Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone

Caroline Puente-Lelievre[1,2], Ashar J. Malik[3,4], Jordan Douglas[2,5], David Ascher[3,4], Matthew Baker[6], Jane Allison[1], Anthony Poole[1,2], Daniel Lundin[7], Matthew Fullmer[1,2], Remco Bouckert[2,8], Hyunbin Kim[9], Martin Steinegger[9,10,11], Nicholas Matzke[1,2]

[1]School of Biological Sciences, University of Auckland, Auckland, New Zealand
[2]Centre for Computational Evolution, University of Auckland, New Zealand
[3]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia
[4]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[5]Department of Physics, University of Auckland, New Zealand
[6]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Australia
[7]Department of Biochemistry and Biophysics, Stockholm University, Sweden
[8]School of Computer Science, University of Auckland, Auckland, New Zealand
[9]School of Biological Sciences, Seoul National University, South Korea
[10]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea
[11]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, South Korea

Corresponding author: n.matzke@auckland.ac.nz

Abstract

Protein structure is more conserved than protein sequence, and therefore may be useful for phylogenetic inference beyond the "twilight zone" where sequence similarity is highly decayed. Until recently, structural phylogenetics was constrained by the lack of solved structures for most proteins, and the reliance on phylogenetic distance methods which made it difficult to treat inference and uncertainty statistically. AlphaFold has mostly overcome the first problem by making structural predictions readily available. We address the second problem by redeploying a structural alphabet recently developed for Foldseek, a highly-efficient deep homology search program. For each residue in a structure, Foldseek identifies a tertiary interaction closest-neighbor residue in the structure, and classifies it into one of twenty "3Di" states. We test the hypothesis that 3Dis can be used as standard phylogenetic characters using a dataset of 53 structures from the ferritin-like superfamily. We performed 60 IQtree Maximum Likelihood runs to compare structure-free, PDB, and AlphaFold analyses, and default versus custom model sets that include a 3DI-specific rate matrix. Analyses that combine amino acids, 3Di characters, partitioning, and custom models produce the closest match to the structural distances tree of Malik et al. (2020), avoiding the long-branch attraction errors of structure-free analyses. Analyses include standard ultrafast bootstrapping confidence measures, and take minutes instead of weeks to run on desktop computers. These results suggest that structural

phylogenetics could soon be routine practice in protein phylogenetics, allowing the re-exploration of many fundamental phylogenetic problems.

Introduction

Protein structural phylogenetics is the "missing chapter in molecular evolutionary biology" (Lai 2020). It is well-known that protein structure is often remarkably conserved (Chothia & Leak 1986; Flores et al. 1993) even in the "twilight zone" (Doolittle 1986; Chung & Subbiah 1986), where amino acid (AA) sequence similarity decays below 30%. Very low sequence similarity limits our ability to detect homology, accurately align sequences, and resolve deep relationships. It would therefore be advantageous to incorporate protein structural information into phylogenetic inference.

Structural phylogenetics: progress and limitations. Researchers have taken a variety of approaches to link structure to phylogenetic inference. Lake (1984) used visual interpretation of low-resolution quaternary structures of the ribosome to identify gross characteristics that seemed to support the placement of eukaryotes within archaea. More generally, except at the coarsest levels, structural classifications of proteins suggest common ancestry (e.g. SCOP Superfamily, CATH Homology; Andreeva et al. 2020; Sillitoe et al. 2019), but they do not resolve within-group relationships, nor do they constitute a formal estimate of a phylogenetic tree (Lundin et al. 2012, Malik et al. 2020).

Secondary structure (SS) and residue accessibility have been used to modify amino acid substitution processes (Thorne et al. 1996; Le & Gascuel 2010), including joint estimation of the secondary structure (Goldman et al. 1996; Goldman et al. 1998). A probabilistic model of amino acid dihedral angle evolution has been used to model pairwise evolution between structures (Golden et al. 2017), but has not been generalized for phylogenetic inference. Modifications to substitution rates based on tertiary structure have also been explored (Choi et al. 2007). Lai et al. (2020) extended this work to include seven DSSP-defined (Kabsch & Sander 1983) secondary structure (SS) states as character data, estimating substitution matrices from prealigned PDB structures.

An alternative approach to these methods is to use 3D structural superposition of structures. Structural superposition is already popular in the related problem of structure-informed sequence alignment (e.g., DALI, Holm & Sander 1995; TM-Align, Zhang & Skolnick 2005; SSM, Krissinel & Henrick 2004; US-align, Zhang et al. 2022; SWAMPNN, Trinquier et al. 2022). For phylogenetics, a matrix of pairwise distances between all structures is calculated using a metric such as root mean square positional deviation (RMSD), or a welter of others (Hasegawa & Holm 2009), including more robust metrics such as TM-Score (Zhang & Skolnick 2005), LDDT (Mariani et al. 2013), or Q_{score} (Krissinel & Henrik 2004; Malik et al. 2020). Following earlier papers using small numbers of structures (≤18; Bujnicki 2000; Breitling et al. 2001; Garau et al. 2005), Lundin et al. (2012) used distance matrices to estimate a phylogenetic network using NeighborNet with 83 PDB structures of the ferritin-like superfamily and broadly replicated previous sequence-based results, as well as structural database classifications and functionally-delineated groups.

Structural phylogenetics with distance methods is a simple approach, and thanks to the AlphaFold2 revolution (Jumper et al. 2021; Skolnick et al. 2021), is no longer hindered by a shortage of structural data. However, such approaches are restricted by the use of distance methods to create trees, such as Neighbor Joining (NJ). For phylogeneticists, distance methods are often considered rather "quick and dirty": NJ provides a reasonable first approximation, but its statistical consistency is contingent on the accuracy of the distance matrix (Bryant 2005, Gascuel and Steel 2006). The accuracy of distances is well-defined for nucleic and amino acids, where it measures the number of nucleotide or amino acid substitutions. However, for structural distances, it is not known whether LDDT (Mariani et al. 2013), TM-score (Zhang & Skolnick 2004), or others are fair metrics of structural change, where the complex true process would include not only changes in AAs, but also other changes such as binding partner or environment. Another issue is that protein conformations vary, sometimes substantially, on nanosecond timescales. Nevertheless, work to date (Chothia & Lesk 1986; Flores et al. 1993; Lundin et al. 2012; Lai et al. 2020; Malik et al. 2020) suggests that structural distances correlate with the better understood metric of sequence divergence.

A further fundamental limitation of NJ used on structural distances (unlike nucleic or amino acid sequences) is the lack of probabilistic models describing generative processes. Probabilistic models have many advantages, including statistical model comparison and selection, and standard procedures to quantify statistical uncertainty, such as bootstrapping and posterior probabilities. Moreover, generative models are explanatory and capture real biological processes, allowing use of *in silico* simulations for model validation. Lastly, probabilistic models open the pathway towards “total-evidence” approaches (Hulsenbeck et al. 1996) allowing multiple sources of data to be combined into a single analysis. It is challenging to conceptualize a generative model that realistically describes how secondary, tertiary, and quaternary structure co-evolves over evolutionary time, but models of some aspects exist (Boomsma et al. 2008)

A method to ameliorate some of the above issues was proposed by Malik et al. (2020). Here, Molecular Dynamics (MD) simulations were used to sample different conformations of each leaf structure. Superposition, distance calculation, and NJ trees were calculated for each sample of tip conformations, and MD-based bootstrap scores were calculated for each branch, providing a measure of the effect of structure variability on the stability of the estimated relationships. The method was used on a subset of Lundin et al.'s ferritin-like superfamily dataset (53 PDB structures), retrieving similar relationships but with the advantage of reporting bootstrap supports. Disadvantages of this approach include the slowness of the MD simulation step, its reliance on distance methods, and the fact that sources of uncertainty other than conformational variation are not addressed.

In summary, research to date indicates that structural data could have great utility in phylogenetics, in particular beyond the twilight zone, if the challenges of probabilistic modeling, uncertainty estimation, and computational runtime can be overcome (Bujnicki 2000; Breitling et

al. 2001; Garau et al. 2005; Lundin et al. 2012; Lai et al. 2020; Malik et al. 2020). Combined with reasonably accurate AlphaFold structures now available for many proteins, such a technique could rapidly make structural phylogenetics a routine part of protein phylogenetics.

3Di characters. A potential solution to the above challenges is provided by Foldseek (van Kempen et al. 2023), a program designed for efficient structure-informed homology searches deep in the twilight zone. Foldseek's 3Di alphabet encodes 3-D tertiary structure interactions in a 1-D string of discrete states, enabling deployment of highly efficient dynamic programming algorithms for alignment and search.

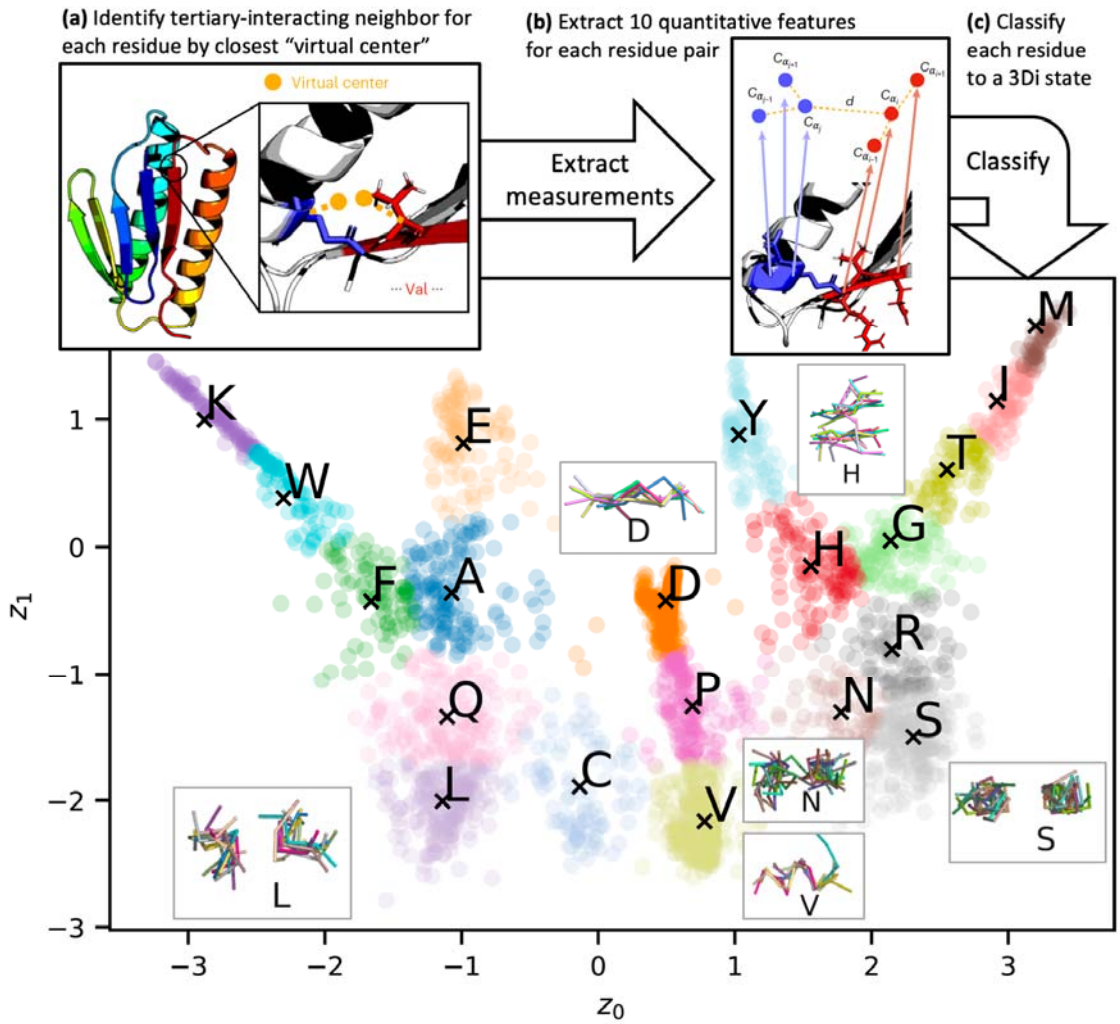


Figure 1. Summary of Foldseek's system for classifying tertiary interactions into the 20-state 3Di alphabet. For each residue in a protein structure **(a)** a virtual center is defined, and each residue is assigned a tertiary neighbor based on the closest virtual center. **(b)** 10 quantitative traits are collected for each residue+neighbor interaction, and then **(c)** assigned to the best-matching 3Di state using the classifier pre-trained by Van Kempen et al. 2023. The main panel shows a two-dimensional representation of the 20 states. Examples of 3Di states are inset; each of these represents 10 pairs of backbone fragments showing the typical arrangement for a 3Di state. Graphics modified from Van Kempen et al. 2023.

A summary of Foldseek's procedure classifying residues into 1 of 20 3Di states is given in Figure 1. For a full description of the classification procedure, and the derivation and training of the 3Di alphabet, see Van Kempen et al. (2023). The 3Di alphabet was originally designed to enable fast and sensitive homology search and structural alignment, and the alphabet was trained for that purpose. The encoding of tertiary structure interactions seems to retain the signal of core conserved folds in a protein, and to be robust to changes outside the core (such as indels) that challenge traditional superposition metrics. Sensitivity in homology search suggests that the 3Di states could be interpreted as highly conserved, phylogenetically informative *characters*. In combination with the AlphaFold database, this is an immediately appealing idea for phylogeneticists, as the Foldseek encoder could then be used to create a 3Di character dataset to compare to, or supplement, any traditional amino acid dataset. The fact that the 3Di alphabet has 20 states is particularly convenient, as it allows 3Di characters to be input into traditional bioinformatic and phylogenetics programs. It is tempting to think this brings us a step closer to the "Holy Grail": a set of easily-obtainable, highly conserved characters that can be added to any traditional protein phylogenetics project, enabling greater resolution deeper in the tree.

The creators of Foldseek have already taken one step in this direction with FoldTree, which uses Foldseek's 3Di alphabet to structurally align protein structures and AA sequences, and then calculate a "Fident" (fraction identical) distance between the AA sequences between pairs of structures (Moi et al. 2023). Although they obtain good results with this approach (higher taxonomic congruence and more ultrametricity), trees are estimated from distance matrices using FastME (NJ followed by NNI/SPR optimization for the minimum evolution tree; Lefort et al. 2015), and therefore subject to the limitations discussed above.

We propose that 3Di codes can be interpreted as conventional phylogenetic characters, and thus can be readily included in the suite of standard model-based phylogenetic estimation procedures, including statistical model comparison, maximum likelihood phylogenetic estimation, and bootstrapping. If it is true that 3Di characters are indeed the "Holy Grail," a source of slowly-evolving characters, then inferences including 3Di characters should outperform standard, structure-free AA-based inferences in the twilight zone.

As a proof-of-concept, we use both PDB and Alphafold2-predicted structures from the ferritin-like superfamily, a group that shares the ferritin fold. From the perspective of evolutionary enzymology, and hence phylogenetics, the ferritin fold is interesting. It likely evolved to control the Fenton reaction (Fenton, 1894), in which ferrous iron reacts with hydrogen peroxide to form highly reactive hydroxyl radicals. From this starting point it evolved into several major types (Lundin et al. 2012): rubrerythrins, and similar small proteins, control reactive oxygen species and are hence likely close to the ancestral function; the ferritins, bacterioferritins and “DPS” proteins store insoluble ferric iron in the cell or protect from radicals; and lastly, a large group of enzymes utilize highly reactive oxygen species for catalysis (fatty acid desaturases, bacterial multicomponent monooxygenases, BMMs) or radical formation (ribonucleotide reductase R2 components, RNR R2). This superfamily has previously been used as a test dataset for

structural phylogenetics (Lundin et al. 2012; Malik et al. 2020) as homology is likely, but inter-family relationships are in the twilight zone.

Methods

Structure Data. The 53 PDB structures used by Malik et al. (2020; henceforth M20) were downloaded from RCSB.org (Berman et al. 2000; RCSB Protein Data Bank 2023). They were clipped to the single chains used by Malik et al. using *bio3d::pdbsplit* (Grant et al. 2006; Grant et al. 2020). The list of PDB chain files was processed to AA and 3Di FASTA files using the custom R function *str_to_3di* (SI), which runs Foldseek's *structureto3didescriptor* command and parses the output to FASTA files.

Reference tree. The M20 structural distances NJ tree was replicated from the original distances matrix using SplitsTree (Huson & Bryant 2006). We used the M20 tree as the comparator for the trees estimated by Maximum Likelihood (ML). While there is no guarantee that the M20 tree is the exact true tree, it at least constitutes the state-of-the-art published estimate which represents the tree obtained from the traditional superposed structural distances approach, its clades reproduced previously-hypothesized groups recognized in structural databases, and bootstrap confidence estimates were available for the deeper branches. M20's clade labels and MD-based bootstraps were manually transferred to edited Newick files for automated comparison with ML trees.

AlphaFold-predicted structures. To test whether 3Di characters derived from AlphaFold structures would result in similar phylogenies to empirical structures, AlphaFold structures were downloaded corresponding to each M20 structure. This was accomplished by manually searching AlphaFold DB (https://alphafold.ebi.ac.uk/; Jumper et al. 2021; Varadi et al. 2022; Varadi et al. 2023) against the AA sequence for each M20 structure. All but 9 AA sequences had 100% matches in AlphaFold DB; when no 100% match was found, the closest-hits AlphaFold structure was chosen. The most distant match was 98% identical to the target AA sequence. While 100% identical sequences would be preferable, close-but-inexact matches represent a realistic situation for researchers with a list of study sequences who are aiming to add 3Di characters to their dataset. The average AlphaFold structure has a mean pLDDT of 94.1%, with 90% of all sites scoring over 85.1%, suggesting that structural predictions have high levels of confidence (likely reflecting the existence of close homologs in the protein data bank). See SI for a breakdown of pLDDT summaries.

Phylogenetic analyses. To test the effect of various analysis decisions, and to see whether or not 3Di codes function like standard phylogenetic characters, we ran a total of 60 maximum likelihood inference runs in IQtree v. 2.2.0 (Minh et al. 2020). Analysis variations used at each step (structures, data, alignment, trimming, models, partitioning) are described below.

Structures and Data. Analyses (alignment + phylogeny) were run using either no structures (AA-only alignment and phylogeny, matching standard phylogenetic analysis), PDB structures, or AlphaFold structures. For a given alignment, phylogenetic analyses were run on AAs-only, 3Dis-only, or AA+3Di datasets.

Alignment. Alignment for AA-only, structure-free runs was conducted with FAMSA (Deorowicz et al. 2016), which uses an MIQS substitution matrix useful for distantly-related proteins (Yamada & Tomii 2014). The first set of structure-informed runs used the Universal Structural Alignment (US-align; Zhang et al. 2022) program's Multiple-Structure Alignment (MSTA) mode to align AAs using the 3D structures themselves, with no involvement of the 3Di characters. The second set of structure-informed runs used FAMSA3di, a version of FAMSA recompiled with Foldseek's substitution cost matrix for the 3Di alphabet. FAMSA3di was used to align the 3Di characters. As Foldseek produces 3Di sequences of identical length to the AA sequence of a structure, a 3Di alignment can be transferred position-by-position to unaligned AA characters and vice versa, a procedure we implemented in the custom R function *align_3dis_to_AAs*.

Trimming. The alignment trimmer trimAl 1.4 (Capella-Gutierrez et al. 2009) was used to trim all columns of AA and 3Di alignments that were <35% data. This tended to cut regions of the alignment that were poorly-aligned regions (which were often low-confidence regions of the AlphaFold structures) or represented elements present in only a few structures. ML analyses were run on trimmed as well as full, untrimmed alignments.

Substitution models. ML analyses of AA, 3Di, and AA+3Di datasets were run under "allmodels", using IQtree's extensive default model set, or under a custom list which contained models *a priori* likely to be relevant to AA or 3Di data as well as computationally efficient. The custom list included a custom 3Di rate matrix, which we calculated from Foldseek's 3Di substitution cost matrix, following the method described by Tammi (2021). The custom models list (IQtree *-mset* input) was: Blosum62, Dayhoff, DCMut, JTT, JTTDCMut, LG, Poisson, PMB, WAG, EX2, EX3, EHO, EX_EHO, 3DI (non-3DI model descriptions available in Minh et al. 2022). The base frequency options used were *-mfreq* FU,F (base frequencies specified by the protein matrix, or empirical base frequencies), and the among-site rate variations used were *-mrate* E,G,R (Equal, Gamma, or FreeRate models of sitewise variation).

Partitioning. *A priori* it is extremely likely that AA and 3Di characters will evolve under different models, that AA models will be poor fits to 3Di data, that the 3Di matrix will be a superior fit, and that AA+3DI datasets should be partitioned so that each partition can have an appropriate model. However, because this is the first ML analysis using 3Di characters, and as a sanity check of the entire pipeline described above, we decided to run both partitioned and unpartitioned analyses of the AA+3Di alignments. Tree topology and relative branch lengths were shared across partitions.

All IQtree runs were conducted with 1000 ultrafast bootstraps (UFBoot), which approximate Bayesian posterior probabilities, and 1000 replicates of SH-like approximate likelihood ratio test (Guindon et al., 2010). Bootstraps were run with the *-bnni* option, which uses a hill-climbing Nearest-Neighbor Interchange (NNI) on each bootstrap tree to reduce the risk of overestimating UFBoot supports when the data severely violates the model (Minh et al. 2022). All runs were conducted on a desktop 4.2 GHz Quad-Core Intel Mac.

Analysis of phylogenies. Using custom R scripts (Supporting Information) we extracted from each IQtree run the best fit models and corresponding log-likelihood, AIC, AICc, BIC, number of data and number of parameters. Bootstrapped consensus trees were extracted from each run, midpoint-

rooted and ladderized for visualization, and compared to the M20 NJ tree by a variety of metrics. These included: number of matching bipartitions (51 possible); number of matching "deep" bipartitions (20 possible, corresponding to named clades of the M20 tree displayed in Figure 10b of Malik et al. 2020), and groupings of those; number of "shallow" bipartitions (all other M20 bipartitions); total tree length (sum of all branch lengths); sum of bootstrap scores across all branches; and tree-to-tree distances available in phangorn (*treedist* function; Schliep 2010), including Robinson-Foulds distance (symmetric difference), Subtree Pruning-Regrafting (SPR) distance, approximating the number of edits required to match tree topologies; and branch-score distance (calculated on trees where branch lengths had been divided by the total tree length, to avoid artifacts due to comparing trees of very different total lengths). RF and SPR distances consider only topology and ignore branch lengths, unlike the branch-score distance. Nonmetric multidimensional scaling was used to visualize the tree-to-tree distances (Matzke et al. 2015).

Results

Alignment. AA and 3Di alignments were visualized and compared using Jalview's percent-identity color scheme (Waterhouse et al. 2009). 3Di characters show substantially more conservation (116 sites are over 50% conserved) than AAs (only 17 sites) (Figure 2). The conservation pattern is particularly clear in the ferritin superfamily's 4 core alpha helices (Lundin et al. 2012).

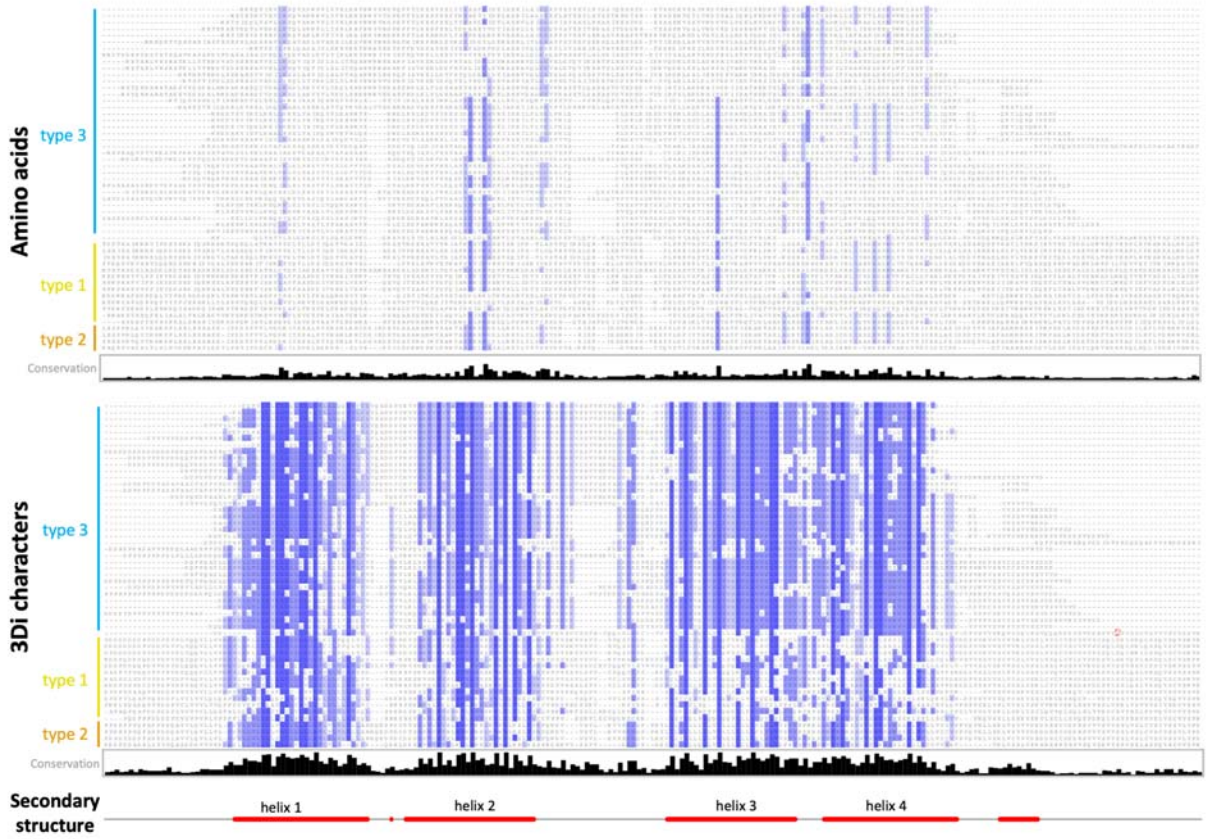


Figure 2. AA and 3Di alignments for 53 AlphaFold structures of the ferritin-like superfamily, displayed with Jalview's pairwise identity color scheme, where characters matching the majority character (if any) are colored, and darker colors are closer to 100% agreement. The 3Di alignment was generated with famsa3di; the AA alignment was generated by one-to-one replacement of 3Di states with AAs. Sites with <35% data were trimmed. FASTA files are available in SI.

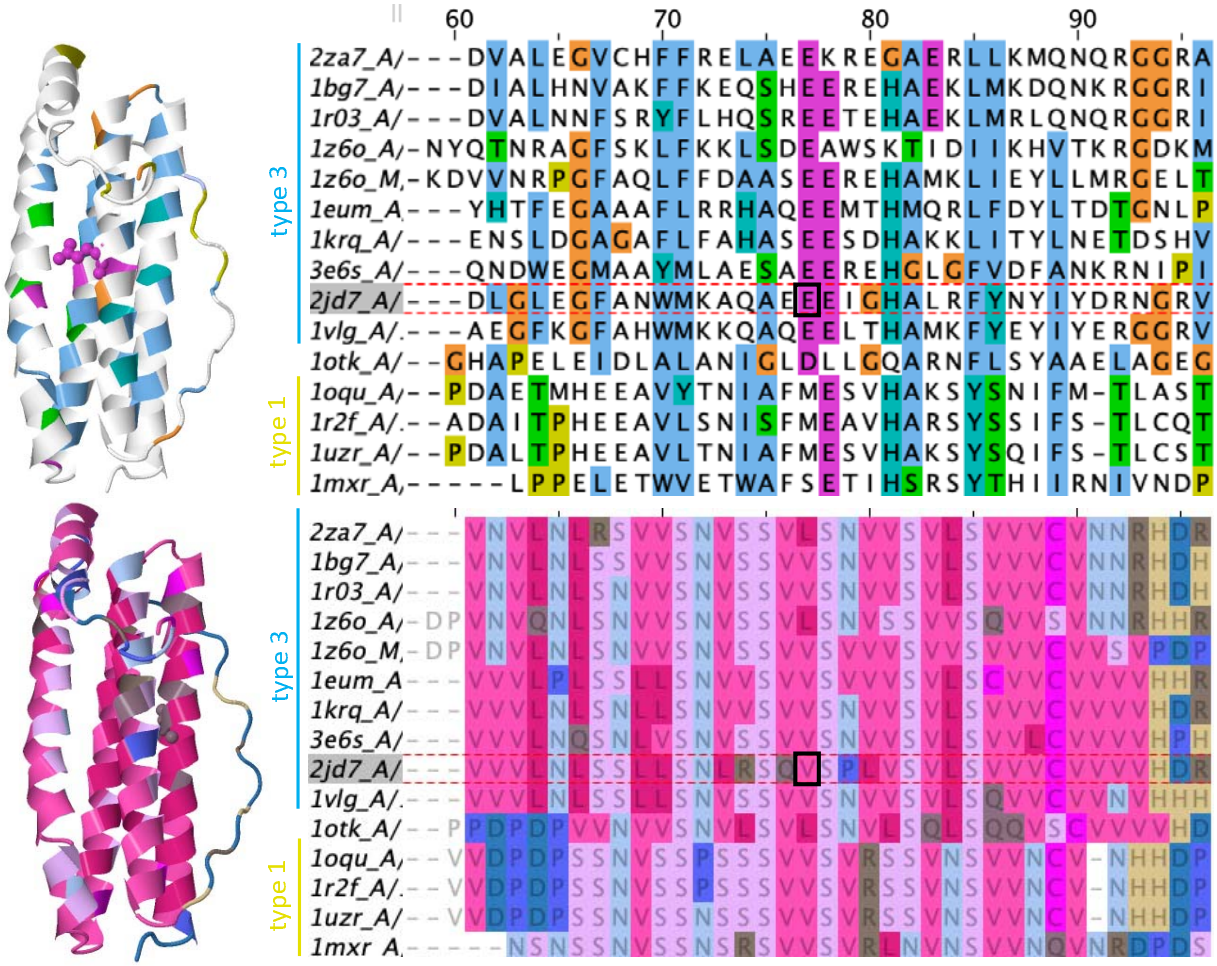


Figure 3. Illustrating the differences between AA and 3Di characters by zooming in on a portion of the alignment in Fig. 2, corresponding to the end of helix 2 and beginning of the subsequent loop. Glutamic acid E77 is displayed on the structure for reference. **Top:** AA alignment next to a ribbon diagram of PDB entry 2JD7:A (Tatur et al. 2007), both colored with the ClustalX color scheme (Jeanmougin et al. 1998). **Bottom:** 3Di alignment using Jalview's gecos-Sunset color scheme.

While both alignments have some repetitive structure in the depicted alpha-helical region, the 3Di characters are more conserved and contain information that is distinct from the AAs (Figure 3). For example, in the section of the complete alignment shown in Figure 3, sites 70-71 are both hydrophobic AAs, but do not show obvious grouping information between type 3 and type 1 groups. The corresponding 3Di states show a clear fixed difference at position 70 (V/L vs. S), and invariance at 71 (S). Sites 79-80 show little AA conservation, but the corresponding 3Di sites show fixed differences between the type 1 and type 3 groups. These "fixed" differences do not strictly hold for the full alignment, but type 3 vs. type 1/2 grouping is evident in the 3Di alignment (Figure 2).

Model selection. AA data are typically best fit by Blosum62 or LG, standard matrices for highly diverged AA sequences, or by EX3+FU, a mixture model containing three matrices modeling different transition rates and AA frequencies for exposed, intermediate, and buried sites (Le et al. 2008). Interestingly, LG is selected only for AA data aligned by a structure-free method. All

3Di alignments are best-fit by the 3Di substitution matrix with empirical base frequencies, with minor variations in the treatment of among-site rate variation (G4, R3-R5) (Table 1, Supporting Information).

Table 1. Best-fit substitution models for IQtree runs, according to log-likelihood (lnL) and BIC. IQtree inferences were conducted using the BIC-best model. Runs are sorted by deep & shallow similarity to the M20 tree (first line). Results are shown only for IQtree runs using the custom model list (with partitioning for AA+3Di datasets) which included the 3Di substitution matrix. *k* represents the number of free parameters in the substitution model; all runs also included 103 free branch-length parameters in the BIC calculation. Note: lnL & BIC values can only be compared within lines, not between them, as they are only comparable for identical alignments.

Structure source	Alignment data	Alignment method	Phylogeny data	# bipartitions matched			Best log-likelihood models:			Best BIC models:			
				all	deep	shal-low	lnL	*k*	Model	BIC	lnL	*k*	Model
PDB	structures	Superpose	*Q* score	51	22	29							
alphafold	struct.+AAs	USalign	AA+3di, trim	40	20	20	-24477.7	27	EX3+FU+G; 3DI+F+R3	49728.3	-24477.7	27	EX3+FU+G; 3DI+F+R3
PDB	3di	famsa3di	AA+3di	39	20	19	-39984.6	48	Blosum62+F+G4; 3DI+F+R5	81081.6	-39984.6	48	Blosum62+F+G4; 3DI+F+R5
alphafold	struct.+AAs	USalign	AA+3di	39	20	19	-43208.6	47	Blosum62+F+R3; 3DI+F+R3	87609.2	-43208.6	47	Blosum62+F+R3; 3DI+F+R3
PDB	3di	famsa3di	AA+3di, trim	37	20	17	-29156.4	41	Blosum62+F+G4; 3DI+F+G4	59179.9	-29156.4	41	Blosum62+F+G4; 3DI+F+G4
PDB	struct.+AAs	USalign	AA	37	19	18	-15160.7	25	Blosum62+F+R4	30976.0	-15208.7	4	Blosum62+R3
alphafold	3di	famsa3di	AA+3di	36	19	17	-42057.8	44	Blosum62+F+G4; 3DI+F+R3	85195.4	-42057.8	44	Blosum62+F+G4; 3DI+F+R3
PDB	struct.+AAs	USalign	AA+3di	34	19	15	-40816.9	47	Blosum62+F+R3; 3DI+F+R3	82772.5	-40816.9	47	Blosum62+F+R3; 3DI+F+R3
alphafold	struct.+AAs	USalign	AA, trim	38	18	20	-16270.9	25	Blosum62+F+R4	33187.1	-16315.2	3	EX3+FU+G
PDB	3di	famsa3di	AA	36	18	18	-24783.6	25	Blosum62+F+R4	50404.9	-24792.0	20	Blosum62+F+G4
alphafold	struct.+AAs	USalign	AA	36	18	18	-27095.9	25	Blosum62+F+R4	55106.9	-27107.4	20	Blosum62+F+G4
PDB	struct.+AAs	USalign	AA	36	18	18	-24877.6	25	Blosum62+F+R4	50631.3	-24881.1	23	Blosum62+F+R3
PDB	struct.+AAs	USalign	AA+3di, trim	33	18	15	-23679.1	25	Blosum62+R3; 3DI+F+G4	48115.0	-23679.1	25	Blosum62+R3; 3DI+F+G4
alphafold	3di	famsa3di	AA+3di, trim	36	17	19	-30459.0	27	EX3+FU+G; 3DI+F+R3	61715.6	-30459.0	27	EX3+FU+G; 3DI+F+R3
alphafold	3di	famsa3di	3di	32	17	15	-14950.2	25	3DI+F+R4	30745.2	-14953.6	23	3DI+F+R3
alphafold	struct.+AAs	USalign	3di	31	16	15	-15985.7	25	3DI+F+R4	32886.3	-15986.2	23	3DI+F+R3
alphafold	3di	famsa3di	AA	32	14	18	-26932.0	27	Blosum62+F+R5	54691.6	-26993.3	3	EX3+FU+G
alphafold	3di	famsa3di	3di, trim	31	14	17	-10259.3	27	3DI+F+R5	21218.1	-10260.7	25	3DI+F+R4
PDB	3di	famsa3di	AA, trim	30	14	16	-18633.2	25	Blosum62+F+R4	37938.1	-18641.4	20	Blosum62+F+G4
PDB	struct.+AAs	USalign	3di, trim	23	14	9	-8173.4	25	3DI+F+R4	17006.9	-8174.6	23	3DI+F+R3
PDB	struct.+AAs	USalign	3di	22	14	8	-15612.5	23	3DI+F+R3	32094.1	-15612.5	23	3DI+F+R3
alphafold	struct.+AAs	USalign	3di, trim	28	13	15	-8037.0	25	3DI+F+R4	16740.9	-8039.5	23	3DI+F+R3
PDB	3di	famsa3di	3di, trim	25	13	12	-10215.4	25	3DI+F+R4	21107.5	-10218.1	23	3DI+F+R3
none	AAs	famsa	AA, trim	30	12	18	-24136.9	23	LG+F+R3	49084.3	-24139.0	20	LG+F+G4
alphafold	3di	famsa3di	AA, trim	29	12	17	-20019.9	25	LG+F+R4	40665.8	-20044.4	3	EX3+FU+G
PDB	3di	famsa3di	3di	24	12	12	-14851.9	25	3DI+F+R4	30553.0	-14856.1	23	3DI+F+R3
none	AAs	famsa	AA	29	10	19	-18766.9	25	LG+F+R4	38190.2	-18813.5	1	LG+G4

Standard AA models fit best to AA data, with comparatively small differences between Blosum62, LG, and EX3. A much worse fit is the Poisson model, a Jukes-Cantor-style model which assumes that transition rates are equal between all residues. Finally, the least fitting model for AA data is the 3Di rate matrix. On the other hand, for 3Di data, the pattern is reversed, where the 3Di model dominates, followed by Poisson and then the AA models (Table 2).

Table 2. Illustrative models compared on AA and 3Di datasets. Alignment source: USalign on AlphaFold structures+AAs, trimmed sites with <35% data; phylogeny estimated with IQtree on custom model list including 3DI substitution matrix.

Data	Model	lnL	ΔlnL	*k*	BIC
amino acids (*n* =191 sites)	EX3+FU+G	-16315.2	35.9	106	33187.1
	Blosum62+F+G4	-16279.2	0.0	123	33204.5
	LG+F+G4	-16289.1	9.8	123	33224.2
	EX3+F+G	-16299.1	19.8	125	33254.6
	Poisson+F+G4	-16888.9	609.7	123	34423.9
	3DI+F+G4	-17950.8	1671.5	123	36547.6
3Di characters (*n* =191 sites)	3DI+F+G4	-8054.9	0.0	123	16755.9
	Poisson+F+G4	-8329.4	274.5	123	17304.9
	Blosum62+F+G4	-8485.0	430.0	123	17616.0
	EX3+F+G	-8620.9	566.0	125	17898.3
	LG+F+G4	-8734.1	679.1	123	18114.1
	EX3+FU+G	-10415.9	2361.0	106	21388.6

Phylogenetic analyses. IQtree custom-model runs by the number of deep and shallow bipartitions that match the reference M20 tree are shown in Table 1. Results for all 60 ML runs are shown in Figure 4. Our results indicate that runs using both AA+3Di do best at matching the M20 tree topology, although with substantial overlap with structure-aligned AA runs. 3Di-only runs are consistently lower ranked.

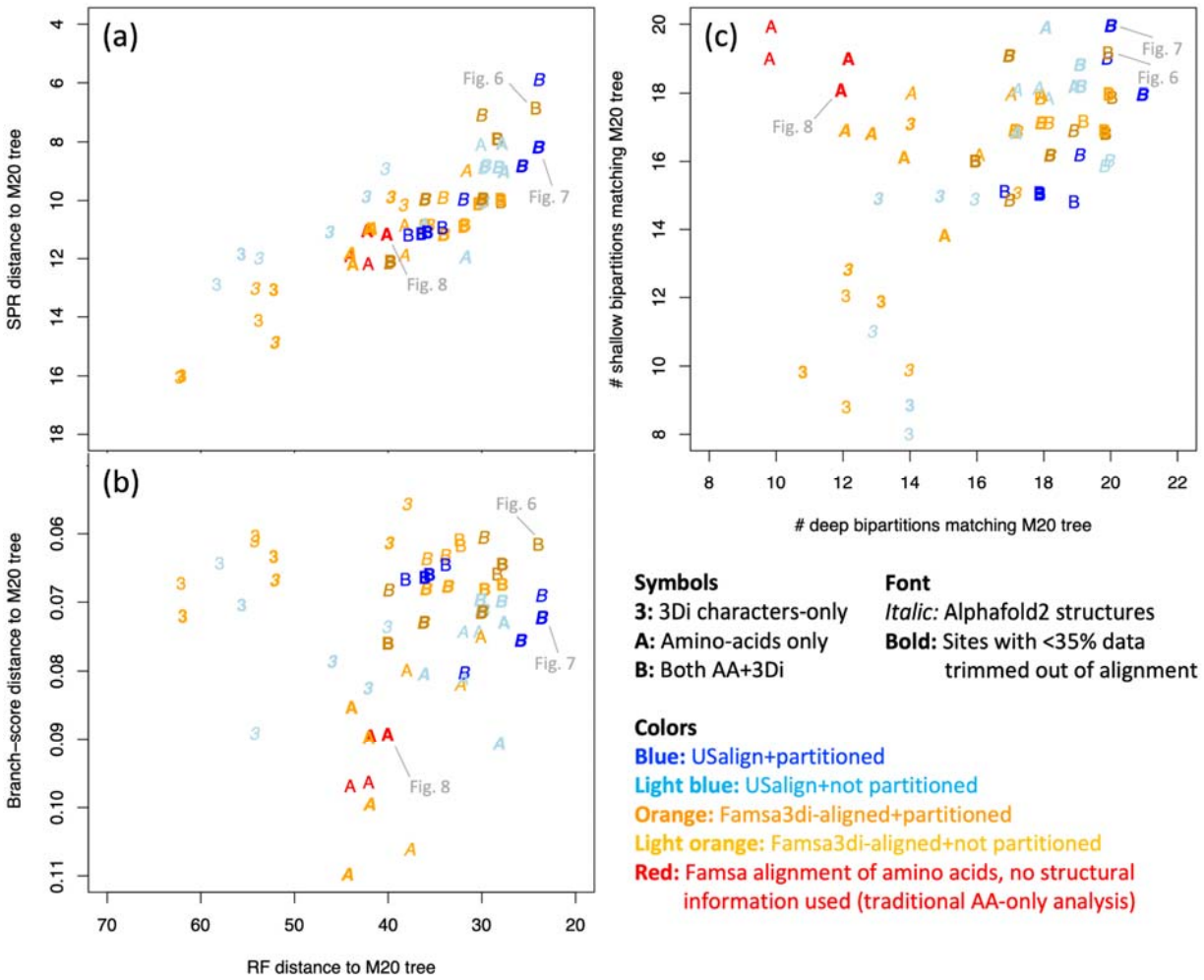


Figure 4. Similarity of all 60 IQtree runs to the Malik et al. (2020) structural-distances NJ tree (M20) for the ferritin-like superfamily. **(a)** Robinson-Foulds (RF) distance vs. Subtree Pruning-Regrafting (SPR) distance. SPR approximates the number of edits required to match the M20 topology. The closest matches to M20 on both metrics tend to be AA+3Di analyses with partitioning. AA-only and 3Di-only analyses tend to fare worse. **(b)** RF distance vs. branch-score distance, which takes branch lengths into account, unlike RF and SPR. 3Di-only match M20 well on the branch-score measure, but poorly on RF; however, AA+3Di performs well on both. **(c)** Number of M20 deep bipartitions matched vs. shallow bipartitions matched. "Deep" bipartitions are those annotated by Malik et al. to indicate major well-recognised categories, and groupings of those. Shallow bipartitions are all others, e.g. groups of 2-5 structures near the tips of the unrooted tree. AA-only runs retrieved shallow bipartitions but not deep ones; 3Di-only was weak on both, but AA+3Di retrieved both shallow and deep. *Note:* Points have been jittered for visibility.

Phylogenetic analyses using 3Di-only tend to be competitive in branch score difference. Structure-aligned AA-only runs are somewhat competitive in RF distance; only both AA+3Di runs are competitive in both (Figure 4b).

Amino-acid-only (structure-free) analyses are competitive for resolving shallow bipartitions, but are systematically poorer for deep bipartitions. 3Di-only analyses are poor for both. However, in combined analyses, the AA and 3Di data together seem to mutually compensate for the

weaknesses of each individually and produce the best matching of both shallow and deep bipartitions (Figure 4c).

Also of interest are cases where differences in method do not seem to produce major differences in performance. It appears that PDB and AlphaFold structural data are equivalent in terms of phylogenetic results matching the M20 tree. Trimming, or not, also does not appear to make a major difference. Model partitioning did not seem to produce a consistent difference in terms of tree similarity to the M20 tree. However, lack of partitioning produces higher bootstrap supports. When bootstrap percent scores are summed across the tree for all 60 models (see SI), all 12 of the AA+3Di runs without partitioning were ranked in the top 15 (sum of bootstraps: 4219-4374). The next 16 runs contain 13 AA+3Di partitioned runs and 3 structure-aligned AA-only runs (sum of bootstraps: 4091-4216). The following 15 runs are all AA-only (3815-4087), with structure-free AA runs representing the last 4 (3815-3924), and the final 14 runs are all 3Di-only (3471-3799). While declining bootstraps for the noisier AA-only and 3Di-only datasets accords with expectations, elevated bootstraps for non-partitioned AA+3Di datasets probably represents artificial elevation of confidence due to a single model poorly fitting two very different partitions.

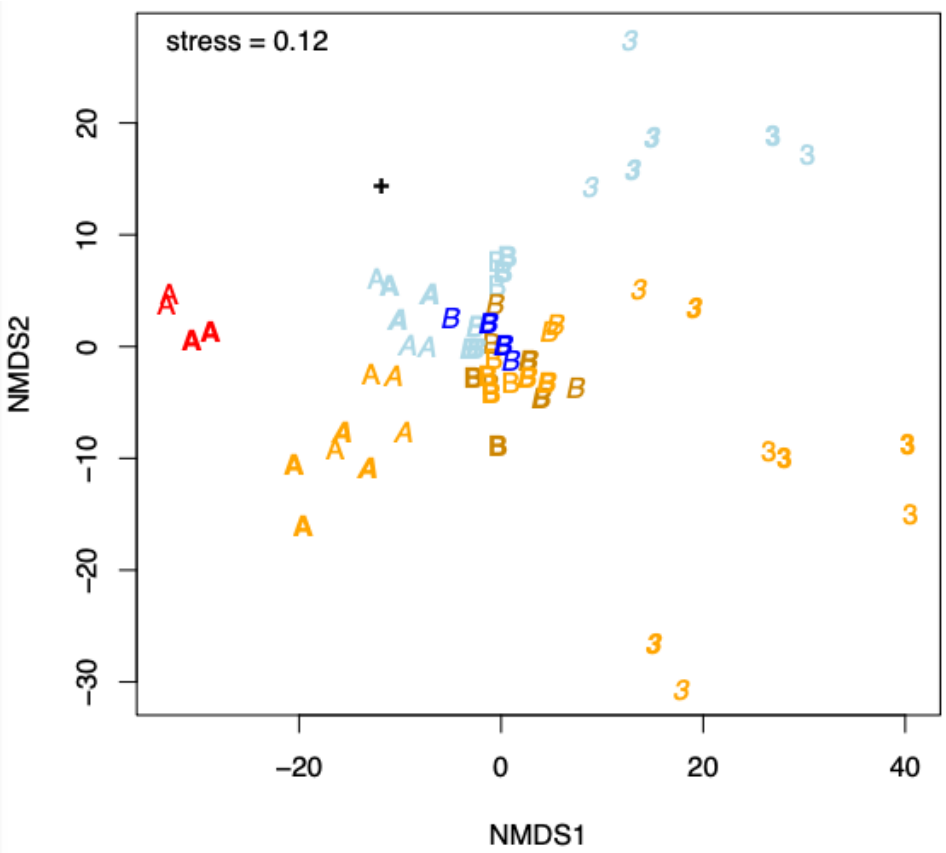


Fig. 5. Nonmetric Multidimensional Scaling (NMDS) representation of the 61x61 matrix of RF distances between all ML trees (60 IQtree runs + the M20 tree). Values <0.1 are typically taken to indicate "excellent," and <0.2 "good," 2-dimensional approximations of the original distances (Clarke 1993). Legend as in Figure 4. M20 tree = **+**. The trees using both AA+3Di data ("B") show a merger of both signals and cluster tightly, compared to the variability between the AA-only or (particularly) 3Di-only trees.

The nonmetric multidimensional scaling (NMDS) analysis of the RF distances between all ML trees and the M20 tree ("+") is shown in Figure 5. An NMDS stress value of 0.12 indicates that AA-only and 3Di-only runs clearly explore different parts of tree space and supports the contention that they are not redundant data sources. The structure-free AA-only runs (Figure 5, red "A") cluster apart from the structure-aligned AA-only runs, suggesting that AA alignments based on structure may pick up phylogenetic signal that structure-free AA alignments miss. 3Di-only analyses are much more scattered than other runs, suggesting larger variation in inferred tree topology, perhaps due to the fact that some 3Di runs were using the default IQtree model list, which does not include a 3Di-specific model. The AA+3Di runs (Figure 5,"B") appear to combine both signals and cluster more tightly than AA-only or 3Di-only runs. The NMDS plot also indicates a difference between alignment methods. USalign-based AA and 3Di alignments (Figure 5, light blue "A" and "3"), all tend to be closer to the M20 tree than famsa3di-based alignments (light orange). This may indicate that USalign, by simultaneously aligning AAs and the 3D backbone of the structures, without considering 3Di states, is more directly capturing structural distance information that is also used by the M20 tree. However, this difference between USalign and famsa3di methods appears to be much reduced in the AA+3Di analyses (Figure 5, "B").

For PDB structural data, the phylogenetic tree that best matched the M20 tree (Figure 6) was generated using famsa3di on 3Di characters generated from the PDB structures. The BIC best-fit models for the AA and 3Di partitions were Blosum62+F+G4 and 3DI+F+R5. This tree shared 19 deep and 20 shallow bipartitions with the M20 tree. Similarity is apparent in both the unrooted (top) and midpoint rooted/re-rooted (bottom) depictions (Figure 6). The most notable topology differences are in the placement of two structures on long, isolated branches in M20: 2fzf_A, a *Pyrococcus furiosus* rubrerythrin domain-containing protein, and 1z60_M, a ferritin chain from the cabbage looper butterfly *Trichoplusia ni* which interacts with 1z60_A from the same species. In the PDB AA+3Di tree, 2fzf_A groups with the other rubrerythrins (albeit with low support, 57%), and 1z60_M is sister to 1z60_A (93% support), suggesting within-structure gene duplication. These are both highly plausible placements, improving on the M20 tree.

The best-match IQtree analysis overall (Figure 7) used AlphaFold structure predictions, jointly aligned structures and AAs using USalign, with Alphafold-derived 3Di characters then mapped to the aligned AA positions. The alignment was trimmed, and the BIC best-fit models for the AA and 3Di partitions were EX3+FU+G and 3DI+F+R3. This tree shared 20 deep and 20 shallow bipartitions with the M20 tree. This analysis also groups 2fzf_A with other rubrerythrins (similar low support, 60%) and 1z60_M with 1z60_A (70% support). The other major difference is that the dimer geometry type 2 group of fatty acid desaturases is nested within the type 1 group, as sister to RNR R2, although with low support (69%).

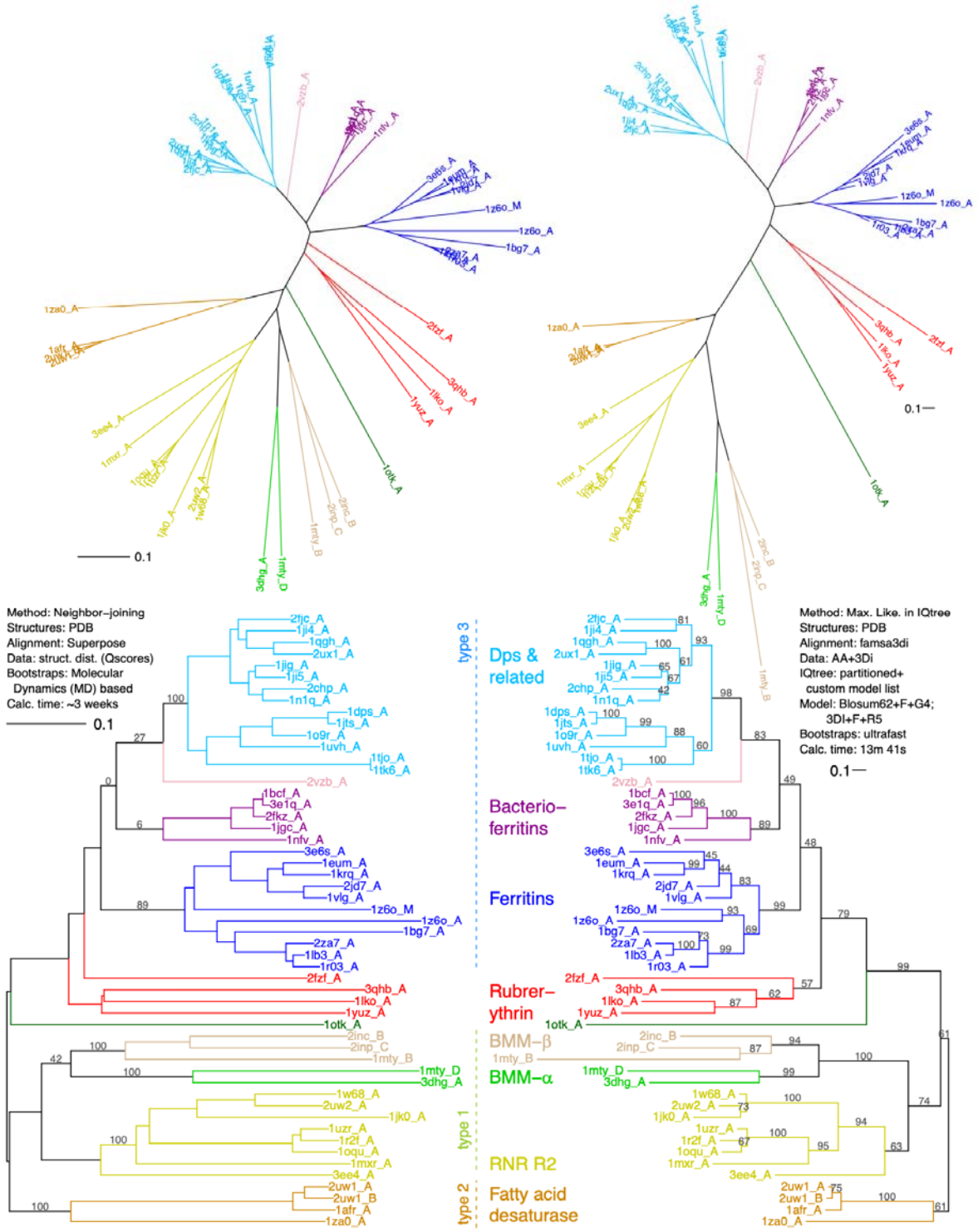


Figure 6. Comparison of trees using PDB structures. **Left:** M20 structural distances tree, with MD-based bootstraps where available in Malik et al. 2020. **Right:** the closest-match PDB-based IQtree analysis using AAs+3Dis from PDB structural data. This tree shares 20 deep and 19 shallow bipartitions with M20. **Top row:** unrooted trees. **Bottom:** arbitrary rooting for visual comparison. Branch length units for the M20 tree are Q_{score}-based distances; for the ML tree they are expected substitutions per site. The types refer to dimer geometries, likely slowly evolving structural characters (Lundin et al. 2012).

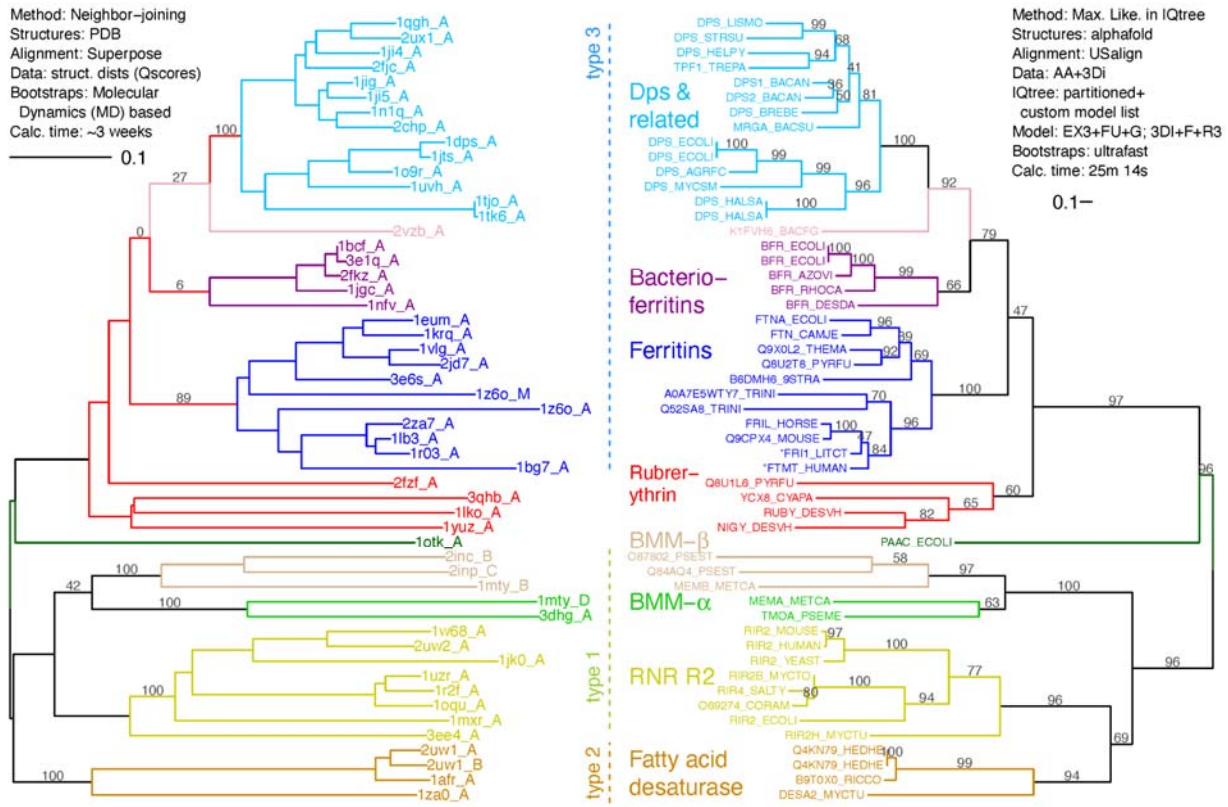


Figure 7. AlphaFold-based ML AA+3Di tree (right) compared to the M20 structural distances tree (left). This ML run was the closest overall topology match to the M20 tree, and shares 20 deep and 20 shallow bipartitions with M20. Tip labels are Uniprot IDs for which AlphaFold predictions were downloaded. Each Uniprot label on the right represents the closest AA match to the corresponding PDB structure on the left, with the exception of two tips marked with asterisks; here, FRI1_LITCT corresponds to 1bh7_A, and FTMT_HUMAN corresponds to 1r03_A. The types refer to dimer geometries, likely slowly evolving structural characters (Lundin et al. 2012).

The resulting topology from a “standard”, sequence-only, structure-free alignment (Figure 8) shares 19 shallow bipartitions with M20, but only 12 deep ones. Severe long-branch attraction effects are evident; instead of 2fzf_A grouping with the rubrerythrins and 1z60_A with 1z60_M and other ferritins, they instead form a group of oddballs along with the long-branched 1otk_A. The BMM-β group breaks apart and renders fatty acid desaturases and ferritins paraphyletic.

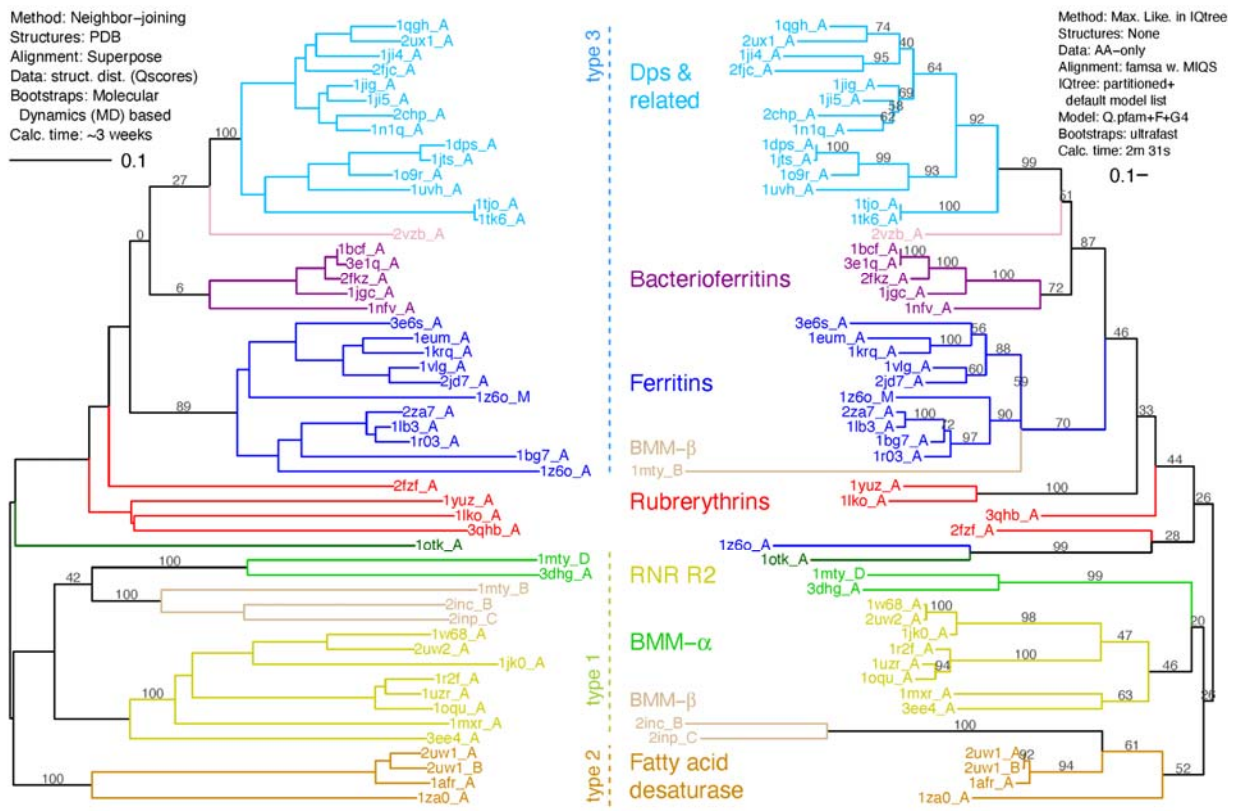


Figure 8. Structure, AA-only ML tree (right) compared to the M20 structural distances tree (left). ML tree is from the closest match "standard" AA-only analysis. The tree shares 12 deep and 19 shallow bipartitions with M20. The types refer to dimer geometries, likely slowly evolving structural characters (Lundin et al. 2012).

In terms of computational run-time, these results are a dramatic improvement. The Malik et al. 2020 MD-based bootstrapping of ferritins took approximately 3 weeks of wall-clock time to run, even making use of parallel processing, as each of the 53 structures had simulate 100 nanoseconds (ns) of conformational change with Molecular Dynamics. Even longer simulations (>500 ns) would be preferred in a new analysis, which would imply a 5x increase in runtime. Although this may be ameliorated with newer technology and greater cluster resources when available. On the other hand, the focal PDB and AlphaFold AA+3Di IQtree runs, complete with partitioning, automated model selection, and 1000 ultrafast bootstraps, took 13.6 and 25.25 minutes on a standard desktop computer. The latest approach makes structural phylogenetics more accessible since it can be performed with small computational resources.

Discussion

Although protein structural phylogenetics has been explored for decades, its use has been limited by the availability of structural data, the lack of ready-to-use inference methods other than distance-based neighbor-joining, and the challenges of integrating distance-based methods into the model-based framework of modern statistical phylogenetics. Highly desirable tools like parametric model fitting, statistical model comparison, Maximum Likelihood and

Bayesian inference, statistical measurement of uncertainty through bootstrapping, posterior probabilities or other techniques, integration of multiple data sources in a unified statistical framework, and simulation-based *in silico* tests of inference methods, all remained challenging in a distance-based framework.

The combination of two developments, namely AlphaFold and the Foldseek tertiary interaction encoder, raises the prospect that this barrier has suddenly fallen, and the way is open for the emergence of model-based structural phylogenetics. Our study represents a proof-of-concept: a complete pipeline stepping from amino acid sequences, to AlphaFold structures, to structure-informed alignments, to 3Di character datasets that appear to be well-behaved, and helpful, in statistical model selection (Table 2) and phylogenetic inference (Figs. 4-8).

In the case of the ferritin-like superfamily, the application of this pipeline produced estimated phylogenies, with formal ultrafast bootstrapping estimates of uncertainty, in minutes instead of the weeks required in the previous state-of-the-art analysis of Malik et al. (2020). Across the different IQtree run conditions, the phylogenies that were the best topological matches to the Malik et al. (2020) were the ones that one might predict *a priori* to be phylogenetic best-practice (Table 1): combining both AA and 3Di alignments, using structure to inform the alignment, using a custom model list that included a 3DI-specific substitution model, and using statistical model comparison on the partitioned dataset to find the best-fit model separately for AA and 3Di characters. Combining AA+3Di characters seemed to ameliorate long-branch attraction artifacts found in AA-only analyses (Figs. 6-8), and also produced closer matching to the Malik et al. tree on the branch-score distance (Fig. 4b). Probably these are related phenomena, suggesting that AA+3Di analyses are less affected by saturation of sequence divergence on deep branches. All of these results are just what would be expected when a set of characters even more conserved than amino acids are added to an amino acid dataset.

In summary, our results suggest that the inferred structure-informed phylogenies are comparable, and perhaps superior, to a previously-published phylogeny of the ferritin-like superfamily using structural distances. The addition of 3Di characters to ML phylogenetic inference is appealing because of the approach's ease-of-use, computational speed, and potential to add highly-conserved, slowly-evolving characters just where they are typically needed: deep in phylogenies, where AA sequence similarity is highly decayed. Therefore, it is tempting to suggest that this approach can, and should, be tried on almost every tough problem in protein phylogenetics. While we do encourage researchers to explore and build on this approach, our study highlights a number of considerations that should be kept in mind.

In the case of the ferritin-like superfamily, AlphaFold-predicted structures have high confidence scores (see Supporting Information) and performed equally well to PDB structures in our analyses. However, while AlphaFold has been a huge step forward in structural prediction, some limitations have become evident. AlphaFold predictions will tend to be more accurate on structures more similar to its (admittedly vast) training set. Protein types that have fewer solved high-resolution structures (membrane-embedded proteins or those with no stable fold) will have more uncertain AlphaFold predictions and therefore will presumably have noisier 3Di

characters. Intrinsically disordered proteins/regions/loops are also poorly predicted by AlphaFold (Ruff & Pappu 2021; Bertoline et al. 2023), although such regions are likely to be less conserved anyway. The AlphaFold database also has the limitation of holding only structural predictions of monomers. Interactions with cofactors, chaperones, and/or within multimers will all potentially change a predicted structure (Perrakis & Sixma 2021) and thus any extracted characters. Interactions can be modeled with AlphaFold, but for the moment this is a bespoke process rather than a precalculated database.

Unresolved questions about 3Di characters. The prospect of estimating phylogenies with the assistance of a 3Di character dataset that might change based on how protein structures are modeled and predicted may give pause to many molecular phylogeneticists who are used to sequence data being fixed and highly confident (leaving aside cases of sequencing error and contamination). A broader focus on the kinds of characters that have been used in phylogenetics generally might help bridge this gap. In organismal biology and paleontology, sources of characters have included morphology of bones as well as soft tissue, ultrastructure, characters observable only in a particular sex (antlers) or embryonic developmental stage (post-anal tail), characters that are revealed only after substantial processing or imaging of a specimen (biochemical fossils, internal bone structure), or even ephemeral events like bird songs or stereotyped mating behaviors. Just like 3Di characters, many of these other data sources produce characters that are dependent on a particular ascertainment method. As long as an ascertainment method can be described and replicated by other researchers, it is a potentially valid data source, and its utility for phylogenetics becomes an empirical question. For a novel data source, tree structure in character data, phylogenetic congruence with other evidence, and reliability of character coding under perturbation of the input material are all standard research questions that can be answered statistically (Matzke 2015).

A case can be made that 3Di states are actually easier to accept as phylogenetic characters than routinely accepted data sources like discrete morphological characters. Morphologists often encounter difficult questions such as: How should continuous data be discretized? Is a character state not observed because it is not present in the organism, or is it missing due to specimen incompleteness, decay or damage (taphonomy; Sansom et al. 2010), the specimen being a juvenile, or population-level, environmental, or random variation in development? Sometimes there is an element of subjectivity in the judgment calls made in morphological character coding that can be difficult to completely remove. In comparison, 3Di characters have no such difficulties. For a given experimental or AlphaFold-predicted structure file, the output 3Di characters are a deterministic result that is easily repeatable by any researcher. It is, of course, likely that future improvements in structural prediction and in tertiary-interaction encoding will produce variant datasets, but these will also be replicable and subject to statistical tests for phylogenetic signal. Replicability will require that any future iterations on structural character encoding schemes are given clear version numbers that are reliably reported by researchers, and also that old versions of encoders are kept available in software packages for backwards comparability.

Despite the good arguments for the validity of 3Di states as characters, many questions do need further research. We caution that it remains unclear how sensitive the 3Di encoding is to small perturbations in the protein structure, such as indels, or minor errors made during protein structure solving or prediction. As 3Di characters are based on minimum distances between residues, small changes in one region of the protein might sometimes induce rippling effects that perturb the 3Di characters elsewhere, potentially disrupting the phylogenetic signal. However, the success of Foldseek for remote homology search, and our results showing high conservation over vast evolutionary distances (Fig. 2), as well as congruence between PDB- and AlphaFold-derived phylogenies (Figs. 6-7), suggest that 3Di characters are likely to be fairly robust to perturbations. We would, however, caution that it may be hazardous to "mix and match" structures from different sources: for example, a dataset consisting of some PDB experimental structures, some AlphaFold-database monomer predictions, and some bespoke AlphaFold predictions of chains in a multimeric conformation might result in systematic biases that end up grouping structures by method rather than by history. (On the other hand, it is also possible that in some structures, the evolutionary-conserved tertiary structure is best observed/predicted in the multimeric state.)

Other caveats. Model-based phylogenetic estimation using discretized structural characters is novel, and several classic questions in statistical phylogenetics need to be revisited when the methods are applied to 3Di characters. First, programs like IQtree employ the standard phylogenetic assumption of site independence in their likelihood calculations. This assumption allows the log-likelihood of each site in the alignment to be calculated independently; summing across sites allows rapid calculation of the total log-likelihood. The very definition of 3Di characters as descriptors of tertiary interactions between pairs of residues identified as closest neighbors arguably violates this assumption, although the question is complex as neighbor relationships do not have to be symmetric. Violations of the site independence assumption in other data are already well-known; for example, in coevolution of AA sites (Korber et al. 1993), or in co-variation of morphological characters due to a shared developmental pathway or convergence on an ecological role. These are usually not considered fatal for phylogenetic inference, but co-dependence may be even greater for 3Di characters. Failure to account for site correlation may lead to spurious results by placing overconfidence in the wrong tree (i.e., inflated bootstrap or posterior supports), as information is being double counted (Nasrallah et al. 2011).

IQtree also assumes independence between AA and 3Di sequences. While inspection of AA and 3Di alignments proves that they are far from identical, and clearly have some nonoverlapping grouping information, ideally any nonindependence between the datasets could be measured and corrected for.

Another limitation concerns phylogenetic information that 3Di characters may be missing. 3Dis describe the tertiary relationships between segments of a structure's backbone. This approach does not directly capture structural changes that only affect side chains. In addition, the dihedral angles that directly describe the changes in backbone orientation from one residue to the next are not directly captured by 3Di. Whether these additional kinds of structure-derived datasets

could be combined with AAs and 3Di in a total-evidence analysis, or this would only exacerbate the non-independence problem, is an open question.

Future outlook. Alongside the methodological questions above, 3Di characters present phylogeneticists with a number of additional research opportunities. Just as DNA models advanced from Jukes-Cantor to GTR and codon models, and just as AA models advanced from PAM and BLOSUM matrices to LG and mixture models, character-based structural phylogenetics is likely to advance from the Poisson and 3DI matrices to matrices optimized for different kinds of proteins and protein regions. It is likely that the frequencies of different 3Di character states change between alpha-helix and beta-sheet regions of proteins, so an obvious first step would be to explore the potential of partitioning a 3Di dataset on a protein family containing both (the ferritin fold is largely alpha-helical).

The speed of ML analyses including 3Di structural information also means that Bayesian structural phylogenetics is clearly feasible. This would allow the incorporation of additional modeling flexibility and sources of information, including relaxed clocks, absolute and relative (Shih and Matzke, 2013) time constraints, joint inference of substitution model and tree, and potentially even uncertainty in structural prediction and the derived 3Di characters.

Finally, while *a priori* considerations and our proof-of-concept study strongly suggest that the addition of 3Di characters to AA datasets will improve phylogenetic inference, especially deep in the twilight zone, the ideal study would test this hypothesis in a situation where the true phylogenetic tree is known. While an *in silico* test would be relatively easy to perform under the model assumed by IQtree, where 3Dis are just another 20-state character where every site evolves independently under the 3DI rate matrix, this form of simulation-inference experiment is only of limited interest as it does not model the evolution of the 3-dimensional protein structure. While models of structural evolution are beginning to be developed, a realistically complex forward model of protein structural evolution remains a major challenge. An alternative source of known phylogenies could be experiments, but this would require a system where AA sequence identity could decay from 100% to <30% in the course of the experiment, all while retaining substantial structural similarity. It may be that this could be achieved in a rapidly-evolving virus system, although whether this would imitate protein structural evolution at billion-year timescales could be debated.

Apart from methodological investigations, the major research opportunity presented by character-based structural phylogenetics is a potential revolution in the study of deep phylogenetics questions stretching back before the Last Universal Common Ancestor. The origin and evolution of protein families and fundamental cellular machinery can potentially be addressed or re-addressed with more data and more confidence. The deepest, hardest to resolve relationships – the relationships of the bacterial phyla, of archaea and eukaryotes, and of different classes of viruses, may all become more accessible. However, it may also be that structural characters will be useful even for improving the resolution of shallower relationships, as reported by Moi et al. (2023). We think it likely that structural phylogenetics will soon move from revolutionary to routine.

Acknowledgements and Funding

This collaboration emerged from the inaugural Australasian Structural Phylogenetics Meeting (ASPM2023) that took place at the University of Auckland on October 24th and 25th, 2023. The meeting was supported by the Centre for Computational Evolution (CCE) and the Centre for Cellular and Molecular Physiology (CMP), School of Biological Sciences at the University of Auckland. CPL, MABB and NJM are supported by Human Frontier Science Program Grant No. RGY0072/2021. NJM was additionally supported by NZ RSTA grants 21-UOA-040 & 18-UOA-034. JD is supported by the Alfred P. Sloan Foundation Matter-to-Life program Grant number G2021-16944. DBA is supported by the investigator grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405] and the Victorian government Operational Infrastructure Support program. DL is supported by grants from Vetenskapsrådet 2021-03992 and Knut och Alice Wallenbergs stiftelse 2019.0436 to Martin Högbom.
MSF is supported by Marsden Fast Start #3726031. MS acknowledges the support by the National Research Foundation of Korea, grants [2020M3-A9G7-103933, 2021-R1C1-C102065, 2021-M3A9-I4021220], Samsung DS research fund and the Creative-Pioneering Researchers Program through Seoul National University.

Supporting Information

Data files, analysis scripts, results tables, and alignment and IQtree run inputs and outputs are available at https://github.com/nmatzke/3diphy or as the 191 MB zipfile *3diphy.zip*. The README file in the main directory points to key files/analyses mentioned in the main text: *_00_README_SuppInf_files_v2.txt*.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-3402.

Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research,* 48(D1), D376-D382. doi: 10.1093/nar/gkz1064

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research,* 28: 235-242 doi: 10.1093/nar/28.1.235.

Bertoline LMF, Lima AN, Krieger JE, Teixeira SK (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Front Bioinform*. 3:1120370. doi: 10.3389/fbinf.2023.1120370

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., & Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26), 8932-8937.

Bryant, D. (2005). On the Uniqueness of the Selection Criterion in Neighbor-Joining. *Journal of Classification* 22, 3-15. doi: 10.1007/s00357-005-0003-x

Bujnicki, J. M. (2000) Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures. *J. Mol. Evol*. 50, 39-44.

Breitling, R., Laubner, D., and Adamski, J. (2001) Structure-based phylogenetic analysis of short-chain alcohol dehydrogenases and reclassification of the 17-hydroxysteroid dehydrogenase family. *Mol. Biol. Evol*. 18, 2154-2161.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25(15), 1972-1973. doi: 10.1093/bioinformatics/btp348

Chandonia JM, Guan L, Lin S, Yu C, Fox NK, Brenner SE. (2022). SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning *Nucleic Acids Research* 50:D553-D559. doi: 10.1093/nar/gkab1054.

Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol*. 24(8), 1769-1782. doi: 10.1093/molbev/msm097

Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5(4), 823-826. doi: 10.1002/j.1460-2075.1986.tb04288.x

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* 18: 117-143. doi:10.1111/j.1442-9993.1993.tb00438.x

Chung, Su Yun; Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Current Biology*, 4:1123-1127.

Doolittle, R.F. (1986). *Of Urfs and Orfs: Primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA, USA.

Fenton, H.J.H. (1894) Oxidation of tartaric acid in presence of iron. *Journal of the Chemical Society,* Transactions, 65, 899-910. doi: 10.1039/ct8946500899

Flores, T.P.; Orengo, C.A.; Moss, D.S.; Thornton, J.M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science* 2, 1811-1826.

Garau G, Di Guilmi AM, Hall BG. Structure-based phylogeny of the metallo-beta-lactamases. *Antimicrob Agents Chemother*. 49(7), 2778-2784. doi: 10.1128/AAC.49.7.2778-2784.2005

Gascuel O, Steel M. Neighbor-joining revealed. *Mol Biol Evol*. 23(11), 1997-2000. doi: 10.1093/molbev/msl072

Golden, M., García-Portugués, E., Sørensen, M., Mardia, K. V., Hamelryck, T., & Hein, J. (2017). A generative angular model of protein structure evolution. *Molecular Biology and Evolution*, *34*(8), 2085-2100.

Goldman N, Thorne JL, Jones DT. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol*. 263(2), 196-208. doi: 10.1006/jmbi.1996.0569

Goldman N, Thorne JL, Jones DT. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*. 149(1), 445-458. doi: 10.1093/genetics/149.1.445

Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*. 22(21), 2695-2696. doi: 10.1093/bioinformatics/btl461.

Grant BJ, Skjaerven L, Yao XQ. The Bio3D packages for structural bioinformatics. *Protein Sci*. 30(1), 20-30. doi: 10.1002/pro.3923

Guindon, Stéphane; Dufayard, Jean-François; Lefort, Vincent; Anisimova, Maria; Hordijk, Wim; Gascuel, Olivier (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), 307-321. doi: 10.1093/sysbio/syq010

Hasegawa H, Holm L. Advances and pitfalls of protein structural alignment. *Curr Opin Struct* Biol. 19(3), 341-348. doi: 10.1016/j.sbi.2009.04.003

Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research* 25(1), 231-234. doi: 10.1093/nar/25.1.231

Huelsenbeck, J. P., Bull, J. J., & Cunningham, C. W. (1996). Combining data in phylogenetic analysis. *Trends in Ecology & Evolution*, 11(4), 152-158. doi: 10.1016/0169-5347(96)10006-9

Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol*., 23(2), 254-267.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., & Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends in Biochemical Sciences*, *23*(10), 403-405.

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. Highly accurate protein structure prediction with AlphaFold. *Nature*. 596(7873), 583-589. doi: 10.1038/s41586-021-03819-2

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22(12), 2577-2637. doi: 10.1002/bip.360221211

Katoh, Kazutaka, et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucleic Acids Research* 30(14) (2002), 3059-3066.

Korber, Bette T., et al. "Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis." *Proceedings of the National Academy of Sciences* 90(15), 7176-7180.

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr.* 60(Pt 12 Pt 1), 2256-2268. doi: 10.1107/S0907444904026460

Lake JA, Henderson E, Oakes M, Clark MW. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proceedings of the National Academy of Sciences* 81(12), 3786-3790. doi: 10.1073/pnas.81.12.3786.

Lai, Jhih-Siang 2020. *Protein structural phylogeny, a missing chapter in molecular evolutionary biology*. Ph.D. thesis, University of Queensland. doi: 10.14264/uql.2020.984

Lai, J-S, Rost, B, Kobe, B, Bodén, M. (2020). Evolutionary model of protein secondary structure capable of revealing new biological relationships. *Proteins*. 88, 1251-1259. https://doi-org/10.1002/prot.25898

Le, Si Quang; Lartillot, Nicolas; Gascuel, Olivier (2008). Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B*. 363, 3965-3976. http://doi.org/10.1098/rstb.2008.0180

Le SQ, Gascuel O. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol*. 59(3), 277-287. doi: 10.1093/sysbio/syq002

Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol Biol Evol*. 32(10), 2798-2800. doi: 10.1093/molbev/msv150

Lundin D, Poole AM, Sjöberg BM, Högbom M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *J Biol Chem*. 287(24), 20565-20575. doi: 10.1074/jbc.M112.367458.

Malik, A. J., Poole, A. M., & Allison, J. R. (2020). Structural Phylogenetics with Confidence. *Molecular Biology and Evolution*, 37(9), 2711-2726. doi: 10.1093/molbev/msaa100

Matzke, Nicholas J.; Shih, Patrick M.; Kerfeld, Cheryl A. (2014). "Bayesian Analysis of Congruence of Core Genes in *Prochlorococcus* and *Synechococcus* and Implications on Horizontal Gene Transfer." *PLoS One*, 9(1), e85103. doi: 10.1371/journal.pone.0085103

Matzke, Nicholas (2015). The Evolution of Antievolution Policies After *Kitzmiller v. Dover*. *Science*, 351(6268), 10-12. doi: 10.1126/science.aad4057

Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 29(21), 2722-2728. doi: 10.1093/bioinformatics/btt473

Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci*. 32(11), e4792.

Minh, B. Q., Lanfear, R., Ly-Trong, N., Trifinopoulos, J., Schrempf, D., & Schmidt, H. A. (2022). IQ-TREE version 2.2.0: Tutorials and Manual. Phylogenomic software by maximum likelihood. March 25, 2022. http://www.iqtree.org/doc/iqtree-doc.pdf

Moi, David; Bernard, Charles; Steinegger, Martin; Nevers, Yannis; Langleib, Mauricio; Dessimoz, Christophe (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *bioRxiv* 2023.09.19.558401. doi: 10.1101/2023.09.19.558401

Nasrallah, C. A., Mathews, D. H., & Huelsenbeck, J. P. (2011). Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Systematic Biology,* 60(1), 60-73.

Perrakis A, Sixma TK. AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Rep.* 22(11), e54046. doi: 10.15252/embr.202154046

RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning (2023) *Nucleic Acids Research* 51: D488–D508 doi: 10.1093/nar/gkac1077

Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. *J Mol Biol.* 433(20), 167208. doi: 10.1016/j.jmb.2021.167208

Salvador Capella-Gutierrez; Jose M. Silla-Martinez; Toni Gabaldon. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25: 1972-1973.

Rajapaksa, Sandun; Konagurthu, Arun S.; Lesk, Arthur M. (2023). Sequence and structure alignments in post-AlphaFold era, *Current Opinion in Structural Biology*, 79, 102539. doi: 10.1016/j.sbi.2023.102539.

Sansom RS, Gabbott SE, Purnell MA. Non-random decay of chordate characters causes bias in fossil interpretation. *Nature*. 463(7282), 797-800. doi: 10.1038/nature08745

Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics*. 27(4) 592-593. doi: 10.1093/bioinformatics/btq706

Shih, Patrick M.; Matzke, Nicholas J. (2013). "Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins." *Proceedings of the National Academy of Sciences*, 110(30), 12355-12360.

Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, Tolulope A, Scholes HM, Senatorov I, Bujan A, Ceballos Rodriguez-Conde F, Dowling B, Thornton J, Orengo CA. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res*. 47(D1), D280-D284. doi: 10.1093/nar/gky1097

Skolnick J, Gao M, Zhou H, Singh S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. *J Chem Inf Model*. 61(10), 4827-4831. doi: 10.1021/acs.jcim.1c01114

Tammi, Martti T. (2021). Construction of substitution matrices. Bioinformaticshome.com. Accessed 2021-06-01. https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices.html

Tatur, J.; Hagen, W. R.; Matias, P. M. (2007). Crystal structure of the ferritin from the hyperthermophilic archaeal anaerobe *Pyrococcus furiosus*. *JBIC Journal of Biological Inorganic Chemistry*, *12*, 615-630. doi: 10.1007/s00775-007-0212-3

Trinquier, J.; Petti, S.; Feng, S.; Söding, J.; Steinegger, M.; Ovchinnikov, S. (2022). SWAMPNN: End-to-end protein structures alignment. *Machine Learning for Structural Biology Workshop*, NeurIPS, 2022. https://www.mlsb.io/papers_2022/SWAMPNN_End_to_end_protein_structures_alignment.pdf

Thorne JL, Goldman N, Jones DT. Combining protein evolution and secondary structure. *Mol Biol Evol*. 13(5), 666-673. doi: 10.1093/oxfordjournals.molbev.a025627.

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. doi: 10.1038/s41587-023-01773-0

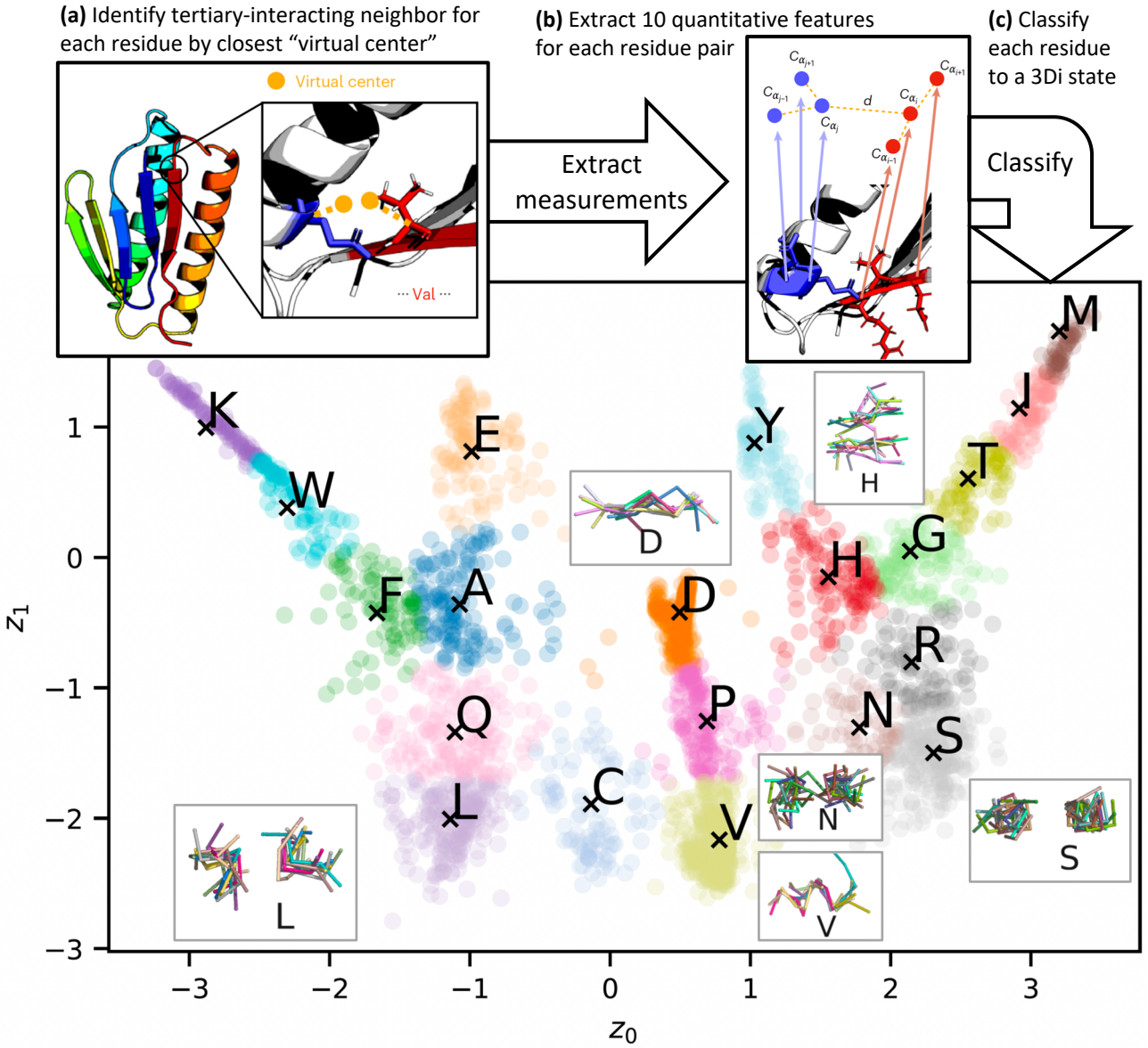
Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*. 50(D1), D439-D444. doi: 10.1093/nar/gkab1061

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. (2023). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, gkad1011. doi: 10.1093/nar/gkad1011

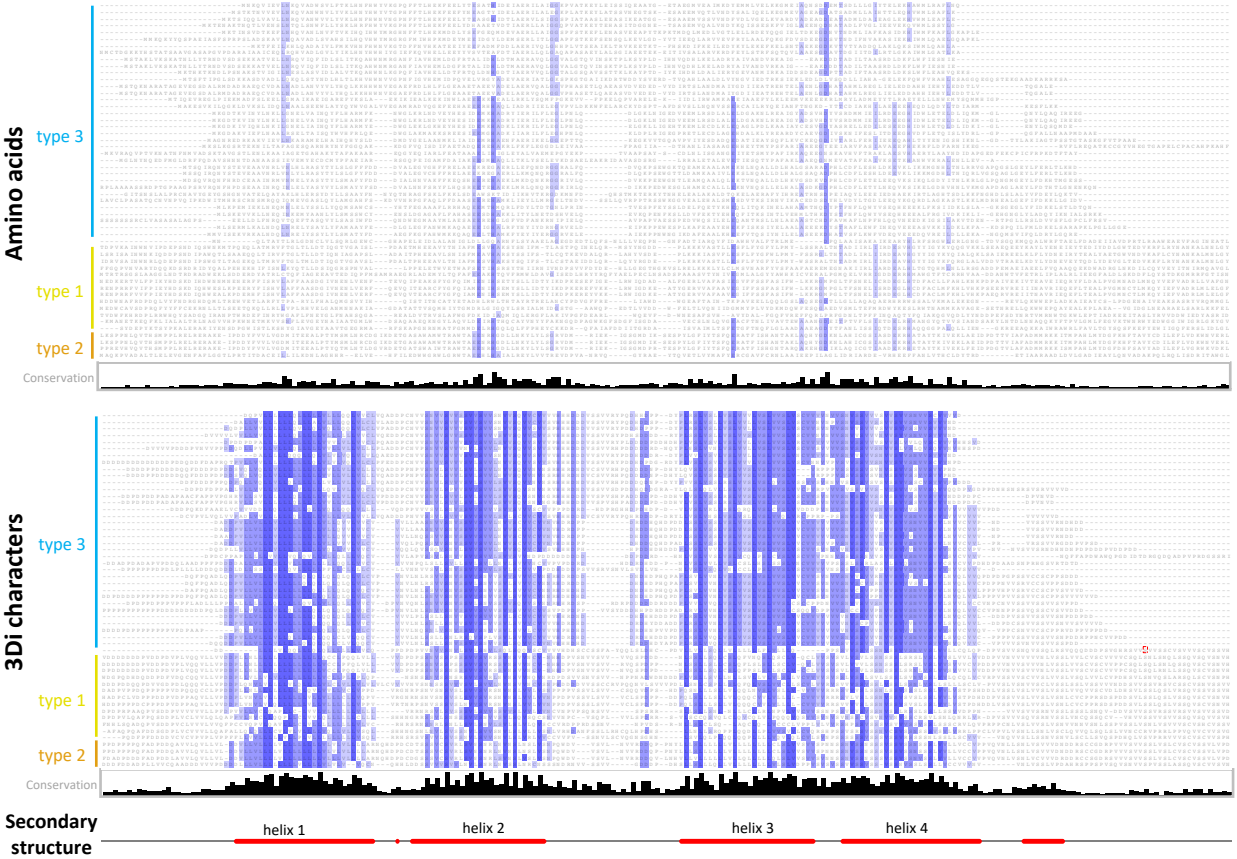
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009). Jalview Version 2 - A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25 1189-1191. doi:10.1093/bioinformatics/btp03

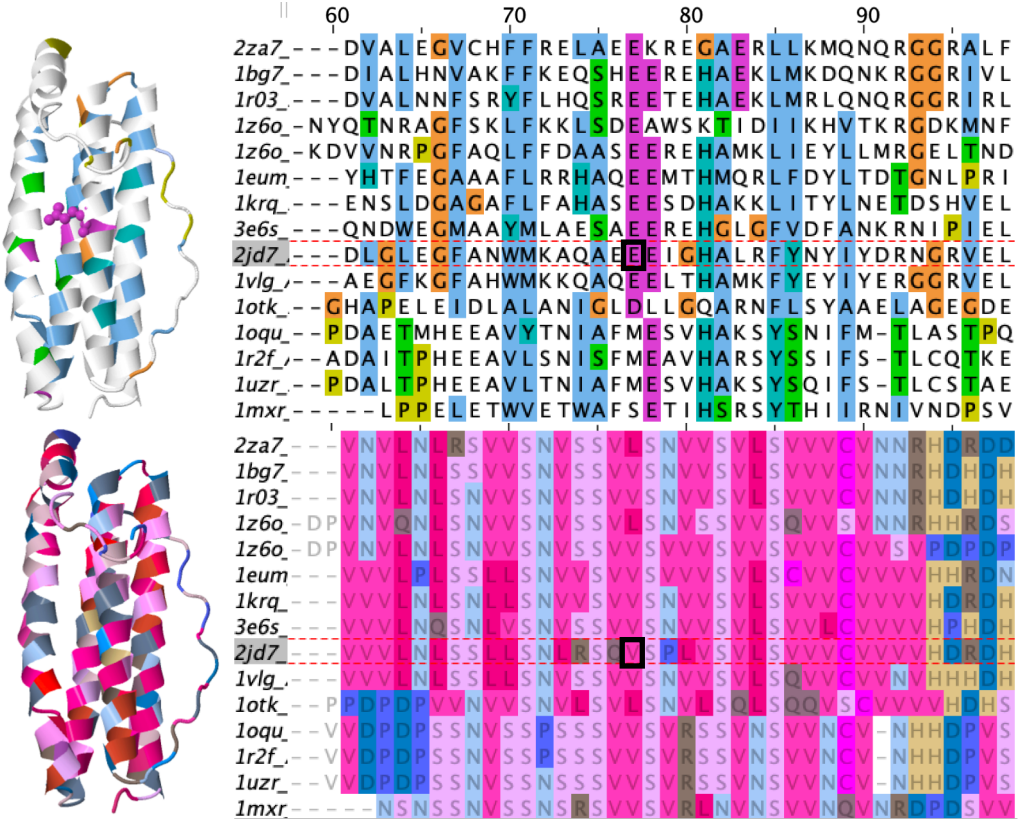
Zhang C, Shine M, Pyle AM, Zhang Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods*. 19(9), 1109-1115. doi: 10.1038/s41592-022-01585-1.

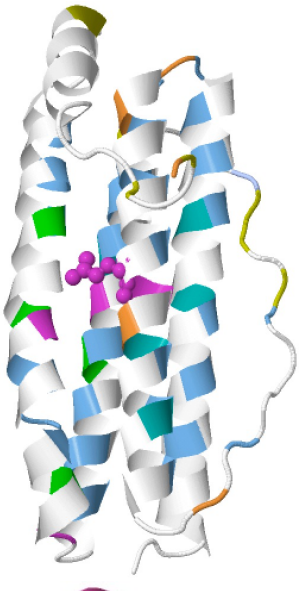
Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7), 2302-2309.

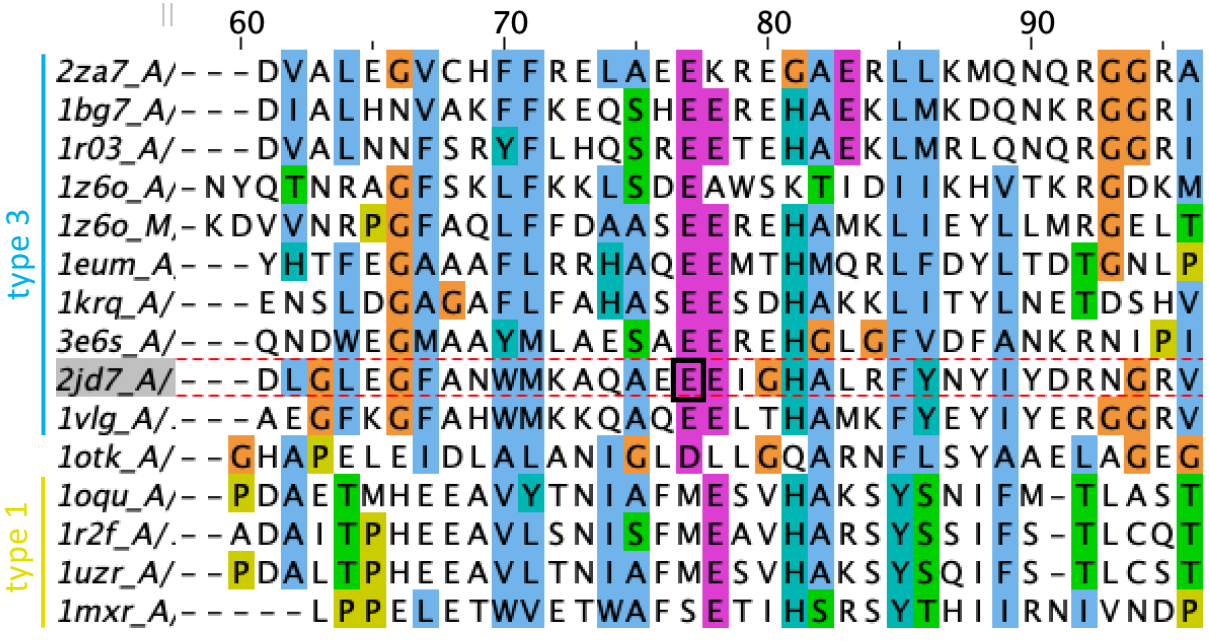


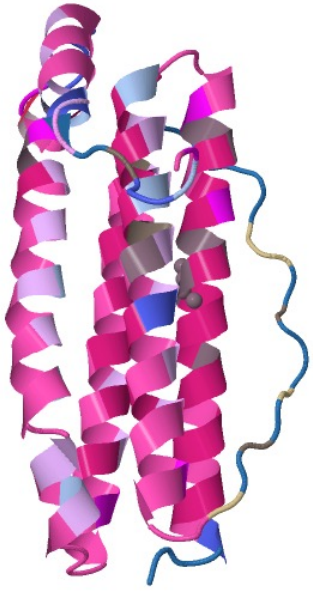
(a) Identify tertiary-interacting neighbor for each residue by closest "virtual center"
Virtual center
··· Val ···
(b) Extract 10 quantitative features for each residue pair
Extract measurements
(c) Classify each residue to a 3Di state
Classify
K
W
E
F
A
Q
L
C
D
P
V
Y
H
G
R
N
S
T
I
M
z_1
z_0


Amino acids
type 3
type 1
type 2
Conservation
3Di characters
type 3
type 1
type 2
Conservation
Secondary structure
helix 1
helix 2
helix 3
helix 4


60
70
80
90
2za7_ DVALEGVCHFFRELAEEKREGAERLLKMQNQRGGRALF
1bg7_ DIALHNVAKFFKEQSHEEREHAEKLMKDQNKRGGRIVL
1r03_ DVALNNFSRYFLHQSREETEHAEKLMRLQNQRGGRIRL
1z6o_ NYQTNRAGFSKLFKKLSDEAWSKTIDIIKHVTKRGDKMNF
1z6o_ KDVVNRPGFAQLFFDAASEEREHAMKLIEYLLMRGELTND
1eum_ YHTFEGAAAFLRRHAQEEMTHMQRLFDYLTDTGNLPRI
1krq_ ENSLDGAGAFLFAHASEESDHAKKLITYLNETDSHVEL
3e6s_ QNDWEGMAAYMLAESAEEREHGLGFVDFANKRNIPIEL
2jd7_ DLGLEGFANWMKAQAEEEIGHALRFYNYIYDRNGRVEL
1vlg_ AEGFKGFAHWMKKQAQEELTHAMKFYEYIYERGGRVEL
1otk_ GHAPELEIDLALANIGLDLLGQARNFLSYAAELAGEGDE
1oqu_ PDAETMHEEAVYTNIAFMESVHAKSYSNIFM-TLASTPQ
1r2f_ ADAITPHEEAVLSNISFMEAVHARSYSSIFS-TLCQTKE
1uzr_ PDALTPHEEAVLTNIAFMESVHAKSYSQIFS-TLCSTAE
1mxr_ LPPELETWVETWAFSETIHSRSYTHIIRNIVNDPSV
2za7_ VNVLNLRSVVSNVSSVLSNVVSVLSVVVCVNNRHDRDD
1bg7_ VNVLNLSSVVSNVSSVVSNVVSVLSVVVCVNNRHDHDH
1r03_ VNVLNLSNVVSNVSSVVSNVVSVLSVVVCVNNRHDHDH
1z6o_ DPVNVQNLSNVVSNVSSVLSNVSSVVSQVVSVNNRHHRDS
1z6o_ DPVNVLNLSNVVSNVSSVVSVVVSVVSVVVCVVSVPDPDP
1eum_ VVVLPLSSLLSNVVSVVSVVVSVLSCVVCVVVVHHRDN
1krq_ VVVLNLSNLLSNVVSVVSNVVSVLSVVVCVVVVHDRDH
3e6s_ VVVLNQSNLVSNVSSVVSNVVSVLSVVLCVVVVHPHDH
2jd7_ VVVLNLSSLLSNLRSQVSPLVSVLSVVVCVVVVHDRDH
1vlg_ VVVLNLSSLLSNVSSVVSNVVSVLSQVVCVVNVHHHDH
1otk_ PPDPDPVVNVVSNVLSVLSNVLSQLSQQVSCVVVVHDHS
1oqu_ VVDPDPSSNVSSPSSSVVSVRSSVNSVVNCV-NHHDPVS
1r2f_ VVDPDPSSNVSSPSSSVVSVRSSVNSVVNCV-NHHDPVS
1uzr_ VVDPDPSSNVSSNSSSVVSVRSSVNSVVNCV-NHHDPVS
1mxr_ NSNSSNVSSNSRSVVSVRLNVNSVVNQVNRDPDSVV




60
70
80
90
type 3
type 1
2za7_A/
1bg7_A/
1r03_A/
1z6o_A/
1z6o_M/
1eum_A
1krq_A/
3e6s_A/
2jd7_A/
1vlg_A/
1otk_A/
1oqu_A/
1r2f_A/
1uzr_A/
1mxr_A/



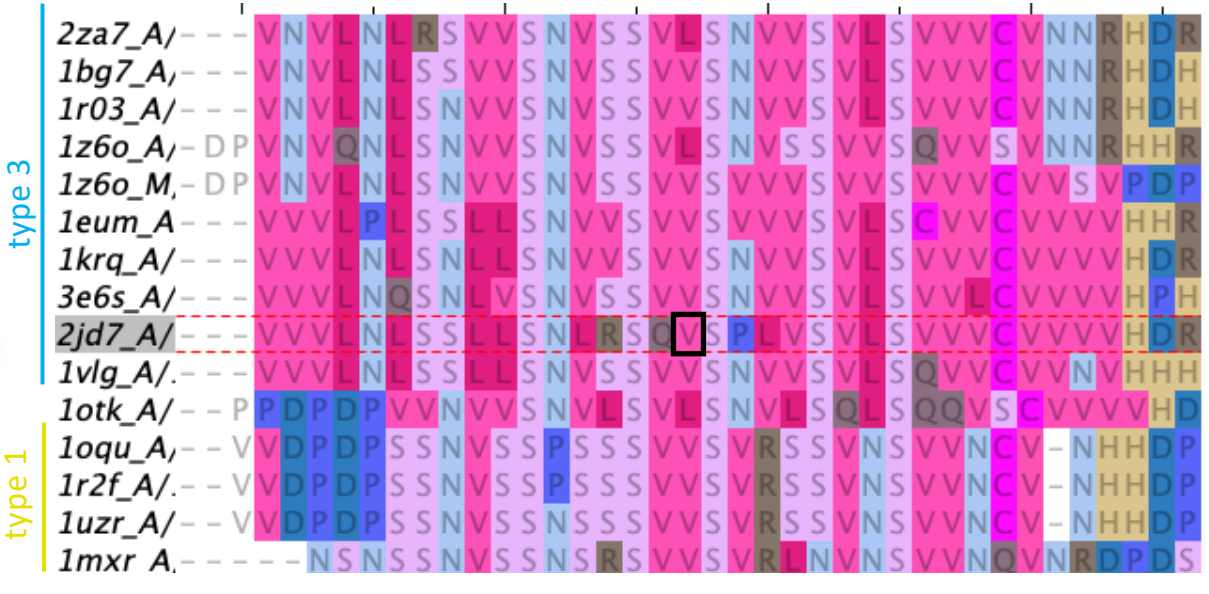
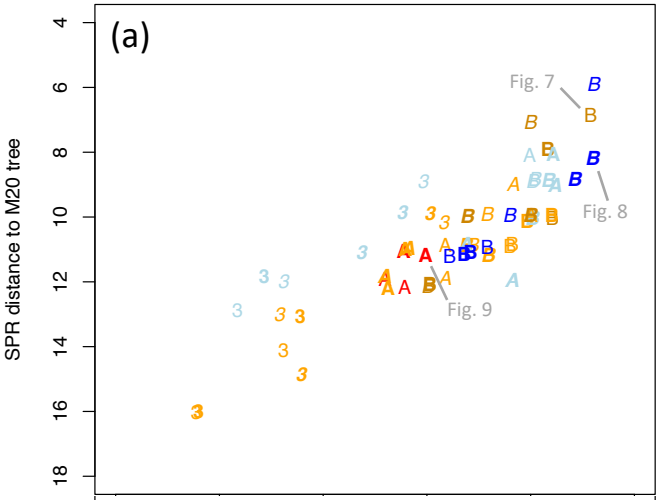

type 3
type 1
2za7_A/
1bg7_A/
1r03_A/
1z6o_A/
1z6o_M/
1eum_A
1krq_A/
3e6s_A/
2jd7_A/
1vlg_A/
1otk_A/
1oqu_A/
1r2f_A/
1uzr_A/
1mxr_A/

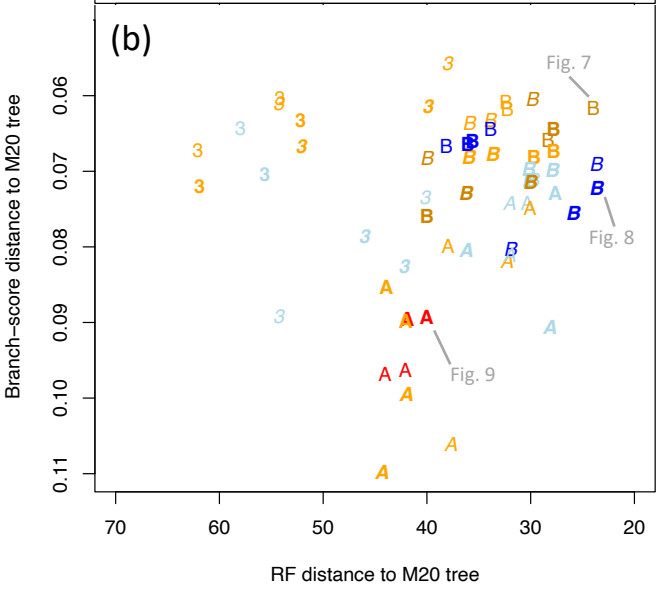
Table 1. Best-fit substitution models for IQtree runs, according to log-likelihood (lnL) and BIC. IQtree inferences were conducted using the BIC-best model. Runs are sorted by deep & shallow similarity to the M20 tree (first line). Results are shown only for IQtree runs using the custom model list (with partitioning for AA+3Di dataset) which included the 3Di substitution matrix. *k* represents the number of free parameters in the substitution model; all runs also included 103 free branch-length parameters in the BIC calculation. Note: lnL & BIC values can only be compared within lines, not between them, as they are only comparable for identical alignments.

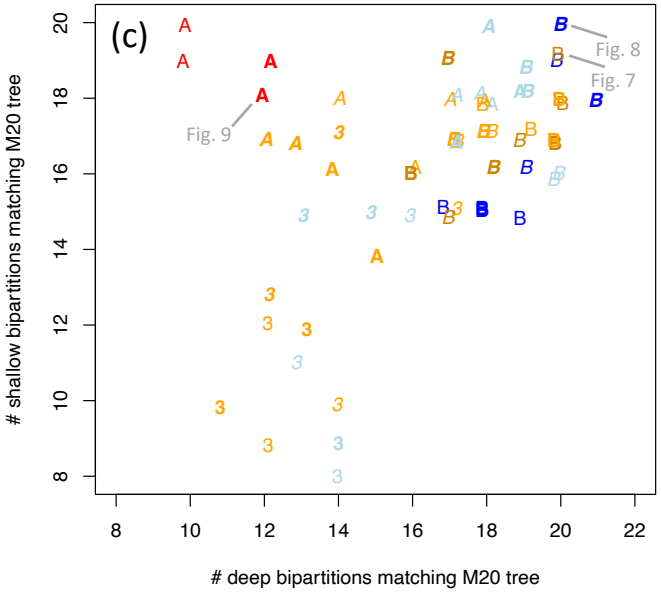
Structure source	Alignment data	Alignment method	Phylogeny data	# bipartitions matched			Best log-likelihood models:			Best BIC models:			
				all	deep	shal-low	lnL	*k*	Model	BIC	lnL	*k*	Model
PDB	structures	Superpose	*Q* score	51	22	29							
alphafold	struct.+AAs	USalign	AA+3di, trim	40	20	20	-24477.7	27	EX3+FU+G; 3DI+F+R3	49728.3	-24477.7	27	EX3+FU+G; 3DI+F+R3
PDB	3di	famsa3di	AA+3di	39	20	19	-39984.6	48	Blosum62+F+G4; 3DI+F+R5	81081.6	-39984.6	48	Blosum62+F+G4; 3DI+F+R5
alphafold	struct.+AAs	USalign	AA+3di	39	20	19	-43208.6	47	Blosum62+F+R3; 3DI+F+R3	87609.2	-43208.6	47	Blosum62+F+R3; 3DI+F+R3
PDB	3di	famsa3di	AA+3di, trim	37	20	17	-29156.4	41	Blosum62+F+G4; 3DI+F+G4	59179.9	-29156.4	41	Blosum62+F+G4; 3DI+F+G4
PDB	struct.+AAs	USalign	AA	37	19	18	-15160.7	25	Blosum62+F+R4	30976.0	-15208.7	4	Blosum62+R3
alphafold	3di	famsa3di	AA+3di	36	19	17	-42057.8	44	Blosum62+F+G4; 3DI+F+R3	85195.4	-42057.8	44	Blosum62+F+G4; 3DI+F+R3
PDB	struct.+AAs	USalign	AA+3di	34	19	15	-40816.9	47	Blosum62+F+R3; 3DI+F+R3	82772.5	-40816.9	47	Blosum62+F+R3; 3DI+F+R3
alphafold	struct.+AAs	USalign	AA, trim	38	18	20	-16270.9	25	Blosum62+F+R4	33187.1	-16315.2	3	EX3+FU+G
PDB	3di	famsa3di	AA	36	18	18	-24783.6	25	Blosum62+F+R4	50404.9	-24792.0	20	Blosum62+F+G4
alphafold	struct.+AAs	USalign	AA	36	18	18	-27095.9	25	Blosum62+F+R4	55106.9	-27107.4	20	Blosum62+F+G4
PDB	struct.+AAs	USalign	AA	36	18	18	-24877.6	25	Blosum62+F+R4	50631.3	-24881.1	23	Blosum62+F+R3
PDB	struct.+AAs	USalign	AA+3di, trim	33	18	15	-23679.1	25	Blosum62+R3; 3DI+F+G4	48115.0	-23679.1	25	Blosum62+R3; 3DI+F+G4
alphafold	3di	famsa3di	AA+3di, trim	36	17	19	-30459.0	27	EX3+FU+G; 3DI+F+R3	61715.6	-30459.0	27	EX3+FU+G; 3DI+F+R3
alphafold	3di	famsa3di	3di	32	17	15	-14950.2	25	3DI+F+R4	30745.2	-14953.6	23	3DI+F+R3
alphafold	struct.+AAs	USalign	3di	31	16	15	-15985.7	25	3DI+F+R4	32886.3	-15986.2	23	3DI+F+R3
alphafold	3di	famsa3di	AA	32	14	18	-26932.0	27	Blosum62+F+R5	54691.6	-26993.3	3	EX3+FU+G
alphafold	3di	famsa3di	3di, trim	31	14	17	-10259.3	27	3DI+F+R5	21218.1	-10260.7	25	3DI+F+R4
PDB	3di	famsa3di	AA, trim	30	14	16	-18633.2	25	Blosum62+F+R4	37938.1	-18641.4	20	Blosum62+F+G4
PDB	struct.+AAs	USalign	3di, trim	23	14	9	-8173.4	25	3DI+F+R4	17006.9	-8174.6	23	3DI+F+R3
PDB	struct.+AAs	USalign	3di	22	14	8	-15612.5	23	3DI+F+R3	32094.1	-15612.5	23	3DI+F+R3
alphafold	struct.+AAs	USalign	3di, trim	28	13	15	-8037.0	25	3DI+F+R4	16740.9	-8039.5	23	3DI+F+R3
PDB	3di	famsa3di	3di, trim	25	13	12	-10215.4	25	3DI+F+R4	21107.5	-10218.1	23	3DI+F+R3
none	AAs	famsa	AA, trim	30	12	18	-24136.9	23	LG+F+R3	49084.3	-24139.0	20	LG+F+G4
alphafold	3di	famsa3di	AA, trim	29	12	17	-20019.9	25	LG+F+R4	40665.8	-20044.4	3	EX3+FU+G
PDB	3di	famsa3di	3di	24	12	12	-14851.9	25	3DI+F+R4	30553.0	-14856.1	23	3DI+F+R3
none	AAs	famsa	AA	29	10	19	-18766.9	25	LG+F+R4	38190.2	-18813.5	1	LG+G4

Table 2. Illustrative models compared on AA and 3Di datasets. Alignment source: USalign on alphafold structures+AAs, trimmed sites with <35% data.

Data	Model	lnL	ΔlnL	*k*	BIC
amino acids (*n* =191 sites)	EX3+FU+G	-16315.2	35.9	106	33187.1
	Blosum62+F+G4	-16279.2	0.0	123	33204.5
	LG+F+G4	-16289.1	9.8	123	33224.2
	EX3+F+G	-16299.1	19.8	125	33254.6
	Poisson+F+G4	-16888.9	609.7	123	34423.9
	3DI+F+G4	-17950.8	1671.5	123	36547.6
3Di characters (*n* =191 sites)	3DI+F+G4	-8054.9	0.0	123	16755.9
	Poisson+F+G4	-8329.4	274.5	123	17304.9
	Blosum62+F+G4	-8485.0	430.0	123	17616.0
	EX3+F+G	-8620.9	566.0	125	17898.3
	LG+F+G4	-8734.1	679.1	123	18114.1
	EX3+FU+G	-10415.9	2361.0	106	21388.6



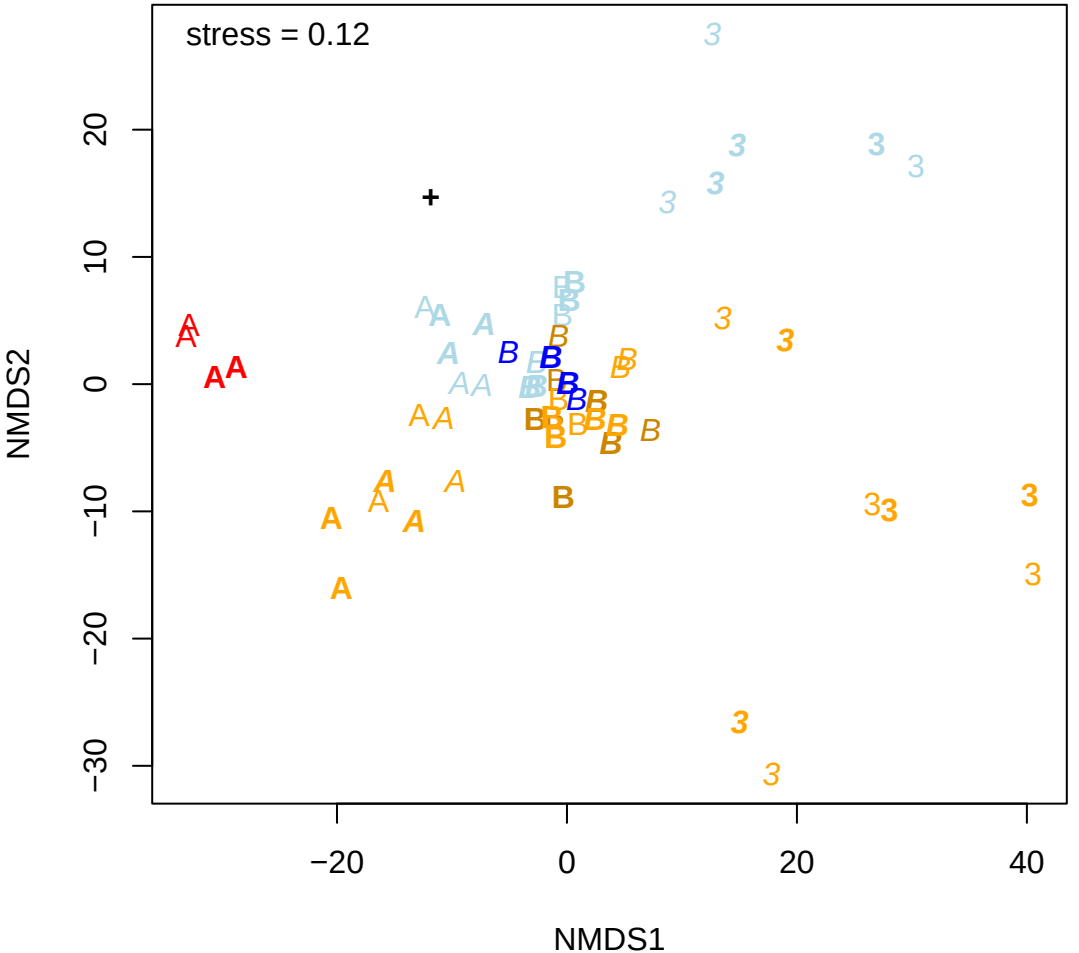


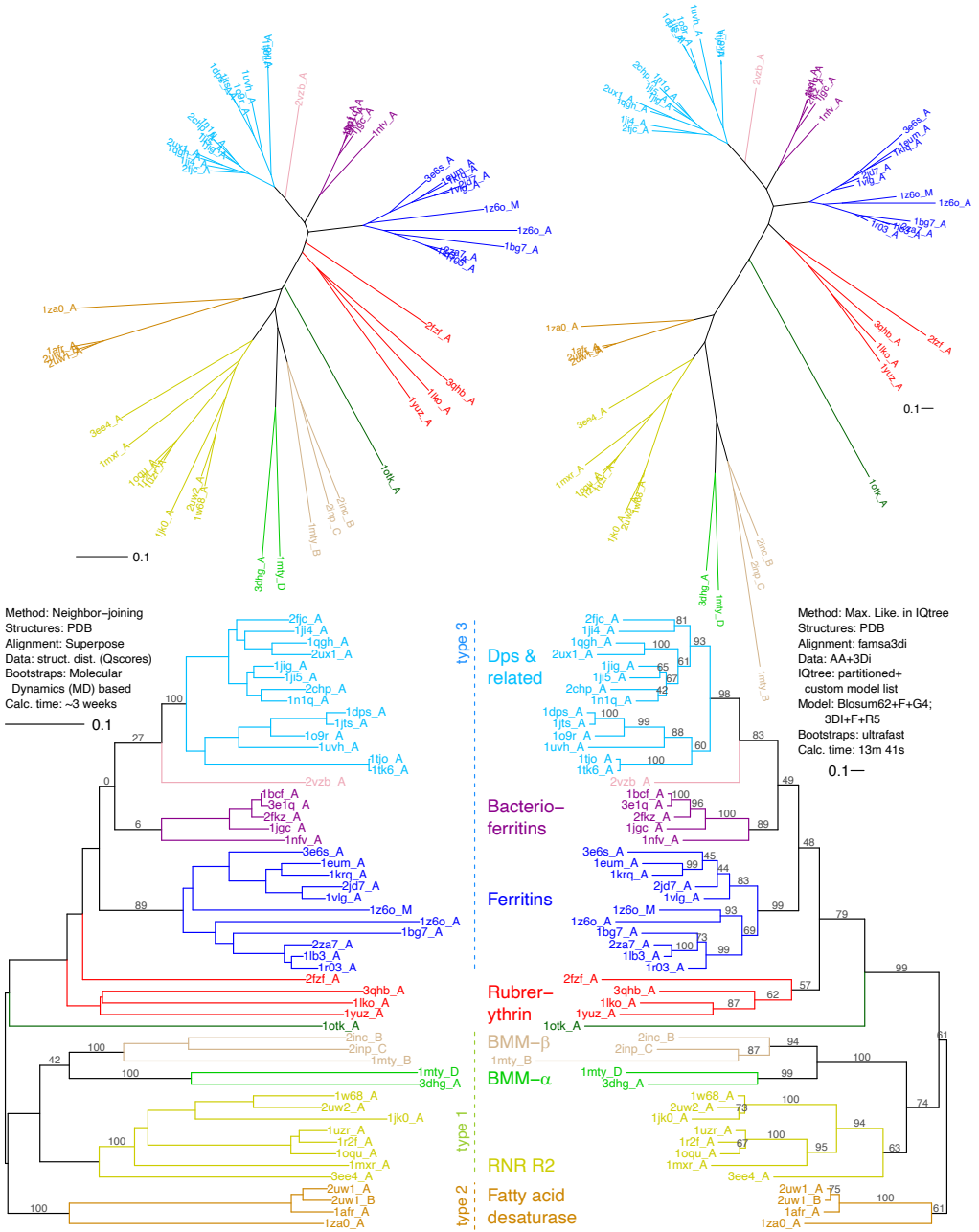


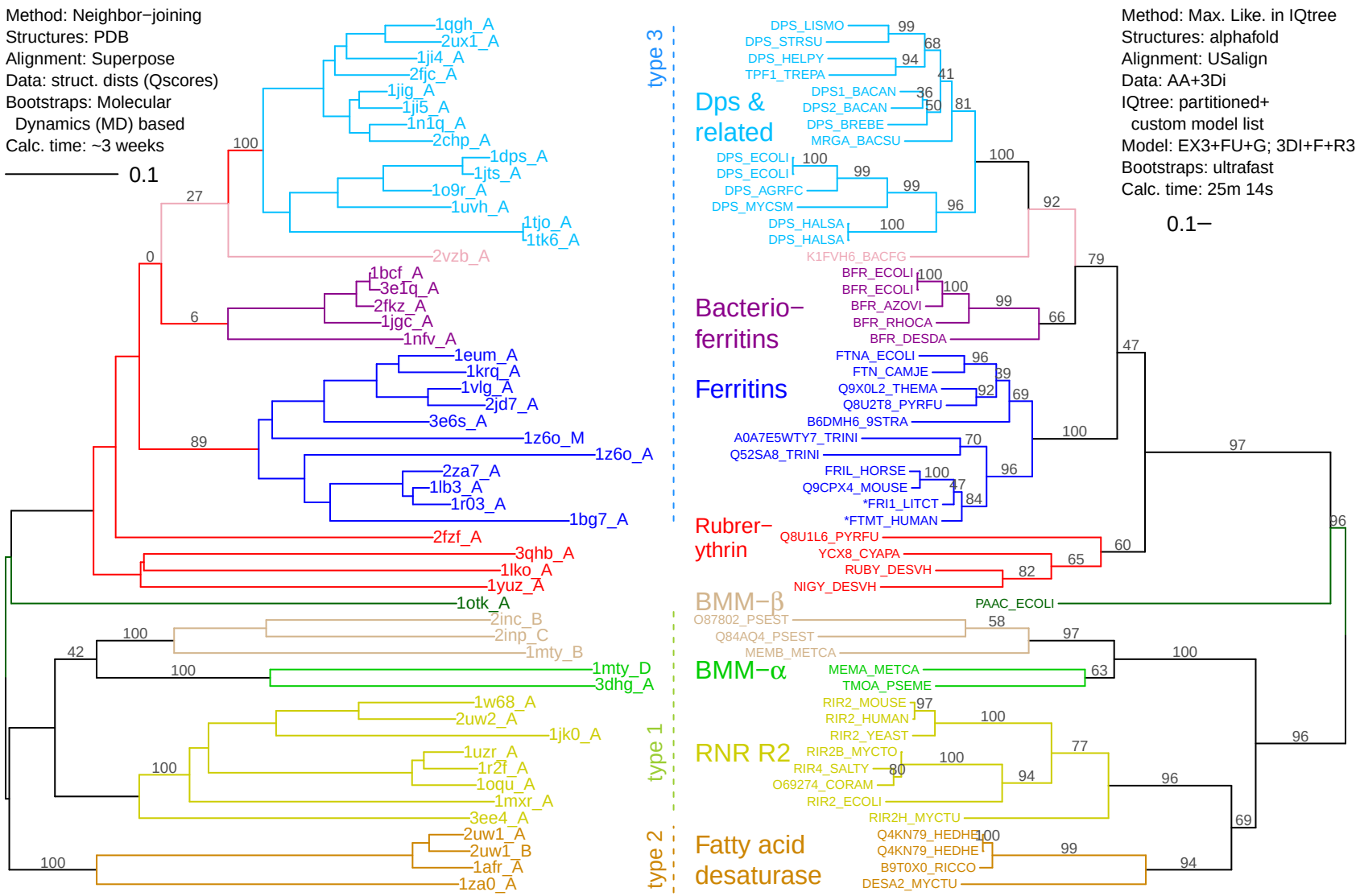
Symbols
3: 3Di characters-only
A: Amino-acids only
B: Both AA+3Di

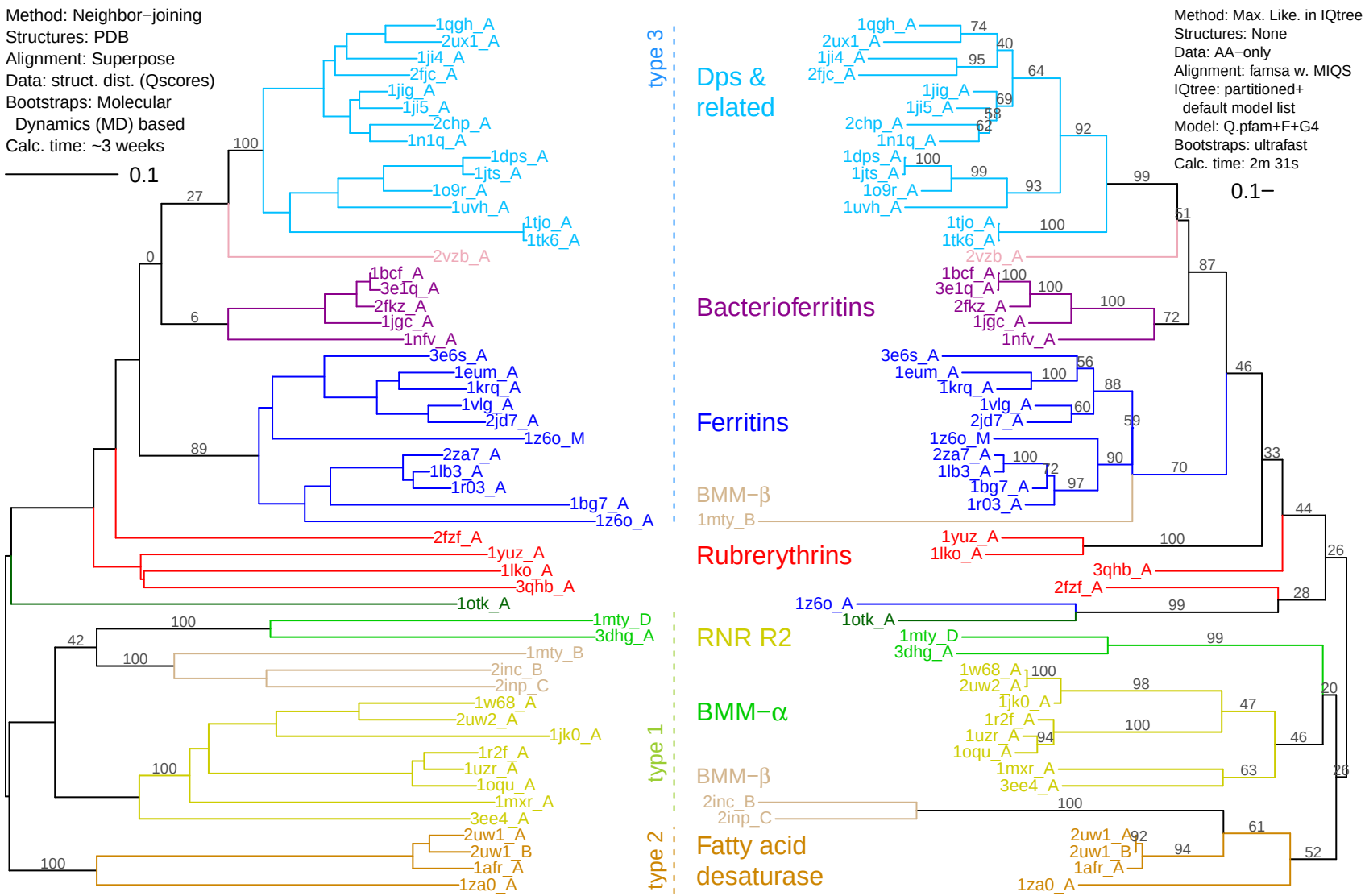
Font
Italic: Alphafold2 structures
Bold: Sites with <35% data trimmed out of alignment

Colors
Blue: USalign+partitioned
Light blue: USalign+not partitioned
Orange: Famsa3di-aligned+partitioned
Light orange: Famsa3di-aligned+not partitioned
Red: Famsa alignment of amino acids, no structural information used (traditional AA-only analysis)


stress = 0.12
NMDS2
NMDS1
20
10
0
−10
−20
−30
−20
0
20
40


Method: Neighbor−joining
Structures: PDB
Alignment: Superpose
Data: struct. dist. (Qscores)
Bootstraps: Molecular Dynamics (MD) based
Calc. time: ~3 weeks
0.1
Method: Max. Like. in IQtree
Structures: PDB
Alignment: famsa3di
Data: AA+3Di
IQtree: partitioned+ custom model list
Model: Blosum62+F+G4; 3DI+F+R5
Bootstraps: ultrafast
Calc. time: 13m 41s
type 3
type 1
type 2
Dps & related
Bacterio−ferritins
Ferritins
Rubrer−ythrin
BMM−β
BMM−α
RNR R2
Fatty acid desaturase


Method: Neighbor−joining
Structures: PDB
Alignment: Superpose
Data: struct. dists (Qscores)
Bootstraps: Molecular Dynamics (MD) based
Calc. time: ~3 weeks
0.1
Method: Max. Like. in IQtree
Structures: alphafold
Alignment: USalign
Data: AA+3Di
IQtree: partitioned+ custom model list
Model: EX3+FU+G; 3DI+F+R3
Bootstraps: ultrafast
Calc. time: 25m 14s
0.1
type 3
type 1
type 2
Dps & related
Bacterio−ferritins
Ferritins
Rubrer−ythrin
BMM−β
BMM−α
RNR R2
Fatty acid desaturase
1qgh_A
2ux1_A
1ji4_A
2fjc_A
1jig_A
1ji5_A
1n1q_A
2chp_A
1dps_A
1jts_A
1o9r_A
1uvh_A
1tjo_A
1tk6_A
2vzb_A
1bcf_A
3e1q_A
2fkz_A
1jgc_A
1nfv_A
1eum_A
1krq_A
1vlg_A
2jd7_A
3e6s_A
1z6o_M
1z6o_A
2za7_A
1lb3_A
1r03_A
1bg7_A
2fzf_A
3qhb_A
1lko_A
1yuz_A
1otk_A
2inc_B
2inp_C
1mty_B
1mty_D
3dhg_A
1w68_A
2uw2_A
1jk0_A
1uzr_A
1r2f_A
1oqu_A
1mxr_A
3ee4_A
2uw1_A
2uw1_B
1afr_A
1za0_A
DPS_LISMO
DPS_STRSU
DPS_HELPY
TPF1_TREPA
DPS1_BACAN
DPS2_BACAN
DPS_BREBE
MRGA_BACSU
DPS_ECOLI
DPS_ECOLI
DPS_AGRFC
DPS_MYCSM
DPS_HALSA
DPS_HALSA
K1FVH6_BACFG
BFR_ECOLI
BFR_ECOLI
BFR_AZOVI
BFR_RHOCA
BFR_DESDA
FTNA_ECOLI
FTN_CAMJE
Q9X0L2_THEMA
Q8U2T8_PYRFU
B6DMH6_9STRA
A0A7E5WTY7_TRINI
Q52SA8_TRINI
FRIL_HORSE
Q9CPX4_MOUSE
*FRI1_LITCT
*FTMT_HUMAN
Q8U1L6_PYRFU
YCX8_CYAPA
RUBY_DESVH
NIGY_DESVH
PAAC_ECOLI
O87802_PSEST
Q84AQ4_PSEST
MEMB_METCA
MEMA_METCA
TMOA_PSEME
RIR2_MOUSE
RIR2_HUMAN
RIR2_YEAST
RIR2B_MYCTO
RIR4_SALTY
O69274_CORAM
RIR2_ECOLI
RIR2H_MYCTU
Q4KN79_HEDHE
Q4KN79_HEDHE
B9T0X0_RICCO
DESA2_MYCTU


Method: Neighbor−joining
Structures: PDB
Alignment: Superpose
Data: struct. dist. (Qscores)
Bootstraps: Molecular Dynamics (MD) based
Calc. time: ~3 weeks
0.1
Method: Max. Like. in IQtree
Structures: None
Data: AA−only
Alignment: famsa w. MIQS
IQtree: partitioned+ default model list
Model: Q.pfam+F+G4
Bootstraps: ultrafast
Calc. time: 2m 31s
0.1
type 3
type 1
type 2
Dps & related
Bacterioferritins
Ferritins
BMM−β
Rubrerythrins
RNR R2
BMM−α
BMM−β
Fatty acid desaturase
1qgh_A
2ux1_A
1ji4_A
2fjc_A
1jig_A
1ji5_A
2chp_A
1n1q_A
1dps_A
1jts_A
1o9r_A
1uvh_A
1tjo_A
1tk6_A
2vzb_A
1bcf_A
3e1q_A
2fkz_A
1jgc_A
1nfv_A
3e6s_A
1eum_A
1krq_A
1vlg_A
2jd7_A
1z6o_M
2za7_A
1lb3_A
1r03_A
1bg7_A
1z6o_A
2fzf_A
1yuz_A
1lko_A
3qhb_A
1otk_A
1mty_D
3dhg_A
1mty_B
2inc_B
2inp_C
1w68_A
2uw2_A
1jk0_A
1r2f_A
1uzr_A
1oqu_A
1mxr_A
3ee4_A
2uw1_A
2uw1_B
1afr_A
1za0_A