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

List of Figures

List of Tables