Phylogenetic analysis that models compositional heterogeneity over the tree

Peter G. Foster
Department of Life Sciences, Natural History Museum, London SW7 5BD, UK

October 13, 2021
Running head: Models for compositional tree-heterogeneity
Key Words: Phylogenetics, models of evolution, compositional heterogeneity, non-homogeneous, non-stationary, maximum likelihood, Bayesian MCMC, model fit, model adequacy, marginal likelihood, posterior predictive simulation

1 Abstract

That we find that molecular sequences in a phylogenetic analysis can differ in composition shows that the process of evolution can change over time. However, models of evolution in common use are homogeneous over the tree, and if used in a phylogenetic analysis with compositionally tree-heterogeneous datasets these models can recover incorrect trees. The NDCH or Node-Discrete Compositional Heterogeneity model is able to model such data by accommodating differences in composition over the tree. Usage, problems, and limitations of this model are discussed, and a modification, the NDCH2 model, is described that can ameliorate some of these problems and limitations. Using these models can greatly increase the fit of the model to the data, and can find better tree topologies. These models and various statistical tests are illustrated using a bacterial SSU rRNA dataset. These models are implemented in the software P4, and files for the analyses described here are made available.

2 Introduction

Most models used in molecular phylogenetics are time-homogeneous, meaning that the same model characteristics apply over the tree. However, compositional variation of the same gene in different organisms has been seen for some time [1–9]. That we see differences in the proportions of the bases of DNA or the amino acids of the protein tells us that the process of evolution can change over the tree, and if this variation is pronounced it can affect phylogenetic reconstruction when using tree-homogeneous models [2, 4, 10].

One possible solution to the problem of compositional heterogeneity over the tree is to only use molecular sequences that are homogeneous in composition [11]. This might require removal of compositionally divergent sequences so that the remaining sequences are more tree-homogeneous. A similar strategy would be to identify and remove compositionally divergent sites in the aligned sequences [12, 13]. It is possible that adding sequences may overcome compositionally-attracted sequences [2]. Use of protein sequences has been suggested as a way to avoid compositionally-biased DNA sequences [4], however protein-based alignments can also be compositionally heterogeneous over the tree [10].

There has been considerable research on methods and models to deal with compositional tree-heterogeneity without adding or removing sequences or alignment sites. Use of LogDet distances is often effective in distance-based analysis with such data [5, 6, 14]. Also, compositional tree heterogeneity has been modelled [15–21]. In this chapter I describe modelling compositional tree-heterogeneity using the NDCH (Node-Discrete Compositional Heterogeneity) and NDCH2 models as implemented in P4 using maximum likelihood and using an MCMC in a Bayesian framework [17]. I illustrate usage of these models and various statistical tests with an example bacterial SSU rRNA dataset.

2.1 Compositional effects

Compositional attraction, where unrelated taxa that share a compositional bias tend to group together in a tree-homogeneous phylogenetic analysis, has been appreciated for some time [2–4, 22]. However, this may be only one facet of a spectrum of possible compositional effects, examples of which are shown in Figure 1. For this Figure, DNA datasets were simulated on the trees in the simulation column, with compositional biases placed on the simulation trees as shown by the different colours. These datasets were analysed using the F81 model using IQ-TREE [23, 24], and the resulting trees are shown in the homogeneous model column with bootstrap support. In the control tree in Figure 1-homog, the simulation composition is homogeneous over the tree (25% for each nucleotide), and the F81 model recovers the correct tree with good support (93%) on the single split. However, when various compositional challenges are introduced to the simulation trees the recovered trees may differ from the simulation trees.

One way that might help us to visualize compositional effects is to make a distance-based tree using pairwise compositional distances, as shown in the tree based on composition column of Figure 1. For phylogenetic distances we can use Euclidean distances as measured using Equation 4 in Lockhart et al. [6], and in the examples here these distances were made into a tree using the program bionj [25]. A measure of confidence for such an analysis can be obtained by making many composition trees in a bootstrap loop, where composition trees are made from each of many pseudoreplicates of the dataset, and where a consensus of those trees shows support for internal splits due to composition alone. If there were no compositional effects then the supports for internal splits would be small. In the control tree in Figure 1-homog, the composition, being homogeneous, results in an equivocal composition tree as expected. The consensus composition tree is not the simulation tree, but it has poor bootstrap support (48%). However, other simulation scenarios have stronger effects, as described below.

In evolution, if there are compositional differences over the tree, then those differences would often follow or parallel the true evolution, such that members of an evolutionary group would tend to become compositionally biased in a similar way to each other, and so sister taxa will have similar compositions. This pattern is shown in the Figure 1-follow simulation tree. When we simulate data on this tree, subsequent analysis with the F81 model shows the correct tree, but with elevated bootstrap support (100%). The composition tree shows full support for the correct tree as well. There appears to be compositional attraction between sister taxa in the F81 tree, tending to raise the support higher than it would be due to phylogenetic signal alone. This is somewhat topologically benign because it reinforces the true simulation tree, although the overly-strong support is likely not warranted.

Compositional attraction is more of an obvious problem when we have parallel compositional biases in unrelated taxa (Figure 1-attract). In this evolutionary scenario unrelated taxa become biased in a similar way, and a tree-homogeneous model tends to put those taxa together. This effect is widely-appreciated, and the distortion caused to the tree is easy to understand.

The simulation scenario in Figure 1-repel shows sister taxa that have diverged in composition. The composition distances between the A-taxa and the B-taxa are smaller than the distances between the two A-taxa, although the distance between the two B-taxa are smaller still. The compositional effects here might be described as attraction of either of the two A-taxa to both of the B-taxa, or these effects might be described as repulsion of the two A-taxa from each other as the label "repel" implies. Compositional effects involving repulsion due to closely-related taxa becoming compositionally divergent (Figure 1-repel), or a complex interplay of attractions and repulsions (Figure 1-complex), are generally less well appreciated compared to compositional attraction. In these scenarios predicting what the effect will be on a homogeneous analysis or on a composition tree is not at all obvious.

2.2 NDCH and NDCH2 models

The NDCH model was made to accommodate compositional heterogeneity over the tree [17]. It was first used in a maximum likelihood framework, and one motivation of its design was to keep the number of model parameters low, to avoid over-parameterization. Other, similar, models had different composition parameters on each branch of the phylogenetic tree, and would not scale well to bigger trees [15, 16]. In the NDCH model, a small but fixed number of composition vectors are used, and each branch in the tree is assigned one of these vectors to make the rate matrices for that branch. This model was described in Foster [17], but later given the name NDCH (Node-Discrete Composition Heterogeneity) model in Cox et al. [26]. It is implemented in the software P4; see below. There is also a similar NDRH (Node Discrete Rate matrix Heterogene-

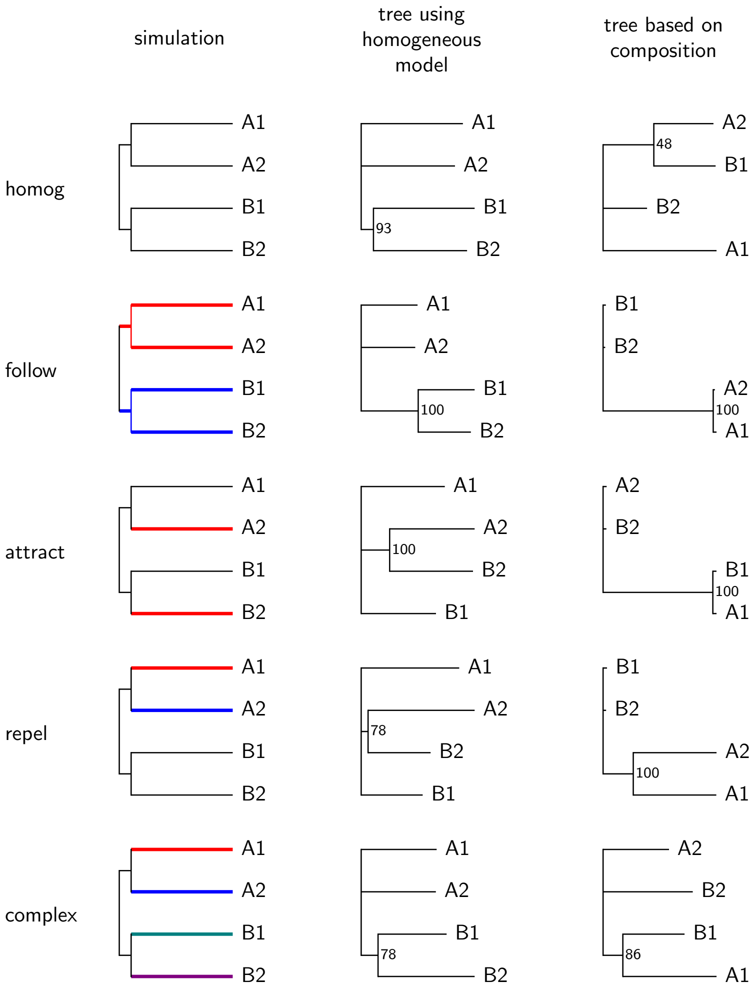

Figure 1: Compositional effects. See the text for explanation.

ity) model as well, which allows the rate matrix to change over the tree [27]. The NDCH and NDRH models can be used at the same time.

The ML (Maximum Likelihood) implementation of the NDCH model has limitations. The tree must be specified, and changes in tree topology are not implemented. One consideration with the ML implementation is that there are many possible ways that two or more composition vectors can be placed on the branches of a tree (one vector for each branch, and another for the root), and this arrangement is fixed for a given analysis. It would be computationally expensive to try all the possible arrangements even on a single tree. Furthermore, different root positions will affect the likelihood. The Bayesian implementation of the NDCH model in P4 addresses these problems. The Bayesian MCMC implementation includes topology moves, root position moves (as an internal node of degree two or three), and moves to change the disposition of the composition vectors on the tree.

P4 implements simulation of datasets based on the NDCH model. This is useful for making tree-heterogeneous datasets, and also for testing purposes, such as *posterior predictive simulations* for model adequacy testing in a Bayesian MCMC.

However, both the ML and Bayesian implementations of the NDCH model currently are limited due to having a fixed number of composition vectors. This is awkward because we will generally not know in advance how many composition vectors are needed to adequately model a dataset. One might start a series of analyses with a single composition vector (which would be a tree-homogeneous model) and then add composition vectors and re-analyse iteratively, perhaps until the model fits by some criterion, but this is computationally expensive. The NDCH2 model addresses this problem; it is described below.

3 An example analysis using bacterial SSU rRNA genes

Here I will use a small dataset from Embley, Thomas, and Williams [2] to illustrate the use of the compositional tree-heterogeneous models in P4, and some associated statistical tests. This dataset is a 1273-site alignment of bacterial SSU rRNA genes. The full alignment has six taxa, although we will mostly be using a five-taxon subset from it.

We first look at analyses with homogeneous models, and then use NDCH in maximum likelihood. The NDCH and NDCH2 models are then used in a Bayesian setting, including posterior predictive simulations to assess model fit, and model comparison using Bayes factors. Finally, these models are used in an attempt to root the phylogeny.

3.1 Analysis with homogeneous models

We can begin by analysing this dataset using tree-homogeneous models. It would be common to use the ModelFinder functionality of IQ-TREE to help in choosing a model, and that is done here (although I did not look at free rates models of among-site rate variation) [24, 28]. IQ-TREE names models in a laudably explicit manner, where for example +F (meaning empirical composition parameters) is used together with the GTR (General Time Reversible) model explicitly as GTR+F, when in common usage of GTR the +F attribute is implied. Here I will follow the common usage where GTR means GTR+F. Furthermore, all the models that were used on this dataset used gamma-distributed among-site rate variation, with four discrete categories, and so the +G4 attribute is implied. When all six bacterial SSU rRNA sequences are included in the analysis, the TIM2 model was chosen (the TIM2 rate matrix is similar to the GTR rate matrix except that in TIM2 the rate of exchanges between the nucleotides A and C is made the same as A and T exchanges, and the rate of CG exchanges is made the same as GT exchanges, giving it two fewer free parameters compared to the GTR model). The TIM2 model was the best choice using the AIC (Akaike Information Criterion), the AICc (AIC corrected for small sample size), or the BIC (Bayesian Information Criterion), making the choice unambiguous. In phylogenetic analysis of the six-taxon dataset with the TIM2 model we get what is believed to be the true biological tree (Figure 2a) [2]. The analysis was repeated with the JC (Jukes-Cantor) and GTR models with similar results (Figure 2a)

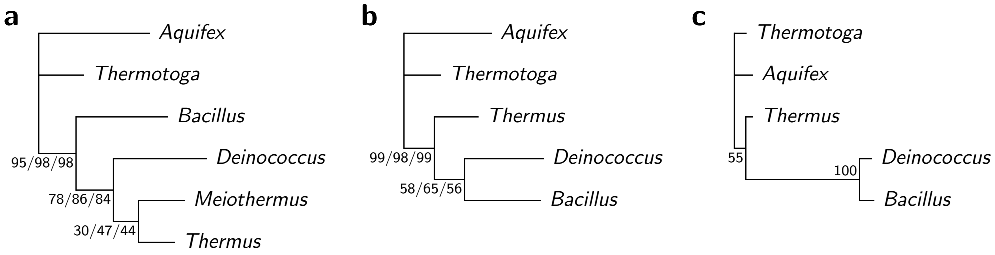

Figure 2: Panel **a**: Bacterial SSU rRNA analysis including *Meiothermus*. This is believed to be the correct biological tree. Bootstrap support is shown for models JC/TIM2/GTR. Panel **b**: Bacterial SSU rRNA analysis excluding *Meiothermus*, showing compositional attraction between *Bacillus* and *Deinococcus*, with Bootstrap support for models JC/TIM2/GTR. Panel **c**: Tree based on pairwise compositional distances; bootstrap support is shown.

However, when the *Meiothermus* sequence was removed the results were quite different. Model choice now had some ambiguity, with the GTR model being best by the AIC and AICc, while the TIM2 model was best by the BIC. Phylogenetic analysis of this dataset resulted in the tree shown in Figure 2b, again regardless of the model used. In this tree *Bacillus* groups with *Deinococcus* with weak support (56 – 65%).

3.2 Compositional heterogeneity

The GC composition of the bacterial SSU rRNA sequences (Table 1) provides an explanation for the unexpected tree shown in Figure 2b. The sequences differ substantially in their GC content, and when the constant sites were removed those differences became more apparent. *Aquifex*, *Thermotoga*, and *Thermus* are GC-rich (64 – 65% GC),

Table 1: Proportion (%) G + C in the SSU sequences. Differences are more evident when constant sites are removed. The alignment length is 1273 sites, of which 815 are constant.

	all sites	no constant sites
Meiothermus	58	57
Aquifex	65	76
Deinococcus	56	52
Thermotoga	64	75
Bacillus	56	51
Thermus	64	73

while *Deinococcus* and *Bacillus* are less so (both 56% GC). With these five sequences there would be a tendency for taxa to group together based on composition, which in this case is strong enough to overwhelm the true phylogenetic signal. The *Meiothermus* sequence is in between these two extremes at 58% GC, and we can speculate that the presence of *Meiothermus* moderates the compositional attraction in favour of the true phylogenetic signal, when analysed with a tree-homogeneous model [2].

A tree based on pairwise composition distances is shown in Figure 2c, showing a strong compositional attraction between *Deinococcus* and *Bacillus*, with 100% bootstrap support, consistent with Figure 2b.

To ask whether an aligned sequence pair violates stationarity we can use the symmetry tests explored and advocated by Jermiin and colleagues [29, 30]. It is a family of three tests — Stuart's or marginal symmetry test, the Ababneh or internal symmetry test, and Bowker's test, which is a sum of the other two. These test for stationarity, reversibility, and homogeneity, or SRH. These tests are implemented in P4 and IQ-TREE, and the marginal test is used in the software BMGE for identification of compositionally divergent sites [9, 13]. Stationarity in the dataset is tested using the marginal symmetry test. This test is for sequence pairs, and so a complete test of an alignment would be a matrix of pairwise statistics and P-values. IQ-TREE has opted to report the results as a summary of the matrix, with the P-value reported only for the most highly-diverged sequence pair [9]. P4 implements these tests, reporting the full matrices, the most significant, or the results from the most highly-divergent pair as in IQ-TREE. Results for the marginal symmetry for the five-taxon bacterial SSU rRNA dataset (we will only use the five-taxon dataset from now on) show that in the ten pairwise tests there were six significant pairs ($P < 0.001$), and four non-significant pairs.

Perhaps the most common way to test for compositional homogeneity is Pearson's Chi-squared test. This uses an $R \times C$ table, or contingency table, where the rows are taxa and the columns are character counts. Assuming homogeneity, we can infer expected counts. We can then use these expected counts to calculate a statistic, X^2, the

Table 2: Chi-squared tests for the bacterial SSU rRNA dataset. Individual sequences were tested, as well as the alignment overall

	X^2	P using χ^2	P using simulations
Aquifex	9.04	0.0288	0.0
Deinococcus	14.33	0.0025	0.0
Thermotoga	6.18	0.1030	0.0
Bacillus	14.47	0.0023	0.0
Thermus	4.89	0.1803	0.0
All	48.91	0.0000	0.0

significance of which is assessed using a χ^2 curve with appropriate degrees of freedom ((R - 1) × (C - 1)). Using the SSU dataset, $X^2 = 48.9$, and $P_{\chi^2,\mathrm{dof}=12} = 0.000002$. This tells us that compositional homogeneity is easily rejected at the 5% level.

However, this test is used inappropriately on these data, or indeed any phylogenetic data, because the sequences are related, and that leads to a high probability of type-II error. While the X^2 statistic is useful because it shows heterogeneity, it is inappropriate to use the χ^2 curve to measure its significance. A better null distribution can be obtained by simulation [17]. A tree and homogeneous model can be used to generate simulated datasets, the measured X^2 from which can be used as a valid null distribution by which to assess significance of the X^2 from the original dataset. In our case a test using this method shows $P < 0.005$, which is no surprise given that these data failed the Chi-squared test.

It is common to look at the Chi-squared contribution of individual sequences to the total Chi-squared (Table 2). These values from individual sequences can be assessed for significance using a χ^2 curve or, better, using a null distribution from simulations, as above. This might be useful to point out the “worst offenders”, which would be good candidates for removal. However, in our bacterial SSU rRNA alignment this would not be useful because all sequences reject homogeneity when assessed using simulations, and three of the five sequences reject homogeneity when assessed using χ^2.

3.3 Maximum likelihood with the NDCH model

That the bacterial SSU rRNA dataset contains high and low GC sequences suggests that this dataset could be modelled with a two-part composition model, and this will be done here with the NDCH model. A limitation of this model as currently implemented in an ML framework in P4 is that both the tree topology and the disposition of the composition vectors over the tree are fixed, and must be defined. Defining the tree topology is done in the usual way using a tree file or Python string in Newick format. Placement of composition vectors on the tree is described in Figure 3. For this

Table 3: Maximum likelihood analysis of the bacterial SSU rRNA using the GTR model and a two-composition NDCH model.

Model	Tree	log L	AU	TAMCFT
GTR	True	-4481.97	0.40	0.0
	Attract	-4479.57	—	0.0
NDCH	True	**-4396.44**	—	0.87
	Attract	-4414.17	0.02	0.64

particular dataset we can place one composition vector on the whole tree and then the second composition vector on the leaf branches for the two mesophiles, *Bacillus* and *Deinococcus*. This is done for both the True and Attract trees (Figure 3d and e).

Using the GTR model, the Attract tree had a better likelihood, differing by 2.4 log units (Table 3). We can use Shimodaira's AU (Approximately Unbiased) test and `consel` software to see if that difference is significant [31, 32], and for the GTR model the difference between the True and the Attract tree is not significant ($P = 0.40$). Using a second composition vector, making an NDCH model, gave an improvement in likelihood of about 85 log units for the True tree and 65 for the Attract tree. (In choosing the homogeneous model, IQ-TREE found the likelihood difference between the GTR model with and without among-site rate variation to be 94.4 log units, meaning that for this dataset accommodation of heterogeneous composition is of similar importance to accommodation of among-site rate variation.) The cost of the three additional free parameters in the NDCH model is therefore worthwhile given the increase in model fit. Using the NDCH model the likelihood difference of 17.73 log units between the True and the Attract trees becomes significant by the AU test ($P = 0.022$).

The fit of the composition of the data to the composition of the model can be assessed in a maximum likelihood context using the tree- and model-based composition fit test (TAMCFT, Table 3) [17]. The test uses the quantity X_m^2, computed by $\sum[(Obs - Exp)^2/Exp]$, as is the X^2 statistic used in the Chi-squared test, except that the expected value comes from the model, not from the data. This can therefore test the fit of a compositionally tree-heterogeneous model to compositionally tree-heterogeneous data. Assessment of significance is done with a null distribution made using maximum likelihood X_m^2 values from simulations under the model being tested. When this test is done with the GTR model using either tree (True or Attract) the model does not fit the composition of the sequences overall, or for any of the sequences individually ($P < 0.01$ for all tests). However, when the TAMCFT is done using the NDCH model above, neither the data overall nor any of the sequences individually reject model fit for composition (P from $0.27 - 0.95$). Therefore two composition vectors are sufficient for model adequacy of the composition component of the model by this test.

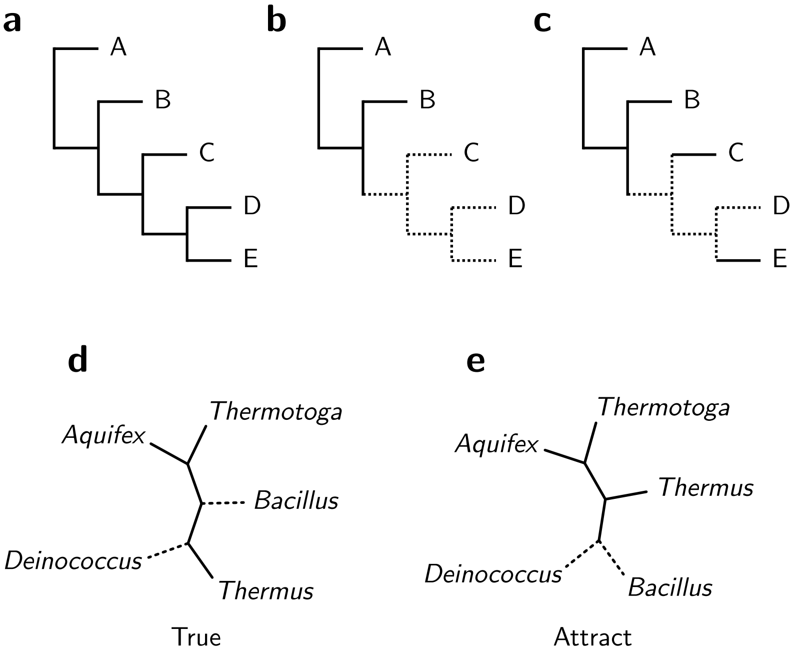

Figure 3: Placement of composition vectors on the tree. Panels **a** – **c** describe how two composition vectors, indicated by solid and dotted lines, might be placed on a tree. Composition vectors can be placed on any branch or on the root, and each placement can be clade-wise or be specific for a single branch. Placements are made in order, with later placements overriding earlier placements. We might begin by placing a composition vector on the root as shown in Panel **a**, clade-wise, indicated by the solid lines over the entire tree, including the root. Panel **b** shows a placement of the second composition vector on the branch for the parent of the taxon C, also clade-wise. Finally in Panel **c** the first composition vector is placed on leaf nodes C and E; being leaf nodes it does not matter in this case whether it is done clade-wise or not. Panels **d** and **e** show the two trees used in the bacterial SSU rRNA example; the tree in Panel **d** is the True biological tree, while the tree in Panel **e** is the Attract tree as found in analyses using homogeneous models. Panels **d** and **e** also show placement of composition vectors on these trees for the NDCH model — one composition vector is placed on the whole tree, including the root, and a second composition vector is assigned to the leaf branches for *Bacillus* and *Deinococcus*.

Table 4: Bayesian MCMC results for bacterial SSU rRNA

Model	*Deinococcus* + *Bacillus*	*Deinococcus* + *Thermus*	Posterior predictive P_t	Marginal L
GTR	0.90	0.10	0.0	-4518.9
NDCH-Fixed [a]	0.0	1.0	0.36	**-4444.0**
NDCH-Free [b]	0.0	1.0	0.37	-4449.5
NDCH-Fully [c]	0.0	1.0	0.48	-4460.7
NDCH2	0.0	1.0	0.29	-4450.5

[a] Fixed placement of two composition vectors on the tree
[b] Free placement of two composition vectors on the tree
[c] Fully parameterized, with a composition vector on each branch, plus the root

3.4 Bayesian analysis

One limitation of the ML implementation of the NDCH model discussed above is that there are many different possible placements of some number of composition vectors on a tree, and to find the ML placement by trying them all would be computationally expensive. Our bacterial SSU rRNA example above looked at only one such placement of only two composition vectors, on only two trees. Bigger, more realistic phylogenetic problems with free tree topology might be intractable. Furthermore, although this was not discussed in the ML analysis above, the position of the root affects the likelihood (but see the "rooting" section, below). These problems are addressed in the Bayesian implementation of the NDCH model — the Bayesian MCMC implementation in P4 allows free tree topology, free placement of a fixed number of composition vectors on the tree, and free root position.

We can initiate a Bayesian MCMC, with either the GTR or NDCH model, with a random tree. When the MCMC is run using the GTR model, the consensus result is the Attract tree, with the *Deinococcus* + *Bacillus* split with posterior probability of 0.90 (Table 4), consistent with the ML analysis above.

With the NDCH model we can run the MCMC either with the composition vectors assigned to the same branches throughout the analysis (NDCH-Fixed in Table 4), or we can allow each branch to change the composition vector that it is associated with during the run (NDCH-Free). When we run the analysis with fixed composition vector assignment, we define two composition vectors, and assign them to the tree as shown in Figure 3 such that one composition is assigned to the branches leading to *Bacillus* and *Deinococcus*, and the other composition is assigned to all the other branches, and the root. The consensus tree for this analysis is the True tree, with all splits with full support (Table 4, NDCH-Fixed).

A measure of the absolute fit of the composition component of the model to the data can be obtained using posterior predictive simulations [33]. We can use the X^2 statistic from the Chi-squared test as the test quantity. As implemented, at intervals during the MCMC, a dataset is simulated using the current state of the chain, and the test quantity X^2 is calculated for that simulated dataset. At the end of the run the simulated X^2 values form a distribution by which the X^2 value from the original dataset (48.9) can be assessed as a tail-area probability P_t. The GTR model is compositionally tree-homogeneous, and simulated datasets from samples taken from an MCMC using this model will also be compositionally tree-homogeneous and so have small X^2 values. In our test the X^2 values ranged from 0.4 to 9.9, and since $X^2 = 48.9$ in the original dataset, this model does not fit the data ($P_t = 0.0$). Using the NDCH model, however, the simulated X^2 values are bigger, ranging from 18 to 90. The value of the test quantity from the original data sits within this distribution, and so the composition component of this model plausibly fits the data ($P_t = 0.36$).

We can compare the two models, GTR and NDCH, using Bayes factors. Bayes factor is the ratio of marginal likelihoods for the models being compared, and we can estimate the marginal likelihood using the Stepping Stone method [34]. The difference of 75 log units between the marginal likelihoods of the GTR and NDCH models (NDCH-Fixed) is the log Bayes factor, and is decisive in favour of the NDCH model (Table 4).

If we allow free disposition of the composition vectors on the tree (NDCH-Free), then the NDCH MCMC run is started with two composition vectors placed randomly on the random starting tree. An MCMC proposal for the NDCH model allows changing which vector is assigned to a branch (or the root). In this way, the disposition of the composition vectors can change over the tree, even as the topology changes. These associations of composition vector to branch are saved in sampled trees. When making a consensus tree from an NDCH MCMC, we can use this information to show consensus placement of composition vectors. The results of such a run (Table 4, NDCH-Free) are almost the same as when the composition vectors are fixed in place. Again the consensus tree is the True tree, with *Deinococcus* + *Thermus* with 100% support, and posterior predictive simulations using X^2 had a tail-area probability of 0.37. In these MCMC runs, with free placement of the composition vectors, the branches to *Bacillus* and *Deinococcus* always had one vector and the other three taxa always had the other vector, regardless of the starting random disposition of the two vectors on the starting tree; in this way it was the same as the runs with fixed placement. However, there was some variation in placement of the two vectors on internal branches and on the root. The acceptance rate of proposals to change the composition vector disposition on the tree was low or zero, which likely indicates poor mixing in this regard. Here poor mixing means that the MCMC chain tends to become stuck in a probability region, unable to sample the full probability landscape. Consistent with this speculation, marginal likelihood estimation was lower than that for the NDCH runs with fixed composition location above (-4449.5), and marginal likelihood estimates from two replicates differed from each other, again pointing to poor mixing.

If we fully parameterize the NDCH model with a composition vector on each branch and the root (NDCH-Fully in Table 4), then results are again similar to NDCH-Fixed, but with a lower marginal likelihood, possibly due to over-parameterization or poor fit of the uninformative prior probability for the composition vectors.

3.4.1 NDCH2

The NDCH2 model addresses the problem of poor mixing of composition placements in the NDCH-Free MCMC above. In the NDCH2 model the composition is fully parameterized, with a separate composition vector on each branch of the tree, and an additional composition vector for the root. In this way the NDCH2 model is similar to the NDCH-Fully variant described above, except that the composition vectors in the NDCH2 model are more constrained by their prior probabilities. As implemented, the NDCH2 model has two prior probabilities on the values of the composition vectors — one prior for the leaf compositions, and another for the internal compositions. These priors are Dirichlet, or multivariate beta, distributions. The amount of constraint given by the prior is determined by the *concentration parameter* for the distribution. The concentration parameters of these priors can be fixed to reasonable values (chosen using preliminary runs with free hyperparameters, or by bracketing fixed values), but can also be sampled in the MCMC. Using such *hyperparameters*, sampled parameters of the prior, the model can tune itself to the given dataset. (The composition vectors in the NDCH model also have a Dirichlet prior, but are set, using a fixed concentration parameter of unity, to be non-informative and so provide no constraint).

When the MCMC is run with fixed hyperparameters, the resulting consensus tree is the True tree, with 100% support for all splits (Table 4, NDCH2). Posterior predictive simulations show that the model composition fits the data ($P = 0.29$). The log marginal likelihood is estimated to be -4450.5, which is somewhat lower than that for the NDCH model with two composition vectors, which possibly indicates some overparameterization or overly uninformative priors. However, the log marginal likelihood for NDCH2 is higher than that for the NDCH-Fully model; this difference is likely due to the prior probabilities of the NDCH2 model as that is the only difference between them.

When run with free hyperparameters, the prior affecting leaf compositions was well-behaved, with samples in a fairly narrow range. However, the sampled prior affecting internal compositions was much more scattered, possibly because there was not enough information in the small bacterial SSU rRNA dataset for decisive MCMC sampling of this prior. However, the Stepping Stone log marginal likelihood estimation, which is sensitive to the model prior, was -4450.6, approximately the same as that using fixed hyperparameters. From these analyses it appears that the NDCH2 model provides a workable way to model compositionally tree-heterogeneous datasets, and it is the most recommendable of the models discussed.

Finally, for the datasets in Figure 1, the simulation tree was obtained in all cases when analysed with the NDCH2 model.

3.5 Rooting using tree-heterogeneous models

Although this was not the focus of the analyses above, the root position of a tree-heterogeneous model affects the likelihood. This presents the possibility of being able to root the tree using such models. We can look at doing this with the bacterial SSU rRNA dataset with the NDCH model in a maximum likelihood framework. Because this is a small tree and dataset, we can evaluate all possible rootings. We can start with the True tree of Figure 3, place root nodes on all seven branches, and evaluate each by ML. If we use the same 2-composition NDCH model as was used above, then there is very little difference (< 0.02 log units) between any of the root positions. If we add a third composition vector for the root node in each tree, then there is somewhat more difference between trees (about 2.4 log units between best and worst), and the ML root is the tree with the root placed on the branch to *Thermatoga*. The worst root tested is the root on the branch leading to *Bacillus*, which is unexpected as it is an outgroup to the other four taxa and would be expected to be the best root.

We can use the NDCH2 model on these data in another MCMC to attempt to root the tree. For this we start with a random tree with a bifurcating root, rather than the trifurcating root that was used in the NDCH2 analysis above. The position of the bifurcating root is free and sampled, and so counts of each root position visited by the MCMC can be obtained as an estimate of the posterior probability of each root. When this is done with the bacterial SSU rRNA dataset, the consensus tree topology is the True tree with full support, and the consensus root position was on the branch leading to *Aquifex*, although rootings on all branches were sampled during the MCMC. Using the bacterial SSU rRNA dataset together with these models results in a root position that is both noisy and likely incorrect (the branch leading to the outgroup *Bacillus* would be the biologically expected root position), possibly because there is not enough information in this small dataset to be decisive, but also there may be other data attributes or biases that we have not modelled.

We can simulate datasets where such an analysis can recover the correct root more decisively, showing that this approach can succeed (an example is given in the scripts made available; see below). However, the poor rooting results for the bacterial SSU rRNA dataset shows that it is not always possible.

4 Software and scripts

The models and methods discussed above are implemented in the phyloinformatic toolkit P4, hosted at GitHub (`https://github.com/pgfoster/p4-phylogenetics`). It is written for the most part in the Python language, although computationally intensive parts are written as a Python module in the C language. The user interface is also Python. It is possible to use P4 interactively, but in practice it would generally be used with short Python scripts. Using Python as the interface has advantages provided by the language, such as the ability to store named values, and the ability to have loops and conditionals.

P4 reads and writes various phylogenetic formats. It flexibly handles multi-partition data and multi-partition models, where the partitions are separate from each other, and can have different models or different datatypes. Another advantage provided by using Python is the ability to *pickle* Python objects to files. P4 uses this ability to pickle ML optimized tree-model objects. Pickling also allows checkpoints of MCMC objects, which can be read and queried, and can be used for restarting the MCMC.

In an MCMC, some proposals have a tuning parameter that controls how close a proposed state will be to the current state. This is reflected in the proposal acceptance rates, which are commonly assumed to reflect mixing in the chain. In P4 the tunings are periodically adjusted during the MCMC, based on running acceptance rates, in order to encourage good mixing.

Metropolis-coupled MCMC, (MCMCMC, $(\mathrm{MC})^2$) also known as parallel tempering, is useful for better mixing in difficult MCMC analyses [35, 36]. It uses several MCMC chains run in parallel, all but one of which is *heated* to different temperatures, which tends to allow higher acceptance rates of proposals, and better mixing. Only the unheated or cold chain is sampled. During the MCMCMC, chains can *swap* with some probability, allowing the cold chain to have access to regions sampled by the heated chains. In the P4 implementation of the MCMCMC the chain temperatures are adaptively tuned during the MCMC in order to maintain a good swapping rate between all chains. In the current implementation, swaps are proposed only between adjacent chains, as it was noted that swaps between non-adjacent chains were rarely accepted.

P4 has a class for handling tree bipartitions or splits. It can make majority-rule consensus trees, but can also show consensus placement of composition vectors in the NDCH model, and can show the consensus root position from MCMC sampled output. Posterior predictive simulations to assess model fit in an MCMC can be made during an MCMC run, but can be done after the run as well. A few different test quantities for posterior predictive simulations are provided by the MCMC class.

Scripts for all the analyses described here for the bacterial SSU rRNA example can be found at GitHub (https://github.com/pgfoster/BacterialSSUrRNAExample)

References

[1] Muto A and Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Nat Acad Sci (USA)* 84:166–9. DOI: 10.1073/pnas.84.1.166.

[2] Embley TM, Thomas RH, and Williams RAD (1993) Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. *Syst Appl Microbiol* 16:25–29. DOI: 10.1016/S0723-2020(11)80247-X.

[3] Steel M, Lockhart P, and Penny D (1993) Confidence in evolutionary trees from biological sequence data. *Nature* 364:440–442. DOI: 10.1038/364440a0.

[4] Hasegawa M and Hashimoto T (1993) Ribosomal RNA trees misleading? *Nature* 361:23. DOI: 10.1038/361023b0.

[5] Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc Nat Acad Sci (USA)* 91:1455–1459. DOI: 10.1073/pnas.91.4.1455.

[6] Lockhart PJ, Steel MA, Hendy MD, and Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol* 11:605–612. DOI: 10.1093/oxfordjournals.molbev.a040136.

[7] Foster PG, Jermiin LS, and Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* 44:282–288. DOI: 10.1007/PL00006145.

[8] Mooers AØ and Holmes EC (2000) The evolution of base composition and phylogenetic inference. *Trends Ecol Evol* 15:365–369. DOI: 10.1016/S0169-5347(00)01934-0.

[9] Naser-Khdour S, Minh BQ, Zhang W, Stone EA, and Lanfear R (2019) The prevalence and impact of model violations in phylogenetic analysis. *Genome Biol Evol* 11:3341–3352. DOI: 10.1093/gbe/evz193.

[10] Foster PG and Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* 48:284–290. DOI: 10.1007/PL00006471.

[11] Collins TM, Fedrigo O, and Naylor GJ (2005) Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. *Syst Biol* 54:493–500. DOI: 10.1080/10635150590947339.

[12] Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, and Philippe H (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol* 56:389–399. DOI: 10.1080/10635150701397643.

[13] Criscuolo A and Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:1–21. DOI: 10.1186/1471-2148-10-210.

[14] Hirt RP, Logsdon JM, Healy B, Dorey MW, Doolittle WF, and Embley TM (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Nat Acad Sci (USA)* 96:580–585. DOI: 10.1073/pnas.96.2.580.

[15] Yang Z and Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol* 12:451–458. DOI: 10.1093/oxfordjournals.molbev.a040220.

[16] Galtier N and Gouy M (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* 15:871–879. DOI: 10.1093/oxfordjournals.molbev.a025991.

[17] Foster PG (2004) Modeling compositional heterogeneity. *Syst Biol* 53:485–495. DOI: 10.1080/10635150490445779.

[18] Gowri-Shankar V and Rattray M (2007) A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol Biol Evol* 24:1286–1299. DOI: 10.1093/molbev/msm046.

[19] Blanquart S and Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol* 25:842–858. DOI: 10.1093/molbev/msn018.

[20] Heaps SE, Nye TM, Boys RJ, Williams TA, and Embley TM (2014) Bayesian modelling of compositional heterogeneity in molecular phylogenetics. *Stat Appl Genet Mol Biol* 13:589–609. DOI: 10.1515/sagmb-2013-0077.

[21] Williams TA, Heaps SE, Cherlin S, Nye TM, Boys RJ, and Embley TM (2015) New substitution models for rooting phylogenetic trees. *Phil Trans Roy Soc B: Biol Sci* 370:20140336. DOI: 10.1098/rstb.2014.0336.

[22] Jermiin LS, Ho SY, Ababneh F, Robinson J, and Larkum AW (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol* 53:638–643. DOI: 10.1080/10635150490468648.

[23] Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17:368–376. DOI: 10.1007/BF01734359.

[24] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Haeseler A von, and Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 37:1530–1534. DOI: 10.1093/molbev/msaa015.

[25] Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695. DOI: 10.1093/oxfordjournals.molbev.a025808.

[26] Cox CJ, Foster PG, Hirt RP, Harris SR, and Embley TM (2008) The archaebacterial origin of eukaryotes. *Proc Nat Acad Sci (USA)* 105:20356–20361. DOI: 10.1073/pnas.0810647105.

[27] Foster PG, Cox CJ, and Embley TM (2009) The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Phil Trans Roy Soc B: Biol Sci* 364:2197–2207. DOI: 10.1098/rstb.2009.0034.

[28] Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, and Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Meth* 14:587–589. DOI: 10.1038/nmeth.4285.

[29] Ababneh F, Jermiin LS, Ma C, and Robinson J (2006) Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* 22:1225–1231. DOI: 10.1093/bioinformatics/btl064.

[30] Jermiin LS, Jayaswal V, Ababneh FM, and Robinson J (2016) Identifying Optimal Models of Evolution. In: *Methods in Molecular Biology.* Springer New York, pp. 379–420. DOI: 10.1007/978-1-4939-6622-6_15.

[31] Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508. DOI: 10.1080/10635150290069913.

[32] Shimodaira H and Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247. DOI: 10.1093/bioinformatics/17.12.1246.

[33] Bollback JP (2002) Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol* 19:1171–1180. DOI: 10.1093/oxfordjournals.molbev.a004175.

[34] Xie W, Lewis PO, Fan Y, Kuo L, and Chen M.-H (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol* 60:150–160. DOI: 10.1093/sysbio/syq085.

[35] Geyer CJ (1991) Markov chain Monte Carlo maximum likelihood.

[36] Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755. DOI: 10.1093/bioinformatics/17.8.754.