Table of Contents

[Introduction 2](#_Toc37780627)

[Background 2](#_Toc37780628)

[Aims 2](#_Toc37780629)

[Literature Review 2](#_Toc37780630)

[Background to Project 2](#_Toc37780631)

[Genetics 2](#_Toc37780632)

[Proteins 2](#_Toc37780633)

[Sequence alignment 3](#_Toc37780634)

[Similarity v Homology 3](#_Toc37780635)

[Phylogenetics 3](#_Toc37780636)

[Using protein sequence to build trees 4](#_Toc37780637)

[Distance Based Methods 4](#_Toc37780638)

[Maximum Parsimony 5](#_Toc37780639)

[Maximum Likelihood 5](#_Toc37780640)

[Twilight Zone 5](#_Toc37780641)

[Using Protein Structure to build trees 6](#_Toc37780642)

[Sequence-Structure Alignments 6](#_Toc37780643)

[Structure-Structure Alignment 6](#_Toc37780644)

[Limitations 7](#_Toc37780645)

[Evaluating Phylogenetic Trees 7](#_Toc37780646)

[Consensus Trees 7](#_Toc37780647)

[Bootstrapping 7](#_Toc37780648)

[Bayesian Posterior Distributions 7](#_Toc37780649)

[1.6 Gaps in Literature 7](#_Toc37780650)

[Methods 8](#_Toc37780651)

[Constructing Sequence Alignment 8](#_Toc37780652)

[Constructing Structural Alignment 8](#_Toc37780653)

[Combining Sequence and Structural Information 8](#_Toc37780654)

[Linear combination of distance matrices 8](#_Toc37780655)

[Grafted trees 8](#_Toc37780656)

[Results 8](#_Toc37780657)

[Case Study 1: Zinc Hydrolyse Family 8](#_Toc37780658)

[Evaluating Tree 8](#_Toc37780659)

[Case study 2 8](#_Toc37780660)

[Discussion 8](#_Toc37780661)

[Implications of Research 8](#_Toc37780662)

[Future Research 8](#_Toc37780663)

[Conclusion 8](#_Toc37780664)

# Introduction

## Background

## Aims

The aims for this thesis project are to: 1. create an application or pipeline that utilizes three-dimensional protein structure data to improve accuracy of phylogenetic trees created using sequences with low homology. 2. Evaluate this application in comparison to existing methods for performance and efficiency. Due to structure being more highly conserved than sequence it is expected to improve the reliability of the phylogenetic tree.

# Literature Review

## Background to Project

A brief overview of topics important to phylogenetic analysis of proteins are outlined below, this includes genetic sequence and how it influences protein sequence and protein structures. How these sequences are aligned, the first step to phylogenetic analysis, is explained along with the theory of homology.

### Genetics

Deoxyribonucleic Acid (DNA) encodes genetic information that determines the function of cells and ultimately the function of the entire organism. There are four nucleotides that comprise DNA: Guanine (G), Cystosine (C), Adesine (A) and Thyamine (T). These nucleotides are connected together in a strand of DNA by the sugar phosphate backbone. Two strands of DNA then form a double helix structure with nucleotides G/C and A/T pairing up and bonding across the strands. Organisms may also encode their genetic information in Ribonucleic Acid (RNA) instead of DNA. The difference is that RNA has the nucleotide Uracile (U) instead of T and stays single stranded instead of forming a double helix. The sequence of DNA is important because it determines the expression of proteins through genes. Genes are sections of DNA that code for specific proteins, the sequence of these genes determines the sequences of the protein and thus the function of the protein (Genetic Alliance, 2009).

### Proteins

Proteins are essential for the function of organisms and have complex configurations and shapes that determine their functionality. Protein structures can be broken down into four component parts which combine together to reach the final protein product. The primary structure of a protein is its amino acid sequence (Sanger & Tuppy, 1951). There are 22 different amino acids that can be strung together to form the amino acid sequence, this sequence is determined by the DNA sequence with three nucleotides determining one amino acid. The secondary structure of proteins involves the formation of local structures due to the formation of hydrogen bonds between amino acids in the protein sequence These structures can be alpha helixes, beta strands or beta sheets (Pauling, et al., 1951). Tertiary protein structure is the formation of compact protein structures driven by hydrophobicity of amino acids and quaternary structure

### Sequence alignment

A simple sequence alignment involves taking two sequences, for example an amino acid sequences, and aligning the positions of the amino acids in the sequences to minimize amino acid mismatches and gaps between amino acids while maximizing amino acid matches. Alignments can be used to find the most similar region of two sequences with a local alignment algorithm such as Smith-Waterman (Smith & Waterman, 1981). Or the best overall alignment can be found using a global alignment algorithm such as Needleman-Wunsch (Needleman & Wunsch, 1970). More than two sequences can be aligned together, this is called a multiple sequence alignment, and is created by aligning sequences to the sequence alignment. Programs such as ClustalW (Larkin, et al., 2007) produce multiple sequence alignments which can then be used in phylogenetic analysis.

### Similarity v Homology

Sequences are said to be homologous when they share a common ancestor. Homology can be estimated from the level of similarity in a sequence using evolutionary models. Finding homologous sequences and estimating the evolutionary distance between them can be accomplished by reconstructing a phylogenetic tree.

## Phylogenetics

Tree diagrams have long been used to visualize the evolutionary relationships between organisms, from Darwin’s Tree of Life (Darwin, 1859) to modern phylogenetic trees. Advances in gene sequencing has seen the rise of using molecular sequence data reconstruct evolutionary trees. Models for how sequences evolve and change overtime are used to reconstruct phylogenetic trees by investigating the differences between sequences and then applying the model to calculate evolutionary distance between the sequences (Page & Holmes, 1998). For example, DNA sequences of organisms can be used to reconstruct a tree that shows the evolutionary relationship between these organisms.

The trees themselves consist of branches, internal nodes and leaves. Each leaf represents a single sequence. Internal nodes represent unknown common ancestors of the leaf nodes. Branches connect the internal nodes and leaves indicating how the sequences are related to one another through common ancestry. Tree topology refers to the way in which nodes, branches and leaves are arranged, multiple tree topologies can exist for the same dataset of sequences. Therefore, the tree displays an evolutionary record for the organisms indicating how distantly they diverged from common ancestors. A screen shot of a computer

Description automatically generated

Figure Rooted and unrooted phylogenetic tree example showing features of trees

Phylogenetic trees can be rooted or unrooted as shown in Figure 1, rooted trees have a single internal node designated as the root and therefore the oldest common ancestor in the set. Unrooted trees do not indicate which node is the oldest and have an impact on the interpretability of the tree (Williams, et al., 2015). Unrooted trees can be rooted using an outgroup i.e. an organism that is known to be distantly related to the other organisms in the set, by using the common ancestor of the outgroup and other organisms as the root. Outgroup rooting generally performs better at rooting trees than other methods however, using highly divergent outgroups can cause inaccuracies in the tree which must be taken into account (Li, et al., 2012) or other methods used (Williams, et al., 2015).

## Using protein sequence to build trees

Phylogenetic trees can be re-constructed based on protein sequences. As proteins evolve deletions, additions and substitutions can occur to the amino acids that comprise the protein sequence. Thus, by comparing protein sequences through multiple sequence alignments or other methods these changes can indicate how closely or distantly different proteins are related. Phylogenetic trees of homologous proteins can identify new protein families and the members of the families (Gabaldon, 2007).

### Distance Based Methods

Closely related sequences are more likely to have high similarity and therefore quantifying the distance between sequences can be used to generated phylogenetic trees. The first step to this process is doing a multiple sequence alignment for all sequences in a sample using programs such as TCoffee (Notredame, et al., 2000) or MUSCLE (Edgar, 2004). Then for each pair of sequences in the alignment a dissimilarity score is calculated by adding up penalties for each mismatched nucleotide or amino acid. Methods such as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal & Michener, 1958) and Neighbor-Joining (NJ) (Saitou & Nei, 1987) are clustering algorithms that take in a dissimilarity matrix and create an evolutionary tree. UPGMA generates an unrooted tree while the tree generated by NJ is rooted.

UPGMA is a bottom up algorithm that first finds the two sequences with lowest dissimilarity score and creates a branch in the phylogenetic tree with depth equaling half the dissimilarity score. These sequences are then merged in the dissimilarity matrix and scores are averaged where a merge occurs. These steps are repeated until no more merges can occur. In contrast to UPGMA the phylogenetic tree created is unrooted. NJ works similarly by finding pairs of most closely related sequences and forming a new node in the tree. It also minimized the lengths of all its branches thereby minimizing evolutionary distance between sequences. NJ is a fast algorithm even for large datasets and generally effective at creating correct tree topologies. (Gabaldon, 2007)

UPGMA and NJ are widely used and available in tree building packages sucW2 (Larkin, et al., 2007) and PHYLIP (Felsenstein, 1993) but may not always perform optimally. The algorithms display some chaotic behavior when the order of sequences inputted is varied. Multiple tree topologies are produced by the same program and algorithm for the same dataset when the order of the input is changed (Backeljau, et al., 1996). Furthermore NJ and UPGMA rely on the assumption that mutations rates remain constant at a particular positions in a sequence and performs badly when this assumption does not hold (Som & Fuellen, 2009).

### Maximum Parsimony

Maximum parsimony methods of constructing phylogenetic trees relies on the assumption that less changes to a sequence over time is more likely. Therefore, trees are reconstructed to simulate a scenario where the least amount of amino acid changes occur to get from common ancestors to the sequences given. Fitch’s algorithm (Fitch, 1971) applies maximum parsimony to build phylogenetic trees by assigning a score for a substitutions and then minimizing that score to produce a tree topography with minimal substitutions between sequences. Maximum parsimony generally performs well and can be improved with heuristics such as described by Weng, et al. (Weng, et al., 2012) who use probabilities to improve the performance of Fitch’s algorithm (Fitch, 1971). However, maximum parsimony methods can produce inconsistent results even when the substation rate is stable because the length of the sequence is influential in how many substitutions it accrues. The longer a sequence is the more likely it is for a substitution to occur and therefore the length of a sequence obscures actual evolutionary relationships (Takezaki & Nei, 1994).

### Maximum Likelihood

Maximum likelihood methods attempt to find the most likely tree topography from the given data by computing all possible tree configurations for the dataset. This method is computationally intense but is the most accurate at producing phylogenetic trees for cell type analysis using RNA (Nair, et al., 2016) and for DNA sequences (Paul & Sahoo, 2014). Maximum likelihood methods also allow the method of modelling evolution to be explicitly specified unlike other methods where the model is implicit to the method such as maximum parsimony. This characteristic enables more fit models to be chosen based on circumstance. To reduce the computational intensity estimations for maximum likelihood can be used in place of calculating the actual likelihood (Paul & Sahoo, 2014).

### Twilight Zone

When reconstructing a phylogenetic tree based on protein sequence similarity the assumption is made that similarity implies homology however this does not always hold true. In the twilight zone, where similarity between sequences is less than 20-30%, homology can no longer be reliably inferred. This is because random similarities in protein sequences cannot be distinguished from similarities due to homologous structures (Chung & Subbiah, 1996).

## Using Protein Structure to build trees

Protein structures are more highly conserved than protein sequence or DNA sequences (Flores, et al., 1993) because protein structure is more important for the function of a protein and variations in sequence can result in the same or similar structures. Rodriguez-Rivas, et al. found proteins that interact with each other and co-evolve likely to have highly conserved structures which can be used in the twilight zone to determine homology between these proteins in prokaryotes and identify protein interactions in eukaryotes (Rodriguez-Rivas, et al., 2016). Furthermore, proteins can diverge to having low sequence similarity while maintaining similarity of sequence and function (Todd, et al., 2001). Therefore incorporating 3-dimensional structural information when there is low sequence similarity may improve protein phylogenetic trees (Balaji & Srinivasan, 2007).

### Sequence-Structure Alignments

Johnson et al. was the first to incorporate protein structure into protein sequence alignments by aligning sequences to a template based on structural information or homologous sequence information (Johnson, et al., 1993). Protein structure data can be used to create multiple sequence alignments with programs such as 3D Coffee (O'Sullivan, et al., 2004) which compiles a library from sequence-sequence, structure-sequence and structure-structure pairwise alignments. The library is then used to assemble a progressive alignment, producing the multiple sequence alignment. Incorporating known structural data improves the quality of the multiple sequence alignment (O'Sullivan, et al., 2004). Predicted protein structures can also be used to improve multiple sequence alignments, SPEM (Zhou & Zhou, 2005) is a method that uses preprocessed secondary protein structure predictions to improve the final multiple sequence alignment. Therefore, when protein structures are not available predicted structures can be used to improve a multiple sequence alignment. PROMALS3D (Pei, et al., 2008) is a multiple sequence aligning program which uses both available protein structural data and predicted protein structure from homolog searches to produce high quality multiple sequence alignments.

### Structure-Structure Alignment

Three-dimensional protein structure alignments produce multiple sequence alignments by comparing protein structure and not proteins sequence. STAMP (Russel & Barton, 1992) is a multiple structural aligner that uses an initial multiple sequence alignment to create a similarity tree for structures and then uses that tree to align the other protein structures. Structure based multiple sequence aligners can perform better than sequence or structure/sequence based aligners with a recent study finding MAMMOTH and MATRAS performing best of all multiple alignment algorithms tested (Carpentier & Chomilier, 2019). MAMMOTH (Ortiz, et al., 2002) is a completely sequence independent algorithm that uses a heuristic capable of accurately aligning even low-resolution three-dimensional protein structures. MATRAS (Kawabata, 2003) uses a structural similarity score to priorities alignment of homologous structures in the three-dimensional protein structure space.

### Limitations

Protein sequence data is much more available and extensive than three-dimensional protein structure data, but protein structure is more highly conserved and therefore more accurate at low sequence similarities (Chung & Subbiah, 1996).

## Evaluating Phylogenetic Trees

Evaluation of phylogenetic trees needs to be undertaken to identify how likely a topography is to represent actual evolutionary processes and relationships.

### Consensus Trees

The choice of algorithm used to build multiple sequence alignments and phylogenetic trees can have a massive impact on the topology of the final tree. Even changes to the order in which protein sequences are provided to an algorithm may change to topography of the phylogenetic tree produced. One way to overcome this issue and find a more optimal tree topography is through a consensus tree. A consensus tree algorithm takes in multiple tree topographies and converges them into one tree, the consensus tree (Bryant, 1991). These algorithms perform well on homologous trees however, consensus trees based on heterologous trees can often be poorly resolved or based on arbitrary choices (Bonnard, et al., 2006).

### Bootstrapping

Bootstrapping (Efron, 1979) can be used to evaluate the robustness of a phylogenetic tree (Felsenstein, 1985). Bootstrapping uses a resampling with replacement on the input data to create multiple trees, a majority consensus tree is then created which indicates the frequency of groups appearing in the bootstrapped trees. Groups with frequencies 95% and above are considered statistically significant (Felsenstein, 1985). This method creates a consensus tree and gives an indication of how much confidence can be placed in a grouping on the tree.

### Bayesian Posterior Distributions

Another method of evaluating phylogenetic trees is by using the Bayesian posterior distribution as outlined in (Huelsenbeck & Ronquist, 2001). The posterior probability of a tree is the probability that the tree is correct given that the model used is correct, this can be used like bootstrapping to indicate how robust a tree topography is. Bayesian methods are more reliable but are more heavily influenced if the model used to build the tree is inappropriate (Huelsenbeck & Rannala, 2004).

## 1.6 Gaps in Literature

Many methods for reconstructing protein phylogenetic trees are available in the literature including methods that utilize sequence data (Felsenstein, 1993), structural data (Ortiz, et al., 2002) or both (Pei, et al., 2008). However streamlined methods that take a hybrid approach using protein sequence data at high similarity and structural data at low similarity were less apparent. The effectiveness of such methods also needs further investigation and comparison to existing phylogenetic tree construction methods.

# Methods

## Constructing Sequence Alignment

## Constructing Structural Alignment

## Combining Sequence and Structural Information

### Linear combination of distance matrices

### Grafted trees

# Results

## Case Study 1: Zinc Hydrolyse Family

### Evaluating Tree

## Case study 2

# Discussion

## Implications of Research

## Future Research

# Conclusion