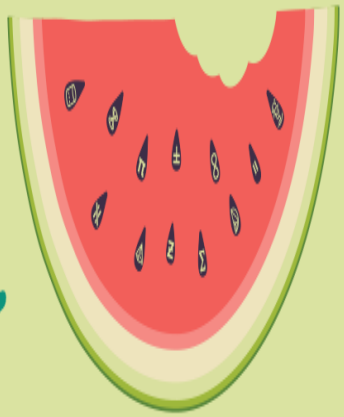
AMSI SUMMER RESEARCH SCHOLARSHIPS 2024-25

Get a TASTE for Research this Summer



Uncertainty in Structural Phylogenetic Analysis

Li Fu Zhang

Supervised by Dr. Heejung Shim

The University of Melbourne

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

AMSI CONNECT

Abstract

Phylogenetic analysis is the study of the evolutionary relationships among entities, aiming to build a phylogenetic tree to describe these relationships. Until recently, most of the phylogenetic analyses were conducted using DNA or amino acid sequences. Because of the advent of AI-based protein structure prediction tools (e.g., AlphaFold 2), more and more people choose to perform structural phylogenetic methods instead of the traditional sequence-based methods, as the structure of a protein is more conserved than its underlying amino acid sequence. Recent structure-based methods often involve translating the protein structure to a sequence called 3Di sequence using Foldseek, where different 3Di letters correspond to different conformations, and performing alignment, trimming, and phylogenetic inference. Due to the diversity of protein structures and the limited number of 3Di letters, there is uncertainty within the translation. In this research project, we manage to quantify such uncertainty. Specifically, for each position of the protein structure, we calculate the likelihood of the position belonging to different 3Di letters. Then, in order to assess whether the uncertainty information we calculate is useful, we perform downstream analyses including visualisation of uncertainty information and structural phylogenetic analysis using uncertainty information. Results suggest that structural phylogenetic analysis using uncertainty information outperforms the original structural phylogenetic method when Foldseek is confident in translating the secondary structures of proteins.

1 Introduction

The recent advances in protein structure prediction (e.g., AlphaFold 2) have had a profound impact on the field of bioinformatics. With AlphaFold 2, protein structures can now be predicted with near-experimental accuracy in a remarkably short time. This breakthrough has made it possible to use the structure-based phylogenetic analysis method, which is considered more reliable than traditional sequence-based methods, as protein structures typically evolve more slowly than their corresponding amino acid sequences.

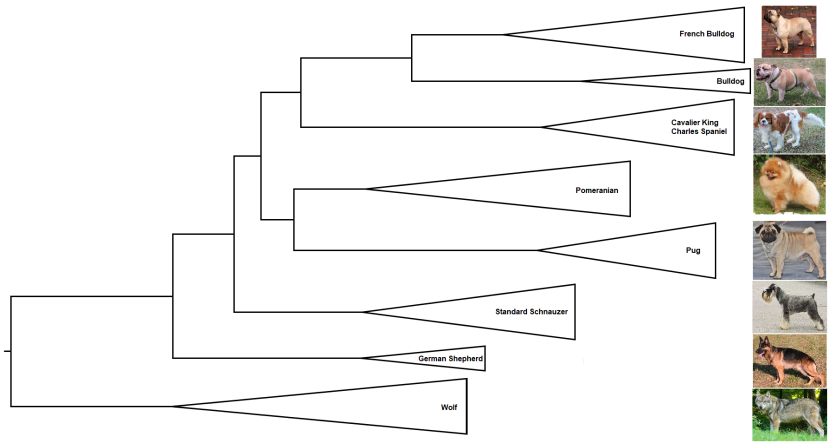


Figure 1: Phylogenetic tree of dogs
https://en.wikipedia.org/wiki/Phylogenetic_tree/media/File:Phylogenetic_tree_of_dogs.png.

Phylogenetic analysis is a method used to study the evolutionary relationships among organisms, inferring their common ancestors. Figure 1 shows an example phylogenetic tree, where we can read out the evolutionary relationships of those dogs. Such a phylogenetic tree is produced by performing phylogenetic analysis on either genetic, protein, or morphological data of the species we are interested in. In this research, we focus on the phylogenetics using protein data. For the two phylogenetic analysis methods mentioned above (sequence-based method and structure-based method), the diagrams of them are shown below (Figure 2). In contrast to the sequence-based method, recent structure-based methods often use a software called Foldseek [1] to translate the protein structure to a sequence called 3Di sequence, which is then used to perform alignment, trimming, and phylogenetic inference [2].

Figure 3 shows how Foldseek translates protein structure to 3Di sequence (hereafter referred to as the Translation Method). For each residue of a protein, Foldseek firstly looks for its

neighbouring residues. Then, ten features that characterise the local structural information are extracted. These features are then embedded into a two-dimensional continuous latent space. After that, the embedding (the coordinate in the 2D latent space) is discretised by mapping each point to its nearest centroid, with each centroid corresponding to a unique 3Di letter. And there are 20 3Di letters in total.

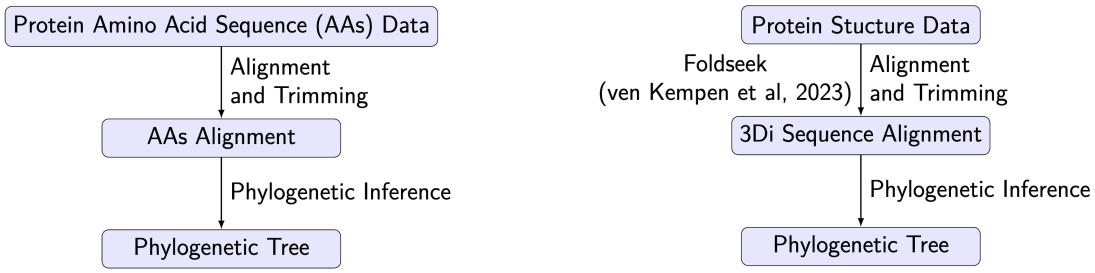


Figure 2: **Left**: Diagram of traditional phylogenetic analysis. **Right**: Diagram of structural phylogenetic analysis from [2].

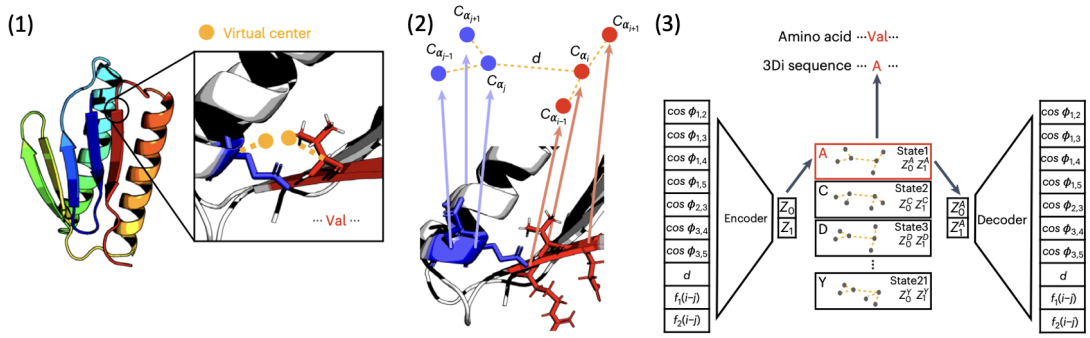


Figure 3: Visualisation of Translation Method from [1]

Alignment, trimming, and phylogenetic inference are three common steps in phylogenetic analysis. Alignment is the process of lining up multiple sequences so that similar or homologous positions in the sequences are properly aligned (Figure 4). For trimming, it refers to the process of cutting off the columns in the alignment where there is a low proportion of data (Figure 5). The last step, phylogenetic inference, refers to employing statistical methods such as maximum likelihood and Bayesian inference to build phylogenetic trees. The software I use here is IQ-TREE2 [5], which builds the phylogenetic trees by maximum likelihood.

Figure 4: Alignment from [1]

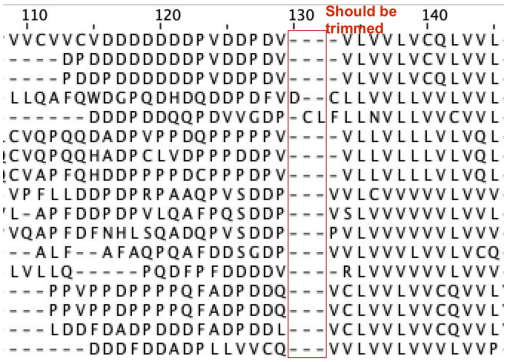


Figure 5: Visualisation of trimming

However, for the structural phylogenetic analysis above, the downside of it is that in step (3) of the Translation Method (Figure 3), we are losing information due to the limited size of the 3Di alphabet. For example, in the case that is shown by Figure 6, even though 'A' is selected as the 3Di letter for the two positions, position 1's and position 2's local structures may exhibit some differences. Motivated by this, I explore the uncertainty in structural phylogenetic analysis, specifically the uncertainty in the Translation Method. Then, I assess the impact of this uncertainty on downstream analysis.

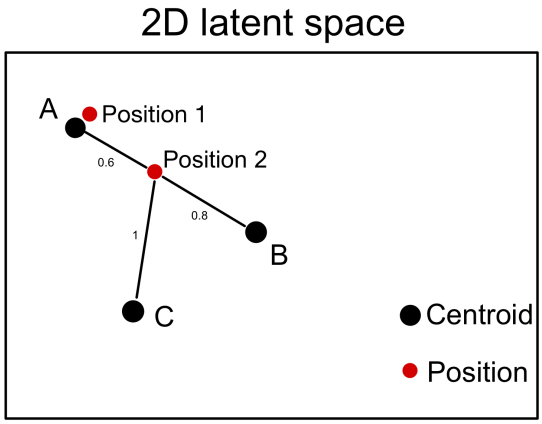


Figure 6: Visualisation of motivating example

2 Datasets

The datasets used in the downstream analysis include protein structures of the ferritin-like superfamily and glycoproteins. All of the proteins are clipped to single chains in order to be used to conduct phylogenetic analysis.

2.1 Protein structure of the ferritin-like superfamily

Ferritin-like superfamily is a group that shares the ferritin fold. The protein structure data of the ferritin-like superfamily, obtained from [2], consist of structures predicted by AlphaFold 2.

2.2 Protein structure of glycoproteins

Glycoproteins are proteins that are possessed by the Flaviviridae family, which is a highly diverse family of RNA viruses. There are a total of three types of glycoproteins: E, E1, and E2. The protein structure data of glycoproteins, obtained from [3], consists of structures predicted using AI-based structure prediction tools.

3 Methods

As mentioned above, we explore uncertainty in structural phylogenetic analysis, specifically uncertainty in the Translation Method. To do that, we modified the Translation Method from [1]. After that, in order to assess whether the uncertainty information we obtained is useful, we perform two downstream analyses: visualisation of uncertainty information and structural phylogenetic analysis with uncertainty information.

3.1 Uncertainty calculation

The uncertainty in the Translation Method arises from its last step, where the embedding is discretised into 3Di letters by the nearest centroid. In this research, we calculate a 20-dim vector for each residue, termed uncertainty, which represents the likelihood of the residue belonging to different 3Di letters (Figure 7). Detailed calculation is shown below:

1. For each residue r, extract 10 features v_r that describe the local structural information of the residue using the code from [1].

2. Encode its local structural information to a coordinate in the 2D latent space $f_{Encoder}(v_r)$ using the Encoder from [1].

3. Then for each centroid in the latent space $c \in [A, ..., Y]$, we calculate its distance d_{r_c} to $f_{Encoder}(v_r)$.

4. For all $c \in [A, ..., Y]$, map the distance d_{r_c} to $z_{r_c} = -2d_{r_c}$.

5. Apply the softmax function to the closeness (z_{r_c}), where $softmax(z_{r_c}) = \frac{exp(z_{r_c})}{\sum_{c\in[A...Y]} exp(z_{r_c})}$.

After step 4 and step 5, all of the normalised closeness sum to 1. Therefore, for each residue, we can use the vector of the normalised closeness to represent the likelihoods of the residue belonging to different 3Di states.

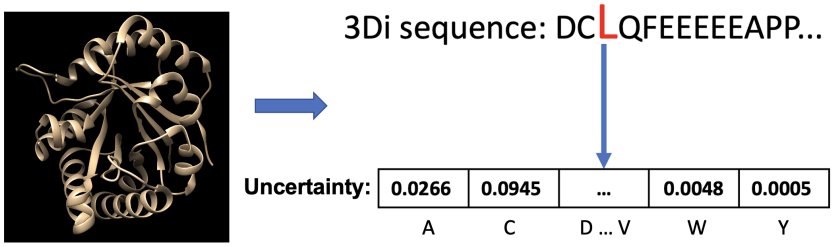


Figure 7: Visualisation of uncertainty calculation

3.2 Visualisation of uncertainty information

For each 3Di sequence we obtained using the Translation Method, we calculate the uncertainty of each position of it. Then for each position, among the 20 likelihoods, we retrieve the one that corresponds to the 3Di letter of the position. Since 3Di sequences are aligned by adding gaps to the sequences, which can be viewed as a mapping, we then have the uncertainty of each position of the aligned 3Di sequences (the sequences that appear in alignment), by mapping the uncertainty of each position of the 3Di sequences.

3.3 Structural phylogenetic analysis using uncertainty information

To conduct structural phylogenetic analysis that incorporates uncertainty information, we modified the structural phylogenetic analysis method from [2] (Figure 8). On top of their trimming method, we also trim the columns that have high uncertainty. In particular, for each column in an alignment, we calculate the median uncertainty across all species. Then, we filter out columns with low median uncertainty. The cut-off value of uncertainty is selected manually according to the data. The "doubly trimmed" alignment is subsequently used to perform phylogenetic inference.

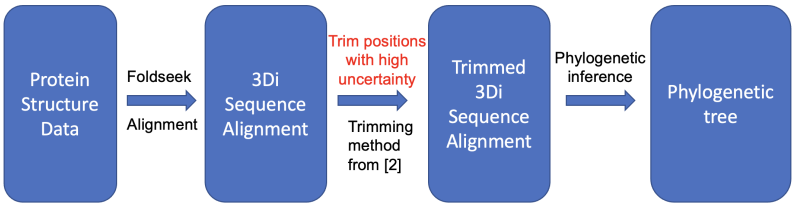


Figure 8: Diagram of structural phylogenetic analysis using uncertainty information

4 Results

The two downstream analyses are conducted using the two datasets mentioned above. For the first downstream analysis, we mainly focus on discovering which part of the protein has relatively high certainty. For the second downstream analysis, phylogenetic trees are inferred using our structural phylogenetic analysis method. We then use it to evaluate the performance of our structural phylogenetic analysis method.

4.1 Visualisation of uncertainty information

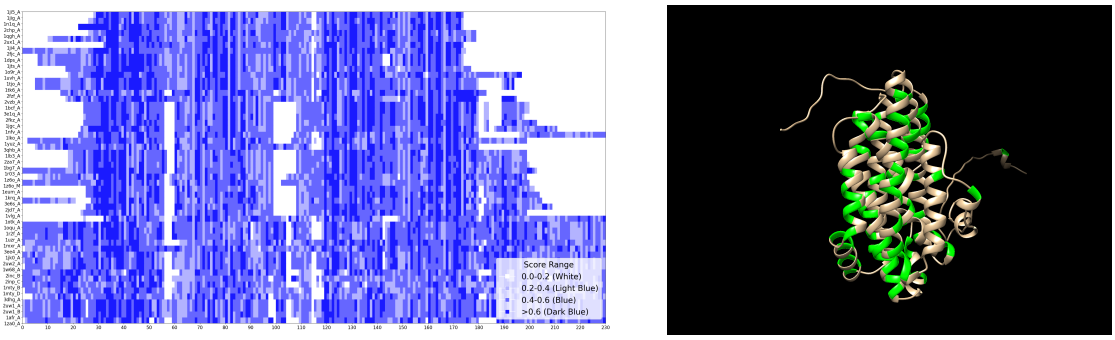


Figure 9: The left figure is the visualisation of the uncertainty information of 3Di alignment of ferritin-like superfamily, where each row corresponds to a protein, and each column corresponds to a position of aligned sequences. Also the darker the colour of the position is, the more certain it is. The right figure is the visualisation of the structure of a protein within ferritin-like superfamily. The coloured area consists of positions with high certainty.

Interpretation of uncertainty information for ferritin-like superfamily. Uncertainty information of 3Di sequences of the ferritin-like superfamily is visualised using the method above (Figure 9) and then reorganised to the form of an alignment. We can see from the visualisation that there are four distinct regions in the middle of the alignment that have high certainty (low uncertainty). Comparing this result with the visualisation of the structure of the example protein (from the ferritin-like superfamily), I found that these regions correspond to the positions of the protein structure that are located within the helix structures. These regions also overlap with the highly conserved regions found in [1] (Definition of the highly conserved position from [1]: If the 3Di character of a position of the amino acid sequence remains largely unchanged across different species, we refer to this position as a highly conserved position). The visualisation of highly conserved positions for the ferritin-like superfamily is in Appendix A.

Interpretation of uncertainty information for E. We also visualise the uncertainty information of aligned 3Di sequences of E (visualisations of 3Di alignments of E1 and E2 are in Appendix B). As shown in Figure 10, there are multiple regions with high certainty. The visualisation of the protein structure shows that the majority of these regions lie in the secondary structure (the diagrams of the structure of α-helix and β-sheet can be found in Appendix C).

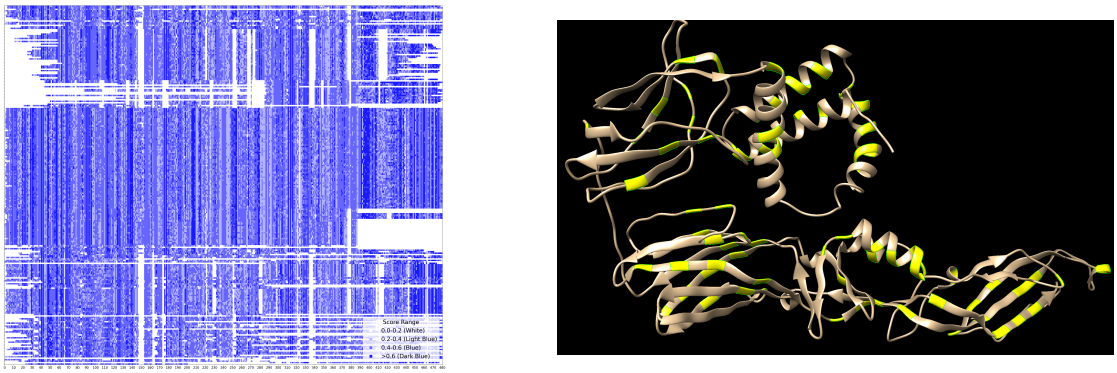


Figure 10: The left figure is the visualisation of the uncertainty information of 3Di alignment of E. The total number of columns of the alignment is much more the previous one, and this is because the structure of E is much more complex than the ferritin-like superfamily.

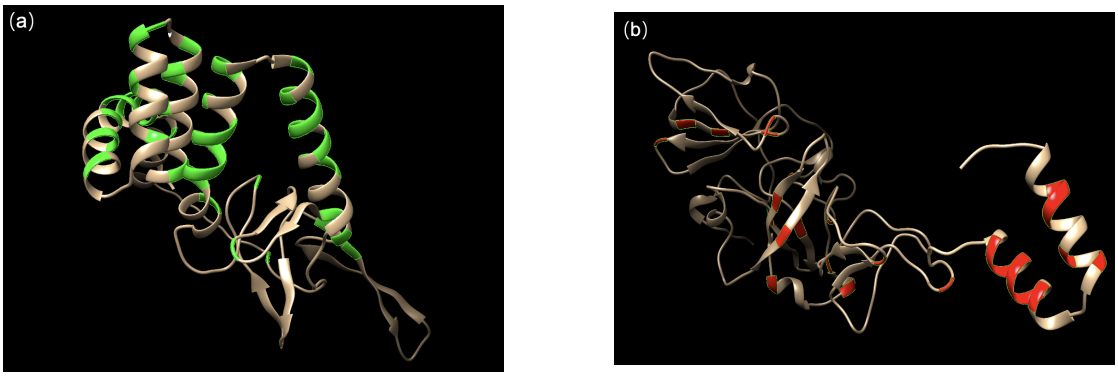


Figure 11: (a): Visualisation of high-certainty positions of an example protein for E1. (b): Visualisation of high-certainty positions of an example protein for E2

Findings. Figure 11 shows more visualisations of high-certainty positions of proteins (for E1 and E2). From all these visualisations (Figure 9-11 and Appendix B), we can find that the majority of the high-certainty positions of proteins lie in secondary structure, either α-helix or β-sheet (diagrams for them are in Appendix C). We can also find that Foldseek is confident in translating α-helix structure, in the sense that almost all α-helix have some positions assigned

high certainty. While for β-sheet, Foldseek seems not to be confident in translating this kind of secondary structure, as a lot of beta sheets don't have any positions with high certainty. One possible explanation is that the more flexible structure of β-sheets makes them harder to be translated by Foldseek.

4.2 Structural phylogenetic analysis using uncertainty information

We use the method mentioned in the section 3.3 to perform structural phylogenetic analysis using uncertainty information. Based on the distribution of the uncertainty of columns in alignment (Appendix D), we choose the cut-off values of uncertainty to be 0.4, 0.35, 0.4, 0.4 for ferritin-like superfamily, E, E1, and E2 respectively. These cut-off values are selected to remove columns that fall within the lower tail of the distributions (the columns that may provide noise) while not losing too much information (we will lose a lot of information if we cut off too many columns).

Evaluation metric from [4]. To assess the performance of our method, we compare the trees produced by our method with the one produced by the original structural phylogenetic analysis method. Following the approach described in [4], we used the variance of root-to-tip distances (RTT variance) as an evaluation metric for phylogenetic trees. Root-to-tip distances refer to the evolutionary distances from the common ancestor of all species (the root) to tips. Trees with smaller RTT variance tend to be better [4]. The detailed formula for calculating RTT variance is shown in Appendix E, and an example of using RTT variance to compare phylogenetic trees is in Appendix F.

Interpretation of the result. Using this evaluation metric, the inferred phylogenetic trees of our method are compared to that of the structural phylogenetic analysis method from [2] (Figure 12). After trimming out the columns that have high uncertainty, the RTT variance decreases or at least maintains the original level except for E1. This result shows that our method that trims high-uncertainty columns in the alignment could help remove columns that are less informative. An intuitive explanation for such a result is that, the visualisation of trimmed positions of the example E2 protein (Figure 13) shows that the majority of the positions trimmed by our method don't lie in secondary structure. Since the residues that don't lie in secondary structure contain less information (compared to those within a secondary structure) in general, they may provide noise if they are also considered to have high uncertainty. Therefore, it is good to trim them.

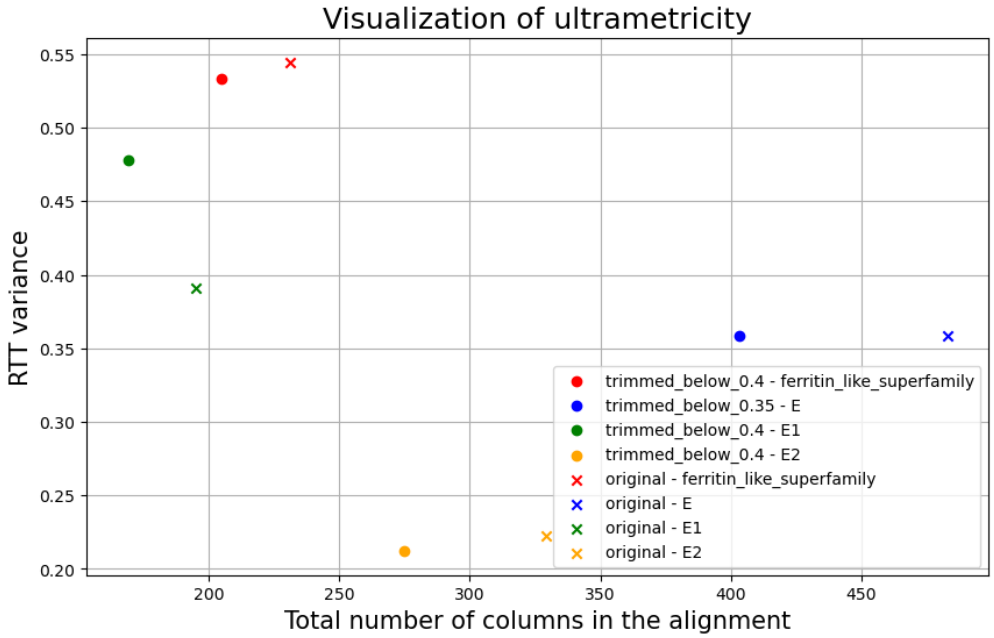


Figure 12: Comparison between the tree inferred by our method and the tree inferred by structural phylogenetic analysis method from [2]. The dot refers to the phylogenetic trees inferred by our method, the cross refers to the phylogenetic trees inferred by structural phylogenetic analysis method from [2].

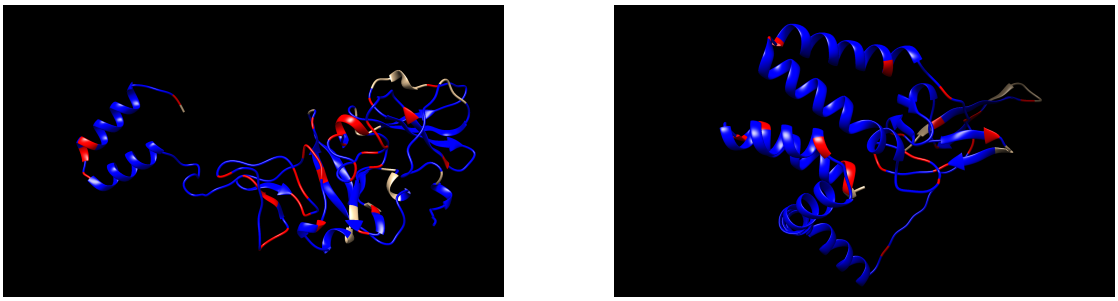


Figure 13: **Left:** Visualisation of trimmed positions of an example E2 protein. The red area consists of positions that are filtered out by our method but are originally kept by the trimming method from [2]. The blue area consists of positions that are kept by our method . The grey area consists of positions that are filtered out by the trimming method from [2]. **Right:** Visualisation of trimmed positions of an example E1 protein. The meanings for the red, blue, and grey area are the same as the left figure

Possible causes of high RTT variance for E1. For E1, from the visualisation of trimmed positions of the example protein (Figure 13), we can find that a large proportion of the trimmed positions belong to secondary structure. Since the positions within a secondary structure are believed to carry more information than others, trimming those positions may lead to losing information. Having quite a few positions belonging to secondary structures with low certainty shows that Foldseek seems not to be confident in translating the secondary structures for E1. This finding can also be seen from Figure 11(a), as a large proportion of β–sheet secondary structures of the example protein for E1 don't have any positions with high certainty. Therefore, for E1, it is no longer useful to trim the columns based on the uncertainty, as the uncertainty we calculate in this case may not be useful anymore.

5 Discussion

In this section, we will discuss potential improvements that can be made to our analysis.

Better scoring method. Even though we calculate a score for the 3Di letter of each position of the 3Di sequence and use it to represent the uncertainty of the position (step 4 and 5), we haven't assessed the performance of the scoring method. Besides, having a few low-certainty area that belongs to secondary structure for E1 may be due to an inappropriate scoring method. Hence, we could try different scoring methods (e.g., map the distance d_c to $z_c = -3d_c$), and compare their performance.

Incorporate the uncertainty information into phylogenetic inference. In our research, we incorporate the uncertainty information into structural phylogenetic analysis by trimming the alignment. We could also incorporate the uncertainty information directly into phylogenetic inference. Since we are using maximum likelihood phylogenetic inference, we can modify the likelihood calculation so that for each position of the aligned 3Di sequence, we consider the likelihood of the position belonging to different 3Di letters, rather than assuming it to be a single determined 3Di letter.

Conducting downstream analyses on more datasets. Since our downstream analyses are conducted on only two datasets, the generalisability of our findings in the first downstream analysis and the performance of our structural phylogenetic method on other datasets are unclear. Therefore, we need to run the two downstream analyses on more datasets.

Rescuing high-certainty columns. In our structural phylogenetic analysis method, we firstly

trim the columns that have a low proportion of data, and then trim the columns that have high uncertainty. Therefore, columns with high certainty but a low proportion of data will be filtered out by our method. However, these columns could possibly provide meaningful information, so it would be great if we could rescue them. Therefore, we will modify our method to retain columns in the alignment with high certainty but that are filtered out due to a low proportion of data by our current method.

6 Conclusion

In this project, we modified the Translation Method from [1]. For each position of an amino acid sequence, in addition to the 3Di letter for the position, we also calculate the uncertainty regarding the position. We also perform two downstream analyses, which are visualisation of uncertainty information and structural phylogenetic analysis with uncertainty information. Results suggest that most of the high-certainty positions of proteins are located in the secondary structure. More than that, Foldseek seems to be more confident in translating α-helix than β-sheets. We also find that trimming columns in the alignment that have high uncertainty before doing the phylogenetic inference could help remove columns that are less informative. However, this method (trimming high-uncertainty columns) might not work well for proteins whose secondary structures have quite a few low-certainty positions.

Acknowledgement. I am grateful to Dr. Heejung Shim (supervisor) for the instruction and supervision during the research program, and Yulin Wu and Dr. Jiangrong Ouyang for their assistance in the discussion about the research area and writing the report.

Appendix A Visualisation of highly conserved positions for ferritin-like superfamily

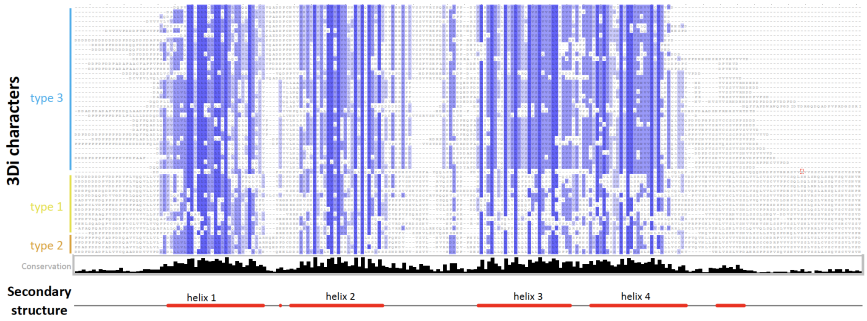


Figure 14: Visualisation of highly conserved positions for ferritin-like family from [2]. Characters matching the majority character in their respective columns are highlighted. Darker shades indicate a higher level of agreement, approaching 100%.

Appendix B Visualisations of 3Di alignments of E1 and E2

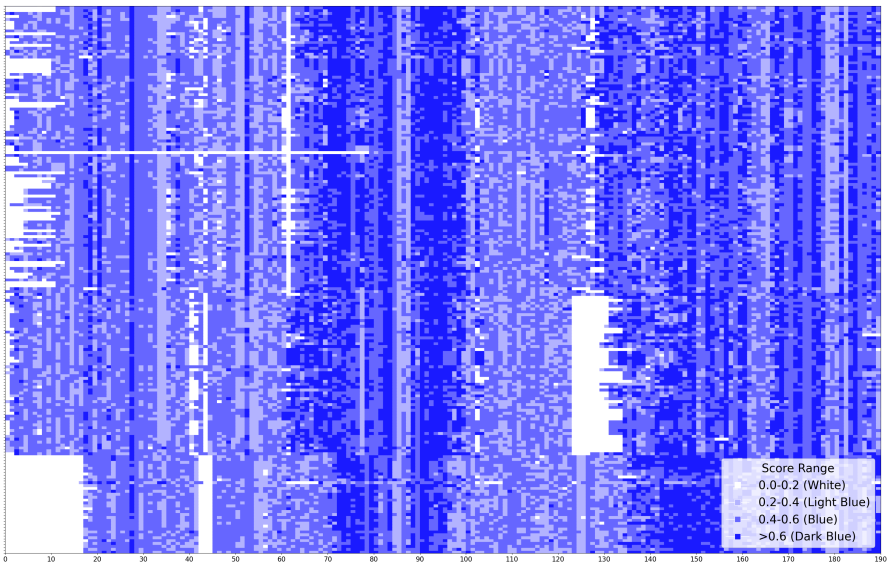


Figure 15: Visualisation of 3Di alignments of E1

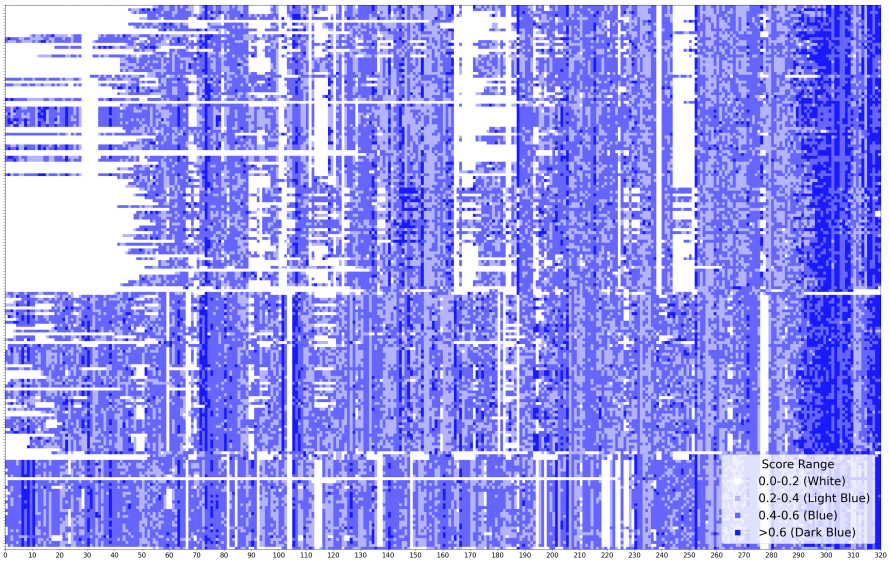


Figure 16: Visualisation of 3Di alignments of E2

Appendix C Diagrams of the structure of α-helix and β-sheet

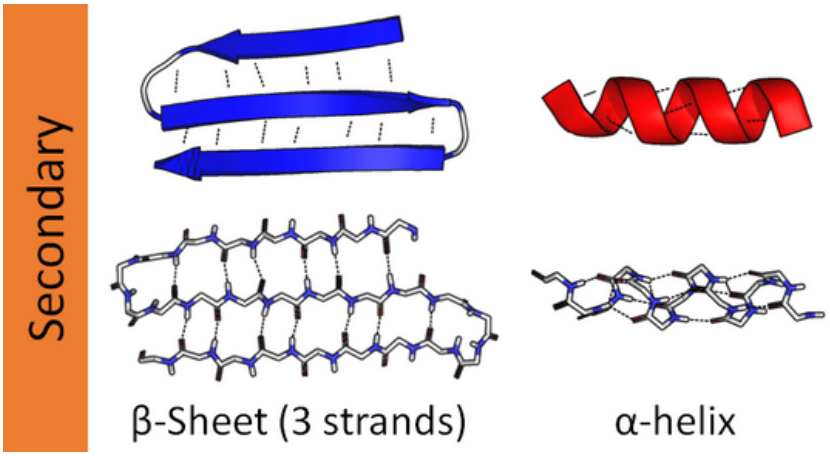


Figure 17: Diagrams of the structure of α-helix and β-sheet

Appendix D Distribution of uncertainty of positions in alignment

Appendix D.1 Ferritin-like superfamily

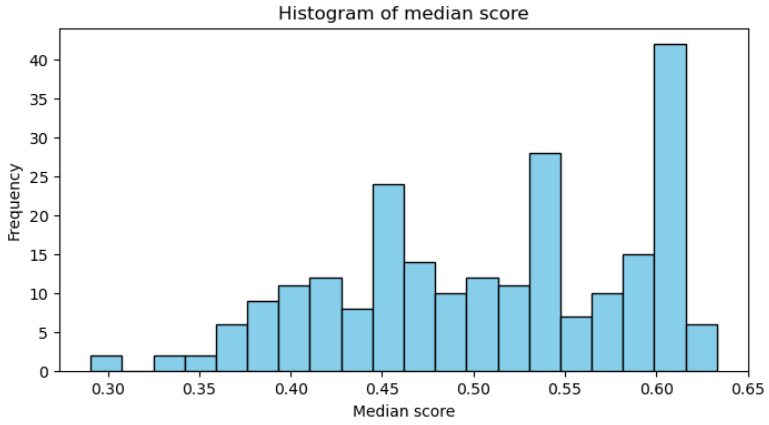


Figure 18: Distribution of uncertainty of positions in alignment for ferritin-like superfamily

Appendix D.2 E

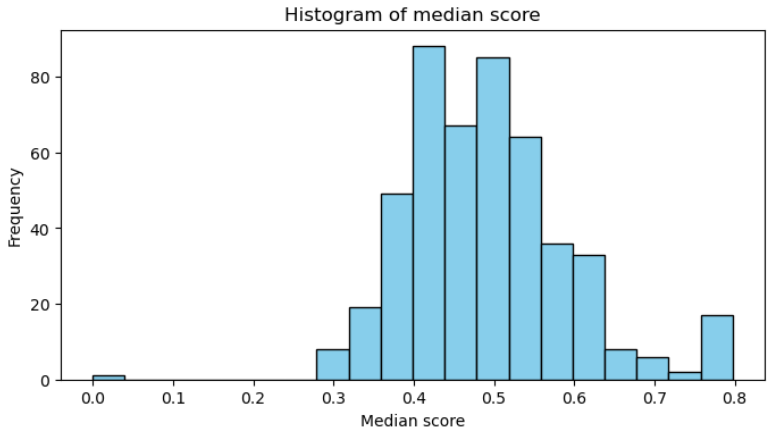


Figure 19: Distribution of uncertainty of positions in alignment for E

Appendix D.3 E1

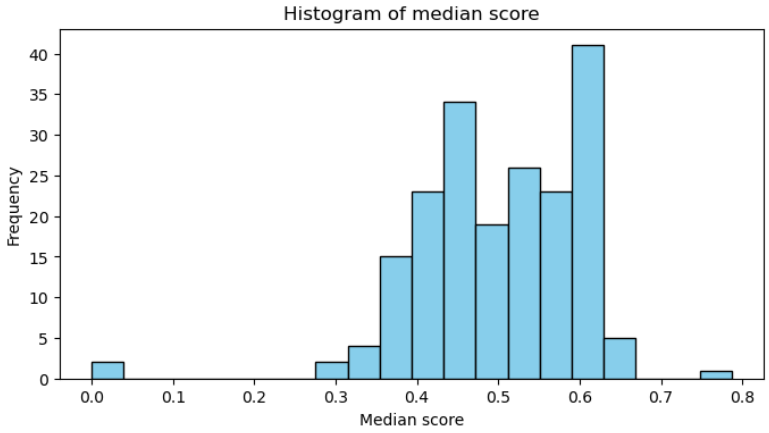


Figure 20: Distribution of uncertainty of positions in alignment for E1

Appendix D.4 E2

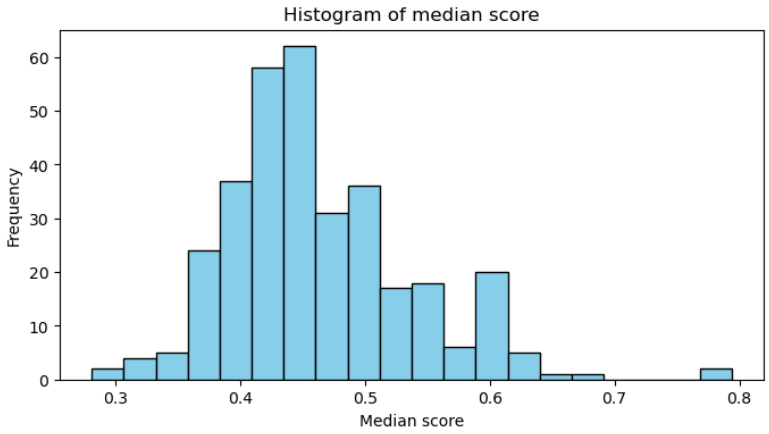


Figure 21: Distribution of uncertainty of positions in alignment for E2

Appendix E Formula for calculating RTT variance from [4]

$$E(rootdist) = \sum_{i=1}^{nleaves} dist(l_i, root)/nleaves$$

$$S_{norm}(rootdist) = \sum_{i=1}^{nleaves} (dist(l_i, root)/E(rootdist) - 1)^2/(nleaves - 1)$$

Figure 22: Formula for calculating RTT variance from [4]

Appendix F Example of comparing phylogenetic trees based on RTT variance

In this example, the top phylogenetic tree has smaller RTT variance compared to the bottom one. Hence we can say that the top phylogenetic tree is better than the bottom one.

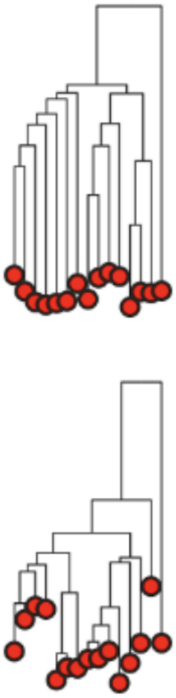


Figure 23: Example of phylogenetic trees. They are inferred by different methods but using the same data.

References

[1] van Kempen, M., Kim, S.S., Tumescheit, C. et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* vol. 42, pp. 243–246 (2024). https://doi.org/10.1038/s41587-023-01773-0

[2] Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534 (2020).

[3] Puente-Lelievre, C., Malik, A. J., Douglas, J., Ascher, D., Baker, M., Allison, J., ... Matzke, N. (2023). Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *BioRxiv*, doi: 10.1101/2023.12.12.571181

[4] Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., Dessimoz, C. (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *BioRXiv*, doi:10.1101/2023.09.19.558401

[5] Mifsud, J.C.O., Lytras, S., Oliver, M.R. et al. Mapping glycoprotein structure reveals Flaviviridae evolutionary history. *Nature* vol. 633, pp. 695–703 (2024). https://doi.org/10.1038/s41586-024-07899-8