

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.



**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, Aparajita 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

|  |            |
|--|------------|
| <b>Abstract</b>  | <b>i</b>   |
| <b>Acknowledgements</b>  | <b>v</b>   |
| <b>List of Figures</b>   | <b>xi</b>  |
| <b>List of Tables</b>  | <b>xvi</b> |
| <b>1 Introduction</b>  | <b>1</b>   |
| 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 <i>Pairwise alignments using dynamic programming</i> . . . . . | 17         |
| 1.5.2.2 <i>Multiple sequence alignment</i> . . . . .                   | 21         |
| 1.5.3 Inferential method . . . . .                                     | 22         |
| 1.5.3.1 <i>Distance methods</i> . . . . .                              | 23         |
| 1.5.3.2 <i>Character methods</i> . . . . .                             | 24         |
| 1.5.4 Phylogenetic tree . . . . .                                      | 24         |
| 1.5.4.1 <i>Parametric and non-parametric bootstrap</i> . . . . .       | 25         |

|          |  |    |
|----------|--|----|
| 1.6      | Structure-based phylogenetics . . . . .                              | 27 |
| 1.6.1    | Hybrid sequence-structure methods . . . . .                          | 27 |
| 1.6.2    | Molecular phylogenetics: 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  | <i>Algorithm summary</i> . . . . .                                   | 38 |
| 1.8.7.2  | <i>Pairwise protein comparison: Sequence and structure</i> . . . . . | 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 | <i>Statistical ensemble</i> . . . . .                                | 53 |
| 1.11.4.2 | <i>Simulation environment</i> . . . . .                              | 53 |
| 1.11.4.3 | <i>Boundary conditions</i> . . . . .                                 | 53 |
| 1.11.5   | Molecular dynamics: Method breakdown . . . . .                       | 54 |
| 1.11.5.1 | <i>Energy minimization</i> . . . . .                                 | 54 |
| 1.11.5.2 | <i>Molecular dynamics: Method breakdown</i> . . . . .                | 54 |
| 1.11.6   | MD simulation: An example . . . . .                                  | 56 |
| 1.12     | Summary . . . . .  | 57 |

|   |            |
|---|------------|
| Bibliography . . . . .  | 61         |
| <b>2 Method Development</b>   | <b>73</b>  |
| 2.1 Overview . . . . .  | 75         |
| 2.2 Secondary structure matching-based $Q_{score}$ . . . . .            | 75         |
| 2.3 Method . . . . .  | 76         |
| 2.3.1 <i>Part 1</i> : The size effect . . . . .                         | 76         |
| 2.3.2 <i>Part 2</i> : The shape effect . . . . .                        | 80         |
| 2.3.3 The MD-based bootstrap method . . . . .                           | 81         |
| 2.4 Results . . . . .   | 83         |
| 2.4.1 <i>Part 1</i> : The size effect . . . . .                         | 83         |
| 2.4.2 <i>Part 2</i> : 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        |
| <b>3 Aminoacyl-tRNA synthetases</b>                                     | <b>113</b> |
| 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        |
| <b>4 The histone fold</b>   | <b>145</b> |
| 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        |
| <b>5</b> | <b>The ferritin-like superfamily</b>   | <b>181</b> |
| 5.1      | Introduction . . . . .   | 183        |
| 5.1.1    | PFam, SCOP and CATH . . . . .  | 183        |
| 5.1.2    | Structural methods in phylogenetics . . . . .                                    | 184        |
| 5.1.2.1  | <i>Structural alignments and scoring</i> . . . . .                               | 184        |
| 5.1.2.2  | <i>Robustness of phylogenetic relationships</i> . . .                            | 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        |
| <b>6 Summary</b>                      | <b>215</b> |
| 6.1 Method development . . . . .      | 217        |
| 6.2 Protein structural data . . . . . | 218        |
| 6.3 Protein databases . . . . .       | 219        |
| 6.4 Future directions . . . . .       | 219        |
| <b>Appendices</b>                     | <b>221</b> |
| 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<br>calculate $Q_{score}$ . . . . . | 39 |
| 1.14 | Superposition of structure using secondary structure matching-<br>based $Q_{score}$ . . . . . | 41 |
| 1.15 | Pairwise sequence alignment of $\alpha$ and $\beta$ -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 | $\alpha$ and $\beta$ -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<br>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-<br>net . . . . .      | 196 |
| 5.5  | Structural phylogeny of the ferritin-like superfamily with sup-<br>port . . . . .        | 197 |
| 5.6  | Conserved structural core amongst members of the ferritin-<br>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 |

