Puolamäki, Kai; Fortelius, Mikael; Mannila, Heikki

Seriation in Paleontological Data Using Markov Chain Monte Carlo Methods

Published in:
PLoS computational biology

DOI:
10.1371/journal.pcbi.0020006

Published: 01/01/2006

Please cite the original version:
Seriation in Paleontological Data Using Markov Chain Monte Carlo Methods

Kai Puolamäki 1*, Mikael Fortelius 2, Heikki Mannila 3

1 Laboratory of Computer and Information Science, Helsinki University of Technology, Espoo, Finland, 2 Department of Geology and Institute of Biotechnology, University of Helsinki, Helsinki, Finland, 3 HIIT Basic Research Unit, University of Helsinki and Helsinki University of Technology, Helsinki, Finland

Given a collection of fossil sites with data about the taxa that occur in each site, the task in biochronology is to find good estimates for the ages or ordering of sites. We describe a full probabilistic model for fossil data. The parameters of the model are natural: the ordering of the sites, the origination and extinction times for each taxon, and the probabilities of different types of errors. We show that the posterior distributions of these parameters can be estimated reliably by using Markov chain Monte Carlo techniques. The posterior distributions of the model parameters can be used to answer many different questions about the data, including seriation (finding the best ordering of the sites) and outlier detection. We demonstrate the usefulness of the model and estimation method on synthetic data and on real data on large late Cenozoic mammals. As an example, for the sites with large number of occurrences of common genera, our methods give orderings, whose correlation with geochronologic ages is 0.95.

Introduction

Seriation, the task of temporal ordering of fossil occurrences by numerical methods, and correlation, the task of determining temporal equivalence, are fundamental problems in paleontology. Fossils have been used for both tasks since the very beginnings of modern paleontology [1,2]. However, the recent advent of large fossil databases [3] and the increased emphasis on quantitative analysis of biological patterns in deep time (e.g., [4–7]) has to some extent changed both the nature and the primary purpose of these activities. The rules and procedures of conventional paleontological seriation (biostratigraphy) [8,9] are not easy to apply in a satisfactory way to large datasets compiled from a wide variety of sources, often without associated data concerning the local distribution of fossil taxa in the rock sequences from which they derive. Conversely, occasional but valuable data regarding lithostratigraphic superposition, geochronologic age estimates, etc., are frequently available but difficult to apply in a global setting. The increasing use of large datasets in paleontological research implies a growing need for methods that do not only order the sites into a temporal sequence based on the distribution of taxon occurrences, but also can make use of a variety of other kinds of readily available stratigraphic information. This need is perhaps most acute in situations such as the continental record of the Eurasian Cenozoic (ca 23 million to 2 million years ago), where the majority of localities are from isolated quarries outside a rock-context and stratigraphic superposition therefore cannot be directly determined (e.g., [10,11]).

In the past several decades, both seriation and correlation of fossil occurrences by numerical methods have in fact become practically feasible alternatives to conventional biostratigraphy. The computational solutions that have been developed for correlation and seriation have much in common, but the implementations differ depending on the purpose and the nature of the data (e.g., the CHRONOS [12] and PAST [13,14] initiatives; also see [15]).

Here we are explicitly concerned with the task of seriation, for which methods based on several distinct approaches are available. These include the graph-theoretical unitary associations method by Guex et al. [16–18], parsimony analysis [19–22], and Bayesian methods [23]. John Alroy [5,6,24–27] has developed and applied techniques based on estimating taxon ranges and maximizing the fit of these hypothesized ranges to independent stratigraphic information, including known stratigraphic superposition of localities. ([5] gives the latest review of this approach.)

A fossil site (a collection of fossils remains collected from some location, typically in a sedimentary deposit) may be loosely regarded as a snapshot of the set of taxa that lived at a certain location at approximately the same time. Sites and their taxa may be described as an occurrence matrix, i.e., a 0–1 matrix, where the rows correspond to sites and the columns correspond to taxa: a one in entry (i,j) means that taxon j has been found at site i. The snapshot may capture a smaller or larger proportion of the taxa that actually present, a smaller or larger area, and a shorter or longer time interval, and it may be biased in different ways. It is therefore clear that the ones and zeros in such a matrix are not all equal. Some presences will be weakly founded on single specimens, others on hundreds or thousands of specimens from many sites. Similarly, many absences will be nothing more than missing data, whereas absences in well-sampled sites may carry more meaning. These facts virtually call out for a probabilistic approach to the analysis of paleontological presence-absence data.

Seriation Using MCMC

Synopsis

Seriation, the task of temporal ordering of fossil occurrences by numerical methods, and correlation, the task of determining temporal equivalence, are fundamental problems in paleontology. With the increasing use of large databases of fossil occurrences in paleontological research, the need is increasing for seriation methods that can be used on data with limited or disparate age information. This paper describes a simple probabilistic model of site ordering and taxon occurrences. As there can be several parameter settings that have about equally good fit with the data, the authors use the Bayesian approach and Markov chain Monte Carlo methods to obtain a sample of parameter values describing the data. As an example, the method is applied to a dataset on Cenozoic mammals. The orderings produced by the method agree well with the orderings of the sites with known geochronologic ages.

Here we describe a straightforward probabilistic model that contains parameters for the origination and extinction of taxa, for the ordering of the sites, and for the probabilities of errors (wrong zeros and wrong ones). Given the ordering of the sites, the origination and extinction parameters for a taxon specify the interval in which the taxon is assumed to be present. Any occurrence (a one in the matrix) of the taxon outside this interval is considered to be a false occurrence, and any nonoccurrence (a zero in the matrix) is considered to be a false occurrence as well. Given the parameters, the likelihood of the data depends on the number of false and true ones and zeros. The task we consider is to find parameter vectors that yield high likelihood, i.e., have a small number of false ones and zeros.

In more detail, our probabilistic model is as follows. Given a dataset with \( N \) sites and \( M \) taxa, we consider arbitrary orderings \( \pi \) of the sites. For each taxon \( m \), we have the lifetime of the taxon, specified by the indices of the first and last occurrences \( a_m \) and \( b_m \) of taxon \( m \). Additionally, we have parameters \( c \) for the probability of a false one (i.e., the presence of taxon \( m \) outside the bounds \( a_m \) and \( b_m ) \) and \( d \) for the probability of a false zero (i.e., the absence of taxon \( m \) between the bounds \( a_m \) and \( b_m ) \).

Denoting by \( \theta \) the whole parameter vector \( (\pi, a, b, c, d) \), the likelihood of the data \( X \) has the form

\[
P(X|\theta) = c^x(1-c)^y d^z(1-d)^\gamma,
\]

where \( x \) is the number of false ones, \( y \) is the number of correct zeros, \( \gamma \) is the number of false zeros, and \( \delta \) is the number of correct ones.

We could, in principle, find a parameter vector \( \theta \) that maximizes the likelihood of the data (maximum likelihood solution). This parameter vector would give a total order for the fossil sites, implying a probability of zero or one for a site pre-dating another. However, we know that the data contain pairs of sites from the same time periods. We are interested in finding pairs of sites for which the seriation is uncertain, i.e., the probability of one site pre-dating another is close to one-half.

Therefore, we work in the Bayesian framework, and find a sample of parameter vectors where the probability of a vector is proportional to its posterior probability. To this end, we use the Markov chain Monte Carlo (MCMC) method [28,29] to sample parameter combinations \( \theta \) with probability proportional to their Bayesian posterior likelihood

\[
P(\theta|X) \propto P(X|\theta)P(\theta),
\]

where \( P(\theta) \) is the prior probability of the parameters \( \theta \). MCMC methods yield samples from the posterior distribution of the parameters, and this makes it possible to study the space of the parameters in many different ways. For example, we can determine, for each pair of sites, the probability that one precedes the other. We can also estimate for the probability of false zeros and false ones and find for a particular observation in the data the probability that it is a false zero or a false one.

A further useful property of the model is that it is easy to incorporate additional information. For example, the model allows us to freeze the ordering of certain sites. That is, if we know that site \( i \) is definitely older than site \( j \), we can restrict the MCMC method to accept only permutations that satisfy this constraint.

Results

Generated Data

We first ran the experiment on synthetically generated data, with known “true” ordering and probabilities of false zeros and ones for varying numbers of sites and taxa, shown in Table 1. The data were generated by forming \( N \) ordered sites and \( M \) taxa, assigning each taxon a lifespan (with median \( 0.18N \) sites) starting from a random position in the sitelist, and using the parameters \( c \) and \( d \) to produce false ones and false zeros.

The results on synthetic data show that the method quite accurately determines the parameters of the model: the expected values of \( d \) obtained from the MCMC simulation are close to the ones used in generating the data. The correlations between the original order and the MCMC orderings are also quite high; note that for high values of \( c \) and \( d \) it can be the case that some orderings different from the generating one fit the data better than the generating ordering.

Cenozoic Large Land Mammal Data

In MCMC simulations, different runs can converge to separate regions in the parameter space. This is indeed what happens with the datasets on genera of Cenozoic large land mammals. We ran 100 MCMC chains over the datasets, and computed the variance in negative log-likelihood within the first chain, and then included all chains with the expected negative log-likelihood within one sigma of the best chain to our analysis.

The results are summarized in Table 2. The results show that the probability of a false one is quite low, whereas the probability of a false zero varies from 0.5 to 0.77, with the highest probabilities in the datasets where sites or genera with smaller occurrence frequencies are included. The correlation with the Mammal Neogene (MN) ordering and database age are also high.

The probability that a site \( i \) occurs before site \( j \) can be estimated simply by counting in how many of the samples \( i \) precedes \( j \) in the ordering \( \pi \). The results are shown in Figure 1. The ordering of the sites is in general quite well determined, but for some blocks of observations the ordering...
seems to be not fixed. Note that if two sites have exactly the same taxa, then the probability \( P(p_i, p_j) \) will be 0.5.

The pair-order matrices for all 100 chains are shown on our Web site (http://www.cis.hut.fi/projects/patdis/paleo). The chains outside the highest likelihood typically contain blocks of sites whose orders have been reversed; the MCMC method is fairly sensitive to initialization, and therefore running several chains with different initializations is useful.

For the dataset specified by \( n_t = 10 \) and \( n_s = 10 \), the original data are shown in Figure 2A. The probability of a genus being alive at site \( m \) is shown in Figure 2B. Figure 2C shows for each one in the data the probability that it is a false one, i.e., that it falls outside the interval \((a_m, b_m)\). We note that certain observations are quite strongly assumed to be false ones. The list of the strongest false ones is given on our Web site. Examination of the list shows that some of the observations that the model considers false ones are probably real errors in the data, while others represent true outliers or pertain to genera with unusual species or genera with artificially truncated distributions, as discussed below.

The median lifespan \((b - a) / C_0\) of taxa is 0.18 N sites.

#### Table 1. Results for Artificially Generated Datasets

<table>
<thead>
<tr>
<th>N</th>
<th>M</th>
<th>c</th>
<th>d</th>
<th>(E(c))</th>
<th>(E(d))</th>
<th>(E(\text{corr}(\pi)))</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>0.003</td>
<td>0.2</td>
<td>0.0036</td>
<td>0.2042</td>
<td>0.9982</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.003</td>
<td>0.5</td>
<td>0.0029</td>
<td>0.5207</td>
<td>0.8775</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.03</td>
<td>0.2</td>
<td>0.0289</td>
<td>0.2172</td>
<td>0.9924</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.03</td>
<td>0.5</td>
<td>0.0320</td>
<td>0.4944</td>
<td>0.8164</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.003</td>
<td>0.2</td>
<td>0.0039</td>
<td>0.2052</td>
<td>0.9283</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.003</td>
<td>0.5</td>
<td>0.0031</td>
<td>0.4978</td>
<td>0.9398</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.03</td>
<td>0.2</td>
<td>0.0326</td>
<td>0.2095</td>
<td>0.9279</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.03</td>
<td>0.5</td>
<td>0.0317</td>
<td>0.5099</td>
<td>0.8629</td>
</tr>
<tr>
<td>300</td>
<td>200</td>
<td>0.003</td>
<td>0.2</td>
<td>0.0035</td>
<td>0.2032</td>
<td>0.9998</td>
</tr>
<tr>
<td>300</td>
<td>200</td>
<td>0.003</td>
<td>0.5</td>
<td>0.0033</td>
<td>0.4967</td>
<td>0.9942</td>
</tr>
<tr>
<td>300</td>
<td>200</td>
<td>0.03</td>
<td>0.2</td>
<td>0.0316</td>
<td>0.2032</td>
<td>0.9520</td>
</tr>
<tr>
<td>300</td>
<td>200</td>
<td>0.03</td>
<td>0.5</td>
<td>0.0307</td>
<td>0.5030</td>
<td>0.8837</td>
</tr>
</tbody>
</table>

The median lifespan \((b - a) / C_0\) of taxa is 0.18 \(N\) sites.

\(N\), number of sites; \(M\), number of taxa; \(c\), probability of false one; \(d\), probability of false zero; \(E(c)\), the expected probability of false one; \(E(d)\), the expected probability of false zero; \(E(\text{corr}(\pi))\), the correlation with the original order.

DOI: 10.1371/journal.pcbi.0020006.s001

#### Table 2. Results on the Large Mammal Dataset

<table>
<thead>
<tr>
<th>(n_t)</th>
<th>(n_s)</th>
<th>(N)</th>
<th>(M)</th>
<th>Chains</th>
<th>(E(c))</th>
<th>(E(d))</th>
<th>CORRMN</th>
<th>CORRDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>124</td>
<td>139</td>
<td>8</td>
<td>0.0113</td>
<td>0.518</td>
<td>0.951</td>
<td>−0.946</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>273</td>
<td>202</td>
<td>2</td>
<td>0.0068</td>
<td>0.661</td>
<td>0.925</td>
<td>−0.912</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>501</td>
<td>139</td>
<td>2</td>
<td>0.0093</td>
<td>0.686</td>
<td>0.704</td>
<td>−0.654</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>526</td>
<td>296</td>
<td>4</td>
<td>0.0033</td>
<td>0.768</td>
<td>0.717</td>
<td>−0.674</td>
</tr>
</tbody>
</table>

\(n_t\), the minimum number of occurrences of a genus; \(n_s\), the minimum number of occurrences of genera per site; \(N\), the number of sites; \(M\), the number of taxa; Chains, the number of chains with likelihood within one standard deviation from the highest likelyhood in 100 experiments; \(E(c)\), the expected probability of false one; \(E(d)\), the expected probability of false zero; CORRMN = \(E(\text{corr}(\pi, MN))\), correlation between the predicted order and MN classification; CORRDB = \(E(\text{corr}(\pi, DBAGE))\), correlation between the predicted order and database age.

DOI: 10.1371/journal.pcbi.0020006.s002

Figure 1. The Pair-Order Matrix \(Q_{ij} = P(\pi(i) < \pi(j))\) between Sites for Dataset \(n_t = 10\), \(n_s = 10\) from the Eight Chains with the Best Likelihood. Black denotes probability one, and white denotes probability zero. For most pairs, the probability is close to zero or one, but some blocks of observations have many different orderings with high probability.

DOI: 10.1371/journal.pcbi.0020006.g001
We further verified the detection of false zeros and ones by preparing two datasets, based on data parametrized by $n_t = 10$ and $n_s = 10$. For the first set, we selected 100 random ones, and flipped them to zero (false zeros). For the second set, we randomly selected 100 zeros, and changed them to ones (false ones). We then performed the analysis, and computed median probability for all zeros to be alive in the first dataset and median probability for all ones to be alive in the second datasets. The probabilities corresponding to 92 of 100 added false zeros was below the median, and 88 of 100 added false ones were above median. The differences are statistically significant when compared to the null hypothesis that the false zeros or ones are equally likely to end up above or below the median (Fisher Sign Test). The median probability that the (site,genus) pair an inserted false one is alive is 0.004, and the median probability that an inserted false zero is alive is 0.92.

We also tested a model where each taxon has its own $c$ and $d$ parameters for false one and false zero. The results of the MCMC runs were almost identical to the ones obtained for the model with one $c$ and one $d$ parameter (unpublished data).

**Discussion**

We have described a probabilistic model for paleontological data and shown that MCMC methods can be used to obtain samples from the posterior distribution of the parameters. The parameters of the model have a natural interpretation, and the hard sites enable us to insert existing prior knowledge of the ordering in a natural way.

The task of finding the optimal ordering, or knowing for certain that a given ordering is optimal, is a very difficult problem. MCMC methods have the advantage of being able to explore various parts of the parameter space, but the issue of guaranteeing convergence of the sampling is always present in these methods. We have solved the problem of convergence by sampling 100 chains in parallel, and taking into account only the chains having the best log-likelihood. We have also checked that the pair-order matrices predicted by these best chains are consistent with each other. This way, we can state with reasonable confidence that our results are indeed an accurate description of the posterior distribution of the model. We also tested the method by adding false zeros and ones to the data randomly, and checking that they were identified correctly.

The results show that for generated data the method is able to reconstruct orderings and locate outliers with excellent accuracy. For the data on large late Cenozoic mammals, the results indicate a high level of agreement with existing orderings and correctly capture the basic feature of paleontological data that false absences are likely to be common and false presences rare.

For the past 40 years the main stratigraphic framework for the study of the Cenozoic land mammals from Europe has been the MN system [30–33]. The MN system rests on a complicated base of taxon appearances and associations that has been.
interpreted somewhat differently by different practitioners [33,34], who nevertheless usually agree on the MN assignment of most sites and virtually always agree on the sequence of MN reference localities. The probability pattern seen in Figure 1 is susceptible to a straightforward paleontological interpretation based on the ages assigned to them within the MN system and a general framework of climate change during the interval studied [35,36]. Starting from the lower left, the sites up to number 50 contain a sequence from the beginning of the Miocene at about 23 million years ago to the main faunal turnover event known in western Europe during this interval, the “Vallesian Crisis,” at about 10 million years ago. Sites 51–63 represent mostly the first million years of post-crisis time, while the large block between sites 63 and 99 represents the relatively stable latest Miocene from 8 million to 5 million years ago, known as the Turolian. A major faunal turnover event separates the Turolian from the Pliocene sites 100–117, and sites 118–124 represent the beginning of the last epoch of the Cenozoic, the Pleistocene. Of the two blocks of localities that the model cannot order well internally, the first and largest thus corresponds to an interval during which little change happened in the mammal faunas. In contrast, the second block, spanning the later Pliocene and the early Pleistocene, corresponds to a time of rapid climatic and faunal change, characterized by the increasingly prominent alternation of cold and warm intervals and, especially in Europe, the cyclic alternation of their attendant “cold” and “warm” mammal assemblages. As emphasized by [37,38], standard biochronology based on taxon ranges breaks down under conditions of cyclic change. The same is obviously true for our model, which assumes genus ranges to be continuous. However, the lack of resolution in this case can also be explained by the simple fact that all included localities are of very similar age, in particular since our data are at the genus level. It remains to be seen to what extent species-level data might resolve such blocks.

The structure just described is also evident in the patterns of Figure 2, where the observed and estimated taxon ranges of the ordered localities trace a pattern of three major faunal units separated by events where many genera go extinct and new, long-lived genera appear. Most marked of these is the turnover event associated with the Miocene–Pliocene transition at 5 million years ago, while the establishment of the two earlier faunal blocks appears more gradual. Figure 2 suggests that the model is especially skeptical of the early occurrences of genera, especially when their record is spotty, as for *Tapirus* (genus number 117 on the lower right).

Most genus occurrences considered by the model to be false are either genuine outliers in time and/or space or actual data errors. For example, of the ten cases at the head of the list for the $n_s = 10$, $n_i = 10$ dataset, *Pliovennerrops* at Laugnac, *Thalassicis* at Wintershof-West, *Anancus* at Concud, and Dorn Dürkheim are true early occurrences, *Pliohystrax* at Pasalar and *Otyrchoerus* at Čandir are isolated early European occurrences of a genus of African origin, *Canis* at Concud is a similar occurrence of a genus dispersing from North America, *Plesiogulo* at Pasalar is a misidentification from a published preliminary list, while *Stephanorhinus* at Belometchetskaya and *Amblycyon* at Stavropol are probable curatorial errors in the NOW database.

Some apparent false occurrences reflect the biology of the animal in question. For example, the genus *Tapirus*, noted above for its spotty record, occurs eight times in the first 50 rows of probable errors. Several carnivore genera are also examples of genuinely rare but long-lived lineages, e.g., *Martes*, *Plesiogulo*, and *Mustela*. It thus appears that the model is a powerful tool to detect not only possible errors in the data but also genera with unusual distributional or ecological characteristics. This suggests that the method could also be of value for the study of the evolutionary dynamics of fossil communities [39,40]. It should be also noted that while the method is powerful and the results are useful, they should still be interpreted by an expert.

In [23], Halekoh and Vach considered an analogous problem in archeology: reconstruction of relative chronology of graves, based on the absence or presence information of finds. They construct a probabilistic model, similar in spirit to ours, but having more nuisance parameters. For example, where the appearance distribution of finds is unimodal and find-specific, i.e., the probability of false zero varies depending on the site and find. They solve the model with an MCMC algorithm, heavily tuned to address the convergence problems, and analyze the resulting modalities in detail. We solve the convergence problem by inserting strong prior information in terms of the hard sites (and fixing the direction of time), optimizing the sampling rules, and analyzing the results from 100 independent runs.

Like Alroy’s [5] method, our model uses relative time, and ignores information about taphonomic regimes, sizes of collections, and biochronology; see [5] for further discussion. Computationally, Alroy’s model has a likelihood function that is based on looking at the probability of not observing a conjunction (co-occurrence) of two taxa. For each taxon $i$, the model uses an individual parameter $K_i$ (called the cryptis parameter), which corresponds to our Lazarus probability $c$. The probability that taxa $i$ and $j$ are not observed together in any site is in Alroy’s model $(K_i + K_j - K_iK_j)$, where $e$ is the number of sites for which the life spans of taxa $i$ and $j$ overlap and $d$ is a parameter that typically is one. The taxon-specific probabilities of false ones and zeros can also be built into our model; the code available on our Web site provides for this. As mentioned, in the experiments this modification did not significantly change the results on real data. An interesting extension to the model would be to allow for temporal intervals for the sites.

One major difference between Alroy’s model and ours is that we use MCMC methods to obtain a sample of the possible parameter values instead of looking for the maximum likelihood solution. This provides additional information about the robustness of the estimates. In particular, we tested our model by randomly adding false zeros and ones, and found that they were identified correctly.

**Materials and Methods**

The probabilistic model. Formally, the dataset has $N$ sites and $M$ taxa, which are described with a $0$−$1$ matrix $X_{nm}$, where $n \in \{1, \ldots , N\}$ and $m \in \{1, \ldots , M\}$. $X_{nm} = 1$ signifies that remains of taxon $n$ have been found from site $m$, while $X_{nm} = 0$ means that no such discovery has been made. In the following, we describe in detail a probabilistic model that we propose has generated this dataset.

First, we assume that the sites appear in some temporal order, denoted by permutation $\pi_{\{1, \ldots , N\}} \rightarrow \{1, \ldots , N\}$. We say that the site $i$ is older than site $j$, if $\pi(i) < \pi(j)$.

We further assume that there exists an ordering for all pairs of sites, i.e., for all $i \neq j$ we have either $\pi(i) < \pi(j)$ or $\pi(j) < \pi(i)$. Strictly speaking, this assumption is not always true. We may have two sites $i$ and $j$ that actually appeared more or less simultaneously. Should this be the case, our probabilistic model should predict
**Dataset.** We used a dataset of European late Cenozoic large land mammals derived from the NOW database (http://www.helsinki.fi/science/how) on 17 July 2003. We restricted the dataset to the Eurasian continent and islands in the Mediterranean Sea, excluding localities with greater than 60 degrees eastern longitude. We also restricted the dataset to the large mammal orders Primates, Creodonta, Carnivora, Perissodactyla, Artiodactyla, Proboscidea, Hyracoidea, and Tubulidentata. We considered three different kinds of age: database age, MN age, and geochronologic age. The MN system [30–33] is a classification of late Cenozoic into 18 classes.

For each locality, we calculated a database age as the mean of the minimum and maximum ages given in the original downloaded file. By MN age, we refer to the mean of the temporal boundaries of MN units according to the correlations given in [11]. This is given only for the subset of localities assigned in the database to a single MN unit or an interval expressed in MN units. For the MN 9 type locality Can Llobateres, we entered a regular MN 9 age in addition to the magnetostratigraphic age provided in the original NOW dataset. We also compiled a new age variable by copying all geochronologic (radiometric or magnetostratigraphic) age data in the original dataset to a separate variable. This new variable, referred to here as geochronologic age, was augmented by data taken from Appendix 2.1 of [11] (the main chronology used in the NOW dataset) and from recent updates for a set of Greek localities [41,42]. The dataset is available on our Web site (http://www.cis.hut.fi/projects/paleo).

We selected further data subsets as follows. First we selected the genera that occurred in at least nt in the original dataset; then we selected the sites in which at least nt such genera had been observed. We used the combinations (nt, ns) = (10,10), (5,5), (2,10) and (2,2). Note that, e.g., in the dataset with nt = 10 and ns = 10 several genera occur less than 10 times, as the selection on the number of genera is done first and then the sites are pruned.

We use the hard orderings of the sites, given by the MN reference sites,

\[ \pi = (\pi_a, 1, b, c). \]

We also denote the prior of all parameters collectively by

\[ P(\theta) = P(\pi) \prod_{n=1}^{N} P(a_n, b_n) P(c) P(d). \]

Given the parameters, we can finally specify the likelihood of the data,

\[ P(X|\theta) = c^c(1 - c)^d (1 - d)^{n-c-d}, \]

where \( c = \sum_{n,m} (1 - \epsilon_{nm})X_{nm} \) is the number of false ones, \( \beta = \sum_{n,m} (1 - \epsilon_{nm}) (1 - X_{nm}) \) is the number of correct zeros, \( \gamma = \sum_{n,m} \epsilon_{nm}(1 - X_{nm}) \) is the number of false zeros, and \( \delta = \sum_{n,m} \epsilon_{nm}X_{nm} \) is the number of correct ones, where we have used auxiliary Boolean parameter \( \epsilon_{nm} \) to signify that the taxon is alive, i.e.,

\[ \epsilon_{nm} = \begin{cases} 1 & \text{if } a_n \leq \pi(n) < b_n, \\ 0 & \text{otherwise} \end{cases} \]

### Table 3. Parameters of Our Model, with Prior Distributions

<table>
<thead>
<tr>
<th>Parameter Type</th>
<th>Notation</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permutation</td>
<td>( \pi )</td>
<td>( P(\pi) =</td>
</tr>
<tr>
<td>Time interval</td>
<td>( [a_n, b_n] )</td>
<td>( P(a_n, b_n) = \log_{0.001}(1 + (N + 1)(N + 2))^{-1}, a_n \leq b_n )</td>
</tr>
<tr>
<td>Probability of false one</td>
<td>( c )</td>
<td>( P(\log c) = \text{Uniform(log 0.001; log 0.1)} )</td>
</tr>
<tr>
<td>Probability of false zero</td>
<td>( d )</td>
<td>( P(\log d) = \text{Uniform(log 0.1; log 0.8)} )</td>
</tr>
</tbody>
</table>

We denote all parameters collectively by \( \theta = (\pi, a, b, c, d) \). We denote the prior of all parameters collectively by \( P(\theta) = P(\pi) P(a_n, b_n) P(c) P(d) \).

DOI: 10.1371/journal.pcbi.0020006.t003

PloS Computational Biology | www.ploscompbiol.org

February 2006 | Volume 2 | Issue 2 | e6

---

\( \pi(\pi | X) = \prod_{n=1}^{N} P(a_n, b_n) P(c) P(d) \)
We are interested in computing various interesting expectation values from the parameter distribution. If we know the posterior distribution, we can compute the expectations from integrand
\[
E_{P(X|\theta)}[f(\theta)] = \int d\theta P(\theta|X)f(\theta).
\] (9)

However, the analytic solution or integration of the posterior distribution is infeasible. Instead of solving the integral of Equation 9 directly, we use numerical integration, namely the MCMC method.

The MCMC algorithm allows us to draw samples from the posterior distribution, without the need to actually solving the Bayes equation. The MCMC algorithm gives us \( T \) samples of the parameters \( \theta^t, \ t \in \{1, \ldots, T\} \), that satisfy \( \theta^t \sim P(\theta|X) \) at the limit of large \( T \). Given the samples, we can then approximate the integral of Equation 9 by
\[
E_{P(X|\theta)}[f(\theta)] \approx \frac{1}{T} \sum_{t=1}^{T} f(\theta^t).
\] (10)

The Markov chain in the name of the MCMC algorithm comes from the fact that a posterior sample is a stochastic function of the previous posterior sample and the data, \( \theta^{t+1} \sim g(\theta^t, X) \). Thus, the MCMC samples form a chain in parameter space: \( \theta^1, \theta^2, \ldots, \theta^T \). The consecutive samples in the chain are not independent. If the chain is too long (\( T \) is small), one chain can effectively cover only a small fraction of the full probability mass.

We first initialize each chain with random values, as follows:
- Initial permutation is drawn from the prior, \( \pi^1 \sim P(\pi) \). The initial ordering of the sites is thus totally random, with the restriction that the "hard" orderings given by the set \( \mathcal{P}_d \) are enforced.
- The initial intervals are set to smallest intervals that have no false ones and zeros, initialized to \( c^1 = 0.01 \) and \( d^1 = 0.3 \).

After the initialization, we run the chain for the \( T_0 = 10,000 \) step burn-in period. The posterior samples drawn during the burn-in period are ignored. The purpose of the burn-in period is to initialize the chain and to bring it to the area of large probability mass in the posterior distribution. We use the final state of the burn-in period, \( \theta^T \), to initialize the actual chain that we use to calculate the expectations. We run the chain for \( T^* = 10,000 \) iterations and save every tenth sample, resulting in \( T = 1,000 \) samples per chain. We then use these stored samples to calculate the desired expectations.

In MCMC methods, the question of convergence always arises. The parameter space may have areas of large probability mass that are separate in the sense that it is very unlikely that a chain jumps from one of these regions to another. It is possible that a chain ends up in these regions, resulting in inaccurate expectations due to the fact that the integration effectively takes only a small subset of the posterior mass into account. Indeed, efficient sampling of the full parameter space is in a general case very difficult problem. The problems of finding the maximum likelihood solution for these types of problems are typically NP-hard [43,44]. Therefore, the best we can do is to build our algorithm so that the chains of large probability mass can be identified with a reasonable confidence (see [29] for review of convergence criteria).

We proceed in two steps: first, we run 100 chains in parallel, and compute the expected log-likelihood, \( E[\log P(X|\theta)] \), of the data for each of the chains. We then compute the standard deviation of the log-likelihoods of the chains and then use only the chains with the expected log-likelihood within one standard deviation of the best chain to account. For example, with the dataset with \( n_s = 10 \) and \( n_t = 10 \) we end up selecting eight of the original 100 chains.

If the predictions given by the chains having the high log-likelihood are consistent with each other, we can conclude that the chains have converged well and that the results are reliable. However, if the predictions given by chains would differ, we could conclude that the chains have converged to separate regions in the parameter space and we also get an estimate for the error due to bad convergence.

Specifically, assume that we analyze \( K \) chains, each with unique initialization and \( T \) samples \( \theta^t \), after the burn-in period. The individual chains produce expectation of a function \( f(\theta) \) of the parameters \( \theta 
\[
F = E_{P(X|\theta)}[f(\theta)] \sim \hat{F} = \frac{1}{K} \sum_{k=1}^{K} f(\theta^t).
\] (11)

The expectation given by all \( K \) chains is given by
\[
F = E_{P(X|\theta)}[f(\theta)] \approx \frac{1}{k} \sum_{j=1}^{k} f(\theta^t).
\]
REVERSE(i, i + 1), denoted by “SAMPLEPI4".

Other parameters. After sampling for the permutation, we proceed to sample the parameters $c$ (probability of false one) and $d$ (probability of false zero). This sampling is done with the aid of the Metropolis-Hastings algorithm. We propose an update to log by sampling the proposal from the normal distribution, $log(\theta) 
\text{N}(log(0.15))$. We account the proposal with the Metropolis-Hastings probability $min(1, P(X|\theta)/P(X|\bar{\theta}))$, where $\theta$ denotes the original parameters and $\bar{\theta}$ the proposed parameters—in this case, the parameters with a new value of $c$. The sampling for the probability of false zero, $d$, proceeds analogously.

To sample for the parameters $a_\ln$ (site where the taxon is first alive), for each $m \in [1, \ldots, N]$, we calculate the relative likelihood of the data for all $a_\ln \in [0, b_\ln]$. This calculation can be done efficiently in $O(N)$ steps. We then normalize these likelihoods to unity and sample new $a_\ln$ from Multinomial($\phi_{a_\ln} + 1$). The sampling for $a_\ln$ proceeds analogously.

Summarizing, one sampling iteration consists of one sampling of $c$, $d$, $a$, $b$ and SAMPLEPI4; and five iterations of SAMPLEPI1, SAMPLEPI2, and SAMPLEPI3. The actual sampling consists of 10,000 of these iterations, of which every tenth sample is used for analysis. The C-implementation of the sampling program is available for download at http://www.cis.hut.fi/projects/patdis/paleo.

Experimental setup. We can visualize a MCMC chain with a pair-order matrix, defined by

$$O^b = \frac{1}{T} \sum_{t=1}^{T} \delta(\pi(i) \lt \pi(j)) \sim \text{EP}(X) \{ \delta(\pi(i) \lt \pi(j)) \},$$

where the indices $i$ and $j$ denote sites, $k$ a particular chain and $b(\cdot)$ is a boolean function that equals one when $\Box$ is satisfied, and zero otherwise. Thus $O^b \delta$ equals the probability that the site $i$ pre-dates the site $j$. Furthermore, the pair-order matrix satisfies $O^b + O^b = 1$. We can also compare two chains by defining a distance measure using the averaged Hellinger divergences [45] over two pair-order matrices, i.e.,

$$d^2(\hat{b}_i, \hat{b}_j) = \frac{1}{\sqrt{N}} \sum_{i,j} \left( \sqrt{O^b_i} - \sqrt{O^b_j} \right)^2,$$

where $N_{\text{pairs}} = \frac{1}{2}N(N - 1)$. The distance measure satisfies $d^2(\hat{b}_i, \hat{b}_j) \in [0, 1]$ and it is equal to zero only if the pair-order matrices are equal. The average Hellinger distance between the pair-order matrices of the eight chains used in the analysis of the dataset with $n_i = 10$ and $n_j = 10$ is $0.010$.

Acknowledgments

We thank three anonymous referees for detailed and constructive comments that improved the manuscript significantly.

Author contributions. KP, MF, and HM conceived and designed the experiments. KP performed the experiments. KP, MF, and HM analyzed the data. KP, MF, and HM contributed reagents/materials/analysis tools. KP, MF, and HM wrote the paper.

Funding. This project was funded in part by the Academy of Finland.

Competing interests. The authors have declared that no competing interests exist.

References

1. Smith W (1816) Strata identified by organized fossils, containing prints on colored paper of the most characteristic specimens in each stratum. London: W. Arding.


