Abstract

Significance statement

Main Text

Protein structure 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. 2024), 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 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, 2024).

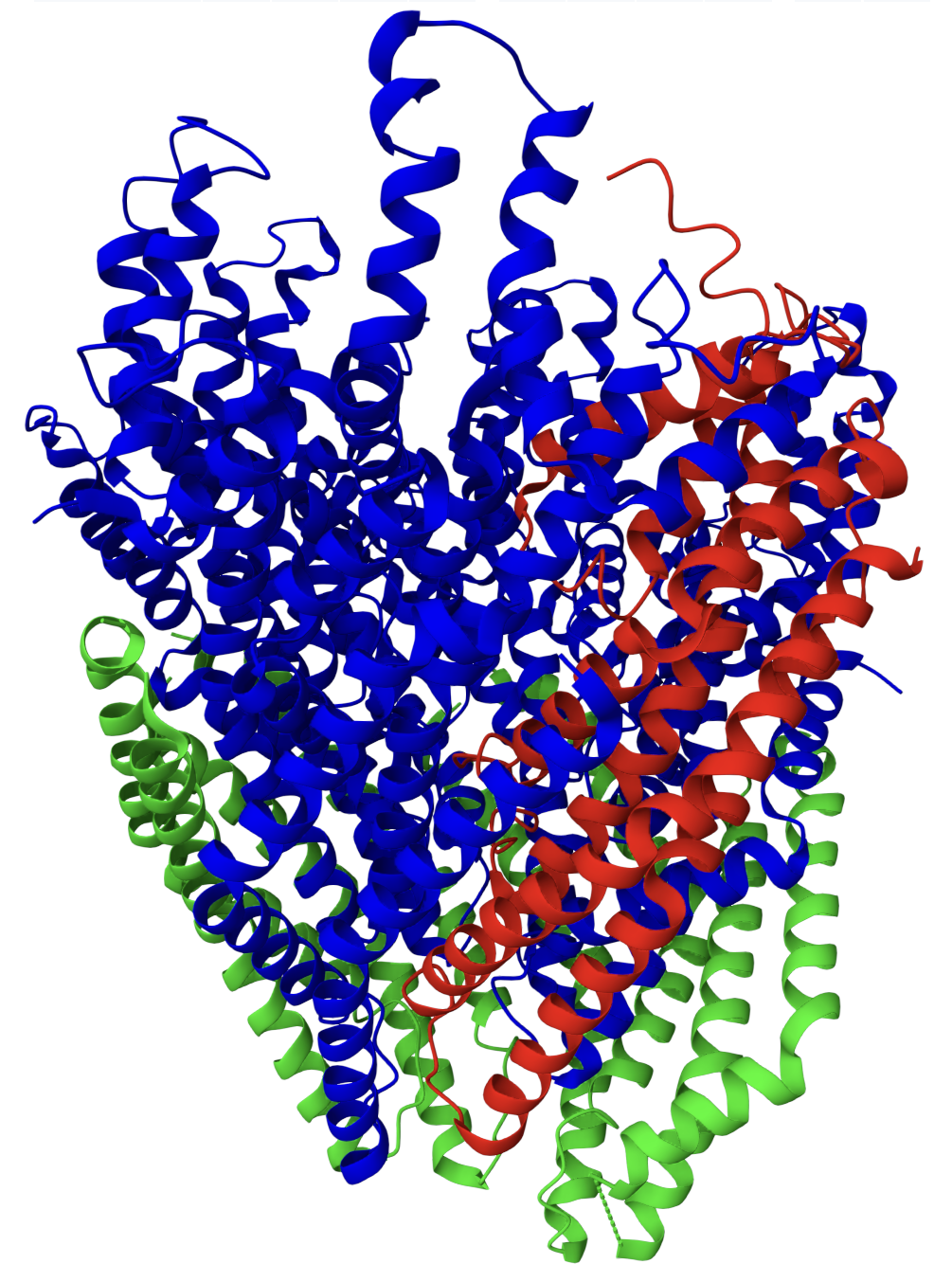
Initial results suggest that adding 3Di characters to an AA dataset can increase support and accuracy for the deepest nodes in a tree, particularly in the “twilight zone” where AA sequence similarity has decayed to very low levels (Puente-Lelievre et al. 2024). Another preliminary study suggests that 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 (2024) 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 2024).

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.

**An homology puzzle: 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 predications 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 FliPQR complex. Blue: FliP (5 copies), Green: FliQ (4 copies), Red: FliR (1 copy). FliQ consists of 2 bent alpha-helices, a structure which is closely superimposable on two regions each of FliP and FliR. Graphic source: <https://www.rcsb.org/3d-view/6S3R> , with manual color schemes. The structure model was 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, 6s3r, 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, perhaps due to insertions/deletions, or posttranslational or other modifications to the chains used in the 6s3r 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 (a larger version is available in Supplementary Material).

A colorful grid with black text

AI-generated content may be incorrect.

**Figure 2.** Identical or near-identical peptides can produce different 3Di characters in different circumstances, as shown by 3Di characters derived from the FliPQR-FlhB structure 6s3r. See text for details.

We are initially confronted by a paradox: although 3Di characters have been shown to be more conserved that AAs in several studies, here we have identical or near-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 amino acid 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.

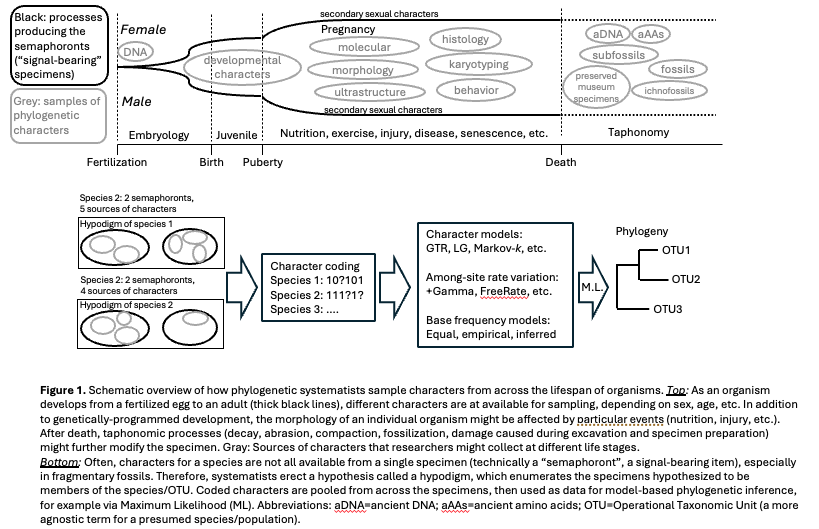
Third, the AlphaFold-predicted structures, while done for identical or near identical sequences, were statistical predictions of monomers structures, with no accounting for the influence of neighbors. It can be seen in Figure 1 that the AlphaFold 3Di sequences share some features with the 3Dis from the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

**A resolution from phylogenetic systematics: homology, characters, 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*.

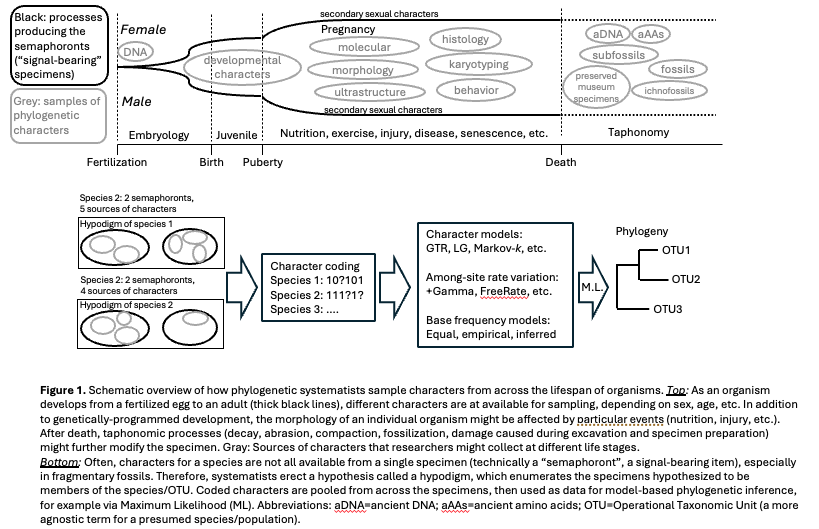
A *semaphoront* (Hennig 1966) is a “signal bearer”, i.e. a specimen of an organism, from which characters may be derived. In the best case, it is the museum-preserved and catalogued specimen of a whole organism, but quite often only a partial specimen is available: the dropped tail of a lizard, the jawbone 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 students of biology know, 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* is the study of how an organism changes after death, including the processes of decay, fossilization and degradation of a fossil, and potentially impacts by human researchers.



The final concept necessary here is the *hypodigm* (Figure 3). A hypodigm is a collection of semaphoronts that researchers have presumed can be grouped in order to provide characters describing a species or, more agnostically, Operational Taxonomic Unit (OTU).

We propose that distant-but helpful analogy can be made between protein synthesis, folding, and sampling/description, and the character sampling process of systematists (Figure 4).

s

Acknowledgements

The work of NJM and his lab members working in structural phylogenetics 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 author acknowledge helpful discussions with Caroline Puente-Lelievre, Jordan Douglas, Ashar Malik, and Anthony M. Poole.

References