Title: Adding 3Di characters to amino acid datasets can improve resolution, but the effect is weaker in shorter and alpha-helical proteins such as histones

Authors

Fullmer, Matthew S.[1]
Puente-Lelievre, Caroline[1]
Matzke, Nicholas J. [1]

[1] School of Biological Sciences, University of Auckland, Auckland, New Zealand

Acknowledgements

The work of MF was supported by 22-UOA-250. The work of NJM, CPL, and their lab group was primarily supported by HFSP grant RGP0060/2021, as well as ARC DP240100462, New Zealand Royal Society RDF 21-UOA-040, Marsden Grant 18-UOA-034, and University of Auckland, Faculty of Science Research Development Fund, FoS RDF #3732317.

The authors acknowledge helpful discussions with Jordan Douglas, Ashar Malik, and Anthony M. Poole. The authors also wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support and training services as part of this research. New Zealand's national facilities are provided by NeSI and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL https://www.nesi.org.nz.

Google docs version:
https://docs.google.com/document/d/1y5Y70d8p3RuvKdcRXmu88HZ6nXmZcAlmWkBQ_ZFr4tY/edit?tab=t.0

Supplementary Material Links:
https://docs.google.com/document/d/1cHpADTxmbqIrQSGIQ5KC-qdd6nzn_3BFqNeE9QAhxLo/edit?usp=sharing

https://github.com/nmatzke/FullmerMatzke25/blob/main/SupplementalMaterials_sansBEAST.zip

GitHub repository:
https://github.com/nmatzke/FullmerMatzke25

ABSTRACT
The recent introduction of Foldseek's 3Di character alphabet to encode 3D protein structure has opened up new possibilities for structural phylogenetics. The high conservation of structure relative to amino acid sequences raises the possibility of better resolving the deepest branches of life's evolution. As 3Di is a 20-letter alphabet of discrete characters, they are readily treatable with off-the-shelf algorithms for model-based phylogenetic inference and application of established methods such as bootstrapping. However, it remains to be seen if 3Di phylogenies are broadly more resolved than sequence-based phylogenies. We present data using samples from ten protein superfamilies showing that 3Di combines with sequence to produce better resolved phylogenies than either sequence or 3Di alone. We also show that information-theoretic measures, applied to superfamily alignments, significantly correlate with resolution in phylogenies derived from these alignments. Further, we identify the proportion of alpha helices in proteins as a major driver in reducing the information carried by 3Di character alignments, explaining the poorer performance of 3Di characters on some datasets with highly-conserved structure, such as Histone-fold proteins. Our results both provide encouragement for the further use of 3Di to address challenging questions in deep history, as well as sound a note of caution about which proteins it is suitable for.

INTRODUCTION
It has long been understood that protein shape may contain phylogenetic information [1]. Protein 3D structure is typically more conserved than amino acid (AA) sequences [2,3]. Attempts have been made to use structural distance measures for neighbor-joining phylogenetic inference (e.g. Lundin et al., 2012 [4]; see Puente-Lelievre et al. 2024a [5] for a review), however these approaches lack an evolutionary model and the associated advantages of probabilistic approaches, such as uncertainty measures. An extension of the structural distances approach [6] uses simulated conformational variability to create a pseudo-support tree set, however it is computationally very expensive. An alternative approach is to encode protein structure as a series of discrete characters, an approach which became possible with the protein structure alphabet developed as part of the Foldseek program [7]. The alphabet is termed "3Di" as it encodes highly conserved 3D tertiary interactions in a protein. The Foldseek program translates any input protein structure model into a series of 3Di characters, one for each residue in the protein backbone. The 3Di alphabet conveniently has 20 characters, enabling usage of BLAST and other standard bioinformatics algorithms for homology search. As the 3Di characters are highly conserved, Foldseek is able to conduct high-sensitivity deep homology searches on databases of structure models in and beyond the "Twilight Zone" of highly decayed sequence similarity. This capability is particularly powerful with the availability of databases of AlphaFold-derived structural predictions.

Beyond deep homology search, having a discrete 3Di structural alphabet that is more conserved than AAs opens new possibilities for structural phylogenetics [5,8–10]. We focus on use of 3Di characters to enhance model-based phylogenetic inference, making use of highly-conserved protein structure data while allowing all the advantages of probabilistic methods, including statistical model comparison and standard measures of statistical confidence such as bootstrapping. While early reports are promising [5,11], the suitability of 3Di characters for phylogenetics needs further exploration.

One of the attractions of structural phylogenetics is the prospect of better resolving the earliest branches of protein evolution. These early branches represent some of the earliest evolution in life's history. In many protein superfamilies, they would predate the Last Universal Common Ancestor (LUCA). We investigate whether 3Di characters are likely to

help with such problems by comparing various measures of phylogenetic resolution on datasets of ancient protein families that have diverged into the Twilight Zone, when making use of AA, 3Di, or AA+3Di character data.

To begin to explore this question, we estimated alignments and phylogenies for 10 protein superfamilies, comparing AA-only, 3Di-only, or AA+3Di datasets with several measures of statistical resolution. We also looked for differences in resolution between deep and shallow branches. We observe that, across the 10 protein superfamilies, there is a substantial and statistically significant increase in Tree Certainty and related measures of phylogenetic resolution for AA+3Di datasets over AA-only or 3Di-only datasets. This suggests that 3Di characters are a useful addition in the Twilight Zone, even though not all deep divergences are resolved. However, the results vary by superfamily, with Histone-fold proteins, which are structurally conserved and probably present in LUCA [12–14], not showing substantial resolution improvements with the addition of 3Di characters.

We expanded our analysis to examine possible correlates that might be responsible for the variation in resolution between superfamilies. Our results suggest that proteins containing large proportions of alpha-helices, such as Histone-fold proteins, may result in drops in the information content of the 3Di-coded alphabet, and a commensurate drop in phylogenetic resolution.

METHODS

Histone data. The initial histone dataset consists of 36 proteins selected from SCOP superfamily a.22.1 and proteins identified as containing the Histone-fold in Structome [15]. They also represent a subset of those used in Malik (2018) [16]. See **Table S1** for full listing.

SCOP superfamilies. SCOP 1.75 supplied the categorization information [17,18]. Superfamilies were initially selected with the goal of evenly sampling the 4 major classes containing alpha helix and beta sheet-containing proteins (alpha-only, beta-only, a+b, and a/b). Another goal was to sample across a variety of size ranges, such that we could interrogate the correlation of sequence length and secondary structure composition with phylogenetic resolution. PDB files for each selected superfamily were downloaded from the PDB database. Each prospective superfamily was processed in a manner based on the filtering and clustering described for STRUCTOME [15]. All proteins below 50 AAs and above 350 AAs were excluded. UCLUST was used to prune the dataset, using centroid clustering at 90% identity to identify a single sequence to represent a group of similar sequences [19], making phylogenetic inference for each superfamily more computationally tractable. However, clustering also meant that there was no guarantee that each family in the superfamily would have a representative, or that component families would have equal or proportional representation. From the superfamilies that, after processing, retained at least 4 remaining members, 9 were chosen: three “a,” and two each of “b/c/d.”

Secondary structure annotation. DSSP, as implemented in XSSP [20,21] was used to quantify the secondary structure proportions of each sampled sequence. All three helix types (H, G, I) were combined into a single proportion. Likewise, both beta-sheet types (B, E) were also combined into a single proportion. The metric "total secondary structure proportion" represents the sum of helix and sheet proportions.

Tree Certainty Measures. Uncertainty of tree collections (e.g. bootstrap support sets) was measured using Internode Certainty (IC). IC represents a quantification of the level of disagreement in a support set for a particular node in a phylogeny; IC decreases as the bootstrap value drops and as the frequency of the most frequent dissenting topology increases. Tree Certainty (TC) is the average Internode Certainty score across all internal branches [22] (see a brief summary of the terminology and metrics in **Supplemental Material**). These scores were calculated by mapping support tree sets against reference trees, as implemented in RAxML v8.1 [23]. The related Internode Certainty All (ICA) score was also calculated. The ICA considers not only the most frequent dissenting topology at a branch, but also whether dissenting trees agree or disagree with each other. I.e., a bipartition supported by 51% of bootstrap trees would score much higher if the conflicting samples represented 49 different alternatives than it would if there were only a single alternative. The Tree Certainty All (TCA) score is the average of ICA values across internal branches, representing an assessment of overall conflict in the support set.

Mean bootstrap values. Bootstrap values across the branches of each tree were extracted using custom R [24] scripts, using the R library *ape* [25]. Scripts and data are available in links supplied in Supplemental Material.

Amino acid (AA) phylogenetic inference. The ClustalO algorithm as implemented in SeaView [26] created initial histone AA alignments. However, for the final analyses presented here, Foldmason [9] was employed to create structurally-informed AA superfamily alignments. Maximum likelihood phylogenetic interference was performed in IQ-TREE2, with substitution models selected by IQ-TREE's model finder using the BIC criterion [27].

3Di phylogenetic inference. Foldmason was employed to create structurally-informed joint 3Di and AA alignments [9]. Maximum likelihood interference of phylogenies was done in IQ-TREE2 using the Foldseek 3Di rate matrix as in in Puente-Lelievre et al. (2024a), and using IQ-TREE's model finder to choose base-frequency and among-site rate variation models [27].

Partitioned analysis of AA+3Di datasets. Partitioned analyses were performed with IQ-TREE2 with the "-p" option [27]. The models for each partition were determined as above for the AA-only and 3Di-only datasets.

Bayesian partitioned analysis. Bayesian analyses in BEAST [28,29] may have some advantages over undated molecular phylogenetics performed by IQ-TREE. While absolute dating of divergence times in ancient protein superfamilies is extremely challenging and is not attempted here, BEAST can be used in a relative-dating framework where no claim is made about absolute divergence times. An advantage of doing so is that BEAST relative-dating analyses automatically perform (1) model-averaging over various possible sequence evolution models; (2) rooting of the phylogeny (all sampled posterior trees are rooted, as are the output summary trees); and (3) Bayesian measures of uncertainty (posterior support values for each branch in the summary tree, and node-height uncertainty for all internal nodes). Model setup was as above for IQ-TREE, but using the BEAST package OBAMA [30] for model averaging, and the Gammaspike model [31] of branchwise rate variation to test for smooth or sudden changes in rate along branches. FoldBeast (https://github.com/jordandouglas/FoldBeast/) was used with the 3Di data. AA and 3Di-only MCMC runs collected 10 million samples, which on these relatively small datasets was

sufficient to obtain 200 ESS for all summary statistics. The Linked tree, using AA+3Di ran for 30 million steps. The run using the GammaSpike model ran for 15 million steps.

3Di rate matrix. The rate matrix from [5] was used with all 3Di alignments. Four superfamilies (a.35.1, a.38.1, a.39.2, and b.122.1) were tested using the alternate rate matrices presented in [11]. Their average bootstrap percentages were both lower (51% vs 45%, and 49%). Linear regression found the difference to be weakly significant (p = 0.018 and 0.0299). This indicated resolution would be better with the Foldseek-derived rate matrix (**Table S2**).

Effective Alphabet Size. Raw character frequencies from alignments were obtained from IQ-TREE2 [27]. Estimated frequencies for the Histone-fold dataset were obtained from BEAST posteriors [28]. Shannon entropies were calculated for each alignment, summed and converted into an effective alphabet size [32].

$$Effective\ alphabet\ size = e\ \hat{}\ - \sum_{base=1}^{20} Freq_{base} * Ln(Freq_{base})$$

sCF. Site concordance factors were calculated in IQ-TREE2 with the -blfix option [27,33]. For3Di alignments, the model was fixed to the3Di rate matrix. IQ-TREE was allowed to re-estimate AA models.

Correlogram. The R function *cor* was used to calculate Pearson correlations. *cor.test* was used to calculate significance. Linear models were calculated with R's *lm* function. The R library *corrplot* [34] visualized the correlogram. Only correlations with significance p =< 0.01 were displayed.

Old/Young nodes. In order to judge some nodes to be relatively older than others, a rooted tree is needed. While BEAST automatically estimates the root position, IQ-TREE does not. Therefore, for IQ-TREE phylogenies, we used midpoint rooting in R package *phytools* [35], which places the root in the middle of the longest path through the tree. Midpoint rooting is known to be uncertain and can be affected by artefacts such as long branches, but outgroups are typically not available for such ancient divergences, and at least midpoint roots plausibly place the root far from the tips of the tree, typically separating major named groups (e.g. protein families) into separate clades. For our purpose, which was merely to roughly distinguish statistical resolution at deeper versus shallower nodes, this was adequate. For each node in a tree, the root-to-node distance was calculated with R library *castor* [36], 2018). The median of these root-to-node distances was calculated for each tree, and nodes closer to root than median were considered "deep," while those above the median were "young." Mean support values are calculated as above.

RESULTS

Comparison of AA-only, 3Di-only, and AA+3Di analyses across 10 superfamilies

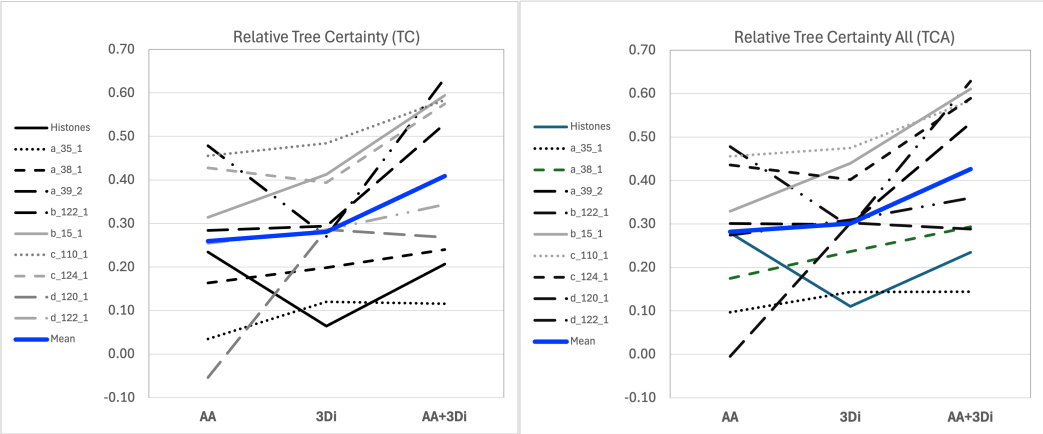

Figure 1. Comparison of full phylogenies of AA-only, 3Di-only, and AA+3Di analyses across all 10 superfamilies. A) Relative Tree Certainty; B) Relative Tree Certainty All.

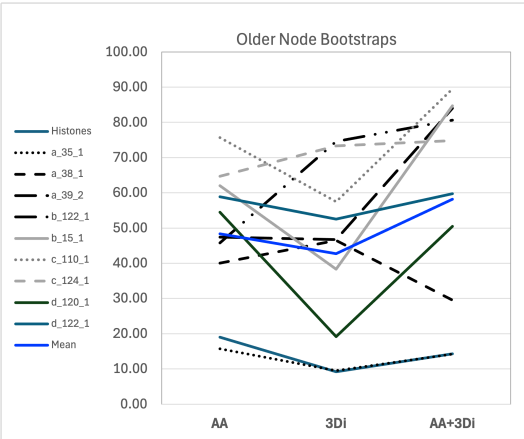

Figure 2. Comparison of only deeper nodes of AA-only, 3Di-only, and AA+3Di analyses across all 10 superfamilies. Measure of resolution is average bootstrap value per node.

Overall, combining AA+3Di offers significant improvements in overall tree resolution ($p <= 0.007$) (**Figure 1**). The AA+3Di partitioned dataset gained ~9% in average bootstrap percentage over AA-only datasets, and ~0.15 TC/TCA. All of these gains were statistically significant (bootstrap $p = 0.021$, TC/TCA $p = 0.007$ and $p = 0.005$). Equally promising, the improvements are not only going to the shallow nodes. The deeper nodes using AA+3Di are also more resolved than AA or 3Di alone (**Figure 2**). The deeper nodes gained 10% and 15% in mean bootstrap percentages. The AA+3DI dataset's bootstrap percentages were significantly higher than the 3Di-only dataset ($p = 0.037$), and slightly higher than AA-only (non-significantly, $p = 0.098$). Although the improvement from adding 3Di characters is stronger on younger nodes, overall, 3Di appears to be improving results on average. For a full record of superfamily statistics, see **Table S3**.

However, these average improvements seem to not apply to all datasets, as can be observed in Figures 1&2, where several datasets show marginal or no improvement from adding 3Di characters. In particular, the Histone superfamily, although it initially seemed promising as a candidate for structural phylogenetics, given the highly conserved core structure, did not

show much improvement from adding 3Di characters. In order to explore the reasons why, we explored the Histone superfamily analysis in more depth.

The Histone-fold dataset's sequences are in the twilight zone. The overall sequence similarity across the alignment is only 19.8%, at the lower bound of the canonical twilight zone [2,37]. Further, IQ-TREE2 estimated pairwise ML substitution rates as high as 9 substitutions per site, with an average of 4.13 substitutions per site. (**See file in Supplemental Materials)**. An amino acid alignment of this nature can probably be assumed to contain limited phylogenetic information. However, the distribution of distances is not uniform. Over 93% of pairwise comparisons were under 30% identity. And 69% were 20% or less. There are multiple clusters, representing 6.7% of the data, with pairwise similarities greater than 40%. Indeed, ~2% of comparisons were over 90% similarity. This suggests that a tree would likely have some resolution for at least a handful of clans.

Inference method	Data	Alignment method	Model	Support set	TC	TCA	Avg. BS / node	Nodes 70%+	Avg. posterior / node	Nodes 0.95+
IQ-TREE2	AA	ClustalO	IQ-TREE2 MFP	iqtree bootstraps	0.2507	0.2880	47.79	11		
IQ-TREE2	AA	foldmason	IQ-TREE2 MFP	iqtree bootstraps	0.2346	0.2795	46.30	14		
IQ-TREE2	3di	foldmason	3di Rate matrix	iqtree bootstraps	0.0643	0.1103	30.85	6		
IQ-TREE2	3di + AA	foldmason	MFP + 3di Rate Matrix	iqtree bootstraps	0.2066	0.2349	46.30	12		
BEAST	AA	foldmason	OBAMA	BEAST posterior	0.4195	0.4659			0.5948	13
BEAST	3di	foldmason	FoldBeast	BEAST posterior	0.4990	0.5128			0.7157	14
BEAST	AA + 3di	foldmason	OBAMA + FoldBeat	BEAST posterior	0.5490	0.5656			0.7245	16

Table 1. Histone-Fold Tree Statistics. Trees inferred with a 3Di-only dataset are worse (maximum likelihood) or only slightly better (Bayesian) than amino acids alone. Partitioned analyses are also either worse (ML) or only slightly better (Bayesian) than the amino acids alone. While Bayesian analyses provided marginally more well resolved nodes, these are concentrated in the crown. In this superfamily, while AA+3Di analyses improve on AA-only analyses in the Bayesian analysis, this is not the case in ML analyses.

Weak resolution at deeper branches of the Histone superfamily tree. The ClustalO aligned tree largely matched these expectations (**Table 1**). Most branches show extremely weak bootstrap support values. However, several clans, mostly just shallow groups of a couple tips, receive 70%+. The net is a mean of ~47.8% per branch. Tree certainty scores calculated from the bootstrap set were 0.251 and 0.288 per branch (IC & TCA, respectively). The Foldmason-aligned tree values were slightly lower, but returned slightly more well-resolved nodes.

Against expectations, for the Histone-fold superfamily, the 3Di-only phylogeny was notably less resolved than the AA (**Table 1**). It also showed no improvements in the deeper nodes (9.18% BS, vs 19% BS for AA). Combining AA+3Di resulted in roughly equivalent resolution to AA-only phylogenies. For an expanded Histone-fold analysis, see **Supplemental Materials**.

The poor performance of 3Di with the Histone-fold proteins, particularly in regards to the deeper nodes, prompted us to ask if 3Di is actually evolving more smoothly than the AA sequence. We employed a GammaSpike model in BEAST on the AA+3Di dataset, on the suggestion that allowing the rate of evolution to change at branching events might fit better (Jordan Douglas, pers. comm.). Against expectation, the AA data did not reject the GammaSpike model, although it also did not strongly accept it. Equally against assumptions,

the 3Di data accepted the model. This suggests the AA data may be evolving in a smoother, more clock-like, manner than the 3Di data.

Histones' brevity and alpha-helix-heavy composition reduce phylogenetic resolution. Histone-fold proteins are short with an average length of only 83.9 amino acids. Having more sequence length simply tends to carry more information [38]. Average sequence length was significantly positively correlated with our measures of phylogenetic resolution among the 10 superfamilies (**Figure 3**).

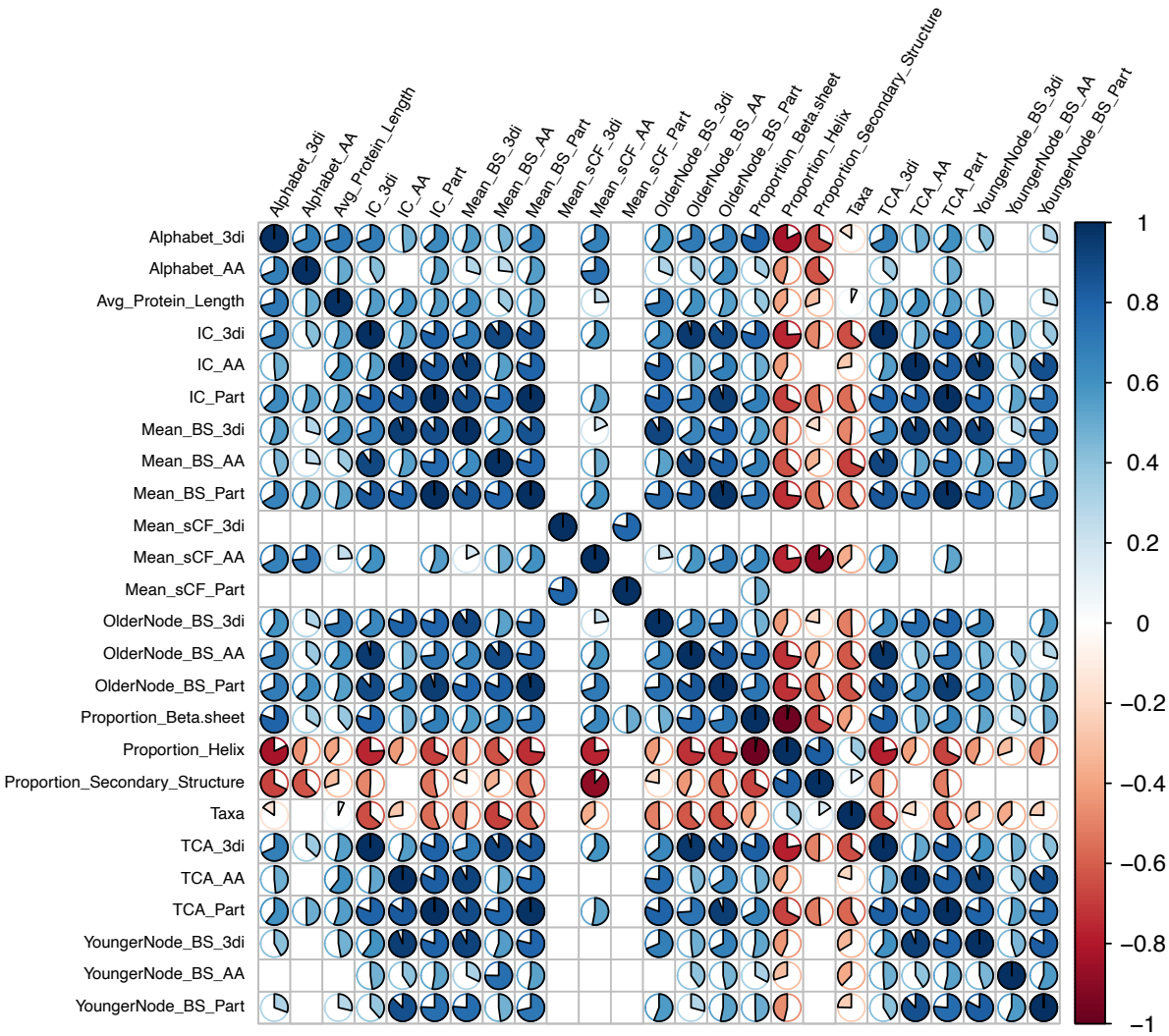

Figure 3. Pearson correlations of protein superfamily characteristics and phylogeny support values. Display cutoff is p = 0.01. Smaller alphabet size is significantly correlated with reduced resolution. Similarly, the proportion of the protein composed of helices is correlated with smaller alphabets and reduced resolution.

The Histone-fold set is also extremely helix-heavy. They are composed of an average of 69.2% percent helices. As a result, the 3Di alignments are dominated by only two character states (V, 50.9% and D, 15.7%); comprising almost exactly two-thirds of all residues. In contrast, it takes their 9 most common amino acids to sum to the same proportion.

	3di Effective Alphabet	AA Effective Alphabet
Empirical	5.493	16.530
BEAST estimated	8.832	17.303

Table 2. Histone-fold dataset effective alphabet sizes. The alphabet of3Di characters is substantially smaller than that of the amino acids.

Smaller alphabet sizes significantly correlate with lower support values. The combination of short and helix-heavy proteins are information-poor alignments - particularly notable in the 3Di data. Calculating entropies to estimate the effective alphabet sizes of the empirical and BEAST-estimated character frequencies reveals that the 3Di sequences are functionally using an alphabet of only 5.5 (empirical) or 8.8 (BEAST estimated) characters while the AA are using a 16.5-17.3 sized alphabet (**Table 2**). Pearson correlations (**Figure 3**) and linear regressions (**Figure 4**) both found this to be a significant relationship across the superfamilies.

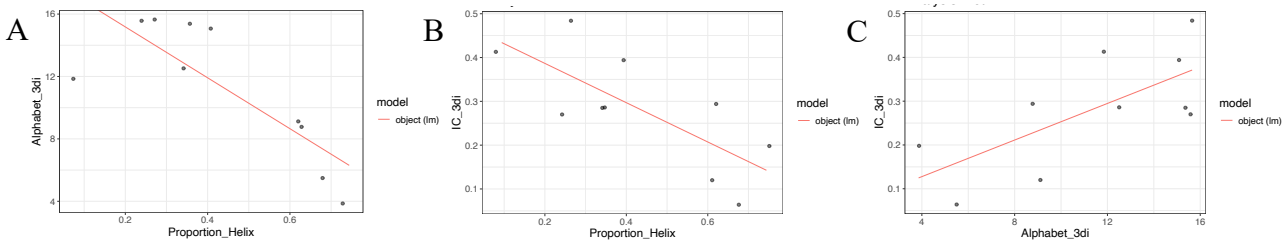

Figure 4. Linear Regressions of relationships between Phylogeny resolution (IC (TC) 3Di), 3Di Alphabet size, and Proportion of Helices. A) 3Di Alphabet vs Proportion of Helix. B) 3Di TC values vs Proportion of Helices; C) 3Di IC (TC) values vs 3Di Alphabet sizes. Increasing proportion of helices appears to drive down the 3di alphabet size. In turn, smaller 3di alphabets appear to drive down phylogenetic resolution.

Helix proportion is significantly anti-correlated with both alphabet size and support values. Helix proportion was by far the largest anti-correlation with effective alphabet size. This was far more emphatic for the 3Di alphabets than AA (although the link is strong in both). This aligns with the conclusion that helix proportion is a driver of reduced alphabets (**Figure 4**).

We were concerned that this result could be a bias from including four alpha-only superfamilies and only 2 of each other fold type. We tested this by removing one, two, three, or all of the alpha-only superfamilies and calculating linear regressions between helix proportion and alphabet size. While the effect size dropped, including only two still shows a significant relationship, indicating that this is a real factor (**Table 3**).

Alpha-only superfamilies	Pr > |t|	Significance	Adjusted R-squared
4	0.00342	**	0.63760
3	0.01253	*	0.55830
2	0.02679	*	0.51720
1	0.10550		0.32530
0	0.33000		0.04341

Table 3. Linear regressions of helix proportion vs 3Di alphabet size. Rows vary the number of alpha-only superfamilies in the regression. When the number of alpha-only

superfamilies is reduced to 2, the same number of superfamilies in the three other fold-types, the relationship remains strong and significant.

Conversely, beta-sheet proportion is positively correlated with alphabet size, as well as higher average support values. As both alphabet sizes are inversely correlated with support values, it seems reasonable to suggest that enrichment in helices, and perhaps paucity of beta-sheets, is a major driver behind poor resolution in Histone-fold proteins and other alpha-rich superfamilies.

DISCUSSION

We present evidence that 3Di combined with sequence both improves overall resolution, as well as resolution in deeper nodes when compared to sequence or 3Di alone. This bolsters the promise of using 3Di in model-based phylogenetics [5,39].

We also present evidence that superfamilies with shorter protein lengths and enrichment in alpha-helices are less suitable for 3Di-based phylogenetics. These two factors combine to reduce the effective alphabet size of their alignments, thereby reducing the amount of information conveyed. These reductions in information content significantly correlate with reduced phylogenetic resolution. These results offer a note of caution, suggesting that researchers doing deep phylogenies should calibrate their expectations by considering the characteristics of proteins they are studying before deploying 3Di characters.

Delving more deeply into the Histone-fold result, we show that the 3Di method performed poorly on this superfamily. Sequence-only phylogenies showed stronger resolution than 3Di-only inferences as well as AA+3Di partitioned analysis. Additionally, both 3Di-only and partitioned analyses failed to show any improvement in resolving deeper nodes. Perhaps most interesting is that the Histone 3Di dataset accepted, albeit by a narrow margin (0.936 vs recommended acceptance threshold of 0.910), the GammaSpike model while the amino acid data failed to accept or reject the model. This would suggest that 3Di has a less clock-like rate of evolution than the amino acids. This is unexpected, as we would anticipate the 3Di alphabet to be capturing the slower changing information of protein shape. Additionally, 3Di-derived measures have been argued to evolve in a more clock-like fashion than sequence data [8]. We are unsure why we would see this. Our speculation is that the Histone-fold dataset has such poor character state diversity and is so dominated by just two states (L and V, both common in 3Di representations of alpha-helices) that very few changes happening along a branch represent a major enough departure for the model to perceive a change in the rate of evolution. Further investigation in other superfamilies is essential to understanding if this is a generalisable result or a one-off created by the very low information content in the alignments.

In this work, we focused on the resolution of inferred phylogenies. We did not focus on the actual topologies the methods produced, as we do not know the “true” history of any of these ancient superfamilies. “Ground truthing” a new set of phylogenetic characters against an already known phylogeny is possible in simulations and in certain laboratory experiments, but these are difficult options to enact as a valid test would have to include realistic protein structure evolution over deep time, as well as sequence divergence into the Twilight zone. However, these problems not unique, as similar problems are present for many sources of phylogenetic characters (fossil morphology, developmental characters, behaviour, etc.).

Nevertheless, even in the absence of ground-truthing, being able to deploy methods from the suite of model-based statistical phylogenetics, such methods to measure uncertaint (bootstrapping, posterior probabilities, etc.), constitutes a major advantage, enabling statistical studies that allow us to at least quantify the precision that structural phylogenetics may achieve, and compare that to more traditional methods.

References

Alva, V., and Lupas, A. N. (2018). Histones predate the split between bacteria and archaea. *Bioinformatics*, 1–5. doi: 10.1093/bioinformatics/bty1000

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., et al. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res* 36, 419–425. doi: 10.1093/nar/gkm993

Bouckaert, R. R. (2020). OBAMA: OBAMA for Bayesian amino acid model averaging. *PeerJ* 8, 1–15. doi: 10.7717/peerj.9460

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 15, 1–28. doi: 10.1371/journal.pcbi.1006650

Doolittle, R. F., Feng, D. F., Johnson, M. S., and McClure, M. A. (1986). Relationships of human protein sequences to those of other organisms. *Cold Spring Harb Symp Quant Biol* 51 Pt 1, 447–455. doi: 10.1101/sqb.1986.051.01.054

Douglas, J., Bouckaert, R., Harris, S. C., Carter, C. W., and Wills, P. R. (2024). Evolution is coupled with branching across many granularities of life. *bioRxiv*, 2024.09.08.611933. Available at: https://www.biorxiv.org/content/10.1101/2024.09.08.611933v1%0Ahttps://www.biorxiv.org/content/10.1101/2024.09.08.611933v1.abstract

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Garg, S., and Hochberg, G. K. (2024). A general substitution matrix for structural phylogenetics. *bioRxiv*.

Gilchrist, C. L. M., Mirdita, M., and Steinegger, M. (2024). Multiple Protein Structure Alignment at Scale with FoldMason. doi: 10.1101/2024.08.01.606130

Gouy, M., Tannier, E., Comte, N., and Parsons, D. P. (2021). "Seaview Version 5: A Multiplatform Software for Multiple Sequence Alignment, Molecular Phylogenetic Analyses, and Tree Reconciliation BT - Multiple Sequence Alignment: Methods and Protocols," ed. K. Katoh (New York, NY: Springer US), 241–260. doi: 10.1007/978-1-0716-1036-7_15

Hocher, A., Laursen, S. P., Radford, P., Tyson, J., Lambert, C., Stevens, K. M., et al. (2023). Histones with an unconventional DNA-binding mode in vitro are major chromatin constituents in the bacterium Bdellovibrio bacteriovorus. *Nat Microbiol* 8, 2006–2019. doi: 10.1038/s41564-023-01492-x

Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. (1999). SCOP: A structural classification of proteins database. *Nucleic Acids Res* 27, 254–256. doi: 10.1093/nar/27.1.254

Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314. doi: 10.1080/10618600.1996.10474713

Illergård, K., Ardell, D. H., and Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* 77, 499–508. doi: https://doi.org/10.1002/prot.22458

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: https://doi.org/10.1002/bip.360221211

Kim, D., Park, S., and Steinegger, M. (2025). Unicore Enables Scalable and Accurate Phylogenetic Reconstruction with Structural Core Genes Open Access. *GBE*. doi: https://doi.org/10.1093/gbe/evaf109

Kobert, K., Salichos, L., Rokas, A., and Stamatakis, A. (2016). Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Mol Biol Evol* 33, 1606–1617. doi: 10.1093/molbev/msw040

Lake, J. A., Henderson, E., Oakes, M., and Clark, M. W. (1984). Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences* 81, 3786–3790. doi: 10.1073/pnas.81.12.3786

Louca, S., and Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics* 34, 1053–1055. doi: 10.1093/bioinformatics/btx701

Lundin, D., Poole, A. M., Sjoberg, B. M., and Hogbom, M. (2012). Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *Journal of Biological Chemistry* 287, 20565–20575. doi: 10.1074/jbc.M112.367458

Malik, A. J. (2018). Exploring deep phylogenies using protein structure. Massey University.

Malik, A. J., Langer, D., Verma, C. S., Poole, A. M., and Allison, J. R. (2023). Structome: a tool for the rapid assembly of datasets for structural phylogenetics. *Bioinformatics Advances* 3, 1–10. doi: 10.1093/bioadv/vbad134

Malik, A. J., Poole, A. M., and Allison, J. R. (2020). Structural Phylogenetics with Confidence. *Mol Biol Evol* 37, 2711–2726. doi: 10.1093/molbev/msaa100

Malik, A. J., Puente-Lelievre, C., and Matzke, N. (2024). On use of tertiary structure characters in hidden Markov models for protein fold prediction. 1–12. Available at: https://doi.org/10.1101/2024.04.08.588419

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Mo, Y. K., Lanfear, R., Hahn, M. W., and Minh, B. Q. (2023). Updated site concordance factors minimize effects of homoplasy and taxon sampling. *Bioinformatics* 39, 1–2. doi: 10.1093/bioinformatics/btac741

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., and Dessimoz, C. (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *bioRxiv*, 2023.09.19.558401. Available at: http://biorxiv.org/content/early/2023/10/02/2023.09.19.558401.abstract

Paradis, E., and Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Puente-Lelievre, C., Malik, A. J., Douglas, J., Ascher, D., Matthew, B., Allison, J., et al. (2024a). Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *BioArXiv*. doi: https://doi.org/10.1101/2023.12.12.571181

Puente-lelievre, C., Ridone, P., Douglas, J., Amritkar, K., Kaçar, B., Baker, M., et al. (2024b). Molecular and structural innovations of the stator motor complex at the dawn of flagellar motility. 1–13.

Rokas, A., and Carroll, S. B. (2005). More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol* 22, 1337–1344. doi: 10.1093/molbev/msi121

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection* 12, 85–94. doi: 10.1093/protein/12.2.85

Shannon, C. E. (1948). A mathematical theory of communication. *e Bell System Technical Journal,* 27, 379–423. doi: doi: 10.1002/j.1538-7305.1948.tb01338.x.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 4, 1–5. doi: 10.1093/ve/vey016

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A. H., Krieger, E., Joosten, R. P., et al. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* 43, D364–D368. doi: 10.1093/nar/gku1028

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., et al. (2024). Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 42, 243–246. doi: 10.1038/s41587-023-01773-0

Wei, T., and Simko, V. (2024). R package "corrplot": Visualization of a Correlation Matrix. Available at: https://github.com/taiyun/corrplot