Peer-reviewed version available at *Biology* **2020**, *9*, 64; doi:10.3390/biology9040064

Article

Phylogenetic Analyses of Sites in Different Protein Structural Environments Result in Distinct Placements of the Metazoan Root

Akanksha Pandey [1] and Edward L. Braun [1,2,*]

[1] Department of Biology, University of Florida; a.pandey@ufl.edu
[2] Genetics Institute, University of Florida; ebraun68@ufl.edu

* Correspondence: ebraun68@ufl.edu

Abstract: Phylogenomics, the use of large datasets to examine phylogeny, has revolutionized the study of evolutionary relationships. However, genome-scale data have not been able to resolve all relationships in the tree of life; this could reflect, at least in part, the poor-fit of the models used to analyze heterogeneous datasets. Some of the heterogeneity may reflect the different patterns of selection on proteins based on their structures. To test that hypothesis, we developed a pipeline to divide phylogenomic protein datasets into subsets based on secondary structure and relative solvent accessibility. We then tested whether amino acids in different structural environments had distinct signals for the topology of the deepest branches in the metazoan tree. The most striking difference in phylogenetic signal reflected relative solvent accessibility; analyses of exposed sites (on the surface of proteins) yielded a tree that placed ctenophores sister to all other animals whereas sites buried inside proteins yielded a tree with a sponge-ctenophore clade. These differences in phylogenetic signal were not ameliorated when we repeated our analyses using the CAT model, a mixture model that is often used for analyses of protein datasets. In fact, the heterogeneous CAT model resulted in several rearrangements that are unlikely to represent evolutionary history. However, analyses conducted after recoding amino acids to limit the impact of deviations from compositional stationarity increased the congruence in the estimates of phylogeny for exposed and buried sites; after recoding amino acids both trees supported placement of ctenophores sister to all other animals. These results provide striking evidence that it is necessary to achieve a better understanding of the constraints due to protein structure to improve phylogenetic estimation.

Keywords: phylogenomics; protein structure; secondary structure; relative solvent accessibility CAT model; non-stationary models; RY coding; metazoan phylogeny; ctenophora; porifera

1. Introduction

The growing availability of very large molecular datasets has transformed the field of phylogenetics. The use of these phylogenomic datasets was suggested to "end incongruence" among phylogenetic estimates by reducing the stochastic error associated with analyses of small datasets [1]. Although phylogenomic analyses have resolved many contentious relationships [2–5], in other cases different analyses using genome-scale data have produced multiple distinct resolutions of problematic nodes, sometimes with strong support [6–11]. This suggests that analyses of these large datasets can be misled by non-historical signals that may not be as apparent in analyses of smaller datasets. Thus, rather than putting an end to incongruence, phylogenomic analyses often highlight the complexity of the phylogenetic signals present in genomic data.

The idea that non-historical signals can overwhelm historical signal in large datasets predates the phylogenomic era (e.g., the early discussion of statistical inconsistency due to long-branch attraction [12,13]). However, the availability of genome-scale data emphasizes the large number of cases where conflicting signals emerge in phylogenetic analysis. The most extreme cases correspond to those where analytical methods are subject to systematic error; these are the cases where

non-historical signal overwhelms historical signal. In that part of parameter space, increasing the amount of data will cause phylogenetic methods to converge on inaccurate estimates of evolutionary history with high support [14]. Thus, phylogenomic analyses should lead either to high support for the true tree (the desired result) or to high support for an incorrect tree (if the analytical method is subject to systematic error). Cases where support is limited despite the use of large amounts of data could reflect one of two phenomena: 1) the data contains a mixture of signals; or 2) the underlying species tree contains a hard polytomy (and, therefore, historical signal is absent). If the former, the mixture of signals could be biological in nature (e.g., a mixture of histories due to reticulation) or it could reflect the existence of both a historical signal and one or more non-historical signals. Understanding the distribution of historical and non-historical signal(s) in large-scale data matrices might provide insights into evolutionary processes and result in better understanding of analytical methods and their limitations with phylogenomic datasets.

The problem of systematic error in phylogenomic analyses has been addressed in several ways; one of the most popular ways to address systematic error has been the use of more complex, and presumably more realistic, models of sequence evolution. These complex models typically assume that phylogenomic data are very heterogeneous. Most of these complex models, such as the CAT model [15], the Thorne-Goldman-Jones structural models [16–18], and structural mixture models [19], introduce this heterogeneity by assuming that distinct patterns of sequence evolution (and therefore different models) characterize different sites in multiple sequence alignments. However, there have also been some efforts to develop models that assume the heterogeneity corresponds to distinct patterns of sequence evolution on different branches in the tree of life [20–22]. Regardless of the specific model under consideration, the degree to which these approaches ameliorate the impact of misleading signals remains a subject of debate (e.g., see the discussion of the CAT model by Whelan and Halanych [23]).

Identifying and examining conflicting signals within the phylogenomic data matrices has the potential to provide information about the biological basis for the heterogeneity and ultimately inform analytical approaches to deal with heterogeneous datasets. It should be possible to identify conflicting signals in phylogenomic data by dividing the data into subsets and asking whether phylogenetic analyses of those subsets support distinct trees. This raises the question of how large-scale datasets should be subdivided. Subdividing data matrices into individual loci is unlikely to be informative because individual loci are short and therefore have limited power to resolve difficult nodes [24]. Moreover, individual loci are expected to be associated with distinct gene trees due to factors such as the multispecies coalescent [25–27]. Therefore, we believe that the most interesting efforts to identify conflicting signal involves datasets that are large enough that the results of analyses do not reflect stochastic error (i.e., cases where a sufficient number of sites sampled from many different genes are available).

There are many different ways to subdivide phylogenomic datasets in two or more subsets that are still large enough to overcome stochastic error. The most logical way to subdivide phylogenomic datasets involve the use of partitions that can be defined *a priori* using non-phylogenetic criteria. The simplest criteria might be functional criteria (e.g., coding versus non-coding [28] or highly transcriptionally-active versus largely untranscribed regions [29]. There are two alternative hypotheses regarding the distribution of signal in any subdivided phylogenomic dataset:

H_0: Conflicting signals are randomly distributed with respect to functionally defined subsets of the data. This hypothesis predicts that separate analyses of those functionally defined subsets of the data matrix will yield trees with the same topology (possibly with lower support than the analysis of the complete dataset due to the smaller size of the subsets).

H_A: Conflicting signals are associated with functionally defined data subsets. Different subsets of the data matrix defined using functional information are associated with distinct signals (i.e., analyses of those subsets support different topologies when the subsets are analyzed separately).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2019 doi:10.20944/preprints201910.0302.v1

Peer-reviewed version available at Biology 2020, 9, 64; doi:10.3390/biology9040064

Failure to reject H_0 could reflect a genuinely random distribution of signal, failure to define the data subsets in an appropriate manner, or a definition of the subsets that reduces their size to the point where stochastic error dominates the analyses (at least for nodes that are difficult to resolve). The last issue is unlikely to be problematic because subdividing the original dataset too finely is likely to result in difficult nodes being resolved randomly with low support. The other two possibilities are essentially impossible to distinguish, so H_0 cannot be corroborated in a global manner. However, it is still possible to corroborate (or falsify) H_A for specific ways of subdividing phylogenomic data matrices if they are sufficiently large. Efforts to do this may provide substantial information about the patterns of sequence evolution in different parts of the genome; this information is also likely to be useful for phylogenetic model development.

Coding sequences are often used to examine phylogenetic relationships deep in the tree of life (e.g., early land plant [5] or metazoan evolution [6,11] and efforts to identify to root of archaea [30] or eukaryotes [31]) and they are good candidates for this type of signal exploration. There are many biologically motivated ways to divide proteins into subsets that may reflect different selective pressures to maintain their structure and function. Ideally the subsets should be stable over evolutionary time, and it should be practical to assign aligned sites to the subsets. Protein structure provides a good criterion for dividing the data into subsets. Protein structures diverge much more slowly than protein sequences [32,33], so it is reasonable to assume that alignments can simply be divided into structural classes that remain relatively constant over evolutionary time. Also, modern protein secondary structure prediction methods are relatively accurate (e.g., the SSPro and ACCPro programs are able to classify ~90% of residues accurately for proteins with homologs in PDB [34]), so it is possible to construct analytic pipelines to define the structural subsets.

In this study, we examined whether different signals were associated with sites in different protein structural environments. Specifically, we subdivided globular proteins based on their structure and explored the distribution of signal supporting different placements of the root for the metazoan tree of life. Deep metazoan phylogeny has been a difficult phylogenetic problem, but there are a limited number of plausible arrangements for the major lineages (Figure 1). The traditional hypothesis (Figure 1a), which is supported by morphology [35] and some phylogenomic analyses (e.g., [9,11,36–38]), places sponges sister to all other metazoans. Other analyses of large-scale datasets (e.g., [7,39–41]) support the hypothesis that ctenophores are sister to all other metazoans (Figure 1b). The third possibility, a ctenophore-sponge clade (Figure 1c), was recovered in analyses the Ryan et al. [6] genomic dataset (hereafter called the "RG dataset"). However, support for the ctenophore-sponge clade is limited in some analyses of the RG dataset. Our analyses revealed that the signal associated with solvent exposed residues differed from associated with buried residues; we believe that this has implications for the fit of models that are currently used for phylogenetic analyses of protein sequences.

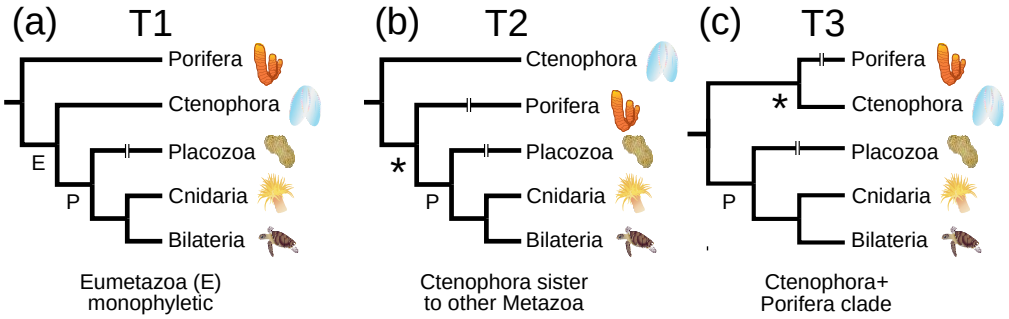

Figure 1. Topologies for the deepest branches in the metazoan tree recovered in phylogenomic analyses. (*a*) Porifera (sponges) sister to all other metazoa. This hypothesis includes a clade designated Eumetazoa (E). (*b*) Ctenophora (comb jellies) sister to all other metazoa. (*c*) A sponge-ctenophore clade sister to all other animals. All trees shown include a clade named Parahoxozoa (P) [42]. The topology for Parahoxozoa was fixed based on King and Rokas [10].

2. Materials and Methods

2.1. Dataset

The RG dataset comprises 242 orthologous protein-coding genes extracted from genomic data for 19 taxa. We used the alignments provided by Ryan et al. [6] without changes; the sequences had been aligned using CLUSTALW [43] with default parameters and poorly-aligned regions had been excluded using Gblocks [44]. We inspected all of the Ryan et al. [6] alignments visually in Geneious v. 9.1.5 (Biomatters Ltd., Auckland, New Zealand); none of the alignments appeared problematic. We used the TopCons prediction server [45] to determine whether there were any transmembrane proteins in the RG data. This resulted in the identification of 10 transmembrane proteins (Supplementary File S1), which were removed because our structural assignment pipeline was not appropriate for those proteins. This resulted in a 232-protein dataset with 102,000 sites and 19.8% missing data.

We conducted BLASTP searches of UNIPROT [46] to determine the identity and function of each gene (using annotation of the human sequence for most genes). There were 22 cases where the human sequence was absent from the RG alignment; in those cases, we used the *Drosophila* sequence to identify the genes. These functional annotations are available in Supplementary File S1. This analysis revealed that remaining proteins represent a diverse set of globular proteins without an overrepresentation of sequences from a particular protein family making it a relatively unbiased dataset. We provide all protein alignments as Nexus files with structural annotation, generated as described below, in Supplementary File S2. We analyzed two taxon samples: 1) the full taxon sample comprising all 19 taxa in the RG dataset; and 2) a reduced taxon sample that excluded four relatively divergent outgroup taxa [two taxa in Holozoa (*Capsaspora* and *Sphaeroforma*) that fall that fall outside of Metazoa and Choanozoa and two fungi (*Saccharomyces* and *Spizellomyces*)].

2.2. Structural Class Assignment

We assigned secondary structures to the MSAs using the SSpro and ACCpro programs in the SCRATCH 1D suite [47]. SSpro classifies sequences into three secondary structural classes (helix, sheet, and coil) [34]. ACCpro assigns each residue to one of the two categories: exposed (e) or buried (-) [48] with the latter defined as amino acids with <25% relative solvent accessibility. We used a weighted consensus sequence representing each protein in the RG dataset as input for SSpro and ACCpro. We generated a weighted consensus sequence for each protein using the Henikoff and Henikoff [49] method; the amino acid residue with the highest weight at each position was used in the consensus sequence. The structural data were extrapolated from the consensus sequence to the whole alignment and then the alignment written along with CHARSETS for each structural subset in a nexus file [50]. We then extracted sites of a given structural class from all the genes and created a concatenated alignment for a given structural class. The perl program for this analysis is available from (https://github.com/aakanksha12/Structural_class_assignment_pipeline).

2.3. Phylogenetic Analyses

We used RAxML v. 8.2.4 for log likelihood estimation and tree searches. We examined a set of standard empirical models (LG [51], WAG [52], VT [53], JTT-DCMUT [54] and rtREV [55]) with ML estimation of amino acid frequency parameters (e.g., the -m PROTGAMMALGX option in RAxML) and we used GTR model parameters optimized on the complete dataset (hereafter, we call the parameter estimates "grand GTR parameters") and various subsets of the data (see below for a description of the structural partitions). In all cases we also used GTRGAMMA (i.e., the GTR model combined with a four-category discrete approximation to the Γ distribution that describes rates across sites). We assessed the nodal support using rapid bootstrap [56] with the bootstopping criterion [57] as implemented in RaxML (i.e., the -N autoMR option).

To examine whether there were any observed differences among partitions in their signal we searched for decisive sites (cf. [58]). Decisive sites are the sites with a very large impact on the likelihood given each a specific topology. To do this we calculated per site log likelihoods for each candidate topology using RaxML (via the '-f G' option in the program). We calculated $\Delta\ln L$ for individual sites; we viewed sites with $\Delta\ln L > 5$ standard deviations from the mean $\Delta\ln L$ as decisive sites, following Kimball et al. [58].

We also conducted preliminary analyses of 232-protein dataset by dividing the alignment into structural subsets; analyses of the exposed and buried subsets resulted in two distinct trees identical to those in Figure 2. We then determined whether any specific gene strongly favored either topology using "outlier gene" analysis [59]. We found that gene 41 had a strong signal favoring the T3 (Figure 1) topology. In an unpartitioned analyses using the LG model the likelihood difference given the two trees in Figure 2 (see below, the two trees shown correspond to T2 and T3 in Figure 1) was more than three-fold greater for gene product 41 than for any of the other proteins in the data matrix (ΔlnL = 106.63 favoring the buried tree for protein 41 compared to a range of ΔlnL = 9.39 to 28.67 for the other proteins; see Supplementary File S3). Because protein 41 was an outlier relative to the other sequences we removed it to yield a 231-protein data matrix called the filtered Ryan genomic (FRG) dataset. We did not explore the basis for the unusual signal associated with gene 41; it could simply represent an incorrect orthology call in Ryan et al. [6]. We also emphasize that removing protein 41 did not alter the trees recovered after dividing sites into subsets based on structure (see Results).

2.4. Model estimation

We obtained ML estimates of the model parameters (amino acid exchangeabilities and amino acid frequencies) for each structural class using RAxML [60]. Our approach relaxes the notion that all sites evolve following the same Markovian process, but it still assumes that all the sites in the same subset of the MSA follow the same stationary and homogeneous Markov process (i.e., each subset has its own set of GTR parameter estimates). We examined the following classes:

1. Two relative solvent accessibility (RSA) based classes (EXPOSED and BURIED).
2. Three secondary structure-based classes (HELIX, SHEET, and COIL).
3. Six classes, combining RSA and secondary structure (HELIX_EXP, HELIX_BUR, SHEET_EXP, SHEET_BUR, COIL_EXP, and COIL_BUR).

We estimated the GTR model parameters for each structural class using the -f e option in RAxML and then performed a tree search. We used the ML tree from Ryan et al. [6] as a starting tree and estimated the model parameters. If the best tree obtained using the estimated parameters differed from the starting tree, we re-estimated model parameters using the new tree, iterating this procedure until the input and output tree converged (cf. [18]). Hereafter, the model parameters optimized for each structural class are further referred as structure-based model estimates; the parameter estimates are available in Supplementary File S4.

We used multidimensional scaling in R [61] to visualize the differences among estimates of exchange rate parameters obtained from the structural partitions as well as the standard empirical models. The exchangeability matrices for time-reversible models or protein sequence evolution models have 190 elements so we treated each model as a 190-element vector. These vectors were normalized so the elements summed to one and they were used to create a matrix of Euclidean distances among the models. Then we used R cmdscale function to reduce this matrix to two dimensions (link to the R script: https://github.com/aakanksha12/Multidimensional-Scaling/).

To test whether the different rate matrix parameter estimates might differ due to sampling variance alone we randomly sampled sites ranging from 500 – 55000 from the original concatenated dataset and estimated the GTR exchangability parameters on these samples. We then calculated the Euclidean distance between the GTR exchangability parameters estimated using the complete concatenated dataset (the grand GTR parameters) and those estimated using the randomly sampled sites. These distances were then compared to the distance between GTR parameters estimated using each structural class.

2.5. Analyses using site-heterogeneous models

To investigate the capability of site-heterogeneous models to explain the differences between exposed and buried residues, we used the ML based version of CAT [62] model integrated into IQ-TREE v. 1.5.3 [63] for various profile mixture classes (C10 – C60), which we ran with exposed and buried alignments using the I+G4+FO options in the program. Nodal support was assessed using ultrafast bootstrap [64] with 1000 replicates (-bb 1000).

2.5. Compositional heterogeneity and data recoding

To examine differences among taxa in their overall amino acid composition we focused on two groups of amino acids: those encoded by GC-rich codons (G, A, R, and P) and those encoded by AT-rich codons (F, Y, M, I, N, and K) [65]. Briefly, we excluded parsimony uninformative sites, counted the numbers of amino acids in each group, and calculated the GARP/FYMINK ratio for parsimony informative sites. To complement our analyses of the GARP/FYMINK ratio we also back translated sequences and examined variation along the three axes that describe nucleotide composition (i.e., the strong-weak (GC-AT) axis, the amino-keto (AC-GT), and the purine-pyrimidine (AG-CT) axis) for the parsimony informative first and second codon positions.

We also recoded the amino acid data matrix in two ways: 1) six-state Dayhoff coding [66,67] and 2) RY (binary) coding [68]. The six-state recoding of amino acids assigns the following codes to the twenty amino acids: 0 = C (cysteine); 1 = A, G, P, S, and T (small residues); 2 = N, D, E, and Q (acidic and amide residues); 3 = R, H, and K (basic residues); 4 = I, L, M, and V (large aliphatic residue); and 5 = F, W, and Y (aromatic residues). The six-state data matrix was analyzed using the MULTIGAMMA model in RAxML, with the other settings described above. Binary coding involved back translating the amino acids but coding purines as 0 and pyrimidines as 1 (nucleotides that were ambiguous at this level were coded as '?'). The binary data matrix was analyzed in RAxML using the BINGAMMA after excluding completely ambiguous columns. This operation is identical to RY coding of nucleotide matrices. We used a perl program to recode the amino acid data matrix (link to program: https://github.com/aakanksha12/recodeAA).

3. Results

3.1 Sites in distinct structural environments have different signals

Different structural classes were associated with different phylogenetic signals based on analyses using standard empirical models. The most obvious difference was evident when we divided the FRG dataset using relative solvent accessibility. Analyses of solvent exposed sites placed ctenophores sister (topology T2 Figure 1) to all other metazoa, whereas analyses of the buried sites recovered a sponge-ctenophore clade (topology T3, Figure1, Figure2). In all cases when compared with standard models, the best-fitting standard model was LG [51], but our results were robust to the use of different models for ML analyses (Table 1). Likewise, the results did not change even when using the 20-state general time reversible (GTR) model, which had a better fit to the structurally-defined subsets of the FRG data than the LG model despite the large number of free-parameters that must be optimized for that model.

Further subdivision of the FRG data based on other secondary structure information (helix, sheet, and coil) generally had less impact on topology, although analyses of buried sheet residues placed ctenophores sister to all other metazoans (unlike analyses of the other buried residues) (Figure 3). However, all analyses of sheet residues, even when those sites were divided into solvent exposed versus buried sites, resulted in an unexpected clade comprising sponges and placozoa (Supplementary File S5). Subdividing the helix and coil sites into exposed versus buried subsets still revealed the different signals evident when the exposed and buried sites were defined globally (Supplementary File S5). Overall, these results indicate that the data subsets with the largest difference in signal for the base of metazoans are the exposed versus buried sites, and the signals in the FRG dataset support either tree T2 or T3 (Figure 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2019 doi:10.20944/preprints201910.0302.v1

Peer-reviewed version available at Biology 2020, 9, 64; doi:10.3390/biology9040064

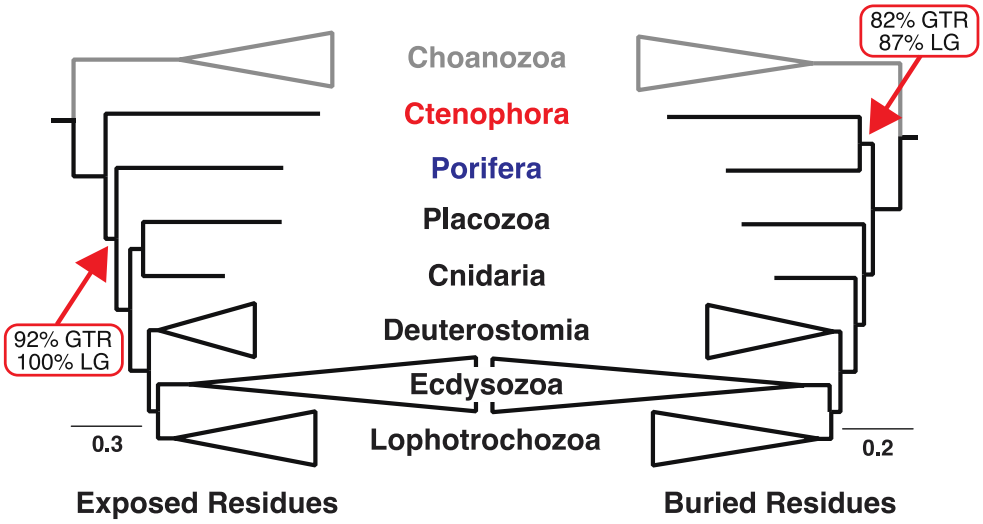

Figure 2. Analyses of sites from different structural environments reveal conflicting phylogenetic signals. We show simplified RAxML trees with both trees are limited to the metazoan ingroups and the choanozoan outgroup (i.e., only Apoikozoa *sensu* [69] are shown). The position of the root drawn in these trees was established by the outgroup taxa (the holozoans *Capsaspora* and *Sphaeroforma* and the fungi *Saccharomyces* and *Spizellomyces*). Bootstrap support for the positions of sponges and ctenophores given the GTR and LG models is indicated next to the arrow.

Table 1. Log likelihood and AIC_c scores for standard empirical models and GTR model optimized for exposed and buried site classes. The GTR model had the best fit (note bold value for AIC_c)

Structural Subset	Model	T2	T3	Cni+Bil[a]	Cni+Pla[b]	ln*L*	AIC_c
Exposed	GTR	92	-	-	100	-1212232.985	**2424956.474**
	LG	100	-	-	100	-1222489.017	2445088.162
	WAG	87	-	-	100	-1225898.553	2451907.235
	VT	96	-	-	98	-1225672.633	2451455.395
	rtREV	90	-	-	98	-1222733.929	2445577.986
	JTTDCMUT	94	-	-	98	-1229028.451	2458167.031
Buried	GTR	-	82	61	-	-1045694.924	**2091880.072**
	LG	-	87	62	-	-1050022.577	2100155.267
	WAG	-	85	58	-	-1054829.256	2109768.626
	VT	-	81	61	-	-1059148.671	2118407.455
	rtREV	-	89	52	-	-1054432.123	2108974.359
	JTTDCMUT	-	81	-	54	-1061103.295	2122316.704

[a] Support for cnidaria + bilateria clade

[b] Support for cnidaria + placozoa clade

3.2 Decisive sites reveal conflicts within each structural class

Strong phylogenetic signals are often limited to a subset of genes [59,70,71] or even to specific sites within genes [58,72]. Examining the sites with the strongest phylogenetic signals can provide a way to determine the distribution and amount of conflict within the data subsets. We examined the number of "decisive sites" (sites that strongly support either of two trees in Figure 2) in the exposed and buried classes, and found significant differences ($P < 0.02$, Fisher's exact test) in the numbers of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2019 doi:10.20944/preprints201910.0302.v1

Peer-reviewed version available at Biology 2020, 9, 64; doi:10.3390/biology9040064

decisive sites that correspond to solvent exposed sites versus those associated with buried residues (Table 2). We also examined the concentration of decisive sites within different genes in exposed and buried residues and found that they were uniformly distributed among the genes. These results indicate that differences in signal in different parts of the FRG dataset that can be detected in comparisons of solvent exposed versus buried sites does not reflect an unusual concentration of decisive sites in any particular gene within the FRG dataset; instead, the contrasting signals appear to be a more universal feature of analyses focused on sites in the two different structural environments.

Table 2. Decisive sites favoring topology T2 vs. T3 in the exposed and buried residues[a].

	Ctenophore sister (T2)	Ctenophore + Porifera (T3)
Exposed	172	150
Buried	167	205

[a] T2 and T3 refer to the arrangement of ctenophores, sponges, and remaining metazoa (Figure 1); other relationships were held constant.

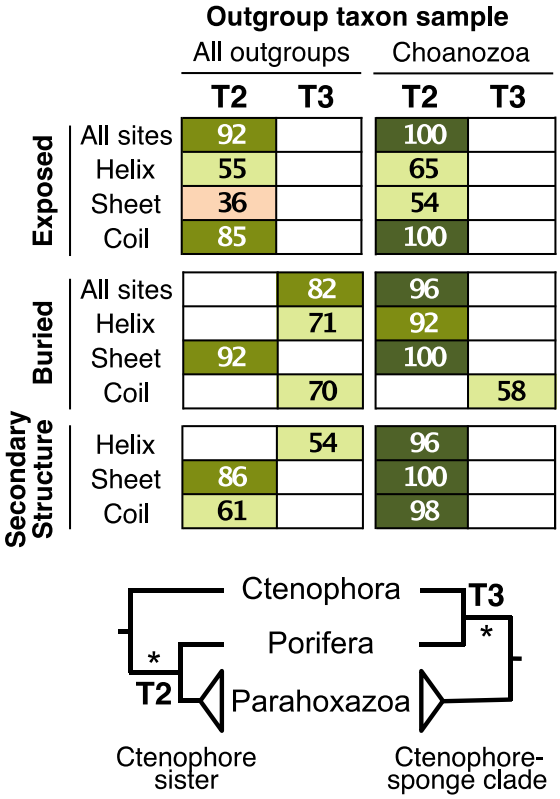

Figure 3. Heat map showing support for tree topologies obtained using various structural classes and taxon samples. In the online version colors indicate bootstrap support values (Dark green > 95, lighter Dark green >75, Yellow > 50 and Pink < 50; No color: Tree was not recovered in the indicated analysis)

3.3. Reduced outgroup sampling also reduces the differences in signal

Highly divergent outgroup taxa are known to affect phylogenetic inference. To examine the impact of outgroup taxon sampling on the signal associated with various structural environments we removed four divergent outgroups [two fungi (*Saccharomyces* and *Spizellomyces*) and two taxa from the classes Mesomycetozoea (*Sphaeroforma*) and Filasterea (*Capsaspora*)] from the FRG dataset. This limits the outgroups to Choanozoa, the sister group of metazoa [69,73,74]. Analyses of exposed and buried residues using the FRG dataset with the reduced taxon sample largely converged on a single basal topology (ctenophores sister to all other metazoa) regardless of whether analyses used GTR with parameters optimized for each subset (Figure 3 and Supplementary File S6) or standard empirical models (Table 1). However, the position of cnidaria and *Trichloplax* varied in analyses using the reduced taxon sample (Table 1), underlining both the complexity of signal in the FRG dataset and the differences among the structural environments.

3.4. Sites in different structural environments exhibit distinct patterns of sequence evolution

In standard models, like the LG, WAG, and Dayhoff models, and GTR optimized for a specific large-scale protein alignment, the rate matrices reflect the patterns of sequence evolution across all structural classes. Thus, the values in the empirical rate matrix might not reflect estimates from structurally divided data. We next tested if patterns of sequence evolution in each structural environment differed. When we examined differences among models using multidimensional scaling, the strongest separation among models was related to the best models for the two solvent accessibility classes (Figure 4). Standard models formed a cluster closer to models estimated using buried residues (Figure 4 and Supplementary Figure S1).

The various structural subsets have different numbers of sites, and the GTR model for amino acids has a large number of free parameters, raising the question of whether the observed differences in exchange rate parameter estimates simply reflect sampling error. Yet the models appear to cluster in a manner that is correlated with structural class in multidimensional scaling space (e.g., note that buried sites and solvent exposed sites cluster in Figure 4 even when they are further subdivided into independent sets of helical, sheet, and coil sites), suggesting that sampling error is unlikely to explain the observed difference. Nevertheless, we also sampled sites from the FRG dataset randomly (i.e., without respect to structure) to generate datasets comparable in size to the structurally defined subsets and estimated model parameters on these random samples. Since the sites were sampled randomly, estimates of GTR model parameters should converge on the values estimated from the dataset as a whole, assuming that the number of sites that were sampled is sufficient to overcome sampling variance. We assessed the distance between the GTR model parameter estimates for the complete FRG dataset to those of randomly selected sites. We found that this distance rapidly decreased as the number of sites in the random sample increased (Supplementary Figure S2); the distances between the "global" model based on the complete FRG dataset and the estimates based on each structural subset are much greater than the distances expected based on sampling variance. This demonstrates that the differences in parameter estimates of structural classes differ much more than expected based on sampling variance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2019 doi:10.20944/preprints201910.0302.v1

Peer-reviewed version available at Biology 2020, 9, 64; doi:10.3390/biology9040064

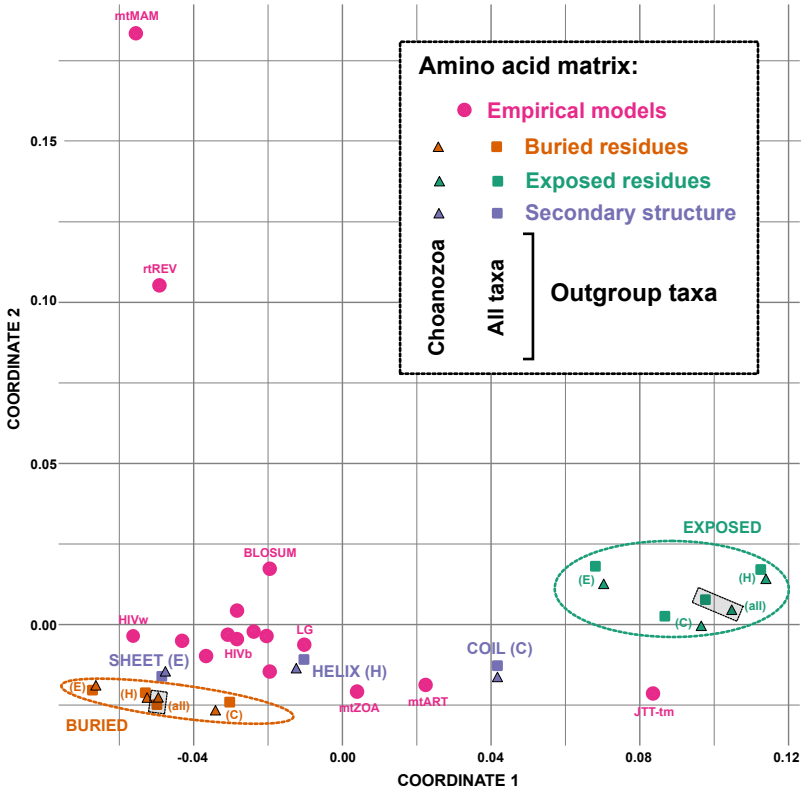

Figure 4. Multidimensional scaling plot showing the Euclidean distances between various amino acids exchange rate matrices. Different colors indicate different categories of the matrices in the online version (Green: exposed residues, Orange: buried residues, Purple: secondary structure, and Pink: standard empirical models).

3.5. Site-heterogeneous profile mixture models can yield surprising topological changes

Site heterogeneous models, like the CAT model [15], represent another way to accommodate heterogeneity in the evolutionary process. CAT-type models assume aligned sites are drawn from a mixture with multiple distinct evolutionary processes (substitutional profiles) that differ in the amino acid equilibrium frequencies of the 20 amino acids. To complement the ML analyses using empirical models and the GTR model, we analyzed the exposed and buried classes using different profile mixture classes (C10-C60). We analyzed the full taxon set as well as reduced taxon set using the six ML variants of the CAT model implemented in IQ-TREE [63] . If variation among sites in their propensity to accept specific amino acids is necessary to obtain accurate estimates of phylogeny, then trees based on the two structural subsets (exposed versus buried) are expected to converge on the same topology.

Analyses of the exposed and buried sites from the full taxon set using the CAT model with various number of profile mixtures converged on a single tree T3, with three exceptions (C10, C30 and C50 with exposed residues), but among C10, C30 and C50 two of these analyses converged on a tree with a sponge-ctenophore clade (Table 3). In contrast, analyses of the reduced taxon sample using CAT models resulted in different topologies for exposed and buried classes (T2 and T3 respectively; Table 3). In two of the three cases where analyses of exposed sites using the CAT model and all outgroup resulted in T2 (C30 and C50) we also recovered a surprising clade uniting *Trichloplax* and sponges (Table 3). The sponge + *Trichloplax* clade was only recovered in analyses that included all outgroups. However, it is important to recognize that analyses that used CAT models sometimes resulted in this surprising clade and the CAT model did not provide clear evidence for increased congruence between the exposed and buried sites.

Peer-reviewed version available at *Biology* **2020**, *9*, 64; doi:10.3390/biology9040064

Table 3. Log likelihood and AIC$_c$ scores for various CAT model profile mixtures as implemented in IQ-TREE. The best-fit model is indicated with a bold AIC$_c$ value.

CAT model	Outgroup	Dataset	T2	T3	Cni+Bil	Cni+Pla	Pori+Pla	ln*L*	AIC$_c$
C10	All outgroups (RG)	Exposed	100			100		-1225103.471	2450339.126
C20		Exposed		63		100		-1216444.36	2433040.964
C30		Exposed	100				72	-1214319.593	2428811.499
C40		Exposed		100		96		-1212974.829	2426142.047
C50		Exposed	100				73	-1212360.091	2424932.657
C60		Exposed		100		93		-1211470.44	**2423173.448**
C10		Buried		83	78			-1053873.212	2107878.588
C20		Buried		97	80			-1048007.721	2096167.66
C30		Buried		94	73			-1046099.005	2092370.288
C40		Buried		92	83			-1044155.469	2088503.283
C50		Buried		93	86			-1043172.44	2086557.3
C60		Buried		93	89			-1042207.261	**2084647.025**
C10	Choanozoa only	Exposed	100			100		-964491.9951	1929100.133
C20		Exposed	98			74		-957706.2984	1915548.793
C30		Exposed	96			69		-956161.8764	1912480.01
C40		Exposed	95			71		-955297.444	1910771.215
C50		Exposed	92			51		-953519.0992	1907234.604
C60		Exposed	90		100			-952582.9203	**1905382.332**
C10		Buried		50	46			-837871.4589	1675859.045
C20		Buried		37	34			-833463.787	1667063.748
C30		Buried		75	63			-832000.2662	1664156.761
C40		Buried		66	63			-830714.1456	1661604.582
C50		Buried		45	45			-829970.749	1660137.858
C60		Buried		74	79			-829121.6382	**1658459.713**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2019 doi:10.20944/preprints201910.0302.v1

Peer-reviewed version available at Biology 2020, 9, 64; doi:10.3390/biology9040064

Analyses of buried residues using the CAT model resulted in trees with a non-monophyletic Deuterostomia; all profile mixtures placed the sea urchin *Strongylocentrotus purpuratus* sister to a clade comprising chordates along with all protostomes in the taxon sample, in contrast to results obtained using the GTR model (Figure 5). To test whether any of the observed topologies reflected differences between model or the programs IQ-TREE and RAxML (i.e., to test whether our results reflected differences between the search or numerical optimization routines in those programs rather than the use of the CAT model) we confirmed that analyses using the site-homogeneous GTR model in IQ-TREE yielded the same results as analyses in RAxML (Figure 5, Supplementary File S7). Although the most complex CAT model did have the best fit to the data based on the AIC_c (Table 3), the other results that emerged from the use of the CAT models could indicate over fitting of the model. Specifically, the CAT models with a larger number of profile mixtures (C40, C50 and C60) also had zero or near zero weights for the mixture components (Supplementary File S7). Regardless of the details, it seems clear that the unexpected signal in the buried residues (deuterostome non-monophyly) revealed by CAT models is very likely to be non-historical.

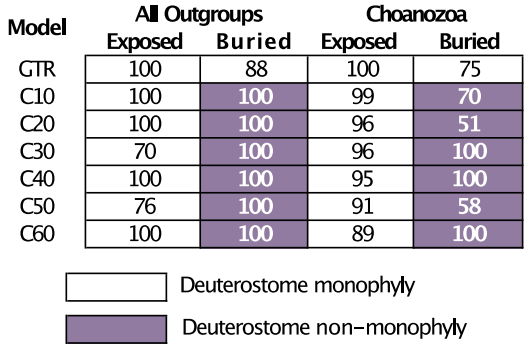

Figure 5. Heat map showing support for deuterostome monophyly for exposed and buried residues using GTR and CAT models. Colors indicate presence or absence of deuterostome monophyly (No color: present, Purple: absent).

3.6. Binary recoding eliminates the observed differences in signal

The GARP/FYMINK ratio, a metric of amino acid composition that correlates with genomic GC-content, differed among taxa for both exposed and buried sites and the mean GARP/FYMINK ratios for the two structural classes also differed (Figure 6a). However, the pattern of variation among taxa in the ratio relative to the means was virtually identical for both structural environments. Specifically, the GARP/FYMINK ratio was lower than the mean for non-bilaterian animals and higher than the mean in most non-metazoan outgroup taxa. Despite the evidence for variation among taxa there was no evidence for large-scale convergence between either ctenophores or sponges and the outgroups (Figure 6a). Although a simple pattern of convergence was not evident, amino acid compositional variation does violate the assumptions of time reversible models and this could have an impact on phylogenetic inference. Thus, limiting compositional variation has the potential to improve our estimates of phylogeny. We examined variation in nucleotide composition among taxa along the three possible axes of nucleotide composition (Figure 6b). Since the largest degree of variation was on the strong-weak axis we reasoned that RY coding might reduce the impact of compositional variation.

Phylogenetic analyses of exposed and buried sites after binary coding resulted in identical trees that placed ctenophores sister to all other animals with strong (≥88%) bootstrap support (Figure 6c). Analyses of the exposed sites after six-state Dayhoff recoding resulted in the same topology, also with high support for placing ctenophores sister to other metazoa (Figure 6c). Similarly, the analyses of the buried sites after six-state coding also resulted in a tree with ctenophores sister to all other animals, albeit with limited (48%) bootstrap support (Supplementary File S8). The buried Dayhoff recoded tree also had a sponge-placozoan clade, although that unexpected clade had limited (28%)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2019 doi:10.20944/preprints201910.0302.v1

Peer-reviewed version available at Biology 2020, 9, 64; doi:10.3390/biology9040064

bootstrap support. Regardless of the details, apply either recoding method to the buried sites results in congruence with the exposed sites and both subsets of the data support the T2 topology.

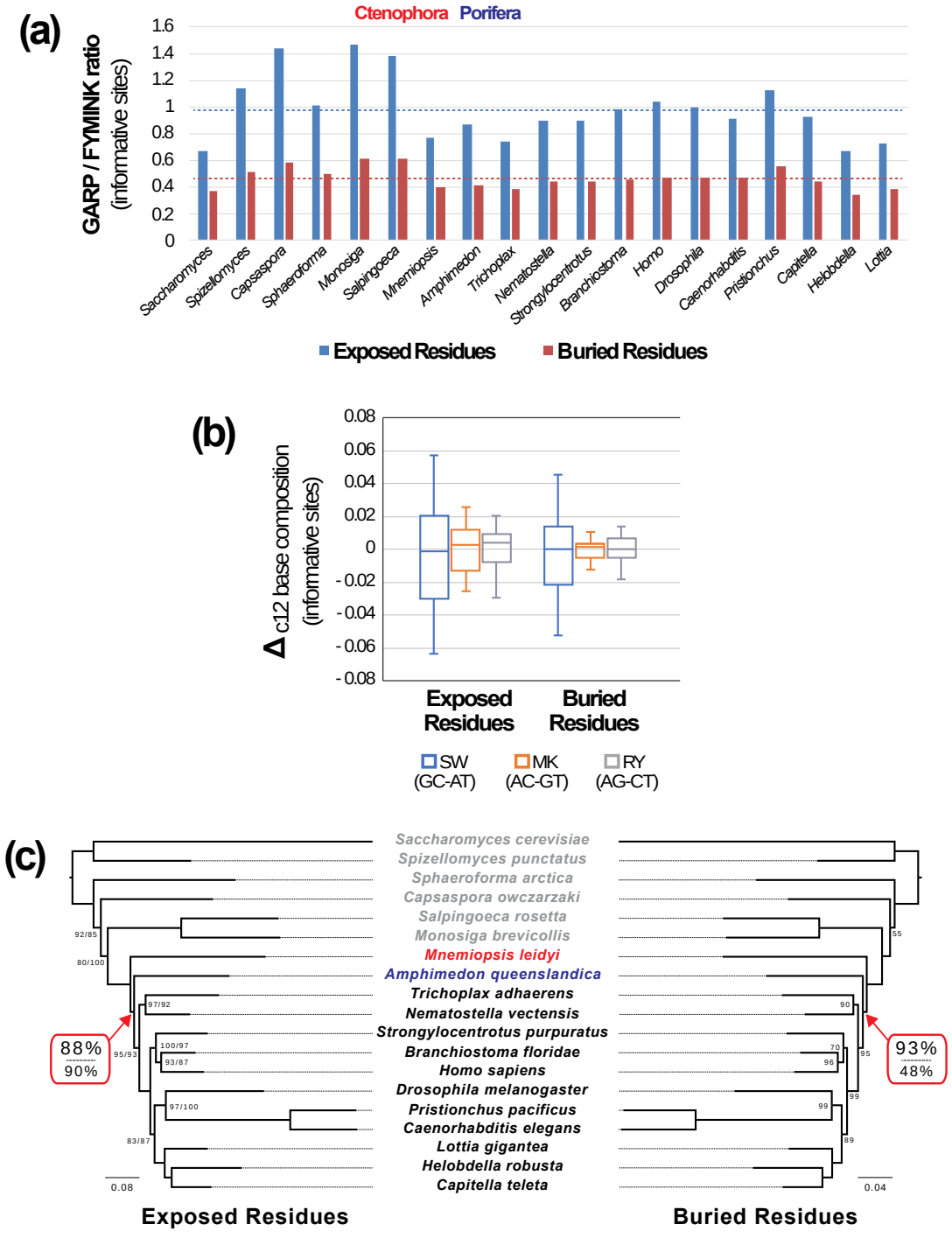

Figure 6. Compositional variation and the impact of recoding on tree estimation. (**a**) Variation across taxa in the ratio of amino acids encoded by GC-rich codons (G, A, R, and P) to those encoded by AT-rich codons (F, Y, M, I, N, and K). To limit the impact of invariant sites we only considered parsimony informative sites. (**b**) Variation in base composition for parsimony informative first and second codon positions after back translation. (**c**) The results of tree searches after recoding as binary (purine-pyrimidine (RY)) characters. The tree topologies were identical, although the tree lengths did differ (note scale bars). Support for the node that defines T2 (i.e., the node that places Ctenophora sister to other Metazoa) is emphasized using a red box; support given the binary data is presented to the top and the value for six-state Dayhoff recoding is presented below. For other nodes, bootstrap values <100% are presented, with values for six-state recoding presented to the right. Since there

were some topological conflicts between trees obtained by binary and six-state coding of buried site the bootstrap support for six-state coding is not presented on that tree (except for the focal node).

4. Discussion

Analysis of the FRG dataset with various structural classes resulted in different tree topologies (e.g., Figure 2), and there were significant differences in the numbers of decisive sites favoring distinct placements of the metazoan root (Table 2). These results strongly indicate that conflicting phylogenetic signals are non-randomly distributed in parts of the dataset that can be defined using protein structure. The difference in signal was most apparent when dividing protein alignments into subsets based on relative solvent accessibility (exposed sites versus buried sites). The different signals in exposed and buried sites were evident regardless of whether the phylogenetic inference was conducted using standard empirical models or with GTR model parameters optimized on each structural subset. Other subdivision strategies (i.e., dividing aligned sites based on secondary structure of a combination of secondary structure and relative solvent accessibility) revealed some differences in signal, but they were not as strong as the differences between solvent exposed versus buried sites. Amino acid recoding, especially binary coding, reduced the differences in the signal for exposed and buried sites.

4.1. Different models for different structural environments

The obvious explanation for the conflicting signals in each structural class, especially the strong difference between the exposed and buried signals, is that the sites in each of these classes exhibit different patterns of sequence evolution. The poor fit of standard empirical models to the data could then result in an incorrect inference. This could result in the observed difference in signal, with one structural subset yielding the correct tree while analyses of the other one result an incorrect estimate of phylogeny. Alternatively, it is possible that analyses of both exposed and buried sites yield different topologies, both of which are inaccurate. GTR model parameter estimates for each class certainly indicate that the patterns of sequence evolution are different in each structural environment (Figure 4 and Supplementary Figure S1). However, conducting analyses using GTR model parameters optimized on each class did not cause analyses in the two classes to converge on a single topology (in fact, the topologies were unchanged relative to those that emerged in the analyses that used standard empirical models). Finally, we note one intriguing aspect of our model comparisons: the rate matrix parameters for most empirical models lie closer to the models based on buried sites (this was true for all empirical models trained on diverse datasets; i.e., all empirical models except those trained using the more limited organellar or viral datasets).

The basis for the differences we observed among the structural classes in their patterns of evolution almost certainly reflects differences in the nature of the purifying selection on sites in different structural environments. It is well known that protein evolution is heterogeneous and depends on the site-specific biochemical constraints like structure, dynamics, and biochemical functions[75].There is substantial evidence that, depending on their position and role in the overall conformation and function sites will only accept a subset of the 20 amino acids, with all other possibilities being selected against [76]. For example, in the case of relative solvent accessibility the sites that are exposed to an aqueous environment will include more polar residues whereas buried sites which form the cores of proteins and are inaccessible to solvent, will include more non-polar residues and be more resistant to changes in side chain volume [77]. Similar differences among the major secondary structural classes have also been appreciated for some time [16–18], although the differences among the secondary structure classes does not appear to be as extreme as the differences based on solvent accessibility. Overall, these differences in evolutionary patterns among the structural classes can lead to different signals in each structural class.

Models of sequence evolution that incorporate protein structure have also been proposed in the context of phylogenetics. For example, Goldman et al. [16] used a hidden Markov model approach to assign distinct rate matrices to sites based on in secondary structure and solvent accessibility (where secondary structure and solvent accessibility are the hidden states). Le and Gascuel [18] developed a

mixture model that uses available protein structure annotations. Both of these studies revealed that efforts to acknowledge the different structural classes for sites in protein multiple sequence alignments can have a major impact on model fit, measured using the improvement in log likelihood values. However, both of those methods estimate a tree topology for the complete alignment; our approach of dividing proteins into structural classes follows the same basic idea but shifts the focus to study the phylogenetic signals in these distinct structural classes separately.

The primary focus of this study was testing whether different signals emerge in analyses of sites in distinct structural environments rather than phylogenetic inference *per se*. This led us to analyze structurally-defined subsets of the data. In principle, analyzing all sites in a sequence alignment using a model that appropriately accommodates heterogeneity should yield the best estimate of phylogeny. However, that approach is not ideal for examining conflicting signals or asking whether the signals are non-randomly distributed. Observing different topologies when data are analyzed using models that are structure-aware vs not structure-aware could result from the existence of distinct signals for sites in different structural environments. Alternatively, it could reflect other aspects of the models used for analyses. In contrast, it is straightforward to determine whether different analyses actually increase (or decrease) the congruence among the data subsets when analyses are conducted after dividing the data into subsets based on protein structure.

4.2. Site-heterogeneous models do not increase congruence in signal for different structural classes

Site-heterogeneous models like CAT has been used in many studies focused on the deep branches in the metazoan tree (e.g., [37,38,78,79]) and it has been asserted that the CAT model is more realistic than empirical models or the GTR model [15,39,80–83]. Analyses using CAT models have been suggested to be less prone to systematic error, such as long-branch attraction, than site-homogeneous models like the standard empirical models when they are applied to heterogeneous data [83–86]. However, structure-aware models of protein evolution [16,17,87] also introduce among-sites heterogeneity. Thus, it is not absolutely clear that the CAT model represents the best way to introduce heterogeneity into phylogenetic analyses. Ultimately, all models are approximations; whether a specific approximating model (e.g. the CAT model) reveals the true historical signal in a specific dataset better than another approximating model (e.g., the GTR model) is an empirical question.

Although the parameter-rich CAT models did have a better fit to the FRG data than the GTR model based on the AIC_c (Table 3) we made several observations that analyses using the CAT model did not behave in an ideal manner. We have two expectations if the CAT model captured the heterogeneity within each subset of the data in a way that improved the results of phylogenetic analyses. First, the trees for the exposed and buried subsets of the data would converge on the same topology. Second, the topology to be T1 or T2, given the arguments that T3 is the least plausible topology. However, our analysis using CAT models with various profile mixtures for exposed and buried residues either shifted back and forth between T2 and T3 or converged towards less plausible T3 topology. We believe that this indicates that CAT models were unable to capture the heterogeneity in these subsets in an appropriate manner. Furthermore, an unexpected clade uniting chordates and protostomes (i.e., the branch that implies deuterostome non-monophyly) also emerged in analysis using buried sites; this unexpected clade has also emerged in analyses of some other datasets using the CAT model [88]. We believe the observed support for the unexpected chordate-protostome clade provides further evidence that CAT models are inappropriate for analyses of these data. This may reflect overestimation of the number substitutional categories appropriate for larger datasets; the zero or near zero estimated weights for specific substitutional profiles provides another line of evidence for this idea. Regardless, our results certainly indicate that CAT models do not represent a panacea for phylogenetic analyses of proteins.

4.3. Amino acid recoding increases congruence in signal for different structural classes

In addition to long-branch attraction, variation in amino acid (or base) composition across the tree is another source of systematic error in phylogenetic analyses [89,90]. However, the impact of variation in base composition on phylogenetic estimation is complex and difficult to predict. This complexity reflects the large number of ways that the proportions of the 20 amino acids can vary. We focused on the GARP/FYMINK ratio (Figure 6a), which correlated with genomic GC-content, because many other studies have found variation along the GC-AT axis [65,91,92]. Indeed, even in studies that have found significant variation along other axes in protein composition space (e.g., in a study focused on prokaryotes [93] an axis of variation correlated with optimal growth temperature and IVYWREL-content was evident) there appears to be much stronger variation on the GC-AT axis (in the Boussau et al. [93] study the GC-AT axis explained 45.4% of the variance among taxa whereas the IVYWREL-axis only explained 13.8%). Back translation of the amino acid sequences confirmed that there was more variation along the GC-AT axis than along the other two axes in nucleotide composition space (AC-GT and AG-CT; Figure 6b). The mean GARP/FYMINK ratios for exposed and buried sites were different, although taxa with an above average GARP/FYMINK ratio in one structural environment also had a higher ratio in the other environment (and vice versa). Thus, the relative pattern of variation in the GARP/FYMINK ratio among taxa did not differ between the structural environments. This makes it unlikely for variation among taxa in amino acid composition to represent the primary explanation for the observed differences in signal for the exposed and buried sites.

Despite the observed similarities in their patterns of compositional variation for exposed and buried sites, conducting analyses after recoding amino acids did increase congruence between exposed and buried sites (Figure 6c). In fact, analyses of exposed and buried data matrices generated by back translation with recoding the nucleotides as binary characters resulted in identical trees with relatively high (≥88%) bootstrap support for T2. We also conducted analyses after six-state recoding using the Dayhoff categories, a commonly used method for amino acid recoding (reviewed by Hernandez and Ryan [94]). Analyses of exposed sites after six-state recoding resulted in a topology identical to the tree generated by binary coding; support for T2 was also relatively high (90%). In contrast, analyses of buried sites after six-state recoding resulted in a tree with limited (<50%) bootstrap support for T2 and several other topological differences from the tree based on exposed sites after recoding. However, many branches in the six-state recoded buried site tree had limited support (Supplementary File S8). Regardless, it was clear that binary coding resulted in completely congruent trees for exposed and buried sites and six-state recoding increased only congruence (albeit with high support only in the analysis of exposed sites).

The increased congruence observed when the data were recoded raises the question of whether the apparent improvements provide evidence that compositional variation explains the different signals evident in exposed and buried sites. The observation that the patterns of variation across taxa in the GARP/FYMINK ratio are quite similar for exposed and buried sites makes it unlikely that the differences in signal reflects a simple case of convergence in GC-content. However, model violations due to changes in amino acid composition might have indirect effects on phylogenetic estimation, perhaps exacerbating other sources of bias (e.g., long-branch attraction). Amino acid recoding could also improve model fit in other ways; regardless of the details, it is clear that amino acid recoding improves the congruence between estimates of phylogeny based on exposed vs. buried sites.

One relatively straightforward way that amino acid recoding might have as positive impact on phylogenetic estimation is by reducing the number of substitutions; this could reduce the potential for phenomena like long-branch attraction. However, reducing the number of substitutions also eliminates phylogenetic information. Indeed, Hernandez and Ryan [94] questioned the utility of six-state recoding, showing that the loss of phylogenetic information outweighs its benefits with respect to compositional heterogeneity. Our observation that analyses of the exposed sites results in relatively strong support after six-state recoding whereas analyses of the buried sites after six-state recoding results in very limited support is consistent with the hypothesis that six-state recoding results in substantial loss of information. This information loss is expected to be less problematic when the substitution rate is high; the higher rate of amino acid substitution in the solvent exposed

environment (note scale bars in Figures 2 and 6c) is likely to make six-state recoding less problematic for the exposed sites. In contrast, binary coding preserves more information because most amino acids are recoded as two or three binary characters; we believe that binary coding of amino acids deserves more consideration in future studies.

4.4. Phylogenetic implications

The primary goal for this study was to ask whether the signal revealed by phylogenetic analyses of a large-scale protein dataset was non-randomly distributed with respect to protein structure. We chose the RG dataset for several reasons, one of which was the fact that basal metazoan relationships had limited support when it was analyzed [6]. This suggested either that conflicting signals were present in the data or that there was very little signal (historical or non-historical) in the data. Our analyses revealed: 1) that conflicting signals were present in the data; and 2) that distinct signals were non-randomly distributed with respect to protein structure. However, those analyses did not establish which of those signals were historical in nature. This raises another fundamental question: what do our analyses reveal about the relationships among deep-branching metazoans?

One important feature of historical signal is that it is expected to be distributed broadly in the genome. Although there are certainly cases where true reticulations are present in the history of life [95,96], many relationships appear to reflect a "tree-like" history combined with discordance among individual gene trees superimposed on that tree for a variety of reasons (e.g., horizontal transfer and incomplete lineage sorting [25]). The structurally defined subsets of data that we examined represent different parts of the same 231 orthologous proteins, although the exact number of gene trees they represent is unclear because intra-locus recombination could lead to a case where one protein may be associated with multiple gene trees [97]. Regardless of the exact number of gene trees in our dataset, it is important to recognize that the exposed and buried sites are found throughout protein sequences, sometimes quite close to each other. Therefore, exposed and buried sites are encoded by sequences that are likely to represent virtually identical sets of gene trees. Thus, the historical signal present in both subsets of the data should reflect similar sets of gene trees so analyses of either subset of the data would be expected to converge on similar topologies (assuming the number of sites analyzed is sufficiently large). This was not what we observed for the majority of the analyses we conducted, but it was true for the analyses that used recoded amino acids. Those analyses converged on T2 (ctenophores sister to all other animals).

The observation that analyses of two non-overlapping phylogenomic datasets result in the same topology is not, in and of itself, sufficient to conclude that the tree topology in question (i.e., T2) is an accurate representation of evolutionary history. However, postulating that ctenophores are sister to all other animal does represent the simplest interpretation of the analyses we conducted; one must explain three different observations if one chooses to postulate that other topologies are the best representation of evolutionary history. First, we observed that most phylogenetic analyses of exposed and buried sites in the FRG dataset yield distinct topologies (specifically, T2 and T3; e.g., Figure 2). However, of the three plausible topologies, T3 is the least plausible because it fails to explain the distribution of character states highlighted by Nielsen [35] any better than T2 but it has not been recovered in other phylogenomic studies. Second, our analyses revealed that the potential for the best-characterized sources of bias in phylogenetic analyses (long-branch attraction and variation in amino acid composition) to have an impact on phylogenetic analyses of exposed vs. buried sites was quite similar (i.e., the relative branch lengths are similar for both structural environments [Figure 1 and Figure 6c] and the amount of variation in for variation in amino acid composition is similar for both [Figure 6a]). Finally, we found that analyses of exposed and buried sites conducted after amino acid recoding converged on a single topology (T2).

It remains possible to postulate that T1 (sponges sister to all other animals) is the correct tree despite the three observations described above. However, postulating that T1 is correct necessitates several assumptions regarding the behavior of phylogenetic analyses of the FRG dataset. First, one must postulate that most analyses of the exposed sites in the FRG dataset are biased toward T2. The

exception to that bias toward T2 would be the subset of the analyses using CAT models with 20, 40, and 60 categories; one must postulate that those analyses are biased toward T3 (Table 3). Second, one must assume that analyses of the buried sites are biased toward T3. Finally, one must assume analyses of the exposed and buried subsets of the data conducted after amino acid recoding are both biased toward T2. Making those assumptions represents a much more elaborate hypothesis than simply hypothesizing that T2 is correct and the only bias is the support for T3 observed in analyses of the buried sites. For that reason, we view our results as additional evidence for the hypothesis that the root of the animal tree lies between ctenophores and all other metazoa.

5. Conclusions

Our signal exploration corroborated the hypothesis that different phylogenetic signals are associated with specific parts of proteins that can be defined using non-phylogenetic criteria (in this case, structural criteria). Most analyses of data from two structural environments resulted in conflicting trees that each had relatively high support. Our analyses also suggest the use of site-homogeneous models (like GTR or empirical models such as LG) should not necessarily be eschewed in favor of the CAT model, despite the observation that the latter models typically have a better fit to the data based on commonly used model selection criteria. This statement may appear to be problematic since it implicitly calls the use of standard model selection criteria into question, but we emphasize that it is actually very difficult to assess the fit of phylogenetic models to empirical dataset in absolute terms [98]. Even the most complex models currently available for phylogenetic analyses (including the site-heterogeneous CAT model), may be quite far from the true underlying processes of molecular evolution and, therefore, every bit as subject to systematic error as simpler models. Sanderson and Kim [99] worried that the very large AIC increases in response to the addition of relatively small numbers of free parameters might indicate that all models under consideration are very far from the (unknown) true model. Thus, the most parameter-rich CAT models could be closer to the true model but still deviate from that true model in ways that actually obscure the historical signal. The available site-heterogeneous models might be very useful in some contexts, but they introduce many free parameters that may not be constrained in a biologically realistic manner. Steel [100] worried that very parameter-rich models would be impractical because they would have to have enough parameters "to fit an elephant" (a colorful metaphor for an unrealistic and overly parameter-rich model that has been attributed to John von Neumann [101]). One way to overcome this "elephant factor" is to place biological constraints on parameters; the association between protein structure and phylogenetic signal that we observed suggests that structure could provide information about those constraints. Alternatively, it may be easier to identify models that reveal historical signal after simplifying the data in some way (e.g., by recoding amino acids). However, there has been limited exploration of the best way to select phylogenetic models when the data can be encoded in multiple ways. Regardless of the specific details, we believe that the identification of additional phylogenomic datasets where different data types are associated with distinct signals could provide a way explore the behavior of phylogenetic methods and their ability to accurately recover true historical signal.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1

Author Contributions: Conceptualization, A.P. and E.L.B.; methodology, A.P. and E.L.B.; software, A.P.. and E.L.B.; validation, A.P. and E.L.B.; formal analysis, A.P. and E.L.B.; investigation, A.P.; resources, E.L.B.; data curation, A.P.; writing—original draft preparation, A.P. and E.L.B.; writing—review and editing, A.P. and E.L.B.; visualization, A.P. and E.L.B.; supervision, E.L.B.; project administration, E.L.B.

Funding: This research received no external funding.

Acknowledgments: We are grateful to Rebecca Kimball, Gordon Burleigh, Joe Ryan, and Gavin Naylor for helpful comments on this manuscript and encouragement throughout the project.

Conflicts of Interest: Authors declare no conflict of interest.

References

1. Gee, H. Ending incongruence. *Nature* **2003,** 425, 782.
2. Rokas, A.; Williams, B.I.; King, N.; Carroll, S.B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **2003**, *425*, 798–804.
3. Nishihara, H.; Okada, N.; Hasegawa, M.; Rokas, A.; Williams, B.; King, N.; Carroll, S.; Soltis, D.; Albert, V.; Savolainen, V.; et al. Rooting the eutherian tree: The power and pitfalls of phylogenomics. *Genome Biol.* **2007**, *8*, R199.
4. Misof, B.; Liu, S.; Meusemann, K.; Peters, R.S.; Donath, A.; Mayer, C.; Frandsen, P.B.; Ware, J.; Flouri, T.; Beutel, R.G.; et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **2014**, *346*, 763–767.
5. Wickett, N.J.; Mirarab, S.; Nguyen, N.; Warnow, T.; Carpenter, E.; Matasci, N.; Ayyampalayam, S.; Barker, M.S.; Burleigh, J.G.; Gitzendanner, M.A.; et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U. S. A.*, **2014**, *111*, E4859-68.
6. Ryan, J.F.; Pang, K.; Schnitzler, C.E.; Nguyen, A.D.; Moreland, R.T.; Simmons, D.K.; Koch, B.J.; Francis, W.R.; Havlak, P.; Smith, S.A.; et al. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **2013**, *342*, 1242592.
7. Moroz, L.L.; Kocot, K.M.; Citarella, M.R.; Dosung, S.; Norekian, T.P.; Povolotskaya, I.S.; Grigorenko, A.P.; Dailey, C.; Berezikov, E.; Buckley, K.M.; et al. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **2014**, *510*, 109–114.
8. Dunn, C.W.; Giribet, G.; Edgecombe, G.D.; Hejnol, A. Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* **2014**, *45*, 371–395.
9. Feuda, R.; Dohrmann, M.; Pett, W.; Philippe, H.; Rota-Stabelli, O.; Lartillot, N.; Wörheide, G.; Pisani, D. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Curr. Biol.* **2017**, *27*, 3864-3870.
10. King, N.; Rokas, A. Embracing uncertainty in reconstructing early animal evolution. *Curr. Biol.* **2017**, *27*, R1081–R1088.
11. Simion, P.; Philippe, H.; Baurain, D.; Jager, M.; Richter, D.J.; Di Franco, A.; Roure, B.; Satoh, N.; Quéinnec, É.; Ereskovsky, A.; et al. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **2017**, *27*, 958–967.
12. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **1978**, *27*, 401–410.
13. Hendy, M.D.; Penny, D. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **1989**, *38*, 297–309.
14. Phillips, M.J.; Delsuc, F.; Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **2004**, *21*, 1455–1458.
15. Lartillot, N.; Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **2004**, *21*, 1095–1109.
16. Goldman, N.; Thorne, J.L.; Jones, D.T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **1998**, *149*, 445–458.
17. Thorne, J.L.; Goldman, N.; Jones, D.T. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **1996**, *13*, 666–673.
18. Le, S.Q.; Gascuel, O. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst. Biol.* **2010**, *59*, 277–287.

19. Le, S.Q.; Lartillot, N.; Gascuel, O. Phylogenetic mixture models for proteins. *Philos. Trans. Royal Soc. B*, **2008**, *363*, 3965-3976.
20. Blanquart, S.; Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **2006**.
21. Blanquart, S.; Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **2008**, *25*, 842–858.
22. Groussin, M.; Boussau, B.; Gouy, M. A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst. Biol.* **2013**, *62*, 523–538.
23. Whelan, N. V; Halanych, K.M. Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Syst. Biol.* **2016**, *66*, 232–255.
24. Patel, S.; Kimball, R.T.; Braun, E.L. Error in phylogenetic estimation for bushes in the tree of life. *J. Phylogenetics Evol. Biol.* **2013**, *1*, 1:110.
25. Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **1997**, *46*, 523–536.
26. Slowinski, J.B.; Page, R.D.M. How should species phylogenies be inferred from sequence data ? *Syst. Biol.* **2007**, *48*, 814–825.
27. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evolution*, **2009**, 1–19.
28. Reddy, S.; Kimball, R.T.; Pandey, A.; Hosner, P.A.; Braun, M.J.; Hackett, S.J.; Han, K.-L.; Harshman, J.; Huddleston, C.J.; Kingston, S.; et al. Why do phylogenomic data sets yield conflicting trees? data type influences the avian tree of life more than taxon sampling. *Syst. Biol.* **2017**, *51*, 588–598.
29. Braun, E.L.; Cracraft, J.; Houde, P. Resolving the avian tree of life from top to bottom: The promise and potential boundaries of the phylogenomic era. In *Avian Genomics in Ecology and Evolution—From the Lab into the Wild*; Kraus, R.H.S., Ed.; Springer: Cham, Switzerland, **2019**, 151-210.
30. Raymann, K.; Brochier-Armanet, C.; Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.*, **2015**, *112*, 6670–6675.
31. He, D.; Fiz-Palacios, O.; Fu, C.J.; Tsai, C.C.; Baldauf, S.L. An alternative root for the eukaryote tree of life. *Curr. Biol.* **2014**, *24*, 465–470.
32. Lesk, A.M.; Chothia, C.H. The response of protein structures to amino-acid sequence changes. *Philos. Trans. Royal Soc.*, **1986**, *317*, 345–356.
33. Illergård, K.; Ardell, D.H.; Elofsson, A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **2009**, *77*, 499–508.
34. Magnan, C.; Baldi, P. SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning, and structural similarity. *Bioinformatics* **2014**, *30*, 1–6.
35. Nielsen, C. Early animal evolution : A morphologist's view. *R. Soc. open sci.* **2019**, *6*, 190638.
36. Pisani, D.; Pett, W.; Dohrmann, M.; Feuda, R.; Rota-Stabelli, O.; Philippe, H.; Lartillot, N.; Wörheide, G. Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl. Acad. Sci. U. S. A.*, **2015**, *112*, 15402-15407.
37. Philippe, H.; Derelle, R.; Lopez, P.; Pick, K.; Borchiellini, C.; Boury-Esnault, N.; Vacelet, J.; Renard, E.; Houliston, E.; Quéinnec, E.; et al. Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **2009**, *19*, 706–712.
38. Pick, K.S.; Philippe, H.; Schreiber, F.; Erpenbeck, D.; Jackson, D.J.; Wrede, P.; Wiens, M.; Alié, A.; Morgenstern, B.; Manuel, M.; et al. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol. Biol. Evol.* **2010**, *27*, 1983–1987.

39. Nosenko, T.; Schreiber, F.; Adamska, M.; Adamski, M.; Eitel, M.; Hammel, J.; Maldonado, M.; Müller, W.E.G.; Nickel, M.; Schierwater, B.; et al. Deep metazoan phylogeny: when different genes tell different stories. *Mol. Phylogenet. Evol.* **2013**, *67*, 223–233.
40. Dunn, C.W.; Hejnol, A.; Matus, D.Q.; Pang, K.; Browne, W.E.; Smith, S.A.; Seaver, E.; Rouse, G.W.; Obst, M.; Edgecombe, G.D.; et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **2008**, *452*, 745–749.
41. Hejnol, A.; Obst, M.; Stamatakis, A.; Ott, M.; Rouse, G.W.; Edgecombe, G.D.; Martinez, P.; Baguna, J.; Bailly, X.; Jondelius, U.; et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B Biol. Sci.* **2009**, *276*, 4261–4270.
42. Ryan, J.F.; Pang, K.; Mullikin, J.C.; Martindale, M.Q.; Baxevanis, A.D. The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *Evodevo* **2010**, *1*, 9.
43. Thompson, J.D.; Gibson, T.J.; Higgins, D.G.; Thompson, J.D.; Gibson, T.J.; Higgins, D.G. Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, **2003**, 2.3.1-2.3.22.
44. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552.
45. Tsirigos, K.D.; Peters, C.; Shu, N.; Käll, L.; Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **2015**, *43*, W401–W407.
46. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699–2699.
47. Cheng, J.; Randall, A.Z.; Sweredoski, M.J.; Baldi, P. SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005**, *33*, 72–76.
48. Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **2002**, *47*, 228–235.
49. Henikoff, S.; Henikoff, J.G. Position-based sequence weights. *J. Mol. Biol.* **1994**, *243*, 574–8.
50. Maddison, D.R.; Swofford, D.L.; Maddison, W.P. NEXUS: An extensible file format for systematic information. *Syst. Biol.* **2006**, *46*, 590-621.
51. Le, S.Q.; Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **2008**, *25*, 1307–1320.
52. 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.
53. Müller, T.; Vingron, M. Modeling amino acid replacement. *J. Comput. Biol.* **2002**, *7*, 761–776.
54. Kosiol, C.; Goldman, N. Different versions of the dayhoff rate matrix. *Mol. Biol. Evol.* **2005**, *22*, 193–199.
55. 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.
56. Stamatakis, A.; Hoover, P.; Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* **2008**, *57*, 758–771.
57. Pattengale, N. D.; Alipour, M.; Bininda-Emonds, O. R. P.; Moret, B. M. E.; Stamatakis, A. How many bootstrap replicates are necessary?. J. Comput. Biol., **2010**, *17*, 337-354.
58. Kimball, R.T.; Wang, N.; Heimer-McGinn, V.; Ferguson, C.; Braun, E.L. Identifying localized biases in large datasets: a case study using the avian tree of life. *Mol. Phylogenet. Evol.* **2013**, *69*, 1021–32.
59. Shen, X.-X.; Hittinger, C.T.; Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* **2017**, *1*, 0126.

60. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313.
61. R Development Core Team. *R:* A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. **2014**, http://www.R-project.org/.
62. Le, S.Q.; Gascuel, O.; Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **2008**, *24*, 2317–2323.
63. 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.
64. Minh, B.Q.; Nguyen, M.A.T.; von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **2013**, *30*, 1188–1195.
65. 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.
66. Embley, T.M.; Van Der Giezen, M.; Horner, D.S.; Dyal, P.L.; Foster, P.; Tielens, A.G.M.; Martin, W.; Tovar, J.; Douglas, A.E.; Cavalier-Smith, T.; et al. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. *Philos. Trans. Royal Soc. B*, **2003**, *358*, 191-203.
67. Hrdy, I.; Hirt, R.P.; Dolezal, P.; Bardonová, L.; Foster, P.G.; Tachezy, J.; Embley, T.M. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **2004**, *432*, 618-622.
68. Woese, C.R.; Achenbach, L.; Rouviere, P.; Mandelco, L. Archaeal Phylogeny: Reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Syst. Appl. Microbiol.* **1991**, *14*, 364-371.
69. Budd, G.E.; Jensen, S. The origin of the animals and a 'Savannah' hypothesis for early bilaterian evolution. *Biol. Rev.* **2017**, *92*, 446-473.
70. Salichos, L.; Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **2013**, *497*, 327–331.
71. Brown, J.M.; Thomson, R.C. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* **2016**, *66*, 571–530.
72. Evans, N.M.; Holder, M.T.; Barbeitos, M.S.; Okamura, B.; Cartwright, P. The phylogenetic position of Myxozoa: exploring conflicting signals in phylogenomic and ribosomal data sets. *Mol. Biol. Evol.* **2010**, *27*, 2733–2746.
73. Schalchian-Tabrizi, K.; Minge, M.A.; Espelund, M.; Orr, R.; Ruden, T.; Jakobsen, K.S.; Cavalier-Smith, T. Multigene phylogeny of Choanozoa and the origin of animals. *PLoS One* **2008**, *3*, e2098.
74. Carr, M.; Leadbeater, B.S.C.; Hassan, R.; Nelson, M.; Baldauf, S.L. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proc. Natl. Acad. Sci.* **2008**, *105*, 16641–16646.
75. Wilke, C.O. Bringing molecules back into molecular evolution. *PLoS Comput. Biol.* **2012**, *8*, e1002572.
76. Crooks, G.E.; Brenner, S.E. An alternative model of amino acid replacement. *Bioinformatics* **2005**, *21*, 975–980.
77. Gerstein, M.; Sonnhammer, E.L.; Chothia, C. Volume changes in protein evolution. *J. Mol. Biol.* **1994**, *236*, 1067–1078.
78. Philippe, H.; Brinkmann, H.; Copley, R.R.; Moroz, L.L.; Nakano, H.; Poustka, A.J.; Wallberg, A.; Peterson, K.J.; Telford, M.J. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* **2011**, *470*, 255.
79. Egger, B.; Lapraz, F.; Tomiczek, B.; Müller, S.; Dessimoz, C.; Girstmair, J.; Škunca, N.; Rawlinson, K.A.;

Cameron, C.B.; Beli, E.; et al. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Curr. Biol.* **2015**, *25*, 1347–1353.

80. Liu, L.; Yu, L.; Kubatko, L.; Pearl, D.K.; Edwards, S. V. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* **2009**, *53*, 320–328.
81. Tsagkogeorga, G.; Turon, X.; Hopcroft, R.R.; Tilak, M.K.; Feldstein, T.; Shenkar, N.; Loya, Y.; Huchon, D.; Douzery, E.J.; Delsuc, F. An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models. *BMC Evol. Biol.* **2009**, *9*, 187.
82. Finet, C.; Timme, R.E.; Delwiche, C.F.; Marlétaz, F. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol.* **2010**, *20*, 2217–2222.
83. Philippe, H.; Brinkmann, H.; Lavrov, D. V; Littlewood, D.T.J.; Manuel, M.; Wörheide, G.; Baurain, D. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol.* **2011**, *9*, e1000602.
84. Lartillot, N.; Brinkmann, H.; Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **2007**, *7 Suppl 1*, S4.
85. Brinkmann, H.; Philippe, H. Animal phylogeny and large-scale sequencing: progress and pitfalls. *J. Syst. Evol.* **2008**, *46*, 274–286.
86. Roure, B.; Baurain, D.; Philippe, H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol. Biol. Evol.* **2013**, *30*, 197–214.
87. Lio, P.; Goldman, N.; Liò, P.; Goldman, N. Models of molecular evolution and phylogeny. *Genome Res.* **1998**, *8*, 1233–1244.
88. Halanych, K.M.; Whelan, N. V; Kocot, K.M.; Kohn, A.B.; Moroz, L.L. Miscues misplace sponges. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E946-E947.
89. Foster, P.G.; Hickey, D.A. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J. Mol. Evol.* **1999**, *48*, 284-290.
90. Katsu, Y.; Braun, E.L.; Guillette, L.J.; Iguchi, T. From reptilian phylogenomics to reptilian genomes: Analyses of c-*Jun* and *DJ-1* proto-oncogenes. *Cytogenet. Genome Res.* **2010**, *127*, 79-93.
91. Wang, H.C.; Singer, G.A.C.; Hickey, D.A. Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* **2004**, *21*, 90-96.
92. Savard, J.; Tautz, D.; Richards, S.; Weinstock, G.M.; Gibbs, R.A.; Werren, J.H.; Tettelin, H.; Lercher, M.J. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* **2006**, *16*, 1334-1338.
93. Boussau, B.; Blanquart, S.; Necsulea, A.; Lartillot, N.; Gouy, M. Parallel adaptations to high temperatures in the Archaean eon. *Nature* **2008**, *456*, 942–945.
94. Hernandez, A.M.; Ryan, J.F. Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses. *BioRxiv* **2019**, 729103.
95. Mallet, J.; Besansky, N.; Hahn, M.W. How reticulated are species? *BioEssays* **2016**, *38*, 140–149.
96. Rivera, M.C.; Lake, J.A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **2004**, *431*, 152-155.
97. Springer, M.S.; Gatesy, J. On the importance of homology in the age of phylogenomics. *Syst. Biodivers.* **2018**, *16*, 210–228.
98. Gatesy, J. A tenth crucial question regarding model use in phylogenetics. *Trends Ecol. Evol.* **2007**, *22*, 509–10.
99. Sanderson, M.J.; Kim, J. Parametric phylogenetics? *Syst. Biol.* **2000**, *49*, 817–829.

100. Steel, M. Should phylogenetic models be trying to "fit an elephant"? *Trends Genet.* **2005**, *21*, 307–309.

101. Dyson, F. A meeting with Enrico Fermi. *Nature* **2004**, *427*, 297.