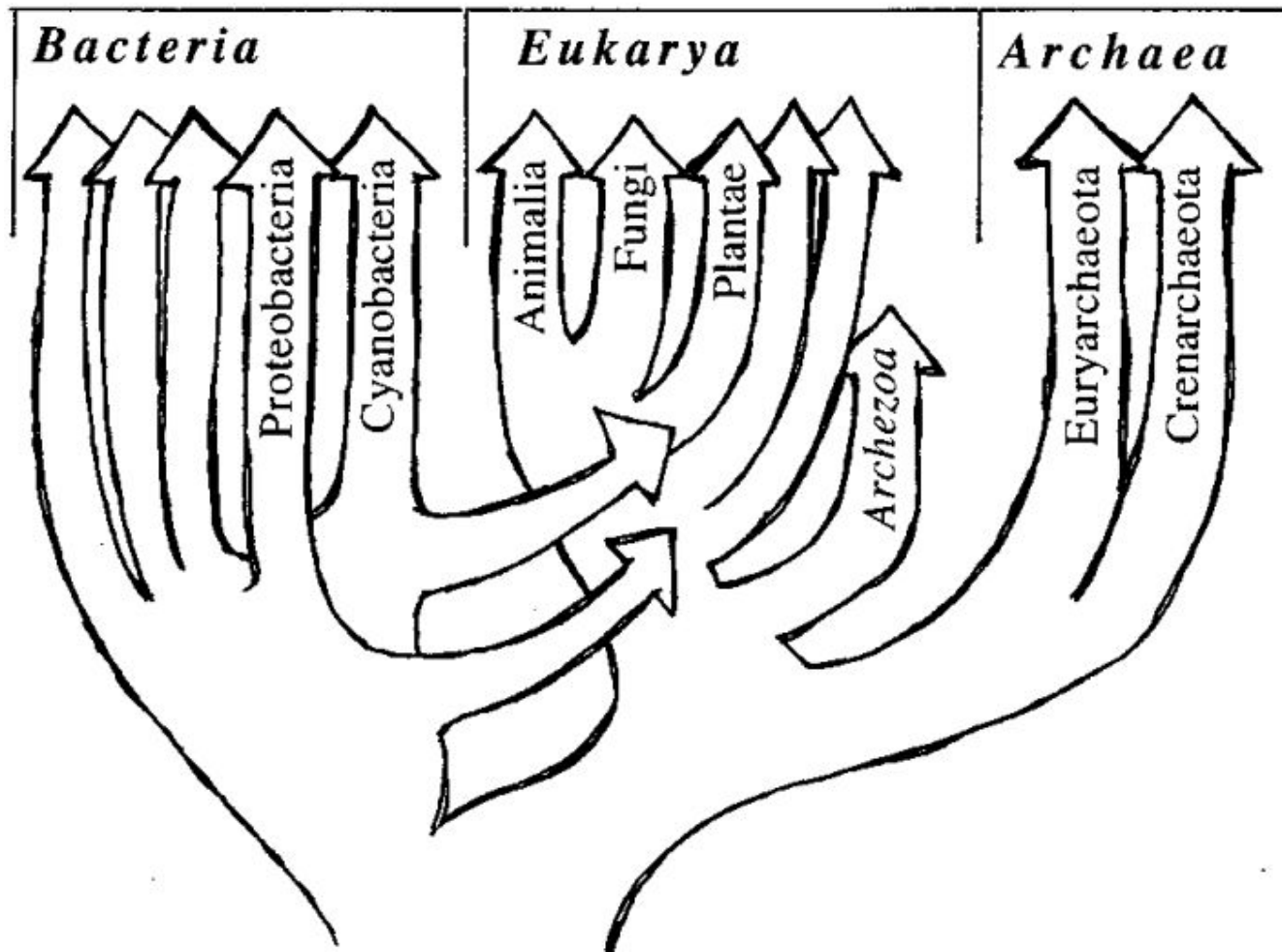


**Open lab:**  
**Bayesian methods for deep species tree inference**



## Bayesian methods for deep species tree inference

- The concatenate/supermatrix is a popular approach particularly in the field of deep species phylogenies
- ML methods (IQ-TREE/RAXML) are probably the best in the trade-off between accuracy / computational time
- If one doesn't have a particular preference ML vs Bayesian, why to use Bayesian tools then?
  - Results confirmed (or not) with an alternative robust method
  - Posterior probabilities: an alternative assessment of branch support
  - Some complex models are unavailable in ML framework: the CAT model
- Bayesian methods are computationally challenging ... Few tools scale well with large amount of data ( $n^\circ$  of sites). One of these tools is Phylobayes

## CAT : Empirical profile mixture models for phylogenetic reconstruction.

Le S.Q., Gascuel O., Lartillot N. Bioinformatics. 2008 Oct 15;24(20):2317-23.

Please cite [THESE](#) papers if you use CAT.

---

CAT ([Lartillot and Philippe 2004](#)) is a model especially devised to account for **site-specific features of protein evolution**. In general, each position of a protein is under a very specific selective constraint, and as a result, only a subset of the 20 amino-acids is likely to be accepted at this position during evolutionary times. As we have shown in previous works, accounting for such site specific features is crucial, both to obtain a better statistical fit ([Lartillot and Philippe 2006](#)), and to alleviate phylogenetic artefacts, due to **long branch attraction** phenomena ([Lartillot et al 2007](#)). Technically, **CAT is a mixture model, assuming a given number  $K$  of components (or site classes). Each component specifies a biochemical profile, which is a probability vector over the 20 amino-acids.** Such a profile in turns defines a very simple amino-acid replacement process : each time a substitution event occurs, a new amino-acid is chosen at random, according to the probabilities defined by the profile. We call this a *Poisson* process, although it is also known as a *Felsenstein1981*, or *proportional*, amino-acid replacement process. The likelihood at each site of the alignment is then an average over all available Poisson processes defined by the mixture.

## CAT : Empirical profile mixture models for phylogenetic reconstruction.

Le S.Q., Gascuel O., Lartillot N. Bioinformatics. 2008 Oct 15;24(20):2317-23.

Please cite [THESE](#) papers if you use CAT.

---

### Papers

- If you use **CAT**, please cite:  
*"Empirical profile mixture models for phylogenetic reconstruction."*  
**Le S.Q., Gascuel O., Lartillot N.**  
Bioinformatics. 2008 Oct 15;24(20):2317-23.  
Click [here](#) to download supplementary information.
- *"Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model."*  
**Lartillot N., Brinkmann H., Philippe H.**  
BMC Evolutionary Biology. 2007 Feb 8;7 Suppl 1:S4.
- *"Computing Bayes factors using thermodynamic integration."*  
**Lartillot N., Philippe H.**  
Systematic Biology. 2006 55:195-207.
- *"A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process."*  
**Lartillot N., Philippe H.**  
Molecular Biology and Evolution. 2004 21(6):1095-1109.



# A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process

*Nicolas Lartillot and Hervé Philippe*

Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec Canada

Most current models of sequence evolution assume that all sites of a protein evolve under the same substitution process, characterized by a  $20 \times 20$  substitution matrix. Here, we propose to relax this assumption by developing a Bayesian mixture model that allows the amino-acid replacement pattern at different sites of a protein alignment to be described by distinct substitution processes. Our model, named CAT, assumes the existence of distinct processes (or classes) differing by their equilibrium frequencies over the 20 residues. Through the use of a Dirichlet process prior, the total number of classes and their respective amino-acid profiles, as well as the affiliations of each site to a given class, are all free variables of the model. In this way, the CAT model is able to adapt to the complexity actually present in the data, and it yields an estimate of the substitutional heterogeneity through the posterior mean number of classes. We show that a significant level of heterogeneity is present in the substitution patterns of proteins, and that the standard one-matrix model fails to account for this heterogeneity. By evaluating the Bayes factor, we demonstrate that the standard model is outperformed by CAT on all of the data sets which we analyzed. Altogether, these results suggest that **the complexity of the pattern of substitution of real sequences is better captured by the CAT model**, offering the possibility of studying its impact on phylogenetic reconstruction and its connections with structure-function determinants.

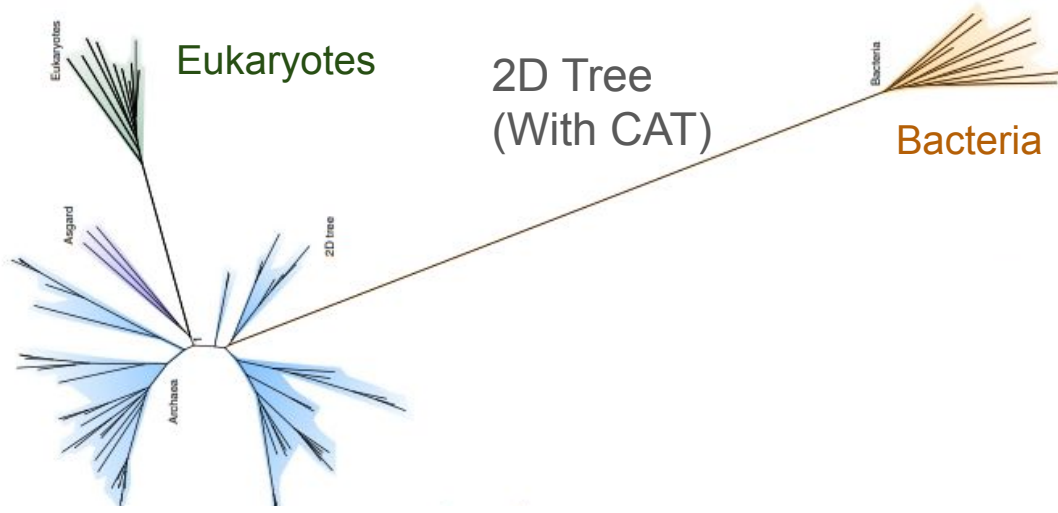
[http://www.atgc-montpellier.fr/download/papers/cat\\_2004.pdf](http://www.atgc-montpellier.fr/download/papers/cat_2004.pdf)

# Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model

Nicolas Lartillot<sup>1</sup>, Henner Brinkmann<sup>2</sup>, Hervé Philippe<sup>2</sup>

**Conclusions:** The CAT model is more robust than WAG against LBA artefacts, essentially because it correctly anticipates the high probability of convergences and reversions implied by the small effective size of the amino-acid alphabet at each site of the alignment. More generally, our results provide strong evidence that site-specificities in the substitution process need be accounted for in order to obtain more reliable phylogenetic trees, although other evolutionary heterogeneities, such as compositional biases and heterotachy, should also be handled.

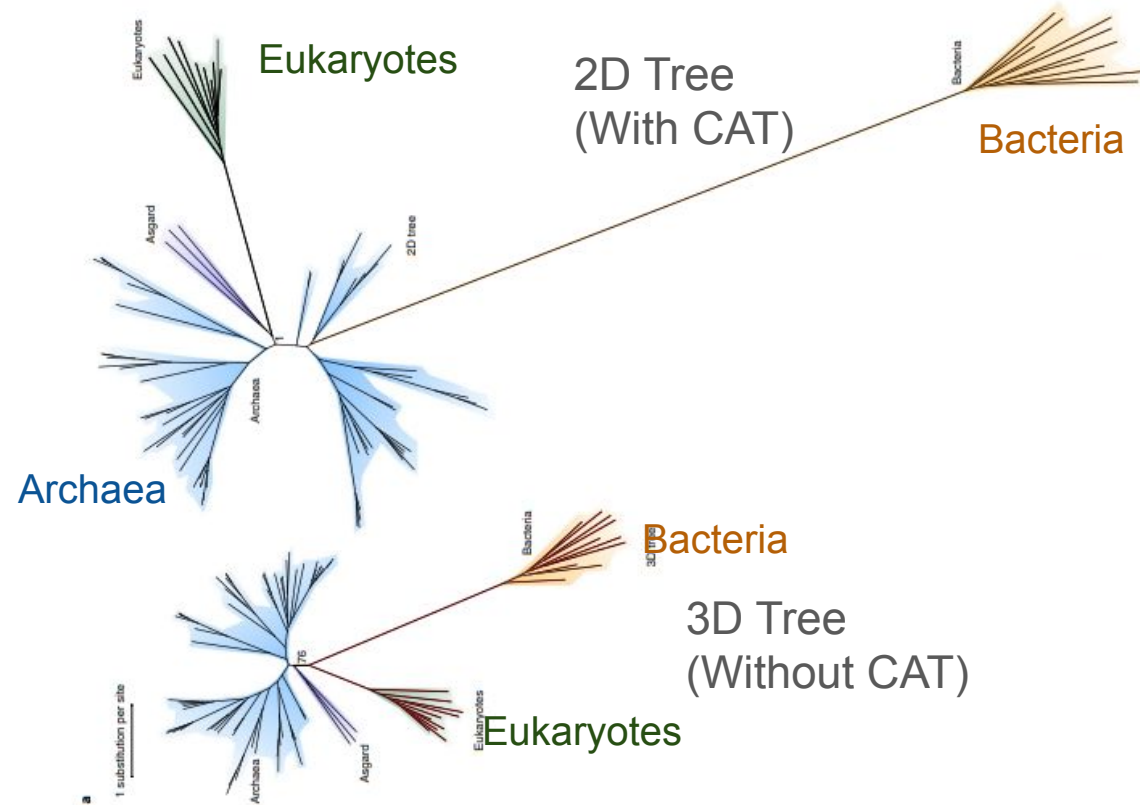
[http://www.atgc-montpellier.fr/download/papers/cat\\_2007.pdf](http://www.atgc-montpellier.fr/download/papers/cat_2007.pdf)



## Phylogenomics provides robust support for a two-domains tree of life

[Tom A. Williams](#) , [Cymon J. Cox](#), [Peter G. Foster](#), [Gergely J. Szöllősi](#) & [T. Martin Embley](#) 

*Nature Ecology & Evolution* **4**, 138–147 (2020) | [Cite this article](#)

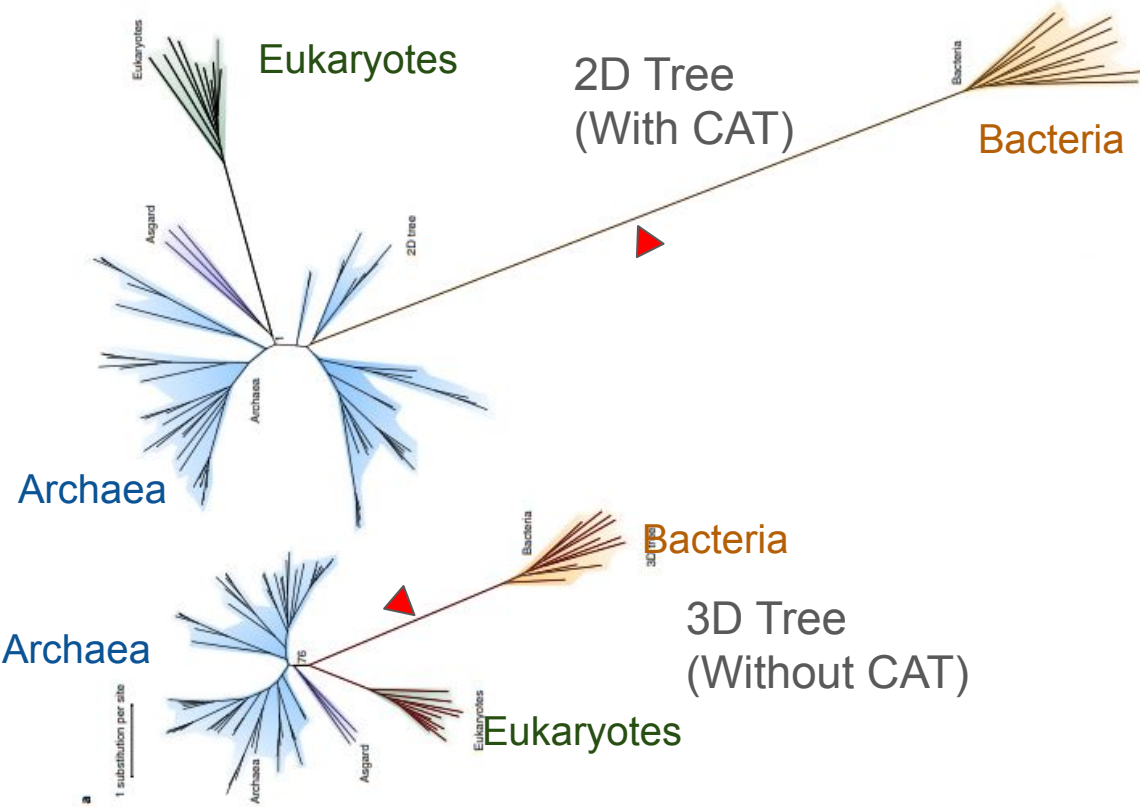


## Phylogenomics provides robust support for a two-domains tree of life

[Tom A. Williams](#) , [Cymon J. Cox](#), [Peter G. Foster](#), [Gergely J. Szöllősi](#) & [T. Martin Embley](#) 

*Nature Ecology & Evolution* **4**, 138–147 (2020) | [Cite this article](#)

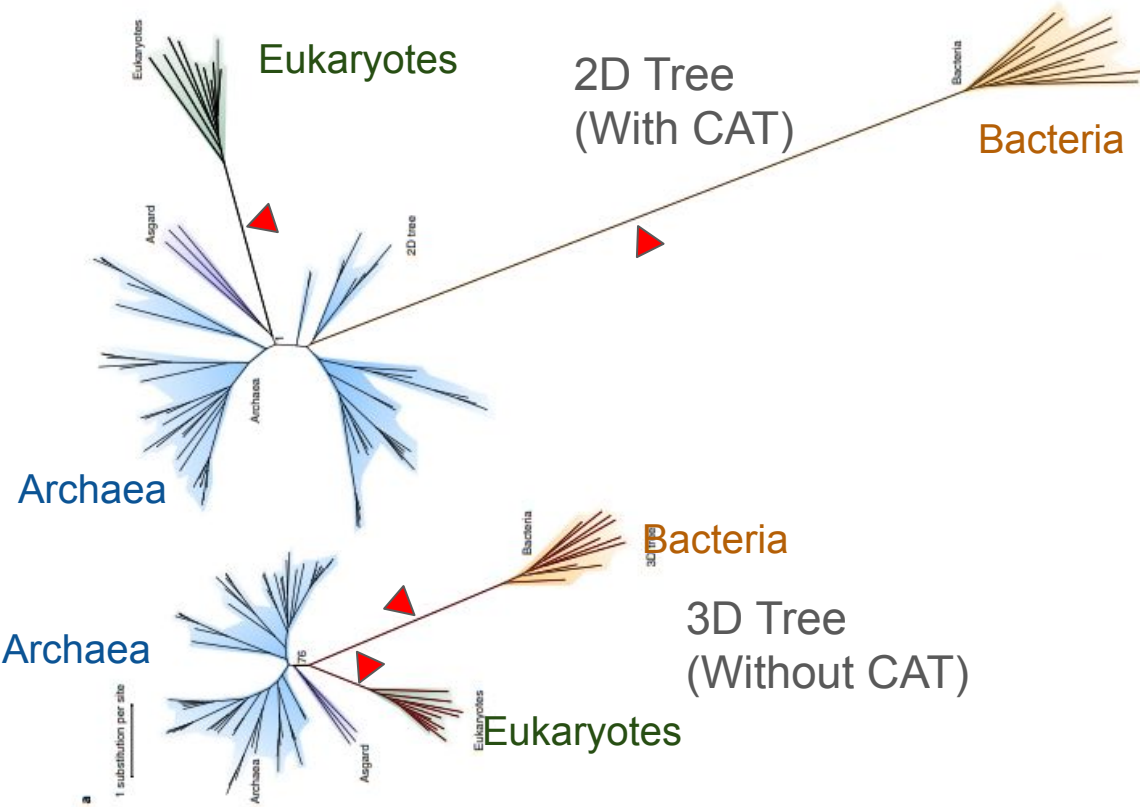




## Phylogenomics provides robust support for a two-domains tree of life

[Tom A. Williams](#) , [Cymon J. Cox](#), [Peter G. Foster](#), [Gergely J. Szöllősi](#) & [T. Martin Embley](#) 

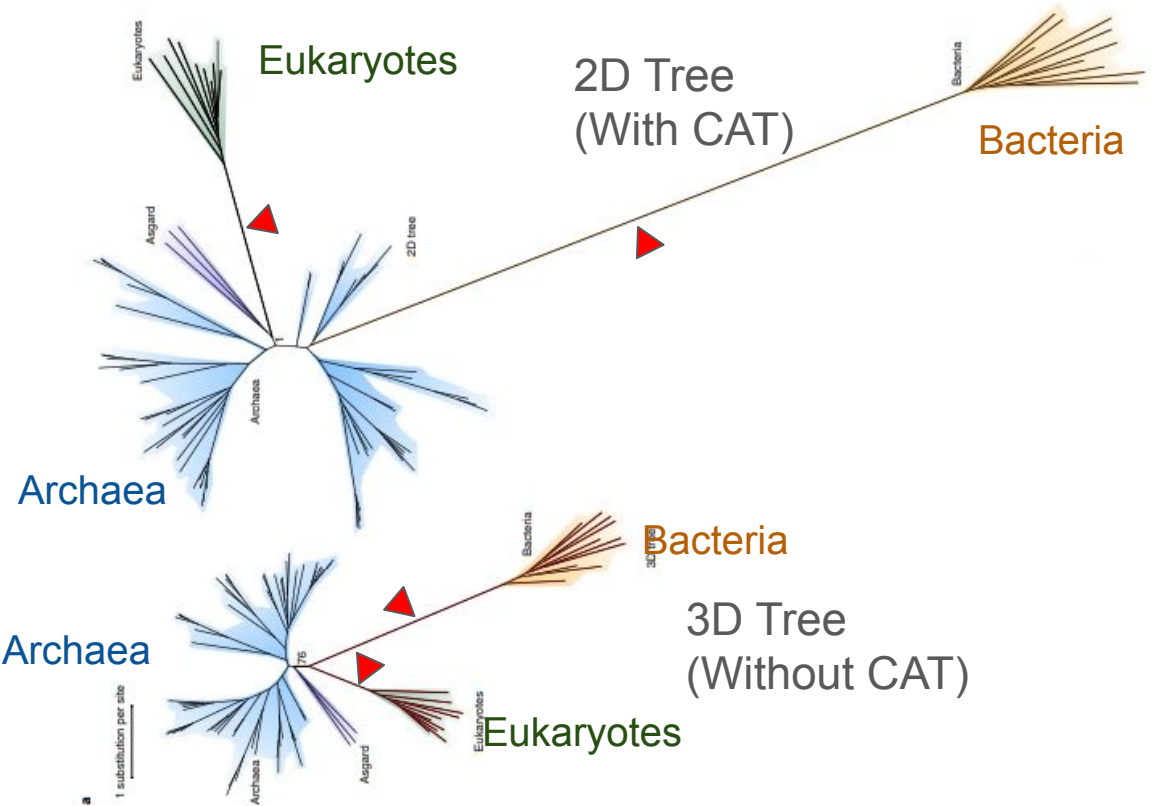
*Nature Ecology & Evolution* **4**, 138–147 (2020) | [Cite this article](#)



## Phylogenomics provides robust support for a two-domains tree of life

Tom A. Williams [✉](#), Cymon J. Cox, Peter G. Foster, Gergely J. Szöllősi & T. Martin Embley [✉](#)

*Nature Ecology & Evolution* 4, 138–147 (2020) | [Cite this article](#)

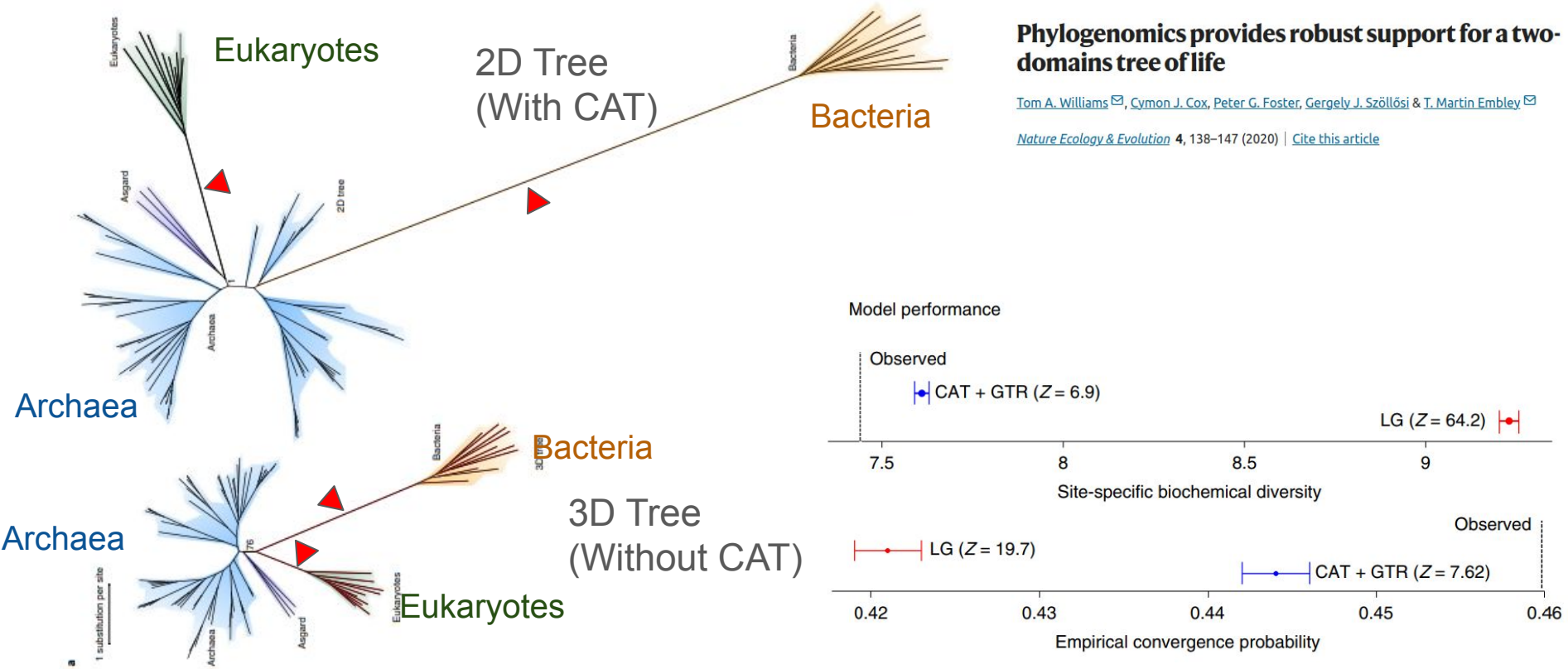


## Phylogenomics provides robust support for a two-domains tree of life

Tom A. Williams [✉](#), Cymon J. Cox, Peter G. Foster, Gergely J. Szöllősi & T. Martin Embley [✉](#)

*Nature Ecology & Evolution* 4, 138–147 (2020) | [Cite this article](#)

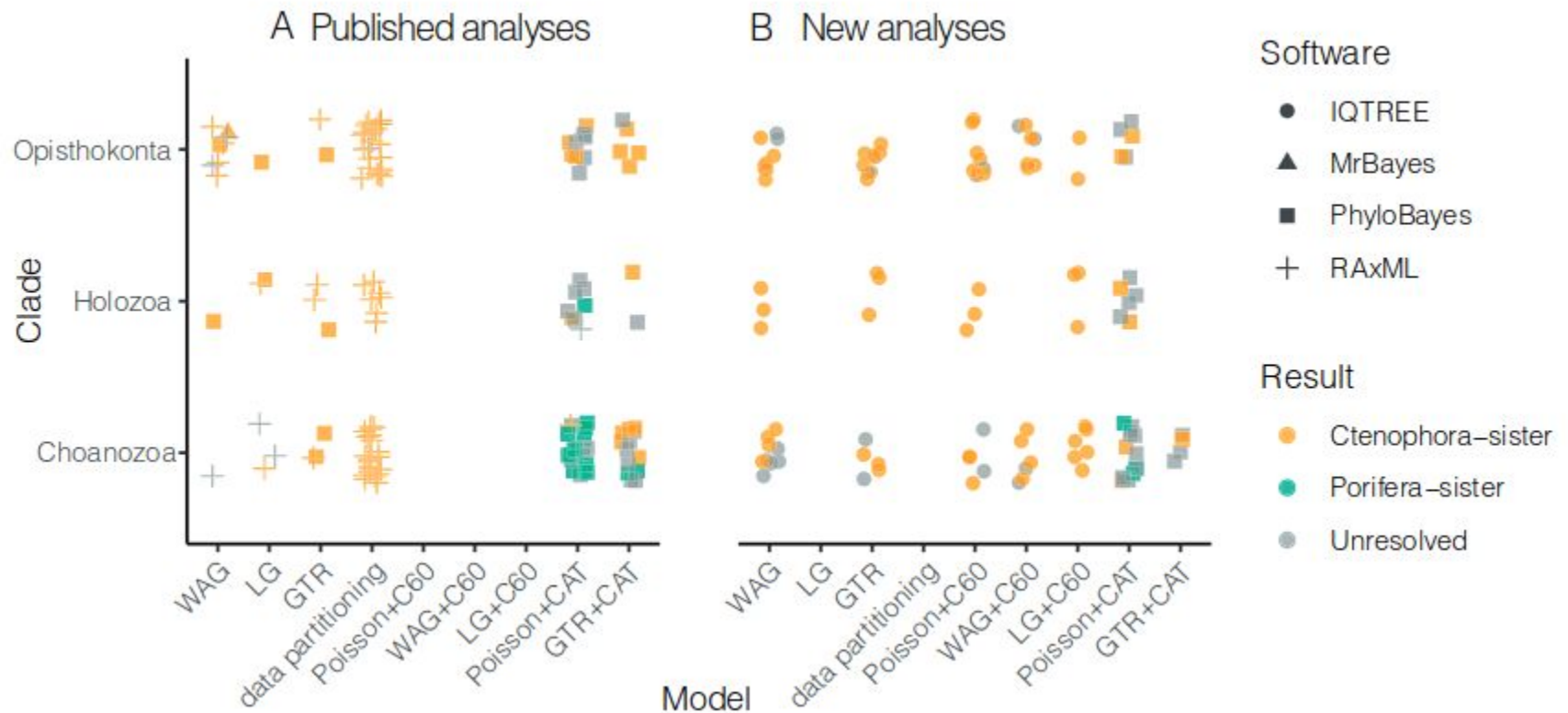
**Fig. 2 | Evidence that the 3D tree is an artefact of long-branch attraction.** **a**, Da Cunha et al.<sup>22</sup> analysed a dataset of 35 core protein-coding genes under the LG + G4 + F model and obtained a 3D tree; the better-fitting (Supplementary Table 4) CAT + GTR + G4 model recovers a 2D tree. **b**, Posterior predictive tests indicate that CAT + GTR + G4 performs significantly better than LG + G4 + F in capturing the site-specific evolutionary constraints reflected by lower biochemical diversity approaching that of the empirical data. This results in more realistic estimates of substitutional saturation and convergence found in the data. The longest branches on both the 3D and 2D trees in **a** are the stems leading to the bacteria and eukaryotes (in yellow and green, respectively). CAT + GTR + G4 identifies many more convergent substitutions on these branches than does LG + G4 + F, as can be seen by comparing the branch lengths in **a**. This failure to detect convergent substitutions under LG + G4 + F has the effect of drawing the bacterial and eukaryotic branches together because convergences are mistaken for homologies (synapomorphies), resulting in a 3D tree. Bootstrap support (**a**) and Bayesian posterior probability (**b**) are indicated for the key nodes defining the 3D and 2D trees. Asgard refers to a clade of Heimdallarchaeota and *Lokiarchaeum*. Plotting these trees to the same scale (in terms of substitutions per site) illustrates major differences in these analyses. The 3D/LG + G4 + F analysis suggests that, on average, 30.77 changes have taken place per site; the 2D/CAT + GTR + G4 analysis suggests that 47.4 changes per site have occurred. This difference amounts to ~128,511 additional substitutions in total inferred under the CAT + GTR + G4 model.







**Fig. 2 | Evidence that the 3D tree is an artefact of long-branch attraction.** **a**, Da Cunha et al.<sup>22</sup> analysed a dataset of 35 core protein-coding genes under the LG + G4 + F model and obtained a 3D tree; the better-fitting (Supplementary Table 4) CAT + GTR + G4 model recovers a 2D tree. **b**, Posterior predictive tests indicate that CAT + GTR + G4 performs significantly better than LG + G4 + F in capturing the site-specific evolutionary constraints reflected by lower biochemical diversity approaching that of the empirical data. This results in more realistic estimates of substitutional saturation and convergence found in the data. The longest branches on both the 3D and 2D trees in **a** are the stems leading to the bacteria and eukaryotes (in yellow and green, respectively). CAT + GTR + G4 identifies many more convergent substitutions on these branches than does LG + G4 + F, as can be seen by comparing the branch lengths in **a**. This failure to detect convergent substitutions under LG + G4 + F has the effect of drawing the bacterial and eukaryotic branches together because convergences are mistaken for homologies (synapomorphies), resulting in a 3D tree. Bootstrap support (**a**) and Bayesian posterior probability (**b**) are indicated for the key nodes defining the 3D and 2D trees. Asgard refers to a clade of Heimdallarchaeota and *Lokiarchaeum*. Plotting these trees to the same scale (in terms of substitutions per site) illustrates major differences in these analyses. The 3D/LG + G4 + F analysis suggests that, on average, 30.77 changes have taken place per site; the 2D/CAT + GTR + G4 analysis suggests that 47.4 changes per site have occurred. This difference amounts to ~128,511 additional substitutions in total inferred under the CAT + GTR + G4 model.



# CAT vs no CAT: Porifera vs Ctenophora



## Rooting the Animal Tree of Life

Yuanning Li <sup>1,2</sup> Xing-Xing Shen <sup>2,3</sup> Benjamin Evans,<sup>4</sup> Casey W. Dunn <sup>1,\*</sup> and Antonis Rokas <sup>2,\*</sup>



# Bayesian Cross-Validation Comparison of Amino Acid Replacement Models: Contrasting Profile Mixtures, Pairwise Exchangeabilities, and Gamma-Distributed Rates-Across-Sites

Thomas Bujaki<sup>1</sup> · Nicolas Rodrigue<sup>1,2</sup>

Received: 20 July 2022 / Accepted: 21 September 2022 / Published online: 7 October 2022  
© The Author(s) 2022

## Abstract

Models of amino acid replacement are central to modern phylogenetic inference, particularly so when dealing with deep evolutionary relationships. Traditionally, a single, empirically derived matrix was utilized, so as to keep the degrees-of-freedom of the inference low, and focused on topology. With the growing size of data sets, however, an amino acid-level general-time-reversible matrix has become increasingly feasible, treating amino acid exchangeabilities and frequencies as free parameters. Moreover, models based on mixtures of multiple matrices are increasingly utilized, in order to account for across-site heterogeneities in amino acid requirements of proteins. Such models exist as finite empirically-derived amino acid profile (or frequency) mixtures, free finite mixtures, as well as free Dirichlet process-based infinite mixtures. All of these approaches are typically combined with a gamma-distributed rates-across-sites model. In spite of the availability of these different aspects to modeling the amino acid replacement process, no study has systematically quantified their relative contributions to their predictive power of real data. Here, we use Bayesian cross-validation to establish a detailed comparison, while activating/deactivating each modeling aspect. For most data sets studied, we find that amino acid mixture models can outrank all single-matrix models, even when the latter include gamma-distributed rates and the former do not. We also find that free finite mixtures consistently outperform empirical finite mixtures. Finally, the Dirichlet process-based mixture model tends to outperform all other approaches.

## CAT vs no CAT: Overfitting?

# Bayesian Cross-Validation Comparison of Amino Acid Replacement Models: Contrasting Profile Mixtures, Pairwise Exchangeabilities, and Gamma-Distributed Rates-Across-Sites

**Table 1** Cross-validation scores

	Broughton	Brown	Delsuc	Lartillot-2007	Lartillot-2012
F81	$-3641.6 \pm 60.4$	$-22518.4 \pm 96.3$	$-20009.6 \pm 342.4$	$-24521.0 \pm 482.9$	$-5579.5 \pm 95.9$
C60-Poisson	$-1520.8 \pm 83.2$	$4341.4 \pm 44.8$	$-5084.2 \pm 128.2$	$-5881.2 \pm 139.5$	$-2733.2 \pm 120.4$
C60-GTR	$-912.4 \pm 77.2$	$-2003.2 \pm 207.9$	$-3454.0 \pm 139.0$	$-3996.6 \pm 240.3$	$-871.5 \pm 77.3$
UDM <sub>256</sub> -Poisson	$-616.7 \pm 69.1$	$1227.2 \pm 52.2$	$36.4 \pm 91.0$	$-95.6 \pm 206.4$	$-1524.5 \pm 87.3$
UDM <sub>256</sub> -GTR	$-238.9 \pm 64.7$	$1746.6 \pm 73.4$	$589.2 \pm 121.2$	$610.6 \pm 249.1$	$-75.0 \pm 45.0$
CAT <sub>f=100</sub> -Poisson	$-66.7 \pm 39.6$	$2859.8 \pm 37.8$	$2028.2 \pm 137.1$	$1692.6 \pm 214.0$	$-69.3 \pm 65.4$
CAT <sub>f=90</sub> -GTR	$216.5 \pm 47.7$	$3650.0 \pm 176.8$	$2684.4 \pm 126.0$	$2878.2 \pm 229.5$	<b><math>560.4 \pm 31.1</math></b>
CAT <sub>f=100</sub> -GTR	$262.4 \pm 32.7$	$3617.4 \pm 205.8$	$2716.8 \pm 127.5$	$2878.0 \pm 274.7$	$607.0 \pm 18.1$
CAT-Poisson	$-78.3 \pm 37.6$	$2988.8 \pm 44.8$	$2228.2 \pm 147.0$	$1852.8 \pm 211.1$	$-65.2 \pm 70.1$
CAT-GTR	$251.9 \pm 34.9$	$3315.6 \pm 63.8$	$2961.0 \pm 151.6$	$3096.8 \pm 249.7$	<b><math>610.6 \pm 15.8</math></b>
F81+ $\Gamma$	$-2194.9 \pm 52.5$	$-15321.4 \pm 215.7$	$-13044.0 \pm 213.4$	$-16612.6 \pm 441.8$	$-3607.9 \pm 65.3$
C60-Poisson+ $\Gamma$	$-170.9 \pm 44.7$	$2227.6 \pm 48.5$	$1435.8 \pm 107.5$	$1061.2 \pm 165.0$	$-1038.1 \pm 84.0$
C60-GTR+ $\Gamma$	$253.6 \pm 16.6$	$3193.2 \pm 34.0$	$2686.4 \pm 105.2$	$2869.6 \pm 162.8$	$370.5 \pm 34.8$
UDM <sub>256</sub> -Poisson+ $\Gamma$	$-39.9 \pm 35.2$	$2962.8 \pm 46.9$	$2306.0 \pm 123.6$	$1996.4 \pm 166.5$	$-868.8 \pm 83.7$
UDM <sub>256</sub> -GTR+ $\Gamma$	$323.4 \pm 20.6$	$3651.2 \pm 41.6$	$3204.0 \pm 139.8$	$3473.6 \pm 169.9$	$427.0 \pm 41.7$
CAT <sub>f=100</sub> -Poisson+ $\Gamma$	$65.0 \pm 40.1$	$3294.2 \pm 62.7$	$2627.2 \pm 158.6$	$2529.6 \pm 186.3$	$13.7 \pm 58.4$
CAT <sub>f=40</sub> -GTR+ $\Gamma$	<b><math>353.1 \pm 23.6</math></b>	$3599.4 \pm 27.2$	$3237.0 \pm 102.6$	$3563.8 \pm 179.7$	$590.0 \pm 36.4$
CAT <sub>f=100</sub> -GTR+ $\Gamma$	<b><math>374.9 \pm 20.8</math></b>	$3789.2 \pm 43.1$	$3446.8 \pm 109.7$	$3824.8 \pm 203.7$	<b><math>624.0 \pm 31.9</math></b>
CAT-Poisson+ $\Gamma$	$57.5 \pm 40.6$	$3404.8 \pm 56.1$	$2820.0 \pm 150.5$	$2638.8 \pm 188.1$	$30.5 \pm 66.5$
CAT-GTR+ $\Gamma$	<b><math>370.9 \pm 24.2</math></b>	<b><math>3943.2 \pm 36.8</math></b>	<b><math>3678.6 \pm 134.5</math></b>	<b><math>4069.8 \pm 196.2</math></b>	<b><math>619.7 \pm 36.6</math></b>

Models with an instance of the highest performance in at least one replicate are displayed in bold. For empirical mixtures, only results for the top-performing model are displayed. For free finite mixture models, results for 100 components are displayed, as well as any free finite mixture having the best performance on at least one of five replicates



## CAT vs no CAT: Overfitting?

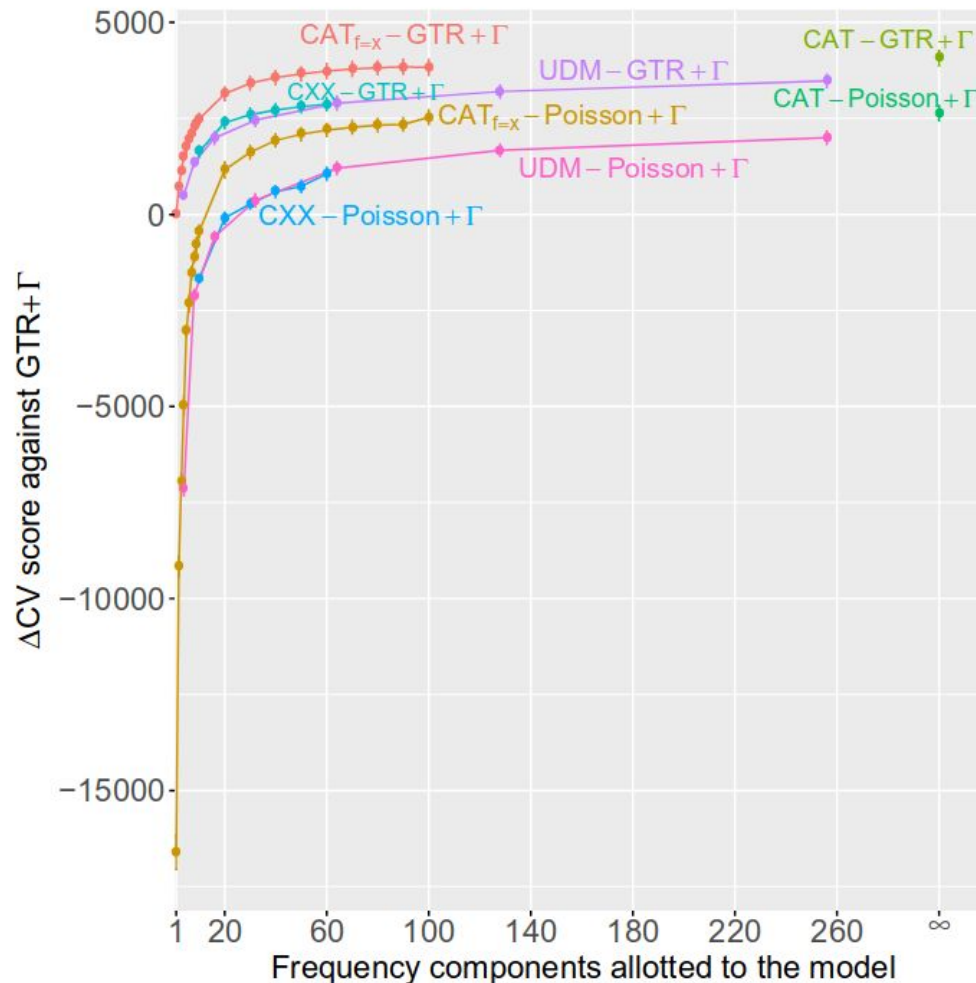
# Bayesian Cross-Validation Comparison of Amino Acid Replacement Models: Contrasting Profile Mixtures, Pairwise Exchangeabilities, and Gamma-Distributed Rates-Across-Sites

**Table 2** Number of replicates where a model had the best performance

	Broughton	Brown	Delsuc	Lartillot-2007	Lartillot-2012
CAT-GTR+ $\Gamma$	2	5	5	5	2
CAT-GTR					1
CAT <sub><math>f=100</math></sub> GTR+ $\Gamma$	2				1
CAT <sub><math>f=40</math></sub> GTR+ $\Gamma$	1				
CAT <sub><math>f=90</math></sub> GTR					1

## CAT vs no CAT: Overfitting?

### Bayesian Cross-Validation Comparison of Amino Acid Replacement Models: Contrasting Profile Mixtures, Pairwise Exchangeabilities, and Gamma-Distributed Rates-Across-Sites



## CAT vs no CAT: Overfitting?

### 6 Cross-Validation

Cross-validation (CV) is a general method for evaluating the fit of alternative models. The rationale is as follows: the dataset is randomly split into two (possibly unequal) parts, the training (or learning) set and the test set. The parameters of the model are estimated on the learning set (i.e. the model is 'trained' on this subset of empirical observations), and these parameter values are then used to compute the likelihood of the test set (which measures how well the test set is 'predicted' by the model). The overall procedure has to be repeated (and the resulting log likelihood scores averaged) over several random splits.

[See section 6 in the manual](#)



## CAT vs no CAT: Overfitting?

### 6 Cross-Validation

Cross-validation (CV) is a general method for evaluating the fit of alternative models. The rationale is as follows: the dataset is randomly split into two (possibly unequal) parts, the training (or learning) set and the test set. The parameters of the model are estimated on the learning set (i.e. the model is 'trained' on this subset of empirical observations), and these parameter values are then used to compute the likelihood of the test set (which measures how well the test set is 'predicted' by the model). The overall procedure has to be repeated (and the resulting log likelihood scores averaged) over several random splits.

On the post-burn-in cycles, we used PhyloBayes to compute site-specific likelihood values over the sample on the testing data set, taking the averages for each site, and finally summing the logarithm of these site-specific likelihood posterior averages to produce the cross-validation score of each replicate. Supposing a sample of  $K$  (post-

**Bayesian Cross-Validation Comparison of Amino Acid Replacement Models: Contrasting Profile Mixtures, Pairwise Exchangeabilities, and Gamma-Distributed Rates-Across-Sites**

## CAT vs no CAT: Overfitting?

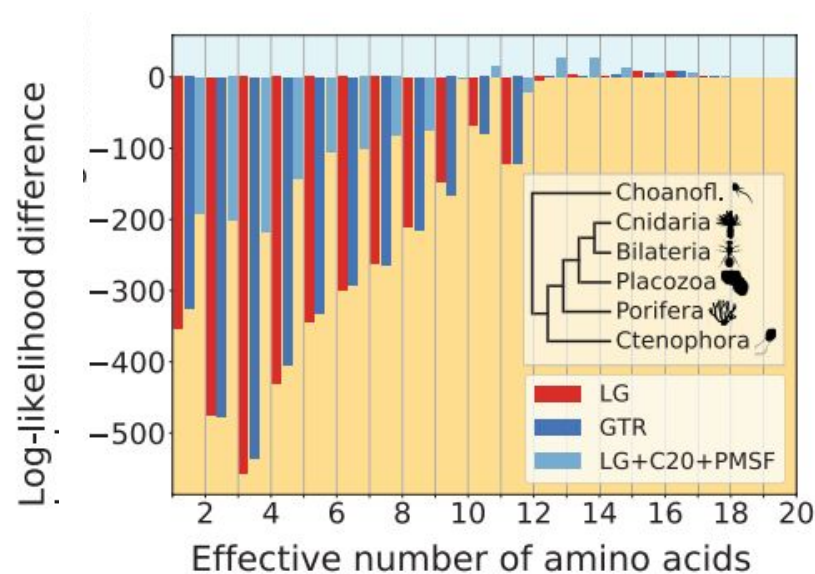
### 6 Cross-Validation

Cross-validation (CV) is a general method for evaluating the fit of alternative models. The rationale is as follows: the dataset is randomly split into two (possibly unequal) parts, the training (or learning) set and the test set. The parameters of the model are estimated on the learning set (i.e. the model is 'trained' on this subset of empirical observations), and these parameter values are then used to compute the likelihood of the test set (which measures how well the test set is 'predicted' by the model). The overall procedure has to be repeated (and the resulting log likelihood scores averaged) over several random splits.




Note that this measure automatically takes into account dimensionality issues and will not intrinsically favor models that have more parameters. Intuitively, overfit means that a model focusses too much on irrelevant (random) features of the training set. By definition, these random features will not be consistently reproduced in the test set, and thus, an overfitted model will typically show less good performance once evaluated on the test dataset.

[See section 6 in the manual](#)

# The long branch attraction problem: Porifera vs Ctenophora

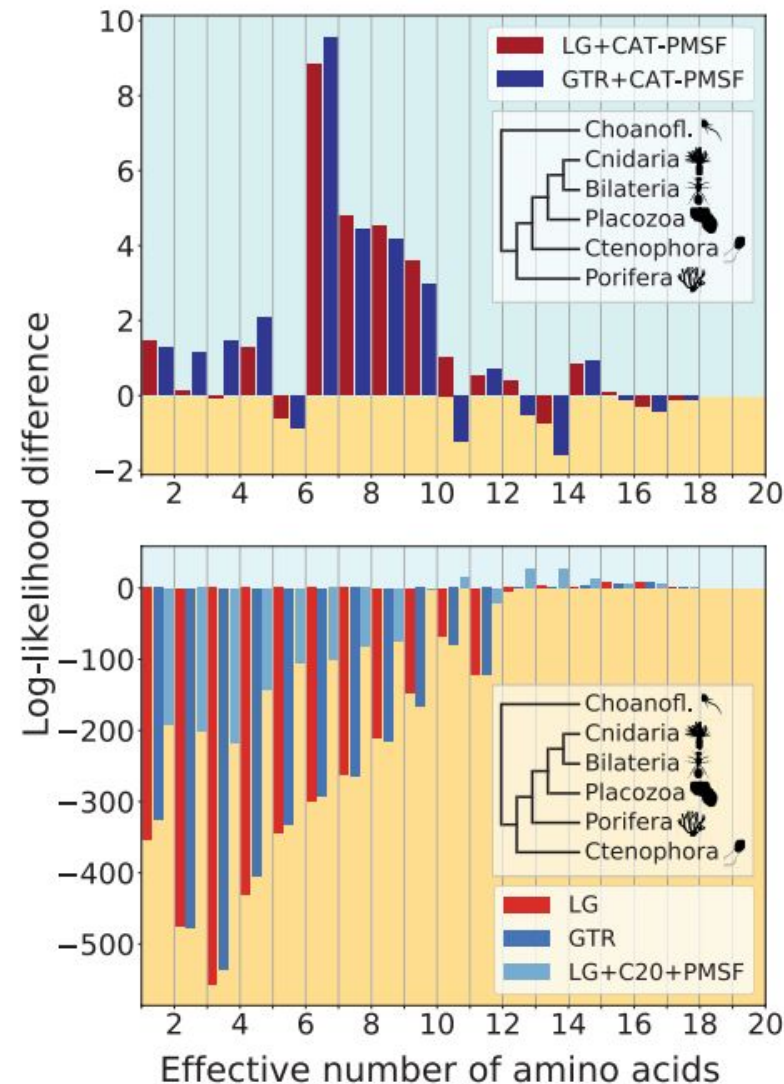


## Compositionally Constrained Sites Drive Long-Branch Attraction



LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup>, NICOLAS LARTILLOT<sup>4</sup>, GERGELY J. SZÖLLÖSI<sup>1,2,3,\*</sup> AND DOMINIK SCHREMPF<sup>1,\*</sup>



# The long branch attraction problem: Porifera vs Ctenophora



## Compositionally Constrained Sites Drive Long-Branch Attraction

LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup> , NICOLAS LARTILLOT<sup>4</sup> , GERGELY J. SZÖLLÖSI<sup>1,2,3,\*</sup>  AND DOMINIK SCHREMPF<sup>1,\*</sup> 

# CAT vs no CAT: The long branch attraction problem

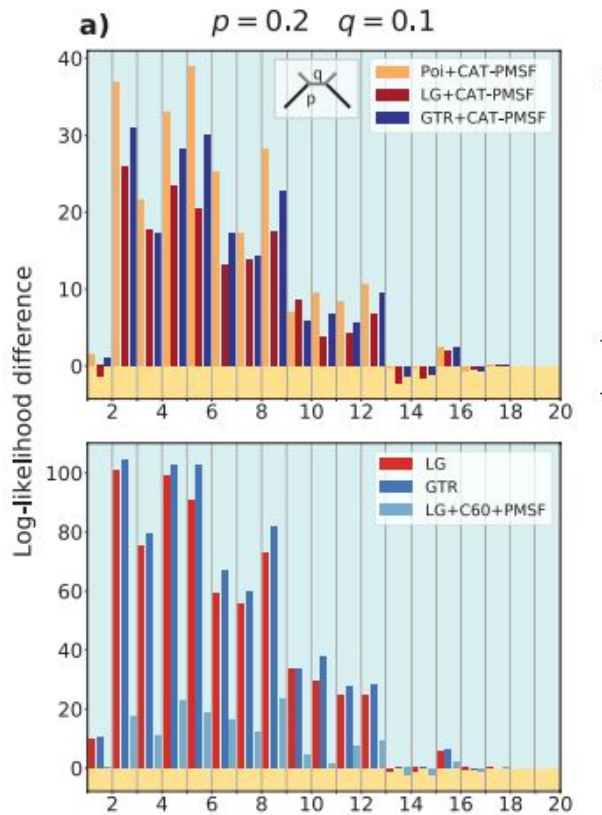






FIGURE 2. Highly constrained sites drive long-branch attraction artifacts in the Felsenstein zone. We simulated amino acid alignments with 10,000 sites exhibiting across-site compositional heterogeneity (Schrempf et al. 2020) along Felsenstein-type trees (insets in the top row; Felsenstein 1978) with 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 Poisson (Felsenstein 1973; Nei 1987), the LG (Le and Gascuel 2008), and the GTR (Tavaré 1986) models constrained to the correct topology as well as to an incorrect topology (inset in the bottom row; Farris 1999) with IQ-TREE 2 (Minh et al. 2020). The site-specific log-likelihood differences between the maximum likelihood trees of the two competing topologies binned according to the site-specific effective number of amino acids are shown. A positive value (blue background) indicates support for the true topology, a negative value (yellow background) indicates support for the incorrect topology exhibiting long-branch attraction. The LG and GTR models incorrectly infer Farris-type trees if  $p \geq 0.8$ .

## Compositionally Constrained Sites Drive Long-Branch Attraction

LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup> , NICOLAS LARTILLOT<sup>4</sup> , GERGELY J. SZÖLLŐSI<sup>1,2,3,\*</sup>  AND DOMINIK SCHREMPF<sup>1,\*</sup> 



# CAT vs no CAT: The long branch attraction problem

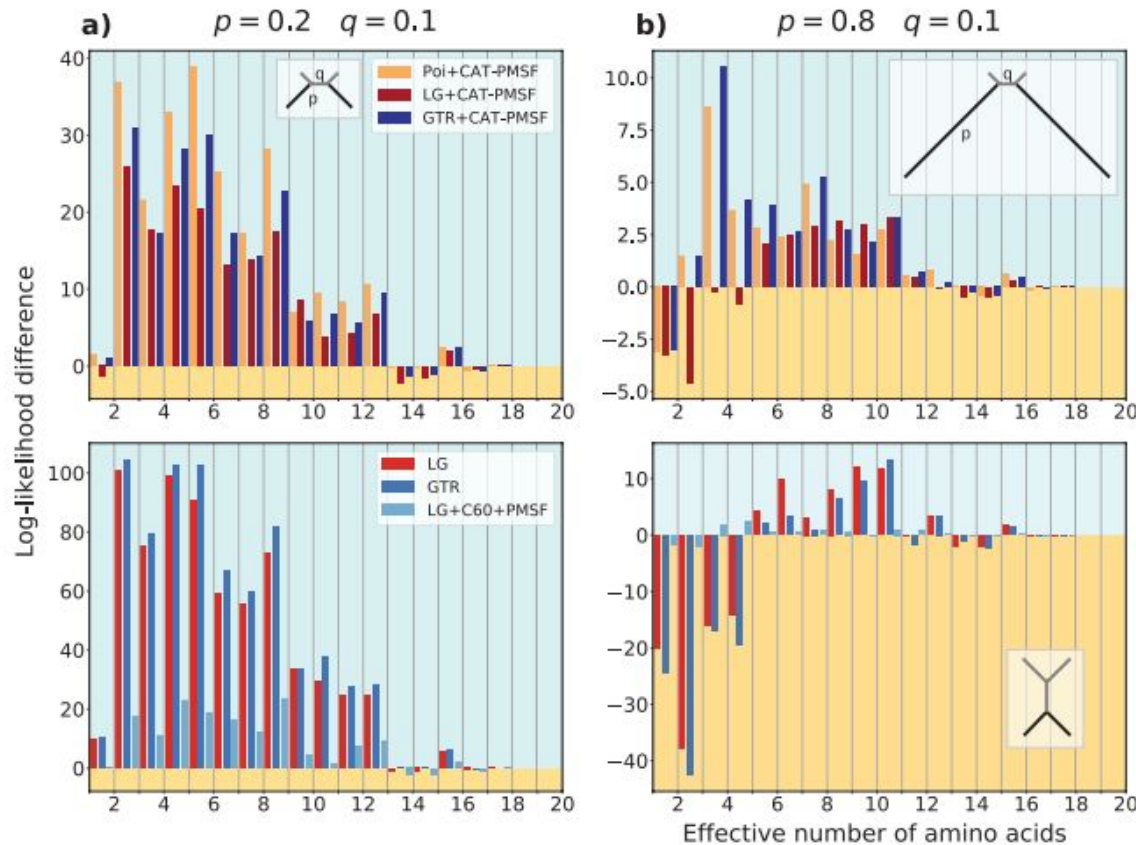


FIGURE 2. Highly constrained sites drive long-branch attraction artifacts in the Felsenstein zone. We simulated amino acid alignments with 10,000 sites exhibiting across-site compositional heterogeneity (Schrempf et al. 2020) along Felsenstein-type trees (insets in the top row; Felsenstein 1978) with 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 Poisson (Felsenstein 1973; Nei 1987), the LG (Le and Gascuel 2008), and the GTR (Tavaré 1986) models constrained to the correct topology as well as to an incorrect topology (inset in the bottom row; Farris 1999) with IQ-TREE 2 (Minh et al. 2020). The site-specific log-likelihood differences between the maximum likelihood trees of the two competing topologies binned according to the site-specific effective number of amino acids are shown. A positive value (blue background) indicates support for the true topology, a negative value (yellow background) indicates support for the incorrect topology exhibiting long-branch attraction. The LG and GTR models incorrectly infer Farris-type trees if  $p \geq 0.8$ .

## Compositionally Constrained Sites Drive Long-Branch Attraction

LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup>, NICOLAS LARTILLOT<sup>4</sup>, GERGELY J. SZÖLLÖSI<sup>1,2,3,\*</sup> AND DOMINIK SCHREMPF<sup>1,\*</sup>

# CAT vs no CAT: The long branch attraction problem

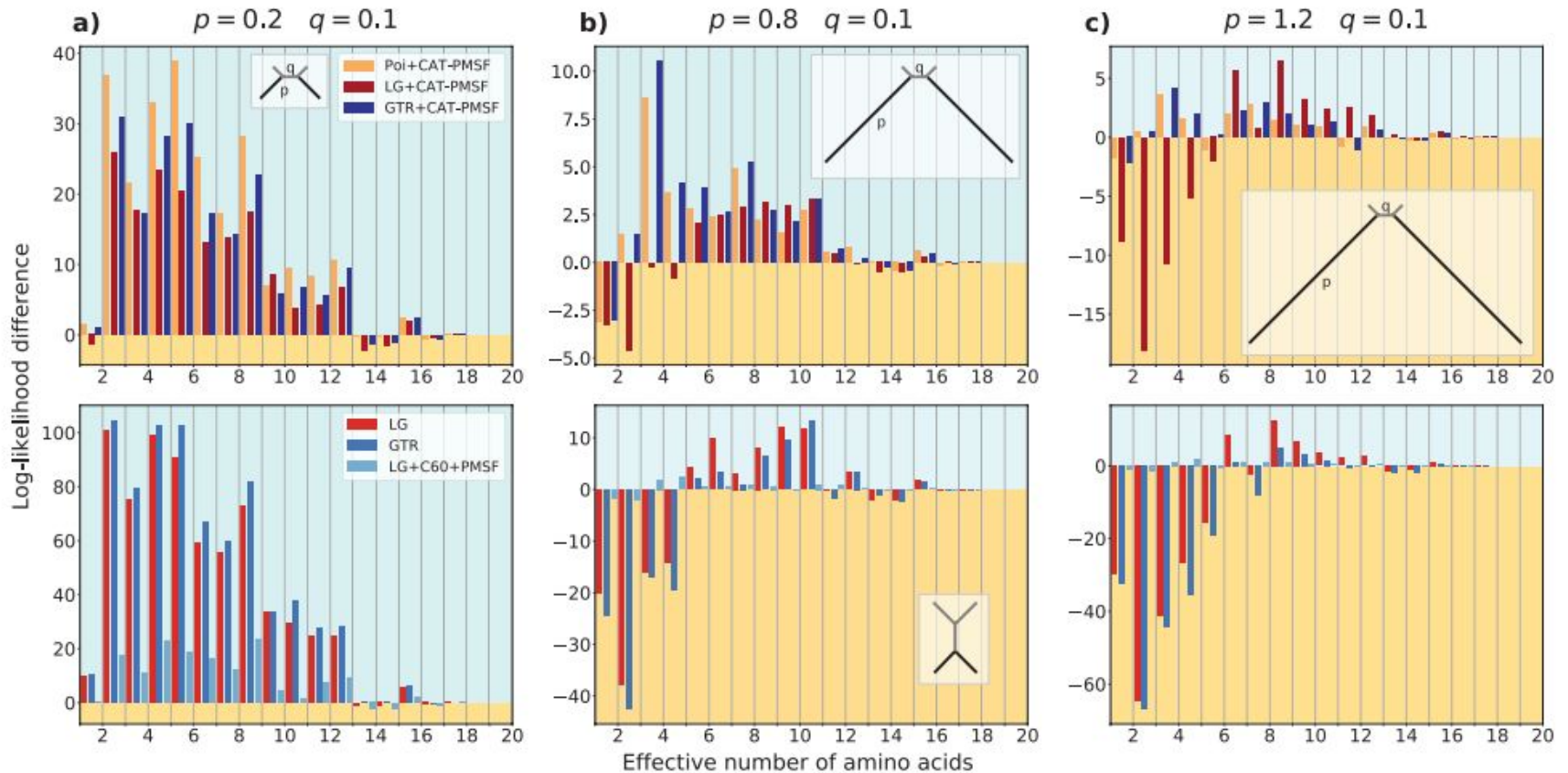


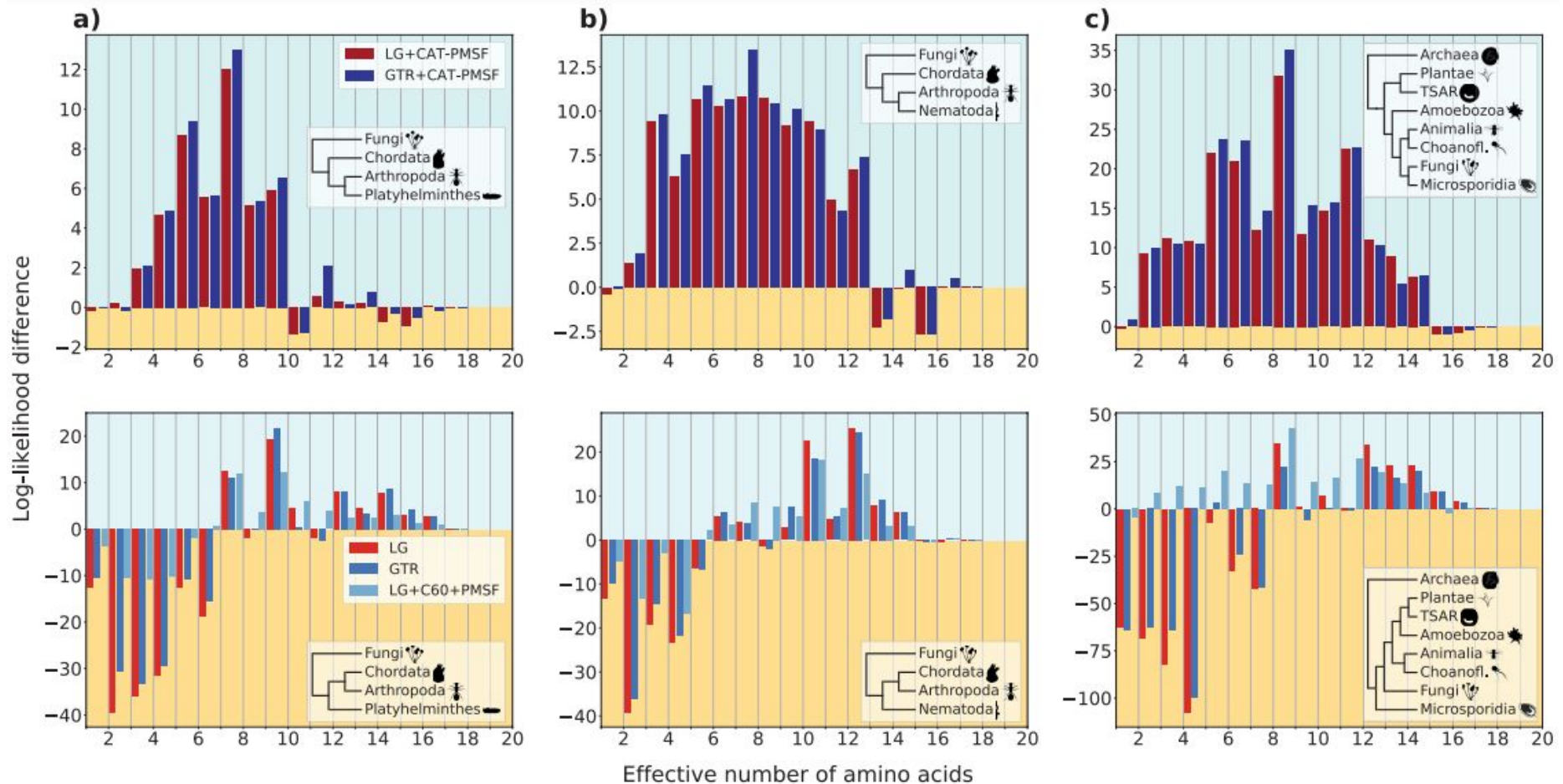
FIGURE 2. Highly constrained sites drive long-branch attraction artifacts in the Felsenstein zone. We simulated amino acid alignments with 10,000 sites exhibiting across-site compositional heterogeneity (Schrempf et al. 2020) along Felsenstein-type trees (insets in the top row; Felsenstein 1978) with 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 Poisson (Felsenstein 1973; Nei 1987), the LG (Le and Gascuel 2008), and the GTR (Tavaré 1986) models constrained to the correct topology as well as to an incorrect topology (inset in the bottom row; Farris 1999) with IQ-TREE 2 (Minh et al. 2020). The site-specific log-likelihood differences between the maximum likelihood trees of the two competing topologies binned according to the site-specific effective number of amino acids are shown. A positive value (blue background) indicates support for the true topology, a negative value (yellow background) indicates support for the incorrect topology exhibiting long-branch attraction. The LG and GTR models incorrectly infer Farris-type trees if  $p \geq 0.8$ .

## Compositionally Constrained Sites Drive Long-Branch Attraction



LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup>, NICOLAS LARTILLOT<sup>4</sup>, GERGELY J. SZÖLLÖSI<sup>1,2,3,\*</sup> AND DOMINIK SCHREMPF<sup>1,\*</sup>



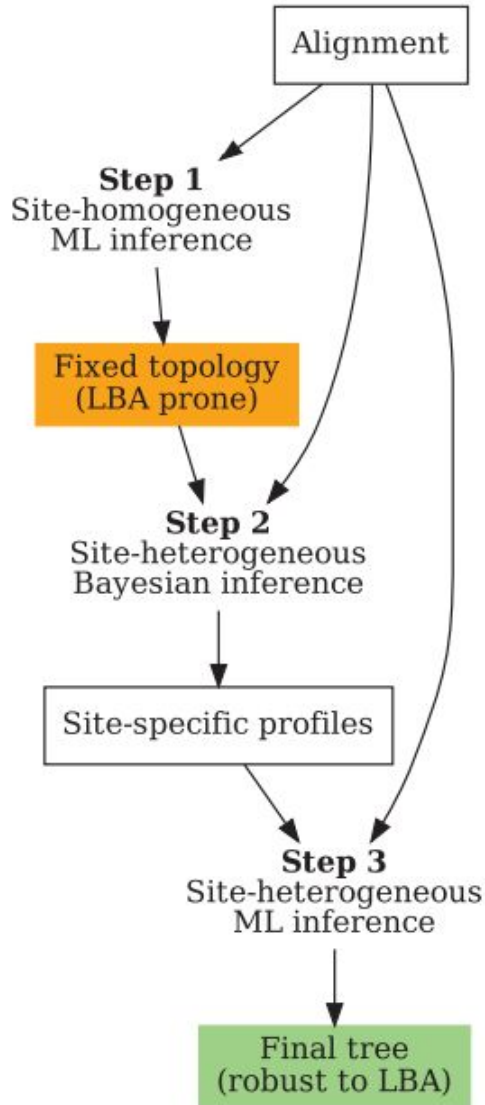
# CAT vs no CAT: The long branch attraction problem



## Compositionally Constrained Sites Drive Long-Branch Attraction




LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup> , NICOLAS LARTILLOT<sup>4</sup> , GERGELY J. SZÖLLÖSI<sup>1,2,3,\*</sup>  AND DOMINIK SCHREMPF<sup>1,\*</sup> 

## CAT model in ML framework: CAT-PMSF



<https://github.com/drenal/cat-pmsf-paper/tree/main>

### Compositionally Constrained Sites Drive Long-Branch Attraction

LÉNÁRD L. SZÁNTHÓ<sup>1,2,3,\*</sup> , NICOLAS LARTILLOT<sup>4</sup> , GERGELY J. SZÖLLŐSI<sup>1,2,3,\*</sup>  AND DOMINIK SCHREMPF<sup>1,\*</sup> 

CAT-PMSF = PMSF run (IQ-TREE) using the site profiles -rates and compositions- sampled with the CAT model (Phylobayes)

## Site-specific frequency models

Starting with version 1.5.0, IQ-TREE provides a new posterior mean site frequency (PMSF) model as a rapid approximation to the time and memory consuming profile mixture models C10 to C60 (Le et al., 2008a; a variant of PhyloBayes' CAT model). The PMSF are the amino-acid profiles for each alignment site computed from an input mixture model and a guide tree. The PMSF model is much faster and requires much less RAM than C10 to C60 (see table below), regardless of the number of mixture classes. Our extensive simulations and empirical phylogenomic data analyses demonstrate that the PMSF models can effectively ameliorate long branch attraction artefacts.

If you use this model in a publication please cite:

**H.C. Wang, B.Q. Minh, S. Susko and A.J. Roger** (2018) Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.*, 67:216-235. <https://doi.org/10.1093/sysbio/syx068>

Here is an example of computation time and RAM usage for an Obazoa data set (68 sequences, 43615 amino-acid sites) from Brown et al. (2013) using 16 CPU cores:

Models	CPU time	Wall-clock time	RAM usage
LG+F+G	43h:38m:23s	3h:37m:23s	1.8 GB
LG+C20+F+G	584h:25m:29s	46h:39m:06s	38.8 GB
LG+C60+F+G	1502h:25m:31s	125h:15m:29s	112.8 GB
LG+PMSF+G	73h:30m:37s	5h:7m:27s	2.2 GB

## Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation

HUAI-CHUN WANG<sup>1,2,3</sup>, BUI QUANG MINH<sup>4</sup>, EDWARD SUSKO<sup>1,3</sup>, AND ANDREW J. ROGER<sup>2,3,\*</sup>



## Phylobayes (tutorials)

- Phylobayes: [tutorial for non-mpi version](#)
- Phylobayes (parallel computing): [tutorial for mpi version](#)
- Step-by-step practical introduction: [\*PhyloBayes: Bayesian Phylogenetics Using Site-heterogeneous Models\*](#)

## Phylobayes (how to run a chain)

- `pb_mpi_icc -cat -gtr -d MSA.phylip -T TREE.nw chain_name`
  - 'pb\_mpi\_icc' for mpi, otherwise 'pb'
  - -cat: CAT model
  - -gtr: GTR model
  - -d: input MSA. It has to be in phylip format
  - -T: topology to constrain tree inference <- optional (e.g., for CAT-PMSF)
  - chain\_name: name given to the chain. Can be any name. Usually people run two chains, ideally more chains should run if dataset is complex (it usually is)

For further information see 'running a chain' section in the tutorials

## Phylobayes (convergence assessment)

[See section 3.2 in the manual](#)

*“Generally, a run under PhyloBayes provides good results for a total number of points of the order of 10 000 to 30 000, although again, this really depends on the datasets.”*

## Phylobayes (convergence assessment)

```
bpcomp -x 1000 10 <chain1> <chain2>
```

Here, using a burn-in of 1000, and sub-sampling every 10 trees, the `bpcomp` program will output the largest (`maxdiff`) and mean (`meandiff`) discrepancy observed across all bipartitions. It will also produce a file (`bpcomp.con.tre`) with the consensus obtained by pooling all the trees of the chains given as arguments.

Note that `bpcomp` can be run on a single chain (in which case it will simply produce the consensus of all trees after burn-in). However, using `bpcomp` on multiple chains usually results in more stable MCMC estimates of the posterior consensus tree.

Some guidelines:

- $\text{maxdiff} < 0.1$ : good run.
- $\text{maxdiff} < 0.3$ : acceptable: gives a good qualitative picture of the posterior consensus.
- $0.3 < \text{maxdiff} < 1$ : the sample is not yet sufficiently large, and the chains have not converged, but this is on the right track.
- if  $\text{maxdiff} = 1$  even after 10,000 points, this indicates that at least one of the runs is stuck in a local maximum.

[See section 3.2 in the manual](#)

*“Generally, a run under PhyloBayes provides good results for a total number of points of the order of 10 000 to 30 000, although again, this really depends on the datasets.”*



# Phylobayes (convergence assessment)

iter	time	topo	loglik	length	alpha	Nmode	statent	statalpha	rrent	rrmean		
0	0	0	-24816728.96		21.03440888	1	12	2.609582388	20	4.846542237	0.9789432874	
1	102.663	68	-14938895.09		47.30268994	0.7032755141	24	2.428581517	19.83146126	5.010751946	1.104091772	
2	107.461	69	-14484436.28		63.52103493	0.6965370084	36	2.406584817	19.17682061	5.004314059	0.9837841347	
3	108.464	68	-14200837.66		74.10647911	0.6804169489	53	2.36840494	17.7892709	4.989205088	1.054846996	
4	109.133	67	-13957482.21		82.13904775	0.6781909863	64	2.346440434	17.02456772	4.970272722	0.9483297879	
5	108.816	66	-13818810.89		88.70843007	0.6707846456	71	2.316263339	16.34504322	4.952039101	0.9954768919	
6	112.347	67	-13696846.88		92.87217491	0.6697586554	81	2.293644481	15.51841521	4.935247644	0.9360783702	
7	114.216	67	-13565222.16		96.40457272	0.6791375646	90	2.280649006	14.27870158	4.917740444	1.024921606	
8	113.79	66	-13344163.54		98.98936664	0.6783715098	112	2.262438308	13.9235364	4.901699867	1.064848813	
9	117.886	66	-13282029.7		100.2657998	0.6875833193	143	2.239272874	13.5360229	4.888706498	1.013636292	
10	119.548	65	-13164562.61		101.4262875	0.699427432	162	2.216661036	12.43812108	4.876897468	0.972924239	
11	119.15	65	-13016621.62		103.0806337	0.7143763391	199	2.186342971	12.44453652	4.865087209	1.095977136	
12	116.426	66	-12877836.15		104.9087596	0.734205057	216	2.1560795	11.44117899	4.853763221	1.030928462	
13	118.082	66	-12825641.73		105.9723256	0.7533143589	243	2.133626486	10.75089686	4.847514729	0.9501537838	
14	122.461	66	-12785982.04		107.3104918	0.7664960451	269	2.11237619	10.3767552	4.840715901	1.078626762	
15	124.033	65	-12691719.16		107.5932328	0.7917682439	300	2.090852982	9.999152597	4.834203637	0.9106905798	
16	125.012	65	-12658244.34		110.0151941	0.8083505838	304	2.073920177	9.850512115	4.830421535	0.9236124413	
17	129.132	65	-12631282.37		112.4521593	0.8229264894	331	2.056160522	9.679694104	4.825829785	1.091898872	
18	127.473	64	-12539003.11		114.1437839	0.8418767269	339	2.045422338	9.452308412	4.819782618	1.053363008	
19	127.784	64	-12486870.21		114.4701117	0.8655256335	369	2.033010747	9.716045644	4.818190981	0.9035808909	
20	129.449	64	-12467381.34		115.5257564	0.8890852984	383	2.017141961	9.245915399	4.817431442	0.8869008654	
21	127.933	65	-12449540.93		116.9055839	0.9080263589	415	2.003894768	8.850847593	4.814717668	1.031635407	
22	127.28	65	-12386623.35		118.2592583	0.9252138624	428	1.994612362	8.688964371	4.809493033	0.9122955715	
23	129.706	65	-12284896.65		120.1322911	0.9397635086	441	1.988583443	8.659772038	4.806869259	0.9986886477	
24	130.212	64	-12260530.99		121.0856437	0.9552285368	449	1.977823285	8.773522691	4.806465236	1.016538085	
25	131.285	64	-12237546.2		121.4090526	0.973296348	478	1.970391324	8.541553853	4.802632378	1.043708125	
26	132.497	63	-12155679.49		121.8743574	0.9890490654	497	1.971389912	9.10237348	4.79641444	1.126574093	
27	128.23	62	-12110590.16		120.9728054	1.014319489	500	1.964096468	8.886074602	4.794610589	1.048662475	
28	130.625	62	-12084733		120.1799888	1.034061488	517	1.964479875	8.69355605	4.789474101	1.035845787	
29	134.746	62	-12064272.9		123.0537704	1.045962559	533	1.960732937	8.879037474	4.784519198	1.016993877	
30	137.252	62	-12047945.38		121.7527363	1.060978081	558	1.959752543	8.665261573	4.783186563	0.9654922072	
31	136.84	62	-12029783.35		120.7988472	1.084060454	583	1.957924395	8.859051401	4.779647157	0.9826847873	
32	132.402	63	-12015688.06		120.2865196	1.10692327	599	1.958054132	8.735711144	4.780154852	0.9784138602	
33	135.349	63	-12004316.02		120.3412265	1.122904801	614	1.955022302	8.488236416	4.778753	0.9614008374	
34	135.606	63	-11965901.62		119.5116148	1.136099803	636	1.955977918	8.778570495	4.77770899	1.013620027	
35	137.083	63	-11903240.45		118.8252804	1.149948499	640	1.95891399	8.655395487	4.774175523	0.8671151248	
36	135.328	62	-11876672		119.3411117	1.173343412	665	1.956690653	8.599501635	4.771885679	1.08189758	
37	137.354	62	-11822065.3		117.2066143	1.202751303	685	1.956590842	8.613601727	4.76972147	1.043460314	

pb spTree chl.trace

## Phylobayes (convergence assessment)

```
tracecomp -x 1000 <chain1> <chain2>
```


will produce an output summarizing the discrepancies and the effective sizes estimated for each column of the trace file. The discrepancy  $d$  is defined as  $d = 2|\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$ , where  $\mu_i$  is the mean and  $\sigma_i$  the standard deviation associated with a particular column and  $i$  runs over the chains. The effective size is evaluated using the method of Geyer (1992). The guidelines are:

- maxdiff < 0.1 and minimum effective size > 300: good run;
- maxdiff < 0.3 and minimum effective size > 50: acceptable run.

name	effsize	rel_diff
loglik	76	0.014258
length	563	0.0521663
alpha	637	0.0636774
Nmode	296	0.0335166
statent	2214	0.226332
statalpha	234	0.0154511
rrent	1955	0.1029
rrmean	12798	0.00651254

[See section 3.2 in the manual](#)

# Phylobayes (convergence assessment)

 BEASTdoc


[Home](#) [Nav](#) [Contents](#) [FAQ](#) [Help](#) [Source Code](#) [News](#)

## Contents

- Getting Started ▼
- Software Packages ▲
- BEAUi, BEAST and other programs
- BEAGLE Library
- Tracer

## Tracer


<https://beast.community/tracer>



Tracer is a graphical tool for visualization and diagnostics of MCMC output.



# Phylobayes (convergence assessment)

 BEASTdoc

Nav Contents FAQ Help Source Code News

search...

Contents


Getting Started

Software Packages

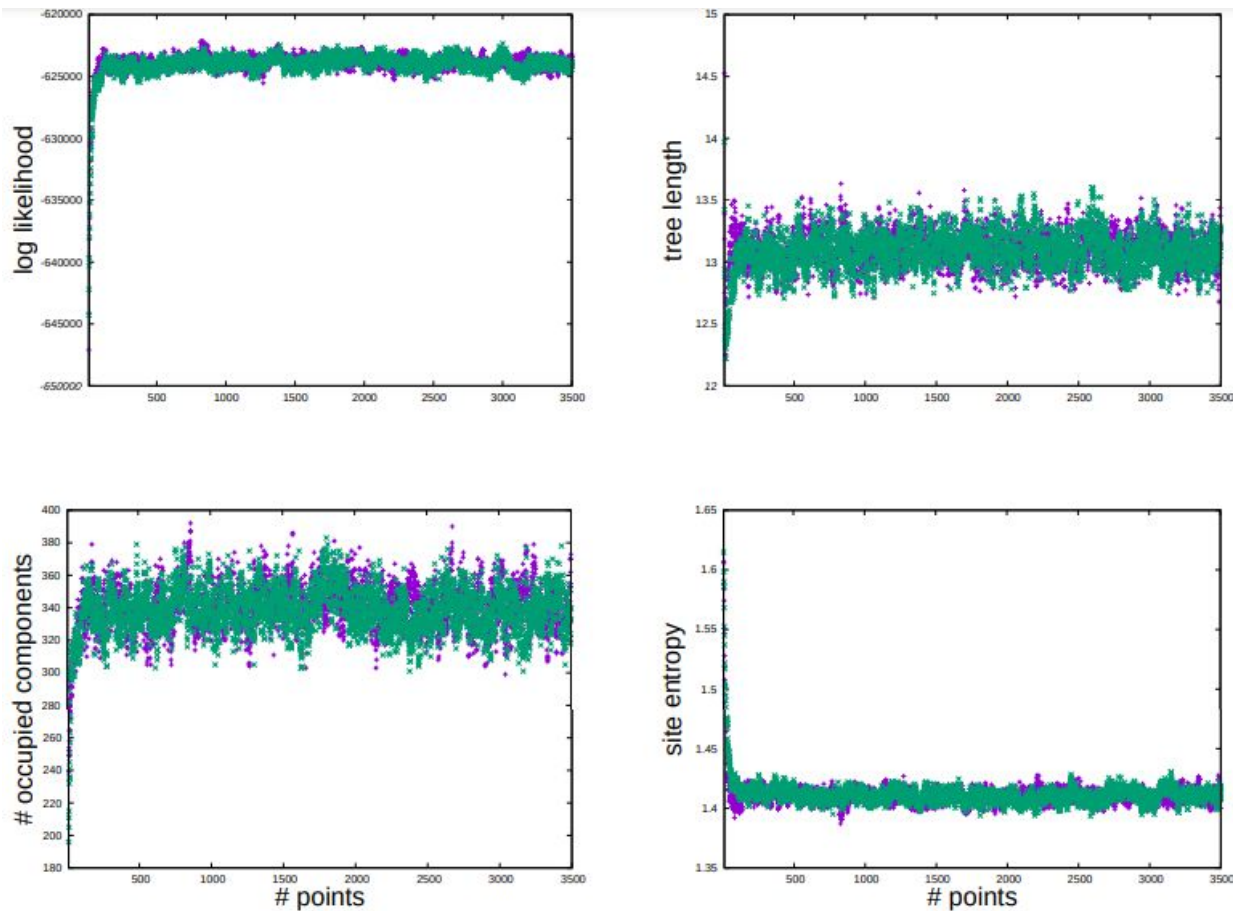
BEAUti, BEAST and other programs

BEAGLE Library

Tracer

 Tracer is a graphical tool for visualization and diagnostics of MCMC output.

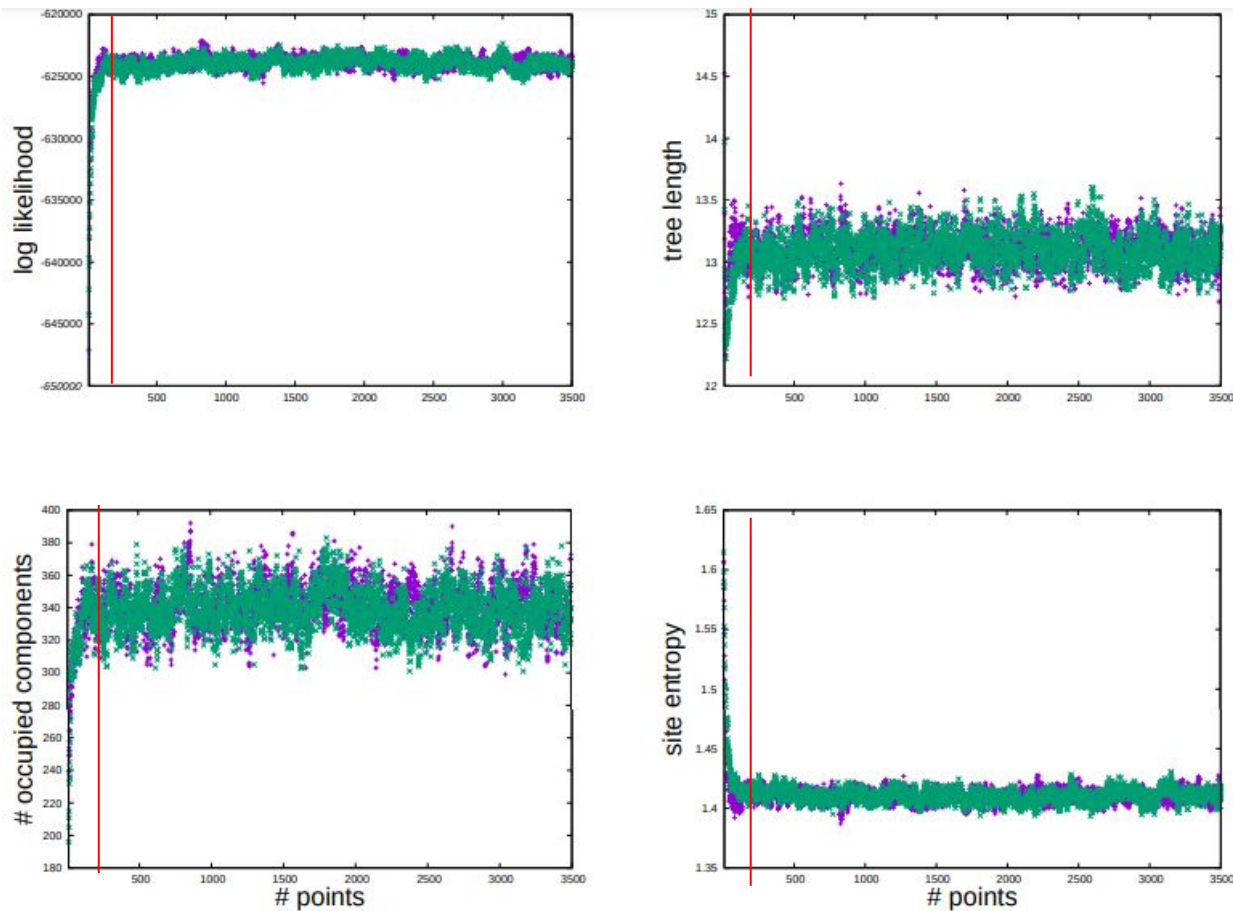
<https://beast.community/tracer>





# Phylobayes (convergence assessment)

“Visual assessment is essential, in particular, for getting a reliable **estimate of the burn-in**, i.e the number of points before the chain has reached stationarity”



# Phylobayes (convergence assessment)

“Visual assessment is essential, in particular, for getting a reliable estimate of the burn-in, i.e the number of points before the chain has reached stationarity”

“In general, it is particularly important to visualize at least the **log likelihood** (loglik, 4th column of the trace file), the **total tree length** (length, column 5), the **number of occupied components of the mixture** (Nmode, column 6) and the **mean site entropy** (statent, column 7), which is a measure of the strength of site-specific amino acid preferences”

