OXFORD UNIVERSITY PRESS | Genome Biology and Evolution

Protein structure characters in the light of phylogenetic systematics

Journal:	*Genome Biology and Evolution*
Manuscript ID	Draft
Manuscript Type:	Perspective
Date Submitted by the Author:	n/a
Complete List of Authors:	Matzke, Nick; University of Auckland - City Campus, School of Biological Sciences Li, Changhao ; University of Auckland, School of Biological Sciences
Keywords:	protein structure, structural phylogenetics, phylogenetic systematics, semaphoronts, Foldseek, AlphaFold

Protein structure characters in the light of phylogenetic systematics

Nicholas J. Matzke[1,*]

Changhao Li[1]

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

[*] Corresponding author. n.matzke@auckland.ac.nz

Abstract (214 words)

Protein structure characters have great potential for improving phylogenetic inference, especially for deep nodes where amino acid sequences are highly diverged. The combination of AlphaFold structure predictions and Foldseek's "3Di" structural alphabet makes it relatively easy to conduct model-based phylogenetic inference that includes a partition of slow-evolving 3Di characters. However, we show that even identical amino acid sequences can produce substantially different 3Di codes, depending on the source of structural model and whether inter-chain interactions are considered. We argue that such variability can be addressed with key concepts from traditional organism-based phylogenetic systematics: semaphoront, hypodigm, and character ascertainment method. To illustrate this, we develop an analogy between organismal development, taphonomy, and subsequent description and character coding by a systematist, and the process of protein synthesis, folding, and interaction and subsequent extraction, experimentation, and structural modeling by a biochemist. We conclude that differences in 3Di codes between semaphoronts and are not

intrinsically a problem, but they do require that the researcher uses the same replicable method on all proteins in the phylogenetic analysis. The guiding principle should be to maximize the chance that character differences in the data matrix are the results of underlying evolutionary changes, rather than artefactual differences between proteins due to differences in the methods used for obtaining semaphoronts and coding characters.

Significance Statement (149 words)

We show how identical amino acid chains can produce different three-dimensional (3Di) protein structure characters. This is concerning because previous work suggests that 3Di characters are more conserved than amino acid sequences, and thus may be useful for deep phylogenetic inference. We address this puzzle with reference to the careful terminology that phylogenetic systematists developed to describe specimens and characters. Similar to organisms and fossils, proteins can exhibit different characters depending on developmental stage, the environment, and impacts from experimental procedures, as well as character coding methods. We suggest that researchers should focus on replicability and uniformity of character data acquisition methods used to generate a particular matrix, and be guided by the principle of maximizing the chance that character differences in the data matrix are the results of underlying evolutionary changes, rather than artefactual differences between proteins due to differences in methods used for obtaining structures and characters.

Main Text (2387 words, without references)

Protein structural phylogenetics is an attempt to use information about protein structure, which is typically highly conserved in evolutionary history (Illergård & Ardell, 2009), to estimate phylogenetic trees, or to improve the estimation of phylogenetic trees. After diverse early efforts with substantial practical limitations (reviewed by Puente-Lelievre et al. 2025), the field may be poised to take off due to the combination of two innovations. The first is easily accessible and reasonably accurate computationally-predicted protein structure models produced by algorithms like AlphaFold (Jumper et al. 2021; Varadi et al. 2024), which (in monomer form) are pre-calculated for the entirety of UniProt, and which can be predicted for novel sequences using CoLabFold (Mirdita et al., 2022). The second is the invention of a 20-character "structural alphabet" of tertiary-interaction characters, termed "3Di" (for 3-dimensional), by the authors of the deep homology search program Foldseek (van Kempen et al., 2024). Foldseek can generate a 3Di structural character at every residue position of an amino acid (AA) chain, and the fact that the 3Di alphabet has 20 letters means that these codes can easily be used in readily available algorithms designed for AA-based homology searches, multiple sequence alignment, and phylogenetic modelling and inference (Puente-Lelievre et al. 2024).

The combination of these tools raises the prospect that almost any protein phylogenetics exercise can be enhanced with a simple workflow (Matzke 2025): (1) download or generate AlphaFold structural predictions for each protein, (2) generate 3Di codes with Foldseek, (3) combine AAs and 3Dis into a partitioned data matrix with simple R functions (Malik et al. 2024; Matzke 2025), and (4) conduct inference using

standard programs such as IQ-TREE (Minh et al. 2022), including statistical model comparison to select optimal substitution models for the AA and 3Di partitions, adding special transition rate matrices for the 3Di characters using a transition matrix derived from the Foldseek 3DI cost matrix (van Kempen et al., 2024; Puente-Lelievre et al. 2024) or similar inferred rate matrices (Garg & Hochberg, 2025).

Initial results suggest that adding 3Di characters to an AA dataset can increase support and accuracy for the deepest nodes in a tree, particularly in the “twilight zone” where AA sequence similarity has decayed to very low levels (Puente-Lelievre et al. 2024). Another preliminary study suggests that adding 3Di characters statistically significantly improves average resolution of nodes in a sample of 10 protein superfamilies, although some of these superfamilies show little improvement, particularly if sequences are short and dominated by alpha-helices (Fullmer et al., 2025). Garg & Hochberg (2025) show that 3Di characters have much slower rates of change than amino acids, potentially substantially improving our confidence in estimates of classic puzzles such as rooting the Tree of Life using ancient gene duplications. However, they also note that 3Di characters can inherit information from structures that might bias phylogenetic inference. For example, in a clade of proteins where some members oscillate between two functional folds, while other members specialize on only one of the two folds, the 3Di-based phylogenetic inference will group a bistable protein based on which conformation was used to generate the 3Di characters, perhaps contradicting the phylogenetic history inferred from AAs (Garg & Hochberg 2025).

Such issues suggest that, as structural phylogenetics comes into view, we need to do some careful thinking about fundamental concepts such as homology, and how they

relate to new datasets such as 3Di characters, and likely additional alphabets and encodings of structure that may emerge in the future.

Homology puzzles: same sequence, different 3di characters

Homology puzzles are not difficult to come across while working with 3Di characters. We give an example from our own work. The bacterial flagellum has long been a model system in biochemistry, and the question of how such a complex system might evolve is a longstanding research topic (Pallen & Matzke, 2006). As the system is very ancient, likely dating back to the last bacterial common ancestor, the relationships between flagellar and nonflagellar proteins, or between pairs of flagellar proteins in the case of ancient duplications, are often in the twilight zone. It is therefore appealing to attempt to use structure predictions and 3Di characters to improve resolution.

We examined the flagellar export gate proteins, consisting of FliP, FliQ, FliR, and FlhB, which form a complex at the bottom of the Type 3 Secretion System export apparatus which serves as both the rotating base of the flagellum, and the system which selects and exports the protein subunits to progressively construct the flagellar rod, hook, and filament. When an experimental structure was resolved with cryo-electron microscopy (Kuhlen et al. 2018; Kuhlen et al. 2020), it was revealed that FliPQR are structural homologs, and FlhB may share a region of homology as well (Figure 1).

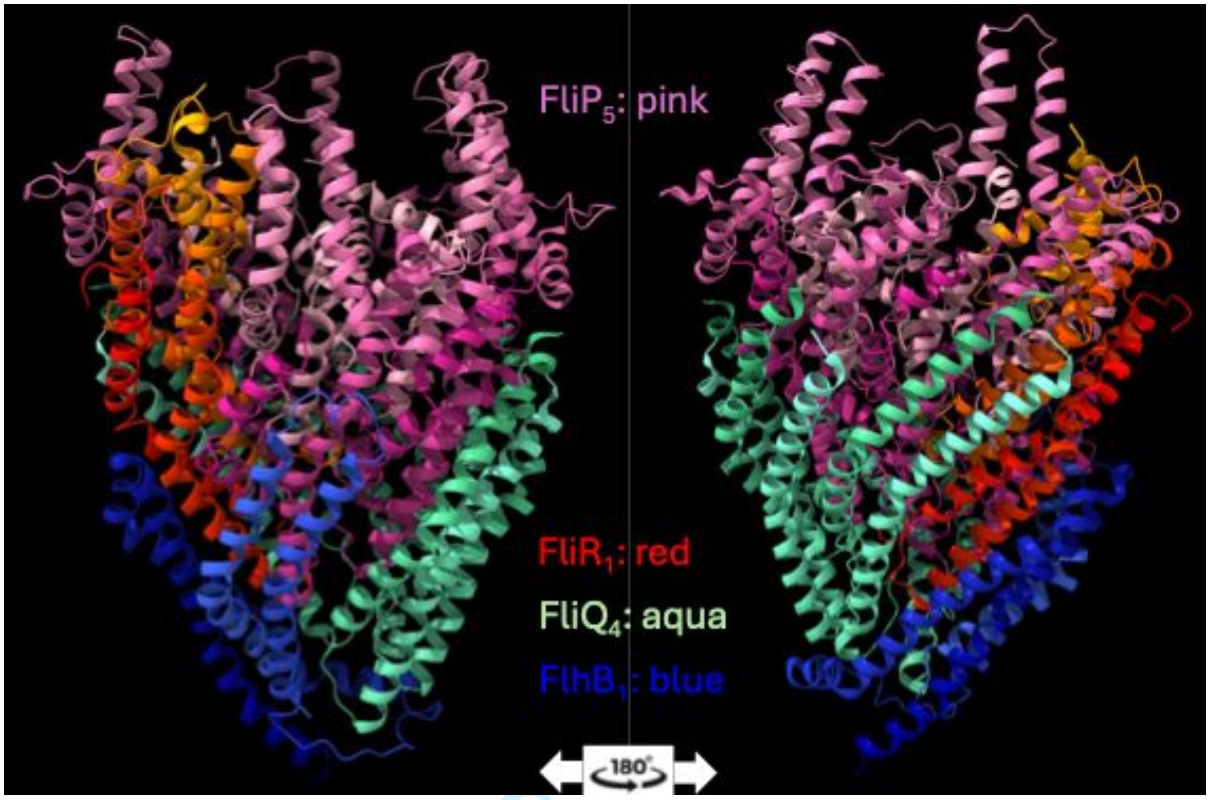

Figure 1. Ribbon model of cryoelectron microscopy structure of the FliPQR-FlhB complex. FliQ consists of 2 bent alpha helices, a structure which is closely superimposable on two regions each of FliP and FliR. Colors and rotations done in ChimeraX. Structure source: https://www.rcsb.org/3d-view/6S3L, model created by Kuhlen et al. (2020).

To test the behavior of Foldseek's *structureto3didescriptor* function, which converts a protein structure backbone chain into a series of 3Di characters equal to the number of AA residues, we generated 3Di characters on a number of input files:

1. Kuhlen et al. (2020)'s FliPQR-FlhB structure, 6S3L, as downloaded from RCSB unmodified.

2. The same structure, but with each chain split out into a separate PDB file. These produced identical 3Di characters to #1, so were not examined further.

3. The same structure, but with the PDB files modified so that all chains were labelled "Z" instead of chains A-K, and with all the residues numbered consecutively according to their order in the PDB file. This procedure changed

no amino acids or coordinates, but apparently makes *structureto3didescriptor* treat all the sequences as one large chain, allowing encoding of tertiary interactions between them.

4. Using the ChimeraX *Fetch* command, we downloaded the databased AlphaFold structure models for the most similar sequences to the FliP, FliQ, FliR, and FlhB sequences used in Kuhlen et al. (2020)'s structure. These were all 100% identical in shared sequence, however some sequences had small differences in length compared to the same protein in the structure, perhaps due to posttranslational or other modifications to the chains used in the experimental 6S3L structure.

All of the resulting 3Di characters were manually aligned by eye, taking care to attempt to align the strongest structural homology region (the last 2 alpha helices in the FliP and FliR structures, which are approximately superimposable on the FliQ structure). Our purpose was not to conduct a phylogenetic analysis, but simply to display similarities and differences in the 3Di characters produced by these different input files. The resulting 3Di characters are shown in Figure 2.

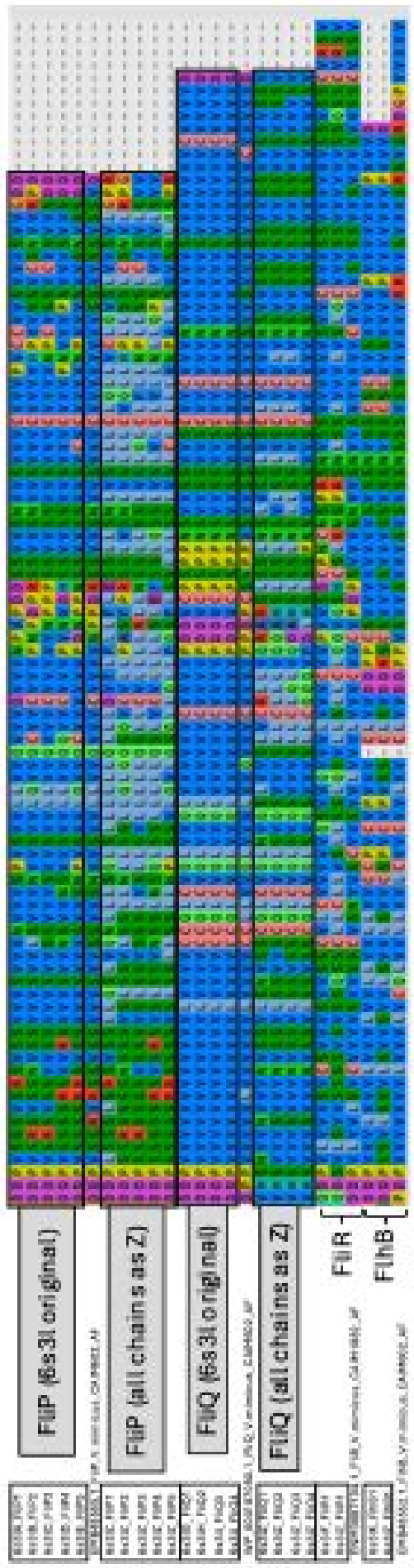

Figure 2. Identical peptides can produce different 3Di characters in different circumstances, as shown by 3Di characters derived from the FliPQR-FlhB structure 6S3L. See text for details.

We are initially confronted by a paradox: although 3Di characters have been shown to be more conserved than AAs in several studies, here we have identical amino acid sequences producing similar but decidedly nonidentical 3Di characters.

This puzzle can be resolved by consideration of the Foldseek algorithm and the input structures. First, in an experimentally resolved structure, identical peptides can have somewhat different structures based on their position in the overall quaternary structure and how they pack with their neighbors.

Second, Foldseek's *structureto3didescriptor* algorithm works by assigning each residue in a chain a geometric "theoretical center", and then identifying which neighboring amino acid in the backbone has the closest theoretical center. This pair of amino acids then has a series of measurements taken, describing the distance and geometric relationship of the backbones containing the pair of residues. These measurements are then fed into the 3Di alphabet classifier, which was trained on thousands of aligned proteins to produce an alphabet that maximized sensitivity for retrieving deep homologs in database searches. Each residue is thus classified into a 3Di structural character.

If we relabel all of the chains as "Z", Foldseek then treats the whole structure as one large chain, and it is allowed to find closest-neighbor residues from other chains, which is a quite frequent occurrence in an integrated structural complex such as FliPQR. This could be expected to change some 3Di codes.

Third, the AlphaFold-predicted structures, while done for identical or near identical sequences, were statistical predictions of monomer structures, with no accounting for the influence of neighbors. It can be seen in Figure 1 that the AlphaFold 3Di sequences share some features with the 3Dis from the original FliPQR chain labelling, and some with the all-labeled-Z chains.

Concepts from phylogenetic systematics to the rescue: semaphoronts and hypodigms

At this point, some researchers might be tempted to throw up their hands and declare that 3Di characters have been debunked, seeing as minor differences in source files can produce different 3Dis from identical AAs. We argue that this would be an incorrect conclusion. While these kinds of ambiguities are quite rare in sequence-based phylogenetics, where the character states are determined by the genetic code, they are quite common in more traditional fields of biology such as morphological systematics and paleontology. These fields have traditionally relied on extensive comparative analyses of bones and other morphology to assign homology and then to semi-subjectively code observed variations as discrete characters. This is a large area, but a few basic concepts are useful to resolve the 3Di puzzle: *semaphoront* and *hypodigm*.

The term *semaphoront* was coined by Hennig in 1966 (Hennig 1999) to denote a “signal bearer”, i.e. a specimen of an organism, from which characters may be derived. In the best case, it is the museum-preserved and catalogued specimen of a whole organism, but quite often only a partial specimen is available: the dropped tail of a lizard, the jawbone of a fossil, etc. Sometimes the semaphoronts can be some other item, such as

a photograph of an endangered species that cannot be collected, or a recording of a bird song.

As any biology student knows, organisms are not static. Instead, especially for multicellular organisms, they go through an elaborate process of development, through which many morphologies may be displayed during growth (Figure 3, black lines), or due to sexual differentiation or environmental effects. All of these stages might be potential sources of phylogenetic characters to be sampled by researchers (Figure 3, grey circles). Furthermore, even at the point of death or sampling, the "development" of the semaphoronts may not cease. *Taphonomy* (Martin, 1999) is the study of how an organism changes after death, including the processes of decay, fossilization and degradation of a fossil, as well as potential impacts by human researchers during the process of fossil preparation and study. This can be important for phylogenetics, as there are known cases where taphonomic processes can bias phylogenetic inference (Sansom and Wills, 2013).

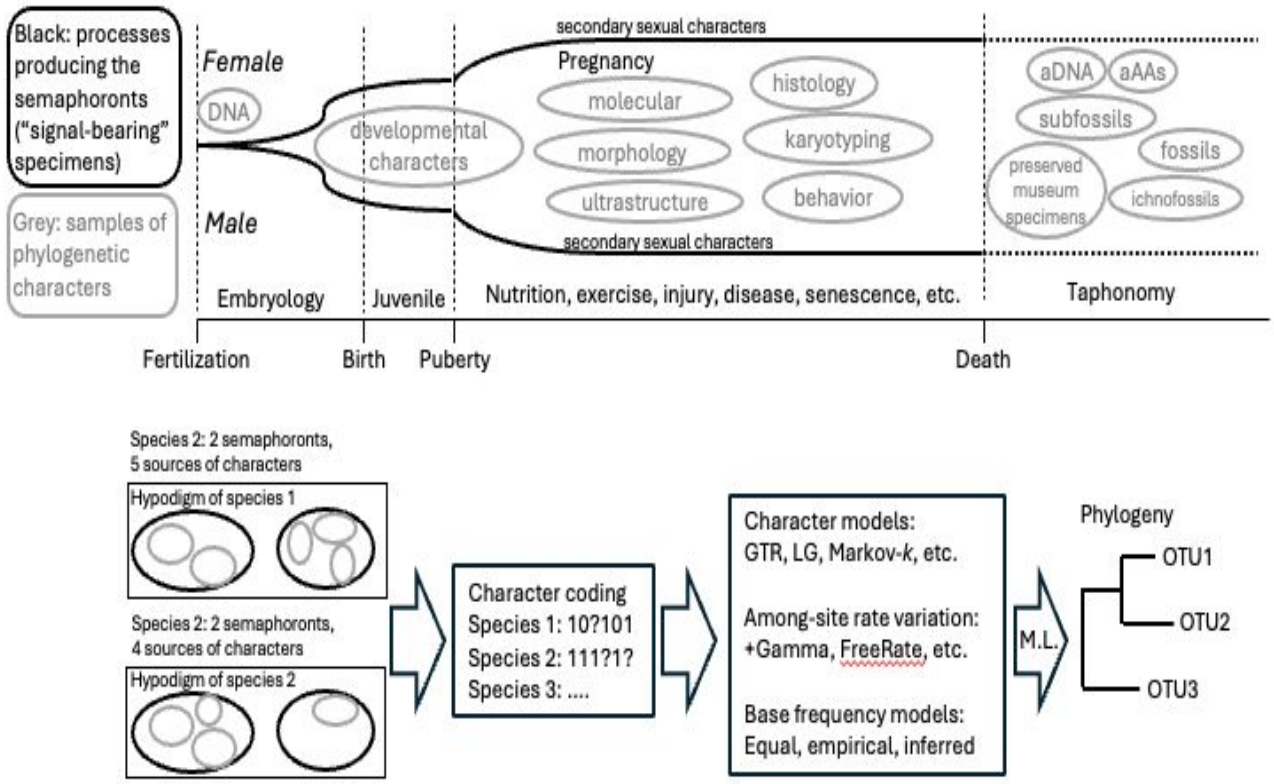

Figure 3. Schematic overview of how phylogenetic systematists sample characters from across the lifespan of organisms. Top: As organisms develop (black lines), different characters are at available for sampling. After death, taphonomic processes might further modify the specimen. *Gray circles*: Sources of characters that researchers might collect at different life stages. Bottom: Often, characters for a species are not all available from a single specimen (technically a "semaphoront", a signal-bearing item). Therefore, systematists propose a hypodigm. Coded characters are pooled from across the specimens for phylogenetic inference. Abbreviations: aDNA=ancient DNA; aAAs=ancient amino acids; OTU=Operational Taxonomic Unit; ML=Maximum Likelihood inference.

The final concept necessary here is the *hypodigm* (Figure 3), a concept invented by the paleontologist George Simpson (Simpson 1940) to describe a hypothesis about which fossil specimens belong to which species. A hypodigm, in modern terms, is a collection of semaphoronts that researchers have presumed can be grouped in order to provide characters describing a species or, more agnostically, an "Operational Taxonomic Unit" (OTU). These semaphoronts might all come from a single very well-studied specimen, but very often, systematists will describe characters from females, males, juveniles, embryos, as well as characters obtained using advanced technology on small

parts of specimens (e.g., sequencing, chromosome structure, sperm ultrastructure) or using visual or auditory observation of live organisms in the wild (behavior, song recordings).

How does this very brief review of classic concepts from systematics help to resolve the puzzles of 3Di characters? We propose that a distant, but helpful, analogy can be made between the processes of organismal growth, development, and taxonomy, and the subsequent character sampling process of systematists (Figure 3), and the processes of protein synthesis, folding, and subsequent sampling and experimental description of proteins (Figure 4).

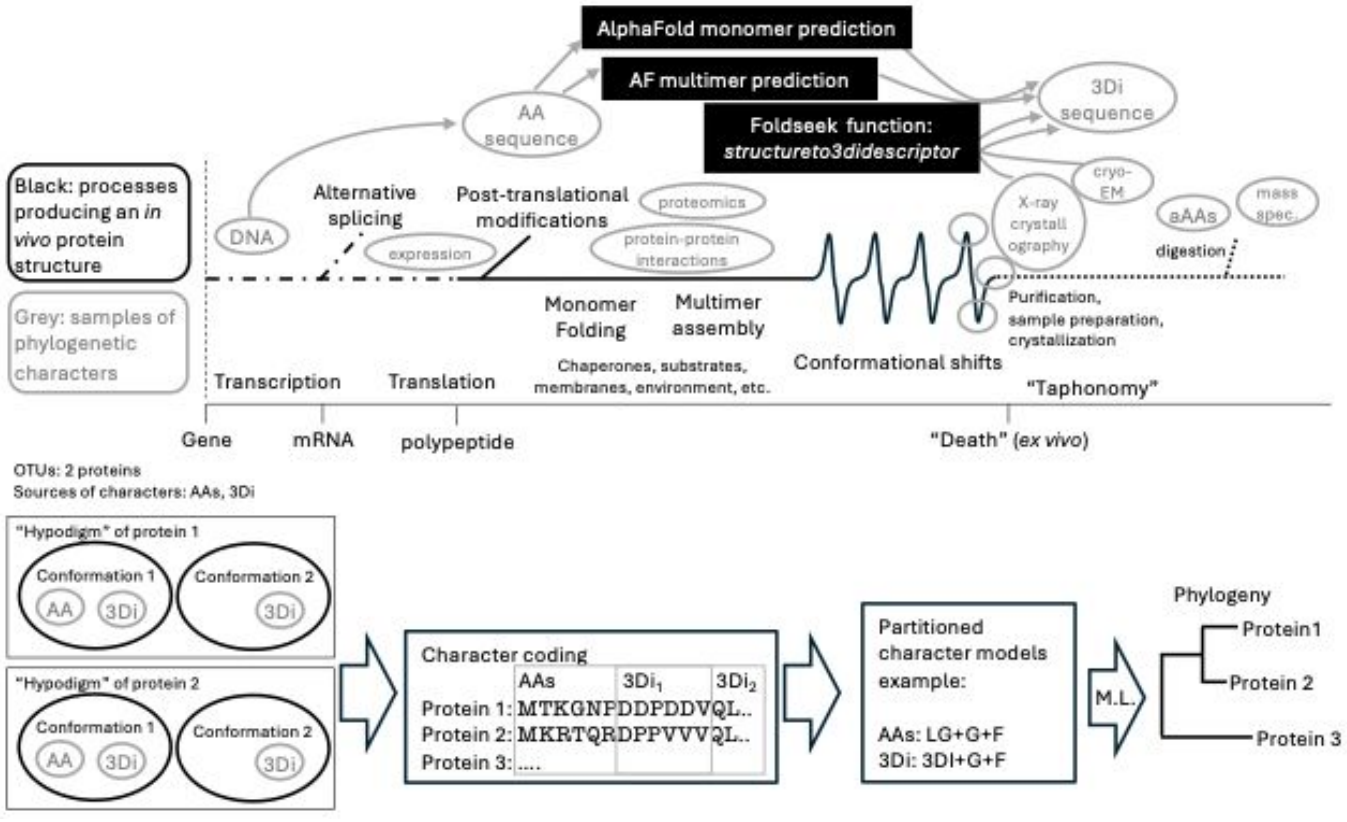

Figure 4. Proposal for how protein phylogeneticists might make use of concepts from organismal phylogenetics (Figure 3) in order to help interpret different sources of characters. Colors as in Figure 3. AlphaFold and Foldseek's 3Di classifier are, from the

point of view of character collection, black boxes that provide another source of characters from the amino acid sequence input.

While the analogies between proteins and developing organisms are only approximate, there are some relevant similarities. A mature, folded protein is the result of a series of processes: gene expression, transcription, translation, possibly post-transcriptional and post-translational modifications, and folding as influenced by a variety of interactions with the environment, chaperones, and cooperative partners. The mature protein is not itself static and may oscillate between multiple conformations for functional or other reasons. Any of the stages from gene expression to folding to function might be a source of phylogenetic characters. Finally, in an analogy to organismal death and taphonomy, proteins in an experimental setting are extracted, purified, and subjected to elaborate preparations such as crystallization, which hopefully enables an approximation of one of the biologically-relevant conformations. AlphaFold (AF) predictions provide a shortcut to structural models, but the models may differ depending on (for example) whether a monomer or multimer prediction is used. From the point of view of phylogenetic systematics, all of the resulting experimental and computational models constitute possible semaphoronts, signal bearers, from which characters for a particular protein might be coded. Foldseek provides one method for coding characters from these 3D structures

The key point is that in protein structural phylogenetics, as with organismal phylogenetics, we should move away from the assumption, derived from sequences, that there is one and only one set of “correct” characters that can be coded from protein. What we actually have is a number of possible semaphoronts that give us various windows into a protein’s structural conformations during its life cycle, and then

we have a number of ways of encoding conserved structural characters for phylogenetic inference – and the number is sure to increase as researchers invent additional structural alphabets beyond 3Di.

This point of view suggests that researchers should not focus on whether there is a "correct" set of structural characters for a particular protein, and instead should put effort into ensuring that their methodology for producing structural characters is replicable, and is the same for all of the proteins under study (Puente-Lelievre et al. 2024). For example, due to the variability in 3Di coding documented above, it would likely be hazardous run a phylogenetic analysis where half the proteins had 3Dis derived from experimental structures, and half from AlphaFold predictions. Similarly, it could be hazardous to combine monomer and multimer structural predictions, or within-chain 3Di codes with within-complex 3Di codes (where all chains are given the same chain label, allowing Foldseek to recognize between-chain interactions in its coding).

The guiding principle should be to maximize the chance that character differences in the data matrix are the results of underlying evolutionary changes, rather than artefactual differences in the methods used for character ascertainment.

Author contributions

NJM conducted the research and wrote the manuscript. As part of his Masters project, CL worked with NJM to explore the FliPQR-FlhB structure models and 3Di character datasets derived from them.

Acknowledgements

This work was supported by Human Frontier Science Program (HFSP) grant RGP0060/2021, as well as Australian Research Council (ARC) DP240100462, New Zealand Royal Society Rutherford Discovery Fellowship (RDF) 21-UOA-040, Marsden Grant 18-UOA-034, and University of Auckland, Faculty of Science Research Development Fund, FoS RDF #3732317. The author acknowledges helpful discussions with Caroline Puente-Lelievre, Jordan Douglas, Ashar Malik, and Anthony M. Poole.

References

Fullmer M, Puente-Lelievre C, Matzke N 2025. Adding 3Di characters to amino acid datasets can improve resolution, but the effect is weaker in shorter and alpha-helical proteins such as histones. *bioRxiv*, submitted to *GBE* special issue June 30, 2025. https://doi.org/

Garg SG, Hochberg GKA 2024 A general substitution matrix for structural phylogenetics. *Molecular Biology and Evolution*, 426. https://doi.org/10.1093/molbev/msaf124

Hennig W. 1999. *Phylogenetic Systematics*. Translated by DD Davis and R Zangerl. University of Illinois Press.

Illergård K, Ardell DH, Elofsson A 2009 Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. *Proteins* 3:499-508. https://doi.org/10.1002/prot.22458

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D 2021 Highly accurate protein structure prediction with AlphaFold. *Nature* 596 7873:583-589. https://doi.org/10.1038/s41586-021-03819-2

Kuhlen L, Abrusci P, Johnson S, Gault J, Deme J, Caesar JJE, Lea SM. 2018. Structure of the core of the type III secretion system export apparatus. *Nature Structural & Molecular Biology*, 257, 583–590. https://doi.org/10.1038/s41594-018-0086-9s

Kuhlen L, Johnson S, Zeitler A, Baurle S, Deme JC, Caesar JJE, Debo R, Fisher J, Wagner S, Lea SM 2020. The substrate specificity switch FlhB assembles onto the export gate to regulate type three secretion. *Nature Communications* 11: 1296-1296. https://doi.org/10.1038/s41467-020-15071-9

Malik AJ, Puente-Lelievre C, Matzke NJ, Ascher DB 2024. On use of tertiary structure characters in hidden Markov models for protein fold prediction. *bioRxiv*, 588419. https://doi.org/10.1101/2024.04.08.588419

Matzke NJ, Puente-Lelievre C, Baker MAB 2025. A Pipeline for Generating Datasets of 3-Dimensional Tertiary Interaction Characters for Model-Based Structural Phylogenetics. "A Pipeline for Generating Datasets of 3-Dimensional Tertiary Interaction Characters for Model-Based Structural Phylogenetics." Chapter in: *Evolutionary Genomics: Methods and Protocols*. Series Title: *Methods in Molecular Biology*. Series Editor: Gustavo Caetano-Anollés. https://osf.io/preprints/osf/5uhkx_v1

Minh BQ, Lanfear R, Ly-Trong N, Trifinopoulos J, Schrempf D, Schmidt HA 2022 IQ-TREE version 2.2.0: Tutorials and Manual. Phylogenomic software by maximum likelihood. *User Manual*, http://www.iqtree.org/doc/iqtree-doc.pdf

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M 2022 ColabFold: making protein folding accessible to all. *Nature Methods* 19 6:679-682. https://doi.org/10.1038/s41592-022-01488-1

Pallen MJ, Matzke NJ 2006 From *The Origin of Species* to the origin of bacterial flagella. *Nature Reviews Microbiology* 4 10:784-790. https://doi.org/10.1038/nrmicro1493

Martin RE 1999. *Taphonomy: a process approach*. Cambridge University Press.

Puente-Lelievre C, Malik AJ, Douglas J 2025 Protein Structural Phylogenetics. *Genome Biology and Evolution* in press.

Puente-Lelievre C, Malik AJ, Douglas J, Ascher D, Baker M, Allison J, Poole A, Lundin D, Fullmer M, Bouckert R, Kim H, Steinegger M, Matzke N 2024 Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *bioRxiv*, 2023.12.12.571181. https://doi.org/10.1101/2023.12.12.571181v2

Sansom R, Wills M 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. *Scientific Reports* 3, 2545. https://doi.org/10.1038/srep02545

Simpson GG 1940. Types in modern taxonomy. *American Journal of Science*. 2386: 413-426. https://doi.org/10.2475/ajs.238.6.413

van Kempen MKS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. 2024 Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* 42:243-246. https://doi.org/10.1038/s41587-023-01773-0

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, Kovalevskiy O, Tunyasuvunakool K, Laydon A, Žídek A, Tomlinson H, Hariharan D, Abrahamson J, Green T, Jumper J, Birney E, Steinegger M, Hassabis D, Velankar S 2024 AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* 52 D1:D368-D375. https://doi.org/10.1093/nar/gkad1011