



HAL
open science

Compositionally constrained sites drive long branch attraction

Lenard L Szantho, Nicolas Lartillot, Gergely Szöllősi, Dominik Schrempf

► **To cite this version:**

Lenard L Szantho, Nicolas Lartillot, Gergely Szöllősi, Dominik Schrempf. Compositionally constrained sites drive long branch attraction. 2022. hal-04272278v1

HAL Id: hal-04272278

<https://hal.science/hal-04272278v1>

Preprint submitted on 30 Aug 2022 (v1), last revised 6 Nov 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Compositionally constrained sites drive long branch
2 attraction

3 Lénárd L. Szánthó^{1,3,4}, Nicolas Lartillot², Gergely J. Szöllősi^{1,3,4,*,+},
4 and Dominik Schrempf^{1,*,+}

5 ¹Dept. Biological Physics, Eötvös University, Pázmány P. stny. 1A., H-1117 Budapest, Hungary

6 ²Laboratoire de Biométrie et Biologie Evolutive UMR 5558, CNRS, Université de Lyon,
7 Villeurbanne, France

8 ³ELTE-MTA “Lendület” Evolutionary Genomics Research Group, Pázmány P. stny. 1A.,
9 H-1117 Budapest, Hungary

10 ⁴Institute of Evolution, Centre for Ecological Research, Konkoly-Thege M. u 29-33, Budapest,
11 Hungary

12 *Corresponding authors: ssolo@elte.hu and dominik.schrempf@gmail.com

13 +Equal contribution

14 February 23, 2022

15 **Abstract**

16 Accurate phylogenies are fundamental to our understanding of the pattern
17 and process of evolution. Yet, phylogenies at deep evolutionary timescales,
18 with correspondingly long branches, have been fraught with controversy re-
19 sulting from conflicting estimates from models with varying complexity and
20 goodness of fit. Analyses of historical as well as current empirical datasets,
21 such as alignments including Microsporidia, Nematoda or Platyhelminthes,
22 have demonstrated that inadequate modeling of across-site compositional het-
23 erogeneity, which is the result of biochemical constraints that lead to varying
24 patterns of accepted amino acid along sequences, can lead to erroneous topolo-
25 gies that are strongly supported. Unfortunately, models that adequately
26 account for across-site compositional heterogeneity remain computationally
27 challenging or intractable for an increasing fraction of contemporary datasets.
28 Here, we introduce “compositional constraint analysis”, a method to investi-

29 gate the effect of site-specific amino acid diversity on phylogenetic inference,
30 and show that more constrained sites with lower diversity and less constrained
31 sites with higher diversity exhibit ostensibly conflicting signal for models ig-
32 noring across-site compositional heterogeneity. We demonstrate that more
33 complex models accounting for across-site compositional heterogeneity can
34 ameliorate this bias. We present CAT-PMSF, a pipeline for diagnosing and
35 resolving phylogenetic bias resulting from inadequate modeling of across-site
36 compositional heterogeneity based on the CAT model. Our analyses indicate
37 that CAT-PMSF is unbiased. We suggest using CAT-PMSF when conver-
38 gence of the CAT model cannot be assured. We find evidence that compo-
39 sitional constrained sites are driving long branch attraction in two metazoan
40 datasets and recover evidence for Porifera as the sister group to all other
41 animals.

42 1 Introduction

43 Understanding the biological foundations of contemporary life on Earth requires
44 detailed knowledge of evolutionary history. The history of speciation events informs
45 us about the appearance of advantageous innovations and the loss of dispensable
46 traits in a continuously changing environment. Consequently, development of phy-
47 logenetic models inferring the history of speciation events has continued at an
48 impressive pace during the past decades.

49 Phylogenetic models do not reflect the full complexity of evolution but in-
50 evitably present a simplified picture of the generating processes underlying sequence
51 evolution. Unfortunately, overly-simplistic models can lead to model misspecifica-
52 tion and long branch attraction (LBA; e.g., [Felsenstein, 1978](#); [Hendy and Penny,](#)
53 [1989](#); [Zharkikh and Li, 1993](#); [Tateno et al., 1994](#); [Bergsten, 2005](#); [Brinkmann et al.,](#)
54 [2005](#); [Philippe et al., 2011b](#)). LBA is a bias in the inferred topology, as a result of
55 which long branches are more likely to branch together than with short ones, inde-
56 pendent of the true evolutionary history. Substitution models ([Jukes and Cantor,](#)
57 [1969](#)) account for the possibility of multiple substitutions per site, and reduce LBA
58 ([Felsenstein, 1973](#); but see [Farris, 1999](#)) compared to models based on maximum
59 parsimony. Probabilistic substitution models describe the evolution of a site as a
60 series of transitions between states. Every possible transition occurs at a specific

61 rate which is the expected number of transitions from one state to another per unit
62 time. Standard substitution models perform well on most datasets (Ripplinger and
63 Sullivan, 2010).

64 However, LBA can be an issue even when using probabilistic substitution mod-
65 els. For example, if variation of evolutionary rate between sites (across-site rate
66 heterogeneity) is ignored. Models accounting for across-site rate heterogeneity have
67 been proposed early on (Yang, 1993), and have been shown to ameliorate LBA in
68 some cases (Kuhner and Felsenstein, 1994; Philippe et al., 2011b). However, across
69 site variation is not restricted to rates. Also, the relative abundance of nucleotide
70 or amino acid characters is variable. For example, sites buried deeply in folded pro-
71 teins tend to be more constrained than sites that end up on the surface (Koshi and
72 Goldstein, 1995; Yeh et al., 2014; Jimenez et al., 2018), and depending on the pro-
73 tein structure may exhibit stronger preferences in hydrophobicity than other sites.
74 Models ignoring across-site compositional heterogeneity are prone to LBA because
75 they underestimate the probability of convergent substitutions at compositionally
76 constrained sites. In particular, the probability of independent substitutions to
77 the same state depends on the number of acceptable amino-acids, which differs
78 across sites. Models ignoring across-site compositional heterogeneity pool all sites,
79 and ignore the variation of the evolutionary process across sites. Indeed, analyses
80 of a series of datasets exhibiting previously contentious evolutionary relationships
81 provide evidence that ignoring across-site compositional heterogeneity can lead to
82 LBA (Phillips et al., 2004; Brinkmann et al., 2005; Philippe et al., 2005b,a; Delsuc
83 et al., 2006; Lartillot et al., 2007; Philippe et al., 2009, 2011a; Brown et al., 2013;
84 Ryan et al., 2013; Cannon et al., 2016; Simion et al., 2017).

85 Thus, there is accumulating evidence that accounting for across-site hetero-
86 geneities is key to an accurate reconstruction of deep evolutionary relationships.
87 The classic approach to modeling such heterogeneities in the phylogenetic inference
88 process are mixture models that combine substitution models specifically tailored
89 to the evolutionary processes observed in the data. In order to model across-site
90 rate heterogeneity, we use a mixture of substitution models with the same relative
91 but different absolute substitution rates (e.g., Yang, 1993; Kalyaanamoorthy et al.,
92 2017).

93 Modeling across-site compositional heterogeneity, however, requires construct-

94 ing a model describing the evolution of sites subject to different compositional
95 constraints. To do so, for time-reversible substitution models, we can leverage the
96 separation of substitution rates into the product of (1) symmetric exchangeabilities
97 describing differences in rates of exchange between pairs of states and (2) station-
98 ary frequencies of the target states, or figuratively, the probability of sampling the
99 target states after waiting for a long time. Any time-reversible substitution model
100 is fully specified by a set of symmetric exchangeabilities and the set of stationary
101 frequencies (sometimes also referred to as a profile or a stationary distribution).
102 Assuming that biochemical constraints primarily affect site-specific amino acid pref-
103 erences in the long term, across-site compositional heterogeneity can be accounted
104 for by composing a number of substitution models sharing a single set of exchange-
105 abilities but differing in their stationary distributions (distribution mixture models;
106 [Quang et al., 2008](#); [Schrempf et al., 2020](#))

107 We distinguish between general distribution mixture models estimated from cu-
108 rated training databases, and distribution mixture models directly estimated from
109 the datasets at hand. For example, [Wang et al. \(2008\)](#) directly estimate mixture
110 model components using principal component analysis. [Susko et al. \(2018\)](#) use a
111 composite likelihood approach and additional methods such as taxon weighing. In
112 contrast, the rationale behind providing and using general mixture models is the
113 assumption that the underlying evolutionary processes share universal features.
114 [Quang et al. \(2008\)](#) use the expectation maximization algorithm to infer general
115 mixture models consisting of 10, 20, . . . , 60 components (C10, C20, . . . , C60, col-
116 lectively CXX models). [Schrempf et al. \(2020\)](#) used a clustering approach together
117 with different compositional transformations to provide a set of general mixture
118 models, termed universal distribution mixtures (UDM), with the number of com-
119 ponents ranging from four up to several thousand. They also provide the clustering
120 method EDCluster to infer dataset specific distribution mixture models.

121 Finite mixtures can be used in the context of maximum likelihood (ML) phy-
122 logenetic inference. However, statistical analyses of model fit and investigation of
123 known cases of LBA indicate that a large number of components are necessary for
124 robustness against LBA ([Schrempf et al., 2020](#)), which is computationally intensive
125 and memory intense.

126 Bayesian approaches can more easily accommodate richer mixtures. In partic-

127 ular, nonparametric Bayesian methods do not require explicit specification of the
128 number of mixture components nor their stationary distributions. In particular, the
129 CAT model (Lartillot and Philippe, 2004) uses a Dirichlet process prior to approxi-
130 mate an arbitrary mixture of stationary distributions across sites. The CAT model
131 was shown to be better fitting and less prone to LBA than site-homogeneous models
132 for a series of classic datasets including Nematoda and Platyhelminthes (Lartillot
133 et al., 2007) as well as in phylogenomic analyses of the tree of life (Williams et al.,
134 2020). The impediment of nonparametric Bayesian methods and specifically, the
135 CAT model, is that it compounds two computationally challenging, but separately
136 tractable problems, the non-parametric inference of the underlying distribution
137 across sites and the exploration of tree space. The composition of these two prob-
138 lems is challenging and can lead to convergence problems.

139 In all cases, however, mixture modeling approaches accounting for across-site
140 compositional heterogeneity are complex and require considerable computational
141 resources (e.g., Whelan and Halanych, 2016). In order to reduce computational
142 cost, Wang et al. (2018) proposed a two-step approximation. First, site-specific
143 stationary distributions are estimated using a reference mixture model and a fixed
144 guide tree. Second, the tree is reconstructed using the fixed stationary distributions
145 obtained in the first step. Thereby, run time is reduced while robustness against
146 LBA is improved compared to using the reference mixture model alone. In par-
147 ticular, the site-specific stationary distributions are set to the posterior mean site
148 frequencies (PMSF) of the reference mixture model. As a result, the phylogenetic
149 accuracy of the PMSF approach is inherently limited by how well the reference
150 mixture model captures across-site compositional heterogeneity. There is no rea-
151 son to restrict the use of the PMSF approach to empirical mixture models: any
152 random-effect model meant to account for pattern heterogeneity could in principle
153 be used here as a reference mixture model for computing the posterior means of
154 the site-specific stationary distributions.

155 In this work we follow a multistep procedure similar to the PMSF model and
156 address two points: First, on the computational side, we extend the PMSF ap-
157 proach by using the CAT model instead of an empirical profile mixture model as
158 the reference model for computing the profiles. Importantly, the CAT model is used
159 under a fixed tree topology. Thus, the problem of simultaneous inference of both,

160 the site-specific stationary distributions and the tree including the tree topology is
161 reduced to a search of site-specific stationary distributions, and tree branch lengths
162 only. We termed our approach CAT-PMSF.

163 Second, we investigate the contribution of sites with different degrees of com-
164 positional constraints to LBA. In particular, we test to what extent more severely
165 constrained sites exhibit bias towards LBA trees under models that do not ad-
166 equately account for across-site compositional heterogeneity. We employ “com-
167 positional constraint” analysis, which examines phylogenetic signal for alternative
168 topologies as a function of per site compositional diversity measured by the effective
169 number of amino acids.

170 Examining simulated alignments as well as classic empirical datasets including
171 Microsporidia (Brinkmann et al., 2005), Nematoda, and Platyhelminthes (Philippe
172 et al., 2005a), we find conflicting phylogenetic signal across sites with different
173 degrees of compositional constraints. Based on these results we apply compositional
174 constraint analysis to two recent datasets (Ryan et al., 2013; Simion et al., 2017)
175 aiming to resolve the early diversification of animal lineages.

176 2 Results

177 In the following, we use the terms *site-homogeneous* and *site-heterogeneous* when
178 referring to models ignoring and accounting for across-site compositional hetero-
179 geneity, respectively. Further, we use the term *tree* to denote a directed acyclic
180 graph with node labels and branch lengths, in which exactly one branch connects
181 any two nodes. We use the term *topology* to denote a tree without information on
182 branch length. We specify an evolutionary model with exchangeabilities EX, and
183 across-site compositional heterogeneity model ASCH as EX+ASCH. All discussed
184 evolutionary models used for simulations as well as inferences implicitly use dis-
185 crete gamma rate heterogeneity with four components. We add a flag +PMSF to
186 denote usage of the posterior mean site frequency model (Wang et al., 2018).

187 In brief, CAT-PMSF comprises three steps: (1) Estimating a guide topology
188 using a site-homogeneous model, (2) Estimating site-specific stationary distribu-
189 tions with the CAT model in PhyloBayes (Lartillot and Philippe, 2004) using the
190 guide topology, and (3) phylogenetic inference in a ML framework with a distri-

191 bution mixture model sharing one set of exchangeabilities, and using the obtained
192 site-specific stationary distributions (see Methods).

193 **Simulation study.** We assessed and compared the accuracy of CAT-PMSF with
194 other site-homogeneous and site-heterogeneous models. To this end, we simulated
195 amino acid sequence alignments with a length of 10 000 sites along Felsenstein-type
196 quartet trees (insets of Figure 1; Felsenstein, 1978). We used uniform exchange-
197 abilities (Poisson; Felsenstein, 1973) and an across-site compositional heterogeneity
198 model with site-specific stationary distributions based on a UDM model (see Meth-
199 ods; Schrempf et al., 2020). We set the branch length of the short branch q to 0.1,
200 and varied the length of the long branch p between 0.1 and 2.0.

201 The true topology (a Felsenstein-type quartet) was not recovered with site-ho-
202 mogeneous models when $p \geq 0.8$ (Figures S7, S8, and Tables S1, S2). Figure 1
203 shows the results of the compositional constraint analysis for different values of
204 p , contrasting the site-wise log-likelihood differences between the maximum like-
205 lihood trees constrained to the genuine Felsenstein-type and incorrect Farris-type
206 topologies prone to LBA (see Methods) for $p = 0.2, 0.8$ and 1.2. Figures S7,
207 and S8 show results for other values of p . We binned sites according to their
208 effective number of amino acids (K_{eff} , see Methods). Lower values of K_{eff} corre-
209 spond to sites under stronger compositional constraint. “Compositional constraint
210 analysis” compares per site phylogenetic signal for two alternative topologies as a
211 function of compositional constraint (i.e., different values of K_{eff}). Here, positive
212 log-likelihood differences indicate support for the true (Felsenstein-type quartet)
213 topology. Conversely, negative values indicate support for the LBA (Farris-type
214 quartet) topology. In the absence of model misspecification we expect consistent
215 phylogenetic signal across sites, and independent of the true value of K_{eff} .

216 At odds with this expectation, site-homogeneous evolutionary models exhibit
217 conflicting phylogenetic signal between sites with low and high K_{eff} values (Fig-
218 ures 1, S8 and S7). For the site-homogeneous evolutionary models, more con-
219 strained sites with a low value of K_{eff} exhibit bias towards the incorrect (Farris-
220 type) topology. For $p \geq 0.8$, the bias outweighs the correct signal of less constrained
221 sites with high values of K_{eff} , and the incorrect (Farris-type) topology has higher
222 support than the true (Felsenstein-type) topology. In contrast, the site-heteroge-

223 neous LG+C60+PMSF model as well as the CAT-PMSF method show consistent
224 support for the true topology across all sites and values of p .

225 To ascertain the statistical significance of the compositional constraint analy-
226 sis we calculated Pearson correlation coefficients and associated p-values between
227 the log-likelihood differences and the site-specific K_{eff} values (Table S3). Site-ho-
228 mogeneous models exhibit large and significant correlation for $p \geq 0.8$, whereas
229 the log-likelihood differences and K_{eff} values of site-heterogeneous models are not
230 correlated.

231 Approximately unbiased (AU) tests (Shimodaira, 2002) of maximum likelihood
232 trees inferred by the GTR+CAT-PMSF model constrained to the two alternative
233 topologies reject the (Farris-type) LBA topology in favor of the true topology for
234 $p < 1.3$ (Table S5). AU tests of the Poisson+CAT-PMSF, and LG+CAT-PMSF
235 models show similar results (Tables S6, and S7).

236 Finally, we note that site-heterogeneous models with LG exchangeabilities per-
237 form well, although the genuine exchangeabilities are uniformly distributed. CAT-
238 PMSF is even more accurate when using the true topology or the true site-specific
239 stationary distributions (Figures S7 and S8).

240 **Applications to empirical data.** Similar to the simulation study above, for
241 empirical alignments the site-specific stationary distributions obtained in Step 2
242 of the CAT-PMSF pipeline can be used to quantify the strength of compositional
243 constraints (i.e., to estimate K_{eff} values per site) and perform compositional con-
244 straint analysis. Figure 2 shows results for three datasets exhibiting classic LBA
245 artifacts when we use site-homogeneous models for inference: The placement of
246 Platyhelminthes and Nematoda (Philippe et al., 2005a), as well as the placement
247 of Microsporidia (Brinkmann et al., 2005; Lartillot et al., 2007).

248 For site-homogeneous models, the site-specific log-likelihood differences between
249 the maximum likelihood trees constrained to the two competing topologies (insets
250 of Figure 2; top vs. bottom) show conflicting phylogenetic signal as a function of the
251 strength of compositional constraint. The bias towards the topologies exhibiting
252 LBA artifacts of more constrained sites outweighs the signal of less constrained
253 sites in all three datasets.

254 The site-heterogeneous LG+C60+PMSF model shows reduced, but still appar-

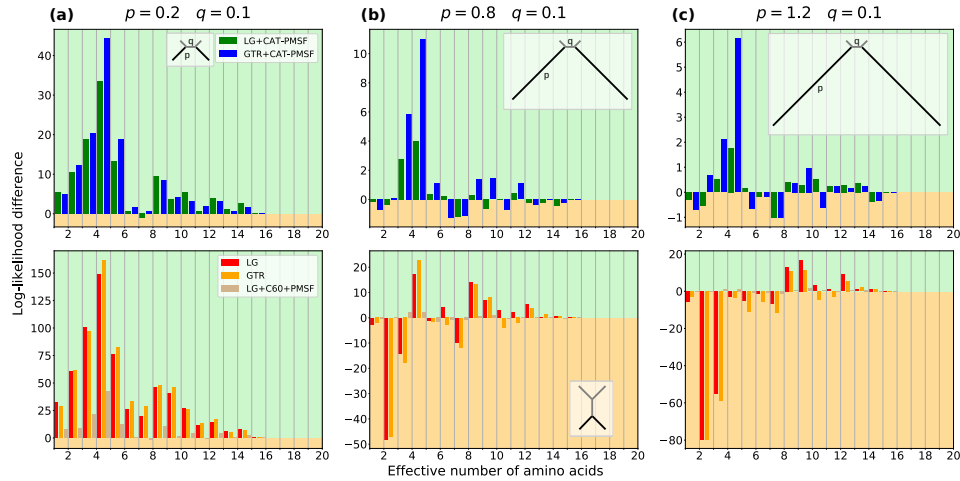


Figure 1: Highly constrained sites drive long branch attraction artifacts in the Felsenstein zone. We simulated amino acid alignments with 10 000 sites exhibiting across-site compositional heterogeneity (Schrempf et al., 2020) along Felsenstein-type trees (insets in top row; Felsenstein, 1978) with different branch lengths $q = 0.1$, and $p = 0.3, 0.8$, and 1.2 from (a) to (c). We performed analyses with CAT-PMSF, the LG (Le and Gascuel, 2008) and the GTR (Tavaré, 1986) models constrained to the correct topology as well as the incorrect Farris-type topology (insets in bottom row; Farris, 1999) with IQ-TREE 2 (Minh et al., 2020). The site-wise log-likelihood differences $\Delta \log L$ between the maximum likelihood trees of the two competing topologies binned according to their effective number of amino acids K_{eff} are shown. A positive value (green background) indicates support for the Felsenstein-type topology, a negative value (orange background) indicates support for the Farris-type topology prone to long branch attraction. We do not expect a uniform distribution across the bins, because they are of different size. The LG, and GTR models incorrectly infer Farris-type trees if $p \geq 0.8$.

255 ent conflict compared to site-homogeneous models and the LG+C10+PMSF model
256 (Figure S9). For Platyhelminthes, the bias is strong enough to that the total like-
257 lihood across all sites is higher for the LBA topology, while for the datasets involv-
258 ing Nematoda and Microsporidia, the LG+C10+PMSF model recover the correct
259 topology, albeit with reduced support. In general, the results of the LG+C10+PMSF
260 and LG+C60+PMSF models are consistent with the observation (Schrempf et al.,
261 2020) that increasing the number of components decreases the bias introduced by
262 more constrained sites. Pearson correlation coefficients are greater for site-homo-
263 geneous models than for models LG+C10+PMSF and LG+C60+PMSF models,
264 but significant for each of these (Table S4).

265 In contrast, CAT-PMSF exhibits consistent signal towards the correct topolo-
266 gies across all sites and datasets with no significant correlation between log-likelihood
267 difference and site-specific K_{eff} value (Table S4). The maximum likelihood trees
268 inferred by CAT-PMSF are consistent with the accepted phylogenetic relationships
269 and AU tests confirm the rejection of trees with LBA topologies (Tables S8-S10).

270 **The phylogenetic position of Ctenophora.** Finally, we used CAT-PMSF on
271 two metazoan datasets (Ryan et al., 2013; Simion et al., 2017) to investigate early
272 evolutionary relationships on the animal tree of life. It is currently a matter of
273 intense debate whether sponges (Porifera) or comb jellies (Ctenophora) are the
274 sister group to all other animals (e.g., Kapli and Telford, 2020; Li et al., 2020).
275 We refer to the competing hypotheses as Porifera-sister and Ctenophora-sister,
276 respectively.

277 Compositional constraint analysis under site-homogeneous models, as well as
278 combinations of PMSF and site-heterogeneous mixture models with 20 and 60
279 components exhibit patterns of conflicting phylogenetic signal for sites with differ-
280 ent degrees of compositional constraints for both the alignments from Simion et al.
281 (2017) and Ryan et al. (2013). The conflicting signal is consistent with LBA driving
282 the placement of Ctenophora as the first animal group to emerge (cf. Table S4).

283 Under site-homogeneous models, sites with K_{eff} values up to approximately
284 10 – 12 exhibit strong preference for Ctenophora-sister (Figures 3 and S12). Sites
285 with higher K_{eff} values, however, switch their preference toward Porifera-sister. In
286 contrast, under the CAT-PMSF models the Simion et al. (2017) dataset exhibits

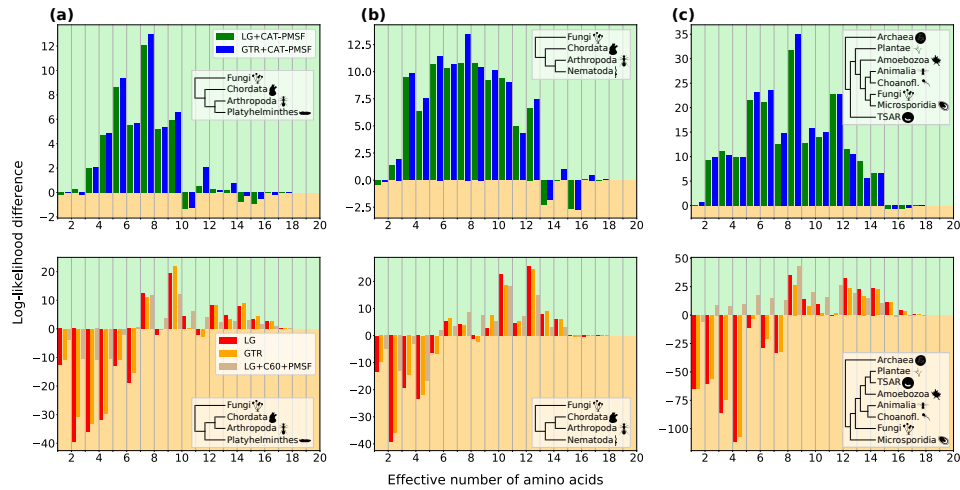


Figure 2: Highly constrained sites explain classic examples of long branch attraction. We analyzed three empirical datasets including (a) Platyhelminthes and (b) Nematoda (Philippe et al., 2005a), and (c) Microsporidia (Brinkmann et al., 2005). We performed analyses with CAT-PMSF, the LG (Le and Gascuel, 2008), the GTR (Tavaré, 1986), and the LG+C60+PMSF (Quang et al., 2008; Wang et al., 2018) models constrained to either one of two competing topologies (insets in top versus bottom rows) with IQ-TREE 2 (Minh et al., 2020). The site-specific log-likelihood differences $\Delta\log L$ between the maximum likelihood trees of the two competing topologies binned according to their effective number of amino acids K_{eff} estimated by PhyloBayes (Lartillot and Philippe, 2004) are shown. A positive value (green background) indicates support for the now accepted topology, a negative value (orange background) indicates support for the topology prone to long branch attraction. We do not expect a uniform distribution across the bins, because they are of different size. Site-homogeneous models infer the wrong topology for all three datasets.

287 consistent phylogenetic signal (Table S3) favoring a Porifera-sister topology and re-
288 jecting the Ctenophora-sister hypothesis (AU test p-values between 3.1×10^{-4} and
289 7.7×10^{-4} ; Table S11) with the closest out group, Choanoflagellata (Figure S4a,
290 Figure S10). For the Ryan et al. (2013) alignment, the total log-likelihood differ-
291 ence of CAT-PMSF between the two hypotheses is marginal at only 0.8, suggesting
292 a lack of resolution in this dataset. None of the models we investigated exhibit
293 phylogenetic consistent signal across sites with different degrees of compositional
294 constraints.

295 3 Discussion

296 We introduce CAT-PMSF, a method for phylogenetic inference from alignments
297 exhibiting across-site compositional heterogeneity. The CAT-PMSF pipeline uses
298 the site-specific amino-acid preferences estimated by a non-parametric Bayesian
299 approach in the context of a downstream maximum likelihood analysis. Doing
300 so combines the benefits of both approaches: a more accurate inference of the
301 patterns across sites with a computationally more efficient and more reproducible
302 inference of the tree topology. In addition to phylogenetic inference the CAT-PMSF
303 pipeline can also be used to investigate the consistency of phylogenetic signal for
304 sites under different degrees of compositional constraints. Using compositional
305 constraint analysis, we elucidate inconsistent signal when using site-homogeneous
306 and site-heterogeneous mixture models for phylogenetic inference from simulated
307 data exhibiting across-site compositional heterogeneity as well as empirical data.

308 **Simulation study and empirical results** In the simulation study (Figure 1),
309 site-homogeneous models favored the LBA (Farris-type) topology when the length
310 p of the terminal branches was long enough. By separating the contribution of sites
311 as a function of compositional constraint, we demonstrated that sites under strong
312 compositional constraints drive the bias leading to the LBA.

313 The threshold K_{eff} value separating sites supporting the correct topology and
314 sites supporting the LBA topology depended on the length p of the terminal
315 branches: The longer the terminal branches, the higher the threshold K_{eff} value.
316 We expect this observation holds more generally.

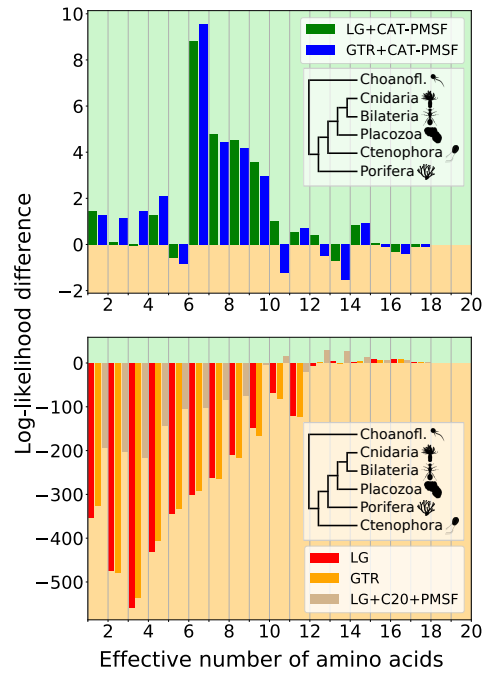


Figure 3: **CAT-PMSF shows consistent signal for Porifera as the sister group to all other animals.** We performed analyses with CAT-PMSF, the LG (Le and Gascuel, 2008), the GTR (Tavaré, 1986), and the LG+C20+PMSF (Quang et al., 2008; Wang et al., 2018) models constrained to either one of two competing topologies (insets in top versus bottom rows) with IQ-TREE 2 (Minh et al., 2020) on the alignment from Simion et al. (2017). The site-specific log-likelihood differences $\Delta\log L$ between the maximum likelihood trees of the two competing topologies binned according to their effective number of amino acids K_{eff} estimated by PhyloBayes (Lartillot and Philippe, 2004) are shown. We do not expect a uniform distribution across the bins, because they are of different size. Site-homogeneous models and the site-heterogeneous LG+C20+PMSF model show inconsistent signal between more versus less constrained sites and favor Ctenophora at the animal root. CAT-PMSF favors Porifera at the animal root, although this result is only significant when using the closest outgroup exclusively.

317 In our simulations, total support of site-homogeneous models shifted from a
318 Felsenstein-type topology towards a Farris-type topology when $p \geq 0.8$. In this
319 case, sites with K_{eff} values above the mentioned separating threshold failed to
320 compensate for the bias introduced by sites with K_{eff} values below the threshold.

321 We observe no bias when using site-heterogeneous models such as CAT-PMSF
322 (Figure 1). Although we expect such a result, it is satisfying that inferences of
323 CAT-PMSF lack bias even for large values of $p \leq 1.2$ (Figures S7 and S8)

324 We can discover bias towards one of the topologies in simulation studies be-
325 cause we know the true parameters and trees. Bias is harder to detect in analyses
326 of empirical data. The compositional constraint analyses detect conflicting sig-
327 nal between more and less constrained sites. Detection of such inconsistencies is
328 a strong indicator for bias: Knowing the stationary distribution of a site alone
329 should not provide us with information about the favored evolutionary history. In
330 mathematical terms, the log-likelihood difference of a site between two hypotheses
331 should be conditionally independent given the stationary distribution of that site.
332 In contrast, we expect the signal obtained from more and less constrained sites be
333 consistent up to random statistical error.

334 In our analyses of empirical data we observed strong inconsistencies between
335 more and less constrained sites for site-homogeneous models and hardly any incon-
336 sistencies when using CAT-PMSF (Figure 2). Pearson correlation coefficients and
337 p-values confirm this observation across a wide range of simulated and empirical
338 datasets (Tables S3, and S4).

339 The results are more nuanced for the alignments involving Ctenophora. In the
340 case of site-homogeneous models, we observe the value of K_{eff} correlates strongly
341 with the log-likelihood difference between the two competing topologies (Figure 3).
342 Moreover, for the dataset provided by Simion et al., CAT-PMSF supports Porifera-
343 sister — similar to the results reported by the original authors, who applied the
344 CAT model to sub-sampled alignments comprising 100 000 sites. The support of
345 CAT-PMSF for Porifera-sister is significant, when we use the closest outgroups
346 exclusively (Table S11). If we add more distant outgroups, the results are less
347 conclusive (Table S11). Long branch attraction provides an explanation for this
348 observation: more distant outgroups attract the outgroups closer related to the
349 species of interest. In turn, the elongated basal branch of animals increases bias

350 due to LBA for branches leading to the metazoan root.

351 We interpret these findings as a confirmation for sponges being the sister group
352 to all other animals (dataset of [Simion et al., 2017](#)), and believe that the incon-
353 clusive results obtained from the dataset of [Ryan et al. \(2013\)](#) reflect a lack of
354 phylogenetic resolution. Irrespective of the final evolutionary history of Metazoa,
355 our results add important evidence that ignoring across-site compositional hetero-
356 geneity leads to LBA ([Phillips et al., 2004](#)).

357 **Further notes** The results of CAT-PMSF are conservative because the CAT
358 model estimates the site-specific stationary distributions using guide topologies
359 prone to LBA artifacts. That is, the guide topologies are obtained with site-
360 homogeneous models. Even so, CAT-PMSF correctly infers the genuine trees in the
361 simulation study (Table S1), and trees that we are convinced to be free from LBA
362 artifacts in the analyses comprising empirical datasets (Figures S3, S2, S1, S6, S5
363 and S4). This observation justifies the usage of site-homogeneous models in Step
364 1 of the CAT-PMSF pipeline.

365 In the simulation study, we observe severe bias when using site-homogeneous
366 models, and no bias or reduced bias when using CAT-PMSF. Further, the absolute
367 values of the log-likelihood differences are greater for site-homogeneous models than
368 for site-heterogeneous models. That is, site-heterogeneous models have reduced
369 power in discriminating between competing hypotheses (bias-variance tradeoff).

370 In general, site-homogeneous models show conflicting signal between more and
371 less constrained sites, but we observe hardly any such inconsistencies when using
372 CAT-PMSF. In any case, even when the signal across sites is consistent, evidence
373 obtained from highly constrained sites should be examined carefully, especially
374 when highly constrained sites weigh more heavily than less constrained sites. We
375 are convinced that inconsistencies between more and less constrained sites are a
376 strong indicator for the presence of LBA.

377 [Li et al. \(2020\)](#) argue that only the most parameter rich models favor Porifera-
378 sister, and so Porifera-sister is not a likely scenario. In contrast, [Schrempf et al.](#)
379 (2020) report that statistical tests favor models using more stationary distributions.
380 This point is confirmed here, where we see that CXX models, in spite of being
381 generally more robust against LBA than site-homogeneous models, may still be

382 insufficient and result in conflicting signal (Figures S8, S9 and S10). In practice,
383 each site is different, and we can not expect all sites to share a universal stationary
384 distribution. In fact, we do not even expect stationarity. In our opinion, we should
385 analyze complex models and decide about which parameters are necessary to grasp
386 the complexities of evolution. With CAT-PMSF we further explored this path.
387 The CAT-PMSF method uses site-specific stationary distributions and therefore is
388 a parameter-rich model.

389 In comparison, the site-specific posterior mean stationary distributions of the
390 classical PMSF approach are a superposition of a finite set of stationary distribu-
391 tions of the underlying mixture model. Consequently, the stationary distribution
392 with the lowest K_{eff} value constitutes a hard, lower limit. Further, we expect even
393 the richest distribution mixture models do not offer adequate variability of compo-
394 nents with stationary distributions exhibiting low K_{eff} values. For example, there
395 are twenty different stationary distributions with K_{eff} values close to 1.0, $190 = \binom{20}{2}$
396 stationary distributions with K_{eff} values close to 2.0, and so on.

397 Finally, the speed benefit of CAT-PMSF originates from fixing the topology
398 during the Bayesian analysis with the CAT model. Of course, estimating the site-
399 specific stationary distributions is still by far the most time-consuming step. In the
400 future, we aim to design improved methods estimating site-specific stationary dis-
401 tributions. Specifically, we are thinking about methods based on machine learning
402 such as AlphaFold (Jumper et al., 2021).

403 In conclusion, our results provide evidence for a potential LBA caused by model
404 misspecification, and thus, an independent qualitative argument for choosing the
405 adequate model for phylogenetic inference.

406 4 Methods

Effective number of amino acids. Let $\pi = (\pi_A, \pi_R, \dots, \pi_V)$ be a distribution
of amino acid frequencies, and

$$G(\pi) = \sum_{i \in \{A, R, \dots, V\}} \pi_i^2. \quad (1)$$

$G(\pi)$ is the probability of sampling the same amino acid twice, which is equivalent to a random event of character homoplasy. The effective number of amino acids of distribution π is defined as

$$K_{\text{eff}}(\pi) = G(\pi)^{-1}. \quad (2)$$

407 K_{eff} is a convenient measure because it ranges from 1.0, when one amino is used
408 exclusively, to 20.0 for a uniform distribution.

409 **Simulations.** In order to assess the accuracy of CAT-PMSF, we simulated align-
410 ments of 10000 amino acids under a distribution mixture model (Schrempf et al.,
411 2020). We used Poisson exchangeabilities (Felsenstein, 1973; Nei, 1987) and a
412 discrete gamma rate model (Yang, 1993) with 4 categories with shape parameter
413 $\alpha = 0.78$. The distribution mixture model has site-specific stationary distributions.
414 For each site, we sampled a random distribution from a universal set of distribu-
415 tions (Schrempf et al., 2020) obtained from the HOGENOM (Dufayard et al., 2005)
416 and HSSP (Schneider et al., 1997) databases.

417 We used Felsenstein-type topologies with four leaves. The quartet trees had
418 different branch length proportions between short (q) and long (p) branches (insets
419 of Figure 1). We fixed q to 0.1, and changed p between 0.1 and 2.0. We stored
420 the randomly sampled site-specific stationary distributions used for the simulation.
421 For the simulations we used the Elynx suite (Schrempf, 2021). The scripts and the
422 simulated data are available at <https://github.com/drenal/cat-pmsf-paper>.

423 **CAT-PMSF.** Figure 4 shows an overview of the CAT-PMSF pipeline. The input
424 to CAT-PMSF is an alignment. The output of the CAT-PMSF pipeline is a tree
425 robust to LBA.

426 Step 1: Use a site-homogeneous model to infer a ML tree with IQ-TREE 2 (Minh
427 et al., 2020). Specifically, we used LG exchangeabilities (Le and Gascuel, 2008),
428 the empirical stationary distribution of amino acids, and a discrete gamma rate
429 model with 4 categories (LG+G4 in IQ-TREE 2 terminology).

430 Step 2: Use the obtained tree, which is prone to LBA artifacts, in a subse-
431 quent Bayesian analysis with the CAT model (Lartillot and Philippe, 2004) in
432 PhyloBayes (Lartillot et al., 2013). Fix the topology of the tree for the second

433 step. Thereby, we reduce the computational requirements of the CAT model.
434 Then, extract the posterior mean site-specific stationary distributions of amino
435 acids. In our analyses, we either used Poisson, LG or GTR (Tavaré, 1986) ex-
436 changeabilities, and a discrete gamma rate model with 4 categories. We ran
437 two Markov chains until either the effective sample size of all parameters was
438 above 100, or after visual inspection with Tracer (Rambaut et al., 2018) indi-
439 cated convergence. For the GTR model, we also extracted the posterior mean ex-
440 changeabilities from the results of PhyloBayes. All scripts are available at <https://github.com/drenal/cat-pmsf-paper>.

442 Step 3: Use the custom site-specific stationary distributions in IQ-TREE 2.
443 To this end, use capabilities of IQ-TREE 2 implemented as part of the PMSF
444 method (Wang et al., 2018). The PMSF method has two steps. First, infer the site-
445 specific stationary distributions. Second, use the inferred site-specific stationary
446 distributions for phylogenetic inference. Here, we use the second step of the PMSF
447 method together with the custom site-specific stationary distributions obtained in
448 Step 2 of the CAT-PMSF pipeline.

449 **Preparation of figures.** We calculated the site-specific likelihood differences
450 between two analyses constrained to topologies A , and B , respectively (`-g` flag in
451 IQ-TREE 2). For example, in the simulation study, topology A was of the Farris-
452 type and topology B was of the Felsenstein-type. For each site i , we calculated the
453 log-likelihood difference as

$$\Delta \log L_i = \log L_i^B - \log L_i^A. \quad (3)$$

454 A positive value of $\Delta \log L_i$ indicates that site i supports topology B . A negative
455 value indicates support for topology A .

456 We ordered and binned the sites according to their K_{eff} values. For the simula-
457 tion study, we used the genuine K_{eff} values. For the analyses of empirical datasets,
458 we used the K_{eff} values calculated from the site-specific stationary distribution ob-
459 tained in Step 2 of the CAT-PMSF pipeline. We performed binning with windows
460 sizes of $1.0 K_{\text{eff}}$ and summed the site-specific log-likelihood differences within each
461 bin.

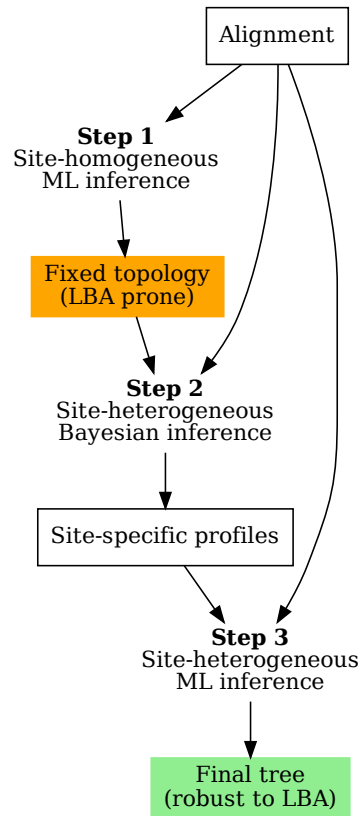


Figure 4: **The CAT-PMSF pipeline.** (1) Apply a site-homogeneous maximum likelihood (ML) model (LG+G4; Le and Gascuel, 2008; Yang, 1993) with IQ-TREE 2 (Minh et al., 2020). The obtained tree may still suffer from long branch attraction (LBA). (2) Fix the topology of this tree in a site-heterogeneous inference with the Bayesian CAT model (Lartillot and Philippe, 2004) and extract the posterior mean site-specific stationary distributions of amino acids. (3) Estimate a tree robust to LBA with the obtained site-specific stationary distributions in IQ-TREE 2.

462 Exemplary taxa on the inset trees are represented using Phylopic ([http://](http://phylopic.org/)
463 phylopic.org/), the silhouette for Microsporidia is based on Tosoni et al. (2002,
464 Figure 7).

465 **Dataset involving Platyhelminthes and Nematoda.** Philippe et al. (2005a)
466 address a well-known LBA artifact concerning the placement of Platyhelminthes
467 and Nematoda on the tree of Bilateria. Lartillot et al. (2007) revisit the same
468 dataset and provide two reduced, and overlapping alignments which contain 37
469 species for Nematoda and 32 species for Platyhelminthes, respectively. Both align-
470 ments have a length of 35371 amino acids. Figure 2 (a) and (b), S2 and S1 show
471 simplified and complete species trees, respectively.

472 **Dataset involving Microsporidia.** The dataset provided by Brinkmann et al.
473 (2005) comprises 40 species with 24294 amino acids. It contains an archaean out-
474 group and eukaryotic taxa. Of particular interest are the Microsporidia, a group
475 of unicellular parasites which lack mitochondria and instead possess mitosomes.
476 Microsporidia evolve fast, and site-homogeneous methods fail to correctly classify
477 them. Application of site-heterogeneous methods confirms that Microsporidia are
478 the closest sister species of Fungi (Brinkmann et al., 2005). For these reasons, the
479 dataset containing Microsporidia is ideal as a proof of concept for CAT-PMSF.
480 Figure 2 (c) and S3 show simplified and complete species trees, respectively.

481 **Metazoan Datasets.** The placement of Ctenophora on the tree of Metazoa is
482 still a matter of debate. We apply CAT-PMSF to two datasets. First, the alignment
483 provided by Ryan et al. (2013) contains 61 species with 88384 amino acids. Sec-
484 ond, the alignment provided by Simion et al. (2017) contains 97 species with 401632
485 amino acids. The complete set of outgroups comprises 2 Filasterea, 5 Ichthyosporea,
486 and 18 Choanoflagellata. The Choanoflagellata are the closest outgroup. Fig-
487 ure 3 shows results obtained from a reduced alignment in which we retained only
488 the Choanoflagellata. The reduced alignment yields 90 species. Figure 3, and
489 Figure S4 show simplified and complete species trees, respectively.

490 5 Acknowledgments

491 This work was supported by the Gordon and Betty Moore Foundation through
492 grant GBMF9741 to G.J.Sz and L.L.Sz. D.S. and G.J.Sz. received funding from
493 the European Research Council under the European Union’s Horizon 2020 Research
494 and Innovation Program (Grant Agreement No. 714774).

495 The authors thank Tom A. Williams, Edu Ocaña-Pallarès and László G. Nagy
496 for constructive criticism of the manuscript.

497 References

498 Johannes Bergsten. A review of long-branch attraction. *Cladistics*, 21(2):163–193,
499 4 2005. doi: 10.1111/j.1096-0031.2005.00059.x.

500 Henner Brinkmann, Mark Van Der Giezen, Yan Zhou, Gaëtan Poncelin De Rau-
501 court, and Hervé Philippe. An empirical assessment of long-branch attraction
502 artefacts in deep eukaryotic phylogenomics. *Systematic Biology*, 54(5):743–757,
503 2005. doi: 10.1080/10635150500234609.

504 Matthew W. Brown, Susan C. Sharpe, Jeffrey D. Silberman, Aaron A. Heiss,
505 B. Franz Lang, Alastair G. B. Simpson, and Andrew J. Roger. Phylogenomics
506 demonstrates that breviate flagellates are related to opisthokonts and apusomon-
507 ads. *Proceedings of the Royal Society B: Biological Sciences*, 280(1769):20131755,
508 10 2013. doi: 10.1098/rspb.2013.1755.

509 Johanna Taylor Cannon, Bruno Cossermelli Vellutini, Julian Smith, Fredrik Ron-
510 quist, Ulf Jondelius, and Andreas Hejnol. Xenacoelomorpha is the sister group
511 to nephrozoa. *Nature*, 530(7588):89–93, 2 2016. doi: 10.1038/nature16520.

512 Frédéric Delsuc, Henner Brinkmann, Daniel Chourrout, and Hervé Philippe. Tu-
513 nicates and not cephalochordates are the closest living relatives of vertebrates.
514 *Nature*, 439(7079):965–968, 2 2006. doi: 10.1038/nature04336.

515 Jean-François Dufayard, Laurent Duret, Simon Penel, Manolo Gouy, François
516 Rechenmann, and Guy Perrière. Tree pattern matching in phylogenetic trees: au-
517 tomatic search for orthologs or paralogs in homologous gene sequence databases.
518 21(11):2596–2603, 2005. doi: 10.1093/bioinformatics/bti325.

- 519 James S. Farris. Likelihood and inconsistency. *Cladistics*, 15(2):199–204, 6 1999.
520 doi: 10.1111/j.1096-0031.1999.tb00262.x.
- 521 Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimat-
522 ing evolutionary trees from data on discrete characters. *Systematic Biology*, 22
523 (3):240–249, 9 1973. doi: 10.1093/sysbio/22.3.240.
- 524 Joseph Felsenstein. Cases in which parsimony or compatibility methods will be
525 positively misleading. *Systematic Biology*, 27(4):401–410, 12 1978. doi: 10.1093/
526 sysbio/27.4.401.
- 527 Michael D. Hendy and David Penny. A framework for the quantitative study of
528 evolutionary trees. *Systematic Zoology*, 38(4):297, 12 1989. doi: 10.2307/2992396.
- 529 María José Jimenez, Miguel Arenas, and Ugo Bastolla. Substitution rates predicted
530 by stability-constrained models of protein evolution are not consistent with em-
531 pirical data. *Molecular Biology and Evolution*, 35(March):743–755, 2018. doi:
532 10.1093/molbev/msx327.
- 533 Thomas H. Jukes and Charles R. Cantor. Evolution of protein molecules. In H. N.
534 Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press,
535 1969.
- 536 John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov,
537 Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek,
538 Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J.
539 Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub
540 Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy,
541 Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Bergham-
542 mer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior,
543 Kavukcuoglu Koray, Kohli Pushmeet, and Hassabis Demis. Highly accurate
544 protein structure prediction with alphafold. 596(7873):583–589, 2021. doi:
545 10.1038/s41586-021-03819-2.
- 546 Subha Kalyaanamoorthy, Bui Qyuan Minh, Thomas K. F. Wong, Arndt von Hae-
547 seler, and Lars S. Jermin. Modelfinder: fast model selection for accurate phyloge-
548 netic estimates. *Nature Methods*, 14:587–589, 2017. doi: doi:10.1038/nmeth.4285.

- 549 Paschalia Kapli and Maximilian J. Telford. Topology-dependent asymmetry in sys-
550 tematic errors affects phylogenetic placement of ctenophora and xenacoelomor-
551 pha. *Science Advances*, 6(50):eabc5162, 12 2020. doi: 10.1126/sciadv.abc5162.
- 552 Jeffrey M. Koshi and Richard A. Goldstein. Context-dependent optimal substitu-
553 tion matrices. *Protein Engineering Design and Selection*, 8(7):641–645, 7 1995.
554 doi: 10.1093/protein/8.7.641.
- 555 Mary K. Kuhner and Joseph Felsenstein. A simulation comparison of phylogeny
556 algorithms under equal and unequal evolutionary rates. *Molecular biology and*
557 *evolution*, 11(3):459–468, 1994.
- 558 Nicolas Lartillot and Hervé Philippe. A bayesian mixture model for across-site
559 heterogeneities in the amino-acid replacement process. *Molecular Biology and*
560 *Evolution*, 21(6):1095–1109, 2004. doi: 10.1093/molbev/msh112.
- 561 Nicolas Lartillot, Henner Brinkmann, and Hervé Philippe. Suppression of long-
562 branch attraction artefacts in the animal phylogeny using a site-heterogeneous
563 model. *BMC Evolutionary Biology*, 7(SUPPL. 1):1–14, 2007. doi: 10.1186/
564 1471-2148-7-S1-S4.
- 565 Nicolas Lartillot, Nicolas Rodrigue, Daniel Stubbs, and Jacques Richer. PhyloBayes
566 MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel
567 Environment. *Systematic Biology*, 62(4):611–615, 04 2013. doi: 10.1093/sysbio/
568 syt022.
- 569 Si Quang Le and Olivier Gascuel. An improved general amino acid replacement
570 matrix. *Molecular Biology and Evolution*, 25(7):1307–1320, 4 2008. doi: 10.1093/
571 molbev/msn067.
- 572 Yuanning Li, Xing-Xing Shen, Benjamin Evans, Casey W. Dunn, and Antonis
573 Rokas. Rooting the animal tree of life. 2020. doi: [https://doi.org/10.1101/2020.](https://doi.org/10.1101/2020.10.27.357798)
574 [10.27.357798](https://doi.org/10.1101/2020.10.27.357798).
- 575 Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf,
576 Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. Iq-tree

577 2: New models and efficient methods for phylogenetic inference in the ge-
578 nomic era. *Molecular Biology and Evolution*, 37(5):1530–1534, 2 2020. doi:
579 10.1093/molbev/msaa015.

580 Masatoshi Nei. *Molecular Evolutionary Genetics*. 1987.

581 Hervé Philippe, Nicolas Lartillot, and Henner Brinkmann. Multigene analyses of
582 bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and
583 protostomia. *Molecular Biology and Evolution*, 22(5):1246–1253, 2005a. doi:
584 10.1093/molbev/msi111.

585 Hervé Philippe, Yan Zhou, Henner Brinkmann, Nicolas Rodrigue, and Frédéric
586 Delsuc. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolu-
587 tionary Biology*, 5:1–8, 2005b. doi: 10.1186/1471-2148-5-50.

588 Hervé Philippe, Romain Derelle, Philippe Lopez, Kerstin Pick, Carole Borchellini,
589 Nicole Boury-Esnault, Jean Vacelet, Emmanuelle Renard, Evelyn Houliston, Eric
590 Quéinnec, Corinne Da Silva, Patrick Wincker, Hervé Le Guyader, Sally Leys,
591 Daniel J. Jackson, Fabian Schreiber, Dirk Erpenbeck, Burkhard Morgenstern,
592 Gert Wörheide, and Michaël Manuel. Phylogenomics revives traditional views
593 on deep animal relationships. *Current Biology*, 19(8):706–712, 4 2009. doi: 10.
594 1016/j.cub.2009.02.052.

595 Hervé Philippe, Henner Brinkmann, Richard R. Copley, Leonid L. Moroz, Hiroaki
596 Nakano, Albert J. Poustka, Andreas Wallberg, Kevin J. Peterson, and Maximil-
597 ian J. Telford. Acoelomorph flatworms are deuterostomes related to xenoturbella.
598 *Nature*, 470(7333):255–258, 2 2011a. doi: 10.1038/nature09676.

599 Hervé Philippe, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood,
600 Michael Manuel, Gert Wörheide, and Denis Baurain. Resolving difficult phy-
601 logenetic questions: Why more sequences are not enough. *PLoS Biology*, 9(3):
602 e1000602, 3 2011b. doi: 10.1371/journal.pbio.1000602.

603 Matthew J. Phillips, Frédéric Delsuc, and David Penny. Genome-scale phylogeny
604 and the detection of systematic biases. *Molecular Biology and Evolution*, 21(7):
605 1455–1458, 7 2004. doi: 10.1093/molbev/msh137.

- 606 Le Si Quang, Olivier Gascuel, and Nicolas Lartillot. Empirical profile mixture mod-
607 els for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323, 8 2008.
608 doi: 10.1093/bioinformatics/btn445.
- 609 Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A
610 Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7.
611 67(5):901–904, 2018. doi: 10.1093/sysbio/syy032.
- 612 Jennifer Ripplinger and Jack Sullivan. Assessment of substitution model adequacy
613 using frequentist and bayesian methods. *Molecular Biology and Evolution*, 27
614 (12):2790–2803, 7 2010. doi: 10.1093/molbev/msq168.
- 615 Joseph F. Ryan, Kevin Pang, Christine E. Schnitzler, Anh-Dao Nguyen, R. Travis
616 Moreland, David K. Simmons, Bernard J. Koch, Warren R. Francis, Paul Havlak,
617 Stephen A. Smith, Nicholas H. Putnam, Steven H. D. Haddock, Casey W. Dunn,
618 Tyra G. Wolfsberg, James C. Mullikin, Mark Q. Martindale, and Andreas D.
619 Baxevanis. The genome of the ctenophore *mnemiopsis leidyi* and its implications
620 for cell type evolution. *Science*, 342(6164):1242592–1242592, 2013. doi: 10.1126/
621 science.1242592.
- 622 Reinhard Schneider, Antoine de Daruvar, and Chris Sander. The HSSP database
623 of protein structure-sequence alignments. 25(1):226–230, 1997. doi: 10.1093/
624 nar/25.1.226.
- 625 Dominik Schrempf. The elynx suite — a step towards reproducible research in
626 phylogenetics. *Manuscript in preparation*, 2021.
- 627 Dominik Schrempf, Nicolas Lartillot, and Gergely Szöllősi. Scalable empirical mix-
628 ture models that account for across-site compositional heterogeneity. *Molecular*
629 *Biology and Evolution*, 9 2020. doi: 10.1093/molbev/msaa145.
- 630 Hidetoshi Shimodaira. An approximately unbiased test of phylogenetic tree
631 selection. *Systematic Biology*, 51(3):492–508, May 2002. ISSN 1063-
632 5157. doi: 10.1080/10635150290069913. URL [http://dx.doi.org/10.1080/
633 10635150290069913](http://dx.doi.org/10.1080/10635150290069913).
- 634 Paul Simion, Hervé Philippe, Denis Baurain, Muriel Jager, Daniel J. Richter,
635 Arnaud Di Franco, Béatrice Roure, Nori Satoh, Éric Quéinnec, Alexander

- 636 Ereskovsky, Pascal Lapébie, Erwan Corre, Frédéric Delsuc, Nicole King, Gert
637 Wörheide, and Michaël Manuel. A large and consistent phylogenomic dataset
638 supports sponges as the sister group to all other animals. *Current Biology*, 27
639 (7):958–967, 2017. doi: 10.1016/j.cub.2017.02.031.
- 640 Edward Susko, Léa Lincker, and Andrew J Roger. Accelerated estimation of fre-
641 quency classes in site-heterogeneous profile mixture models. *Molecular Biology
642 and Evolution*, 35(February 2018):1–53, 2018. doi: 10.1093/molbev/msy026.
- 643 Yoshio Tateno, Naoko Takezaki, and Masatoshi Nei. Relative efficiencies of the
644 maximum-likelihood, neighbor-joining, and maximum-parsimony methods when
645 substitution rate varies with site. *Molecular Biology and Evolution*, 11(2):261–
646 277, 1994.
- 647 Simon Tavaré. Some probabilistic and statistical problems in the analysis of DNA
648 sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.
- 649 Antonella Tosoni, Manuela Nebuloni, Angelita Ferri, Sara Bonetto, Spinello Anti-
650 nori, Massimo Scaglia, Lihua Xiao, Hercules Moura, Govinda S. Visvesvara, Luca
651 Vago, and Giulio Costanzi. Disseminated microsporidiosis caused by encephal-
652 itozoon cuniculi iii (dog type) in an italian aids patient: a retrospective study.
653 *Modern pathology*, 15(5):577–583, 2002.
- 654 Huai Chun Wang, Karen Li, Edward Susko, and Andrew J. Roger. A class frequency
655 mixture model that adjusts for site-specific amino acid frequencies and improves
656 inference of protein phylogeny. *BMC Evolutionary Biology*, 8(1):1–13, 2008. doi:
657 10.1186/1471-2148-8-331.
- 658 Huai Chun Wang, Bui Quang Minh, Edward Susko, and Andrew J. Roger. Mod-
659 eling site heterogeneity with posterior mean site frequency profiles accelerates
660 accurate phylogenomic estimation. *Systematic Biology*, 67(2):216–235, 3 2018.
661 doi: 10.1093/sysbio/syx068.
- 662 Nathan V. Whelan and Keneth M. Halanych. Who let the cat out of the bag?
663 accurately dealing with substitutional heterogeneity in phylogenomic analyses.
664 *Systematic Biology*, 66(2):232–255, 9 2016. doi: 10.1093/sysbio/syw084.

- 665 Tom A. Williams, Cymon J. Cox, Peter G. Foster, Gergely J. Szöllősi, and T. Mar-
666 tin Embley. Phylogenomics provides robust support for a two-domains tree of
667 life. 4:138–147, 2020. doi: 10.1038/s41559-019-1040-x.
- 668 Ziheng Yang. Maximum-likelihood estimation of phylogeny from dna sequences
669 when substitution rates differ over sites. *Molecular Biology and Evolution*, 11
670 1993. doi: 10.1093/oxfordjournals.molbev.a040082.
- 671 So Wei Yeh, Jen Wei Liu, Sung Huan Yu, Chien Hua Shih, Jenn Kang Hwang,
672 and Julian Echave. Site-specific structural constraints on protein sequence evo-
673 lutionary divergence: Local packing density versus solvent exposure. *Molecular*
674 *Biology and Evolution*, 31(1):135–139, 2014. doi: 10.1093/molbev/mst178.
- 675 Andrey Zharkikh and Wen-Hsiung Li. Inconsistency of the maximum-parsimony
676 method: the case of five taxa with a molecular clock. *Systematic Biology*, 42(2):
677 113–125, 6 1993. doi: 10.1093/sysbio/42.2.113.