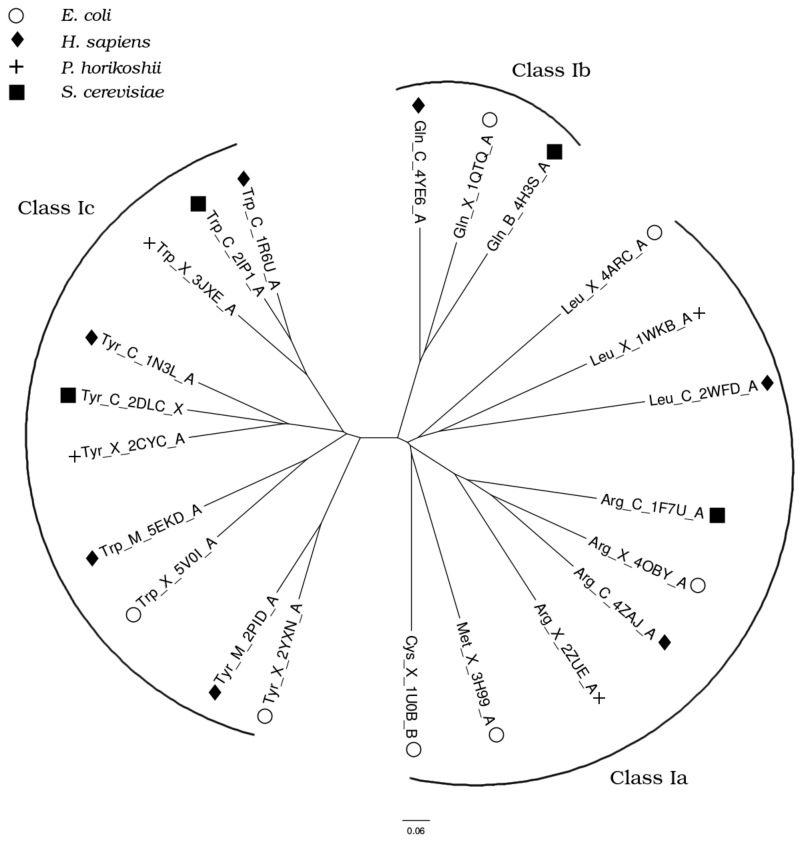


Figure 3.2: Class I aaRSs: An unrooted neighbor-joining phylogenetic tree where each label includes the amino acid synthesized, the sub-cellular localization (Cytoplasm(C), Mitochondria(M), non-organeller species (X)), PDB identifier and the chain used.

Deviation from tree like networks opens the distance data to alternate inferences. A quick view of Figures 3.3 and 3.5 reveal nearly tree like networks for both classes of aaRSs. This network appearance both for class I and II precludes indications of significant alternative phylogenetic inferences. A comprehensive analysis, which follows, recovers most of the relationships established from sequence analysis. A macro observation, for example, reveals that in cases where multiple aaRSs responsible for generating the same amino acid were present, a grouping is observed across the three domains of life i.e. clustering is observed for same aaRSs across the three domains of life, with the few exceptions discussed in detail in sections 3.3.1 and 3.3.2.

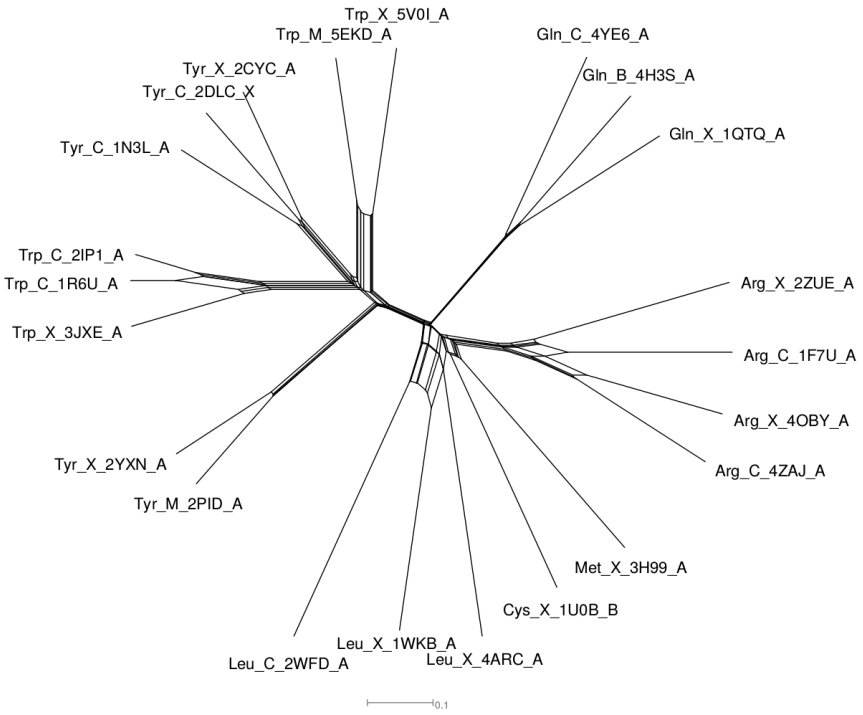


Figure 3.3: Class I aaRSs: A representation of the structural data as a neighbornet network generated using Splitstree. Each label includes the amino acid synthesized, the sub-cellular localization (Cytoplasm(C), Mitochondria(M), non-organeller species (X)), PDB identifier and the chain used.

3.3.1 Subclasses of aaRSs

The purpose of this study was to use the structural methods to recover known relationships, inferred from sequence-based analysis. The structure-based method should be able to successively recover, based on structural information only, three subcategories in each of the classes I and II. Figure 3.2 reveals that from amongst class I aaRSs, the Ia, Ib and Ic clusters are successfully recovered. This is illustrated by the arcs with labels around the leaves of the tree where the clades can be traced back to a rooting. This is the same result as recovered by sequence-based analysis where each of the three subcategories, in class I, descends from its own ancestor. These subcategory ancestors in turn share a common ancestor. Meanwhile for class II, see Figure 3.3, only IIb is recovered, whereas the remaining two subcategories

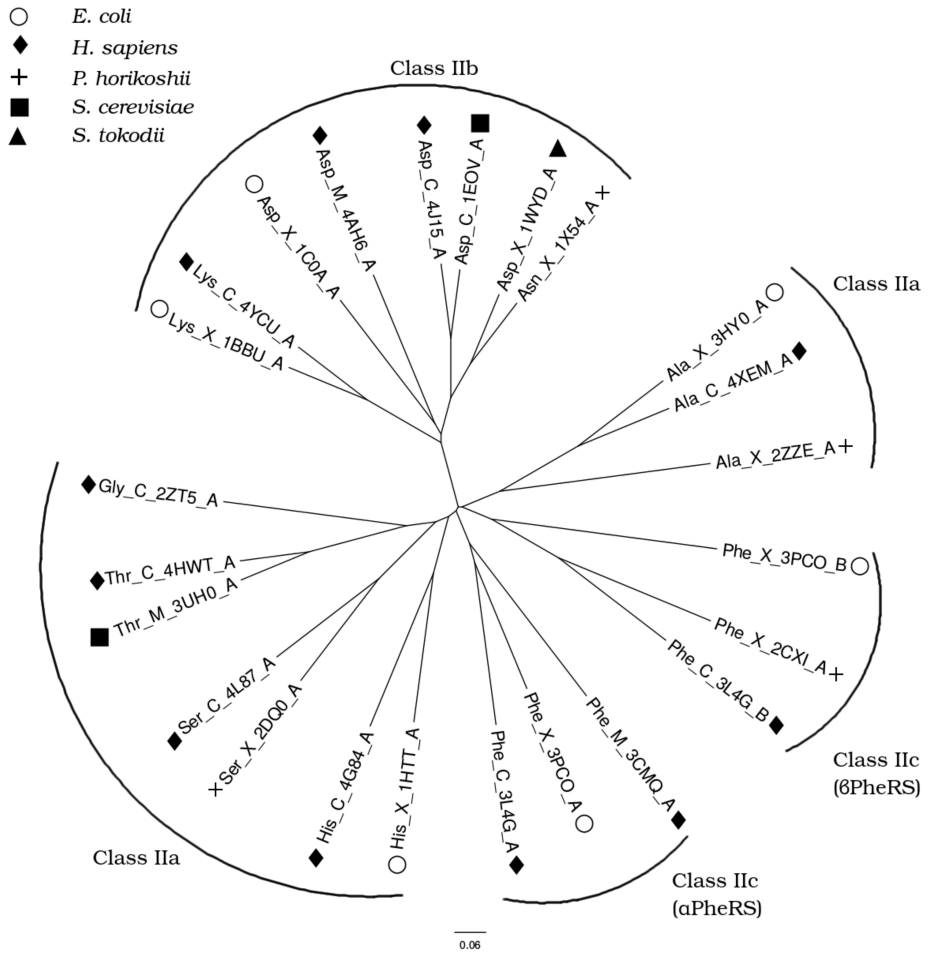


Figure 3.4: Class II aaRSs: An unrooted neighbor-joining phylogenetic tree where each label includes the amino acid synthesized, the sub-cellular localization (Cytoplasm(C), Mitochondria(M), non-organeller species (X)), PDB identifier and the chain used.

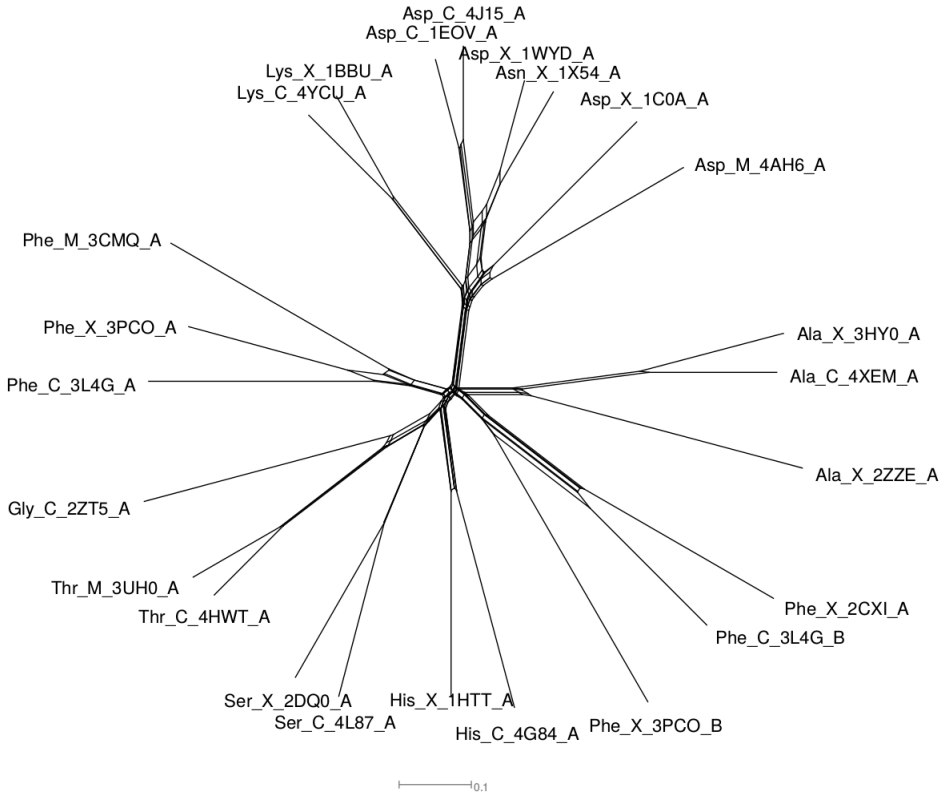


Figure 3.5: Class II aaRSs: A representation of the structural data as a neighbornet network generated using Splitstree. Each label includes the amino acid synthesized, the sub-cellular localization(Cytoplasm(C), Mitochondria(M), non-organeller species (X)), PDB identifier and the chain used.

IIa and IIc show some discrepancy; resolution of which requires further discussion as follows.

Tables 3.2 and 3.3, list canonical placements of aaRSs which do not hold when the breadth of data considered in sequence analysis is increased to include more species. Almost all of the aaRSs in class II are homodimers, see Table 3.1. PheRS is one of the aaRSs that is an exception to this and comprises two α and two β subunits each. These are shown in the subclass IIc (αPheRS and βPheRS) cluster in Figure 3.4. Sequence-based analysis qualifies PheRS to be divergent enough to separate it into an individual subclass (IIc). The classification of GlyRS is also subject to change depending on the quaternary structure, i.e. a tetramer ($\alpha_2\beta_2$, canonical) or homodimer

(α_2, non-canonical), where non-canonical form is placed in IIa [74] and the canonical one in IIc [75]. Conflict also exists in the placement of AlaRS in subclass IIa [76] or IIc [1, 77] because of its homotetrameric nature. The primary reason for the conflict in the placement of three of the class II aaRSs (AlaRS, GlyRS and PheRS) is because of significant sequence divergence, in these three cases. The subcategories are originally driven by sequence analysis which successfully classifies proteins to a subclass. However this sequence analysis fails depending on the source organism for the three aaRSs mentioned above as a result of which they are grouped into a separate subclass (IIc). There is some other discussion [78] around the formation of a new subclass IId to account for proteins not sharing a common ancestor i.e. in the case of PheRS and AlaRS.

The groupings seen here are due to the use of purely structural analysis and the choice of subunits. Figure 3.4 place GlyRS (α unit) with other aaRSs from subclass IIa and separates out AlaRS and PheRS in a way which indicates a separate group formation. These observations are consistent with [75]. The tree in Figure 3.4 shows an AlaRS to be closer to βPheRS forming a separate group as compared to αPheRS and class IIa aaRSs. This observation should not be over interpreted for two reasons (a) βPheRSs are not well-studied and hence no results, to the best of my knowledge, are present in literature for comparison purposes and (b) the splits network in Figure 3.5 is not completely resolved where the branches for the substructure emerge which implies more than one inference can be made. Figure 3.2 shows one of a number of possibilities when resolving ancestral splits.

This method of using structure-based analysis therefore successfully recovers relationships for class I aaRSs and captures some non-canonical ones for class II.

3.3.2 Cytoplasmic, Mitochondrial and Bacterial aaRS

This dataset included protein structures from all three domains of life. Moreover in the case of eukaryotes in some instances, mitochondrial aaRSs were included along with cytoplasmic ones. Figures 3.2 and 3.4 show bacterial and mitochondrial aaRSs sharing a clade which is distant from synthetases localised in the cytoplasm of eukaryotes or from archaea. An example of this can be seen for TrpRS and TyrRS in subclass Ic cluster in Figure 3.2. In case of TrpRS mitochondrial structure (5EKD , chain A from

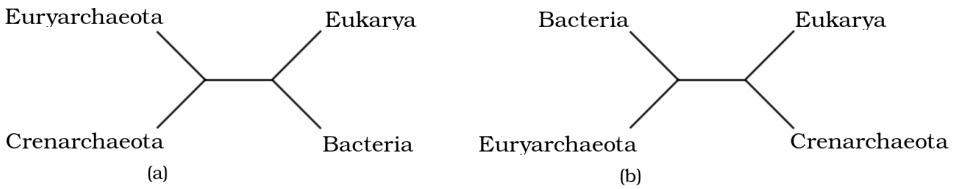


Figure 3.6: The three domain and eocyte trees. (a) The unrooted three domain tree treats Euryarchaeota and Crenarchaeota (eocytes) as monophyletic (b) the eocyte tree divides the archaeal domain with Crenarchaeota related to Eukarya and Euryarchaeota closely grouping with Bacteria.

H. sapiens) is seen closely related to its bacterial counterpart (5V0I, chain A from *E. coli*) and more distantly from its eukaryotic cytoplasmic (1R6U, chain A from *H. sapiens*; 2IP1, chain A from *S. cerevisea*) and archaeal (3JXE, chain A from *P. horikoshii*) counterparts. The same is true for TyrRS.

Mitochondrial aaRSs have been extensively studied [79, 80] and their evolutionary origin probed [20, 81–83], however in the context of structural similarity, literature presents clear evidence of a higher degree of similarity between mitochondrial and bacterial aaRSs [84] compared to mitochondrial and their cytoplasmic counterparts. A similar case can be observed for class II aaRSs in the subclass IIb cluster. This method therefore successfully recovers this evolutionary signal considering only the structural information of these proteins.

3.3.3 Eocyte hypothesis

Further to the satisfactory recovery of known evolutionary signals, to validate the structural phylogeny method, this work can be used to generate insight on the Eocyte hypothesis. Competing arguments exist regarding the structure of the tree of life with two being the three monophyletic domains, according to the Woese system [85] and an alternate topology proposed by Lake [86], i.e. the Eocyte hypothesis. Lake argued that only crenarchaeota, previously known as eocytes, from domain archaea were monophyletic with eukaryotes as opposed to the canonical Woese classification, where the complete domain archaea (comprising Euryarchaeota and Crenarchaeota) are monophyletic, see Figure 3.6. Considerable evidence is present in the lit-

erature both for [87–90] and against [91, 92] with some studies unable to support [93, 94] one or the other.

The class II aaRS presented an opportunity to probe support for one of the two topologies in Figure 3.6. The six structures in the Figure 3.4 class IIb cluster (i.e. AspRS and AsnRS) originate from *E. coli* (PDB ID: 1C0A), *H. sapiens* mitochondria (PDB ID: 4AH6), *H. sapiens* cytoplasm (PDB ID: 4J15), *S. cerevisiae* cytoplasm (PDB ID: 1EOV), *S. tokodii* (Crenarchaeota, PDB ID: 1WYD) and *P. horikoshii* (Euryarchaeota, PDB ID: 1X54). In this case AspRS from *P. horikoshii* (euryarchaeota) and AsnRS from *S. tokodii* (crenarchaeota) are considered equivalent. The primary reason for this equivalence is the structural similarity, as reflected in the small distance between them. Moreover *S. tokodii* does not have an AsnRS and instead uses the mischarging of AspRS to charge the Asn-tRNA [95, 96]. It would also be worth noting here that AsnRS is absent in a majority of bacteria and archaea, where the same function is performed by non-discriminating AspRS [3, 4]. Results shown in Figure 3.4 are consistent with this observation.

As discussed earlier, mitochondrial aaRSS show higher similarity to bacterial synthetases as opposed to their cytoplasmic counterparts which is also seen here where *E. coli* AspRS and *H. sapiens* mitochondrial AspRS are monophyletic, Figure 3.7. This grouping is distant from the remaining four structures, where archaeal AspRS and AsnRS (1WYD, chain A from crenarchaeota and 1X54, chain A from euryarchaeota) form a group and share a clade with eukaryotic aaRSs (4J15, chain A from *H. sapiens* and 1EOV, chain A from *S. cerevisea*). This topology of bacterial, archaeal and eukaryotic aaRSs, illustrated in Figure 3.7, lends support to the Woese classification of the three domains where all archaea (euryarchaeota and crenarchaeota) are monophyletic and share a higher degree of similarity with eukarya as opposed to bacteria. In contrast to this the presence of *P. horikoshii* AspRS in the bacterial-mitochondrial group would have presented evidence in favour of eocyte tree, which is not the case.

3.4 Discussion

In this work, the structural method [21] to probe evolutionary relationships using protein structures was used to recover well studied relationships

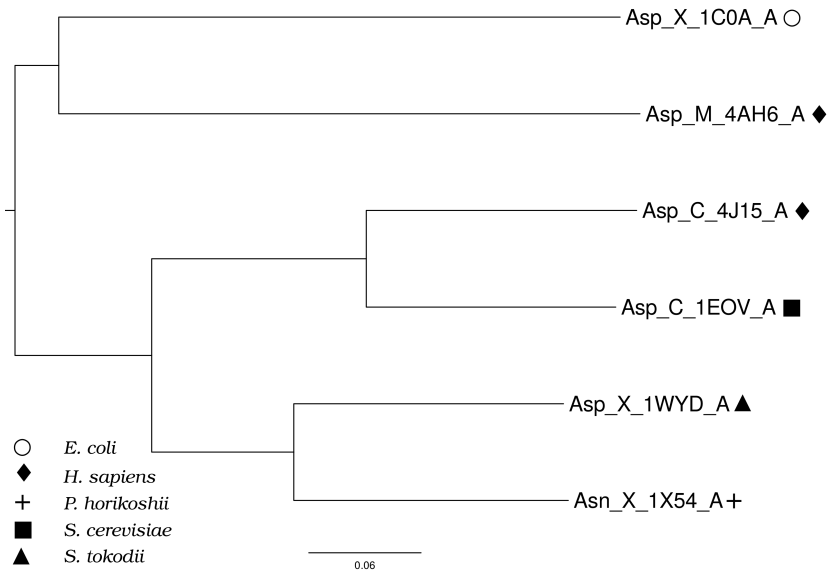


Figure 3.7: A selection of structures from class IIb in Figure 3.4 to illustrate support for the Woese classification of the three domains of life.

between aaRSs. The synthetases are one of the most ancient protein families considering the role they play in living organisms. Due to the conservation of their role, they have been evolutionarily conserved across the three domains of life and provide an opportunity for us to probe deep evolutionary relationships. The primary focus of this method is to assist in the recovery of deep evolutionary relationships from datasets in which the evolutionary signal is too weak to be probed with conventional sequence-based methods. Due to substantial conservation at the sequence level, in this instance the method is used to recover well established relationships derived from sequence-based analysis. This dataset therefore acts as a control and successful recovery of evolutionary signals lends confidence to the predictions made by this method in cases where sequence-based inferences are not possible.

As discussed in the previous section, the method reasonably recovers tree like networks. The structural analysis performed here used SSM-based Q_{score} as the primary metric to quantify distance between protein structures and is coupled to the NJ algorithm to recover phylogenies. This method is successful in terms of (a) recovering the substructure in the aaRSs classification and (b) recovering the known relationships between cytoplasmic,

mitochondrial and bacterial aaRSs. Point (a) is of considerable significance because previously used methods to quantify structural distance for use in phylogenetic inference disagreed with sequence-based methods [19], whereas the choice of Q_{score} agrees well with classifications determined by sequence analysis.

Furthermore, it is well established that mitochondrial aaRSs group closely with bacterial aaRSs, an observation which is also recovered. Moreover, each functional cluster is observed, i.e. aaRSs responsible for charging tRNAs with the same amino acid across species group together. This functional clustering leads into the recovery of well-established substructure of aaRSs. For class I, the presence of the Rossman fold reveals canonical relationships formed from sequence analysis whereas for class II, near canonical relationships are recovered for two of the three subclasses with the deviation in the third explained by sequence and quaternary structure variation, in the previous section.

The success of this method opens up a new area of exploring deep evolutionary relationships which could previously not be analysed at a sequence level due to extreme sequence divergence. An example of this was presented, namely using the structural approach and the class II aaRSs to explore two competing descriptions of the organisation of the domains of life i.e. the three domain Woese tree and eocyte hypothesis. In this case only a single species from each of euryarchaeota and crenarchaeota was present in the data due to a lack of structures. Choosing one classification over the other purely based on a single data point would be incorrect, however, due to the relatively high success rate of recovering established relationships, once more structures of aaRSs are available from both euryarchaeota and crenarchaeota it is anticipated that a better picture will emerge with conclusive evidence either in support or against one of the classifications.

3.5 Future Work

The sequence conservation in aaRSs allowed for a control against which the results from this structural analysis could be validated. More datasets like this need to be probed to further benchmark this method. Secondly, the structural comparison metric used in this analysis performed better than some of the others used in the past [19]. However this metric at present does

not penalize gap opening and extensions in structural alignments. This is something that could be added to recover better informed phylogenies. Thirdly, the topology of the tree needs to be validated. The MD-based bootstrap method developed as a part of this thesis can be employed on this dataset. Simulations, once carried out, could be used to support nodes and search for alternate trees that could result in offering further insight into resolving the phylogenies of class II aaRSs. These are some of the future directions of this work.

Bibliography

[1] Ibba, M. and Söll, D. Aminoacyl-tRNA synthesis. *Annual Review of Biochemistry*, 69(1):617–650, 2000.

[2] Schwartz, M. H. and Pan, T. Determining the fidelity of tRNA aminoacylation via microarrays. *Methods*, 113:27–33, 2017.

[3] Kern, D., Roy, H., and Becker, H. D. Asparaginyl-tRNA synthetases. *Madame Curie Bioscience Database. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6048/*, 2013.

[4] Iwasaki, W., Sekine, S.-i., Kuroishi, C., Kuramitsu, S., Shirouzu, M., and Yokoyama, S. Structural basis of the water-assisted asparagine recognition by asparaginyl-tRNA synthetase. *Journal of Molecular Biology*, 360(2):329–342, 2006.

[5] Eriani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. Partition of tNRA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature*, 347(6289):203, 1990.

[6] Rossmann, M. G., Moras, D., and Olsen, K. W. Chemical and biological evolution of nucleotide-binding protein. *Nature*, 250(463):194–199, 1974.

[7] de Pouplana, L. R., Buechter, D., Sardesai, N. Y., and Schimmel, P. Functional analysis of peptide motif for RNA microhelix binding suggests new family of RNA-binding domains. *The EMBO Journal*, 17 (18):5449–5457, 1998.

[8] Delarue, M. Aminoacyl-tRNA synthetases. *Current Opinion in Structural Biology*, 5(1):48–55, 1995.

[9] Carter Jr, C. W. Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. *Annual Review of Biochemistry*, 62(1): 715–748, 1993.

[10] Brown, J. R. and Doolittle, W. F. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proceedings of the National Academy of Sciences*, 92(7):2441–2445, 1995.

[11] Woese, C. R., Olsen, G. J., Ibba, M., and Söll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, 64(1):202–236, 2000.

[12] de Pouplana, L. R. and Schimmel, P. Aminoacyl-tRNA synthetases: potential markers of genetic code development. *Trends in Biochemical Sciences*, 26(10):591–596, 2001.

[13] Xie, L. and Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. *Proceedings of the National Academy of sciences*, 105(14): 5441–5446, 2008.

[14] Duax, W. L., Huether, R., Pletnev, V., Umland, T. C., and Weeks, C. M. Divergent evolution of a Rossmann fold and identification of its oldest surviving ancestor. *International Journal of Bioinformatics Research and Applications*, 5(3):280–294, 2009.

[15] Laurino, P., Tóth-Petróczy, Á., Meana-Pañeda, R., Lin, W., Truhlar, D. G., and Tawfik, D. S. An ancient fingerprint indicates the common ancestry of Rossmann-fold enzymes utilizing different ribose-based cofactors. *PLoS Biology*, 14(3):e1002396, 2016.

[16] Cusack, S., Härtlein, M., and Leberman, R. Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. *Nucleic Acids Research*, 19(13):3489–3498, 1991.

[17] Wetzel, R. Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code. *Journal of Molecular Evolution*, 40(5):545–550, 1995.

[18] Wolf, Y. I., Aravind, L., Grishin, N. V., and Koonin, E. V. Evolution of aminoacyl-tRNA synthetases - analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Research*, 9(8):689–710, 1999.

[19] O'Donoghue, P. and Luthey-Schulten, Z. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiology and Molecular Biology Reviews*, 67(4):550–573, 2003.

[20] Brindefalk, B., Viklund, J., Larsson, D., Thollesson, M., and Andersson, S. G. E. Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. *Molecular Biology and Evolution*, 24(3):743–756, 2006.

[21] Lundin, D., Poole, A. M., Sjöberg, B.-M., and Högbom, M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *Journal of Biological Chemistry*, 287(24):20565–20575, 2012.

[22] Consortium, U. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158 – D169, 2017.

[23] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. In *International Tables for Crystallography Volume F: Crystallography of Biological Macromolecules*, pages 675–684. Springer, 2006.

[24] Smith, A. and Rosenzweig, A. PDB ID: 4zaj, 2.2 Å crystal structure of a human arginyl-tRNA synthetase. 2014.

[25] Ognjenović, J., Wu, J., Matthies, D., Baxa, U., Subramaniam, S., Ling, J., and Simonović, M. The crystal structure of human GlnRS provides basis for the development of neurological disorders. *Nucleic Acids Research*, 44(7):3420–3431, 2016.

[26] Seiradake, E., Mao, W., Hernandez, V., Baker, S. J., Plattner, J. J., Alley, M. R. K., and Cusack, S. Crystal structures of the human and fungal cytosolic Leucyl-tRNA synthetase editing domains: a structural basis for the rational design of antifungal benzoxaboroles. *Journal of Molecular Biology*, 390(2):196–207, 2009.

[27] Yang, X.-L., Guo, M., Kapoor, M., Ewalt, K. L., Otero, F. J., Skene, R. J., McRee, D. E., and Schimmel, P. Functional and crystal structure analysis of active site adaptations of a potent anti-angiogenic human tRNA synthetase. *Structure*, 15(7):793–805, 2007.

[28] Williams, T. and Carter, C. PDB ID: 5ekd, Binding of Mg2+ATP enhances inhibition of human mitochondrial tryptophanyl-tRNA synthetase by indolmycin. 2015.

[29] Yang, X.-L., Skene, R. J., McRee, D. E., and Schimmel, P. Crystal structure of a human aminoacyl-tRNA synthetase cytokine. *Proceedings of the National Academy of Sciences*, 99(24):15369–15374, 2002.

[30] Bonnefond, L., Frugier, M., Touzé, E., Lorber, B., Florentz, C., Giegé, R., Sauter, C., and Rudinger-Thirion, J. Crystal structure of human mitochondrial tyrosyl-tRNA synthetase reveals common and idiosyncratic features. *Structure*, 15(11):1505–1516, 2007.

[31] Delagoutte, B., Moras, D., and Cavarelli, J. tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. *The EMBO Journal*, 19(21):5599–5610, 2000.

[32] Grant, T. D., Luft, J. R., Wolfley, J. R., Snell, M. E., Tsuruta, H., Corretore, S., Quartley, E., Phizicky, E. M., Grayhack, E. J., and Snell, E. H. The structure of yeast glutaminyl-tRNA synthetase and modeling of its interaction with tRNA. *Journal of Molecular Biology*, 425(14): 2480–2493, 2013.

[33] Malkowski, M. G., Quartley, E., Friedman, A. E., Babulski, J., Kon, Y., Wolfley, J., Said, M., Luft, J. R., Phizicky, E. M., and DeTitta, G. T. Blocking S-adenosylmethionine synthesis in yeast allows selenomethionine incorporation and multiwavelength anomalous dispersion phasing. *Proceedings of the National Academy of Sciences*, 104(16):6678–6683, 2007.

[34] Tsunoda, M., Kusakabe, Y., Tanaka, N., Ohno, S., Nakamura, M., Senda, T., Moriguchi, T., Asai, N., Sekine, M., and Yokogawa, T. Structural basis for recognition of cognate tRNA by tyrosyl-tRNA synthetase from three kingdoms. *Nucleic Acids Research*, 35(13):4289–4300, 2007.

[35] Bi, K., Zheng, Y., Gao, F., Dong, J., Wang, J., Wang, Y., and Gong, W. Crystal structure of *Escherichia coli* arginyl-tRNA synthetase and ligand binding studies revealed key residues in arginine recognition. *Protein & Cell*, 5(2):151, 2014.

[36] Hauenstein, S., Zhang, C.-M., Hou, Y.-M., and Perona, J. J. Shape-selective RNA recognition by cysteinyl-tRNA synthetase. *Nature Structural & Molecular Biology*, 11(11):1134–1141, 2004.

[37] Rath, V. L., Silvian, L. F., Beijer, B., Sproat, B. S., and Steitz, T. A. How glutaminyl-tRNA synthetase selects glutamine. *Structure*, 6(4): 439–449, 1998.

[38] Palencia, A., Crépin, T., Vu, M. T., Lincecum Jr, T. L., Martinis, S. A., and Cusack, S. Structural dynamics of the aminoacylation and proof-reading functional cycle of bacterial leucyl-tRNA synthetase. *Nature Structural & Molecular Biology*, 19(7):677–684, 2012.

[39] Schmitt, E., Tanrikulu, I. C., Yoo, T. H., Panvert, M., Tirrell, D. A., and Mechulam, Y. Switching from an induced-fit to a lock-and-key mechanism in an aminoacyl-tRNA synthetase with modified specificity. *Journal of Molecular Biology*, 394(5):843–851, 2009.

[40] Center for Structural Genomics of Infectious Diseases. PDB ID: 5v0i, Crystal Structure of tryptophanyl-tRNA Synthetase from *Escherichia coli* Complexed with AMP and Tryptophan. 2017.

[41] Iraha, F., Oki, K., Kobayashi, T., Ohno, S., Yokogawa, T., Nishikawa, K., Yokoyama, S., and Sakamoto, K. Functional replacement of the endogenous tyrosyl-tRNA synthetase - tRNA-Tyr pair by the archaeal tyrosine pair in *Escherichia coli* for genetic code expansion. *Nucleic Acids Research*, 38(11):3682–3691, 2010.

[42] Konno, M., Sumida, T., Uchikawa, E., Mori, Y., Yanagisawa, T., Sekine, S.-i., and Yokoyama, S. Modeling of tRNA-assisted mechanism of Arg activation based on a structure of Arg-tRNA synthetase, tRNA, and an ATP analog (ANP). *The FEBS Journal*, 276(17):4763–4779, 2009.

[43] Fukunaga, R. and Yokoyama, S. Crystal structure of leucyl-tRNA synthetase from the archaeon Pyrococcus horikoshii reveals a novel editing domain orientation. *Journal of Molecular Biology*, 346(1):57–71, 2005.

[44] Dong, X., Zhou, M., Zhong, C., Yang, B., Shen, N., and Ding, J. Crystal structure of Pyrococcus horikoshii tryptophanyl-tRNA synthetase and structure-based phylogenetic analysis suggest an archaeal origin of tryptophanyl-tRNA synthetase. *Nucleic Acids Research*, 38(4):1401, 2010.

[45] Kuratani, M., Sakai, H., Takahashi, M., Yanagisawa, T., Kobayashi, T., Murayama, K., Chen, L., Liu, Z.-J., Wang, B.-C., and Kuroishi, C. Crystal structures of tyrosyl-tRNA synthetases from archaea. *Journal of Molecular Biology*, 355(3):395–408, 2006.

[46] Zhou, H., W., H., and YANG, X. PDB ID: 4xem, Crystal Structure of wild type human AlaRS catalytic domain. 2014.

[47] Kim, K. R., Park, S. H., Kim, H. S., Rhee, K. H., Kim, B.-G., Kim, D. G., Park, M. S., Kim, H.-J., Kim, S., and Han, B. W. Crystal structure of human cytosolic aspartyl-tRNA synthetase, a component of multi-tRNA synthetase complex. *Proteins: Structure, Function, and Bioinformatics*, 81(10):1840–1846, 2013.

[48] Neuenfeldt, A., Lorber, B., Ennifar, E., Gaudry, A., Sauter, C., Sissler, M., and Florentz, C. Thermodynamic properties distinguish human mitochondrial aspartyl-tRNA synthetase from bacterial homolog with same 3D architecture. *Nucleic Acids Research*, 41(4):2698–2708, 2012.

[49] Guo, R.-T., Chong, Y. E., Guo, M., and Yang, X.-L. Crystal structures and biochemical analyses suggest a unique mechanism and role for human glycyl-tRNA synthetase in Ap4A homeostasis. *Journal of Biological Chemistry*, 284(42):28968–28976, 2009.

[50] Xu, Z., Wei, Z., Zhou, J. J., Ye, F., Lo, W.-S., Wang, F., Lau, C.-F., Wu, J., Nangle, L. A., and Chiang, K. P. Internally deleted human tRNA synthetase suggests evolutionary pressure for repurposing. *Structure*, 20(9):1470–1477, 2012.

[51] Fang, P., Han, H., Wang, J., Chen, K., Chen, X., and Guo, M. Structural basis for specific inhibition of tRNA synthetase by an ATP competitive inhibitor. *Chemistry & Biology*, 22(6):734–744, 2015.

[52] Finarov, I., Moor, N., Kessler, N., Klipcan, L., and Safro, M. G. Structure of human cytosolic phenylalanyl-tRNA synthetase: evidence for kingdom-specific design of the active sites and tRNA binding patterns. *Structure*, 18(3):343–353, 2010.

[53] Klipcan, L., Levin, I., Kessler, N., Moor, N., Finarov, I., and Safro, M. The tRNA-induced conformational activation of human mitochondrial phenylalanyl-tRNA synthetase. *Structure*, 16(7):1095–1104, 2008.

[54] Xu, X., Shi, Y., and Yang, X.-L. Crystal structure of human Seryl-tRNA synthetase and Ser-SA complex reveals a molecular lever specific to higher eukaryotes. *Structure*, 21(11):2078–2086, 2013.

[55] Teng, M., Hilgers, M. T., Cunningham, M. L., Borchardt, A., Locke, J. B., Abraham, S., Haley, G., Kwan, B. P., Hall, C., and Hough, G. W. Identification of bacteria-selective threonyl-tRNA synthetase substrate inhibitors by structure-based design. *Journal of Medicinal Chemistry*, 56(4):1748–1760, 2013.

[56] Sauter, C., Lorber, B., Cavarelli, J., Moras, D., and Giegé, R. The free yeast aspartyl-tRNA synthetase differs from the tRNA Asp-complexed enzyme by structural changes in the catalytic site, hinge region, and anticodon-binding domain. *Journal of Molecular Biology*, 299(5):1313–1324, 2000.

[57] Ling, J., Peterson, K. M., Simonović, I., Cho, C., Söll, D., and Simonović, M. Yeast mitochondrial threonyl-tRNA synthetase recognizes tRNA isoacceptors by distinct mechanisms and promotes CUN codon reassignment. *Proceedings of the National Academy of Sciences*, 109(9):3281–3286, 2012.

[58] Guo, M., Chong, Y. E., Shapiro, R., Beebe, K., Yang, X.-L., and Schimmel, P. Paradox of mistranslation of serine for alanine caused by AlaRS recognition dilemma. *Nature*, 462(7274):808–812, 2009.

[59] Eiler, S., Dock-Bregeon, A.-C., Moulinier, L., Thierry, J.-C., and Moras, D. Synthesis of aspartyl-tRNA Asp in *Escherichia coli* - a snapshot of the second step. *The EMBO Journal*, 18(22):6532–6541, 1999.

[60] Arnez, J. G., Harris, D. C., Mitschler, A., Rees, B., Francklyn, C. S., and Moras, D. Crystal structure of histidyl-tRNA synthetase from *Escherichia coli* complexed with histidyl-adenylate. *The EMBO Journal*, 14(17):4143, 1995.

[61] Onesti, S., Desogus, G., Brevet, A., Chen, J., Plateau, P., Blanquet, S., and Brick, P. Structural studies of lysyl-tRNA synthetase: conformational changes induced by substrate binding. *Biochemistry*, 39(42): 12853–12861, 2000.

[62] Mermershtain, I., Finarov, I., Klipcan, L., Kessler, N., Rozenberg, H., and Safro, M. G. Idiosyncrasy and identity in the prokaryotic phe-system: Crystal structure of *Escherichia coli* phenylalanyl-tRNA synthetase complexed with phenylalanine and AMP. *Protein Science*, 20 (1):160–167, 2011.

[63] Sokabe, M., Ose, T., Nakamura, A., Tokunaga, K., Nureki, O., Yao, M., and Tanaka, I. The structure of alanyl-tRNA synthetase with editing domain. *Proceedings of the National Academy of Sciences*, 106(27): 11028–11033, 2009.

[64] Iwasaki, W., Sekine, S.-i., Kuroishi, C., Kuramitsu, S., Shirouzu, M., and Yokoyama, S. Structural basis of the water-assisted asparagine recognition by asparaginyl-tRNA synthetase. *Journal of Molecular Biology*, 360(2):329–342, 2006.

[65] Sasaki, H. M., Sekine, S.-i., Sengoku, T., Fukunaga, R., Hattori, M., Utsunomiya, Y., Kuroishi, C., Kuramitsu, S., Shirouzu, M., and Yokoyama, S. Structural and mutational studies of the amino acid-editing domain from archaeal/eukaryal phenylalanyl-tRNA synthetase. *Proceedings of the National Academy of Sciences*, 103(40):14744–14749, 2006.

[66] Itoh, Y., Sekine, S.-i., Kuroishi, C., Terada, T., Shirouzu, M., Kuramitsu, S., and Yokoyama, S. Crystallographic and mutational stud-

ies of seryl-tRNA synthetase from the archaeon Pyrococcus horikoshii. *RNA Biology*, 5(3):169–177, 2008.

[67] Sato, Y., Maeda, Y., Shimizu, S., Hossain, M. T., Ubukata, S., Suzuki, K., Sekiguchi, T., and Takenaka, A. Structure of the nondiscriminating aspartyl-tRNA synthetase from the crenarchaeon *Sulfolobus tokodaii* strain 7 reveals the recognition mechanism for two different tRNA anticodons. *Acta Crystallographica Section D: Biological Crystallography*, 63(10):1042–1047, 2007.

[68] Krissinel, E. and Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12): 2256–2268, 2004.

[69] Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[70] Talevich, E., Invergo, B. M., Cock, P. J. A., and Chapman, B. A. Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13(1):209, 2012.

[71] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., and Wilczynski, B. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[72] Rambaut, A. FigTree, a graphical viewer of phylogenetic trees. *See http://tree.bio.ed.ac.uk/software/figtree*, 2007.

[73] Huson, D. H. and Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[74] Francklyn, C., Perona, J. J., Puetz, J., and Hou, Y.-M. Aminoacyl-tRNA synthetases: versatile players in the changing theater of translation. *RNA*, 8(11):1363–1372, 2002.

[75] Cusack, S. Eleven down and nine to go. *Nature Structural & Molecular Biology*, 2(10):824–831, 1995.

[76] Kavran, J. M., Gundllapalli, S., O'Donoghue, P., Englert, M., Söll, D., and Steitz, T. A. Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation. *Proceedings of the National Academy of Sciences*, 104(27):11268–11273, 2007.

[77] Smith, T. F. and Hartman, H. The evolution of class II aminoacyl-tRNA synthetases and the first code. *FEBS Letters*, 589(23):3499–3507, 2015.

[78] Valencia-Sánchez, M. I., Rodríguez-Hernández, A., Ferreira, R., Santamaría-Suárez, H. A., Arciniega, M., Dock-Bregeon, A.-C., Moras, D., Beinsteiner, B., Mertens, H., and Svergun, D. Structural insights into the polyphyletic origins of glycyl tRNA synthetases. *Journal of Biological Chemistry*, 291(28):14430–14446, 2016.

[79] Martinis, S. A., Plateau, P., Cavarelli, J., and Florentz, C. Aminoacyl-tRNA synthetases: a new image for a classical family. *Biochimie*, 81 (7):683–700, 1999.

[80] Duchêne, A.-M., Pujol, C., and Maréchal-Drouard, L. Import of tRNAs and aminoacyl-tRNA synthetases into mitochondria. *Current Genetics*, 55(1):1–18, 2009.

[81] Small, I., Akashi, K., Chapron, A., Dietrich, A., Duchene, A. M., Lancelin, D., Maréchal-Drouard, L., Menand, B., Mireau, H., and Moudden, Y. The strange evolutionary history of plant mitochondrial tRNAs and their aminoacyl-tRNA synthetases. *Journal of Heredity*, 90 (3):333–337, 1999.

[82] de Pouplana, L. R. and Schimmel, P. A view into the origin of life: aminoacyl-tRNA synthetases. *Cellular and Molecular Life Sciences*, 57 (6):865–870, 2000.

[83] Schneider, A. Does the evolutionary history of aminoacyl-tRNA synthetases explain the loss of mitochondrial tRNA genes? *TRENDS in Genetics*, 17(10):557–559, 2001.

[84] Sissler, M., Putz, J., Fasiolo, F., and Florentz, C. Mitochondrial aminoacyl-tRNA synthetases. *Madame Curie Bioscience Database. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6033/*, 2013.

[85] Woese, C. R., Kandler, O., and Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12): 4576–4579, 1990.

[86] Lake, J. A., Henderson, E., Oakes, M., and Clark, M. W. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences*, 81(12): 3786–3790, 1984.

[87] Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, T. M. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, 105(51):20356–20361, 2008.

[88] Archibald, J. M. The eocyte hypothesis and the origin of eukaryotic cells. *Proceedings of the National Academy of Sciences*, 105(51):20049–20050, 2008.

[89] Foster, P. G., Cox, C. J., and Embley, T. M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1527):2197–2207, 2009.

[90] Williams, T. A., Foster, P. G., Cox, C. J., and Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479):231–236, 2013.

[91] Poole, A. M. and Penny, D. Evaluating hypotheses for the origin of eukaryotes. *Bioessays*, 29(1):74–84, 2007.

[92] Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nature Genetics*, 28(3):281–285, 2001.

[93] Poole, A. M. and Neumann, N. Reconciling an archaeal origin of eukaryotes with engulfment: a biologically plausible update of the Eocyte hypothesis. *Research in Microbiology*, 162(1):71–76, 2011.

[94] Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., and Brochier-Armanet, C. The origin of eukaryotes and their relationship with the

archaea: are we at a phylogenomic impasse? *Nature Reviews Microbiology*, 8(10):743–752, 2010.

[95] Sato, Y., Maeda, Y., Shimizu, S., Hossain, M. T., Ubukata, S., Suzuki, K., Sekiguchi, T., and Takenaka, A. Structure of the nondiscriminating aspartyl-tRNA synthetase from the crenarchaeon *Sulfolobus tokodaii* strain 7 reveals the recognition mechanism for two different tRNA anticodons. *Acta Crystallographica Section D: Biological Crystallography*, 63(10):1042–1047, 2007.

[96] Cardoso, A. M., Polycarpo, C., Martins, O. B., and Soll, D. A nondiscriminating aspartyl-trna synthetase from *Halobacterium salinarum*. *RNA Biology*, 3(3):110–114, 2006.

Chapter 4

The histone fold

Overview

This chapter uses the structure-based method to determine relationships between a set of highly diverged proteins, that present either the histone fold or its proposed ancestral motif. The protein structures used in this analysis span all three domains of life. The relationships, inferred from the dissimilarity between structures, are compared to their classifications by SCOP and PFam. This analysis concludes with the determination of an evolutionary model for the histone fold, something which could not be achieved previously using conventional sequence-based phylogenetic methods.

4.1 Introduction

4.1.1 Histone fold and the core histone proteins

The histone fold is a protein structural motif, Figure 4.1, comprising a triple-helical (HHH) topology. The terminal helices have approximately 15 residues each, whereas the central helix comprises approximately 30 residues. As the name suggests, this fold is present in the histone proteins.

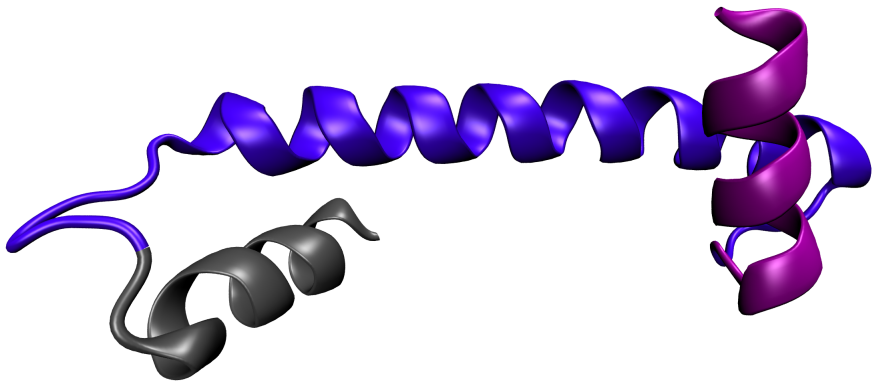


Figure 4.1: The histone fold (H3 from PDB 2nzd). The fold comprises three helices shown here in purple (N terminal), blue and grey (C terminal).

Histones are a family of proteins that are responsible for the formation of octameric nucleosomal assemblies, a unit of higher DNA structural organization called chromatin. The octameric nucleosomal assembly comprises two

copies each of histones H2A, H2B, H3 and H4. Once bound to the DNA, the octamer compacts the DNA by wrapping ∼145 base pairs of DNA around it, Figure 4.2.

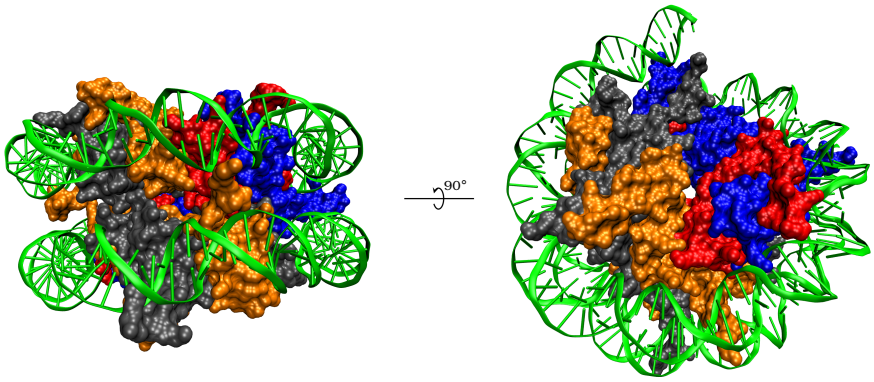


Figure 4.2: Nucleosome structure: (PDB 2cv5). Side and top view of nucleosome octamer comprising two H2A (gray), two H2B (orange), two H3 (blue) and two H4 (red) proteins. 145bp of DNA (green) are shown wrapped around the nucleosome.

4.1.2 Nucleosome formation and properties of the histone fold

The octamer formation starts with a tetramer of H3 and H4 which is deposited on the newly synthesized DNA strand by chromatin assembly factor, CAF-I. This is followed by two H2A-H2B dimers to complete the octamer [1, 2]. Chromatin organization in eukaryotes allows for relatively open (euchromatin) and closed (heterochromatin) regions of the DNA which subsequently allows for or suppresses gene expression [3]. The net positive charge, Table 4.1, on nucleosomal core histones allows for binding the negatively charged DNA and hence forming the higher order chromatin structure. The histone fold, thus, has three characteristic properties i.e. ability to dimerize (contact formation between H3-H4, H2A-H2B etc), DNA binding and compaction (bending DNA).

4.1.3 Prevalence of the histone fold

Some families of proteins are continuously interacting with the DNA, carrying out a variety of functions e.g. replication of the DNA, transcription,

Table 4.1: Nucleosome core histone proteins: The histone types were determined by annotations provided by RCSB (www.rcsb.org). Charges were calculated by counting the unbalanced basic residues. Length refers to the number of amino acids in the particular protein structure.

Species	**PDB**	**Histone**	**Net +ve charge**	**Length**
Xenopus laevis	1aoi	H3	10	98
		H4	12	83
		H2A	15	115
		H2B	15	99
Gallus gallus	1eqz	H3	12	108
		H4	16	90
		H2A	18	125
		H2B	17	108
Saccharomyces cerevisiae	1id3	H3	12	97
		H4	10	79
		H2A	13	110
		H2B	8	93
Homo sapiens	2cv5	H3	10	97
		H4	11	78
		H2A	13	108
		H2B	12	96
Drosophila melanogaster	2nqb	H3	10	98
		H4	11	81
		H2A	11	106
		H2B	11	95

repairing incorrectly replicated DNA, toggling gene expression, controlling the rate at which genes are expressed, maintaining the structure in which DNA is packaged, to name a few [4]. The histone fold, originally identified as a common motif in all eukaryotic nucleosomal histones (H2A, H2B, H3 and H4), has now been shown to be a part of other eukaryotic and archaeal DNA binding proteins [5], e.g. centromere specific proteins (CENP) [6], TATA binding protein associated factors (TAFs) [7] and archaeal DNA binding proteins [8].

This presence of a histone fold in these proteins is not surprising as the fold allows for DNA binding, the first step for any protein before it can carry out DNA related activity. Non-histone proteins, however, that present the histone fold do not necessarily replicate the entire functionality of the histone proteins. Only some of the proteins comprising the histone fold go beyond binding and compact the DNA. For example, TAFs are a class of proteins that bind one another and form a complex with TATA binding

protein, playing a crucial role by binding to the DNA, in the intermediate process leading to protein synthesis [9]. CENP, on the other hand, have been demonstrated to assemble as a heterotetramer, bind DNA and induce compaction [10] in a nucleosome core histone-like manner. Mutational studies have revealed their significant role in chromosome segregation during mitosis. Archaea contain proteins that behave in a manner similar to nucleosomal core histones [11]. This is indicated by their ability to dimerize and form nucleosome-like structures [8].

4.1.4 Histone-like proteins and the phylogenetic history of the histone fold

The sharing of the histone fold between archaeal and eukaryotic domains is suggestive of a deep evolutionary origin. The histone fold motif is thought to be the result of gene duplication and fusion of the simpler helix-loop-helix motif, [12, 13] which doubled the length of the central helix, i.e. helix-loop-**helix-helix**-loop-helix. It can therefore be conjectured that remnants of the helix-loop-helix motif in bacteria capable of binding DNA are likely evolutionary neighbours of the histone fold in eukaryotes and archaea.

Proteins with DNA binding ability are observed in all domains of life. Proteins, apart from the core histones, which exhibit some of the properties characterized by the histone fold are termed histone-like [14]. Previously, protein families from archaea and eukaryotes were listed. From domain bacteria, the HU (heat unstable) protein family presents histone-like properties [15, 16]. The HU family comprises of a multifunctional group of proteins with the ability to bind DNA in a capacity similar to nucleosomal histones. In a rare case, eukaryotes lacking histones were able to compact their DNA using proteins similar to bacterial HU proteins [17]. The bacterial histone-like HU family, however, is argued to share only superficial similarity [14] with the histone fold i.e. only in terms of DNA binding. This is not entirely true, as a closer examination reveals that the superficial similarity in HU family comes from the presence of a simpler helix-loop-helix motif.

The presentation of a simpler structural motif and the DNA binding ability qualifies the HU family as a likely evolutionary neighbour of the histone fold presenting domains. Phylogenetic analysis of a protein family presenting a common structural motif [18] has, in the past, allowed for evolutionary history to be traced back to the last universal common ancestor (LUCA).

The histone fold and its ancestral motif present a similar opportunity [8, 19] to recreate early evolutionary history of this structural motif, a feat that has not been accomplished so far.

A reason for the missing common phylogeny is the significant sequence divergence. The core histones exhibit this problem clearly. Of the four core histone families (i.e. H2A, H2B, H3 and H4), each family shows very strong conservation across its evolutionary history e.g. pea and calf thymus histone H4 are 98% identical [20]. A similar trend is observed within all histone families, however the results are quite different when compared across [21] histone families e.g. sequences of histone H3 and H4 from *Xenopus laevis* (PDB 2nzd) can only be aligned across a region of 30 residues of which just 14 can be considered conserved. This amounts to a decrease in conservation from 98% within histone families (across species) to just 13% across histone families (within [the same] species). Due to this significant divergence between histone families, sequence-based analysis has been carried out [22] only at the family level i.e. separately for each of the core histones H2A, H2B, H3 and H4.

In contrast to poor sequence similarity across histone families, structural analysis shows strong conservation of the histone fold [23]. This conservation is not limited to the core histones as shown in Figures 4.3 and 4.4, but is also seen in non-histone proteins. Figures 4.3 and 4.4 show conserved topologies across histone families and a superposition of the histones with one another further illustrates the strong conservation of the histone fold.

Previous sequence-based studies have shown that H2A and H2B from the same species evolve at identical rates [22]. H3 and H4, on the other hand show a high degree of conservation [22] with little or no change across species. These results are perhaps expected. The nature of interaction between H2A and H2B implies that changes in one may perhaps be compensated with changes in the other leading to identical rates of evolution. In the case of H3-H4, the requirement for dimer formation and binding the DNA introduces multidimensional pressure on their divergence, resulting in little or no change. These results offer meaningful insights however fail to address the question of their origin and early divergence, leaving a gap in our understanding.

The structural method of inferring phylogenetic relationships using protein structures presented in this thesis can be used to fill this gap. This work

therefore aims to infer phylogenetic relationships between protein structures presenting the conserved histone fold or its simpler ancestral motif, across the three domains of life.

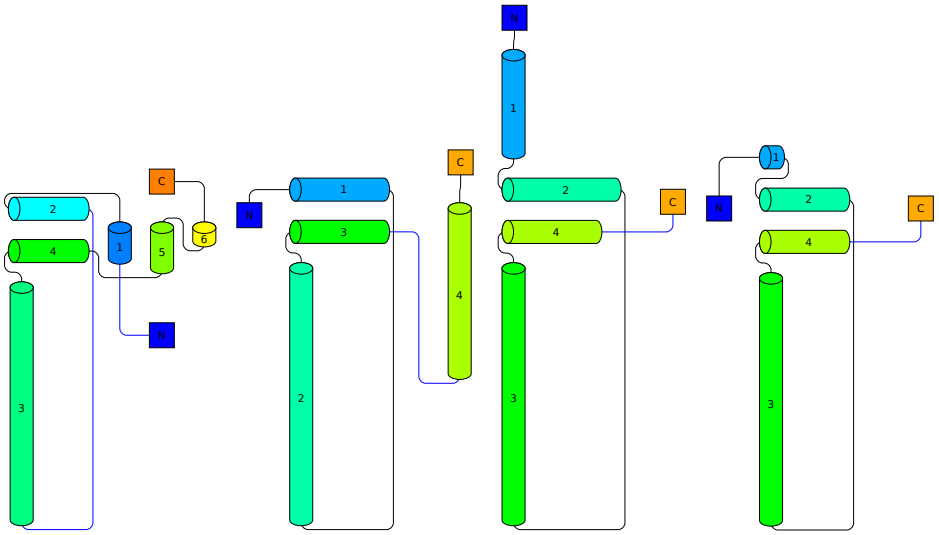


Figure 4.3: Histone H2A, H2B, H3 and H4 (left to right) topologies: Generated from core nucleosome structure (2nzd). Cartoons were generated using pro-origami [24]. Cylinders indicate helical structures. Apart from the N and C terminals the colours are for visual guidance only whereas numbers indicate separate helices.

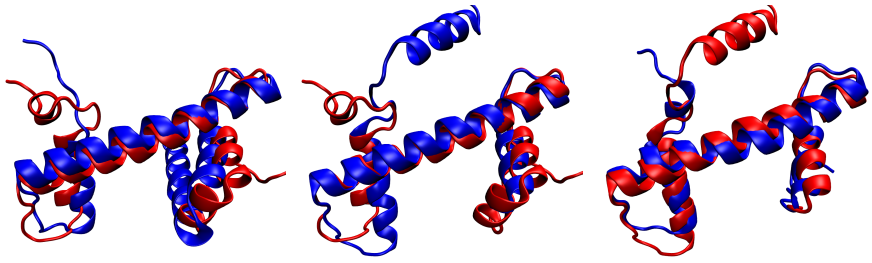


Figure 4.4: Histone structural superimposition: H2A(red)-H2B(blue) [Left], H2A(red)-H3(blue) [middle], H3(red)-H4(blue) [right]. All structures were extracted from the core nucleosome (2nzd). Superposition was created using Q-score and structural visualization done through VMD [25].

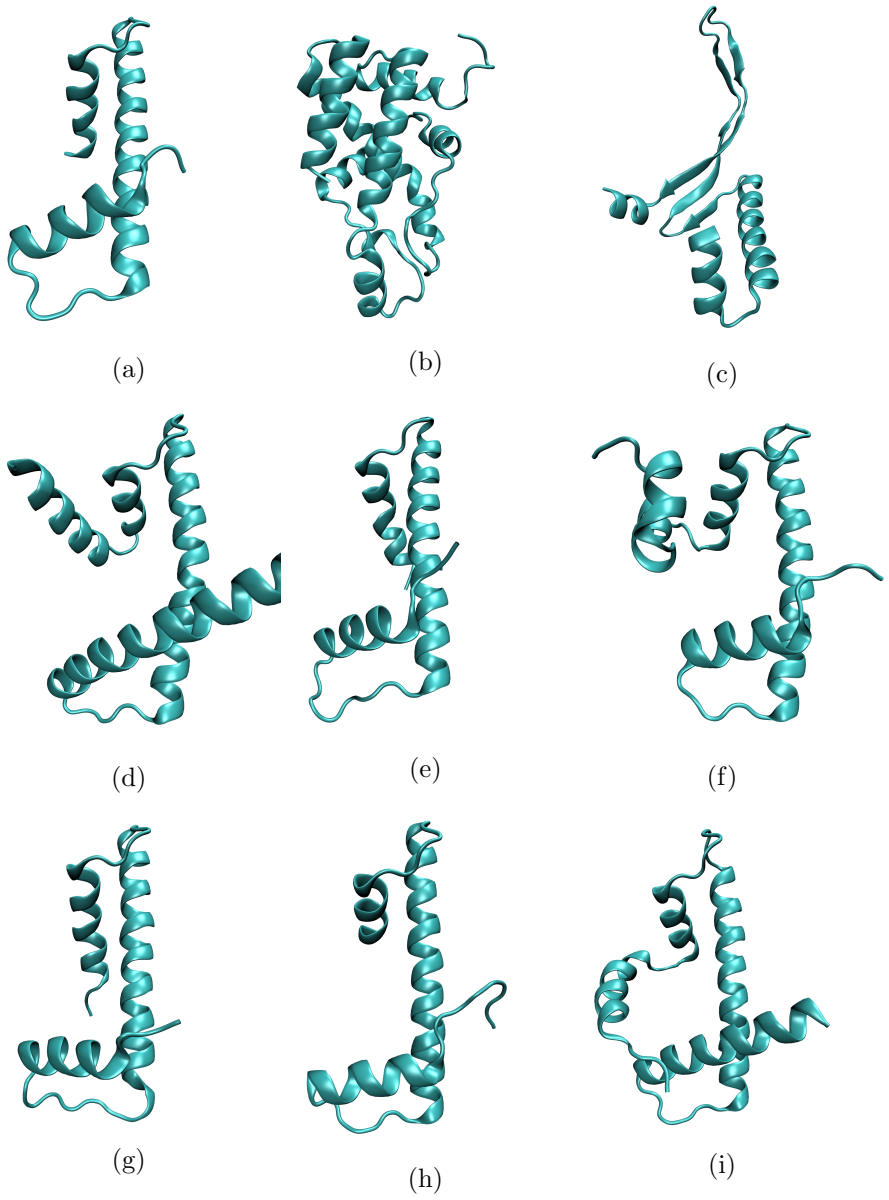


Figure 4.5: Representative proteins which either present the histone fold or the simpler ancestral motif and show histone-like DNA binding. (a) Archaeal DNA binding protein (1b67), (b) Histone-like bacterial DNA binding protein (1r4v), (c) bacterial DNA binding protein (3rhi), (d-g) CENP-S, CENP-X, CENP-T_C and CENP-W (3vh5, chains A, D, T and W), (h-i) TAFs (1taf and 4wv4).

4.2 Method

PFam [26] was used to obtain protein data bank (PDB) identifiers for core histones. The core histones were found to be a part of the histone (CL0012) clan [27] in PFam and therefore the core histone structural dataset was extended by including other protein families in the histone clan, as they shared an evolutionary history [27] with the core histones. This extension included archaeal histones, CENPs and TAFs amongst others, see Tables 4.2 and 4.3 for more details.

PFam groups sequences based on sequence similarity. Due to insufficient similarity, bacterial histone-like proteins were not grouped with eukaryotic and archaeal histones but were present in a separate clan IHF-like DNA binding (CL0548), which included bacterial DNA binding proteins and HU DNA binding proteins and Tra-M. For consistency with the usage of the histone clan, all families in the IHF-like DNA binding clan were also considered.

It is known [14, 28] that bacterial nucleoid-structuring (H-NS) proteins also show histone-like properties and work in a manner similar to HU DNA binding proteins. The H-NS proteins can be found in the Histone_HNS (PF00816) family, however, this family is not part of a PFam clan and hence Histone_HNS (PF00816) was included as a single family in this analysis.

The structural database, SCOP, was also used to recover PDB identifiers for the aforementioned families. Due to the different methods of organizing structural information, protein structure assignment was inconsistent between the two databases. However, all identifiers from SCOP and PFam were retained for the aforementioned protein families.

Following the recovery of the PDB identifiers, protein structures corresponding to them were gathered from RCSB. Tables 4.2 and 4.3 provide more details regarding the identifiers, their source and family classification according to SCOP and PFam.

The protein structure data was then processed in the following way:

- Pairwise comparisons were done for each protein structure using Superpose [65]. Due to the nature of the algorithm comparisons are order specific i.e. A $\cong$ B $\neq$ B $\cong$ A. Both pairwise comparisons were, therefore, performed and the scores were averaged to attain a final score

Table 4.2: PDB, SCOP and Pfam identifiers of protein structures used in this analysis. The blank entries indicate PDBs not classified by the database and are later referred to as "NC". Entries marked with "*" are shown in Figure 4.5. Colours are for visual assistance only. [29–64].

PDB	Chain	SCOP	Pfam	PDB	Chain	SCOP	Pfam
1b8z	A	a.55.1.1	PF00216	3vh5*	D		PF09415
1hue	A	a.55.1.1	PF00216	3vh6	D		PF09415
1ihf	A	a.55.1.1	PF00216	4dra	E		PF09415
1ihf	B	a.55.1.1	PF00216	1r4v*	A	a.22.1.4	PF09123
1p51	A	a.55.1.1	PF00216	1wwi	A	a.22.1.4	PF09123
2iie	A		PF00216	1aoi	A	a.22.1.1	PF00125
2np2	A	a.55.1.0	PF00216	1aoi	C	a.22.1.1	PF00125
3rhi*	A	a.55.1.1	PF00216	1aoi	D	a.22.1.1	PF00125
4dky	A	a.55.1.0	PF00216	1eqz	A	a.22.1.1	PF00125
4n1v	A	a.55.1.0	PF00216	1eqz	B	a.22.1.1	PF00125
4qju	A		PF00216	1eqz	C	a.22.1.1	PF00125
4wv4*	B		PF07524	1id3	A	a.22.1.1	PF00125
1b67*	A	a.22.1.2	PF00808	1id3	C	a.22.1.1	PF00125
1b6w	A	a.22.1.2	PF00808	1id3	D	a.22.1.1	PF00125
1f1e	A	a.22.1.2	PF00808	2cv5	A	a.22.1.1	PF00125
1ku5	A	a.22.1.2	PF00808	2cv5	C	a.22.1.1	PF00125
1n1j	A	a.22.1.3	PF00808	2cv5	D	a.22.1.1	PF00125
1n1j	B	a.22.1.3	PF00808	2nqb	A	a.22.1.1	PF00125
4g91	B	a.22.1.3	PF00808	2nqb	C	a.22.1.1	PF00125
4g91	C	a.22.1.3	PF00808	2nqb	D	a.22.1.1	PF00125
3v9r	A		PF15630	1hns	A	a.155.1.1	PF00816
3vh5*	A		PF15630	2jr1	A		PF00816
3vh6	A		PF15630	2l92	A		PF00816
4dra	A		PF15630	2l93	A	a.155.1.1	PF00816
1aoi	B	a.22.1.1	PF15511	4fmr	A		PF14848
1eqz	D	a.22.1.1	PF15511	1taf*	B	a.22.1.3	PF02969
1id3	B	a.22.1.1	PF15511	1h3o	A	a.22.1.3	PF05236
2cv5	B	a.22.1.1	PF15511	1bh8	B	a.22.1.3	PF04719
2nqb	B	a.22.1.1	PF15511	1h3o	B	a.22.1.3	PF03847
3vh5*	T		PF15511	1bh8	A	a.22.1.3	PF02269
3vh6	T		PF15511	1taf	A	a.22.1.3	PF02291
3vh5*	W		PF15510	1dp3	A	a.55.1.2	PF05261
3vh6	W		PF15510	3d8a	A	a.241.1.1	PF05261
3v9r	B		PF09415				

Table 4.3: Proteins used in this analysis were spread across four SCOP superfamilies comprising nine protein families and two PFam clans comprising 18 families. Histone_HNS was not assigned to a clan.

SCOP SF	**Family**	**Annotation**
a.22.1	a.22.1.1	Nucleosome histones
	a.22.1.2	Archaeal histone
	a.22.1.3	TBP-associated factors
	a.22.1.4	Bacterial histone-fold
a.55.1	a.55.1.0	IHF-like (automated)
	a.55.1.1	Prokaryotic DNA bending
	a.55.1.2	DNA bending
a.155.1	a.155.1.1	H-NS histone-like
a.241.1	a.241.1.1	Tra_M-like
Pfam Clan	**Family**	
CL0012	PF07524	Bromo_TP
	PF00808	CBFD_NFYB_HMF
	PF15630	CENP-S
	PF15511	CENP-T_C
	PF15510	CENP-W
	PF09415	CENP-X
	PF09123	DUF1931
	PF00125	Histone
	PF02969	TAF
	PF05236	TAF4
	PF04719	TAFII28
	PF02269	TFIID-18kDa
	PF02291	TFIID-31kDa
	PF03847	TFIID_20kDa
CL0548	PF00216	Bac_DNA_binding
	PF14848	HU-DNA_bdg
	PF05261	Tra_M
-na-	PF00816	Histone_HNS

"q" of the comparison between structure A and B.

- The similarity score was subtracted from one (d = 1 - q) to generate distance (d).
- A matrix was populated with the pairwise distances.
- The neighbour-joining (NJ) algorithm [66] as implemented by the Phylo [67] package in Biopython cock2009biopython was used to gen-

erate a NJ-tree from the matrix. The tree was then visualized using Figtree [68].

- Splitstree [69] was used to visualize the matrix as a neighbour-net network.

The Kelly palette [70] was used for selecting contrasting colours.

4.3 Results

4.3.1 Long branch attraction

Phylogenetic analysis sometimes plays host to long branch attraction (LBA), a form of systematic error. LBA occurs when fast diverging protein sequences (structures in this case) incorrectly attract other branches impacting the final results. A simple solution to avert this error is long branch extraction [71], where sequences occupying long branches (or where LBA is suspected) are removed to see if the topology of the resulting tree changes. Using long branch extraction, i.e. removing structures occupying long branches, the resulting tree topology was compared to that of the original tree, to ensure no such errors existed in this analysis. A transcription factor (1N1J [44], chain A and B) was removed from this analysis as a result of long branch extraction.

4.3.2 Presence of an evolutionary signal

The goal of this analysis was to detect an evolutionary signal in a group of proteins that either contained the histone fold or a simpler helix-loop-helix motif, which is considered to be the ancestor of the histone fold and has DNA binding properties. The structural dataset collated spanned the three domains of life and allowed for the determination of the evolutionary history of histone fold. To illustrate this history, a phylogenetic tree was constructed.

The un-rooted neighbour joining (NJ) phylogenetic tree representing the relationships between structures in this dataset is shown in Figure 4.6. The same distance data is used to construct a neighbour-net reticulation network in Figure 4.7. The tree likeness of the network informs the quality of the analysis, i.e. inferences articulated from a tree based on distance data whose

network is tree-like will be robust. Figure 4.7 represents a tree-like network which when converted into a tree, Figure 4.6, clearly shows a split between bacteria (green) and non-bacteria (archaea: red and eukarya: blue).

Given that the SSM-based Q_{score} metric considers the relative sizes of the proteins compared, see Chapter 2, to ensure the split between bacterial and eukaryotic proteins was not due to size differences, the distribution of protein sizes was calculated and shown in Figure 4.8. Due to the position of the archaeal proteins on the tree, only bacterial and eukaryotic protein distributions were compared. The overlap between the eukaryotic and bacterial protein size distributions reveal that any segregation present in the tree is enforced primarily by shape differences between proteins and not purely by size. The separation of the domains of life lends confidence to the ability of the method to reveal deep evolutionary signals. This is something that could not be achieved, for the proteins considered in this analysis, if relying solely on protein sequence comparisons due to significant divergence.

4.3.3 SCOP and Pfam organisation

SCOP and Pfam organize proteins according to their structures and sequences, respectively. SCOP achieves a structure-based hierarchy by grouping similar structures into classes, folds, superfamilies and families, with structural similarity between members increasing from the class to family level. Pfam groups proteins based on sequence similarity into protein families and further groups these families into clans. SCOP superfamilies group structures based on a inferred common evolutionary history [72] whereas in Pfam the same is said about the grouping of evolutionarily related protein families into clans [73].

Figures 4.9 and 4.10 show leaves of the tree coloured according to SCOP and Pfam classifications. A superficial view shows a broad recovery of some of the classifications for both SCOP and Pfam. A clustering of each of the nucleosomal core histones, H2A, H2B, H3 and H4 (a.22.1.1) is observed in Figure 4.9 (shown in brown) and explicitly labelled in 4.10. Archaeal histones appear as a cluster with the exception of the histone from *Methanopyrus kandleri* (PDB 1f1e). This protein structure was somewhat symmetrical about the central residue, indicating a gene duplication and fusion, and thus was split into two structures (1f1e_1 and 1f1e_2). The TAF protein family appears to be spread across the non-bacterial split. On the bacterial split

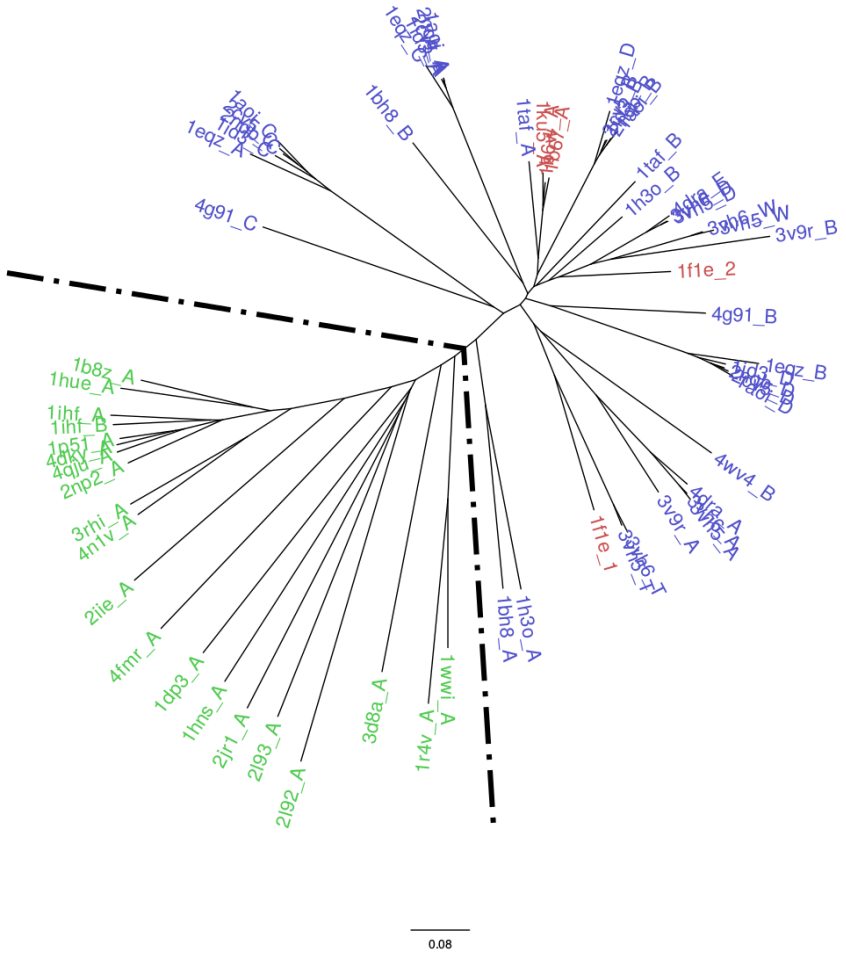


Figure 4.6: Histone fold phylogeny: An unrooted NJ-tree of 66 protein structures spanning three domains of life. The labels indicate PDB and chain identifiers; green are from bacterial histone-like proteins, red from archaeal origin and blue represent structures from eukaryotes. The black dotted line indicates a split between bacterial and non-bacterial proteins. The scale bar represents distance.

(Figure 4.6), a distinct grouping of bacterial histone-fold proteins (a.22.1.4) is observed. Prokaryotic DNA bending (a.55.1.1), IHF-like (a.55.1.0) families and DNA-binding (a.55.1.2) all belong to the IHF-like DNA-binding protein (a.55.1) superfamily in SCOP and shows some clustering with the exception of the singleton from the DNA-binding protein (a.55.1.2) family which is clustered with the H-NS histone-like (a.155.1.1) protein family and

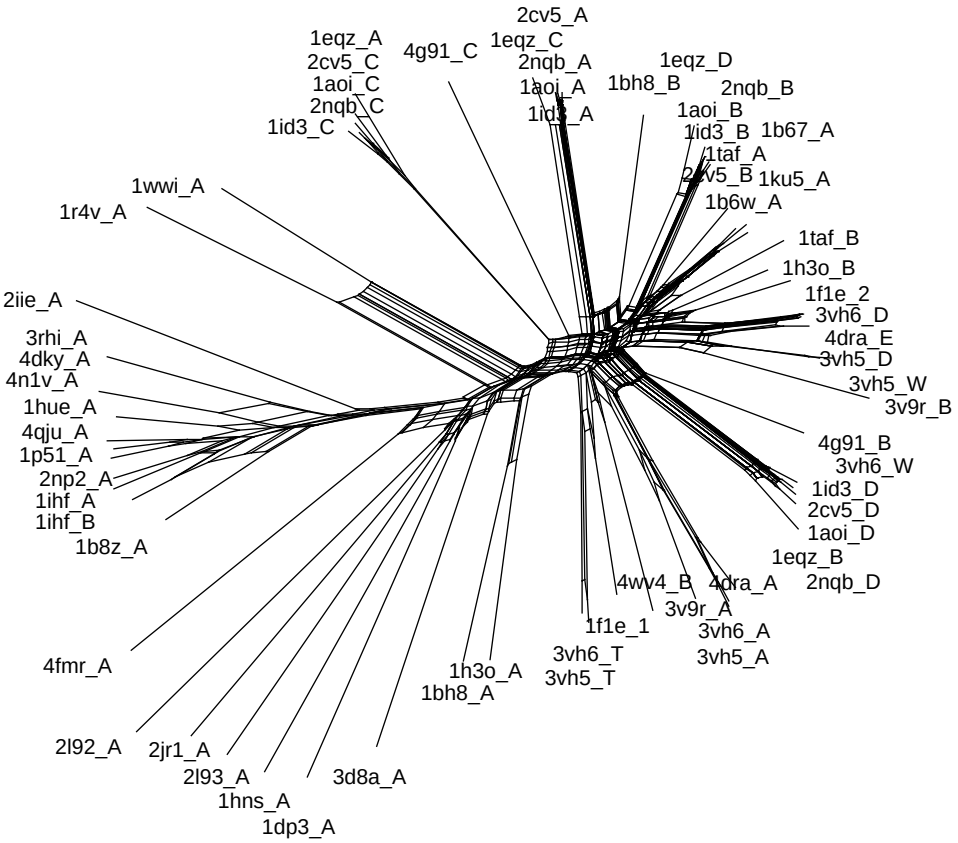


Figure 4.7: Histone fold phylogenetic network. A neighbor-net splits network of the dataset. A tree-like representation, as seen in this figure, indicates fewer alternative relationships. Similar to Figure 4.6 the tree labels include PDB and chain identifiers. The scale bar represents distance.

some unclassified proteins ("NC").

Some of the protein structures lacking a classification (labelled as "NC" in Figure 4.9) can be seen having a classification by Pfam in Figure 4.10. Core histones appear to cluster at the histone family level in both Figures 4.9 and 4.10. It should be noted that the cluster comprising of histone H4 is classified as the centromere associated protein (CENP-T_C) by Pfam. Archaeal histone clustering is observed with the exception of his-

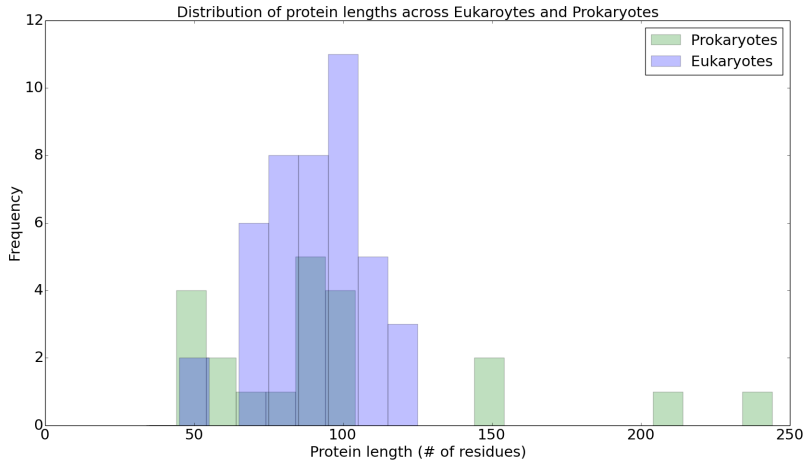


Figure 4.8: Size distribution of protein structures used in this analysis. Eukaryotic (blue) and prokaryotic (green) proteins overlap in size.

tone from *Methanopyrus kandleri* mentioned earlier. The centromere associated proteins (excluding the misclassified H4 cluster) are seen to break up into two clusters, one comprising CENP-W/X and the other CENP-T_C/S. Centromere-associated protein structures had not been classified by SCOP v2.06 and carried the "NC" label in Figure 4.9. In Figure 4.9, the TAFs are spread out on the non-bacterial split. As opposed to SCOP, Pfam shows better clustering on the bacterial clades with bacterial DNA binding proteins (BAC_DNA_binding) and H-NS family (Histone_HNS) forming clusters with some singletons interspersed. Proteins classified as bacterial histone fold proteins and grouped with the histone superfamily in SCOP are classified domains of unknown function (DUF1931) in Figure 4.10.

As explained earlier, the classifications of proteins in SCOP and PFam are determined by structure and sequence similarity. A macro observation of the results in Figures 4.9 and 4.10 recover this. In Figure 4.9, the non-bacterial clade comprises a single SCOP superfamily (a.22.1) whereas the bacterial clade has three (a.55.1, a.241.1 and a.155.1) with the exception of the bacterial histone fold which are a part of the histone superfamily (a.22.1). Apart from the superfamily TraM-like, which is a singleton, the other two clusters, i.e. a.155.1 and a.55.1 can be separated. Similarly in Figure 4.10, the non-bacterial clade comprises protein families that are classified under a single clan (CL0012, see Table 4.3). The bacterial clade comprises another

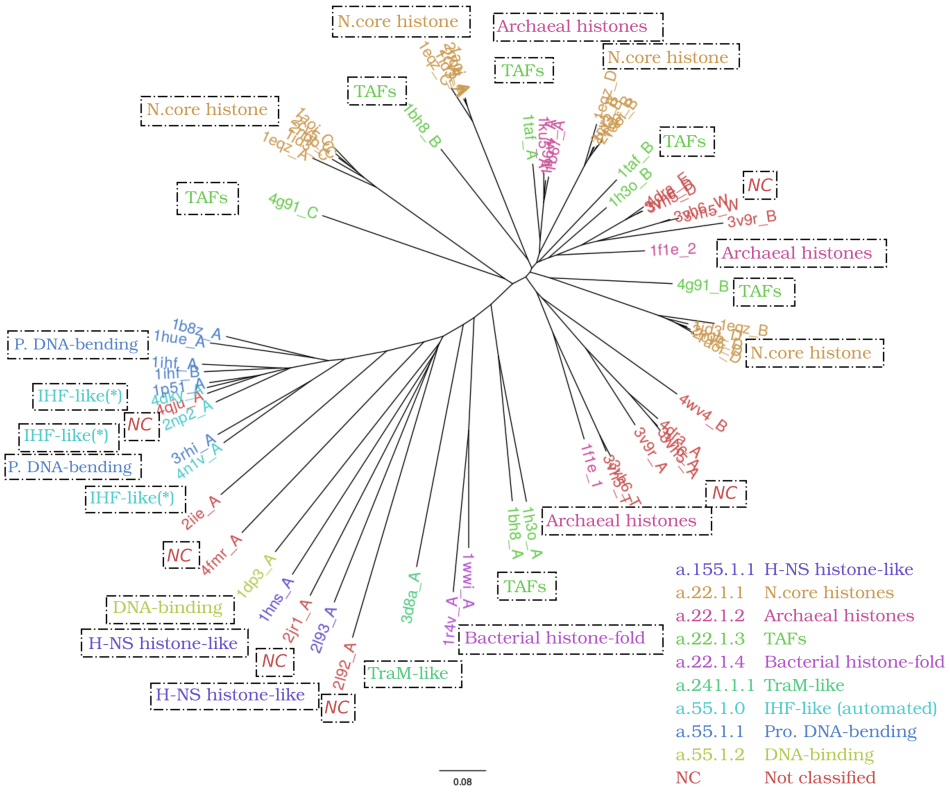


Figure 4.9: Histone fold phylogeny: Tree labels are colour coded according to classification by SCOP. NC is used where structures were not classified by SCOP. The dotted boxes, alongside PDB and chain labels, include SCOP family labels. See Table 4.2 for more details. Prokaryotic DNA-bending and Nucleosomal core histone are shortened to P. DNA-bending and N.core histones respectively.

clan (CL0548), members of which can be separated from the Histone_HNS protein family which is not classified under a clan(see Figure 4.10).

Given the nature of hierarchical databases one would expect proteins collated at the family level to group together, as both PFam and SCOP group evolutionarily-related proteins at this level of the hierarchy. The PFam groupings, which are based on sequence comparisons, may not follow this strictly as the structural and the sequence groupings may be unrelated, however this expectation should hold for SCOP. This does not appear to be the case as can be seen with nucleosomal core histones, Figure 4.9, which all belong to the same family (a.22.1.1) yet are segmented into four clusters instead of appearing as one. Each cluster belongs to a distinct histone family

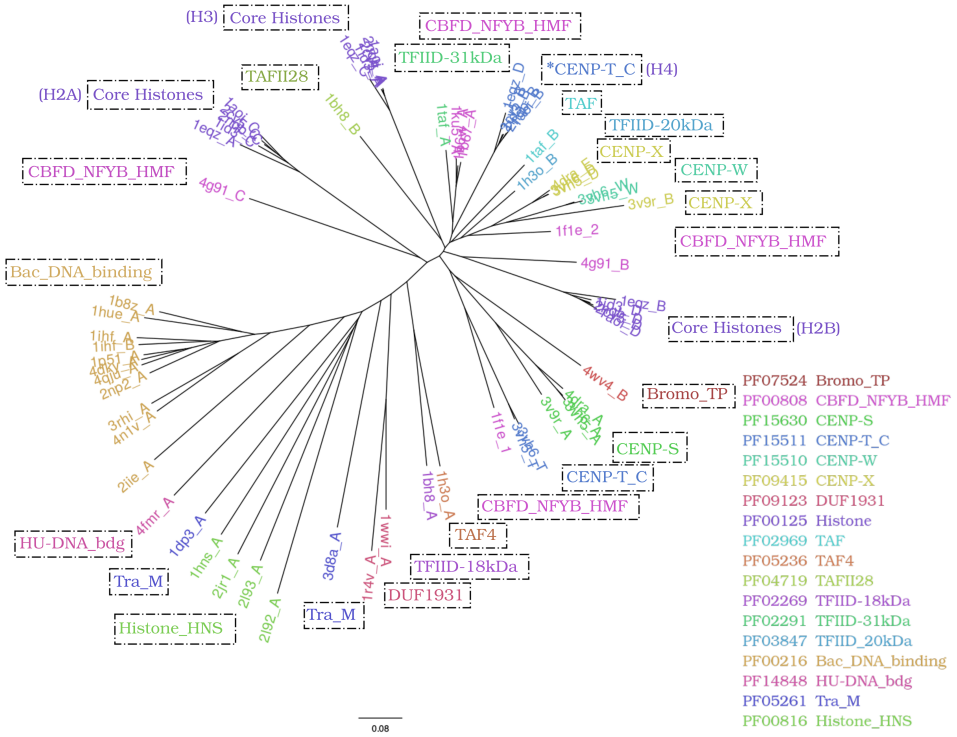


Figure 4.10: Histone fold phylogeny: Labels are colour coded according to classification by Pfam. *CENP-T_C cluster contains H4 but is classified as CENP-T_C by Pfam. See Table 4.2 for more details. Core histones label is shortened to Histone in the figure legend.

(H2A, H2B, H3 and H4), with the groupings interspersed by other protein families, like TAFs. It appears from this structural analysis that protein family classification in SCOP is not rigidly confined by strict quantitative structural similarity thresholds.

4.3.4 TATA binding protein associated factors and the histone fold

Transcription factor II D (TFIID) is a complex that recognises promoter regions on the DNA [74] and initiates the recruitment of RNA polymerase II [75]. TATA binding protein and associated factors make key ingredients in the TFIID pre-initiation complex. TATA binding protein associated factors or TAFs show a high degree of conservation across eukaryotes [76]. The histone fold, a structural motif, has been suggested to be the most important structural part of TFIID, shown to be a part of 9 out of 14 TAFs [13].

Table 4.4: TAFs used in this work along with their PDB and chain identifiers.

PDB	Chain	TAF
1taf	A	TAF9
1taf	B	TAF6
1h3o	A	TAF4
1h3o	B	TAF12
1bh8	A	TAF13
1bh8	B	TAF11
4wv4	B	TAF8

Some of the 14 TAFs are believed to have lost DNA contacting residues [77] and therefore only retain protein-protein dimerization ability similar to core histones. On the other hand, contradicting evidence states that a histone octamer-like structure, contributed by TAFs, inside TFIID [78, 79] is believed to be responsible for the DNA binding ability [80] of the TFIID complex. Regardless of their exact function, either in dimerization or DNA binding, the presence of the histone fold justifies their inclusion in the structural dataset analysed in this work.

The structures used in this work contained seven proteins which function as TAFs and are listed in Table 4.4. Pfam classifies these TAFs as individual protein families with each TAF getting assigned a unique family identifier. In contrast to this SCOP identifies the structural similarity and groups all TAFs in one family, TBP associated factors (a.22.1.3). In this structural analysis (Figures 4.9 and 4.10) it appears that different TAF proteins share varying degrees of similarity with other families as seen from their interspersed patterns. To summarize this observation it can be said that members of the TAF protein family have less similarity amongst themselves then with members of other protein families in their superfamily (a.22.1).

Considering the behaviour of TAFs, it has previously [79, 81] been discussed that TAFs, namely TAF6-TAF9 and TAF12-TAF4 have the ability to form an octamer-like structure similar to that formed by core histones [77]. The presence of this octamer-like structure has been theorised to be responsible for the DNA binding capability of TFIID [78]. Given that experimental evidence exists for the octamer structure formation, we see TAF6 (1taf_A), TAF9 (1taf_B) and TAF12 (1h3o_B) to be close to H3/H4 and TAF4 (1h3o_B) to be close to H2A. This observation also agrees with results

recorded by [82]. A conserved histone fold generates structural proximity to core histones and explains their ability to form octamer-like structures, however key changes in protein-protein interactions and a missing third helix in TAF4 (1h3o_A) may impair DNA binding capacity as stated by Werten et al. [82].

As indicated earlier, like histones, TAF proteins have significantly diverged protein sequences. As a result a phylogeny of TAF proteins is missing. This work attempts to gauge evolutionary profiles based on the presence of a common structural motif, i.e. the histone fold. Pfam does not offer any insight into TAFs as each TAF protein is classified as a singleton. Structural classification by SCOP assigns TAFs to a family in a superfamily also comprising of core histones, archaeal histones and some bacterial histone-fold protein families. In this context TAFs appear to share an evolutionary history with these DNA binding proteins. The distribution of TAFs illustrates multiple points of origins instead of a single ancestor and the positions of TAFs in the phylogenetic tree can be explained by duplication and divergence events most likely from non-TAF proteins. This is simplified with an evolutionary model presented in the following section, Figure 4.11.

4.3.5 Centromere-forming histones

At the time this study was conducted (SCOP v2.06) some structures had not been classified by SCOP. Figure 4.9 illustrated these labelled as "NC". The same structures are classified into four unique centromere associated protein families, namely CENP-X, CENP-W, CENP-T and CENP-S. These proteins are a part of CCAN (Constitutive Centromere Associated Network) [83]. Earlier works have shown that a majority of the proteins constituting CCAN are conserved in eukaryotes [84].

Contrasting evidence is present in the literature regarding the formation of the CENP-T-W-S-X complex and the similarity of this tetramer to the nucleosomal octameric structure [10, 85]. In the phylogenetic tree, Figure 4.10, one can find CENP-W and CENP-X sharing a clade and CENP-T and CENP-S sharing a clade. An interesting observation here is that CENP T/S are close to histones H2A and H2B whereas CENP W/X are close to histones H3 and H4. Complexes between CENP-T and CENP-W and between CENP-S and CENP-X have been reported [10, 86, 87] with the S-X complex varying slightly compared to T-W complex, in terms of structural arrange-

ment, and the four proteins forming an heterotetramer. It is interesting to note here that CENP-T also has been shown to interact with histone H3 [88].

As discussed earlier, significant sequence differences place these proteins into four separate families. While a sequence-based phylogeny cannot be attempted, a structure-based one places these proteins in the structural neighbourhood of nucleosomal core histones, indicating CENP W/X to have a common ancestry with H3/H4 and CENP T/S to share history with H2A/H2B.

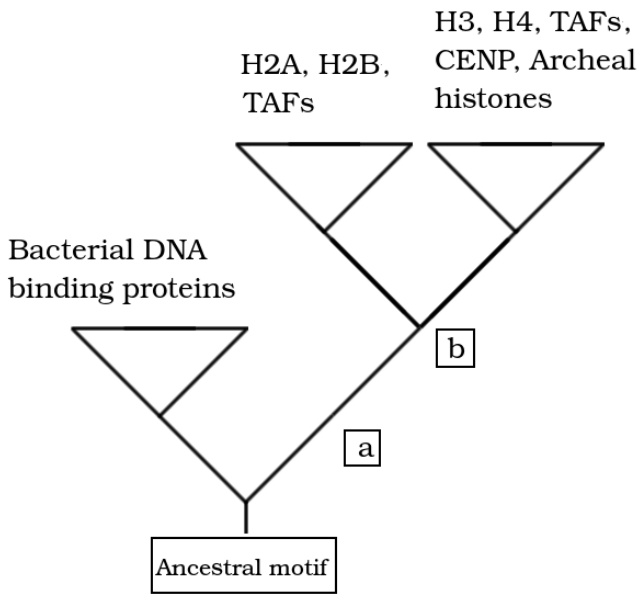


Figure 4.11: Evolutionary model of the histone fold. The simpler helix-loop-helix motif is shown as a possible root. The ancestral motif creates the bacterial and non-bacterial clades. The non-bacterial clade (a) has the fusion event to produce the histone fold. Dinoflagellates, *Crypthecodinium cohnii*, were probably present on (a) prior to the fusion event. (b) indicates the duplication and divergence event that led to the ancestral proteins of H2A/H2B and H3/H4. The tree cannot be further resolved with the current state of knowledge.

4.4 Discussion

Using a structure-based approach this work created a unified phylogeny of proteins, which either comprised the histone fold or its proposed ancestral motif that showed histone-like behaviour, i.e. DNA binding. Due to considerable divergence, the sequences of these proteins could not be analysed based on conventional phylogenetic methods. The structural method allowed for the determination of the relationships between protein structures. The relationships recovered from the tree were also compared to the classification of these proteins by PFam and SCOP.

Firstly the proteins in 18 PFam families, Table 4.3, which were grouped into two clans, showed only enough structural diversity to occupy nine structural families across four superfamilies, as classified by SCOP. Groupings at the SCOP superfamily and PFam clan levels indicate evolutionary relatedness, which was recovered in this analysis, Figure 4.9 and 4.10. The analysis further revealed that the groupings within clans and superfamilies were not robust when analysed quantitatively at a structural level. As discussed earlier multiple origins of the TAF proteins and the histone family subgroups support this.

Secondly, the histone fold appears to be the result of an ancient evolutionary event which is supported, Figure 4.6, by the bacterial non-bacterial split. The histone fold is theorised to be the product of a fusion event of the simpler helix-loop-helix motif [89] which could serve as a root. Furthermore, dinoflagellate *Crypthecodinium cohnii*, a eukaryote, has histone-like proteins similar to bacterial DNA-binding proteins. If this evidence is considered it can be said that the ancestral motif was the last common ancestor before the bacterial non-bacterial split. The non-bacterial split then experienced the fusion event followed by duplication and divergence, resulting in splits leading to the ancestor of H2A/H2B and H3/H4 clades. These branches further see similar events resulting in multiple origins of TAFs, the core histone families (H2A, H2B, H3 and H4) and centromere specific (CCAN) proteins.

The discussion above is illustrated in a proposed evolutionary model, Figure 4.11, based on empirical evidence extracted from this analysis. Due to the nature of the structural analysis and the method being in its infancy, the evolutionary model cannot be further resolved accurately. Due to the poor understanding of the evolution of protein structure, a time line cannot

be associated with events depicted on the model at present. However, the model does indicate the presence of the archaeal histones on only one of the core histone branches leading one to speculate that perhaps in the future more archaeal histone structures may emerge occupying the H2A/H2B clade.

4.5 Conclusion

Structural analysis can be used to bypass the conventional problem of significant divergence that stops sequence analysis from being used to probe deep evolutionary signals. This work attempted, for the first time, to recover evolutionary history of a set of proteins presenting either the histone fold or a simpler structural motif, argued to be the ancestor of the histone fold.

Comparison with known relationships, as classified by SCOP and PFam, concluded that while some relationships evident from significant structural and sequence similarity were easy to recover, namely clusters of core histones H2A, H2B, H3 and H4, recovering divergent relationships, e.g. a single cluster comprising all the core histone presented a challenge.

The difficulty of recovering PFam-based classifications was expected as it is based on sequence similarity. The same cannot be said for classifications from SCOP. The structrual database classifies structures based on structural similarity. It was surprising to find the classifications break down when using a simple quantitative measure to score distances as was seen with TAFs and core histones, from two different SCOP families, grouping together. This work, therefore, indicates that perhaps robust groupings can be achieved if a more stringent quantitative criterion is used based on structural analysis, rather then what is currently implemented by SCOP.

The structural analysis also allowed for the prediction of the existence of more archaeal histones using the evolutionary model that resulted from this structural analysis.

4.6 Future Work

There exist two separate avenues which require further attention. One is the use of the structural method to infer phylogenies, whereas the other is the extension of the work regarding histone fold phylogeny. These are separately discussed below.

4.6.1 Structure-based method for inferring phylogenies

Phylogenetic methods have been around for quite a while and have been extensively investigated to capture possible sources of error. This, however is not the case with this method as it utilizes structures instead of sequences which is one possible line of work; to identify sources of error. The metric used in this work accounts for size and shape differences between proteins compared, but it is well established that protein superposition is a non-trivial problem. How the SSM-based Q_{score} or other metrics handle this non-trivial problem is another avenue of investigation.

4.6.2 Histone fold phylogeny

Inclusion of more data has been shown to change relationships inferred, for the better, from a phylogenetic tree [90, 91]. Including more protein structures, as they become available, comprising either the histone fold or its ancestor may enhance the results and will be looked into further. Another avenue of work includes building the structure-based phylogenetic trees, as shown here, with some form of support akin to bootstrap [92]. This has been tackled in this thesis for another structural dataset, but due to time constraints was not used for the histone fold structural dataset, but will be implemented in the near future to add a measure of support for the inferred evolutionary relationships.

Bibliography

[1] Arents, G., Burlingame, R. W., Wang, B.-C., Love, W. E., and Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proceedings of the National Academy of Sciences*, 88(22):10148–10152, 1991.

[2] Smith, S. and Stillman, B. Stepwise assembly of chromatin during DNA replication in vitro. *The EMBO Journal*, 10(4):971–80, 1991.

[3] Allis, C. D. and Jenuwein, T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics*, 2016.

[4] Appling, D. R., Anthony-Cahill, S. J., and Mathews, C. K. *Biochemistry: Concepts and Connections*. Pearson, 2015.

[5] Baxevanis, A. D., Arents, G., Moudrianakis, E. N., and Landsman, D. A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Research*, 23(14):2685–2691, 1995.

[6] Sullivan, K. F., Hechenberger, M., and Masri, K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *The Journal of Cell Biology*, 127(3):581–592, 1994.

[7] Michel, B., Komarnitsky, P., and Buratowski, S. Histone-like TAFs are essential for transcription in vivo. *Molecular Cell*, 2(5):663–673, 1998.

[8] Pereira, S. L., Grayling, R. A., Lurz, R., and Reeve, J. N. Archaeal nucleosomes. *Proceedings of the National Academy of Sciences*, 94(23): 12633–12637, 1997.

[9] Wassarman, D. A. and Sauer, F. TAFII250. *Journal of Cell Science*, 114(16):2895–2902, 2001.

[10] Nishino, T., Takeuchi, K., Gascoigne, K. E., Suzuki, A., Hori, T., Oyama, T., Morikawa, K., Cheeseman, I. M., and Fukagawa, T. CENP-TWSX forms a unique centromeric chromatin structure with a histone-like fold. *Cell*, 148(3):487–501, 2012.

[11] Reeve, J. N., Bailey, K. A., Li, W. T., Marc, F., Sandman, K., and Soares, D. J. Archaeal histones: structures, stability and DNA binding. *Biochemical Society Transactions*, 32(2):227–230, 2004.

[12] Alva, V., Ammelburg, M., Söding, J., and Lupas, A. N. On the origin of the histone fold. *BMC Structural Biology*, 7(1):17, 2007.

[13] Gangloff, Y.-G., Romier, C., Thuault, S., Werten, S., and Davidson, I. The histone fold is a key structural motif of transcription factor TFIID. *Trends in Biochemical Sciences*, 26(4):250–257, 2001.

[14] Dorman, C. J. and Deighan, P. Regulation of gene expression by histone-like proteins in bacteria. *Current Opinion in Genetics & Development*, 13(2):179–184, 2003.

[15] Kamashev, D. and Rouviere-Yaniv, J. The histone-like protein HU binds specifically to DNA recombination and repair intermediates. *The EMBO Journal*, 19(23):6527–6535, 2000.

[16] Schmid, M. B. More than just "histone-like" proteins. *Cell*, 63(3): 451–453, 1990.

[17] Wong, J. T. Y., New, D. C., Wong, J. C. W., and Hung, V. K. L. Histone-like proteins of the dinoflagellate *Crypthecodinium cohnii* have homologies to bacterial DNA-binding proteins. *Eukaryotic Cell*, 2(3): 646–650, 2003.

[18] Brindefalk, B., Dessailly, B. H., Yeats, C., Orengo, C., Werner, F., and Poole, A. M. Evolutionary history of the TBP-domain superfamily. *Nucleic Acids Research*, 41(5):2832–2845, 2013.

[19] Arents, G. and Moudrianakis, E. N. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proceedings of the National Academy of Sciences*, 92(24):11170–11174, 1995.

[20] DeLange, R. J., Fambrough, D. M., Smith, E. L., and Bonner, J. Calf and pea histone IV III. Complete amino acid sequence of pea seedling Histone IV; comparison with the homologous calf thymus histone. *Journal of Biological Chemistry*, 244(20):5669–5679, 1969.

[21] Draizen, E. J., Shaytan, A. K., Mariño-Ramírez, L., Talbert, P. B., Landsman, D., and Panchenko, A. R. HistoneDB 2.0: a histone database with variants - an integrated resource to explore histones and their variants. *Database*, 2016:baw014, 2016.

[22] Thatcher, T. H. and Gorovsky, M. A. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Research*, 22(2): 174–179, 1994.

[23] Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A., and Landsman, D. Histone structure and nucleosome stability. *Expert Review of Proteomics*, 2(5):719–729, 2005.

[24] Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., and Stuckey, P. J. Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics*, 27(23):3315–3316, 2011.

[25] Humphrey, W., Dalke, A., and Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[26] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., and Forslund, K. The Pfam protein families database. *Nucleic Acids Research*, 38(suppl_1): D211 – D222, 2010.

[27] Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., and Durbin, R. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl_1):D247 – D251, 2006.

[28] Feijoo-Siota, L., Rama, J. L. R., Sánchez-Pérez, A., and Villa, T. G. Considerations on bacterial nucleoids. *Applied Microbiology and Biotechnology*, 101(14):5591–5602, 2017.

[29] Christodoulou, E. and Vorgias, C. E. Cloning, overproduction, purification and crystallization of the DNA binding protein HU from the

hyperthermophilic eubacterium *Thermotoga maritima*. *Acta Crystallographica Section D: Biological Crystallography*, 54(5):1043–1045, 1998.

[30] Vis, H., Mariani, M., Vorgias, C. E., Wilson, K. S., Kaptein, R., and Boelens, R. Solution structure of the HU protein from Bacillus stearothermophilus. *Journal of Molecular Biology*, 254(4):692–703, 1995.

[31] Rice, P. A., Yang, S.-w., Mizuuchi, K., and Nash, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell*, 87 (7):1295–1306, 1996.

[32] Swinger, K. K., Lemberg, K. M., Zhang, Y., and Rice, P. A. Flexible DNA bending in HU–DNA cocrystal structures. *The EMBO Journal*, 22(14):3749–3760, 2003.

[33] Bao, Q., Chen, H., Liu, Y., Yan, J., Dröge, P., and Davey, C. A. A divalent metal-mediated switch controlling protein-induced DNA bending. *Journal of Molecular Biology*, 367(3):731–740, 2007.

[34] Mouw, K. W. and Rice, P. A. Shaping the *Borrelia burgdorferi* genome: crystal structure and binding properties of the DNA-bending protein Hbb. *Molecular Microbiology*, 63(5):1319–1330, 2007.

[35] Osipiuk, J., Makowska-Grzyska, M., Hasseman, J., Anderson, W. F., and Joachimiak, A. PDB ID: 3rhi, DNA-binding protein HU from *Bacillus anthracis*. 2011.

[36] Bhowmick, T., Ghosh, S., Dixit, K., Ganesan, V., Ramagopal, U. A., Dey, D., Sarma, S. P., Ramakumar, S., and Nagaraja, V. Targeting *Mycobacterium tuberculosis* nucleoid-associated protein HU with structure-based inhibitors. *Nature Communications*, 5:4124, 2014.

[37] Boyko, K. M., Gorbacheva, M. A., Rakitina, T. V., Korgenevsky, D. A., Kamashev, D. E., Vanyushkina, A. A., Lipkin, A. V., and Popov, V. O. PDB ID: 4n1v, Structure of micoplasma DNA binding HU protein. 2013.

[38] Kim, D.-H., Im, H., Jee, J.-G., Jang, S.-B., Yoon, H.-J., Kwon, A.-R., Kang, S.-M., and Lee, B.-J. β-Arm flexibility of HU from *Staphylococcus*

aureus dictates the DNA-binding and recognition mechanism. *Acta Crystallographica Section D: Biological Crystallography*, 70(12):3273–3289, 2014.

[39] Trowitzsch, S., Viola, C., Scheer, E., Conic, S., Chavant, V., Fournier, M., Papai, G., Ebong, I.-O., Schaffitzel, C., and Zou, J. Cytoplasmic TAF2–TAF8–TAF10 complex provides evidence for nuclear holo–TFIID assembly from preformed submodules. *Nature Communications*, 6:6011, 2015.

[40] Decanniere, K., Babu, A. M., Sandman, K., Reeve, J. N., and Heinemann, U. Crystal structures of recombinant histones HMfA and HMfB from the hyperthermophilic archaeon *Methanothermus fervidus*. *Journal of Molecular Biology*, 303(1):35–47, 2000.

[41] Decanniere, K., Babu, A. M., Sandman, K., Reeve, J. N., and Heinemann, U. Crystal structures of recombinant histones HMfA and HMfB from the hyperthermophilic archaeon *Methanothermus fervidus*. *Journal of Molecular Biology*, 303(1):35–47, 2000.

[42] Fahrner, R. L., Cascio, D., Lake, J. A., and Slesarev, A. An ancestral nuclear protein assembly: crystal structure of the *Methanopyrus kandleri* histone. *Protein Science*, 10(10):2002–2007, 2001.

[43] Li, T., Sun, F., Ji, X., Feng, Y., and Rao, Z. Structure based hyperthermostability of archaeal histone HPhA from Pyrococcus horikoshii. *Journal of Molecular Biology*, 325(5):1031–1037, 2003.

[44] Romier, C., Cocchiarella, F., Mantovani, R., and Moras, D. The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *Journal of Biological Chemistry*, 278(2):1336–1345, 2003.

[45] Huber, E. M., Scharf, D. H., Hortschansky, P., Groll, M., and Brakhage, A. A. DNA minor groove sensing and widening by the CCAAT-binding complex. *Structure*, 20(10):1757–1768, 2012.

[46] Yang, H., Zhang, T., Tao, Y., Wu, L., Li, H.-t., Zhou, J.-q., Zhong, C., and Ding, J. *Saccharomyces cerevisiae* MHF complex structurally

resembles the histones (H3-H4) 2 heterotetramer and functions as a heterotetramer. *Structure*, 20(2):364–370, 2012.

[47] Nishino, T., Takeuchi, K., Gascoigne, K. E., Suzuki, A., Hori, T., Oyama, T., Morikawa, K., Cheeseman, I. M., and Fukagawa, T. CENP-TWSX forms a unique centromeric chromatin structure with a histone-like fold. *Cell*, 148(3):487–501, 2012.

[48] Tao, Y., Jin, C., Li, X., Qi, S., Chu, L., Niu, L., Yao, X., and Teng, M. The structure of the FANCM-MHF complex reveals physical features for functional assembly. *Nature Communications*, 3:782, 2012.

[49] Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251, 1997.

[50] Harp, J. M., Hanson, B. L., Timm, D. E., and Bunick, G. J. Asymmetries in the nucleosome core particle at 2.5 Å resolution. *Acta Crystallographica Section D: Biological Crystallography*, 56(12):1513–1534, 2000.

[51] White, C. L., Suto, R. K., and Luger, K. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *The EMBO Journal*, 20(18):5207–5218, 2001.

[52] Tsunaka, Y., Kajimura, N., Tate, S.-i., and Morikawa, K. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Research*, 33(10):3424–3434, 2005.

[53] Chakravarthy, S. and Luger, K. PDB ID: 2nqb, Comparative analysis of nucleosome structures from different species. 2006.

[54] Qiu, Y., Tereshko, V., Kim, Y., Zhang, R., Collart, F., Yousef, M., Kossiakoff, A., and Joachimiak, A. The crystal structure of Aq_328 from the hyperthermophilic bacteria *Aquifex aeolicus* shows an ancestral histone fold. *Proteins: Structure, Function, and Bioinformatics*, 62(1):8–16, 2006.

[55] Wang, H., Murayama, K., Terada, T., Chen, L., Liu, Z. J., Wang, B. C., Shirouzu, M., Kuramitsu, S., and Yokoyama, S. PDB ID: 1wwi, Crystal structure of ttk003001566 from *Thermus thermophilus* HB8. 2005.

[56] Shindo, H., Iwaki, T., Ieda, R., Kurumizaka, H., Ueguchi, C., Mizuno, T., Morikawa, S., Nakamura, H., and Kuboniwa, H. Solution structure of the DNA binding domain of a nucleoid-associated protein, H-NS, from *Escherichia coli*. *FEBS Letters*, 360(2):125–131, 1995.

[57] Rosselli, L. K., Sforca, M. L., Souza, A. P., and Zeri, A. C. PDB ID: 2jr1, Solution structure of the DNA binding domain of a nucleoid-associated protein, H-NS, from the phytopathogen *Xylella fastidiosa*. 2007.

[58] Gordon, B. R. G., Li, Y., Cote, A., Weirauch, M. T., Ding, P., Hughes, T. R., Navarre, W. W., Xia, B., and Liu, J. Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proceedings of the National Academy of Sciences*, 108(26):10690–10695, 2011.

[59] Joint Center for Structural Genomics. PDB ID: 4fmr, Crystal structure of a hypothetical protein (BVU_2165) from *Bacteroides vulgatus* ATCC 8482 at 2.25 Å resolution. 2012.

[60] Xie, X., Kokubo, T., Cohen, S. L., and Mirza, U. A. Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature*, 380(6572):316, 1996.

[61] Werten, S., Mitschler, A., Romier, C., Gangloff, Y.-G., Thuault, S., Davidson, I., and Moras, D. Crystal structure of a subcomplex of human transcription factor TFIID formed by TATA binding protein-associated factors hTAF4 (hTAFII135) and hTAF12 (hTAFII20). *Journal of Biological Chemistry*, 277(47):45502–45509, 2002.

[62] Birck, C., Poch, O., Romier, C., Ruff, M., Mengus, G., Lavigne, A.-C., Davidson, I., and Moras, D. Human TAF II 28 and TAF II 18 interact through a histone fold encoded by atypical evolutionary conserved motifs also found in the SPT3 family. *Cell*, 94(2):239–249, 1998.

[63] Stockner, T., Plugariu, C., Koraimann, G., Högenauer, G., Bermel, W., Prytulla, S., and Sterk, H. Solution structure of the DNA-binding domain of TraM. *Biochemistry*, 40(11):3370–3377, 2001.

[64] Lu, J., Wong, J. J. W., Edwards, R. A., Manchak, J., Frost, L. S., and Glover, J. N. Structural basis of specific TraD–TraM recognition during F plasmid-mediated bacterial conjugation. *Molecular Microbiology*, 70 (1):89–99, 2008.

[65] Krissinel, E. and Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12): 2256–2268, 2004.

[66] Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[67] Talevich, E., Invergo, B. M., Cock, P. J. A., and Chapman, B. A. Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13(1):209, 2012.

[68] Rambaut, A. FigTree, a graphical viewer of phylogenetic trees. *See http://tree.bio.ed.ac.uk/software/figtree*, 2007.

[69] Huson, D. H. and Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[70] Kelly, K. L. Twenty-two colors of maximum contrast. *Color Engineering*, 3(26):26–27, 1965.

[71] Bergsten, J. A review of long-branch attraction. *Cladistics*, 21(2): 163–193, 2005.

[72] Chandonia, J.-M., Fox, N. K., and Brenner, S. E. SCOPe: Manual curation and artifact removal in the structural classification of proteins–extended Database. *Journal of Molecular Biology*, 429(3):348–355, 2017.

[73] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., and Sangrador-Vegas, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279 – D285, 2016.

[74] Nogales, E., Fang, J., and Louder, R. K. Structural dynamics and DNA interaction of human TFIID. *Transcription*, 8(1):55–60, 2017.

[75] Hampsey, M. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews*, 62 (2):465–503, 1998.

[76] Sanders, S. L. and Weil, P. A. Identification of Two Novel TAF Subunits of the yeast *Saccharomyces cerevisiae* TFIID Complex. *Journal of Biological Chemistry*, 275(18):13895–13900, 2000.

[77] Shao, H., Revach, M., Moshonov, S., Tzuman, Y., Gazit, K., Albeck, S., Unger, T., and Dikstein, R. Core promoter binding by histone-like TAF complexes. *Molecular and Cellular Biology*, 25(1):206–219, 2005.

[78] Hoffmann, A., Chiang, C.-M., Oelgeschlager, T., and Xie, X. A histone octamer-like structure within TFIID. *Nature*, 380(6572):356, 1996.

[79] Xie, X., Kokubo, T., Cohen, S. L., and Mirza, U. A. Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature*, 380(6572):316, 1996.

[80] Hoffmann, A., Oelgeschläger, T., and Roeder, R. G. Considerations of transcriptional control mechanisms: Do TFIID–core promoter complexes recapitulate nucleosome-like functions? *Proceedings of the National Academy of Sciences*, 94(17):8928–8935, 1997.

[81] Selleck, W., Howley, R., Fang, Q., Podolny, V., Fried, M. G., Buratowski, S., and Tan, S. A histone fold TAF octamer within the yeast TFIID transcriptional coactivator. *Nature Structural & Molecular Biology*, 8(8):695–700, 2001.

[82] Werten, S., Mitschler, A., Romier, C., Gangloff, Y.-G., Thuault, S., Davidson, I., and Moras, D. Crystal structure of a subcomplex of human transcription factor TFIID formed by TATA binding protein-associated factors hTAF4 (hTAFII135) and hTAF12 (hTAFII20). *Journal of Biological Chemistry*, 277(47):45502–45509, 2002.

[83] McAinsh, A. D. and Meraldi, P. The CCAN complex: linking centromere specification to control of kinetochore–microtubule dynamics.

In *Seminars in Cell & Developmental Biology*, volume 22, pages 946–952. Elsevier, 2011.

[84] Meraldi, P., McAinsh, A. D., Rheinbay, E., and Sorger, P. K. Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biology*, 7(3):R23, 2006.

[85] Musacchio, A. and Desai, A. A molecular view of kinetochore assembly and function. *Biology*, 6(1):5, 2017.

[86] Fukagawa, T. and Earnshaw, W. C. The centromere: chromatin foundation for the kinetochore machinery. *Developmental Cell*, 30(5):496–508, 2014.

[87] Foltz, D. R. and Stukenberg, P. T. A new histone at the centromere? *Cell*, 148(3):394–396, 2012.

[88] Hori, T., Amano, M., Suzuki, A., Backer, C. B., Welburn, J. P., Dong, Y., McEwen, B. F., Shang, W.-H., Suzuki, E., and Okawa, K. CCAN makes multiple contacts with centromeric DNA to provide distinct pathways to the outer kinetochore. *Cell*, 135(6):1039–1052, 2008.

[89] Hadjithomas, M. and Moudrianakis, E. N. Experimental evidence for the role of domain swapping in the evolution of the histone fold. *Proceedings of the National Academy of Sciences*, 108(33):13462–13467, 2011.

[90] Nabhan, A. R. and Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, 13(1):122–134, 2012.

[91] Graybeal, A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology*, pages 9–17, 1998.

[92] Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

Chapter 5

The ferritin-like superfamily

Overview

This chapter describes the first ever implementation of the MD-based bootstrap method on the structural phylogeny of the ferritin-like protein superfamily, as classified by SCOP. The chapter starts by summarizing the gaps in our understanding using conventional sequence-based methods, namely the poor resolution of relationships when organizing proteins into hierarchical databases and reconstructing deep evolutionary relationships. This is followed by stressing the importance of the use of a structural method for inferring relationships and presenting a brief summary of the structural method introduced in this thesis. Finally, the complete method is implemented on the ferritin-like superfamily, i.e. using structural information to recover evolutionary relationships between members of this superfamily and using the MD-based bootstrap method to gauge the significance of the results.

5.1 Introduction

5.1.1 PFam, SCOP and CATH

With the growth in molecular data related to proteins, an increasing number of databases have been set up to organize the information relating to these biomolecules. Notable databases that carry out this daunting task are PFam [1], SCOP [2] and CATH [3], which were described in detail in Chapter 1, Section 1.4.

PFam aims to organise protein sequences into families and further clusters families into clans [4] based on homology. SCOP and CATH aim to organise protein 3D structures based on structural similarity. These structural databases work on different philosophies with SCOP organising structures based on classes, folds, superfamilies and families, whereas CATH attempts to organise structural information based on classes, architecture, topology and homology. In the cases of both SCOP and CATH, structures at the bottom of these hierarchies are said to share an evolutionary origin.

All the aforementioned databases aim at grouping proteins to reveal the underlying evolutionary dynamics and do so to a certain extent. However the relationships between members grouped at certain hierarchies remain unresolved, i.e. polytomies exist at clan and family levels in PFam, su-

perfamily and family levels in SCOP and topology and homology levels in CATH. A polytomy is illustrated in Figure 5.1.

One reason behind PFam failing to offer evolutionary insight is the significant divergence between sequences clustered into clans which does not allow for conventional sequence-based phylogenetic methods to recover meaningful evolutionary histories. For structural databases, the case is slightly different. While there exists notable similarity between structures, a complete structure-based phylogenetic method is missing.

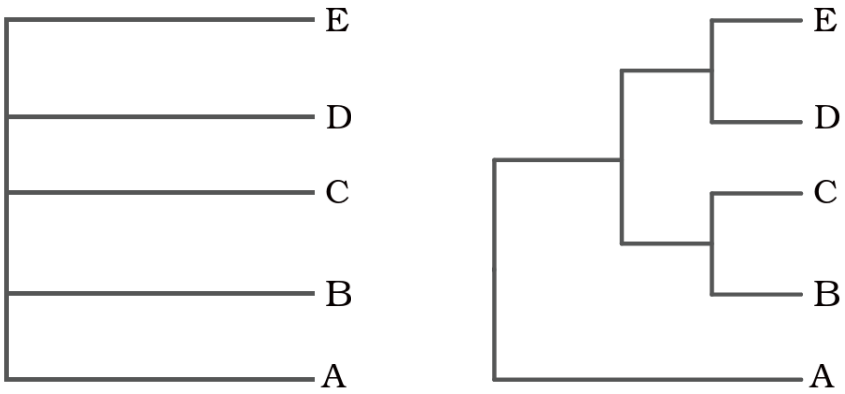


Figure 5.1: Unresolved (left) relationships between a set of five proteins (A-E) clustered at a particular level of hierarchy in a database. Resolved relationships (right) provide evolutionary insight beyond simple clustering, something that is not presently provided by databases.

5.1.2 Structural methods in phylogenetics

The novel structure-based method developed in this thesis aims to use the conservation in structure to infer deep evolutionary relationships and simultaneously resolve polytomies in the structural databases. This method is different from previous attempts to use protein structure for evolutionary analysis, and is explained briefly in the following sections.

5.1.2.1 *Structural alignments and scoring*

The methods used in previous studies to incorporate structural distance are either subjective or incomplete. The subjectivity is incorporated while aligning protein structures for the purpose of determining similarity. To reliably compare two 3D protein structures, a set of equivalent points need

to be determined, amongst which similarity or distance is to be quantified. Previous studies, which were limited to a handful of structures, did this either manually [5, 6] or through the use of a structural comparison metric DALI [7, 8].

While the manual approach is not tractable when comparing more than a handful of structures, the DALI comparison score is not normalised by default, making DALI an incomplete metric for the determination of distance-based phylogenies. The normalization is required to convert the similarity score to a distance score for use with a distance-based method. To normalise the similarity scores from DALI, a background distribution is used. This is questionable as the distribution resulting from a limited number of comparisons is bound to change as more comparisons are incorporated, impacting the normalization process.

The use of secondary structure matching (SSM)-based Q_{score} [9], as done in this method, circumvents both these problems, i.e. selecting equivalent residues in an automated and robust way, discussed at length in Chapter 2, Section 2.2.7, and generating a normalised similarity score which can easily be converted to a distance score.

5.1.2.2 *Robustness of phylogenetic relationships*

Once a phylogenetic tree is constructed from the distance data, determined from comparing structures, the topology of the tree needs to be tested. In previous structural distance-based studies this has either not been attempted [5, 6], has been integrated using sequence-based testing methods [8] or done in a qualitative way [10], see Chapter 1 for details. The introduction of the MD-based bootstrap method, in this thesis, provides a robust quantitative way of testing the tree topology moving away from qualitative assessments and sequence-based approaches.

Conventionally, the bootstrap analysis generates perturbations in the original distance data, by changing the underlying alignments, to see if the relationships on the phylogenetic tree are affected. To this end, replicates of the original sequence alignment are generated using parametric or non-parametric approaches. While a limited understanding of protein structural evolution prevents the use of parametric approaches, the non-parametric approach implemented with protein sequence data is not scalable to protein structures, as outlined in Chapter 1, Section 1.5.4.

The use of a MD-based bootstrap method aims to replicate the effect of the non-parametric bootstrap analysis and associates a measure of significance to the relationships inferred from the structural distance data. The MD-based bootstrap method creates perturbations by employing MD simulations to sample alternative conformations of each structure. While the reference tree utilizes the comparisons between the reference structures, comparisons between alternative conformations are used to generate the replicate trees. Enumerating the number of times the relationships, as shown by the reference tree, are revisited in the replicate trees and expressing counts as a fraction of the total replicate datasets sampled quantifies the robustness of relationships when structural and hence distance data is perturbed.

Between the use of the SSM-based Q_{score} metric for determining equivalence between structures and generating a normalized similarity score and the use of the MD-based bootstrap method to test the final topology of the inferred tree, a complete method to infer phylogenies from protein structure is presented in this work.

A previous structure-based study of the ferritin-like superfamily, by Lundin et al. [10], used the SSM-based Q_{score} to generate distances between protein 3D structures and used these distances to infer evolutionary relationships between them. The structural phylogeny [10] was subsequently tested qualitatively. In this work, the MD-based bootstrap method was employed to associate a quantitative measure of significance to the results inferred in that study. The method, through this association, also aims at resolving polytomies within protein structural databases, using the ferritin-like protein superfamily as a test case. The biological interpretation of the ferritins phylogeny is already extensively covered [10], therefore this work aims to comment on the correctness of those results.

5.2 Methods

5.2.1 Structural data

The protein structures as used in the previous study [10] were obtained from RCSB (www.rcsb.org). These are listed along with their SCOP and PFam classifications in Tables 5.1 and 5.2. The family annotations used to cluster proteins were acquired from descriptions listed by RCSB for each structure and are the same as in the previous study [10].

5.2.2 Structural phylogeny

MD simulations, discussed in the following section, were carried out for each of the structures. An intermediate step while preparing the structural data for MD simulations generated slightly refined structures which were used to generate the reference tree using the following steps.

1. Pairwise comparisons were done for each protein structure using Superpose [9]. Due to the nature of the algorithm comparisons were order specific, i.e. A $\cong$ B $\neq$ B $\cong$ A. The pairwise scores were averaged to attain a final score "q" for the comparison between structures A and B.

2. The similarity score was subtracted from one (i.e. $d = 1 - q$) to generate a distance value (d).

3. A matrix was populated with the pairwise distances.

4. The neighbour-joining (NJ) algorithm [11] as implemented by the Phylo [12] package in Biopython [13] was used to generate an unrooted NJ-tree from the matrix.

5. Splitstree [14] was used to obtain a neighbour-net network from the matrix in 3.

5.2.3 MD simulations and the bootstrap-like analysis

The NAMD [87] program along with the CHARMM36FF [88] force field was used to generate alternative conformations for each of the structures in the dataset. The following steps were followed.

1. Each structure was energy minimized for 200 steps using the default minimiser in NAMD [87].

2. Each structure was solvated using the TIP3P [89] water model. In this step, a minimum cubic box was created with dimensions that fit the protein, i.e. between the minimum $(x_{min}, y_{min}, z_{min})$ and maximum $(x_{max}, y_{max}, z_{max})$ coordinates of the protein. The boundaries were extended by 15 Å in each direction and the newly added volume was filled with the solvent.

Table 5.1: SCOP and Pfam identifiers of protein structures previously used in determining the structural phylogeny [10]. Unclassified entries are left blank. "*" indicates structures that were excluded from this analysis due to poor structural geometries rendering them inedequate for MD simulations. Colours are for visual guidance only. [15–86]

PDB	Chain	SCOP	Pfam	PDB	Chain	SCOP	Pfam
1bcf	A	a.25.1.1	PF00210	1uzr	A	a.25.1.2	PF00268
1bg7	A	a.25.1.1	PF00210	1w68	A	a.25.1.2	PF00268
1dps	A	a.25.1.1	PF00210	*2rcc	A		PF00268
1eum	A	a.25.1.1	PF00210	2uw2	A	a.25.1.2	PF00268
1jgc	A	a.25.1.1	PF00210	*3dhz	A	a.25.1.2	PF00268
1ji4	A	a.25.1.1	PF00210	3ee4	A		PF00268
1ji5	A	a.25.1.1	PF00210	*1z3a	A	c.97.1.2	PF00383
1jig	A	a.25.1.1	PF00210	*1mhy	B	a.25.1.2	PF02332
1jts	A	a.25.1.1	PF00210	*1mhy	D	a.25.1.2	PF02332
1krq	A	a.25.1.1	PF00210	1mty	B	a.25.1.2	PF02332
1lb3	A	a.25.1.1	PF00210	1mty	D	a.25.1.2	PF02332
1n1q	A	a.25.1.1	PF00210	*1xvb	A	a.25.1.2	PF02332
1nfv	A	a.25.1.1	PF00210	*1xvb	C	a.25.1.2	PF02332
1o9r	A	a.25.1.1	PF00210	*2inc	A	a.25.1.2	PF02332
1qgh	A	a.25.1.1	PF00210	2inc	B	a.25.1.2	PF02332
1r03	A	a.25.1.1	PF00210	*2inp	A		PF02332
*1rci	A	a.25.1.1	PF00210	2inp	C		PF02332
*1s3q	A	a.25.1.1	PF00210	3dhg	A	a.25.1.2	PF02332
1tjo	A	a.25.1.1	PF00210	*3dhg	B	a.25.1.2	PF02332
1tk6	A	a.25.1.1	PF00210	1lko	A	a.25.1.1	PF02915
1uvh	A	a.25.1.1	PF00210	*1vjx	A	a.25.1.1	PF02915
1vlg	A	a.25.1.1	PF00210	1yuz	A	a.25.1.1	PF02915
1z6o	A	a.25.1.1	PF00210	2fzf	A	a.25.1.1	PF02915
1z6o	M	a.25.1.1	PF00210	*2oh3	A	a.25.1.8	PF02915
2chp	A	a.25.1.0	PF00210	3qhb	A		PF02915
*2clb	A		PF00210	1afr	A	a.25.1.2	PF03405
2fjc	A	a.25.1.1	PF00210	*1oqb	A	a.25.1.2	PF03405
2fkz	A	a.25.1.1	PF00210	1za0	A	a.25.1.2	PF03405
2jd7	A	a.25.1.0	PF00210	2uw1	A	a.25.1.2	PF03405
*2qqy	A		PF00210	2uw1	B	a.25.1.2	PF03405
2ux1	A	a.25.1.1	PF00210	*2qf9	A		PF03713
2vzb	A		PF00210	*1o9i	A	a.25.1.3	PF05067
2za7	A	a.25.1.1	PF00210	*2cwl	A	a.25.1.3	PF05067
3e1q	A	a.25.1.1	PF00210	1otk	A	a.25.1.2	PF05138
3e6s	A	a.25.1.1	PF00210	*2gs4	A	a.25.1.4	PF05974
*3fvb	A	a.25.1.0	PF00210	*2gyq	A	a.25.1.4	PF05974
1jk0	A	a.25.1.2	PF00268	*2itb	A	a.25.1.7	PF06175
*1jk0	B	a.25.1.2	PF00268	*3fse	A		PF09537
1mxr	A	a.25.1.2	PF00268	*2oc5	A	a.25.1.6	PF11266
1oqu	A	a.25.1.2	PF00268	*3ez0	A		PF13794
1r2f	A	a.25.1.2	PF00268	*2ib0	A	a.25.1.9	PF14530
*1syy	A	a.25.1.2	PF00268				

Table 5.2: The reduced structural dataset used in this analysis was spread across two manually curated SCOP families in one superfamily (SF) and seven protein families in one PFam clan. Families marked with "*" were included in the previously determined structural phylogeny [10] but excluded from this analysis due to poor structural geometries rendering them inedequate for MD simulations.

SCOP SF	**Family**	**Annotation**
a.25.1	a.25.1.1	Ferritin
	a.25.1.2	Ribonucleotide reductase-like
	*a.25.1.3	Manganese catalase (T-catalase)
	*a.25.1.4	YciF-like
	*a.25.1.6	PMT1231-like
	*a.25.1.7	MiaE-like
	*a.25.1.8	AMB4284-like
	*a.25.1.9	Rv2844-like
	a.25.1.0	automated matches
a.97.1	*a.97.1.2	class I lysyl-tRNA synthetase
Pfam Clan	**Family**	
CL0044	PF00210	Ferritin
	*PF03713	DUF305
	*PF14530	DUF4439
	*PF05974	DUF892
	PF00268	Ribonuc_red_sm
	PF02915	Rubrerythrin
	PF02332	Phenol_Hydrox
	PF05067	Mn_catalase
	PF05138	PaaA_PaaC
	PF03405	Fatty acid desaturase
	*PF06175	MiaE
	*PF13794	MiaE_2
	*PF11266	fatty aldehyde decarbonylase
	*PF09537	DUF2383
CL0109	*PF00383	dCMP_cyt_deam_1

3. After minimization, excess charge was neutralized (if present) through the addition of Na^+ and Cl^- as counter ions, for effective calculation of long-range electrostatics using PME [90] summations. The ions were added by randomly selecting a water molecule and substituting it for the ion.

4. After solvation and ionization, the system was minimized (for 300 steps using the default minimizer) to remove clashes and find a new potential energy minimum. These structures were used to generate the reference tree as discussed in the previous section.

5. Following the previous step, a heating phase of the MD simulations was conducted. To achieve the simulation temperature of 310 K, the temperature was increased from 0 K in 5 K increments every 10000 integration steps, with each step being 2 fs apart.

6. After a temperature of 310 K was achieved, each structure was simulated for an additional 20 ns and conformations were recorded every 2 ps.

All the simulations were conducted in an NPT ensemble to mimic real-life conditions, at a pressure of 1 atm. The Lennard-Jones potential was switched to zero between 10 Å and 12 Å whereas a 12 Å cutoff was used for calculating the electrostatic interactions. Langevin dynamics were used for the temperature control and a modified Nosé-Hoover Langevin control for pressure was implemented, using a piston period and decay of 100 fs and 50 fs respectively. Hydrogen bond lengths were constrained using SHAKE [91]. Sampled conformations were recorded into trajectories every 5000 integration steps (10 ps). The trajectories for each of the system were analysed using VMD [92] to extract conformations for further analysis.

Not all the structures used in the previous study were included in this analysis. To reduce the computational time for this work, thinning of the dataset was carried out. This was done by removing some of the families uncharacterised by SCOP, or having one or two members or incomplete structures, i.e. where structures had missing residues. Furthermore, some structures failed the simulation process, resulting in a RATTLE error, indicating problems with the structural geometries. Additional refinements carried out by further energy minimising these structures did not correct the

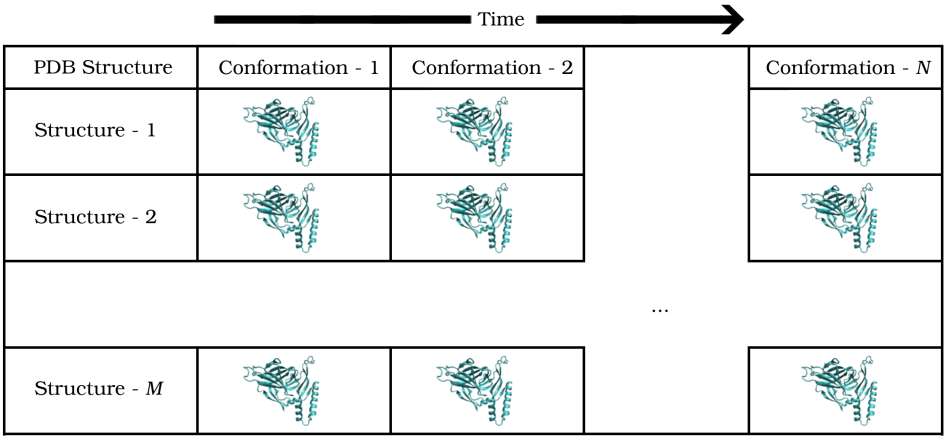


Figure 5.2: MD trajectories of protein structures. The conformations sampled are used for generating replicate datasets when implementing the MD-based bootstrap method. For 1000 conformations recorded, i.e. $N = 1000$, in a dataset comprising 53 structures, i.e. $M = 53$, examples of two replicates are as follows. Replicate-1 = $[S_1^{610}, S_2^{17}, S_3^{333}, ..., S_{53}^{980}]$, Replicate-2 = $[S_1^{1}, S_2^{900}, S_3^{877}, ..., S_{53}^{23}]$, where S_3^{877} represents the 877th conformation from the trajectory of Structure 3.

geometries and hence some structures had to be excluded. In total 53 proteins structures that had clean structural geometries, were well-characterised by various databases and were part of clusters regarding which important inferences had previously [10] been drawn, were included in the analysis. The structures that were excluded from this work are marked with "*" in Table 5.1.

For the remaining structures a total of 100 replicate structure datasets were created, constituting 100 trials. To create a replicate dataset, a conformation was randomly selected for each of the simulated structures from the production phases, Figure 5.3, in their respective trajectories. Each replicate dataset was used to generate a phylogenetic tree, using steps oulined in Section 5.2.2. The relationships between protein structures in the trees from the replicates were compared to those in the reference tree, using a phylogenetic tree summarization program SumTrees [93] as made available by DendroPy [94], a Python library for phylogenetic computing. The recovered relationships were expressed as a fraction of the total number of trials on the nodes in the reference phylogenetic tree. The final tree, with support, was visualized using Figtree [95].

5.3 Results

5.3.1 PFam and SCOP classifications

The primary aim of this work was to implement the MD-based bootstrap method, developed in this thesis, to add a measure of confidence to the relationships inferred by structural phylogenies. The group of proteins to which the MD-based bootstrap method was applied are the ferritin-like superfamily, for which a structural phylogeny was generated in a previous study [10]. This work therefore only aims to assess previously proposed evolutionary relationships based on structure and not exhaustively describe the biology of the ferritin-like superfamily.

The reduced protein structure dataset used in this work spanned one SCOP superfamily, i.e. ferritin-like (a.25.1) comprising two manually curated protein families, i.e. ferritin (a.25.1.1) and ribonucleotide reductase-like (a.25.1.2). The same structural data spanned seven PFam families, i.e. ferritins (PF00210), Ribonuc_red_sm (PF00268), Rubrerythrin (PF02915), Phenol_Hydrox (PF02332), Mn_catalase (PF05067), Fatty acid desaturase (PF03405) and PaaA_PaaC (PF05138). The PFam families were classified into a single clan, i.e. ferritin (CL0044). The structures and their classifications by both SCOP and PFam are listed in Tables 5.1 and 5.2.

There exists consistency between the classifications of ferritin proteins by SCOP (a.25.1.1) and PFam (PF00210). For the remaining proteins, i.e. the SCOP ribonucleotide reductase-like family (a.25.1.2), members of the same cluster are classified into three distinct families by PFam Phenol_Hydrox (PF02332), Ribonuc_red_sm (PF00268) and Fatty acid desaturase (PF03405).

While in one case, a.25.1.1/PF00210, sequence and structural conservation are correlated, i.e. SCOP and PFam agree on the protein family assignment, the other case shows poor agreement between SCOP and PFam. As is the case with ribonucleotide reductase-like proteins, poor sequence similarity results in PFam classifying proteins into three separate families whereas structural similarity leads SCOP to group them into a single family. While both SCOP and PFam indicate shared evolutionary origin, through assignment of the proteins to a superfamily, by SCOP (a.25.1), and a clan, by PFam (CL0044), no further insight is provided by these databases into the relationships shared between member proteins.

The previous results from the structural phylogeny [10] of the ferritin-like superfamily, as classified manually by SCOP, make it an ideal case to which to apply the MD-based bootstrap method. By focusing on structural conservation, deep evolutionary relationships can be uncovered that conventional sequence-based methods are not able to tackle, as seen by break down of PFam classification of the ribonucleotide reductase-like proteins, and a measure of confidence in these relationships is provided by the addition of statistical support. Existing polytomies within protein databases, Figure 5.1, can also be resolved with confidence allowing for further organisation at different level of hierarchies.

5.3.2 MD trajectory stability

While the primary structural phylogeny is generated using the reference structures, see Section 5.2.3 (point 4), to enact the MD-based bootstrap method alternative conformations are extracted from MD simulations. Prior to extraction of the conformations from trajectories the quality of the simulations needs to be assessed, which is done by calculating the all atom-positional RMSD. The RMSD values were plotted against simulation time to gauge the structural stability. Figure 5.3 shows the RMSD trends for all the trajectories.

To calculate the RMSD, each conformation sampled in a trajectory was compared to its starting structure (at 0 K) using Equation 5.1. Successive conformations in the heating phase gradually show the structure moving away from the starting structure as they accumulate kinetic energy. When the temperature equilibrates at 310 K, i.e. in the equlibration phase, the RMSD values reach a plateau in all cases except two (1mty_B and 1z60_M). For small to medium-sized proteins, i.e. a few hundred amino acid residues, as is the case with the ferritin-like superfamily, the observed range of RMSD values correspond to conformational fluctuation without substantial unfolding.

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)} \quad (5.1)$$

Where the starting structure v and the sampled conformation w comprise two sets of n points, each of which has x, y, z 3D Cartesian coordinates.

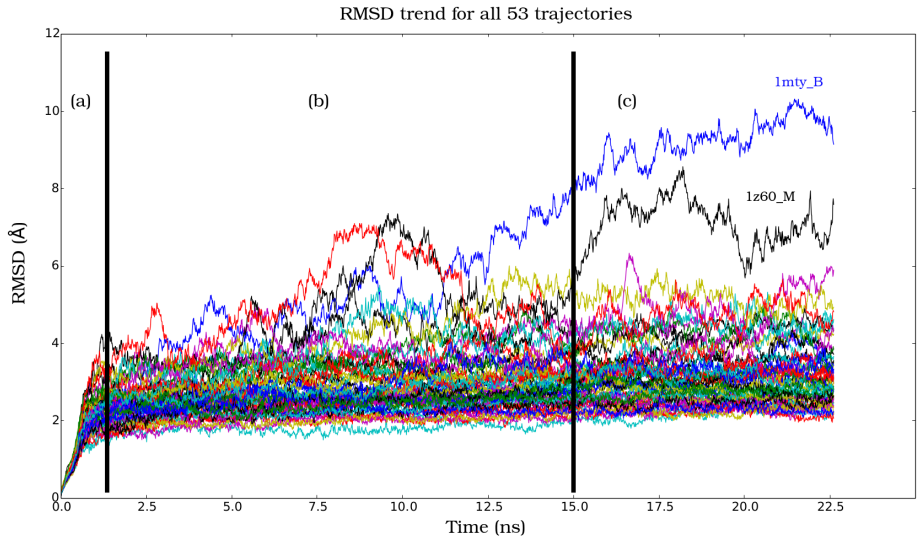


Figure 5.3: RMSD trends for 53 trajectories. RMSD was calculated by comparing each conformation sampled to the starting structure for each simulation. The simulations were broken down into three distinct phases, (a) heating, i.e. the temperature of the structure was increased from 0 K to 310 K, (b) an equilibration phase, i.e. the system was allowed to attain an equilibrium, empirically chosen to have occurred at 15.0 ns (c) a production phase, i.e. the remainder of the trajectory from which conformations were selected for the MD-based bootstrap method. All structures appear to have converged except 1mty_B and 1z60_M.

Although any conformation from the entire trajectory can be ideally used in the MD-based bootstrap method, to maintain consistency with standard MD simulation practices, the simulation was broken down into three phrases, i.e. heating, equilibration and production, Figure 5.3. An empirical time-limit cannot be associated with the equilibration period, however a general consensus is to observe a plateau in the measured quantity indicating equilibration [96], which in this case is RMSD. Here, the RMSD trend indicates that conformations extracted after 15 ns of simulation, i.e. in the phase labelled "production", Figure 5.3, can be considered to be equilibrated. The bootstrap-like trials therefore used 100 conformations for each structure that were randomly extracted from the repertoire of 750 conformations available in the production phase of each of the respective trajectories.

5.3.3 Interpretation of results from the MD-based bootstrap method

Previously, for the ferritin-like superfamily [10], the phylogeny was broadly inferred from the reticulated network and assessed qualitatively, using structural topology, i.e. dimerization types. Figure 5.4 shows the same type of network for the 53 ferritin-like structures studied here. The same distance data is shown as a phylogenetic tree in Figure 5.5 with the nodes carrying percentages indicating their support from the MD-based bootstrap method. The use of this method provides an intuitive quantitative assessment of the phylogeny, which is subsequently used to validate the evolutionary inferences.

The percentages reflected on the nodes in Figure 5.5 need a careful interpretation. A percentage (x), on a node reflects the number of times the structures present on that node were observed on the same node in the trees from the replicate datasets. The number is independent of the arrangement of structures on the node and hence only signifies occupancy. For example, a support of 77 on a particular node would represent that the occupancy, as reflected on the reference tree, was recovered 77 times out of the 100 trials conducted (77%), indicating for the remaining 23 times, one or more of the structures departed that particular node.

5.3.4 Structural phylogeny of the ferritin-like superfamily

Having developed an understanding of the support values from the bootstrap-like method, some of the important conclusions from Lundin et al. [10] are assessed for correctness below.

Ferritins, Bacterioferritins and Dps share a common evolutionary history.

A shared ancestry of ferritins, bacterioferritins and Dps is validated by the strong support, N1 = 100, on the node carrying these structures, Figure 5.5. It is perhaps interesting to note that strong support also exists for the split within the "Dps and related" cluster, N2 = 100, suggesting that there are two separate groupings, Figure 5.5.

On the other hand, as reflected by some departure from tree-likeness in the network, Figure 5.4, the relationships between ferritins and bacterioferritins cannot be easily resolved. This is illustrated by the poor support (i.e. values are lesser than 50%) on the nodes occupied by these structures in

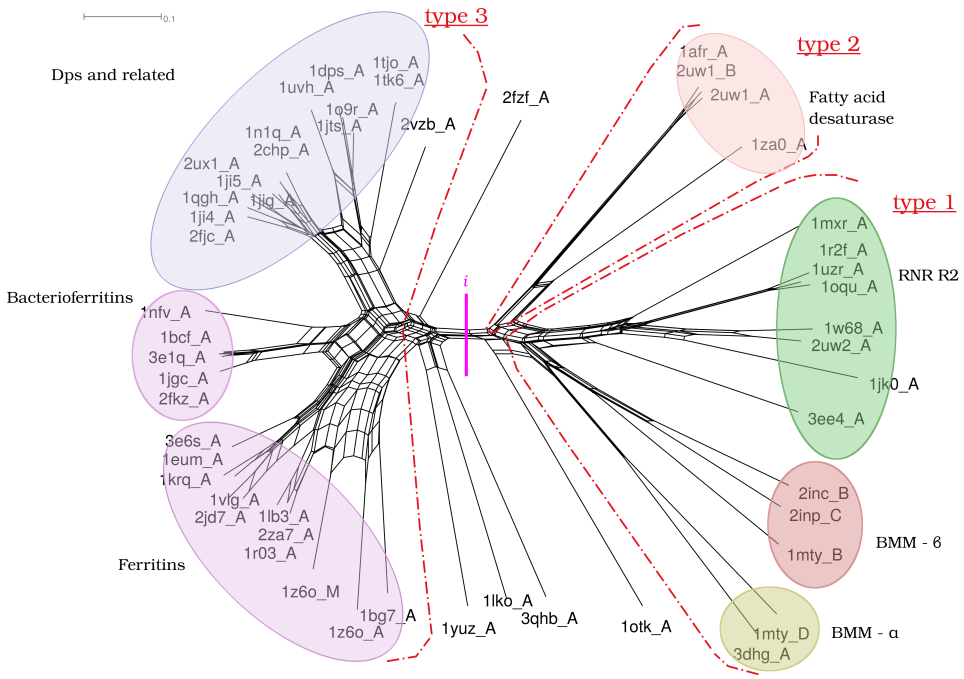


Figure 5.4: The neighbor-net structural network of the ferritin-like protein superfamily comprising 53 structures. The red dot-dashed arcs separate the structures with three different dimerization types, which were used to assess the quality of the phylogenetic tree by Lundin et al. [10]. The vertical line "*i*" marks the broad split between two SCOP families, ferritins (a.25.1.1) and ribonucleotide reductase-like (a.25.1.2). The colour-coded ellipses are consistent with the previous study [10] and labelled with annotations provided by RCSB. The scale bar represents distance.

Figure 5.5.

Two possible explanations can be presented for this poor support. Firstly, Figure 5.6, illustrates strong structural conservation which in this case obfuscates the relationships between structures with nearly identical pairwise-compared distance values, see Chapter 2 for a detailed discussion on the MD-based bootstrap method. Secondly, the reference tree in Figure 5.5 showing the support values may not be entirely accurate. While the neighbour-joining algorithm, often, recovers a tree similar to the true tree, it is known to struggle with increasing number of entities compared [97], implying that the inherent limitations of the neighbor-joining algorithm may limit the accuracy of the tree generated in this case. Nonetheless, strong support, N1 = 100, ratifies the conclusion that these three groups share a common ancestor.

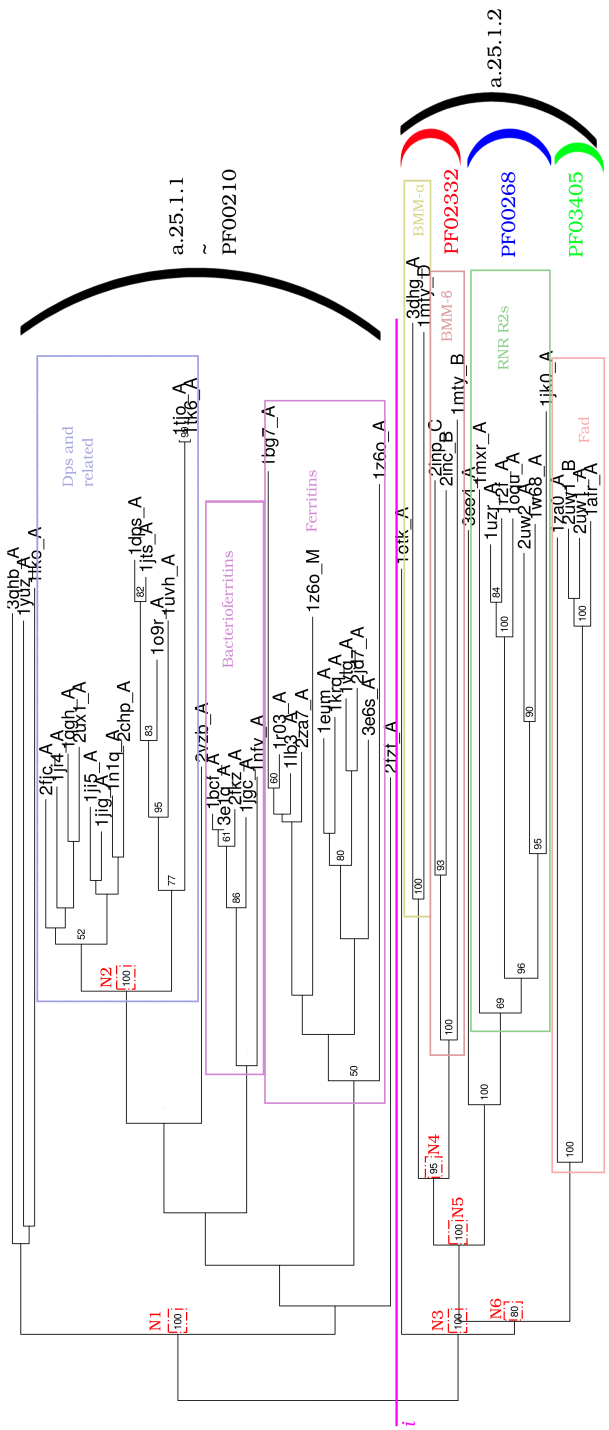


Figure 5.5: Structural phylogeny of the ferritin-like superfamily with statistical support. The colour of the boxes is consistent with Figure 2 in the previous work by Lundin et al. [10]. Only support values above 50% are shown. See text for significance of nodes N1-N6. Broad recovery of protein structures as classified by SCOP and PFam are indicated by arcs. "i" denotes a divide (pink horizontal line) between the two SCOP families, ferritins (a.25.1.1) and ribonucleotide reductase-like (a.25.1.2), as in Figure 5.4. The scale bar represents distance.

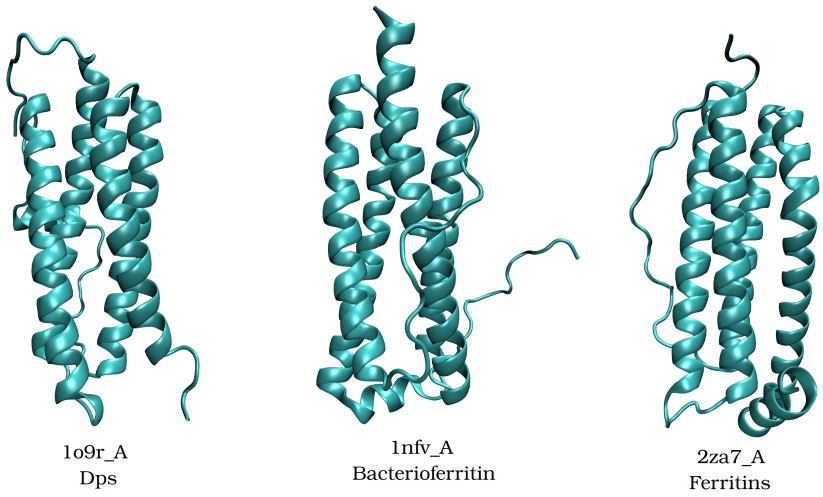


Figure 5.6: Conserved structural core in representative structures from ferritins, bacterioferritins and Dps. The conservation is likely to produce near-identical or overlapping distances as the structures sample conformational space during MD simulations, resulting in poor support for the tree in Figure 5.5, see Appendix-I Figure 2 for details.

Fads share a common ancestor with RNR R2s and BMMs.

There exists strong support, N6 = 80, for a shared ancestor between Fads, RNR R2s and BMMs. Moreover, the clear separation of the ferritin/bacterioferritin/Dps and Fads/RNR R2/BMM groupings, as reflected in their strong support, i.e. N1 and N3, validates the emergence of Fads, RNR R2s and BMMs from the others by, possibly, the gain of the substrate oxidation function [10].

BMMs duplicated into BMM-α and BMM-β following divergence from the common ancestor with RNR R2.

The duplication and divergence of BMMs into α and β groups is supported by strong, N4 = 95, evidence. While this node, N4, indicates that one or a few structures departed from this node, indicated by the 5% loss of support, strong support on the outer node, N5 = 100, indicates that all the structures originally occupying this node were present on it during the replicate trial. Therefore strong support, N5, exists for the BMMs and RNR R2s sharing a common ancestor and for the duplication and divergence of

BMMs, N4, into BMM-α and BMM-β.

5.4 Discussion

The primary focus of this work was to implement the MD-based bootstrap method in a first of a kind approach to determine structural phylogenies. The method was implemented on the ferritin-like protein superfamily. Support from the MD-based bootstrap method was used to assess the correctness of inferences made in a previous study which only used qualitative information to assess the phylogenetic relationships. While the inferences made in that study were supported by the results from the MD-based bootstrap method, several aspects of this method require further consideration. These are generation of a reference tree, resolving relationships in protein databases and sampling conformations through MD simulations. These are briefly discussed below.

Generating a reference tree

Two common ways of representing evolutionary information are through the use of either a reticulated network, Figure 5.4 or a phylogenetic tree, Figure 5.2. The tree-likeness of a network imparts a coherent evolutionary tale to the structures analysed. Departure from tree-likeness indicates that the information in the distance data is not sufficient for a coherent interpretation.

In this analysis, this problem impacts the relationships between Dps, bacterioferritins and ferritin proteins. The network in Figure 5.4 shows departure from tree-likeness. This is also illustrated by poor support for the relationships between these protein families, Figure 5.5. It can be argued that use of a consensus method to summarize the replicate trees may offer a better interpretation of the structural phylogeny. While in this particular case, important relationships between members of the ferritin-like protein superfamily were still able to be ratified using a specific reference tree, the low confidence values ascribed to some nodes of the tree suggest that alternative means of generating a reference tree should be explored.

Organization of protein databases

As discussed earlier, databases that group proteins based on sequences or structures tend to form clusters at each level of the hierarchy. While

the clusters offer some important insight regarding shared ancestry, exact relationships are not determined.

While protein sequence divergence and a missing method of determining structural phylogenies prevent databases from resolving these relationships, the method introduced in this work offers a solution.

In this analysis, the classification of well-characterized proteins, manually curated by SCOP, was successfully recovered. This method therefore creates a way to automate the classification process. Furthermore the structural conservation, as used by SCOP to group proteins into families, was used to resolve polytomies both at the superfamily level in SCOP, uncovering deep evolutionary relationships, and within individual families as well, some of which can be seen strongly supported in Figure 5.5.

Conformational sampling

Conformational sampling using MD simulations is of particular interest, especially in the case of Dps, bacterioferritins and ferritins protein families, Figure 5.6. In this instance, as the structural core between the three families, Figure 5.6, is remarkably similar, the distance data is not sufficient to rigorously classify these protein families into distinct well-formed groups, Figure 5.4. Furthermore, the alternative conformations sampled for the structures in these families are also highly similar to one another. These structural similarities mean that all of the pairwise distances between proteins in these families are of similar magnitude, such that small changes in the distance values change the topology of the replicate trees, thus producing numerous different tree topologies. This results in poor support at the nodes occupied by proteins from these families. Although the structural method delineates relationships for both less diverged and deeply diverged proteins, the MD-based bootstrap method is not sufficiently sensitive to resolve relationships between less diverged proteins, i.e. proteins that are significantly similar, see Chapter 2 for a detailed discussion.

5.5 Conclusion

Relationships between deeply diverging proteins cannot be meaningfully resolved by sequence-based methods. Conservation in structure, as used by the method presented in this work, can be used to fill the gap created by

sequence-based methods, as demonstrated on the ferritin-like protein superfamily. The recovered relationships can inform the unresolved relationships currently present within structural databases. Moreover, the use of a MD-based bootstrap method provides a way in which to assess the reliability of the inferred relationships.

5.6 Future Work

The addition of more structures is known to improve phylogenetic reconstruction. This is one avenue of future work where more structures are added to improve this phylogeny.

Another avenue of future work is to generate longer simulations and evaluate the impact of related and non-related conformations, i.e. with low and high temporal and spatial displacements.

Furthermore as previously noted comparing protein structures that are significantly similar presents a challenge to the MD-based bootstrap method. Further work on this would involve determining support for nodes occupied by significantly similar structures. One way to do this can be excluding some of the structures from the analysis, reducing redundancy. As similar structure can sometimes imply similar sequence, sequence analysis would be sufficient for such cases. One example encountered here was PF00210. Another way would be through extensive sampling using more sophisticated MD methods such as replica exchange.

Bibliography

[1] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., and Sangrador-Vegas, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279 – D285, 2016.

[2] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[3] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[4] Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., and Durbin, R. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl_1):D247 – D251, 2006.

[5] Johnson, M. S., Sutcliffe, M. J., and Blundell, T. L. Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. *Journal of Molecular Evolution*, 30(1):43–59, 1990.

[6] Bujnicki, J. M. Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures. *Journal of Molecular Evolution*, 50(1):39–44, 2000.

[7] Holm, L. and Sander, C. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20(11):478–480, 1995.

[8] Breitling, R., Laubner, D., and Adamski, J. Structure-based phylogenetic analysis of short-chain alcohol dehydrogenases and reclassification of the 17β-hydroxysteroid dehydrogenase family. *Molecular Biology and Evolution*, 18(12):2154–2161, 2001.

[9] Krissinel, E. and Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12): 2256–2268, 2004.

[10] Lundin, D., Poole, A. M., Sjöberg, B.-M., and Högbom, M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *Journal of Biological Chemistry*, 287(24):20565–20575, 2012.

[11] Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[12] Talevich, E., Invergo, B. M., Cock, P. J. A., and Chapman, B. A. Bio. Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13(1):209, 2012.

[13] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., and Wilczynski, B. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[14] Huson, D. H. and Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[15] Cobessi, D., Huang, L. S., Ban, M., Pon, N. G., Daldal, F., and Berry, E. A. The 2.6 Å resolution structure of *Rhodobacter capsulatus* bacterioferritin with metal-free dinuclear site and heme iron in a crystallographic "special position". *Acta crystallographica. Section D, Biological crystallography*, 58(Pt 1):29–38, 2002.

[16] Andersson, M. E., Högbom, M., Rinaldo-Matthis, A., Blodig, W., Liang, Y., Persson, B. O., Sjöberg, B. M., Su, X. D., and Nordlund, P.

Structural and mutational studies of the carboxylate cluster in iron-free ribonucleotide reductase R2. *Biochemistry*, 43(24):7966–7972, 2004.

[17] Eriksson, M., Jordan, A., and Eklund, H. Structure of *Salmonella typhimurium* nrdF ribonucleotide reductase in its oxidized and reduced forms. *Biochemistry*, 37(38):13359–13369, 1998.

[18] Gauss, G. H., Benas, P., Wiedenheft, B., Young, M., Douglas, T., and Lawrence, C. M. Structure of the DPS-like protein from *Sulfolobus solfataricus* reveals a bacterioferritin-like dimetal binding site within a DPS-like dodecameric assembly. *Biochemistry*, 45(36):10815–10827, 2006.

[19] Kim, J., Malashkevich, V., Roday, S., Lisbin, M., Schramm, V. L., and Almo, S. C. Structural and kinetic characterization of *Escherichia coli* TadA, the wobble-Specific tRNA deaminase. *Biochemistry*, 45(20): 6407–6416, 2006.

[20] Lawson, T. L., Crow, A., Lewin, A., Yasmin, S., Moore, G. R., and Le Brun, N. E. Monitoring the iron status of the ferroxidase center of *Escherichia coli* bacterioferritin using fluorescence spectroscopy. *Biochemistry*, 48(38):9031–9039, 2009.

[21] Sazinsky, M. H., Dunten, P. W., McCormick, M. S., DiDonato, A., and Lippard, S. J. X-ray structure of a hydroxylase-regulatory protein complex from a hydrocarbon-oxidizing multicomponent monooxygenase, *Pseudomonas sp.* OX1 phenol hydroxylase. *Biochemistry*, 45 (51):15392–15404, 2006.

[22] Swartz, L., Kuchinskas, M., Li, H., Poulos, T. L., and Lanzilotta, W. N. Redox-dependent structural changes in the *Azotobacter vinelandii* bacterioferritin: New insights into the ferroxidase and iron transport mechanism. *Biochemistry*, 45(14):4421–4428, 2006.

[23] Whittaker, M. M., Barynin, V. V., Igarashi, T., and Whittaker, J. W. Outer sphere mutagenesis of *Lactobacillus plantarum* manganese catalase disrupts the cluster core. Mechanistic implications. *European Journal of Biochemistry*, 270(6):1102–16, 2003.

[24] Högbom, M. and Nordlund, P. A protein carboxylate coordinated oxo-centered tri-nuclear iron complex with possible implications for ferritin mineralization. *FEBS Letters*, 567(2-3):179–182, 2004.

[25] Uppsten, M., Davis, J., Rubin, H., and Uhlin, U. Crystal structure of the biologically active form of class Ib ribonucleotide reductase small subunit from *Mycobacterium tuberculosis*. *FEBS Letters*, 569(1-3):117–122, 2004.

[26] Gauss, G. H., Reott, M. A., Roha, E. R., Young, M. J., Douglas, T., Smith, C. J., and Lawrence, C. M. Characterization of the *Bacteroides fragilis* bfr gene product identifies a bacterial DPS-like protein and suggests evolutionary links in the ferritin superfamily. *Journal of Bacteriology*, 194(1):15–27, 2012.

[27] Yoshizawa, K., Mishima, Y., Park, S. Y., Heddle, J. G., Tame, J. R. H., Iwahori, K., Kobayashi, M., and Yamashita, I. Effect of N-terminal residues on the structural stability of recombinant horse L-chain apoferritin in an acidic environment. *Journal of Biochemistry*, 142(6):707–713, 2007.

[28] Ceci, P., Ilari, A., Falvo, E., and Chiancone, E. The Dps protein of *Agrobacterium tumefaciens* does not bind to DNA but protects it toward oxidative cleavage. X-ray crystal structure, iron binding, and hydroxyl-radical scavenging properties. *Journal of Biological Chemistry*, 278(22):20319–20326, 2003.

[29] Ceci, P., Ilari, A., Falvo, E., Giangiacomo, L., and Chiancone, E. Reassessment of protein stability, DNA binding, and protection of *Mycobacterium smegmatis* Dps. *Journal of Biological Chemistry*, 280(41): 34776–34785, 2005.

[30] Guy, J. E., Whittle, E., Kumaran, D., Lindqvist, Y., and Shanklin, J. The crystal structure of the Ivy δ4-16:0-ACP desaturase reveals structural details of the oxidized active site and potential determinants of regioselectivity. *Journal of Biological Chemistry*, 282(27):19863–19871, 2007.

[31] Moche, M., Shanklin, J., Ghoshal, A., and Lindqvist, Y. Azide and acetate complexes plus two iron-depleted crystal structures of the di-

iron enzyme $\delta 9$ stearoyl-acyl carrier protein desaturase: Implications for oxygen activation and catalytic intermediates. *Journal of Biological Chemistry*, 278(27):25072–25080, 2003.

[32] Papinutto, E., Dundon, W. G., Pitulis, N., Battistutta, R., Montecucco, C., and Zanotti, G. Structure of two iron-binding proteins from *Bacillus anthracis*. *Journal of Biological Chemistry*, 277(17):15093–15098, 2002.

[33] Strand, K. R., Karlsen, S., Kolberg, M., Røhr, Å. K., Görbitz, C. H., and Andersson, K. K. Crystal structural studies of changes in the native dinuclear iron center of ribonucleotide reductase protein R2 from mouse. *Journal of Biological Chemistry*, 279(45):46794–46801, 2004.

[34] Granier, T., D'Estaintot, L. B., Gallois, B., Chevalier, J. M., Précigoux, G., Santambrogio, P., and Arosio, P. Structural description of the active sites of mouse L-chain ferritin at 1.2 Å resolution. *Journal of Biological Inorganic Chemistry*, 8(1-2):105–111, 2003.

[35] Iyer, R. B., Silaghi-Dumitrescu, R., Kurtz, D. M., and Lanzilotta, W. N. High-resolution crystal structures of *Desulfovibrio vulgaris* (Hildenborough) nigerythrin: Facile, redox-dependent iron movement, domain interface variability, and peroxidase activity in the rubrerythrins. *Journal of Biological Inorganic Chemistry*, 10(4):407–416, 2005.

[36] Tatur, J., Hagen, W. R., and Matias, P. M. Crystal structure of the ferritin from the hyperthermophilic archaeal anaerobe *Pyrococcus furiosus*. *Journal of Biological Inorganic Chemistry*, 12(5):615–630, 2007.

[37] Stillman, T., Hempstead, P., Artymiuk, P., Andrews, S., Hudson, A., Treffry, A., Guest, J., and Harrison, P. The high-resolution X-ray crystallographic structure of the ferritin (EcFtnA) of *Escherichia coli*; comparison with human H ferritin (HuHF) and the structures of the Fe3+ and Zn2+ derivatives. *Journal of Molecular Biology*, 307(2):587–603, 2001.

[38] D'Estaintot, B. L., Santambrogio, P., Granier, T., Gallois, B., Chevalier, J. M., Précigoux, G., Levi, S., and Arosio, P. Crystal structure and biochemical properties of the human mitochondrial ferritin and its mutant Ser144Ala. *Journal of Molecular Biology*, 340(2):277–293, 2004.

[39] Hamburger, A. E., West, A. P., Hamburger, Z. A., Hamburger, P., and Bjorkman, P. J. Crystal structure of a secreted insect ferritin reveals a symmetrical arrangement of heavy and light chains. *Journal of Molecular Biology*, 349(3):558–569, 2005.

[40] Ren, B., Tibbelin, G., Kajino, T., Asami, O., and Ladenstein, R. The multi-layered structure of Dps with a novel di-nuclear ferroxidase center. *Journal of Molecular Biology*, 329(3):467–77, 2003.

[41] Trikha, J., Theil, E. C., and Allewell, N. M. High resolution crystal structures of amphibian red-cell L ferritin: potential roles of structural plasticity and solvation in function. *Journal of Molecular Biology*, 248 (5):949–967, 1995.

[42] Zanotti, G., Papinutto, E., Dundon, W., Battistutta, R., Seveso, M., Giudice, G., Rappuoli, R., and Montecucco, C. Structure of the neutrophil-activating protein from *Helicobacter pylori*. *Journal of Molecular Biology*, 323(1):125–30, 2002.

[43] Jin, S., Kurtz, D. M., Liu, Z.-J., Rose, J., and Wang, B.-C. X-ray crystal structures of reduced rubrerythrin and its azide adduct: a structure-based mechanism for a non-heme diiron peroxidase. *Journal of the American Chemical Society*, 124(33):9845–55, 2002.

[44] McCormick, M. S., Sazinsky, M. H., Condon, K. L., and Lippard, S. J. X-ray crystal structures of manganese(II)-reconstituted and native toluene/o-xylene monooxygenase hydroxylase reveal rotamer shifts in conserved residues and an enhanced view of the protein interior. *Journal of the American Chemical Society*, 128(47):15108–15110, 2006.

[45] Sazinsky, M. H. and Lippard, S. J. Product bound structures of the soluble methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): Protein motion in the α-subunit. *Journal of the American Chemical Society*, 127(16):5814–5825, 2005.

[46] Schönafinger, A., Morbitzer, A., Kress, D., Essen, L. O., Noll, F., and Hampp, N. Morphology of dry solid-supported protein monolayers dependent on the substrate and protein surface properties. *Langmuir*, 22 (17):7185–7191, 2006.

[47] Marchetti, A., Parker, M. S., Moccia, L. P., Lin, E. O., Arrieta, A. L., Ribalet, F., Murphy, M. E. P., Maldonado, M. T., and Armbrust, E. V. Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature*, 457(7228):467–470, 2009.

[48] Frolow, F., Kalb, A. J., and Yariv, J. Structure of a unique twofold symmetric haem-binding site. *Nature Structural Biology*, 1(7):453–60, 1994.

[49] Grant, R. A., Filman, D. J., Finkel, S. E., Kolter, R., and Hogle, J. M. The crystal structure of Dps, a ferritin homolog that binds and protects DNA. *Nature Structural Biology*, 5(4):294–303, 1998.

[50] Ilari, A., Stefanini, S., Chiancone, E., and Tsernoglou, D. The dodecameric ferritin from *Listeria innocua* contains a novel intersubunit iron-binding site. *Nature Structural Biology*, 7(1):38–43, 2000.

[51] Macedo, S., Romão, C. V., Mitchell, E., Matias, P. M., Liu, M. Y., Xavier, A. V., LeGall, J., Teixeira, M., Lindley, P., and Carrondo, M. A. The nature of the di-iron site in the bacterioferritin from *Desulfovibrio desulfuricans*. *Nature Structural Biology*, 10(4):285–90, 2003.

[52] Zeth, K., Offermann, S., Essen, L.-O., and Oesterhelt, D. Iron-oxo clusters biomineralizing on protein surfaces: Structural analysis of *Halobacterium salinarum* DpsA in its low- and high-iron states. *Proceedings of the National Academy of Sciences*, 101(38):13780–13785, 2004.

[53] Voegtli, W. C., Ge, J., Perlstein, D. L., Stubbe, J., and Rosenzweig, A. C. Structure of the yeast ribonucleotide reductase Y2Y4 heterodimer. *Proceedings of the National Academy of Sciences*, 98(18): 10073–8, 2001.

[54] Andersson, C. S. and Högbom, M. A *Mycobacterium tuberculosis* ligand-binding Mn/Fe protein reveals a new cofactor in a remodeled R2-protein scaffold. *Proceedings of the National Academy of Sciences*, 106(14):5633–5638, 2009.

[55] Bailey, L. J., McCoy, J. G., Phillips, G. N., and Fox, B. G. Structural consequences of effector protein complex formation in a diiron

hydroxylase. *Proceedings of the National Academy of Sciences*, 105 (49):19194–19198, 2008.

[56] Högbom, M., Galander, M., Andersson, M., Kolberg, M., Hofbauer, W., Lassmann, G., Nordlund, P., and Lendzian, F. Displacement of the tyrosyl radical cofactor in ribonucleotide reductase obtained by single-crystal high-field EPR and 1.4 Å X-ray data. *Proceedings of the National Academy of Sciences*, 100(6):3209–3214, 2003.

[57] Zeth, K., Offermann, S., Essen, L.-O., and Oesterhelt, D. Iron-oxo clusters biomineralizing on protein surfaces: structural analysis of *Halobacterium salinarum* DpsA in its low- and high-iron states. *Proceedings of the National Academy of Sciences*, 101(38):13780–13785, 2004.

[58] Elango, N., Radhakrishnan, R., Froland, W. A., Wallar, B. J., Earhart, C. A., Lipscomb, J. D., and Ohlendorf, D. H. Crystal structure of the hydroxylase component of methane monooxygenase from *Methylosinus trichosporium* OB3b. *Protein Science*, 6(3):556–568, 1997.

[59] Hindupur, A., Liu, D., Zhao, Y., Bellamy, H. D., White, M. A., and Fox, R. O. The crystal structure of the *Escherichia coli* stress protein YciF. *Protein Science*, 15(11):2605 – 2611, 2006.

[60] Dyer, D. H., Lyle, K. S., Rayment, I., and Fox, B. G. X-ray structure of putative acyl-ACP desaturase DesA2 from *Mycobacterium tuberculosis* H37Rv. *Protein Science*, 14(6):1508–1517, 2005.

[61] Havukainen, H., Haataja, S., Kauko, A., Pulliainen, A. T., Salminen, A., Haikarainen, T., Finne, J., and Papageorgiou, A. C. Structural basis of the zinc- and terbium-mediated inhibition of ferroxidase activity in Dps ferritin-like proteins. *Protein Science*, 17(9):1513–21, 2008.

[62] Rosenzweig, A. C., Brandstetter, H., Whittington, D. A., Nordlund, P., Lippard, S. J., and Frederick, C. A. Crystal structures of the methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath): implications for substrate gating and component interactions. *Proteins*, 29(2):141–52, 1997.

[63] Thumiger, A., Polenghi, A., Papinutto, E., Battistutta, R., Montecucco, C., and Zanotti, G. Crystal structure of antigen TpF1 from

Treponema pallidum. *Proteins: Structure, Function and Genetics*, 62 (3):827–830, 2006.

[64] Hogbom, M. The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass. *Science*, 305(5681):245–248, 2004.

[65] Cooley, R. B., Rhoads, T. W., Arp, D. J., and Karplus, P. A. A diiron protein autogenerates a valine-phenylalanine cross-link. *Science*, 332 (6032):929–929, 2011.

[66] Johnson, E., Cascio, D., Sawaya, M. R., Gingery, M., and Schröder, I. Crystal structures of a tetrahedral open pore ferritin from the hyperthermophilic archaeon *Archaeoglobus fulgidus*. *Structure*, 13(4):637–648, 2005.

[67] Lindqvist, Y., Huang, W., Schneider, G., and Shanklin, J. Crystal structure of $\delta 9$ stearoyl-acyl carrier protein desaturase from castor seed and its relationship to other di-iron proteins. *The EMBO Journal*, 15 (16):4081–92, 1996.

[68] Takagi, H., Shi, D., Ha, Y., Allewell, N. M., and Theil, E. C. Localized unfolding at the junction of three ferritin subunits. A mechanism for iron release? *The Journal of Biological Chemistry*, 273(30):18685–8, 1998.

[69] Joint Center for Structural Genomics. PDB ID: 2oh3, Crystal structure of COG1633: Uncharacterized conserved protein (ZP_00055496.1) from *Magnetospirillum magnetotacticum* MS-1 at 2.00 Å resolution. 2007.

[70] Joint Center for Structural Genomics. PDB ID: 3ez0, Crystal structure of NTF2-like protein of unknown function (YP_270605.1) from *Colwellia psychrerythraea* 34H at 1.61 Å resolution. 2010.

[71] Joint Center for Structural Genomics. PDB ID: 2itb, Crystal structure of putative tRNA-(ms(2)io(6)a)-hydroxylase (NP_744337.1) from *Pseudomonas Putida* KT2440 at 2.05 Å resolution. 2006.

[72] Joint Center for Structural Genomics. PDB ID: 3fse, Crystal structure of two-domain protein containing DJ-1/ThiJ/PfpI-like and ferritin-like domains. (YP_324989.1) from *Anabaena variabilis* ATCC 29413 at 1.90 Å resolution. 2009.

[73] Joint Center for Structural Genomics. PDB ID: 2rcc, Crystal structure of putative class I ribonucleotide reductase (NP_241368.1) from *Bacillus halodurans* at 1.90 Å resolution, 2007.

[74] Luo, J., Liu, D., White, M., and Fox, R. PDB ID: 1jts, DNA protection and binding by *Escherichia coli* DPS protein. 2003.

[75] Ebihara, A., Yokoyama, S., and Kuramitsu, S. PDB ID: 2cwl, Structural and functional analysis of pseudocatalase from *Thermus thermophilus* HB8. 2005.

[76] Kim, Y., Joachimiak, G., Wu, R., Patterson, S., Gornicki, P., and Joachimiak, A. PDB ID: 2qqy, Crystal structure of ferritin-like, diiron-carboxylate proteins from *Bacillus anthracis* str. Ames. 2007.

[77] Osipiuk, J., Evdokimova, E., Kudritska, M., Savchenko, A., Edwards, A., and Joachimiak, A. PDB ID: 2gyq, X-ray crystal structure of YcfI protein, a putative structural protein from *Rhodopseudomonas palustris*. 2006.

[78] Ramagopal, U., Rutter, M., Adams, J., Toro, R., Groshong, C., Sauder, J., Burley, S., and Almo, S. PDB ID: 2qf9, Structure of putative secreted protein DUF305 from *Streptomyces coelicolor*. 2007.

[79] Yu, M., Bursey, E., Radhakannan, T., Kim, C., Kaviratne, T., Woodruff, T., Segelke, B., Lekin, T., Toppani, D., Terwilliger, T., and Hung, L. PDB ID: 2ib0, Crystal structure of a conserved hypothetical protein, rv2844, from *Mycobacterium tuberculosis*. 2006.

[80] Zhang, R., Joachimiak, A., Edwards, A., Savchenko, A., and Skarina, T. PDB ID: 1otk, The 2 A crystal structure of protein paaC from *Escherichia coli*. 2003.

[81] Seattle Structural Genomics Center for Infectious Disease. PDB ID: 3ix6, Crystal structure of thymidylate synthase thyA from *Brucella melitensis*. 2009.

[82] Fu, Z.-Q., Liu, Z.-J., Lee, D., Kelley, L., Chen, L., Tempel, W., Shah, N., Horanyi, P., Lee, H., Habel, J., Dillard, B., Nguyen, D., Chang, S.-H., Zhang, H., Chang, J., Sugar, F., Poole, F., Jr, J., F.E, Adams,

M., Rose, J., and Wang, B.-C. PDB ID: 2fzf, Hypothetical protein Pfu-1136390-001 From *Pyrococcus furiosus*. 2006.

[83] Hortolan, L., Saintout, N., Granier, G., Langlois d'Estaintot, B., Manigand, C., Mizunoe, Y., Wai, S., Gallois, B., and Precigoux, G. PDB ID: 1krq, Crystal structure analysis of *Campylobacter jejuni* ferritin. 2002.

[84] Joint Center for Structural Genomics. PDB ID: 4h0a, Crystal structure of a hypothetical protein (SAV1118) from *Staphylococcus aureus* subsp. aureus Mu50 at 1.90 Å resolution. 2012.

[85] Joint Center for Structural Genomics. PDB ID: 1vlg, Crystal structure of Transaldolase (EC 2.2.1.2) (TM0295) from *Thermotoga maritima* at 2.40 Å resolution. 2004.

[86] Welin, M., Ogg, D., Arrowsmith, C., Berglund, H., Busam, R., Collins, R., Edwards, A., Ehn, M., Flodin, S., Flores, A., Graslund, S., Hammarstrom, M., Hallberg, B., Holmberg Schiavone, L., Hogbom, M., Kotenyova, T., Magnusdottir, A., Moche, M., Nilsson-Ehle, P., Nyman, T., Persson, C., Sagemark, J., Sundstrom, M., Stenmark, P., Uppenberg, J., Thorsell, A., Van Den Berg, S., Wallden, K., Weigelt, J., and Norlund, P. PDB ID: 2uw2, Crystal structure of human ribonucleotide reductase subunit R2. 2007.

[87] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[88] Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.

[89] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[90] Darden, T., York, D., and Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[91] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[92] Humphrey, W., Dalke, A., and Schulten, K. Vmd: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[93] Sukumaran, J. and Holder, M. T. SumTrees: phylogenetic tree summarization. 4.0. 0. *available at: https://github.com/jeetsukumaran/Dendrophy*, 2015.

[94] Sukumaran, J. and Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

[95] Rambaut, A. FigTree v1. 4. *Molecular Evolution, Phylogenetics and Epidemiology*, 2012.

[96] Grossfield, A. and Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual Reports in Computational Chemistry*, 5:23–48, 2009.

[97] Strimmer, K. and von Haeseler, A. Accuracy of neighbor joining for n-taxon trees. *Systematic Biology*, 45(4):516–523, 1996.

Chapter 6

Summary

The work included in this thesis illustrates that protein structures can be used to uncover evolutionary signals which sometimes escape sequence-based methods due to notable divergence in compared sequences. While structural phylogenetics has been around for a number of years, absence of a method to gauge the robustness of evolutionary relationships inferred from the structural comparisons impeded its more widespread usage. Previous usage was limited to complimenting the structure-based evolutionary relationships with support from the underlying sequences of the structures compared. The introduction of the novel MD-based bootstrap method utilizes information from protein structure alone using molecular dynamics simulations and therefore completes the structure-based phylogenetics toolkit.

6.1 Method development

Two key methodological aspects of structural phylogenetics as used in this work were the structural comparison metric Q_{score} and the MD-based bootstrap method, the later being the novel addition. Both of these aspects were thoroughly explored through the usage of empirical structural datasets to highlight their strengths and weaknesses. This mechanistic exploration led to the conclusion that while Q_{score} encompasses numerous traits that make it a suitable structural comparison metric, it should be used cautiously when comparing protein structures which are of different sizes, where size is the number of amino acids each of them have.

Another important conclusion formulated was that the Q_{score} metric showed sensitivity to slight changes in the conformation of a structure, meaning that if for a set of structures a relationship is inferred, using slightly different conformations of the structure may yield a completely different evolutionary relationship between them. While this is encouraging in that it will create perturbations in the distance data which is the intended purpose of the MD-based bootstrap, it also serves as a warning, that care needs to go into the selection of structures used for this analysis.

The explorations around the MD-based bootstrap method revealed that the method generated strong statistics when there existed minimal overlap between the sampled conformations and weaker statistics when the opposite was true. This finding led to the conclusion that the method should only be used for structural datasets which have notable structural divergence. While

not established in this work, through conjecture it was concluded that the application of this method should perhaps be limited to protein datasets which are too diverged for analysis with sequence-based methods, as is the intended use of this method.

6.2 Protein structural data

In this work three different structural datasets were explored, namely the aminoacyl-tRNA synthetases, the histone fold and its ancestral precursor and the ferritin-like superfamily. All three of these structural datasets have deep evolutionary origins, however the **aminoacyl-tRNA synthetases** are conserved enough for sequence-based methods to delineate evolutionary relationships between them. This dataset therefore presented an opportunity to compare results from structural analysis to that determined from sequence-based analysis. Recovery of all well-established evolutionary signals in the aminoacyl-tRNA synthetases generated confidence in the ability of Q_{score} to capture evolutionary signals from structural comparisons.

Use of the Q_{score} metric on proteins presenting the **histone fold** or its ancestral precursor led to interesting observations. These included the clear separation between bacterial proteins, which had the ancestral precursor, and the remaining structures. The non-bacterial split showed evidence of multiple points of origin for TATA protein associated factors, core histones and other related proteins. Furthermore broad clusters allowed for development of an evolutionary model. This work therefore uncovered relationships between a set of proteins which would not have been possible using conventional sequence-based methods due to notable divergence between members at a sequence level.

The Q_{score} metric had previously been used to develop a structural phylogeny of the **ferritin-like superfamily** and qualitatively support those inferences. In this work, the MD-based bootstrap method was used to gauge quantitatively the robustness of inferences made previously. The MD-based bootstrap method agreed with all the previous results which had been meticulously derived from the structural network. The close proximity of the ferritins, bacterioferritins and DPS proteins resulted in weaker statistics from the MD-based bootstrap method, which were attributed to structural proximity as previously determined in the method development work.

6.3 Protein databases

At present SCOP and CATH are two prominent protein structural databases which group structures together at certain hierarchical levels based on homology. PFam is a protein sequence database which also groups proteins based on shared evolutionary origin. From amongst the two structural databases, only SCOP was focused on in this work along with PFam.

It was observed in the case of proteins presenting either the histone fold or its ancestral precursor, the PFam and SCOP classifications did not agree. Interestingly, the structural analysis using Q_{score} agreed more with PFam's classification rather than with SCOP's. In the case of the ferritin-like superfamily, while the SCOP and PFam classifications agreed in the ferritin protein family, they disagreed on the classification of the ribonucleotide reductase-like protein family.

Moreover, while databases tend to group proteins at different structural levels, they do not resolve relationships between proteins clustered in a group, resulting in polytomies. The use of structural phylogenetics can resolve these polytomies and associate a measure of significance with the relationships uncovered, as was shown in the work done on the ferritin-like superfamily. Furthermore, as highlighted in cases where SCOP classification broke down, use of Q_{score} and the MD-based bootstrap method can offer a more robust way to classify proteins based on their structures.

6.4 Future directions

Protein structural comparisons using Q_{score}, in this work, has been shown to detect evolutionary signals. The addition of the novel MD-based bootstrap method to add significance to evolutionary relationships determined from structural comparisons equips structural phylogenetics to address a number of important questions. Conventional sequence-based methods sometimes struggle with these deeply diverging protein datasets, however, as shown in this work, protein structure can be used to capture the evolutionary signal that escapes sequence-based methods.

While improving the overall organization of protein structural databases to offer evolutionary relationships is one important issue which can be addressed, another could be related to individual families. For instance, viral

capsid proteins incorporate a conserved structural fold, namely the Jelly-roll fold. The Jellyroll fold is found conserved in 17 viral families including *Picornaviridae*. Viruses from this *Picornaviridae* family cause paralysis, meningitis, hepatitis and poliomyelitis to name a few. The capsid protein is a very important part of the virus as it forms the outer coat the virus. Developing a better understanding of the evolution of viruses may help answer questions like: Did viruses evolve alongside the organisms they infect or do they have their own separate lineage? This question has currently not been addressed. Moreover viruses are currently classified based on their genetic content. Understanding the evolutionary relatedness of these viruses may lead to answers regarding their origin and provide a better classification regime and perhaps even provide indications which may prove useful in medical interventions.

Structural relationships between protein families conserved in archaea, bacteria and eukarya can also be used to find evidence in favour of or against the Woese three-domain classification as discussed in detail in the work on aminoacyl-tRNA synthetases.

Apart from the application of the method to protein families, the method itself can be enhanced. While a number of ways have been discussed already, one possible enhancement could be the use of Monte Carlo based simulations of proteins instead of MD to generate alternative conformations. While this may not offer an enhancement in recovery of evolutionary relationships, it may certainly speed up the acquisition of alternative conformations.

The addition of the novel MD-based bootstrap method to the area of structural phylogenetics has enabled it to effectively address questions related to deeply diverged proteins using only their respective structures.

Appendices

Appendix - I

Miscellaneous

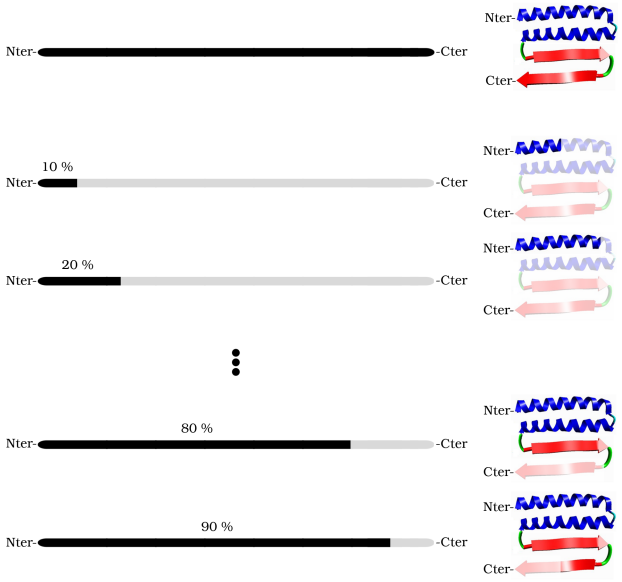


Figure 1: A diagramatic illustration of the protein structure fractions. The left column shows the primary structure and its respective fractions (i.e. 10%, 20% etc.), whereas the right shows the tertiary structure corresponding to those fractions. The fractions are grown from the N-terminal end. The 100% fraction corresponds to the entire structure shown at the top of the figure.

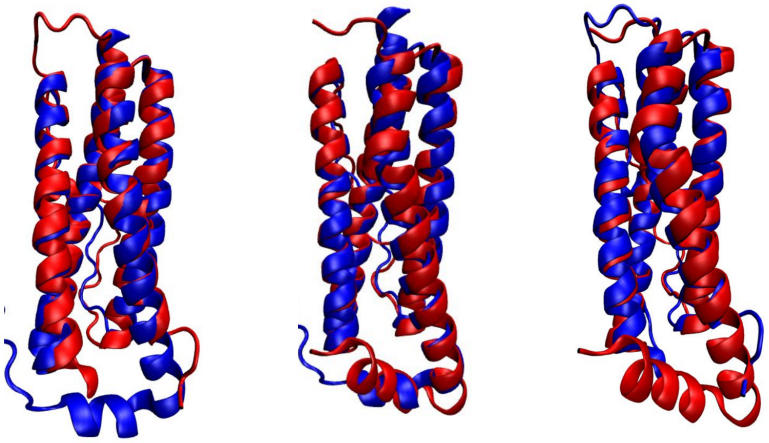


Figure 2: Pairwise comparison between conserved structural cores in representative structures from ferritins, bacterioferritins and Dps, as shown in Figure 5.6. For the comparison on the left, 1nfv_A (red) and 1o9r_A (blue), the Q_{score} is 0.54 with an RMSD of 1.4 Å over 136 aligned residues. The comparison in the middle is between 1nfv_A (red) and 2za7_A (blue), the Q_{score} for which is 0.53 with an RMSD 1.92 Å over 147 aligned residues. The comparison on the right is between 1o9r_A (red) and 2za7_A (blue), for which the Q_{score} is 0.44 with an RMSD 2.07 Å over 134 aligned residues.

Appendix - II

Cytochromes

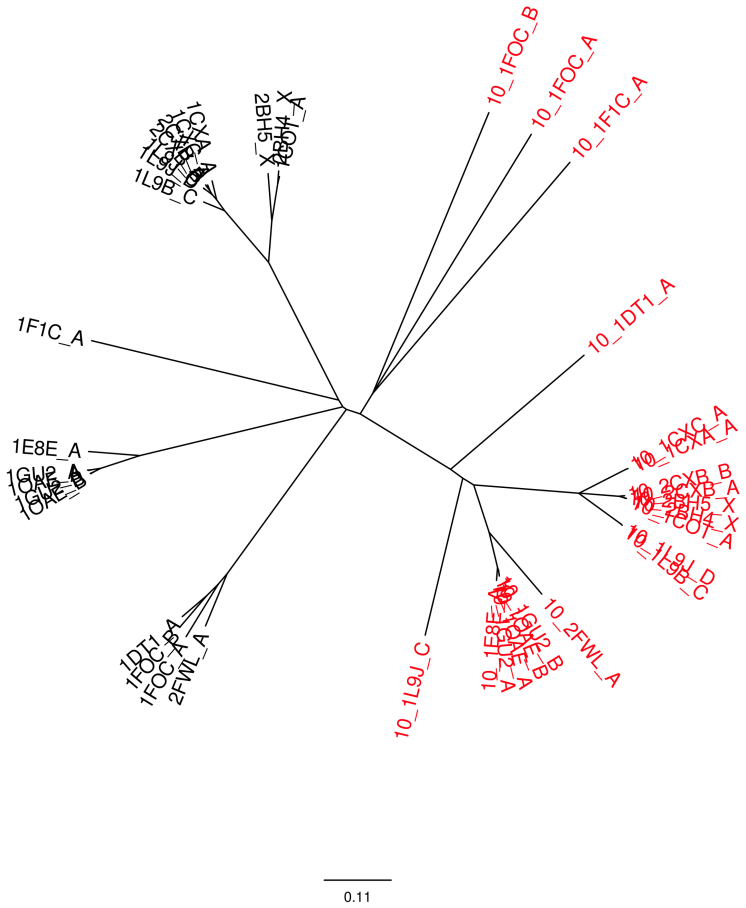


Figure C.1: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (10%), and the complete structures in black.

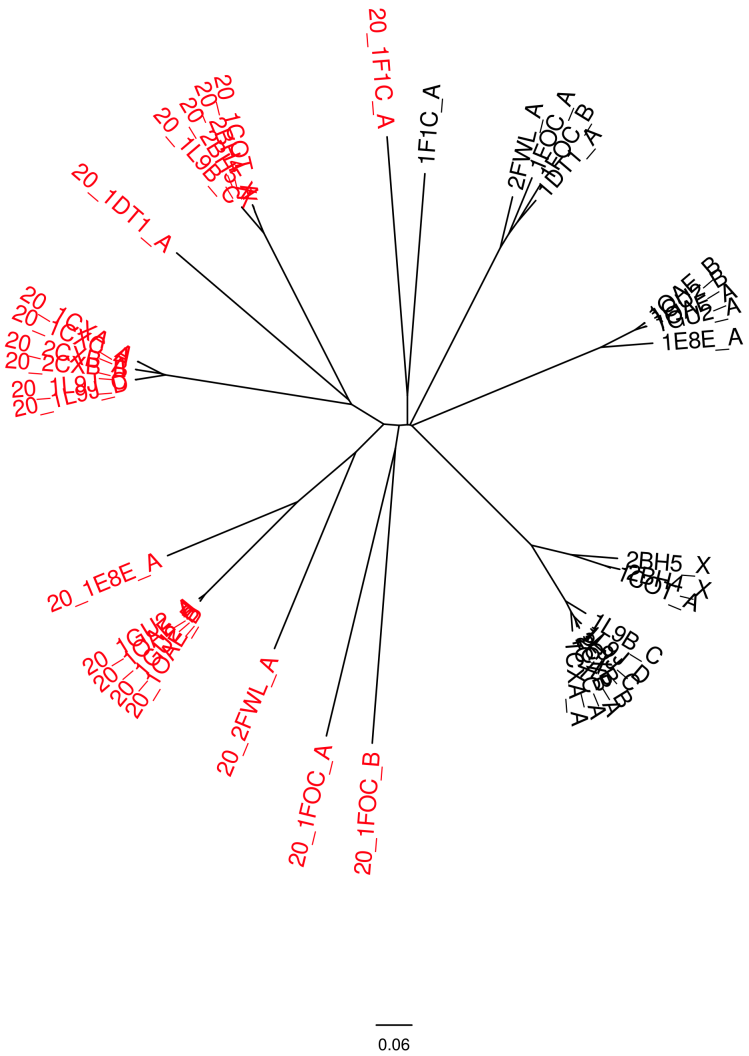


Figure C.2: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (20%), and the complete structures in black.

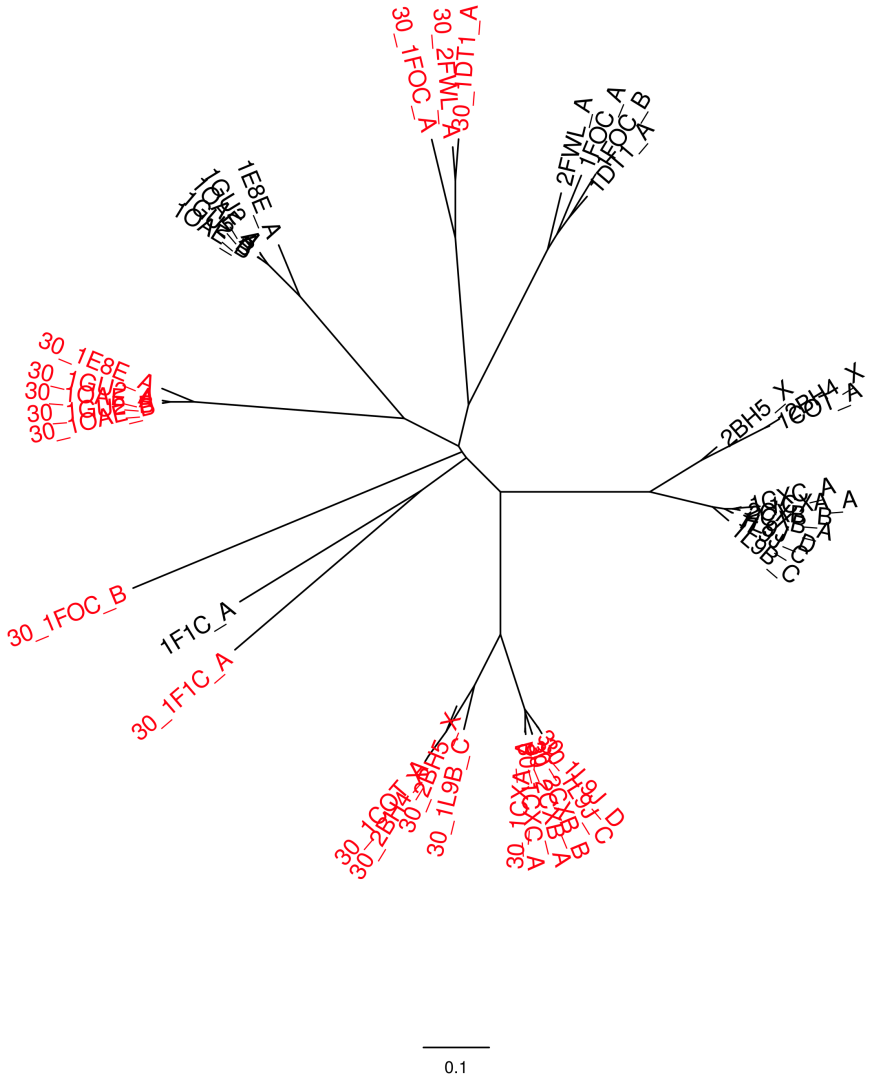


Figure C.3: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (30%), and the complete structures in black.

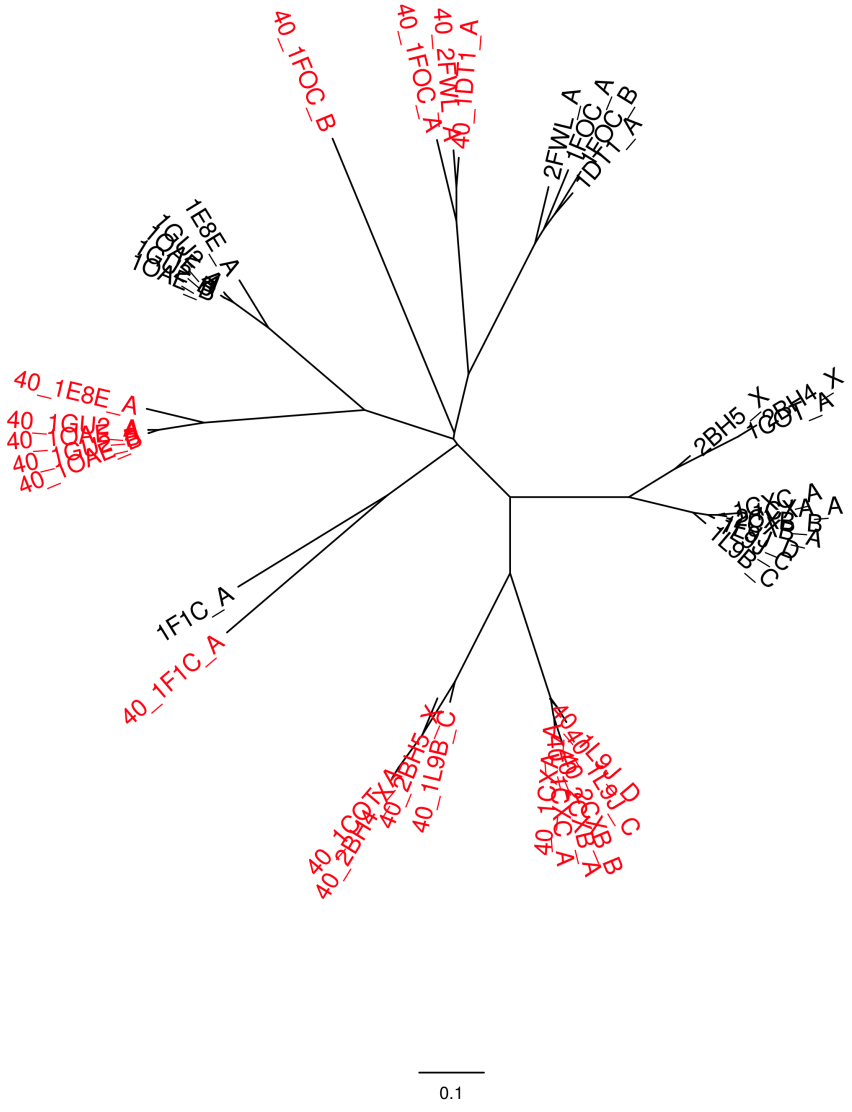


Figure C.4: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (40%), and the complete structures in black.

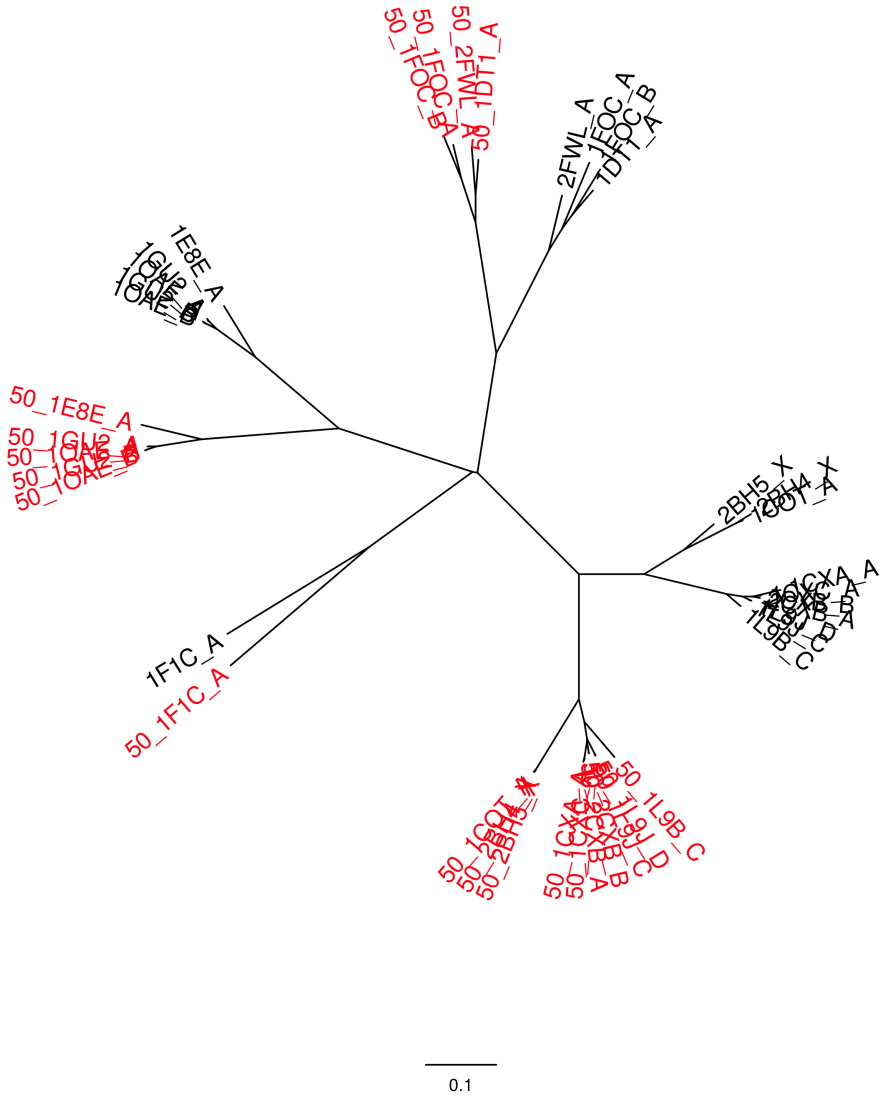


Figure C.5: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (50%), and the complete structures in black.

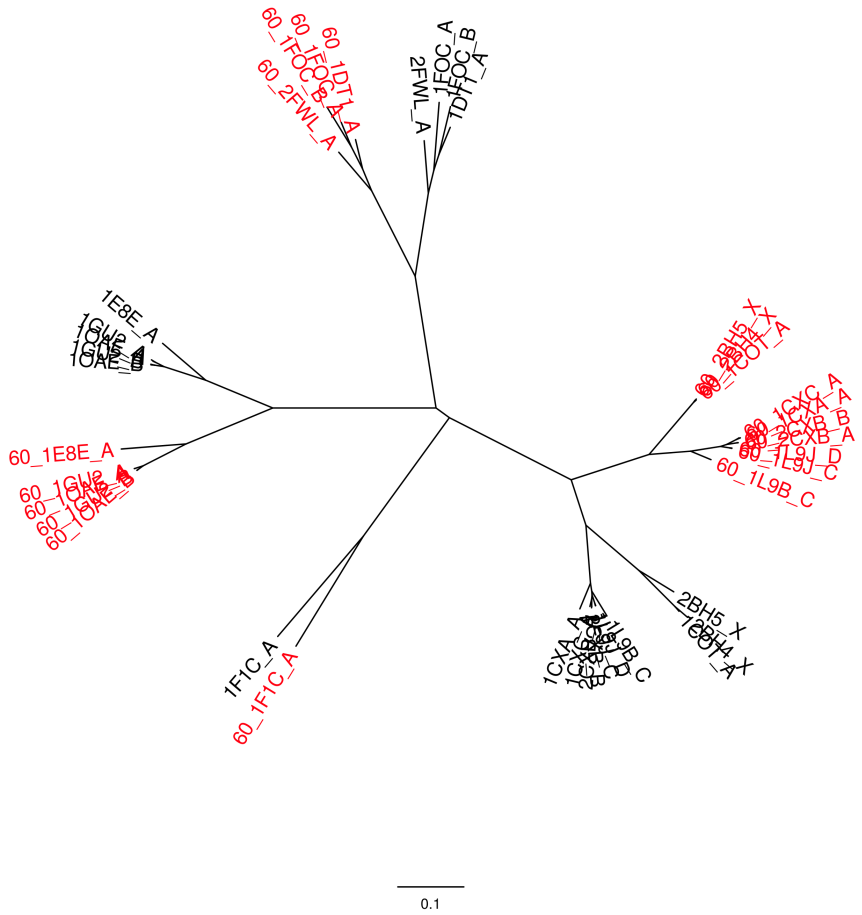


Figure C.6: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (60%), and the complete structures in black.

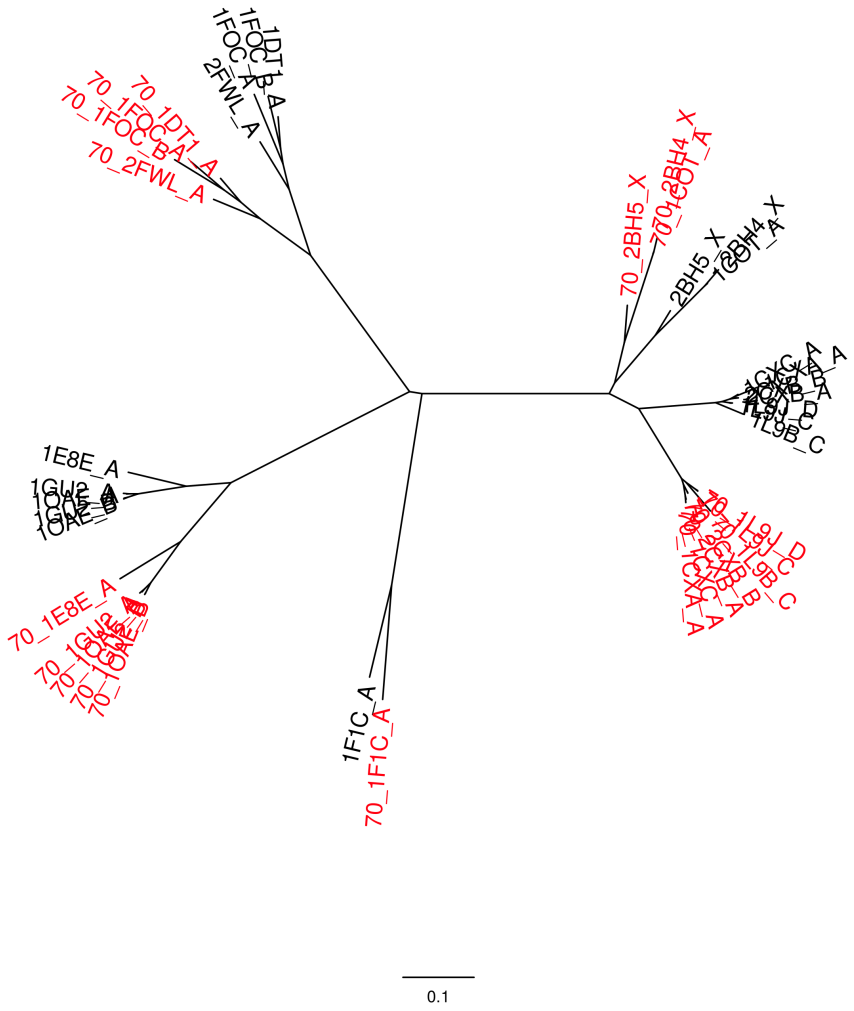


Figure C.7: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (70%), and the complete structures in black.

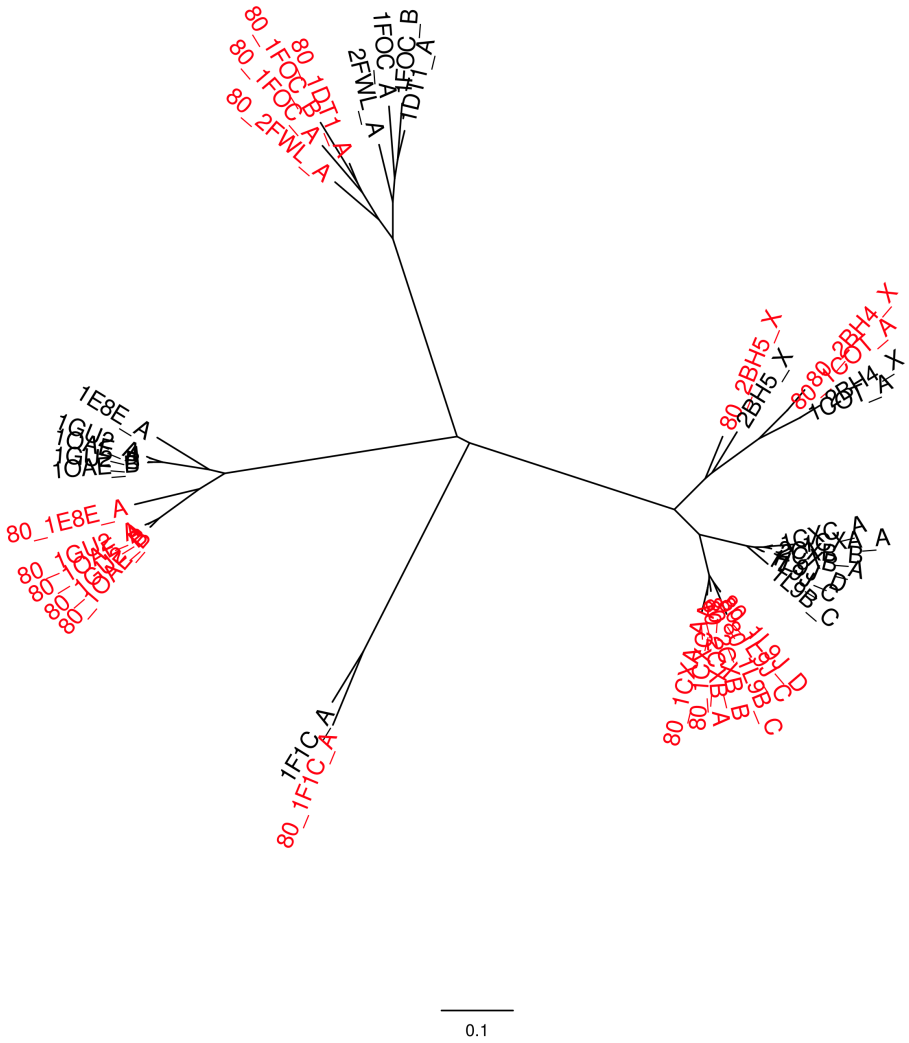


Figure C.8: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (80%), and the complete structures in black.

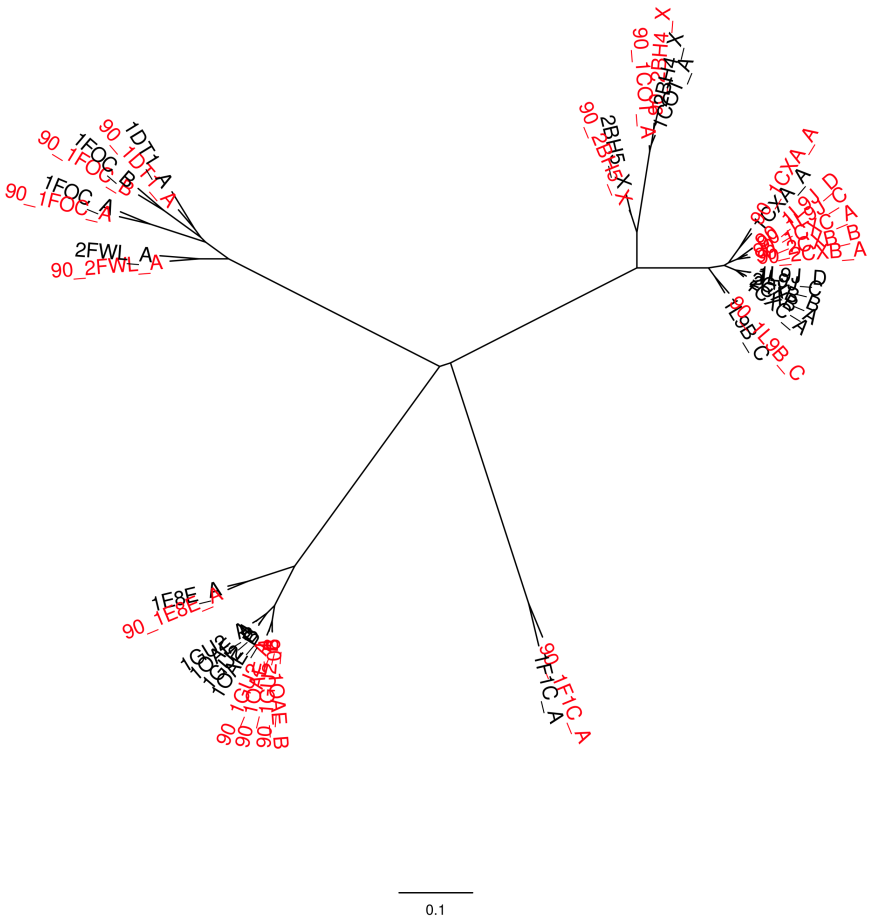


Figure C.9: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (90%), and the complete structures in black.

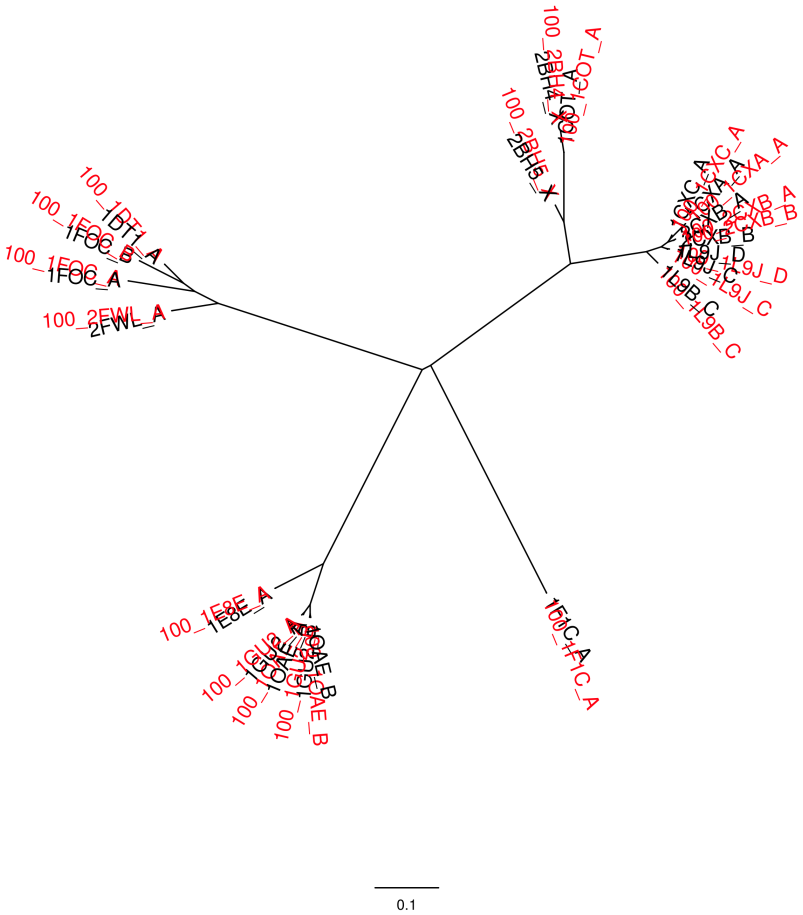


Figure C.10: Fractional structural analysis of proteins in the cytochrome family. The fractional structures are shown in red (100%), and the complete structures in black.

Appendix - III

Ferritins

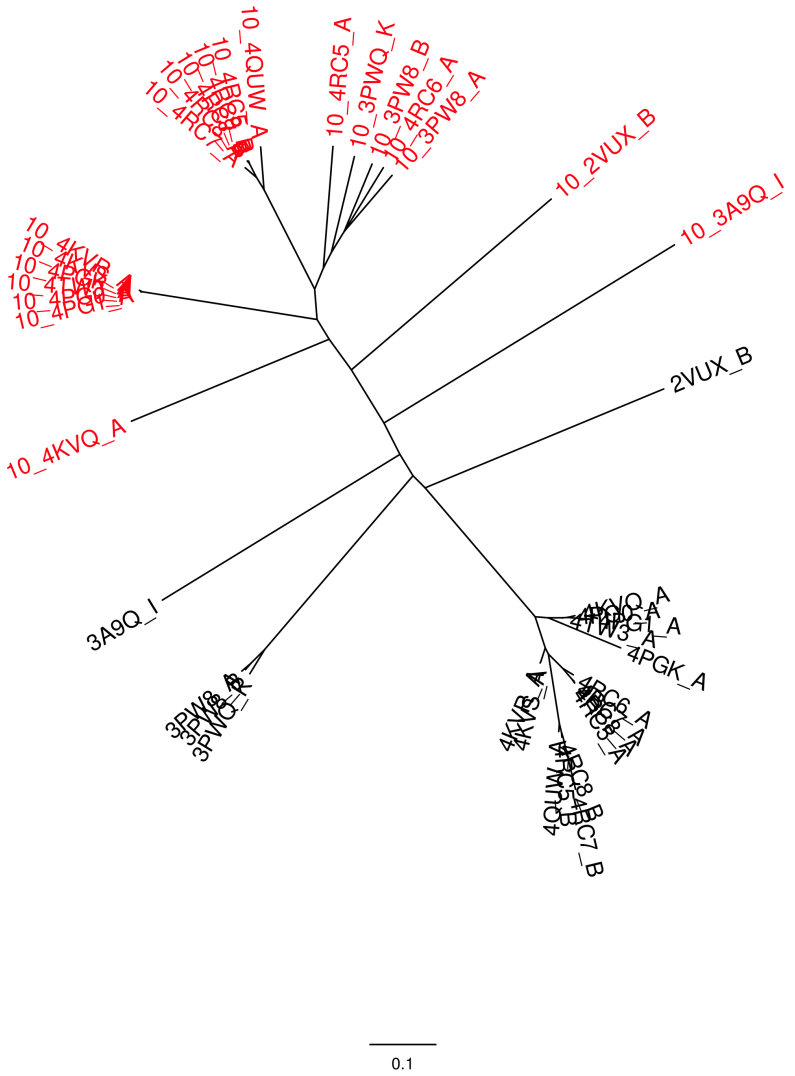


Figure F.1: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (10%), and the complete structures in black.

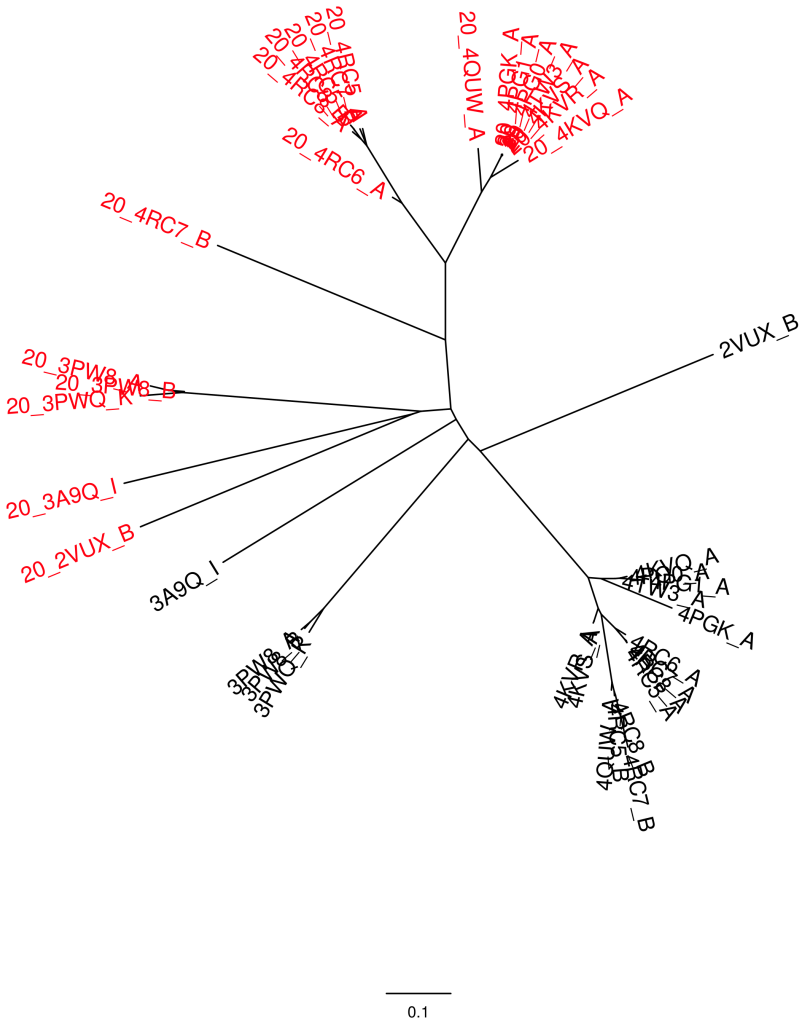


Figure F.2: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (20%), and the complete structures in black.

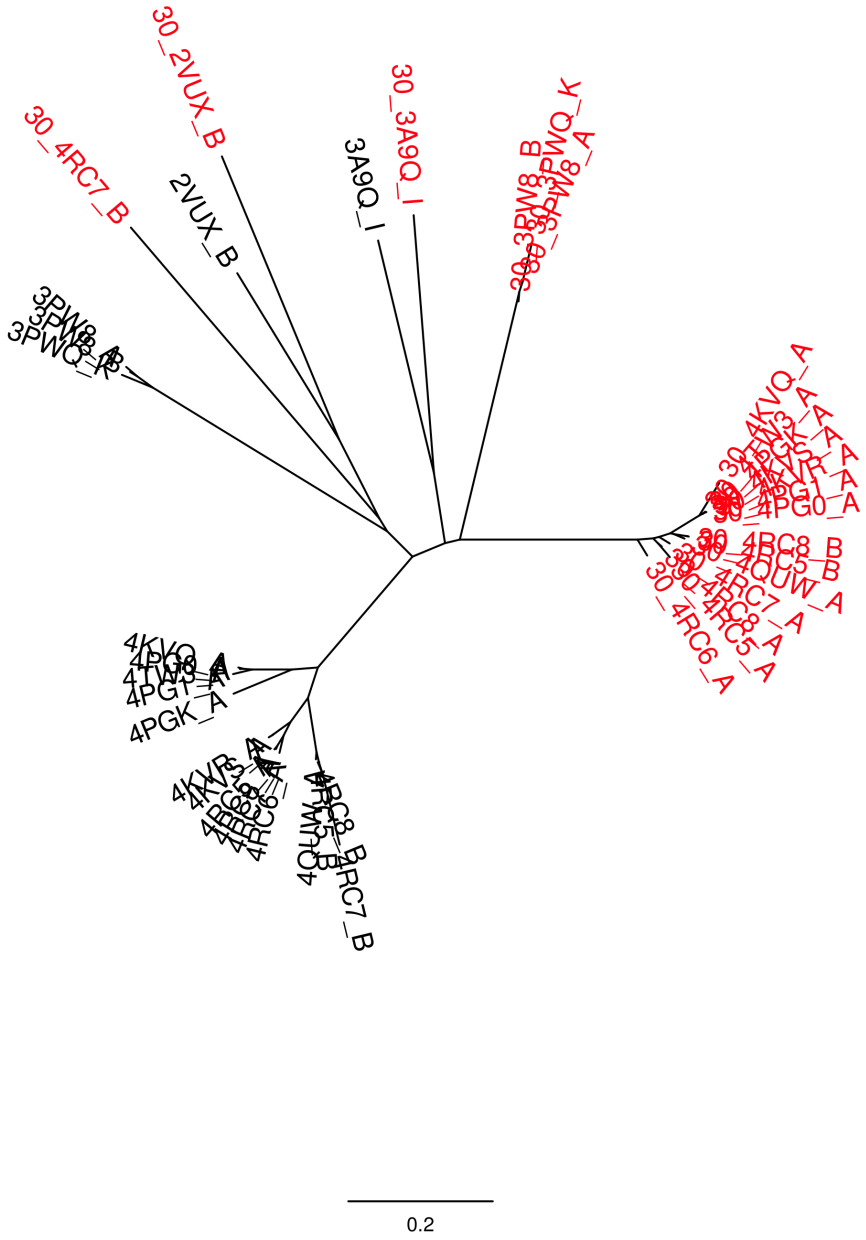


Figure F.3: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (30%), and the complete structures in black.

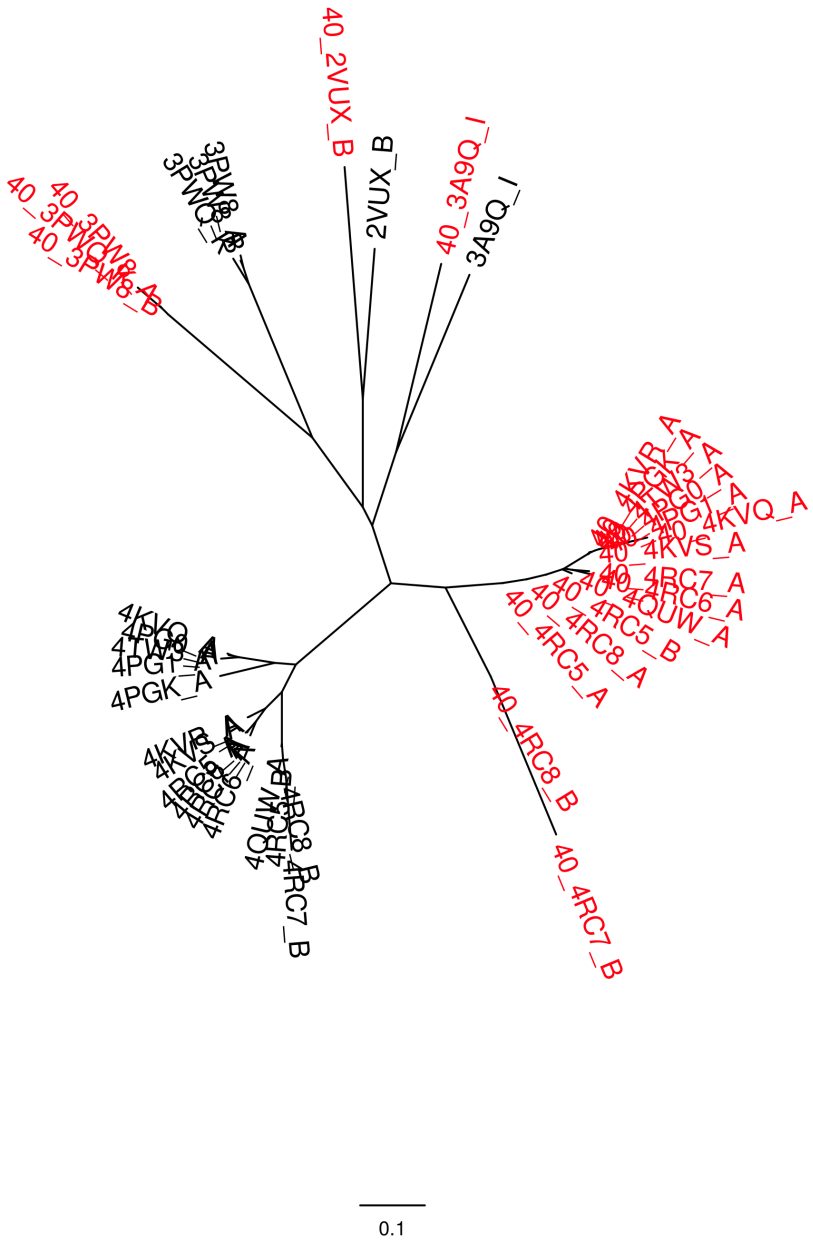


Figure F.4: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (40%), and the complete structures in black.

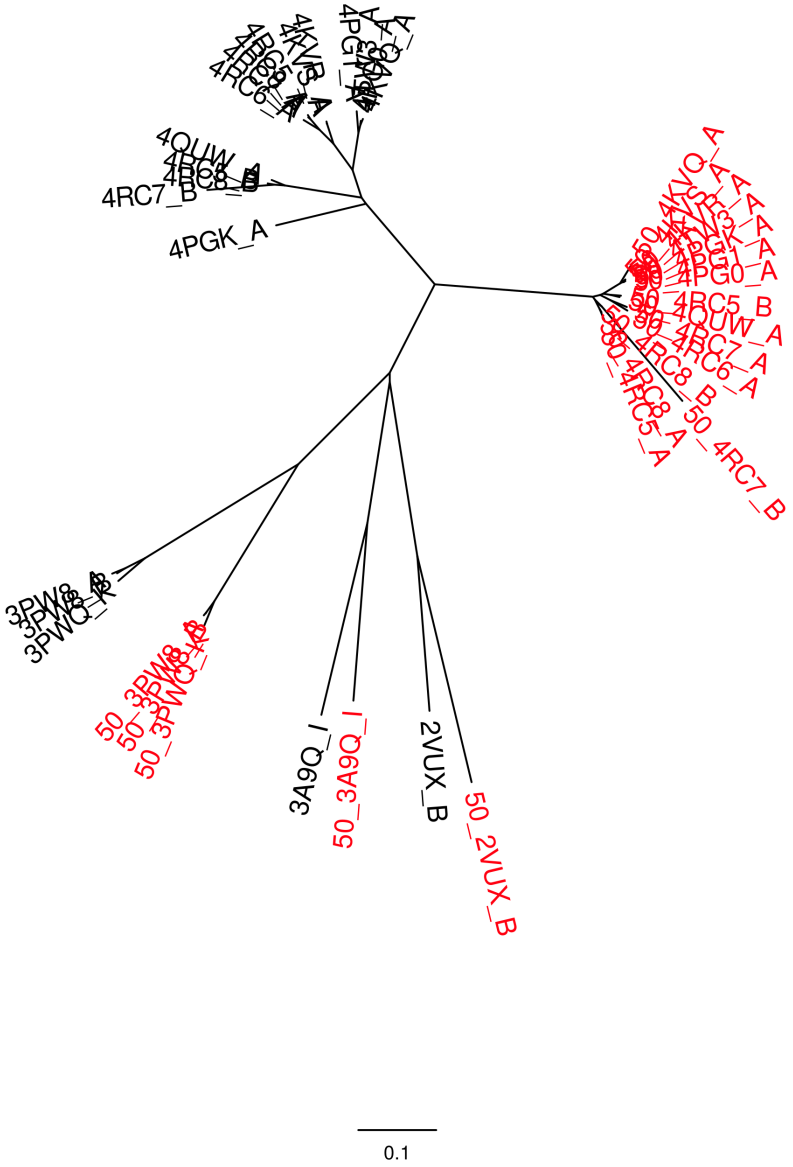


Figure F.5: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (50%), and the complete structures in black.

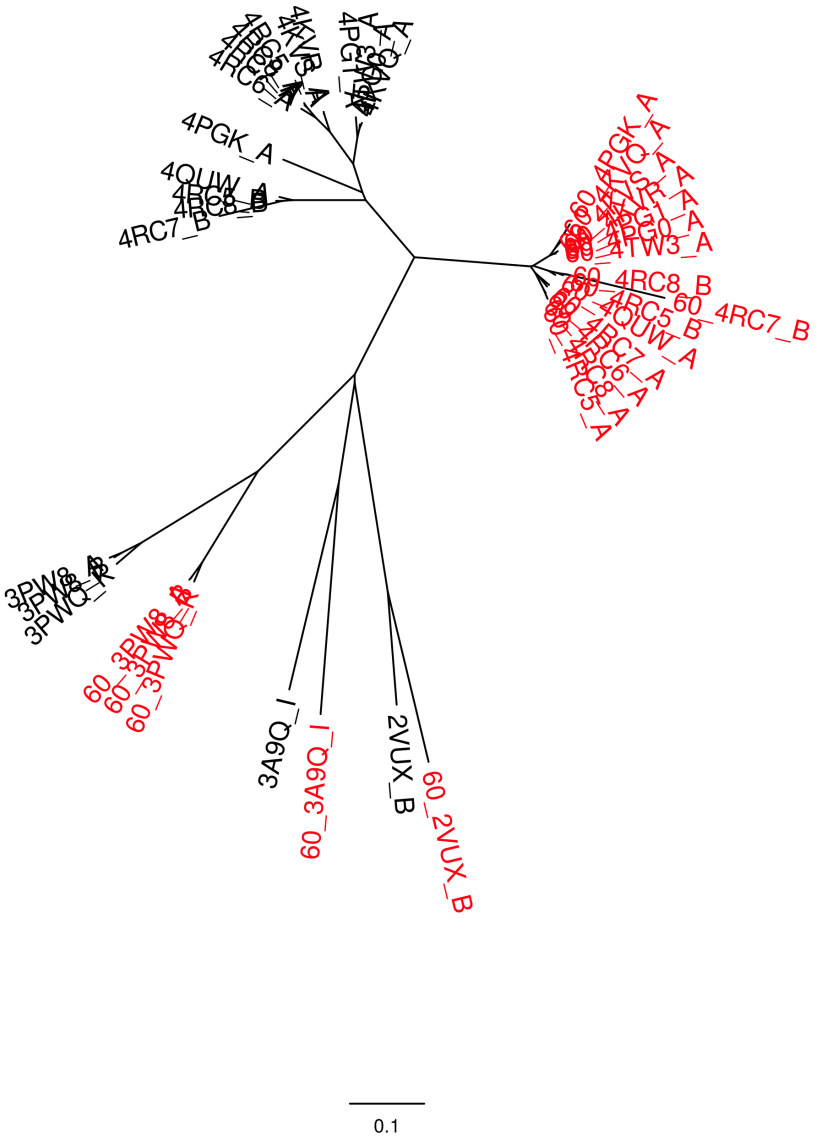


Figure F.6: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (60%), and the complete structures in black.

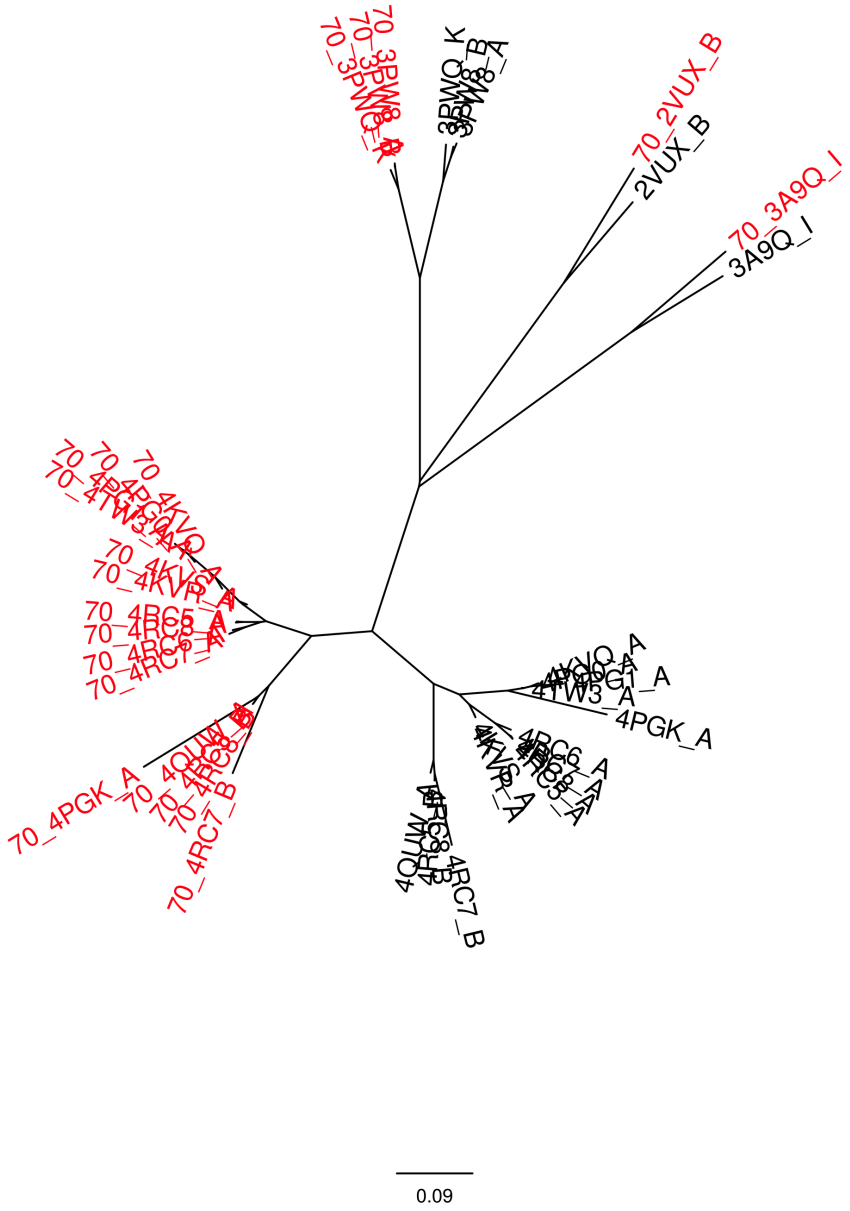


Figure F.7: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (70%), and the complete structures in black.

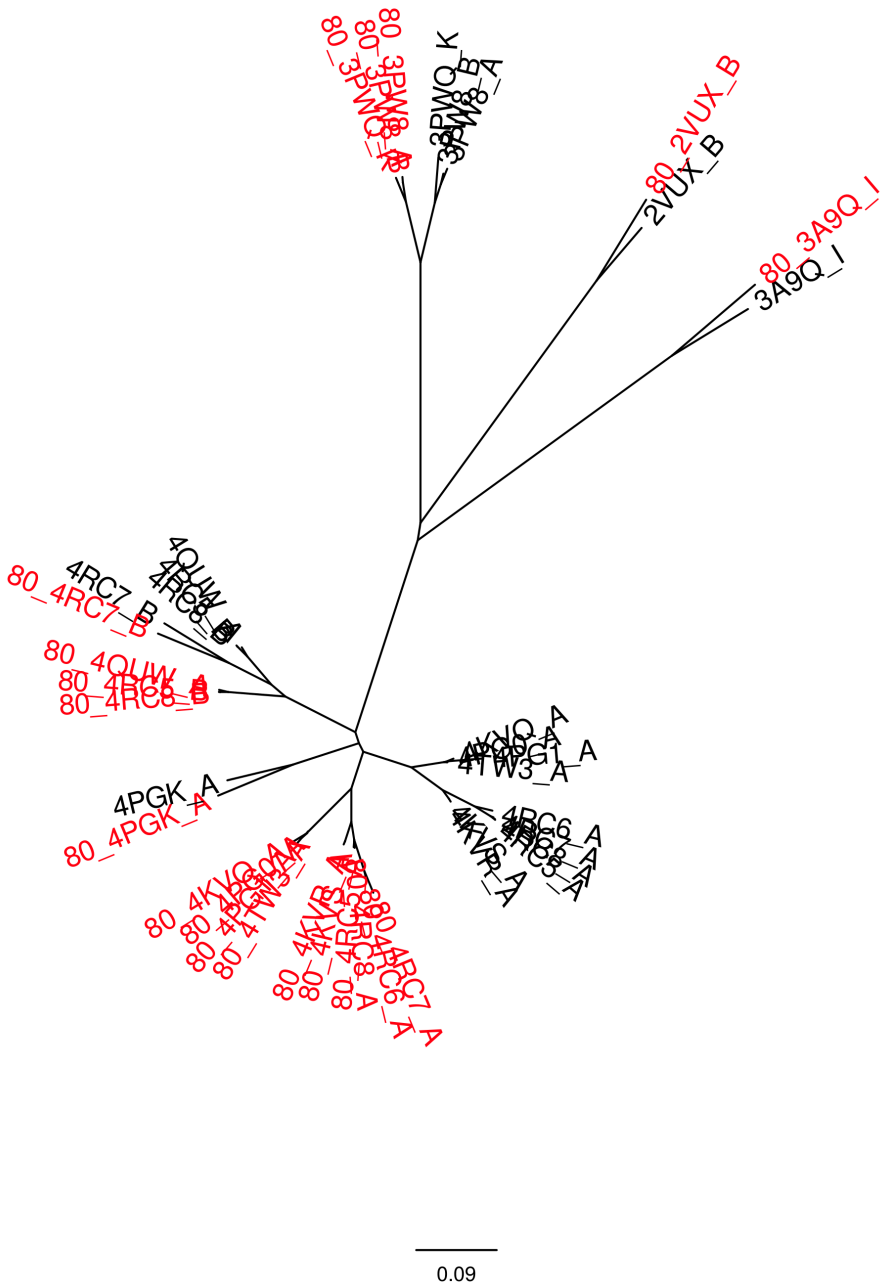


Figure F.8: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (80%), and the complete structures in black.

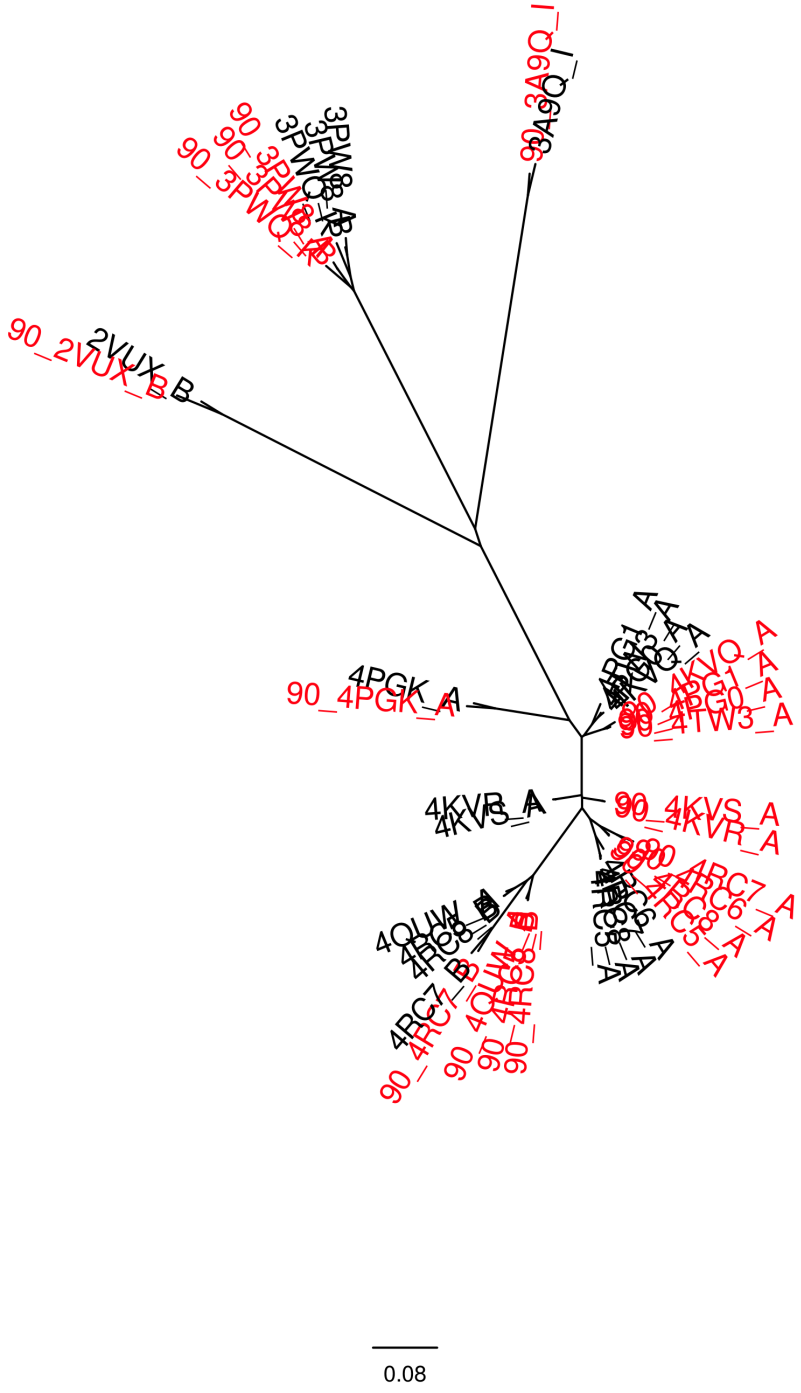


Figure F.9: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (90%), and the complete structures in black.

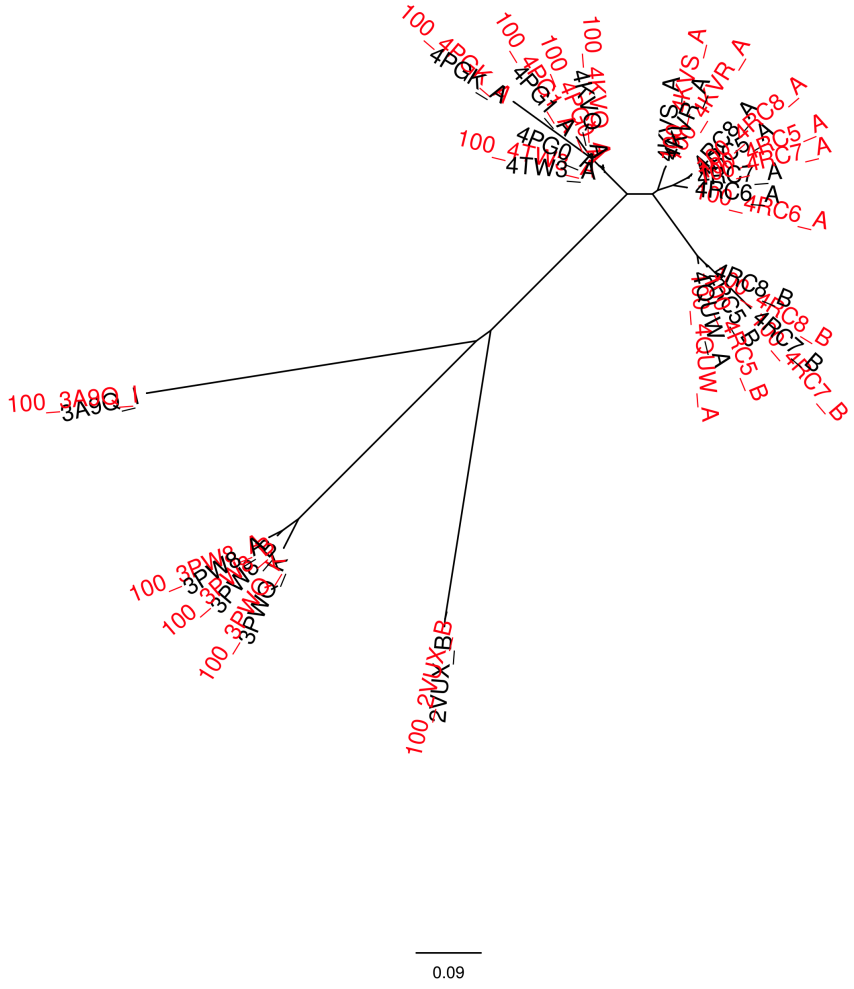


Figure F.10: Fractional structural analysis of proteins in the ferritin family. The fractional structures are shown in red (100%), and the complete structures in black.

Appendix - IV

Globins

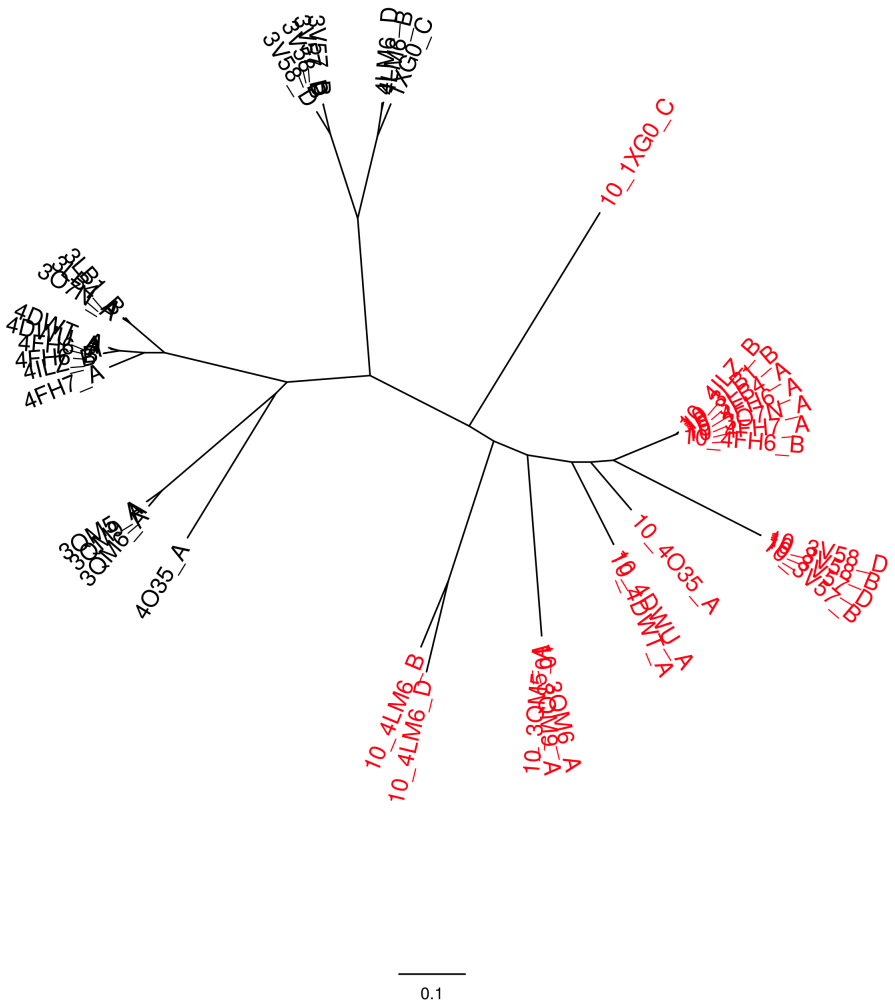


Figure G.1: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (10%), and the complete structures in black.

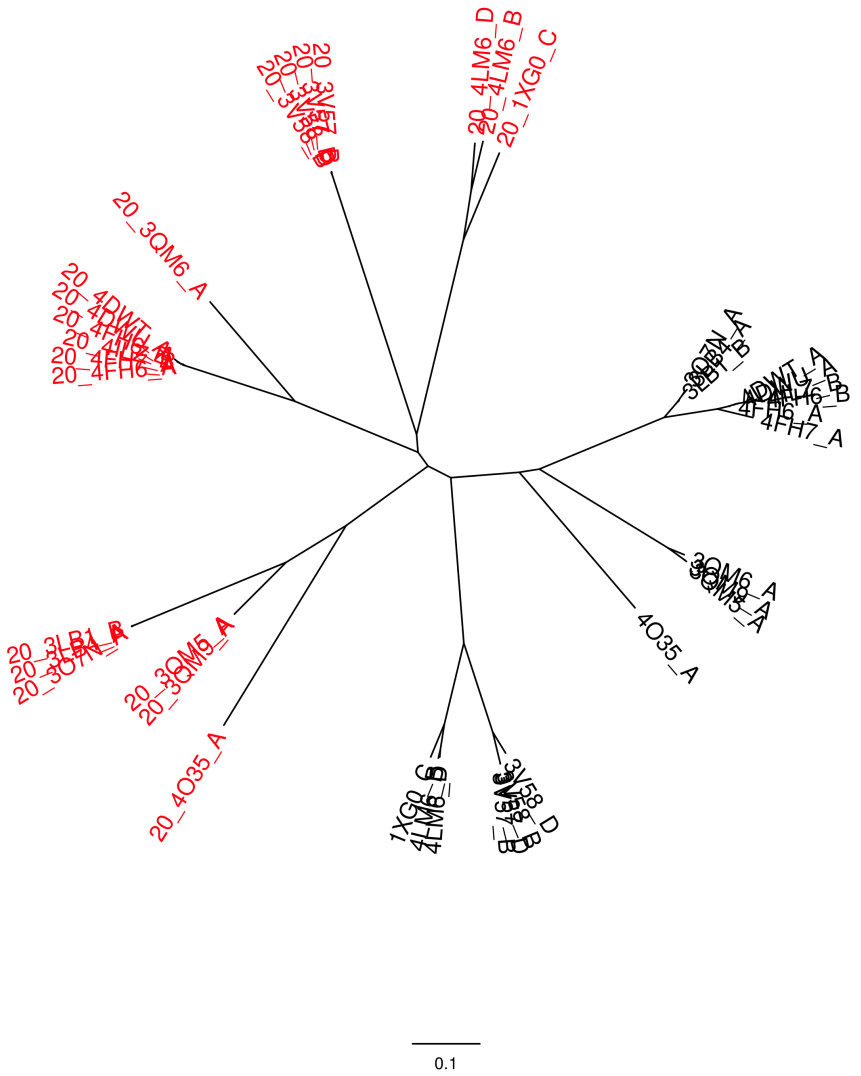


Figure G.2: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (20%), and the complete structures in black.

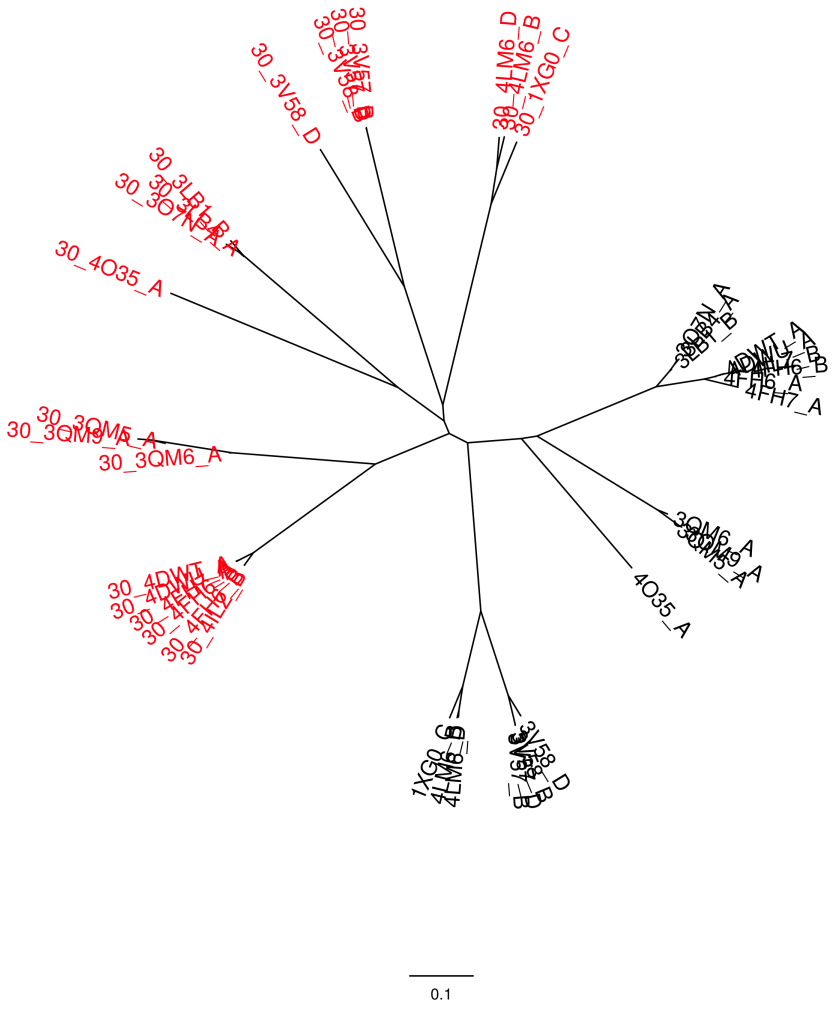


Figure G.3: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (30%), and the complete structures in black.

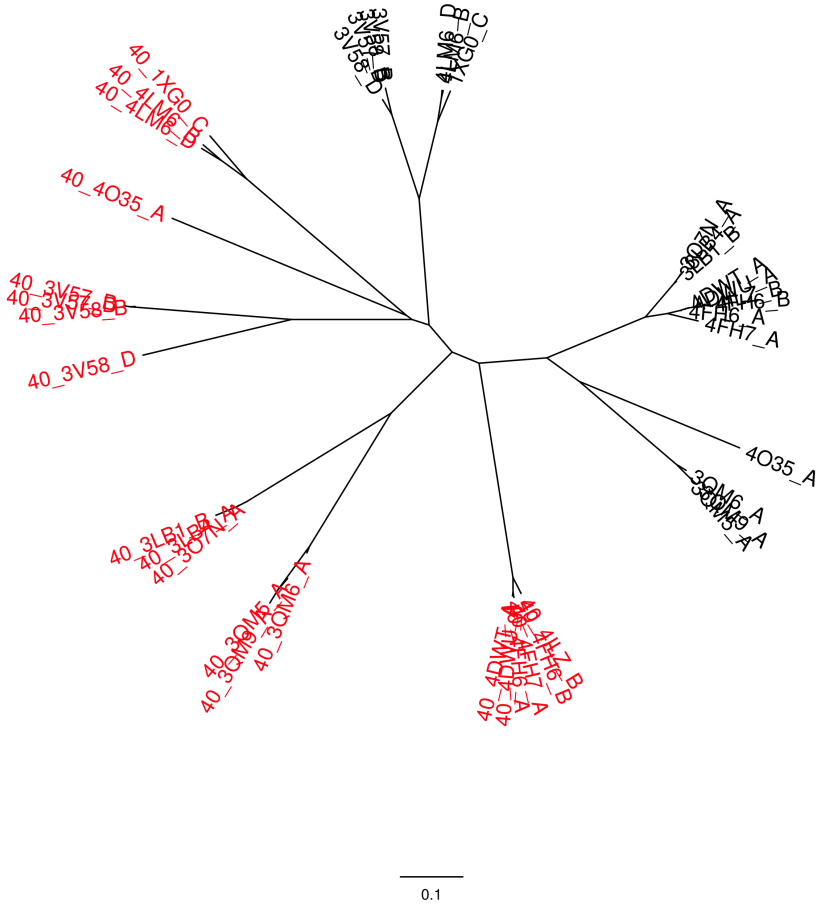


Figure G.4: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (40%), and the complete structures in black.

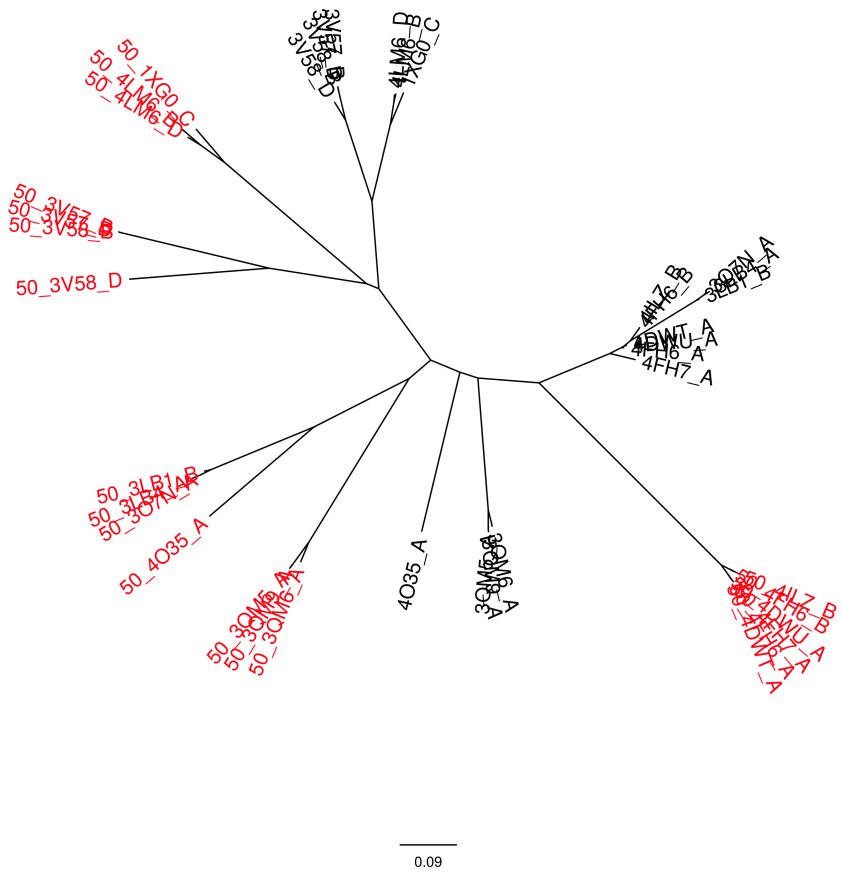


Figure G.5: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (50%), and the complete structures in black.

Figure G.6: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (60%), and the complete structures in black.

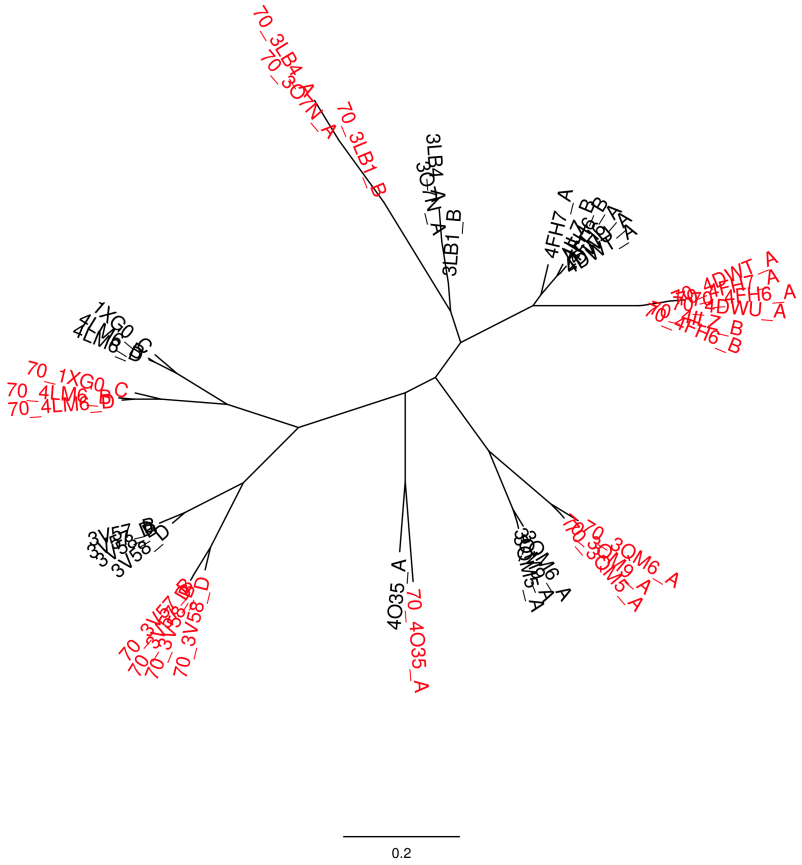


Figure G.7: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (70%), and the complete structures in black.

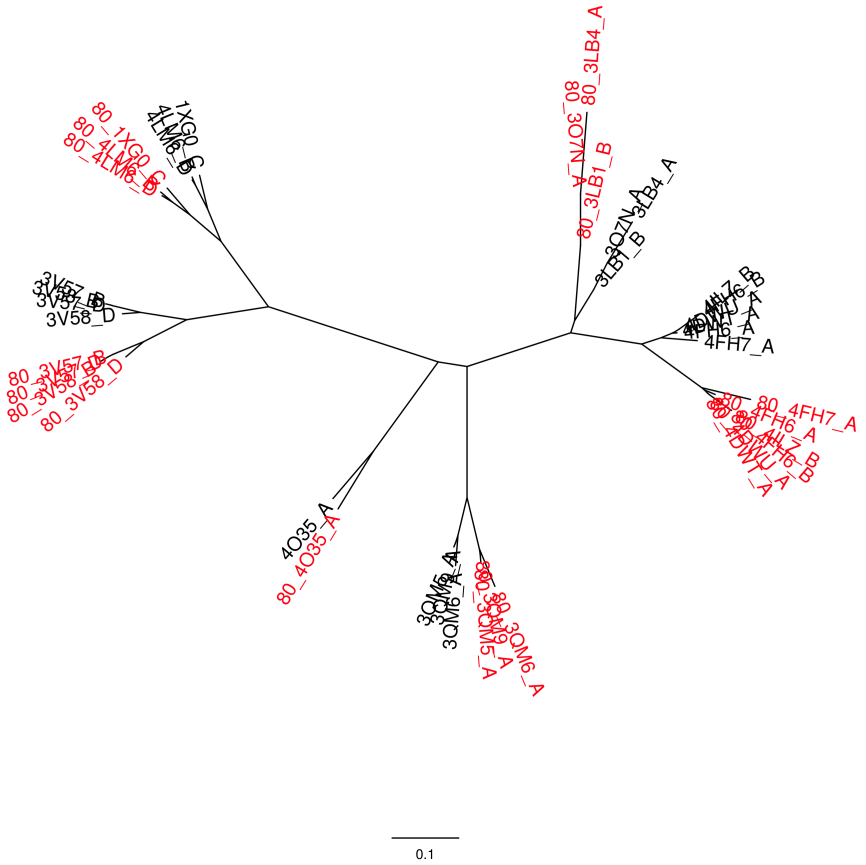


Figure G.8: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (80%), and the complete structures in black.

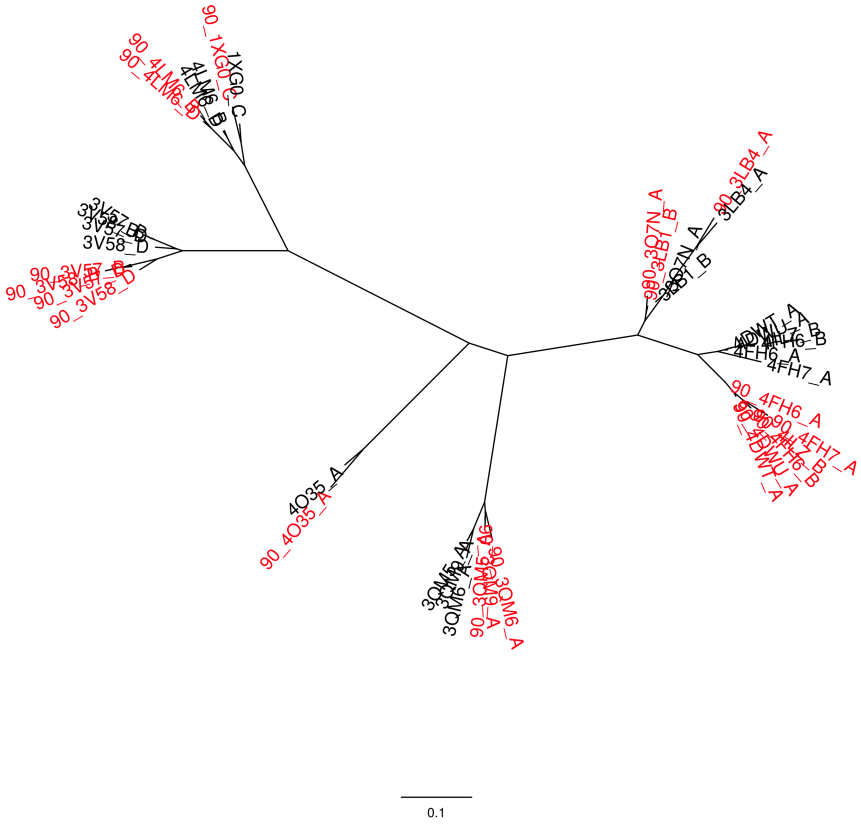


Figure G.9: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (90%), and the complete structures in black.

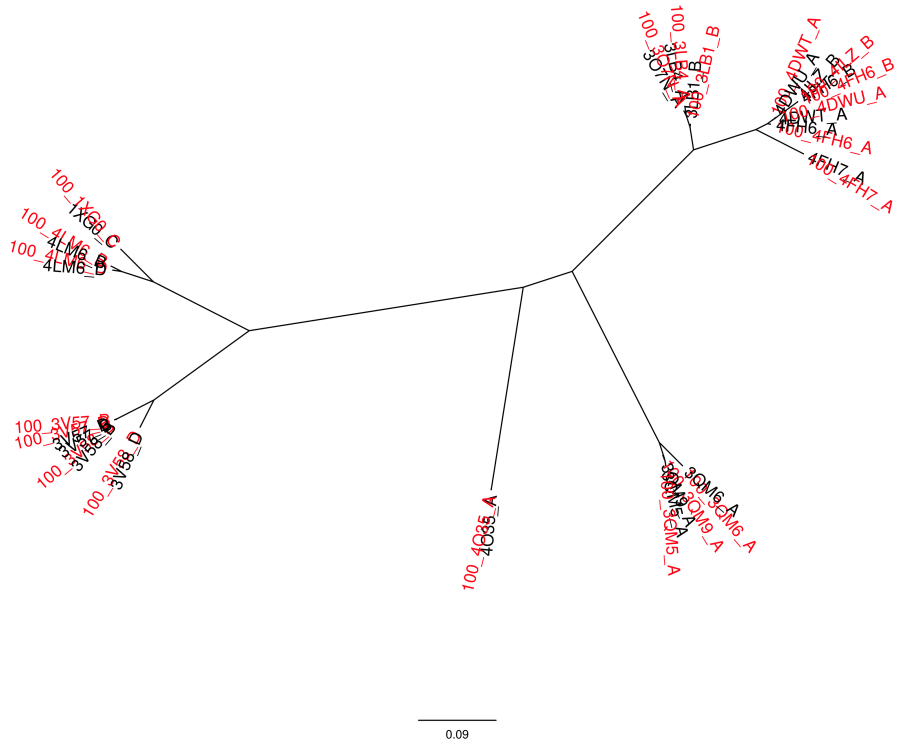


Figure G.10: Fractional structural analysis of proteins in the globin family. The fractional structures are shown in red (100%), and the complete structures in black.