

IMPLEMENTATION OF A MARKOV MODEL FOR PHYLOGENETIC TREES

Erich Bohl
Fakultät für Mathematik
Universität Konstanz
Postfach D 194
78457 Konstanz, Germany
bohl@mathe.biologie.uni-konstanz.de

Peter Lancaster
Department of Mathematics and Statistics,
University of Calgary,
Calgary, Alberta T2N 1N4,
Canada
lancaste@ucalgary.ca

Abstract

A recently developed mathematical model for the analysis of phylogenetic trees is applied to comparative data for 48 species. The model represents a return to fundamentals and makes no hypothesis with respect to the reversibility of the process. The species have been analysed in all subsets of three, and a measure of reliability of the results is provided. The numerical results of the computations on 17,296 triple of species are made available on the internet. These results are discussed and the development of reliable tree structures for several species is illustrated. It is shown that, indeed, the Markov model is capable of considerably more interesting predictions than has been recognised to date.

KEY WORDS: Phylogeny, Phylogenetic tree, Markov process

1 Introduction

A well-known mathematical model of the phenomenon of molecular drift, as a cause for species diversification, uses the notion of a discrete Markov process (Kolmogorov, 1931). After the earliest appearances of this model in the phylogeny literature, research efforts were directed toward improvements obtained by admitting more and

more parameters in the representation of the fundamental “rate” matrix defining the process, and hence in the transition probabilities. This trend began with the one-parameter version of Jukes and Cantor (1969) and was further developed by Kimura (1981), Takahata and Kimura (1981), and Gojobori et al. (1982) (who admitted six parameters). An important milestone in the development of this theory was a paper of Lanave et al. (1984). There, it was claimed that complete freedom in the choice of rate matrix was admitted but, in fact, their analysis was faulty and admits only (the more restrictive) reversible processes. The authors have presented a detailed mathematical discussion of this issue elsewhere (Bohl and Lancaster (2003)).

Notice also that, in the method of Lanave et al., divergence times are computed in terms of eigenvalues of a certain matrix function. This means that, in the case of nucleotides, *three* estimates are obtained for each triple of species, thus imposing unnecessary scatter on the results. Indeed, careful analysis shows that a *unique* estimate is available (see formula (2) below, Zarkikh (1994) and the authors (2003)). Furthermore, this estimate is obtained directly from the estimated divergence matrix.

Two recent accounts of the theory do not progress beyond the 1984 paper of Lanave and co-authors, namely, Chapter 3 of Li’s “Molecular Evolution” (1998) and Chapter 8 of Durbin, et al., (1998). However, Felsenstein (2004) discusses more recent developments in this direction by (among others) Hasegawa, Kishino and Yano (1985), Kishino and Hasegawa (1989), Tamura and Nei (1993), Felsenstein and Churchill (1996), and Schadt, Sinsheimer and Lange (1998). But all of these models are of *reversible* processes and, in our view, this seriously limits their validity. Indeed, Felsenstein asserts that “we cannot rely on reversibility”.

A return to fundamentals is proposed here, in the form of a direct estimation of the rate matrix - *entirely* on the basis of the data: with no intervening simplifying hypothesis. The technique used in this paper does exactly that but, in addition, the process is stabilized by ensuring that estimated divergence matrices have the necessary properties of symmetry and positivity required by the Markov model (see Section 4 and the comments on equation (1)). This strategy implies that no commitment is made with respect to the reversible or irreversible nature of the Markov process. Indeed, we would argue that the use of a *reversible* Markov process is misleading.

Important aspects of the theory include:

- the construction of phylogenetic trees using the relative divergence times of *triples* of species as the basic building block (a direct attack on the construction of trees for many taxa is fraught with difficulties and may require the development of a second numerical processing stage),
- a *data smoothing* technique which is applied to the raw data for each triple, before computing the relative divergence time,
- the introduction of measures of confidence in the predictions of relative divergence times, and

- the fact that the technique can be used with any number of biochemical components (nucleotides, proteins, or codons, for example).

The authors' numerical implementation of the theory (Bohl and Lancaster (2003)) is applied in this paper to nucleotide data from DNA strings for a collection of 48 vertebrate species. Consequently, this presentation and discussion is entirely in the context of molecular drift of *four* nucleotides between DNA sites; although the technique can also be used for broader sets of taxa, as well as different biochemical ingredients of DNA .

Recent decades have also seen considerable development of statistical and simulation methodologies in phylogeny. In this connection we have been reminded of the dangers of “tenuous methodology” involving overlays of confidence intervals associated with statistical techniques (Graur and Martin (2004)). Thus, when properly used, the Markov model of these dynamical systems may provide a firm place to stand in a quagmire of statistical estimates.

Our technique is not a universal cure, of course. There are important issues of phylogeny which cannot be addressed directly with this markovian model. These include integration with paleontological and morphological methods, questions concerning molecular time-scales, and so on. But it is suggested that the Markov model has more to offer than has been recognised previously.

The techniques discussed are made widely accessible by providing numerical codes and directions to an internet site for necessary data. The data consists of comparative counts of nucleotides at DNA sites for 48 vertebrate species. The technique has been applied to all possible triples chosen from these 48 species (17,296 in number), and the results are available in tabular form at the internet site

www.math.ucalgary.ca/~lancaste/dtimes

Here, these results are discussed and are utilised to build some reliable trees with consistent estimates of relative divergence times for up to seven species. The “consistency” referred to here is in the sense that, where different data sets can be used to predict one relative divergence time, they are generally consistent to within four or five per cent (see Section 6) - a quite remarkable property.

The method presents more difficult problems of interpretation when nodes on the same branch of a tree are relatively close in time. Not surprisingly, this tends to occur in those areas where there is ongoing debate concerning the systematics of species, and these are frequently supposed to be related to periods of rapid diversification. These fascinating problem areas are not our immediate concern, but it is possible that this tool is sharp enough to contribute to the resolution of topical problem areas of phylogeny, such as the branching of fish and tetrapods (Arnason et al., 2004), the period of conquest of terrestrial habitats (Brinkman et al., 2004), the human-ape splitting (Hasegawa, et al., 1985, and Tamura and Nei, 1993), or the ordinal relationships of bird species (Meyer and Zardoya, 2003). Section 7, for example, contains a preliminary discussion of the relationship between three bird species.

2 Basic ideas

It is supposed that biological species evolve and diversify through time by molecular drift of nucleotides (or other nuclear materials) among DNA sites or, possibly, by more discontinuous mutational changes in nucleotide distributions among the DNA sites.

A phylogenetic tree has branches which represent biological pathways tracking species changes through time and including branching points, or *nodes*, at which one species may split into two different evolutionary pathways. The branches of a tree link one node to either another node or to an extant species, or *leaf*. The following properties are used in the construction of phylogenetic trees connecting several species:

1. The leaves of the tree (the taxa, or extant species under consideration) have a common (generally unknown) ancestral species. (The fact that the species are extant admits detailed analysis of their DNA properties. The common ancestor is the *root* of the tree, although in our figures it will appear at the top.)
2. For any two leaves, there is a minimal time interval since they had a common ancestor species (i.e. there is a unique associated node and a unique *divergence time* for each pair of leaves).
3. As time increases, two (and only two) pathways emerge from any node. (In the terminology of Li (1998), the nodes are all bifurcations. In the language of combinatorics, only *binary* trees are admitted.)
4. There is a unique pathway (backwards in time) from each leaf to the root of the tree.
5. The tree can be modelled as a Markov process of Kolmogorov type having a fixed (elementwise positive) stationary state vector, q , and an elementwise positive, time dependent, probability transition matrix $P(t)$.

For any pair of species, say A and B , the divergence time mentioned in Property 2 above is denoted by t_{AB} ($= t_{BA}$) and there is an associated *divergence matrix* $S(t_{AB})$ of size 4×4 (in the case of nucleotides). For our purposes, a primitive tree determined by three species has just two associated divergence times (see Figure 1); one associated with each of the two nodes. Their *ratio*, or *relative divergence time* is the information to be estimated by comparing the nucleotide distributions in the DNA of each species of the triple. Thus, the theory depends on experimental estimates of divergence matrices based on comparisons of DNA strings. Corresponding ratios of divergence times are computed from these matrices as in equation (2) below. When all triples of leaves have been compared for, say, n species, the objective is to reconstruct the corresponding n -tree. Note that the number of triples in an n -tree is $\frac{1}{6}(n!/(n-3)!)$; a number which grows rapidly with n (see also Appendix A).

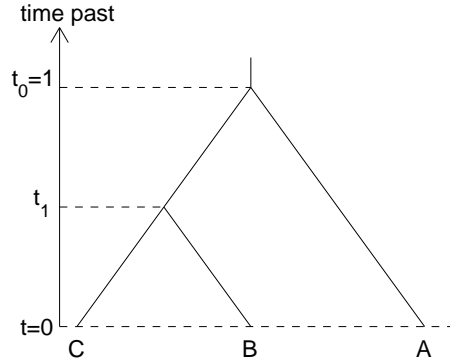


Figure 1: The basic 3-tree

The Markov theory tells us that the experimentally observed transition matrix $S(t)$ can be expressed in terms of the (unknown) stationary vector q and the transition matrix of probabilities, $P(t)$. To write down an explicit formula, first form a diagonal matrix Q from q by writing

$$Q = \text{diag}[q_1^{1/2}, q_2^{1/2}, \dots, q_n^{1/2}],$$

then the model of Property 5 above leads to:

$$S(t_{AB}) = P(t_{AB})Q^2P(t_{AB})^T, \quad (1)$$

where the index T denotes matrix transposition. Clearly, matrix S is both element-wise positive and positive definite. It is convenient to set the time scale *in reverse*, so that $t = 0$ denotes the present time and, generally, time is scaled so that the root of the tree under consideration is set at $t = 1$.

For pairs of nodes which are well-separated in time, the authors' paper of 2003 provides a relatively reliable method for finding the ratios of divergence times from experimental data. In addition, a measure of the reliability of the result is devised in terms of the separation of the nodes. Here, applications of this method are made to the 48 species mentioned above, listed in Appendix C, and for which the raw data from DNA sequences can be found at the website quoted above.

3 The case of three species

As mentioned above, our approach to the development of phylogenetic trees is via the *triples* of species in the investigation. So suppose that three species species A, B , and C are under investigation. It follows from the properties listed above that there is a unique time, t_0 , corresponding to the root of the tree. At this time two pathways were created. Then one of these pathways branches at a subsequent time t_1 to create a total of three pathways terminating at the present time with the three extant species A, B , and C . This situation is sketched in Figure 1 where it is supposed that, at the first branching time t_0 , the pathway to species A does not bifurcate again. Species A is designated the *outlying* species of the triple (in

the terminology of Li (1998), it forms an “outgroup”). Notice that, as sketched, the positions of B and C could be reversed without affecting our conclusions. Also, the whole sketch may be reflected in a vertical line so that the outlier appears on the left. Trees differing only in these ways are treated as the same tree; they are said to be isomorphic. It is important to note that, in our earlier terminology, $t_{AC} = t_{AB} = t_0$, and $t_{BC} = t_1$. Now it follows from equation (1) that $S_{AC} = S_{AB}$, i.e. among the three possible divergence matrices associated with three species, only two are distinct. Thus, observation of these matrices allows us to identify the outlying species. However, the divergence matrices are measured experimentally so they are not known precisely.

The larger “reliable” trees determined below will consist of combinations of triples for which, of the three estimated divergence matrices, two are indeed approximately equal. For the purposes of this study, most of the other triples (the great majority) are disregarded. However, Section 7 contains a closer investigation of a poorly separated triple of bird species.

To summarize the authors’ methods: after averaging and applying a least squares technique to ensure symmetry and a consistent stationary state (see Property 5 above), two *corrected* divergence matrices are obtained. Then relative divergence times are calculated from the corrected divergence matrices using the formula

$$\frac{t_{BC}}{t_{AB}} = \frac{\ln(\Delta^{-1}\det S(t_{BC}))}{\ln(\Delta^{-1}\det S(t_{AB}))} < 1, \quad (2)$$

where $\Delta = \prod_{j=1}^n q_j$. This is all the data required to form the associated phylogenetic 3-tree.

4 Computing with three species

For the benefit of those who may wish to utilise our techniques, some technical discussion is presented in this section.

Consider three extant species A , B , C with three 4×4 estimated matrices of relative nucleotide frequencies N_{BC} , N_{CA} , and N_{AB} . (As noted above, the order in which the pairs of subscripts are written down is irrelevant, i.e. $N_{CB} = N_{BC}$ etc..) Two of these matrices are estimates for the *same* matrix. Knowledge of this matrix tells us which species is the outlier.

First calculate the three spectral norms

$$n_A = \|N_{CA} - N_{AB}\|_s, \quad n_B = \|N_{BC} - N_{AB}\|_s, \quad n_C = \|N_{CA} - N_{BC}\|_s.$$

(The spectral norm is used mainly because MATLAB has a simple command for this norm.) The smallest of these nonnegative numbers determines the outlier. (But note that if t_0 and t_1 are too close to one another, their difference may be completely masked by experimental errors.) Thus, if n_A is the smallest of the three, then species A is assumed to be the outlier, and so on.

To be specific, let us assume that when the outlier is identified it is labelled A (as in Figure 1). This is interpreted as saying that N_{CA} and N_{AB} are estimates of

the same matrix. So (re)define

$$N_1 = \frac{1}{2}(N_{CA} + N_{AB}), \quad N_2 = N_{BC}. \quad (3)$$

(These play the role of the “ N ” matrices of Section 10 of the authors’ earlier paper.) The time from the root to the present is now $t_{CA} = t_{AB}$, and the more recent node has the associated time $t_{BC} < t_{AB}$ (see Figure 1). Now N_1 and N_2 are “normalized” to obtain the first estimates of divergence matrices:

$$T_1 = N_1/(e^T N_1 e), \quad T_2 = N_2/(e^T N_2 e).$$

Then determine improved estimates which are guaranteed to be symmetric and positive definite (cf. equation (13) of Bohl and Lancaster(2003)):

$$S_1 = \frac{1}{4}(T_1 + T_1^T + 2(T_1 T_1^T)^{1/2}), \quad S_2 = \frac{1}{4}(T_2 + T_2^T + 2(T_2 T_2^T)^{1/2}). \quad (4)$$

Now use the least squares procedure of Section 7 of Bohl and Lancaster (2003) to generate matrices S_{AB} ($= S_{CA}$) and S_{BC} from S_1 and S_2 , respectively. These two matrices are our final estimate of the two underlying divergence matrices; as required by the Markov process, they are symmetric and have the same stationary distribution.

Finally, apply (2) to obtain the relative divergence times and hence the phylogenetic 3-tree.

A MATLAB program is included in Appendix B which performs the tasks of this section. It includes the determination of two parameters which give some measure of confidence in the results. Since the Markov model predicts that, without errors, $n_A = 0$, the measured ratio of n_A to $\min(n_B, n_C)$ should be “close” to zero. This ratio is the first of our “confidence ratios”. The second is the ratio of n_B to n_C and should be close to one. The first ratio necessarily lies between zero and one and seems to be the most sensitive. The authors’ numerical experience shows that it can vary from numbers less than 0.1 (which would be seen as very reliable) to more than 0.9, which may be seen as unacceptable, but may also indicate that the nodes in question are not sufficiently well-separated. For the “well-separated” trees to be presented in Section 7, it is required that this ratio is less than 0.61 and, for most triples in those examples, it is considerably less.

One of the disadvantages of this technique is that if, in fact, two nodes occurred relatively close in time, then the determination of an outlier becomes difficult, or impossible. It may be that some of the rejected triples could be resolved by hypothesising a “triple” node and relaxing Property 3 of Section 2. An apparently successful study of such a triple is presented in Section 7.

In Section 5 we consider 4-trees; the next step in a potentially inductive process.

5 Four species

When four species are examined two essentially different tree types are possible and are illustrated in Figure 2. In Figure 2(a) there is, again, a unique outlying species.

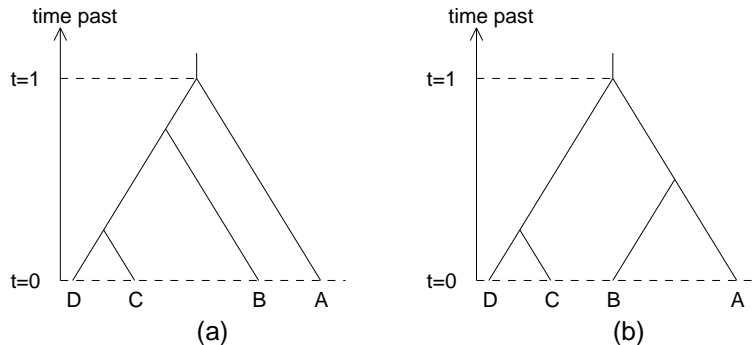


Figure 2: The two possible 4-trees

In the second, species C and D form an outlying pair (an outgroup), and (like A and B) their order can be interchanged in the sketch without affecting the tree structure. Where possible, it is very helpful to identify these two types of 4-tree from the numerical data. Three ways of doing this are suggested. In our experience, the first is the most useful and can immediately be tested using the tabulated results of the website (and there is a nice illustration of this in Section 7).

1. Three of the four possible triples have the same outlier in Case (a). In Case (b), each of the four triples has a different outlier.
2. As there are six different species pairs, six divergence matrices can be computed, but only (with no experimental error) three are distinct. However, in the case of Figure 2(a) and 2(b), respectively, there are either
 - 3 alike, 2 alike, and 1 distinct, or
 - 4 alike, and 2 distinct.
3. Of the 15 possible differences of the six divergence matrices, four will be zero in Case (a) (of Figure 2) and seven will be zero in Case (b); all in the absence of error, of course.

Implementation of these criteria requires that experimental error is not so large that these properties are completely obscured. Since our declared philosophy is to focus on the analysis of *triples*, we do not pursue this line of thought to groupings of five or more taxa, but defer an additional comment to Appendix A.

6 Some well-separated trees

Consider now the tabulated results for all 17,296 triples chosen from the 48 vertebrate species listed in Appendix C. These results can be found at the web-site specified above. Each of the 17,296 rows of data includes:

- Three numbers identifying a triple of species (to be identified via Appendix C of this paper).

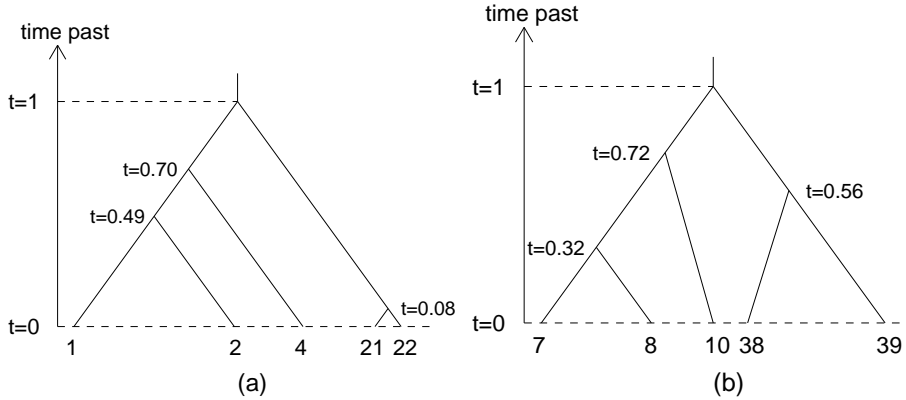


Figure 3: Two 5-trees: (a) conf. level 0.33, (b) conf. level 0.38

- The outlier for the triple.
- The ratio of divergence times (a number which is generally between 0 and 1 but, because of experimental error occasionally exceeds 1 by a small margin).
- Two confidence ratios (preferably close to 0 and 1, respectively.)

Figures 3 and 4 show 5- or 6-trees determined by an (unstructured) examination of these results. As remarked above, many of the triples do not produce favourable confidence levels. However, the trees shown here have the property that every triple of the tree has acceptable confidence levels, thereby increasing our confidence in the whole tree. A confidence level is shown for each tree and refers to the smaller level of the two tabulated (the one which, in an ideal world, would be zero). The level shown for the tree is the largest of these measures for any of the triples in the tree.

The reader is referred to Appendix B for precise identification of the species involved. Informally, note that in Figure 3(a) there are two mammals (1 and 2), a marsupial (4), and two coelacanth (21 and 22). In Figure 3(b) there are two turtles (7 and 8), an ostrich (10), and two fish (38 and 39). In Figure 4a three salmon-like fish appear on one branch (species 36, 37, and 38) and, on the other branch there are two sharks (46 and 47) and a skate (48).

To illustrate the internal consistency of these trees, consider the time ratio 0.70 attached to a node of Figure 3a. This is the relative divergence time for *all* of the triples (1,4,21), (2,4,21), (1,4,22) and (2,4,22) and is an approximation for the four predictions:

$$0.6961, \quad 0.6925, \quad 0.6948, \quad 0.6925,$$

respectively. This kind of agreement reinforces our confidence in the Markov model. Similarly for the time ratio 0.56 appearing in Figure 3(b): This same number should be predicted by three different triples, namely, (7,38,39), (8,38,39) and (10,38,39). The predictions are, in fact, 0.5605, 0.5474, and 0.5596, respectively.

On attempting to add species 15 (an eel) to the tree of Figure 3b, it is found that the triple (7,10,15) has a confidence ratio of 0.82, which may be seen as unacceptably high. An explanation for this may be found in properties of another triple of the

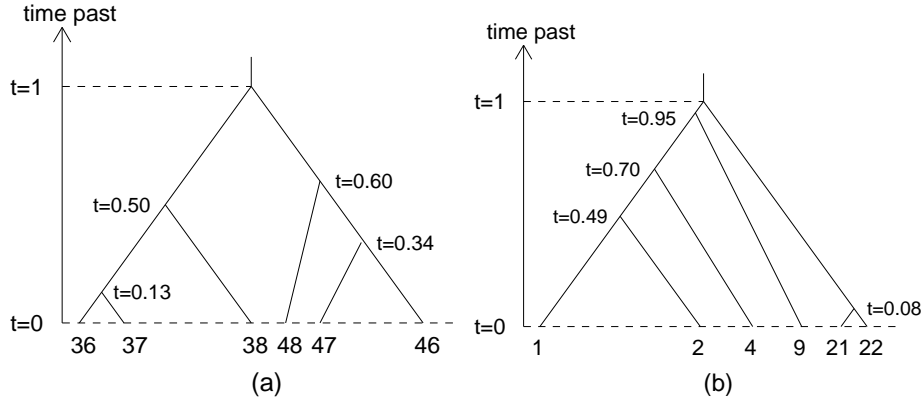


Figure 4: Two 6-trees: (a) conf. level 0.29, (b) conf. level 0.61

extended tree (10,15,39), namely, (turtle, ostrich, porthole fish). Here the relative divergence time is 1.0018 (with relatively good confidence ratio 0.33) which suggest strongly that there is, in effect, a *triple* node at the root of the extended tree.

Figure 4(b) gives the result of adding one more species (species 9, a turtle) to the tree of Figure 3(a). The most obvious effect is a deterioration in the confidence level, from 0.33 to 0.61, which may still be acceptable as a plausible tree is determined. However, although this is not indicated in the sketches, there is now some ambiguity in the time-ratios displayed. For example, the time ratio predicted by the triple 1,2,9 differs from the quotient $0.49/0.95$ by a few percentage points. This kind of inconsistency generally appears as additional leaves are attached to a tree, particularly when confidence levels are poor. This is perhaps the main difficulty encountered in extending trees to larger and larger sets of taxa.

Once again, difficulties with analysis of this 6-tree may be connected with the close proximity of two nodes near the root of the tree (an issue which will arise again in the next section).

The final example of this section is a 7-tree whose leaves are four mammal species (1,2,3,4) and three salmon-like fish (35, 36, 38) (Figure 5a), with a confidence level of 0.53. If species 3 is removed, the remaining 6-tree has an improved confidence level of 0.29 - probably because one of the two neighbouring nodes at times 0.46 and 0.50 has been removed. Concerning the consistency of predicted relative divergence times notice that the nominal figure of 0.42 appearing there is associated with eight different triples and, in fact, the predicted ratios vary between 0.3950 and 0.4207.

More such trees can, of course, be extracted from the tabulated results. These examples are not exhaustive. Also, it may be that larger trees can be constructed by applying quite different techniques as a second numerical processing stage.

7 Nodes in close proximity

For the purpose of illustration, consider first a quadruple of species which gives conflicting results. They are the species (11, 21, 24, 39); a bird, a coelacanth, a

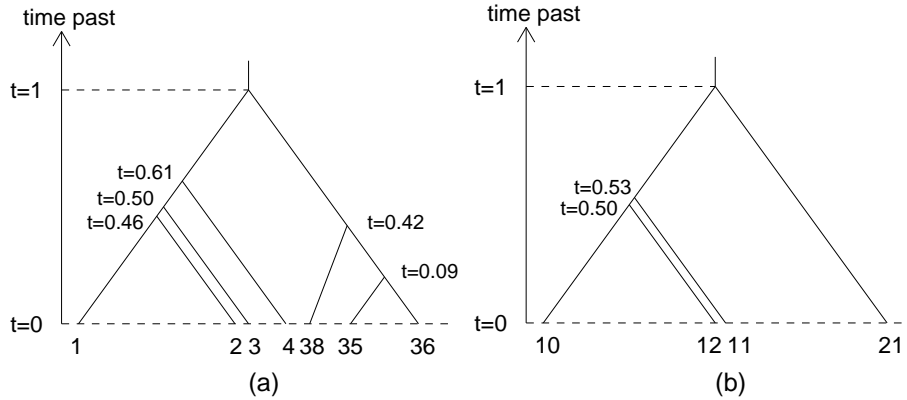


Figure 5: (a) Mammals and salmon-like-fish, conf. level 0.53, (b) Three birds and a coelacanth

sturgeon, and a porthole fish, respectively. The associated figures are tabulated below and the predicted outlier in each triple appears in bold type:

species 1	species 2	species 3	rel. time	conf.ratio
11	21	24	0.659	0.35
11	21	39	1.130	0.70
11	24	39	0.629	0.71
21	24	39	0.910	0.90

Observe first that there must be an error in the predicted outliers; they do not conform to the possible configurations specified in item 1 of Section 5. Observe also that there are two predicted relative times near one. At our present state of knowledge, and without the addition of more information, we do not know of a good interpretation for this data set.

However, it is possible that, by adding further information to that produced by an apparently contradictory data set, useful conclusions can be drawn. In particular, poor confidence ratios are strongly correlated with divergence time ratios near one. Thus, a triple with a poor confidence ratio and a predicted time ratio close to one, may be ripe for deeper investigation.

This idea is illustrated with an example concerning the three bird species appearing in Appendix C (species 10, 11 and 12). (Note that there is some controversy about the ordinal relationships of birds, Meyer and Zardoya (2003).) The time ratio predicted for this triple is 1.10 with a (poor) confidence level of 0.89 and the predicted outlier is species 10. But rather than discarding this result, take it that (relative to other nodes) these three species arose from two nodes which were relatively close together in time (in other words, one may wish to approximate this situation with a *triple* node). The time ratio greater than 1 may arise as an inconsistency between the predicted outlier and the predicted time ratio, and admits the possibility of changing the predicted outlier and taking the reciprocal of the time-ratio.

Let us proceed on this basis and consider an extended tree with four species:

10,11,12 and 21 (a coelacanth). Now the other three triples in this quadruple have acceptable confidence ratios of 0.59, 0.45 and 0.44, and the resulting 4-tree is sketched in Figure 5b. The proposed correction in the predicted outlier of the triple (10,11,12) from 10 to 11 is confirmed by comparing time ratios for the triples (10,11,21), (10,12,21) and (11,12,21) (see Figure 5b). In this way a highly plausible 4-tree is built in spite of the low confidence level associated with the (almost) double-node. Furthermore, the ordering predicted in this Figure is consistent with that predicted by Meyer and Zardoya.

8 Conclusions

A Markov model for the determination of phylogenetic trees from comparative DNA data of triples of species has been developed and implemented numerically. Significant new features of this method include a preliminary data smoothing process, and a model which makes no hypotheses on the rate matrix of the Markov process used to model the dynamical system. Hence, no assumptions are made concerning the reversible/irreversible nature of the process.

The model has been applied to data collected from 48 vertebrate species, and the results have been posted at a website. Measures of reliability of the results have been formulated and shown to be closely linked with the notion of clustered, or multiple nodes. The technique has been applied in the extraction of consistent data sets for some (“well-separated”) trees of modest size (up to seven species). A feature of this analysis is remarkable agreement in the multiple predictions of relative divergence times for the nodes of such trees. These firm predictions of relative divergence times seem to be a unique and valuable attribute of this line of attack. It is no accident that published tree structures rarely, if ever, give careful reference to relative divergence times; their focus is generally on the tree structure only.

We emphasize that the results presented here are the tip of the iceberg. Much more information is to be gained from analysis of the computational results already made available at the website but, more importantly, there is great potential for application of this technique to other sets of taxa using other DNA data banks.

Further research on the technique itself could include closer examination of the treatment of clustered nodes, and implementation of the existing algorithm in the context of the larger protein or codon data sets.

ACKNOWLEDGEMENTS: This work would not have been possible without the generous provision of data by Prof. Axel Meyer, to whom the authors are very grateful. Also the authors are happy to acknowledge the assistance of I. Hendekovic and E. Luik in the preparation of computer programs and, in the case of I. Hendekovic, for patiently explaining biological background to the authors.

Appendix A. Any number of species

Consider briefly binary trees with n leaves (taxa). There are necessarily $n - 1$ nodes the earliest of which (in time) is the *root* of the tree. Let c_n denote the number of distinct trees associated with n leaves. For example, it has been seen above that $c_3 = 1$ and $c_4 = 2$. It is natural to define $c_1 = c_2 = 1$, and we also set $c_0 = 1$. For any integer, n , denote the integer part of $n/2$ by $[n/2]$. By applying a counting procedure to the root of a tree one arrives at the recursion ¹

$$c_n = c_1 c_{n-1} + c_2 c_{n-2} + \dots + c_{[n/2]} c_{[(n+1)/2]}, \quad (5)$$

valid for all positive integers n . Beginning with c_1 this yields the rapidly increasing sequence

$$1, 1, 1, 2, 3, 6, 11, 24, 47, 103, \dots$$

(known as the half-Catalan sequence). Thus, for 8 extant species there are 24 essentially different possible phylogenetic trees. Given the levels of error in the data sets examined to date, the task of deciding *a priori* which tree type is most appropriate will quickly become very difficult. This supports the approach adopted in this paper; namely, to analyse larger trees via the set of all triples of taxa.

Appendix B: A “matlab” program for 3 species

```
COMMENT {Given 3 4-by-4 matrices of positive integers, N23,N31,N12, this program computes the ratio of divergence times (reltime) ordered to give a number less than one (in physically sensible cases). The first steps identify the outlying species (a message is printed identifying the outlier as species 1,2, or 3) and the two data matrices N1 (associated with the outlier) and N2, (independent of the outlier).}
```

```
n1=norm(N31-N12);
n2=norm(N23-N12);
n3=norm(N23-N31);
n=min([n1 n2 n3]);
if n==n1
c=1;
elseif n==n2
c=2;
else
c=3;
end
if c==1
N2=(N31+N12)/2; N1=N23; 'Outlier is species 1'
elseif c==2
N2=(N23+N12)/2; N1=N31; 'Outlier is species 2'
else
```

¹The authors are grateful to Ludwig Elsner for this observation

```

N2=(N23+N31)/2; N1=N12; 'Outlier is species 3'
end
COMMENT {The next steps compute the divergence matrices S1 and S2 adjusted
to be symmetric and pos.def. and to have a consistent stationary state.}
e=[1 1 1 1];
T1=N1/(e*N1*e');
T2=N2/(e*N2*e');
S1=(T1+T1'+2*sqrtm(T1*T1'))/4;
S2=(T2+T2'+2*sqrtm(T2*T2'))/4;
l=(0.5)*(S1-S2)*e';
for i=1:4
for j=1:4
K(i,j)=(l(i,1)+l(j,1))/2;
end
end
S1=-K/2+S1;
S2=-K/2+S2;
COMMENT {Now compute the relative divergence time, reltime}
q=S1*e';
del=q(1)*q(2)*q(3)*q(4);
reltime=(log(det(S1)/del))/log(det(S2)/del)
COMMENT {There are two confidence ratios. The first should be small and less
than one - the smaller the better. The second should be close to +1.}
if n==n1
m=min([n2 n3]);
r2=min([n2/n3 n3/n2]);
elseif n==n2
m=min([n1 n3]);
r2=min([n1/n3 n3/n1]);
else
m=min([n1 n2]);
r2=min([n1/n2 n2/n1]);
end
r ='Confidence ratios are'
[n/m, r2]

```

Appendix C. The 48 species

(GenBank Accession Number, <http://www.ncbi.nlm.nih.gov>)

1. *Felis catus* (cat) (U20753)
2. *Bos taurus* (cow) (V00654)
3. *Oryctolagus cuniculus* (rabbit) (AJ001588)
4. *Macropus robustus* (wallaroo) (Y10524)
5. *Didelphis virginiana* (opossum) (Z29573)
6. *Ornithorhynchus anatinus* (platypus) (X83427)
7. *Chrysemys picta* (eastern painted turtle) (AF069423)
8. *Chelonia mydas* (green sea turtle) (AB12104)
9. *Pelomedusa subrufa* (African helmeted turtle) (AF039066)
10. *Struthio camelus* (ostrich) (Y12025)
11. *Corvus frugilegus* (rook) (Y18522)
12. *Falco peregrinus* (peregrine falcon) (AF090338)
13. *Alligator mississippiensis* (American alligator) (Y13113)
14. *Eumeces egregius* (mole skink) (AB016606)
15. *Typhlonectes natans* (rubber eel) (AF154051)
16. *Xenopus laevis* (African clawed frog) (Y10943)
17. *Mertensiella luschani* (Lycian salamander) (AF154053)
18. *Lepidosiren paradoxa* (S.American lungfish) (AF302934)
19. *Protopterus dolloi* (African lungfish) (L42813)
20. *Neoceratodus forsteri* (Queensland lungfish) (NC_003127)
21. *Latimeria menadoensis* (Sulawesi coelacanth) (AF176901)
22. *Latimeria chalumnae* (African coelacanth) (U82228)
23. *Polypterus ornatipinnis* (ornate bichir) (U62532)
24. *Acipenser transmontanus* (white sturgeon) (AB042837)
25. *Amia calva* (bowfin) (AB042952)
26. *Osteoglossum bicirrhosum* (arawana) (AB043025)
27. *Pantodon buchholzi* (butterfly fish) (AB043068)
28. *Conger myriaster* (conger eel) (AB038381)
29. *Anguilla japonica* (Japanese eel) (AB038556)
30. *Carassius auratus* (goldfish) (AB006953)
31. *Cyprinus carpio* (common carp) (X61010)
32. *Danio rerio* (zebrafish) (AC024175)
33. *Crossostoma lacustre* (tasseled-mouth loach) (M91245)
34. *Sardinops melanostictus* (Japanese sardine) (NC_002616)
35. *Oncorhynchus mykiss* (rainbow trout) (L29771)
36. *Salmo salar* (Atlantic salmon) (U12143)
37. *Coregonus lavaretus* (Lake chud whitefish) (AB034824)
38. *Plecoglossus altivelis* (ayu fish) (AB047553)
39. *Diplophos taenia* (Pacific porthole fish) (NC_002647)
40. *Aulopus japonicus* (Japanese thread-sail fish) (NC_0002674)
41. *Trachurus japonicus* (Japanese jack mackerel) (NC_002813)
42. *Paralichthys olivaceus* (Japanese flounder) (AB028664)
43. *Arctoscopus japonicus* (sailfin sandfish) (AP003090)
44. *Polymixia japonica* (beardfish) (NC_002648)
45. *Gadus morhua* (Atlantic cod) (X99772)
46. *Squalus acanthia* (spiny dogfish) (Y18134)
47. *Mustelus manazo* (shark) (AB015962)
48. *Raja radiata* (starry skate) (AF106038)

SPECIES BY GROUPS:

1-4	mammals	23-25	old freshwater fish
5-6	marsupials	26-27	bonytongues
7-9	turtles	28-29	eels
10-12	birds	30-34	carp
13-17	reptiles/amphibians	35-45	salmon-like
18-20	lungfish	46-47	shark
21-22	coelacanth	48	skate

LIST OF REFERENCES

- Arnason, U., Gullberg, A., Janke, A., Joss, J. A., and Elmerot, C., 2004, Mitogenomic analyses of deep gnathostome divergences: a fish is a fish, *Gene*, **333**, 61-70.
- Brinkman, H., Derk, A., Zitzler, J., Joss, J. A., and Meyer, A., 2004, Complete mitochondrial genome sequences of the South American and Australian lungfish:..., *Journal of Molecular Evolution* **59**, 834-848.
- Bohl, E., and Lancaster, P., 2003, Irreversible Markov processes for phylogenetic models, *Numerical Linear Algebra with Applications*, **10**, pp. 577-593.
- Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., 1998, *Biological Sequence Analysis*, Cambridge University Press.
- Felsenstein, J., 2004, *Inferring Phylogenies*, Sinauer Associates, Inc., Sunderland, Mass., USA.
- Felsenstein, J., and Churchill, G.A., 1996, A hidden Markov model approach to variation among sites in rates of evolution, *Molecular Biology and Evolution*, **13**, 93-104.

- Gojobori, Y., Ishii, K., and Nei, M., Estimation of average numbers of nucleotide substitutions when the rate of substitution varies with nucleotide, *Journal of Molecular evolution*, **18**, 414-423.
- Graur, D., and Martin, W., 2004, Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision, *Trends in Genetics*, **20**, no.2, 80-86.
- Hasegawa, M., Kishino, H., and Yano, T., 1985, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, pp. 160-174.
- Jukes, T. H., and Cantor, C. R., 1969, *Evolution of protein molecules*, In Munro H. N. (Ed.), *Mammalian Protein Metabolism*, vol.3, pp.21-132, Academic press, New York.
- Kimura, M., 1981, Estimation of evolutionary distances between homologous nucleotide sequences, *Proceedings of the National Acadademy of Sciences, USA*, **78**, 454-458.
- Kishino, H., and Hasegawa, M., 1989, Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea, *Journal of Molecular Evolution*, **29**, pp.170-179.
- Kolmogoroff, A., 1931, Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung, *Mathematische Annalen*, **104**, 415-458.
- Lanave, C., Preparata, G., Saccone, C., and Serio, G., 1984, A new method for calculating evolutionary substitution rates, *Journal of Molecular Evolution*, **20**, 86-93.
- Li, Wen-Hsiung, 1998, *Molecular Evolution*, Sinauer Associates, Inc., Sunderland, Mass., USA.
- Meyer, Axel, and Zardoya, Rafael, 2003, Recent advances in the (molecular) phylogeny of vertebrates, *Annu. Rev. Ecol. Evol. Syst.*, **34**, pp.311-338.
- Schadt, E. E., Sinsheimer, J., S., and Lange, K., 1998, Computational advances in maximum likelihood methods for molecular phylogeny, *Genome Research*, **8**, pp. 223-233.
- Takahata, N., and Kimura, M., 1981, A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes, *Genetics*, **98**, pp. 641-657.
- Tamura, K., and Nei, M., 1993, Estimation of the number of nucleotide substitutions in the the control region of mitochondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*, **10**, pp. 512-526.
- Zharkikh, A., 1994, Estimation of evolutionary distances between nucleotide sequences, *Journal of Molecular Evolution*, **39**, pp. 315-329.