Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

The structure of evolutionary model space for proteins across the tree of life

Gabrielle E. Scolaro [1] and Edward L. Braun [1,*]

[1] Department of Biology, University of Florida, Gainesville, FL 32611, USA; gscolar1@jh.edu (G.E.S.); ebraun68@ufl.edu (E.L.B)

* Correspondence: ebraun68@ufl.edu

Simple Summary: Relative rates of amino acid substitution over evolutionary time reflect the chemical properties of the amino acids. Substitutions that result in an amino acid similar to the ancestral residue accumulate more rapidly than those resulting in a dissimilar amino acid. The substitution rates for each amino acid pair are the parameters in models of evolutionary change for proteins. Although the best-fitting model of protein evolution is known to differ among taxa, a comprehensive picture of model changes across the tree of life is not available. In principle, models of protein change might reflect evolutionary history (i.e., closely related taxa have similar models) or the environment (i.e., taxa living in similar environments have similar models). We estimated models of amino acid evolution for organisms across the tree of life, finding evidence that history and the environment have both contributed to model differences. Bacterial models differed from archaeal and eukaryotic models. Models for Halobacteriaceae (Archaea that live in highly saline environments) and Thermoprotei (a group of thermophilic Archaea) were very distinctive. Rates of substitution for pairs of aromatic amino acids were especially variable. Overall, these results paint a picture of the "evolutionary model space" for proteins across the tree of life.

Abstract: The factors that determine the relative rates of amino acid substitution during protein evolution are complex and they are known to vary among taxa. We estimated relative exchangeabilities for pairs of amino acids from clades spread across the tree of life and assessed the historical signal in the distances among these clade-specific models. We trained these models separately on collections of arbitrarily selected protein alignments and on ribosomal protein alignments. In both cases we found a clear separation between the models trained using multiple sequence alignments from bacterial clades and the models trained on archaeal and eukaryotic data. We assessed the predictive power of our novel clade-specific models of sequence evolution by asking whether fit to the models could be used to identify the source of multiple sequence alignments. Model fit was generally able to classify protein alignments correctly at the level of domain (bacterial versus archaeal), but the accuracy of classification at finer scales was much lower. The only exceptions to this were the relatively high classification accuracy for two archaeal lineages: Halobacteriaceae and Thermoprotei. Genomic GC content had a modest impact on relative exchangeabilities despite having a large impact on amino acid frequencies. Relative exchangeabilities involving aromatic residues exhibited the largest differences among models. There were a small number of exchangeabilities that exhibited large differences in comparisons among major clades and between generalized models and ribosomal protein models. Taken as a whole, these results reveal that a small number of relative exchangeabilities are responsible for much of the structure of the "model space" for protein sequence evolution. If we look beyond the information that these clade-specific models reveal about protein evolution the models themselves are likely to be useful tools for phylogenomic inference across the tree of life.

Keywords: Molecular evolution; Substitution matrix; Amino acid exchangeability; Models of sequence evolution; Protein evolution; Archaea; Bacteria

© 2022 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

1. Introduction

The fact that rates at which different pairs of amino acids experience substitutions over evolutionary time vary by orders of magnitude has been appreciated for more than five decades [1,2]. Indeed, Kimura and Ohta [3] included the fundamental pattern observed in those pioneering studies, in which the rate conservative substitutions (exchanges that involve pairs of chemically-similar amino acids) is higher than the rate of more radical substitutions, as one of the five principles governing molecular evolution. The processes that determine the rates of substitution for the various pairs of amino acids can be divided into two fundamental categories: 1) the rate and spectrum of non-synonymous mutations; and 2) the probability that any of those novel non-synonymous changes will increase in frequency to the point that they will be observed as substitutions. The rapid accumulation of sequence data in the genomic and post-genomic eras [4,5] has served to confirm the fundamental patterns observed in those classic studies. However, the explosive growth of sequence databases has provided enough information to show differences among clades in their patterns of amino acid substitution [6–8]. Those more recent results indicate that the processes governing the relative rates of amino acid change have themselves changed over evolutionary time.

Understanding the ways that patterns of protein evolution have changed across the tree of life requires a mathematical framework. Pandey and Braun [7] suggested a simple approach using the 20-state GTR (general time reversible) model (i.e., the extension of the nucleotide GTR model [9,10] to the amino acid alphabet). GTR model parameters are typically written as a diagonal matrix ($\mathbf{\Pi}$) of equilibrium state frequencies and a symmetric matrix of relative exchangeability (RE) parameters [11], often called the $\boldsymbol{R}$ matrix. The biological meaning of the $\mathbf{\Pi}$ matrix, which describes the equilibrium amino acid frequencies, is straightforward. In contrast, the best interpretation of the RE parameters is more complex. Ultimately, the REs are numerical values that capture the rates at which new non-synonymous changes enter populations and the probability that those novel changes become fixed as substitutions. Thus, REs reflect processes at the molecular and cellular level (e.g., the mutational rate and spectrum, the structure of the genetic code, the impact of amino acid changes on protein function) and population-level processes (e.g., the relative impacts of selection and drift on polymorphisms). Despite their complexity, the relationship between REs and differences in physicochemical properties of amino acids is clear [12]. It is also clear that REs estimated using multiple sequence alignments (MSAs) for unrelated proteins from the same clade can be used to identify the clade of origin for protein MSAs that were not used to generate the RE estimates [7,8]. This suggests amino acid substitutions may be more or less conservative depending on the clade under consideration. In principle, it is straightforward to examine differences among taxa in REs (and amino acid frequencies): estimate GTR_{20} model parameters for a set of clades across the tree of life and compare the values of those parameter estimates. We believe that a large-scale effort to leverage the relative ease of obtaining maximum likelihood (ML) estimates of RE values across the tree of life will yield biological insights.

However, estimating REs using the GTR_{20} model presents two fundamental challenges. First, the GTR_{20} model assumes that both the REs and the equilibrium frequencies of amino acids remain constant over evolutionary time (i.e., it assumes time-reversibility). The evidence for changes in the patterns of amino acid substitution [6–8] indicates that the time-reversibility assumption cannot hold across the tree of life. Second, the GTR_{20} model is parameter rich; it has a total of 208 free parameters (19 equilibrium amino acid frequencies and 189 free REs). Individual proteins will not provide enough information to allow accurate estimation of GTR_{20} model parameters (MSAs for most individual proteins are 280–600 amino acids in length [13]). However, neither of these challenges are insurmountable. The time reversibility assumption is irrelevant as long as protein sequence evolution is *locally time reversible*. In other words, time-reversibility may hold (at least approximately) within specific clades despite the fact that it does not hold at the scale of the tree of life. This changes the nature of the challenge posed by the time-reversibility

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

assumption; the assumption is only problematic if violations of time-reversibility within clades are large enough to distort the results of model comparisons. Evidence that REs exhibit a historical signal (i.e., patterns in the data that are congruent with at least some parts of the tree of life), an ecological signal (i.e., strong differences between RE estimates for taxa living in distinct environments, like thermophiles versus mesophiles), or both signals would indicate the violations of local time-reversibility are not especially problematic. Obviously, it is possible for some historical and/or ecological signals to emerge and other signals to be distorted, so it is impossible to rule out the existence of any problems linked to violations of the time-reversibility assumption. However, the existence of either (or both) signals in the RE data would provide evidence that our estimates of those parameters suffice for the goals of this study.

The simplest way to overcome the challenge presented by the high dimension of the GTR_{20} model is to use a large number of proteins to estimate parameters [11,14]. This approach has been used extensively in phylogenetics; in fact, one of the earliest studies that employed ML to estimate a phylogeny using a protein MSA [15] used REs based on the Dayhoff et al. [16] PAM matrix. The PAM matrix was generated using an approximate method to estimate REs from 71 groups of proteins with diverse structures and functions. Subsequent studies used matrices estimated using larger numbers of proteins (e.g., the JTT [17], VT [18], WAG [11], and LG [14] models), the $\boldsymbol{R}$ matrices for those "generalized models" reflect patterns of sequence evolution averaged across proteins and across the tree of life. Thus, despite their utility in protein phylogenetics they cannot provide information about the differences among clades in their REs. Other studies have eschewed this broad sampling of proteins and taxa, focusing on specific taxa and/or proteins, like the proteins encoded by specific viruses [19–22] or proteins encoded by organelle genomes [23–26]. Concerted efforts to estimate REs for diverse proteins from specific groups of free-living taxa have only been undertaken recently [6–8]. The approach we have chosen for this study is to extend those efforts to other taxa spread across the tree of life.

Unlike the RE parameters, proteome-wide variation in amino acid frequencies has received substantial study. These studies have typically used observed proportions of each amino acid instead of ML estimates of the $\boldsymbol{\Pi}$ matrix parameters. The use of observed amino acid proportions makes it possible to examine proteome-wide amino acid compositions for individual taxa (estimating $\boldsymbol{\Pi}$ matrix parameters by ML requires MSAs) and it is less computational burdensome than the full ML approach. This approach has revealed many correlates of proteome-wide amino acid composition, such as genomic GC-content and extremophilic lifestyles [27–31]. As for the differences among proteins, one major axis of variation in protein composition appears to differentiate ribosomal proteins from other proteins [32]. For this reason, we have estimated models for ribosomal proteins and the broader proteome separately and compared those models to determine whether the REs differed between these protein types.

To better understand the structure of "protein model space" we used the GTR_{20} model to estimate REs from multiple lineages of bacteria, archaea, and eukaryotes. We trained clade-specific models for large numbers of proteins using arbitrarily selected MSAs of homologous proteins from each taxonomic group and we also estimated model parameters for ribosomal proteins by selecting subsets of taxa from the Hug et al. [33] MSA. These clade-specific models allowed us to ask several questions. First, what types of signal are present in clade-specific model REs? In principle, there could be a phylogenetic signal, an ecological signal, or some mixture of both types of signals. We addressed this question by examining the structure of a clustering diagram generated using distance among models, comparing that "tree of models" with estimates of the tree of life and looking for any informative clusters. Second, can this set of clade-specific models be used as a classifier for MSAs that were not present in the training data? We used the method described by Pandey and Braun [7] to answer this question. Third, which parameters of the clade-specific models exhibited differences? We examined differences among models at the level of the REs (the $\boldsymbol{R}$ matrices) and the amino acid frequencies (the $\boldsymbol{\Pi}$ matrices). We answered these three major questions both for generalized clade-specific models and for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

clade-specific ribosomal protein models. This led to a third question: do the REs and equilibrium amino acid frequencies for generalized models and ribosomal protein models differ? Finally, we discuss the implications of the RE estimates. We believe that the results of these analyses provide a picture of the structure of protein model space for the tree of life.

2. Materials and Methods

2.1. Estimating clade-specific models

We selected 19 clades spread across the tree of life, using the estimate of the tree of life from Hug et al. [33] (hereafter called the Hug tree) as a guide for taxon selection. Then we trained the generalized clade-specific models using homologous proteins from representative members of those clades. To do this we identified whole genome assemblies for selected taxa with annotated genome assemblies in 19 different clades (five archaea, two eukaryotes, and 12 bacteria). We collected files of annotated protein sequences for 11 to 37 taxa in each clade (see Supplementary File S1 for the complete taxon list), giving preference to taxa in the Hug tree. Then we clustered and aligned proteins in each clade using usearch [34] with a 50% similarity cutoff to cluster, retaining clusters with at least four sequences. The MSAs generated using usearch were then arbitrarily assigned to a training set and a validation set. The training sets were used to estimate clade-specific model parameters using the ReplacementMatrix server [35]. Whenever possible training sets comprised 1000 MSAs (we obtained fewer clusters from Thermoprotei so that training set was limited to 640 MSAs). We refer to the REs (***R*** matrix parameters) and equilibrium amino acid frequencies (the π_X values for each amino acid, since they are the diagonal elements of the $\mathbf{\Pi}$ matrix) estimated by the ReplacementMatrix server using the training data for each clade as the generalized model for that clade.

We estimated parameters for clade-specific ribosomal protein models using MSAs extracted from the Hug et al. [33] data matrix, which is a concatenated matrix of ribosomal proteins. For this analysis we expanded the number of focal clades to 43 MSAs and used alignment from Hug et al. [33] as published (with the exception of eliminating all gap/missing columns). We estimated GTR_{20} rate matrices for each clade by using IQ-TREE v.1.6.12 ([36]), modeling among-sites rate heterogeneity using invariant sites and Γ-distributed rates (i.e., the GTR+I+G4 model).

2.2. Analyzing model space and assessing the performance of model fit as a classifier

To examine the structure of model space we calculated Euclidean distances among the exchangeability matrices (190 values normalized to sum one) from each model using perl scripts (available from https://github.com/ebraun68/protmodels; accessed 31 January 2022). We visualized model space by clustering the Euclidean distances among those models by neighbor-joining [37] in PAUP* [38]. We assessed the congruence between the model clustering trees and the Hug tree using matching distances [39,40] calculated in PAUP*. We tested the hypothesis that the structure of model space resembles the tree of life (i.e., the hypothesis that model parameters have phylogenetic signal) by comparing the matching distance between the Hug tree and the model clustering trees to the matching distances between the Hug tree and random trees (generated by PAUP* using the assumption that trees are equiprobable). This approach is similar to the approach that Penny et al. [41] used to show that the theory of evolution is falsifiable.

We evaluated the ability of model fit to classify validation set MSAs. We identified the best-fitting model by fitting all models to a single tree for each validation set MSA. The tree for each MSA was obtained by conducting a tree search using the Q.pfam model [8], combined with empirical amino acid frequencies and Γ-distributed rates across sites (i.e., the Q.pfam+F+G4 model). Then we obtained BIC [42] scores for each model given that tree, using the ***R*** matrix+F+G4 model in all cases (we parsed the BIC value because it the default criterion for model selection in IQ-TREE, but we note that all models that we examined have the same dimension so the model fit rankings would be identical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

regardless of whether we used BIC, AIC, AIC_c, or likelihood because). The script for this analysis (fitprotFGmodels.sh) is available from github (https://github.com/ebraun68/protmodels, accessed 20 December 2022). We examined the performance of models as classifiers using recall; because we conducted separate analyses of each validation set recall is simply the number of true positives (cases where the MSA had the best fit to the appropriate clade-specific model) divided by the number of MSAs in each validation set. We scored true positives at two taxonomic levels. *Precise true positives* were cases where a validation set MSA from a specific bacterial clade had the model trained on MSAs from that specific clade. In contrast, *domain-level true positives* were cases where a bacterial MSA was called a true positive if the best-fitting model was any of the models trained using bacterial MSAs. We repeated the model fit analyses on the training set MSAs to serve as a comparison to the model fit analysis for the validation set.

We examined the fit of two model sets, the novel clade-specific models generated for this study (both the general models and the ribosomal protein models) and a "complete" model set. The complete model set comprises: 1) the novel clade-specific models; 2) generalized standard models (e.g., the PAM al. [16], JTT [17], VT [18], WAG [11], LG [14], and Q.pfam [8] models; and 3) published clade-specific models [7,8,44] (the complete list of models is available from https://github.com/ebraun68/protmodels as part of the script used for these analyses). All novel clade-specific models are available from Zenodo [43] and from https://github.com/ebraun68/protmodels (accessed 20 December 2022) as PAML format rate matrices. The training data used to generate those models and the validation data are available as from Zenodo [43].

3. Results and Discussion

3.1. The structure of protein model space resembles the tree of life

The trees of models (clustering diagrams based on Euclidean distances among models; Figure 1a,b) showed clear similarities to the tree of life (Figure 1c). The topological distances between the Hug tree and the clustering diagrams based on generalized models and ribosomal protein models were smaller than expected for distances between random trees of similar sizes (Figure 1d,e). The most important similarity between the clustering diagrams for models and the tree of life is the fact that most or all bacterial models were separated from the models from archaeal and eukaryotic models. The existence of a bipartition separating Bacteria from Archaea and eukaryotes is long-standing [45] and, at this point, uncontroversial (reviewed by Eme et al. [46]). The only exception to this division was observed in the clustering diagram for general models: Methanomicrobia (Archaea) and Selanomonadales (Bacteria) formed a cluster separate from either group (Figure 1a). In contrast, the bipartition separating Bacteria from Archaea+eukaryotes was perfect when we clustered ribosomal protein models (Figure 1b), although the position of the Methanomicrobia ribosomal protein model differed from the current best estimate of archaeal phylogeny (Figure 1c), which places the DPANN group sister to the other Archaea (see also Castelle et al. [47] and Williams et al. [48]). The tree based on clustering ribosomal protein models exhibited additional similarities to the tree of life (these similarities involve the taxa presented using bold text in Figure 1b and 1c). Specifically, the clustering diagram for clade-specific ribosomal protein models and the Hug tree both included: 1) the bipartition separating eukaryotes, Thaumarchaeota, and Thermoprotei from other taxa; 2) the subtree comprising Rhodospirillales, Rhodobacteriacae, and Rhizobiales; and 3) the subtree within the "candidate phyla radiation" (CPR [49,50]) comprising Levybacteria, Gottesmanbacteria, and Roizmannbacteria). These results suggest there is phylogenetic signal in protein models, especially the differences between Bacteria and Archaea+eukaryotes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

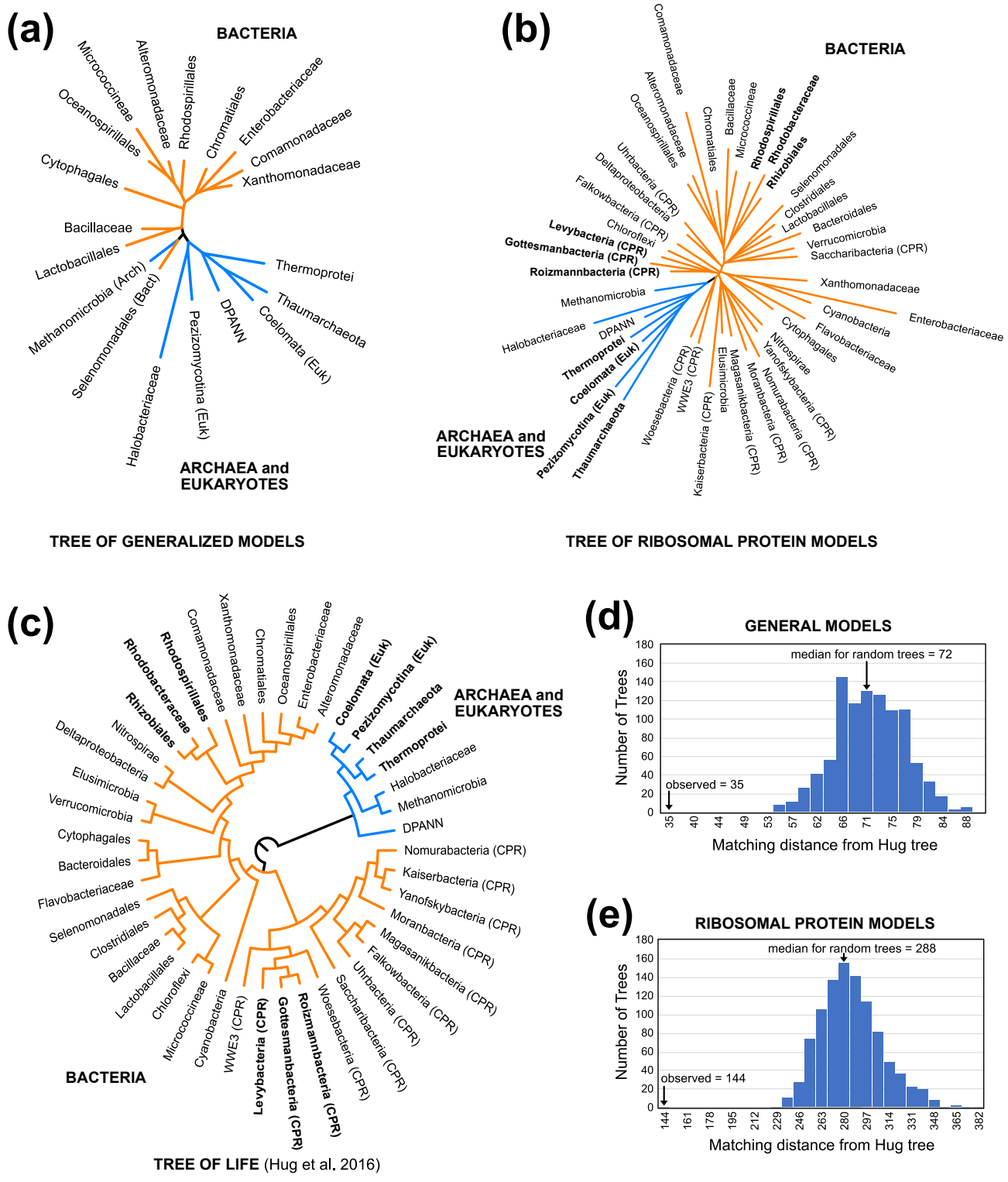

Figure 1. "Trees of models" and the tree of life exhibit topological similarities. Trees of models generated were by clustering Euclidean distances among estimated REs using neighbor-joining. Archaea and eukaryotes are shaded blue in all trees. (**a**) Tree of models for the general models. (**b**) Tree of models for ribosomal protein models. (**c**) Cladogram for the taxa used for this study based on the Hug tree, which is an ML analysis of concatenated ribosomal proteins [33]. "CPR" indicates members of the candidate phyla radiation [33,49,50]. (**d**) Matching distance between the Hug tree and the tree of general models compared to the distribution of distances between the Hug tree and random trees). (**e**) Matching distance between the Hug tree and the tree of ribosomal protein models compared to the distribution of distances between the Hug tree and random trees. The labels for each bin are the minimum matching distance for that bin.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

We have referred to the clustering diagrams for models of protein evolution (Figure 1a,b) as "trees of models" to emphasize that they are not estimates of phylogeny. Trees of models generated by clustering distances among models are only expected to exhibit congruence with phylogeny if three conditions are met. First, the estimates of model parameters must not be strongly biased due to violations of the GTR_{20} model; this can be viewed as a restatement of our local time-reversibility assumption. Second, patterns of protein evolution captured by the models must change over an appropriate timescale and exhibit limited convergence. If proteome-wide REs changed very rapidly (relative to the timescales for the evolutionary history of clades we chose) we are unlikely to recover any historical signals at the scale of the tree of life. Likewise, proteome-wide convergence in REs could have an impact on the topology of the tree of models; in fact, we expected strong ecological signals in the REs to yield a tree that includes bipartitions that divide the taxa into subsets with similar lifestyles (e.g., a thermophile versus non-thermophile bipartition). Finally, the estimates of distances among models (in this case, Euclidean distances among normalized vectors of RE values) must yield useful estimates of actual distances between models. These factors mean that it would have been appropriate to interpret failure to find congruence between trees of models and the Hug tree with caution; failure to find evidence of congruence could have meant the there was no historical signal or it could be evidence for analytical biases that distort trees of models. On the other, evidence of congruence is likely to be meaningful.

The topological congruence between the trees of models and the tree of life (Figure 1) provides evidence that there is historical signal in the RE values. Indeed, these results are consistent with a single large-scale model shift on the branch separating Bacteria from Archaea and eukaryotes combined with modest historical signal elsewhere in the tree. Of course, it remains possible that other methods to extract information about changes in the patterns of protein sequence evolution might reveal additional evidence for historical signal. Regardless, it is clear that there is some historical signal in the patterns of protein sequence evolution that can be extracted using our methodology.

3.2. Models of protein sequence evolution can sometimes be used as domain-level classifiers

A straightforward way to examine the ability of the bacterial models (sets of REs estimated using each training set) to predict patterns of sequence evolution is to examine recall (true positives divided by total number of MSAs from each source) when the models are used to classify validation set MSAs. Pandey and Braun [7] reported that recall when fit to clade-specific models is typically >70% when those models were used to classify MSAs comprising sequences from specific eukaryotic clades (birds, mammals, plants, yeasts, and oomycetes). In fact, Pandey and Braun [7] found that recall as >70% in all cases if one viewed the classification of birds and mammals as vertebrates as a true positive. The recall was >70% only for some of the models estimated as part of this work, and in those cases the >70% recall was only achieved when the domain-level true positives were considered (i.e., when classification of the MSAs as bacterial, archaeal, or eukaryotic was viewed as the endpoints; see bold values in the last three columns of Table 1; also see Supplementary File S2). Moreover, there were eight (out the 19) clades where recall was <70% even when we limited our consideration to domain-level classification, emphasizing the limited utility of model fit as a classifier for MSAs from these taxa. We observed virtually identical results when we classified training set individual MSAs (Supplementary File S2). Recall was much lower when we limited our consideration to precise true positives (i.e., when we attempted to classify MSAs at the level of the specific clades used for model training). The inability to use model fit as a precise classifier was further emphasized by the observation that Thermoprotei and Halobacteriaceae were the only clades with precise matches >50% (Table 1). The high precise match percentages probably reflect the distinctive nature of the selective pressures on those archaeal lineages (e.g., the high intracellular ion concentrations of Halobacteriaceae [51]). Thus, we interpret the distinctive nature of the models for those lineages to be evidence for an ecological signal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

Table 1. Recall for the novel models when they are used as classifiers[1].

Domain	Clade	Median GC %	%Precise Match	%Archaeal Match	%Eukaryotic Match	%Bacterial Match
Archaea	DPANN	34.05	5.07	41.11	9.76	49.12
	Thaumarchaeota	34.20	25.12	37.68	5.68	56.64
	Methanomicrobia	47.40	18.50	55.49	7.52	36.99
	Thermoprotei[2]	49.00	65.94	**76.56**	7.19	16.25
	Halobacteriaceae[2]	63.70	65.39	**75.20**	4.86	19.94
Eukaryotes	Coelomata	40.72	10.74	50.93	15.96	33.11
	Pezizomycotina	48.90	14.80	33.42	39.63	26.95
Bacteria	Lactobacillales	37.70	18.90	24.56	12.01	63.43
	Cytophagales	40.60	13.24	10.29	7.35	**82.35**
	Bacillaceae	42.15	11.84	36.93	11.15	51.92
	Alteromonadaceae	45.90	7.33	13.56	6.31	**80.13**
	Selenomonadales	48.90	14.68	28.67	9.35	61.82
	Enterobacteriaceae	52.18	18.28	18.52	6.98	**74.50**
	Oceanospirillales	54.40	1.52	17.81	7.47	**74.72**
	Chromatiales	62.15	21.06	21.62	6.50	**71.88**
	Rhodospirillales	65.03	29.77	14.91	7.37	**77.72**
	Comamonadaceae	65.38	18.29	9.84	6.71	**83.46**
	Xanthomonadacae	66.03	9.62	11.70	10.78	**77.51**
	Micrococcineae	69.65	36.26	15.36	5.32	**79.32**

[1] The % matches are the sums for matches to the relevant generalized models and the relevant ribosomal protein models. Values >70% are presented in bold. Full data are available in Supplementary File S2.
[2] The % precise matches for these taxa are underlined because they exceed 50%.

Since Thermoprotei and Halobacteriaceae had large percentages of precise matches we examined differences between the REs for each of those taxa and the average REs for other archaea. The specific REs with the largest differences appeared to be related to the equilibrium amino acid frequencies for each model, although the nature of this relationship was complex. The most extreme cases for Thermoprotei included R-K, S-T, and N-G exchanges (elevated in Thermoprotei) and Q-K, R-Q, and C-S (reduced in Thermoprotei; Supplementary File S3). Relative to the other archaea, the equilibrium frequency for arginine was elevated in Thermoprotei (π_R range for Archaea = 0.0342 - 0.0771; Thermoprotei π_R was the maximum, Supplementary File S3) whereas serine and threonine were both reduced in Thermoprotei (π_S range = 0.0568 - 0.0794 and π_T range = 0.0407 - 0.0651; Thermoprotei π_S and π_T were the minimum values; Supplementary file S3). It is possible that these estimated equilibrium frequencies for Thermoprotei are related to the generally elevated arginine frequency and reduced serine frequency in thermostable proteins [52], although we note that we examined other thermophilic lineages (Thaumarchaeota and DPANN also include thermophiles) . In Halobacteriaceae the elevated REs included I-V, H-Y, and Q-K and the reduced REs included P-Y, L-V, and I-Y (Supplementary File S3); the estimated equilibrium frequencies for isoleucine, histidine, lysine, and tyrosine for Halobacteriaceae were all the minimal values for the Archaea we examined (Supplementary File S3). A prominent feature of Halobacteriaceae proteins relative to other taxa is a much higher ratio of acidic to basic residues, which is known to improve protein solubility given high salt concentrations [53]. This difference was evident in our estimated equilibrium frequencies (the acidic to basic residue ratio based on equilibrium amino acid frequencies was 2.43 for Halobacteriaceae but it ranged from 0.792 to 1.06 for the other archaea we examined; Supplementary File S3). However, only one of the outlier REs (if we use top three and bottom three RE differences to define outlier) involved a basic or an acidic residue, suggesting that shifts in those amino acid frequencies did not have a major impact on the REs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

The bacterial clade with the highest precise match percentage was Micrococcineae, an actinobacterium (high-GC Gram-positive bacterium). However, the precise true positive percentage for Micrococcineae was only 36.3% (Table 1). Some validation sets have very low percentages of precise true positives, with the most extreme being 1.5% (for Oceanospirillales, a clade within the larger Gram-negative lineage called γ-proteobacteria). We note, however, that having low proportions of MSAs with precise matches is not a feature of proteins from γ-proteobacteria; the precise match percentages for the other γ-proteobacteria we examined (Alteromonadaceae, Chromatiales, Enterobacteriaceae, and Xanthomonadaceae) range from 7.3% to 21.1%. Regardless of the specific details, it seems likely that the low recall when model fit is used as a classifier (at least when we use precise matches to examine recall) indicates that there is a high degree of variation among individual proteins in their patterns of sequence evolution.

Many different protein models trained on a variety of datasets have been published at this time; most of these models (e.g., the WAG [11] and LG [14] models) are similar to our general models in that they were trained on a wide variety of proteins. Most published models were trained using MSAs that include a wide variety of organisms (i.e., the models are not clade specific). Including those published models in our model fit analyses further emphasized the limited ability to classify protein MSAs using model fit (Supplementary File S2). As with the analyses using new models this is likely to reflect variation among proteins in their patterns of evolution. The percentage of cases where our new models (either general or ribosomal protein) had the best fit to a protein MSA ranged from 85.16% (for Thermoprotei) to 23% (for Coelomata), with a median of 44.38%. The percentage of times that a published generalized model (i.e., models trained on proteins from diverse taxa, like Q.pfam/Q.pfam-gb [8], WAG [11], LG [14], JTT [17], and VT [18]) had the best fit ranged from a minimum of 5.78% (Thermoprotei) to a maximum of 53.38% (coelomate animals). The percentage of times that a published clade-specific model (i.e., the clade-specific models from Pandey and Braun [7] and Minh et al. [8]) ranged from a minimum of 9.06% (Thermoprotei) to a maximum of 50.51% (Pezizomycotina). In general, clades with a high percentage of matches to the new models also had a high recall at the domain level (compare Table 1 and Supplementary File S2).

3.3. Genomic base composition affects amino acid frequencies and relative exchangabilities

The bacterial and archaeal clades with relatively high recall (Halobacteriaceae and Micrococcineae) also have the highest genomic GC content (Table 1). This led us to speculate that models of protein evolution trained using MSAs from GC-rich taxa might be more successful as classifiers than those trained using GC-poor taxa. If true, it suggests that the patterns of protein evolution for GC-rich taxa, which are what we are capturing in our models, might be especially distinctive relative to GC-poor taxa. The hypothesis that models for protein evolution in GC-rich taxa are especially distinctive, bot relative to each other and relative to the models for GC-poor taxa, is consistent with the observation that the other archaeal group with a very high precise match percentage (Thermoprotei) had the second highest genomic GC-content within Archaea (Table 1). This led us to ask whether there was a correlation between GC-content and the precise match percentage; however, we did not find a correlation (Spearman's $\varrho = 0.32383$, $P = 0.17622$). Thus, it seems likely that the observation that the most GC-rich clades in our sample also had a relatively high precise match percentages was coincidental.

Some REs did appear to be related to genomic GC content, despite the absence of a correlation between the precise match percentage and GC content. Focusing on bacteria, where we sampled the largest number of clades, the mean REs for high-GC bacteria (54.4% to 69.7% GC) and those for low-GC bacteria (37.7% to 52.2% GC) differed in specific ways: 1) the three REs most elevated in high-GC taxa all involved aromatic residues (F-Y, W-Y, and H-Y); and 2) the three REs most elevated in high-GC taxa all involved small residues (A-S, A-C, and C-S) (see Supplementary File S4 for all RE comparisons). ML estimates of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

equilibrium amino acid frequencies also exhibited a strong relationship with genomic GC content (Figure 2 and Supplementary File S4). The second result is consistent with earlier work that has shown that GC content has a proteome-wide impact on the observed amino acid composition [27,54]. The correlation between amino acid composition and genomic GC-content is likely to have an impact on estimates of REs; after all, the **Π** matrix (equilibrium amino acid frequencies) and ***R*** matrix (REs) are interrelated. Indeed, phenylalanine and tyrosine are both encoded by GC-poor codons and alanine is encoded by a GC-rich codon; this may explain some of the RE differences between the models for GC-rich and GC-poor taxa. However, the equilibrium frequency of tryptophan (π_W) was only weakly correlated with genomic GC-content (Supplementary Figure S1) and the other amino acids involved in REs with the largest differences between GC-rich and GC-poor taxa are not correlated with genomic GC-content. Overall, these results provide evidence that genomic GC-content has a major impact on amino acid composition and a modest impact on REs. However, genomic GC-content does not affect the overall model in such a way the fit becomes universally better as a classifier for either the high-GC or low-GC subsets of taxa.

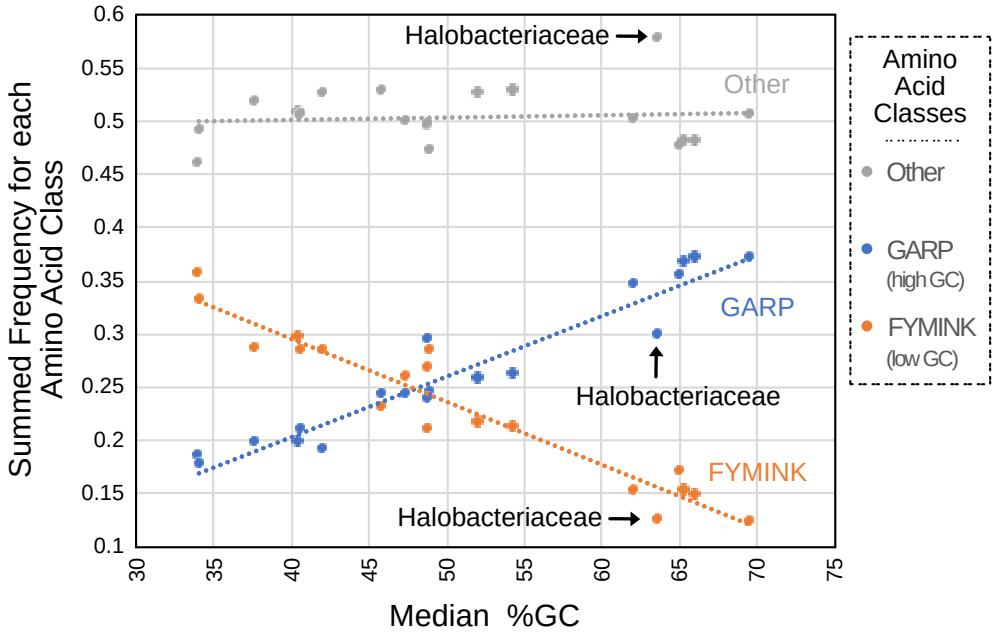

Figure 2. ML estimates of equilibrium amino acid frequencies are related to genomic GC-content. The summed equilibrium frequencies for each amino acid class correspond to values for the general models. We considered three amino acid classes: 1) GARP amino acids (glycine, alanine, arginine, and proline; encoded by GC-rich codons); 2) FYMINK amino acids (phenylalanine, tyrosine, methionine, isoleucine, asparagine, and lysine; encoded by GC-poor codons); and other amino acids. Summed equilibrium frequencies for GARP have a strong positive relationship with GC content (f_{GARP} = 0.0057[%GC] - 0.0245; r^2 = 0.9222) whereas the summed equilibrium frequencies for FYMINK have a strong negative relationship with GC content (f_{FYMINK} = -0.0059[%GC] + 0.5321; r^2 = 0.9183). The sum of the equilibrium frequencies for the other amino acids was unrelated to genomic GC content. Halobacteriaceae, which has the largest deviation from the line for other amino acids, is indicated on the graph (the deviation in the "other" category is positive and this is likely to be the reason for the negative deviations in both the GARP and the FYMINK categories). Similar data for the individual amino acids involved in REs that exhibited large differences between high-GC and low-GC bacteria are presented in Supplementary Figure S1.

3.4. General models and ribosomal protein models exhibit specific differences

The RE values that exhibited the largest differences when general models and ribosomal protein models were compared fit expectations of the Pandey and Braun [7] "rule of opposites." The rule of opposites is the observation that the REs for pairs of amino acids

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

that have low frequencies in specific structural environments are elevated; the rule's name reflects the observation that the behavior of REs is the opposite of equilibrium frequencies. Pandey and Braun [7] explored the relationship between REs and solvent exposure in globular proteins; they found that REs for pairs of polar amino acids are elevated for buried residues and that REs for pairs of hydrophobic amino acids are elevated for solvent exposed residues. Gordon et al. [26] observed a similar pattern in their comparison of transmembrane helices and extramembrane residues. Although this study did not consider protein structure, the observation that the RE value in ribosomal protein models with the greatest elevation relative to the general models involved the acidic amino acids (D-E; Figure 3) is a similar result because acidic residues are rare in ribosomal proteins [55]. Likewise, the RE that was most elevated in general models relative to ribosomal protein models involved basic amino acids (R-K), which are common in ribosomal proteins. Comparison of high- and low-GC bacteria highlighted F-Y as the most elevated RE in high-GC bacteria, and estimated equilibrium frequencies of both phenylalanine and tyrosine are relatively low in high-GC taxa (Supplementary Figure S1). Thus, comparisons between generalized models and ribosomal protein models and, at least to some degree, those between models for high- and low-GC bacteria, indicate that the rule of opposites extends beyond different structural environments to complete MSAs.

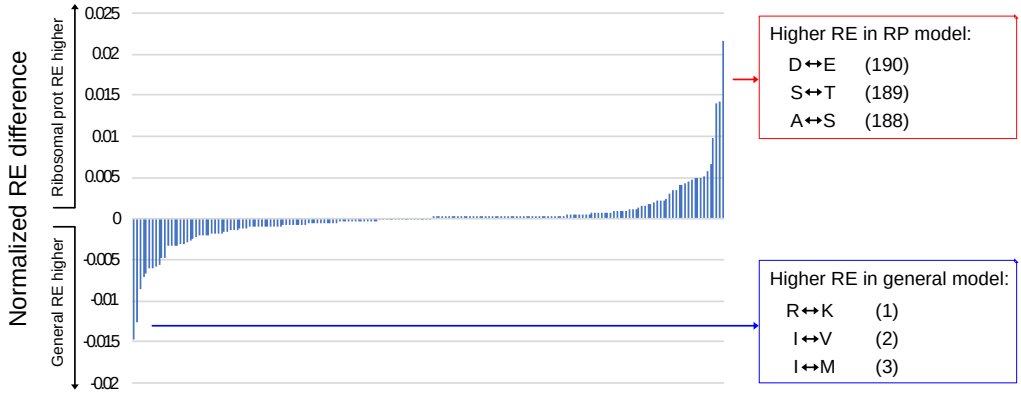

Figure 3. Comparison of REs for ribosomal protein models and general models. Normalized averages of the REs for general models were subtracted from REs for ribosomal protein models and sorted in ascending order. Large negative values correspond to higher REs in general models and large positive values correspond to higher REs in the ribosomal protein models. The parenthetical numbers after each exchange correspond to the rank order of the differences when the RE differences are sorted from smallest to largest (i.e., 1 is the RE that is most elevated in the average general model relative to the average ribosomal protein model, 2 is the second most elevated, and so forth until 190, which are the REs that are most elevated in the average ribosomal protein model relative to the average general model). All RE estimates are available in Supplementary File S5.

The rule of opposites is likely to reflect, at least in part, our assumption that models of protein evolution are locally time reversible. After all, to achieve time reversibility, large REs are necessary to explain substitutions that result in low frequency amino acids. This means that the RE values for rare-rate pairs must be elevated if the MSAs provide evidence for at least some substitutions involving pairs of rare amino acids. However, we believe that finding RE differences that are consistent with the rule in multiple matrices provides evidence that the rule is likely to have a biological basis (as opposed to a purely methodological basis). This can be illustrated by considering two alternative hypotheses regarding substitutions involving rare amino acids: 1) rare amino acids, when they are present, are typically necessary for specific functions and are therefore highly conserved; and 2) rare amino acids are not especially conserved. Hypothesis 1 predicts a very low number of rare-rare exchanges and will therefore result in low RE estimates for those pairs if the dataset used to estimate the $\boldsymbol{R}$ matrix is large enough. In contrast, hypothesis 2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

predicts that the number of rare-rare exchanges relative to their frequencies will be large and it therefore it predicts high RE estimates. Of course, both hypotheses predict that the small number of site patterns that can provide information about rare-rare REs will increase the variance of the rare-rare RE estimates. Thus, one could reconcile hypothesis 1 with the observation that some rare-rare RE estimates are high in specific ***R*** matrices by invoking the high variance expected for those RE estimates. On the other hand, hypothesis 1 cannot be reconciled with the observation that rare-rare RE estimates are high across multiple ***R*** matrices estimated using independent datasets; that is only expected if hypothesis 2 is correct. Both Pandey and Braun [7] and Gordon et al. [26] examined multiple ***R*** matrices trained using multiple independent datasets and they found high rare-rare RE estimates for distinct structural environments. The observation that rare-rare exchanges are elevated in the ribosomal protein versus generalized model comparisons and the results of comparisons between models for GC-rich and GC-poor taxa (see above, section 3.3) further corroborates the second hypothesis.

3.5. Aromatic-aromatic REs differentiate bacterial versus archaeal/eukaryotic models

Most RE values in the average bacterial model and the average archaeal/eukaryotic model were quite similar (note the large number of RE differences that are near zero in Figure 3). A similar pattern emerges when we examine a bivariate plot of bacterial versus archaeal/eukaryotic RE differences for general and ribosomal protein models (Figure 4) The F-Y exchange was the most extreme difference between Bacteria and Archaea+eukaryotes; it was elevated in the average bacterial model of both types relative to the average archaeal/eukaryotic model (Figure 4).

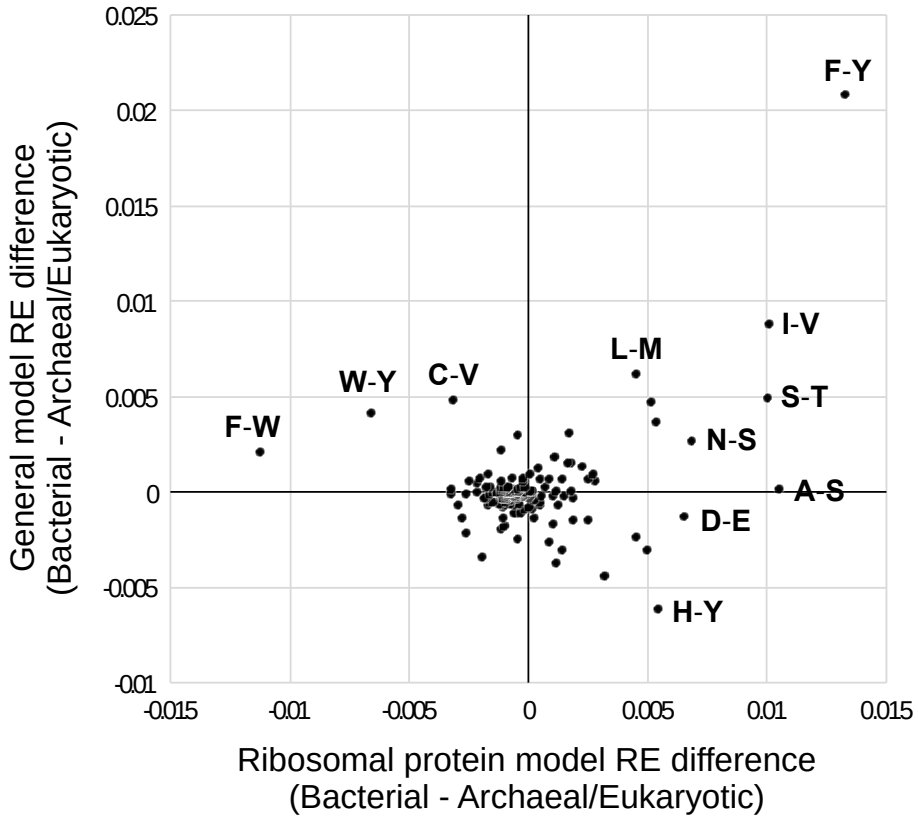

Figure 4. Comparison of RE difference values for bacteria versus archaea/eukaryotes for the ribosomal protein and general models. RE differences were generated by subtracting the average values for Archaea+eukaryotes from those for bacteria. Archaea and eukaryotes were combined based on the near perfect bipartition (for generalized models) or perfect bipartition (for ribosomal protein models) that separated those taxa (see Figure 1). Positive values on each axis correspond to REs that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

are higher in models trained on Bacteria than in those trained using MSAs from Archaea and eukaryotes; negative values are higher in models trained on Archaea+eukaryotes than in those based on bacterial MSAs. REs for all models (along with these RE differences in this graph) are available in Supplementary File S5.

Other RE differences were less extreme, but a number were still outliers. For example, the histidine-tyrosine (H-Y; Figure 4) RE are elevated in the average general archaeal/eukaryotic model (relative to the average general bacterial model) but the opposite is true for ribosomal protein models. In both cases, these exchanges have fairly extreme ranks (this ranking is similar to Figure 3, with a rank of 1 indicating the RE most elevated in the average archaeal/eukaryotic model and a rank of 190 indicating the RE most elevated in the average bacterial model). Specifically, the H-Y RE has rank 1 for general models and rank 184 for ribosomal protein models (in contrast to F-Y, which has rank 190 in both general models). F-W and W-Y are elevated in the average archaeal/eukaryotic ribosomal protein models (ranks 1 and 2, respectively) but less different from the majority of REs for general models (F-W and W-Y had ranks 178 and 184, respectively, for general models). Although some outlier RE differences did not involve aromatic residues (including the very conservative I-V exchange; Figure 4), we emphasize that four of the six possible aromatic-aromatic exchanges were outliers.

3.6. Models trained using Methanomicrobia MSAs are outliers within Archaea

The generalized model tree (Figure 1a) and the ribosomal protein model tree (Figure 1b) both include a bipartition that divides Methanomicrobia+Bacteria from the other Archaea and eukaryotes. In fact, the generalized model for Methanomicrobia clustered to the generalized model for a bacterial clade (Selenomonadales). This suggests that, based on REs, Methanomicrobia had the most distinctive patterns of protein sequence evolution relative to other Archaea. The largest RE difference between Methanomicrobia and the remaining Archaea and eukaryotes was the F-Y pair (Supplementary File S3), with the Methanomicrobia model having a larger value than the mean for the other Archaea and eukaryotes. Several amino acid pairs that include cysteine also had larger REs in the generalized Methanomicrobia model than in the other generalized Archaea+eukaryotes models; more specifically, C-S, C-Y, and C-F had ranks two, three, and four when the differences between the Methanomicrobia are subtracted from the mean normalized REs for other Archaea+eukaryotes. The C-W difference was ranked eighth in the same analysis (Supplementary File S3). Since three of these examples reflect cysteine-aromatic pairs, the generally high variation in the REs for aromatic residues could contribute to the variation in those pairs. However, cysteine has an unusual role in methonogen proteomes; methanogens have cysteine-rich proteomes and they sometimes have an unusual cysteinyl-tRNA synthesis pathway [56]. Our estimate of the equilibrium frequency of cysteine in the Methanomicrobia model (π_C = 0.00953) was the highest value out of all the archaeal models (π_C range for Archaea = 0.00325 - 0.00953; median π_C = 0.0078, see Supplementary File S3). The observation that the high REs involving cysteine occurred in an archaeote with an elevated (rather than reduced) cysteine frequency is not consistent with the rule of opposites, but we view it as likely to be biologically relevant given the role of cysteine in methanogen proteomes.

The unexpected placement of Methanomicrobia in the trees of models might lead one to hypothesize that those models were either especially unique, suggesting a high percentage of precise matches for Methanomicrobia, or that the REs for Methanomicrobia were intermediate between those for Archaea and Bacteria, suggesting a large number of cases where a Bacterial model had the best fit to an MSA from Methanomicrobia. Neither of these predictions were true (Table 1); the percentage of precise matches was relatively low and the percentage of cases where a Bacterial model had the best fit to an MSA from Methanomicrobia was not especially high (it was actually lower than the percentages for two other archaeal lineages, Thaumarchaeota and DPANN; Table 1). Taken as a whole, these results suggest patterns of sequence evolution for Methanomicrobia have some

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

unusual features, but they also indicate that patterns of sequence evolution in that clade are not exceptionally distinctive relative to the other lineages that we examined (at least using the metrics employed in this study).

3.7. Models trained using Methanomicrobia MSAs are outliers within Archaea

The REs that appeared to exhibit the largest degree of variation involved specific types of amino acids, most notably aromatic residues (Figure 4). This observation has several possible interpretations, the simplest of which could be related to the relatively distinctive chemical properties of aromatic amino acids. Specifically, it has long been recognized that aromatic residues play an important role in protein structure, both via aromatic-aromatic interactions [57–60] and through interactions between the aromatic rings and nitrogen and sulfur atoms in other amino acids [61,62]. This might make the REs for those amino acid pairs especially sensitive to global shifts in the patterns of selection on the proteome. Another possibility is that there was a major, ancient shift in REs on the branch separating Bacteria from Archaea+eukaryotes, followed by changes in response to specific factors such as the shifts in genomic base composition or environmental shifts (e.g., high salt concentrations in Halobacteriaceae). Notably, the aromatic amino acids, as well as some of the other amino acids involved in highly variable pairs, are thought to be late additions to the genetic code and to have increased in frequency since the time of the last universal common ancestor of life [63–65]. Regardless of the basis for the RE shifts, it seems clear that the shifts represent an important feature of protein evolution.

Model fit had a relatively limited ability to classify MSAs in this study. It was only able to classify protein MSAs with ≥70% recall in 11 out of the 19 clades we examined (Table 1), and we only achieved this high recall when we considered domain-level classification. This raises an important question: do estimates of GTR_{20} model parameters provide a meaningful picture of evolutionary model space for proteins? After all, it is obvious that the GTR_{20} model simplifies the underlying processes of protein sequence evolution. However, we do not believe that the poor performance of model fit as a classifier is evidence that our approach is a problem for several reasons. Model fit is likely to perform poorly if the variance among individual proteins in their patterns of sequence evolution is greater than the variance among clades. Thus, the ability to classify alignments at the level of domain indicates that our approach is capturing a biological signal. Likewise, it was reasonable to expect Halobacteriaceae proteins to exhibit distinctive patterns of sequence evolution given their high intracellular ion concentrations [51] and that was one of the two clades with a high recall. The basis for the high recall for Thermoprotei is less clear, but it also provides evidence that GTR_{20} model parameters captured a biological signal. Obviously, more "biologically realistic" models might reveal additional signals in the data, but the fact remains that our approach has captured biologically relevant signals. Moreover, a fundamental problem for all efforts to improve the realism of models of sequence evolution is that they may require the addition of free parameters and it is possible that some of those parameters will be non-identifiable (or only weakly identifiable; see Ponciano et al. [66] for discussion). Likewise, improved models might lead to more realistic distances among models. However, what is clear is that these models capture biological signals and that suggests that REs have the potential to reveal meaningful information about changes in patterns of sequence evolution over deep

4. Conclusions

We hypothesized that comparisons of estimated amino acid REs for various clades, obtained using the GTR_{20} model, would reveal a picture of evolutionary model space for proteins. Trees of models generated by clustering distances among the REs estimated using a set of clades spread across the tree of life revealed structures similar to the tree of life, both for REs estimated using arbitrarily selected proteins and ribosomal proteins. There were no obvious clusters that separated taxa that grow in distinct environments or metabolic features, suggesting that phylogeny has a stronger influence on the structure of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

protein evolutionary model space than ecology. However, we did find evidence that the models for two archaeal clades (Halobacteriaceae and Thermoprotei) were especially distinctive. The evidence for the distinctive nature of the Halobacteriaceae and Thermoprotei models was the good performance of model fit as a classifier when used with MSAs from these clades. However, the performance of model fit as a classifier was generally poor; indeed, classification with high recall at the level domain was only possible for a subset of clades. This differs from the results in earlier studies that estimated models for several different eukaryotic groups [7,8,67] and found recall was often quite high. This study focused on a wide variety of clades across the tree of life, so it is tempting to speculate that variation among individual proteins in the patterns of evolution may be greater in Bacteria and Archaea, which dominated the taxon sample for this study. We found that a specific subset of REs tended to exhibit differences among taxa, most notably exchanges involving aromatic amino acids. We also observed differences in REs that appeared to be related to amino acid frequencies that were evident in model comparisons; those comparisons included models estimated using diverse proteins versus ribosomal proteins as well as models estimated using proteins from high-GC versus low-GC bacteria. Taken as a whole, these analyses demonstrate that REs estimated using the GTR_{20} model yield a meaningful picture of evolutionary model space for proteins.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Equilibrium frequencies of selected amino acids compared with genomic GC-content, File S1: Excel file listing clades and taxa used for these analyses, File S2: Excel file with model fit results, File S3: Excel file with comparisons of archaeal exchangeabilities and state frequencies, File S4: Excel file with comparisons of bacterial exchangeabilities and state frequencies, File S5: Excel file with comparisons of model exchangeabilities.

Author Contributions: Conceptualization, G.E.S. and E.L.B..; methodology, G.E.S. and E.L.B.; software, E.L.B.; validation, G.E.S. and E.L.B.; formal analysis, G.E.S.; investigation, G.E.S.; resources, E.L.B..; data curation, G.E.S. and E.L.B.; writing—original draft preparation, G.E.S. and E.L.B.; writing—review and editing, G.E.S. and E.L.B.; visualization, E.L.B.; supervision, E.L.B.; project administration, E.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data used to train models and model parameters are available from Zenodo [43]. The model parameter files in a format that several phylogenetics programs (e.g. IQ-TREE) are also available from github (https://github.com/ebraun68/protmodels, accessed 20 December 2022).

Acknowledgments: We are grateful to Rebecca Kimball for constructive feedback on earlier versions of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zuckerkandl, E.; Pauling, L. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*; Bryson, V.; Vogel, H.J., Ed.; Elsevier, **1965**; pp. 97–166, doi:10.1016/B978-1-4832-2734-4.50017-6.
2. Dayhoff, M.O.; Eck, R.V. The chemical meaning of amino acid mutations. In *Atlas of Protein Sequence and Structure*; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Silver Springs, MD, **1969**; Vol. 4, pp. 85–87.
3. Kimura, M.; Ohta, T. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 2848–2852, doi:10.1073/pnas.71.7.2848.
4. Sayers, E.W.; Cavanaugh, M.; Clark, K.; Pruitt, K.D.; Schoch, C.L.; Sherry, S.T.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2021**, *49*, D92–D96, doi:10.1093/nar/gkaa1023.
5. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489, doi:10.1093/nar/gkaa1100.
6. Zou, Z.; Zhang, J. Amino acid exchangeabilities vary across the tree of life. *Sci. Adv.* **2019**, *5*, eaax3124, doi:10.1126/sciadv.aax3124.
7. Pandey, A.; Braun, E.L. Protein evolution is structure dependent and non-homogeneous across the tree of life. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Virtual Event, 21–24 September 2020; ACM: New York, NY, USA, 2020; pp. 1–11. doi:10.1145/3388440.3412473.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

8. Minh, B.Q.; Dang, C.C.; Vinh, L.S.; Lanfear, R. Qmaker: fast and accurate method to estimate empirical models of protein evolution. *Syst. Biol.* **2021**, *70*, 1046–1060, doi:10.1093/sysbio/syab010.
9. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In Lectures on Mathematics in the Life Sciences; Miura, R.M., Ed.; The American Mathematical Society: Providence, RI, **1986**, 17, 57–86.
10. Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **1994**, *39*, 105–111, doi:10.1007/BF00178256.
11. Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699, doi:10.1093/oxfordjournals.molbev.a003851.
12. Braun, E.L. An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins. *Bioinformatics* **2018**, *34*, i350–i356, doi:10.1093/bioinformatics/bty261.
13. Tiessen, A.; Pérez-Rodríguez, P.; Delaye-Arredondo, L.J. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes* **2012**, *5*, 85, doi:10.1186/1756-0500-5-85.
14. Le, S.Q.; Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **2008**, *25*, 1307–1320, doi:10.1093/molbev/msn067.
15. Kishino, H.; Miyata, T.; Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **1990**, *31*, 151–160, doi:10.1007/BF02109483.
16. Dayhoff, M.O.; Schwartz, R.M.; Orcutt, B.C. A model of evolutionary change in proteins. Dayhoff, M.O. (ed.), Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Silver Springs, MD, **1978**, 5, 345–352.
17. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282, doi:10.1093/bioinformatics/8.3.275.
18. Müller, T.; Vingron, M. Modeling amino acid replacement. *J. Comput. Biol.* **2000**, *7*, 761–776, doi:10.1089/10665270050514918.
19. Dimmic, M.W.; Rest, J.S.; Mindell, D.P.; Goldstein, R.A. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* **2002**, *55*, 65–73, doi:10.1007/s00239-001-2304-y.
20. Nickle, D.C.; Heath, L.; Jensen, M.A.; Gilbert, P.B.; Mullins, J.I.; Kosakovsky Pond, S.L. HIV-specific probabilistic models of protein evolution. *PLoS ONE* **2007**, *2*, e503, doi:10.1371/journal.pone.0000503.
21. Dang, C.C.; Le, Q.S.; Gascuel, O.; Le, V.S. FLU, an amino acid substitution model for influenza proteins. *BMC Evol. Biol.* **2010**, *10*, 99, doi:10.1186/1471-2148-10-99.
22. Le, T.K.; Vinh, L.S. FLAVI: an amino acid substitution model for flaviviruses. *J. Mol. Evol.* **2020**, *88*, 445–452, doi:10.1007/s00239-020-09943-3.
23. Adachi, J.; Waddell, P.J.; Martin, W.; Hasegawa, M. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* **2000**, *50*, 348–358, doi:10.1007/s002399910038.
24. Rota-Stabelli, O.; Yang, Z.; Telford, M.J. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. *Mol. Phylogenet. Evol.* **2009**, *52*, 268–272, doi:10.1016/j.ympev.2009.01.011.
25. Le, V.S.; Dang, C.C.; Le, Q.S. Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. *BMC Evol. Biol.* **2017**, *17*, 136, doi:10.1186/s12862-017-0987-y.
26. Gordon, E.L.; Kimball, R.T.; Braun, E.L. Protein structure, models of sequence evolution, and data type effects in phylogenetic analyses of mitochondrial data: A case study in birds. *Diversity* **2021**, *13*, 555, doi:10.3390/d13110555.
27. Singer, G.A.C.; Hickey, D.A. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* **2000**, *17*, 1581–1588, doi:10.1093/oxfordjournals.molbev.a026257.
28. Singer, G.A.C.; Hickey, D.A. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **2003**, *317*, 39–47, doi:10.1016/S0378-1119(03)00660-7.
29. Fukuchi, S.; Yoshimune, K.; Wakayama, M.; Moriguchi, M.; Nishikawa, K. Unique amino acid composition of proteins in halophilic bacteria. *J. Mol. Biol.* **2003**, *327*, 347–357, doi:10.1016/s0022-2836(03)00150-5.
30. Schmidt, A.; Rzanny, M.; Schmidt, A.; Hagen, M.; Schütze, E.; Kothe, E. GC content-independent amino acid patterns in bacteria and archaea. *J. Basic Microbiol.* **2012**, *52*, 195–205, doi:10.1002/jobm.201100067.
31. Reed, C.J.; Lewis, H.; Trejo, E.; Winston, V.; Evilia, C. Protein adaptations in archaeal extremophiles. *Archaea* **2013**, 2013, 373275, doi:10.1155/2013/373275.
32. Pasamontes, A.; Garcia-Vallve, S. Use of a multi-way method to analyze the amino acid composition of a conserved group of orthologous proteins in prokaryotes. *BMC Bioinformatics* **2006**, *7*, 257, doi:10.1186/1471-2105-7-257.
33. Hug, L.A.; Baker, B.J.; Anantharaman, K.; Brown, C.T.; Probst, A.J.; Castelle, C.J.; Butterfield, C.N.; Hernsdorf, A.W.; Amano, Y.; Ise, K.; Suzuki, Y.; Dudek, N.; Relman, D.A.; Finstad, K.M.; Amundson, R.; Thomas, B.C.; Banfield, J.F. A new view of the tree of life. *Nat. Microbiol.* **2016**, 1, 16048, doi:10.1038/nmicrobiol.2016.48.
34. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, 26, 2460–2461, doi:10.1093/bioinformatics/btq461.
35. Dang, C.C.; Lefort, V.; Le, V.S.; Le, Q.S.; Gascuel, O. ReplacementMatrix: a web server for maximum-likelihood estimation of amino acid replacement rate matrices. *Bioinformatics* **2011**, 27, 2758–2760, doi:10.1093/bioinformatics/btr435.
36. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, 32, 268–274, doi:10.1093/molbev/msu300.
37. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, 4, 406–425, doi:10.1093/oxfordjournals.molbev.a040454.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

38. Swofford, D.L. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).; Sinauer Associates: Sunderland, Massachusetts, **2003**.
39. Bogdanowicz, D.; Giaro, K. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *9*, 150-160, doi:10.1109/TCBB.2011.38.
40. Lin, Y.; Rajan, V.; Moret, B.M.E. A metric for phylogenetic trees based on matching. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1014–1022, doi:10.1109/TCBB.2011.157.
41. Penny, D.; Foulds, L.R.; Hendy, M.D. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* **1982**, *297*, 197–200, doi:10.1038/297197a0.
42. Schwarz, G. Estimating the dimension of a model. *Ann. Statist.* **1978**, *6*, 461–464, doi:10.1214/aos/1176344136.
43. Scolaro, G.E.; Braun, E.L. Date for: The structure of evolutionary model space for proteins across the tree of life (1.0) [Data set]. *Zenodo* **2022**, doi:10.5281/zenodo.7463835
44. Pandey, A.; Braun, E.L. Phylogenetic analyses of sites in different protein structural environments result in distinct placements of the metazoan root. *Biology* **2020**, *9*, doi:10.3390/biology9040064.
45. Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5088–5090, doi:10.1073/pnas.74.11.5088.
46. Eme, L.; Spang, A.; Lombard, J.; Stairs, C.W.; Ettema, T.J.G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **2017**, *15*, 711–723, doi:10.1038/nrmicro.2017.133.
47. Castelle, C.J.; Wrighton, K.C.; Thomas, B.C.; Hug, L.A.; Brown, C.T.; Wilkins, M.J.; Frischkorn, K.R.; Tringe, S.G.; Singh, A.; Markillie, L.M.; Taylor, R.C.; Williams, K.H.; Banfield, J.F. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **2015**, *25*, 690–701, doi:10.1016/j.cub.2015.01.014.
48. Williams, T.A.; Szöllősi, G.J.; Spang, A.; Foster, P.G.; Heaps, S.E.; Boussau, B.; Ettema, T.J.G.; Embley, T.M. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E4602–E4611, doi:10.1073/pnas.1618463114.
49. Rinke, C.; Schwientek, P.; Sczyrba, A.; Ivanova, N.N.; Anderson, I.J.; Cheng, J.-F.; Darling, A.; Malfatti, S.; Swan, B.K.; Gies, E.A.; Dodsworth, J.A.; Hedlund, B.P.; Tsiamis, G.; Sievert, S.M.; Liu, W.-T.; Eisen, J.A.; Hallam, S.J.; Kyrpides, N.C.; Stepanauskas, R.; Rubin, E.M.; Woyke, T. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **2013**, *499*, 431–437, doi:10.1038/nature12352.
50. Brown, C.T.; Hug, L.A.; Thomas, B.C.; Sharon, I.; Castelle, C.J.; Singh, A.; Wilkins, M.J.; Wrighton, K.C.; Williams, K.H.; Banfield, J.F. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **2015**, *523*, 208–211, doi:10.1038/nature14486.
51. Oren, A. Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes. *Front. Microbiol.* **2013**, 4, 315, doi:10.3389/fmicb.2013.00315.
52. Kumar, S.; Tsai, C.J.; Nussinov, R. Factors enhancing protein thermostability. *Protein Eng. Des. Sel.* **2000**, *13*, 179–191, doi:10.1093/protein/13.3.179.
53. Blanquart, S.; Groussin, M.; Le Roy, A.; Szöllosi, G.J.; Girard, E.; Franzetti, B.; Gouy, M.; Madern, D. Resurrection of ancestral malate dehydrogenases reveals the evolutionary history of halobacterial proteins: deciphering gene trajectories and changes in biochemical properties. *Mol. Biol. Evol.* **2021**, *38*, 3754–3774, doi:10.1093/molbev/msab146.
54. Chen, W.; Shao, Y.; Chen, F. Evolution of complete proteomes: guanine-cytosine pressure, phylogeny and environmental influences blend the proteomic architecture. *BMC Evol. Biol.* **2013**, *13*, 219, doi:10.1186/1471-2148-13-219.
55. Lott, B.B.; Wang, Y.; Nakazato, T. A comparative study of ribosomal proteins: Linkage between amino acid distribution and ribosomal assembly. *BMC Biophys.* **2013**, *6*, 13, doi:10.1186/2046-1682-6-13.
56. Klipcan, L.; Frenkel-Morgenstern, M.; Safro, M.G. Presence of tRNA-dependent pathways correlates with high cysteine content in methanogenic Archaea. *Trends Genet.* **2008**, *24*, 59–63, doi:10.1016/j.tig.2007.11.007.
57. Burley, S.K.; Petsko, G.A. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **1985**, *229*, 23–28, doi:10.1126/science.3892686.
58. Singh, J.; Thornton, J.M. The interaction between phenylalanine rings in proteins. *FEBS Lett.* **1985**, *191*, 1–6, doi:10.1016/0014-5793(85)80982-0.
59. McGaughey, G.B.; Gagné, M.; Rappé, A.K. π-Stacking Interactions. *J. Biol. Chem.* **1998**, *273*, 15458–15463, doi:10.1074/jbc.273.25.15458.
60. Chourasia, M.; Sastry, G.M.; Sastry, G.N. Aromatic-Aromatic Interactions Database, A(2)ID: an analysis of aromatic π-networks in proteins. *Int. J. Biol. Macromol.* **2011**, *48*, 540–552, doi:10.1016/j.ijbiomac.2011.01.008.
61. Burley, S.K.; Petsko, G.A. Amino-aromatic interactions in proteins. *FEBS Lett.* **1986**, *203*, 139–143, doi:10.1016/0014-5793(86)80730-x.
62. Zauhar, R.J.; Colbert, C.L.; Morgan, R.S.; Welsh, W.J. Evidence for a strong sulfur-aromatic interaction derived from crystallographic data. *Biopolymers* **2000**, *53*, 233–248, doi:10.1002/(SICI)1097-0282(200003)53:3<233::AID-BIP3>3.0.CO;2-4.
63. Brooks, D.J.; Fresco, J.R.; Lesk, A.M.; Singh, M. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **2002**, *19*, 1645–1655, doi:10.1093/oxfordjournals.molbev.a003988.
64. Trifonov, E.N. The triplet code from first principles. *J. Biomol. Struct. Dyn.* **2004**, *22*, 1–11, doi:10.1080/07391102.2004.10506975.
65. Higgs, P.G.; Pudritz, R.E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **2009**, *9*, 483–490, doi:10.1089/ast.2008.0280.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2022 doi:10.20944/preprints202212.0403.v1

66. Ponciano, J.M.; Burleigh, J.G.; Braun, E.L.; Taper, M.L. Assessing parameter identifiability in phylogenetic models using data cloning. *Syst. Biol.* **2012**, *61*, 955–972, doi:10.1093/sysbio/sys055.
67. Dang, C.C.; Minh, B.Q.; McShea, H.; Masel, J.; James, J.E.; Vinh, L.S.; Lanfear, R. nQMaker: estimating time non-reversible amino acid substitution models. *BioRxiv* **2021**, doi:10.1101/2021.10.18.464754.