# *Phylogenetic Trees Predicted by an Irreversible Markov Process*

Bohl, Erich and Hamilton, Ryan and Lancaster, Peter

2008

MIMS EPrint: **2008.58**

ISSN 1749-9097

# PHYLOGENETIC TREES PREDICTED BY AN IRREVERSIBLE MARKOV PROCESS

Erich Bohl
Fakultät für Mathematik
Universität Konstanz
Postfach D 194
78457 Konstanz, Germany.

Ryan Hamilton and Peter Lancaster
Department of Mathematics and Statistics,
University of Calgary,
Calgary, Alberta T2N 1N4,
Canada

August 1, 2007

**Abstract**

A new Markovian method for the prediction of phylogenetic trees has been developed earlier by the authors. Here, the method is illustrated and applied using mitochondrial DNA data for vertebrate species (and technicalities of the theory are avoided). Discussions include the sensitivity of the results to DNA alignment techniques and the inclusion of polytomy in tree structures. Several comparisons are made with tree structures in the literature which have been predicted using statistical techniques.

*Keywords:* Phylogenetic relationships, irreversible Markov process, vertebrates, rooted trees.

## 1 Introduction

Two recent publications of Bohl and Lancaster (2003, 2006) show that a mathematical model of speciation can produce new and interesting results when applied to a monophyletic group. This model is based on the idea of molecular drift of nucleotides between DNA sites, and utilises a suitable Markov process. It provides an alternative to the use of statistical methodologies for tree-building (maximum likelihood, Bayesian analysis, parsimony, et al.). Some inherent dangers in the use of such methods are underlined by Graur and Martin (2004) and are completely avoided here. It can be argued that the method proposed here is less subjective than the several tree-searching statistical techniques currently available and, in the hands of biologosts, could become a valuable tool.

Early use of a highly simplified Markovian model was made by Jukes and Cantor (1969) and there have been several further developments since then some of which are summarized by Yang (1994) and Li (1997, Chapter 3) (see also the two Bohl-Lancaster papers). However, with few exceptions they focus on the more tractable reversible processes. Here, there is no hypothesis with respect to reversibility. In contrast, the data determines the nature of the process and the structure of the underlying phylogenetic tree, including the determination of its root. Thus, for each species pair, the model includes a "fully connected" matrix of transition probabilities among the four nucleotides. Furthermore, the data leads directly to the topology of the branches and the root of the tree (cf. Huelsenbeck et al. (2002), and the discussion of Durbin et al. (1998), p.68).

Features of this implementation of the Markov model include:

- A smoothing technique for the raw DNA data analysed in terms of nucleotide distributions and designed to minimise the effects of data errors.
- Implementation of an unrestricted Markov model making no assumptions with respect to reversibility.
- Predictions of relative divergence times bring a new quantitative feature into phylogeny.
- A high degree of internal consistency in predicted phylogenetic trees.
- The root of the tree is determined by the data.

It is our objective to give a non-technical account of some illustrative investigations using the Bohl-Lancaster algorithms, and the reader is referred to the earlier papers for technical details. In particular, we demonstrate the usefulness of the theory when applied to mitochondrial DNA data of 48 vertebrate species (listed in Appendix A); where a common alignment process has been applied resulting in a common sequence length of about 2000 base genome pairs. This is minimal by current standards, but it makes a convenient source on which to test the new methodology. We also consider the effects of re-alignment in smaller subsets of species for comparison with predictions made with longer sequence lengths. It should be emphasised that the necessary algorithms and software are now available and can, of course, be implemented on data sets for any species or sequence lengths.

It is not appropriate to venture into the theory here, but some remarks on "reversibility" are in order - based on the discussion of Bohl and Lancaster (2003). Given the fundamental matrix $P(t)$ of transition probabilities mentioned above, a second process with transition matrix $\Pi(t)$ is determined entirely by $P(t)$; this is the "reverse" process. The first process is reversible or irreversible according as $\Pi(t) = P(t)$, or $\Pi(t) \neq P(t)$. However, there is no physical evidence suggesting

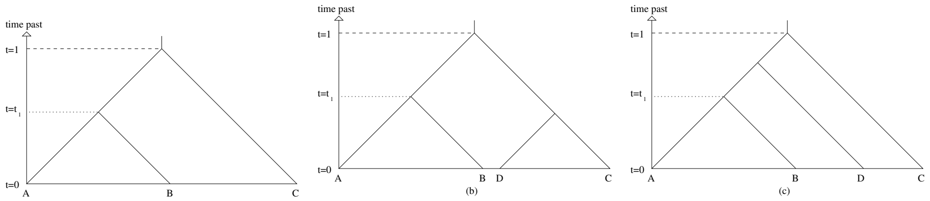

Figure 1: The basic 3-tree

that, in modelling the drift of nucleotides, equality holds. Hence our insistence on the irreversible model, and on careful estimation of $P(t)$[1] .

The comparison of just three species is the basic tool employed. Initially, it is assumed that, of any three species admitted to the analysis, there is a well-defined *outlier* and, for the corresponding tree with two nodes, there is a well-defined *relative divergence time*. For example, in Fig. 1 the three species are denoted by A, B, C, and C is the outlier. A figure like 1(a) would be appropriate when considering three species in isolation. However, if more than three taxa are included in the analysis, the *same triple* may be configured as in 1(b) or 1(c). Higher order nodes can also play an important role, but discussion of this possibility is deferred until Sections 8 and 9.

For a binary tree of $n$ species there are $\frac{1}{6}(n!)/(n-3)!$ possible triples, and for each triple, an outlier and a relative divergence time are predicted by the model. Using DNA data for the 48 species of Appendix A there are 17,296 triples. The Bohl/Lancaster method has been applied to each of these triples and the results are tabulated at the website

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

It will be seen that the results do not clearly determine a unique tree for all of the 48 species. However, the present method can be utilised to examine interesting subsets of species chosen from the 48. In Sections 8 and 9 a more comprehensive approach is taken to the construction of trees.

The results at the website are tabulated as 17,296 rows of numbers, each row having seven entries. The first three denote the three species in question, the fourth and fifth are, respectively, the predicted outlier (denoted by 1,2, or 3) and relative divergence time. The sixth and seventh are denoted (here) by “cr1” and “cr2” and record the two so called *confidence ratios*. This format will also be used in this paper (see Appendix B, for example).

Further illustrations of the method are given with smaller subsets of species admitting *re-alignment* of the DNA sequences. This results in longer common sequence lengths and, after recalculation of divergence matrices, provides different estimates of divergence times. We note also that, with increasing length of DNA sequences, the divergence matrices become increasingly diagonally dominant. Re-alignments have been made here using the “ClustalW” software. “ClustalX” was used in the

[1] The transition matrix $P(t)$ is not estimated directly: it is deduced from experimental measurements of the “divergence matrix”.

initial data set - which was also used by Brinkmann et al. (2004) (see interesting comments on alignment in Grant et al. (2003)).

Although it is difficult to generate an unambiguous tree for the full set of 48 vertebrates, some larger trees generated from this data (without realignments) are presented and discussed in Section 9.

Some questions of current interest in phylogenetics play no role in this work. Because the Markov model used here is "homogeneous in time", it is neutral with respect to calibration of time scales with geological events, morphology, and the possibility of differing "rates of evolution" at different epochs or on different portions of a given tree. The results discussed here are not a complete biological analysis; they are the authors' attempts to illustrate the potential and limitations of the underlying Markov model in phylogeny.

## 2 The confidence ratios

We wish to avoid technicalities as far as possible, but it is necessary to spend some time on the meaning of the confidence ratios. When examining a triple of species, say A, B, and C, it is necessary to compute three matrices of relative nucleotide frequencies (of size 4 × 4) one associated with each pair. Denote these matrices by $N_{BC}$, $N_{CA}$, $N_{AB}$. If the data fits the Markov model precisely two of these matrices will be identical. For example, if C is the outlying species and there is no error, then $N_{BC} = N_{CA}$. The confidence ratios measure to what extent this occurs with the *observed* matrices. Thus, if matrix "sizes" are measured in terms of norms, say

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

then (with C the outlier), $n_C$ should be considerably less than the least of $n_A$ and $n_B$ (ideally, it would be zero). So the first measure of confidence is

$$\text{cr1} = \frac{n_C}{\min(n_A,\, n_B)}.$$

So $0 \leq$ cr1$< 1$ and, ideally, cr1=0 but, in experimental practice it never is. In fact, if all the numbers cr1 generated by all the triples of a given tree are less than 1/3, say, the results are interpreted as being highly significant.

The second confidence ratio (with C the outlier) is cr2 and is defined to be the smaller of the two ratios $n_A/n_B$, $n_B/n_A$. Ideally, cr2 is close to one. In practice this ratio is in the range 0.8 to 1.0. As Appendix B shows, it is not so sensitive as cr1 and we seldom appeal to cr2 in the discussion of results.

It turns out, of course, that the Markov model is capable of limited resolution of events in time. So what is to be expected if we compare three species which emerged from the same ancestral species in a "short" period of time - relative to the power of resolution of the model? In a period of "rapid diversification" of species? It is not hard to see that we should anticipate that the numbers $n_A$, $n_B$, $n_C$ are not well separated, and this implies that cr1 will be close to one. Thus, a value for cr1 close to one (when coupled with a predicited divergence time close to one) does not mean a complete loss of confidence in the results, although *the corresponding prediction of an outlier will not be reliable*. When this phenomenon is observed in the computed results, it may indicate rapid diversification of species and is an invitation to deeper investigation. This is the kind of situation examined in more detail in Sections 4-9 below.

## 3 First experiments; well-separated trees

By definition, a *well-separated tree* has acceptable confidence ratios (i.e. cr1 is considerably less than one) for all possible triples of species represented. It turns out that, for $n$ larger than six, a well-separated tree is hard to find in this data set. One reason for this seems to be that, at least with this data, and with larger groups, there will generally be at least one pair of nodes which are relatively close in time. As suggested above, times close to one are generally linked with a value of cr1 close to one; which is seen as a "poor" confidence ratio leading to an unreliable prediction of an outlier. Such situations suggest a closer examination of, say, three poorly resolved species after re-alignment.[2] In this section we consider some well-separated trees and, subsequently, will analyse some topical cases involving nodes in "close" proximity.

The first example appears in Figure 2a and includes six species, namely three mammals (1,2,4) and three ray-finned fish (36,37,38). The data for the twenty triples of this example (extracted from the website) appears in Appendix B. To illustrate, the second and third rows of the table indicate that in each of these two triples, the fish is the outlier and the predicted divergence times of 0.4696 and 0.4695 are remarkably close. Values of the confidence ratio, cr1, of 0.1413 and 0.1389 give some confidence in these predictions.

Scanning the whole of this table, it is seen that the largest confidence ratio, cr1, among the twenty triples is 0.2926, so this is attached to the whole tree. The two triples, for the mammals 1,2,4, and the fish, 36,37,38 are well-defined with good confidence ratios cr1 0.19 and 0.24, respectively. Figure 2a then represents the results obtained from the twenty triples chosen from the six species and, among these triples, the largest value of cr1 is a modest 0.29; so this qualifies as a "well-separated" tree.

An important property of well-separated trees is as follows: For a tree of more than three species, the time associated with a particular node can be determined using more than one triple. It is found that for a well-separated tree, the divergence times representing the same node agree with each other to a high degree of precision. In effect, different experiments lead to the same conclusion.

To illustrate this phenomenon, consider Figure 2a and, in particular, the node associated with the time $t = 0.67$. There are six estimates of this time from the triples (1,4,36), (1,4,37), (1,4,38), (2,4,36), (2,4,37), and (2,4,38). The six times listed in Appendix B vary between 0.642 and 0.686 and their average appears on

[2] Indeed, it may suggest that the whole exercise be repeated after separately aligning the data for each triple of species.

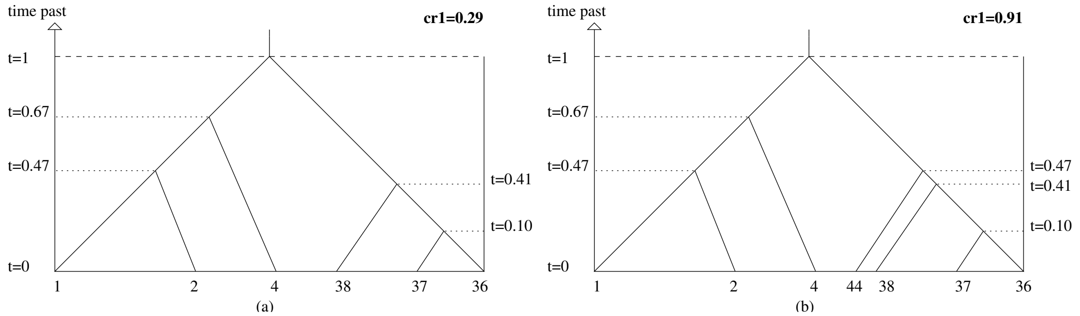

Figure 2: (a) A well-separated 6-tree. (b) The same tree with species 44.

Figure 2a as $t = 0.67$.

Similarly, the node at time $t = 0.47$ is associated with (1,2,36), (1,2,37) and (1,2,38), and the predicted divergence times of these three triples are 0.4696, 0.4695, and 0.4613, respectively. Furthermore, the predicted time for the (1,2,4) triple may be multipled by the predicted times for the (1,4,36), (1,4,37), and (1,4,38) triples giving three additional predicted divergence times for the (1,2,36) node: 0.4770, 0.4746, and 0.4631, respectively. Again, the noted value $t = 0.47$ is an average of six estimates.

Typically, on examination of a set of well-separated species, similar results will be obtained. However, it is found that for our mtDNA data set for 48 species, well-separated trees with $n > 7$ are rare. For example, if we add a beardfish (44) to the tree of Figure 2a, the divergence time predicted for the (1,38,44) triple is 0.5312 and confidence ratio cr1 deteriorates from 0.29 to 0.40. This deterioration generally appears as one continues to add species to an otherwise well-separated tree. Indeed, the overall (displayed) value of cr1 = 0.91 arises at the "difficult" triple (36,38,44). The divergence time (at 0.8760) is consistent with the rest of the tree, but this induces the poor confidence ratio, cr1 = 0.91. If this item is disregarded, the overall confidence ratio cr1 for Figure 2b is 0.53 (attained at the triple (37,38,44)). Thus, unreliable results for one triple do not invalidate the whole tree.

A less dramatic effect is produced by adding another mammal to Figure 2a (the rabbit, species 3, instead of the beardfish, 44). In this case, there is a relative divergence time of 0.8840 for the node (1,2,3), suggesting almost confluent nodes. Nevertheless, the parameter cr1 for this node (and hence the 7-tree) is a more favourable 0.4953 (and all other additional triples have cr1 less than the ratio 0.2926 of the original 6-tree).

It is natural to suppose that a tree consisting of species chosen one-by-one from groups which are morphologically very different will lead to a well-separated tree. To examine this possibility consider first four species, a mammal (represented by the rabbit, 3), a marsupial (represented by the opossum, 5), a turtle (represented by the eastern painted turtle, 7), and a bird (represented by the ostrich, 10). The tree is shown as Figure 3 and has a disappointing confidence ratio, cr1= 0.67; it does not really qualify as "well-separated".

On the other hand a representative of a fifth group, species 39, a fish, can be added to make a 5-tree with no further deterioration of the over-all confidence ratio,

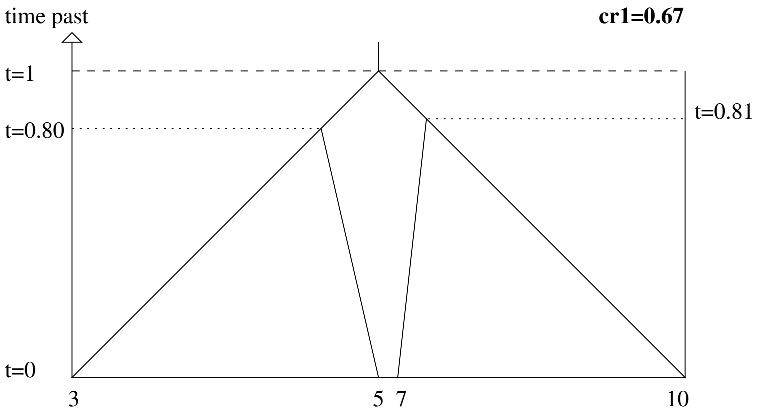

Figure 3: Species from four different groups.

cr1= 0.67, and the fish is the outlier for the 5-tree (i.e. a new "root" is created).

One interpretation of such results is that the power of resolution of the model decreases as the times become more remote.

# 4 Phylogenetic relationships of birds

As a first case study we take up an investigation of Bohl and Lancaster (2005) concerning the ordinal relationships of birds (see also Meyer and Zardoya (2003)). Our data set includes just three birds: species 10 (ostrich), 11 (rook), and 12 (peregrine falcon). The results for the triple (10,11,12) (see table below) show that the relationship of these species is not immediately apparent; the Markov model fails to resolve their ordinal relationships. It is predicted that the ostrich is the outlier, but the time ratio greater than one is not physically sensible, and merely tells us that the model is not able to resolve this issue at first glance.

Our next move is to consider trees of 4 species where the fourth species will serve as an outlier to the three birds. For example, consider the quadruple (10,11,12,21), three bird species and a coelecanth:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| **10** | 11 | 12 | 1 | 1.0961 | 0.8924 | 0.8474 |
| 10 | 11 | **21** | 3 | 0.5155 | 0.5885 | 0.7662 |
| 10 | 12 | **21** | 3 | 0.4980 | 0.4525 | 0.7611 |
| 11 | 12 | **21** | 3 | 0.5358 | 0.4378 | 0.9414 |

Obviously, the data for the triple (10,11,12) is not to be interpreted literally. A divergence time greater than one (with an associated poor confidence ratio) simply indicates relatively rapid diversification between the three species. As noted in Section 7 of Bohl and Lancaster (2005), the presence of the fourth species (21-the coelecanth) suggests that, in the (10,11,12) triple, it is indeed 11 that is the outlier (i.e. the 10,12 bifurcation occurred more recently) . But the confidence ratio of this 4-tree is still 0.59, which is not entirely satisfactory, and by searching for another species as the outlier, it may be possible to improve this.

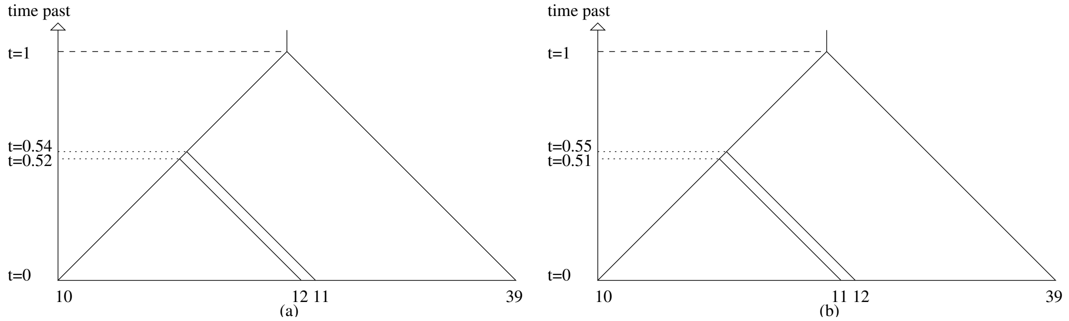

Figure 4: 3 birds and the porthole fish:
(a) From the data for 48 species. (b) With re-aligned data.

Indeed, it is found that if the Pacific porthole fish (39) is used as the outlier, the next table is obtained (see also Fig 4a). The prediction that, of the triple (10,11,12), 11 is the outlier is maintained, but with an improved confidence ratio of 0.41.

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| **10** | 11 | 12 | 1 | 1.0961 | 0.8924 | 0.8474 |
| 10 | 11 | **39** | 3 | 0.5227 | 0.4124 | 0.8762 |
| 10 | 12 | **39** | 3 | 0.5171 | 0.2809 | 0.9637 |
| 11 | 12 | **39** | 3 | 0.5612 | 0.3396 | 0.9616 |

Clearly, the search for an optimal outlier can be made systematically, but we do not go into detail. However, it is emphasised that the identification of the rook (11) as the outlier of the three birds, rook, ostrich, and paregrine falcon, is confirmed with several outliers, thus superseding the prediction made on examination of the (10, 11, 12) triple in isolation. Although the confidence level of the 4-tree of Figure 4a is dominated by that of the (10,11,12) triple we can, nevertheless, associate a confidence level of 0.41 with that tree.

In any case, it is clear that, based on the data set for 48 species, the predictions are delicately balanced. Given the desirability of maximizing the amount of common DNA material employed (as emphasized by Ruvolo (1997)), a re-alignment of the genome sequences for the four species of Figure 4a was performed, resulting in an increase from approximately 2000 to 6000 base-pairs. The results are as follows, and make an interesting contrast with those of the preceding table.

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 10 | 11 | **12** | 3 | 0.9145 | 0.4196 | 0.8470 |
| 10 | 11 | **39** | 3 | 0.5141 | 0.0633 | 0.9979 |
| 10 | 12 | **39** | 3 | 0.5431 | 0.1241 | 0.9944 |
| 11 | 12 | **39** | 3 | 0.5555 | 0.1456 | 0.9970 |

They are summarised in Figure 4b - which must be seen as more reliable than 4a. Observe that the confidence ratios are much more favourable, and the resolution of the 10,11,12 triple is relatively clear. Surprisingly, our earlier conclusion that the rook is the outlier of the three birds is contradicted and the peregrine falcon is (more reliably) predicted to be the outlier of the ostrich - rook- falcon triple.

The triple of bird species investigated here form a twig on larger phylogenetic trees developed by Brinkman et al. (2004). There, several different statistical methodologies and data sets are used and results summarised in their Figures 2, 3,and 4. None are consistent with Figure 4b above. In their Figure 2 the rook (11) is the outlier, as in our Figure 4a. In their Figures 3 and 4 the ostrich (10) is the outlier.

# 5 Tetrapods and reptiles

Stimulated by a discussion of Meyer and Zardoya (2003), this section concerns the evolutionary positions of reptiles and the tetrapods. First, consider three groups of tetrapods: placentals, marsupials, and monotremes, which are represented by species 1 (cat), 5 (opossum), and 6 (platypus), respectively. The website entry for this triple is:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| **1** | 5 | 6 | 1 | 0.9715 | 0.5193 | 0.9581 |

(cf. Fig. 5b). Given that nodes in close-proximity often result in poor confidence ratios, the prospect of ratios less than 0.52 is promising, although the proximity of the two nodes gives pause for thought. Given the value 0.52 for cr1, an error of four or five percent in the predicted relative time would not be surprising, and could produce a change in the predicted outlier.

Indeed, a different conclusion can be drawn if one adds the outlying reptile species (13) (the American alligator) to the triple:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 5 | 6 | **13** | 3 | 0.6418 | 0.3835 | 0.8769 |
| 1 | 5 | **13** | 3 | 0.6203 | 0.3862 | 0.9722 |
| 1 | 6 | **13** | 3 | 0.6832 | 0.4507 | 0.8720 |

If we assume that the first triple is correct (i.e., the monotremes and marsupials diverged from a common ancestor at rel. time = 0.64), the second and third triples give results which are in conflict with the first (1,5,6) findings. The second indicates that placentals diverged before the monotreme-marsupial split, whereas the third indicates that such a divergence took place afterwards (this interpretation appears as Fig.5a). Given that the confidence ratios are quite good for both possibilities, either of the two tree structures seems to be plausible.

Comparisons with the other placentals (using 2,3 instead of 1) yield similar results (in conflict with the (1,5,6) triple). For example:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 2 | 5 | **13** | 3 | 0.6294 | 0.3763 | 0.9070 |
| 2 | 6 | **13** | 3 | 0.6866 | 0.3722 | 0.9491 |
| 3 | 5 | **13** | 3 | 0.6199 | 0.3105 | 0.9403 |
| 3 | 6 | **13** | 3 | 0.6838 | 0.2721 | 0.9091 |

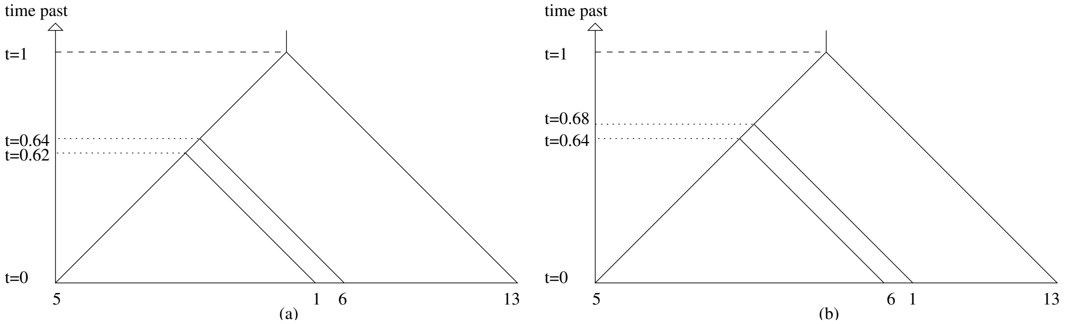

Figure 5: Mammals and a reptile: (a) A misleading tree. (b) Best interpretation.

Since the reptile is the common factor in these calculations, we conclude that there is something amiss with this species data (other reptiles should be investigated in a similar way).

However, the precedence of the three tetrapod subclasses can be investigated further by using other species as an outlier - the techique outlined in the preceding section. Since it is clear that the reptile species is an outlier for the three groups of tetrapods, we may scan the dataset of all possible triples in order to find additional outliers for the triple (1,5,6).

The following species produced reliable results (confidence ratios all less than 0.35) when added to the 1-5-6 tree: rook (11), peregrine falcon (12), Japanese sardine (34), rainbow trout (35), Atlantic salmon (36), Lake Chud whitefish (37), ayu fish (38), and Pacific porthole fish (39). (Numerical data appear in Appendix B.)

In all but one case, the placental is the outlier for the placental-marsupial-platypus triple, which supports the initial data for the (1,5,6) triple - and helps to justify our mistrust of the data for reptile species, 13. It is worth noting that, if we include *all* possible triples involving species 5 and 6, similar results are obtained (albeit with larger confidence ratios, but usually less than 0.5).

Our best interpretation of the results is summarized in Fig. 5b and supports the Marsupionta hypotheses of Figure 8B of Meyer and Zardoya (2003), namely, that "marsupials are closely related to monotremes, and both groups are equidistant to placentals".

As a further check on this conclusion, the DNA data for the four species 5,6,1 of Figure 5 together with 39 (the porthole fish) was re-aligned to get longer sequences of base-pairs. The results are:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| **1** | 5 | 6 | 1 | 0.9715 | 0.6758 | 0.9991 |
| 1 | 5 | **39** | 3 | 0.7016 | 0.3413 | 0.9997 |
| 1 | 6 | **39** | 3 | 0.7277 | 0.3229 | 0.9974 |
| 5 | 6 | **39** | 3 | 0.6759 | 0.2609 | 0.9839 |

This serves to confirm the main conclusion obtained from Figure 5b: the placental is the outlier of the placental, marsupial, montreme triple. It is also noteworthy

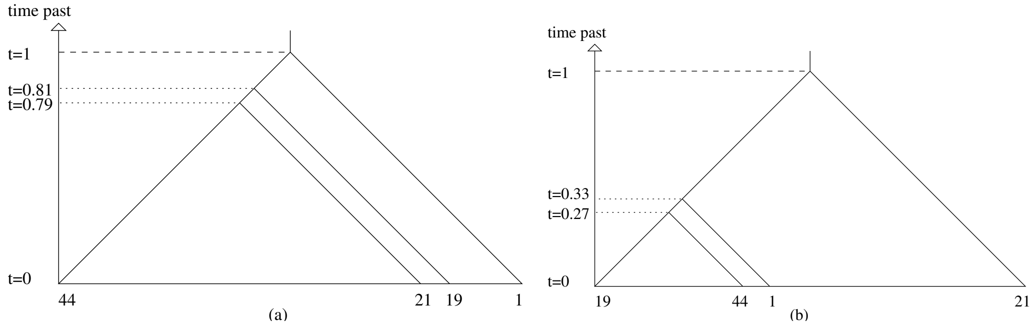

Figure 6: Tetrapods and fish: (a) A misleading tree. (b) Re-aligned data.

that, in spite of the increased sequence lengths, the confidence ratio for the 1,5,6 triple deteriorates, while the other three values of cr1 are improved.

# 6 Tetrapods and fish

Here we consider the ordinal relationships of the major fish groups vis-á-vis the tetrapods. This investigation was stimulated by work of Brinkman et al. (2004). The first choices of representatives for these groups are:

| No. | Species | Group |
|---|---|---|
| 1 | cat | tetrapods |
| 19 | African lungfish | lungfish |
| 21 | Sulawesi coelecanth | coelecanth |
| 44 | beardfish | ray-finned fish |

Our objectives here include comparisons between the original data set for 48 species (based on 2000 genome base pairs) and results obtained by applying our techniques to *realigned* genome sequences for these four species alone; admitting 6000 genome base pairs.

The following results are drawn from the website data base for 48 species:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 1 | 19 | **21** | 3 | 1.1091 | 0.7614 | 0.9987 |
| 1 | 19 | **44** | 3 | 1.1146 | 0.9588 | 0.8874 |
| **1** | 21 | 44 | 1 | 0.7934 | 0.6236 | 0.9199 |
| **19** | 21 | 44 | 1 | 0.8129 | 0.5019 | 0.9571 |

Although the distribution of outliers makes an admissible 4-tree, the high first confidence ratios and relative times larger than one suggest that the tree is not well-defined. A possible tree structure (in which we rely on the last two rows of the table above) is sketched in Figure 6a and it will be seen that, in fact, this structure is almost certainly misleading.

First we ask whether the structure sketched is maintained when we combine the triple 1, 19, 21 with another ray-finned fish, namely, the white sturgeon, 24, the butterfly fish, 27, or the common carp, 31. We do not go into detail but simply report that these three new quadruples support the sketch of Figure 6a, but even with improved confidence ratios. The value cr1=0.96 of the table above is even improved to 0.76, 0.79, or 0.80, respectively. However, in each case two relative divergence times greater than one are predicted.

In this case, when the DNA sequences for the original quadruple 1,19,21,44 are re-aligned (to obtain approximately 4000 base pairs), divergence matrices are re-computed, and the "matlab" program of Bohl and Lancaster, 2005, re-applied, the following revised table is produced:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 1 | 19 | **21** | 3 | 0.3515 | 0.1852 | 0.9890 |
| **1** | 19 | 44 | 1 | 0.7996 | 0.5798 | 0.7344 |
| 1 | **21** | 44 | 2 | 0.3214 | 0.3983 | 0.9802 |
| 19 | **21** | 44 | 2 | 0.2731 | 0.3716 | 0.9803 |

These results determine a unique 4-tree (all ambiguities of the preceding data set are removed). Furthermore, completely different divergence times are produced - and with acceptable confidence ratios. This tree appears as Figure 6b and, with a confidence ratio of 0.58, must be seen as more reliable than Figure 6a. These conclusions were reinforced by performing re-alignment of the four species in question several times, and observing no significant change in the results displayed in Figure 6b. Figure 1 of the paper Brinkman et al (2004) displays three different 4-trees for similar species and they assert that the correct topology "remains to be unambiguously determined". In predicting the coelecanth as the outlier, Figure 6b differs from all three and may resolve this particular issue.

The lessons to be learned from these examples seem to be that predicted divergence times greater than one should be treated with great caution (but see also Section 8), and that broadly based data sets (with abbreviated common sequence length) are not reliable when comparing closely related species.

As a final case-study we use the same four taxa to investigate the effects of the "gap cost penalty" (or GCP) of the Clustal software. It is found that the topology of Figure 6b is not sensitive to the GCP. However there is a significant change in the predicted relative divergence times. The effects on the confidence ratios and the divergence times are summarised in the next table. Here, the "default" GCP is that determined by the software. The authors' understanding of this example suggests that the lowest GCP used here is the most reliable, although this is not strongly reflected in the confidence ratios. The GCP is determined by two parameters ("gap extend" and "gap open") and the last two rows of the table determine the "low", "default", and "high" GCP's in terms of these parameters.

| GCP | low | default | high |
|---|---|---|---|
| 19,1,21 rel.dgce.time | 0.53 | 0.38 | 00.35 |
| cr1 | 0.58 | 0.61 | 0.62 |
| cr2 | 0.76 | 0.72 | 0.73 |
| gap extend | 3 | 5 | 8 |
| gap open | 8 | 10 | 15 |

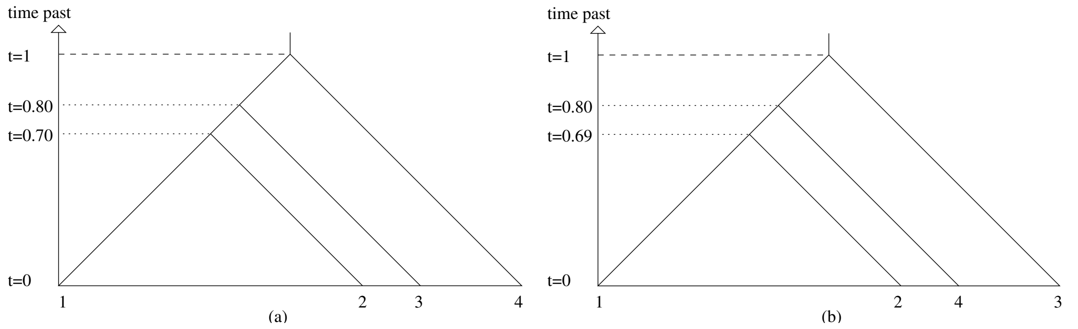

Figure 7: Mammals: (a) Data from the bank of 48 species. (b) Re-aligned data for 4 species.

# 7 Mammalian species and re-alignment

In this section we focus on some mammalian species and, first, investigate the effect of re-alignment of the four mammals from the 48-species set. Secondly, an independent study is made of the ape species: human, chimpanzee, and gorilla, together with the cat as an outlier.

The results for the species 1 (cat), 2(cow), 3(rabbit) and 4(wallaroo) from the 48-species set are:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 1 | 2 | **3** | 3 | 0.8840 | 0.4953 | 0.8467 |
| 1 | 2 | **4** | 3 | 0.6951 | 0.1862 | 0.9627 |
| 1 | 3 | **4** | 3 | 0.7728 | 0.5254 | 0.9540 |
| 2 | 3 | **4** | 3 | 0.8373 | 0.5133 | 0.9083 |

After re-aligning the four species (to obtain about 16,000 base pairs) we obtain:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 1 | 2 | **3** | 3 | 0.6850 | 0.1847 | 0.9515 |
| 1 | 2 | **4** | 3 | 0.8529 | 0.2499 | 0.9625 |
| 1 | **3** | 4 | 2 | 0.7968 | 0.4742 | 0.9196 |
| 2 | **3** | 4 | 2 | 0.8170 | 0.4849 | 0.9909 |

The two predicted 4-trees are compared in Figure 7. The tree for the re-aligned species has marginally improved confidence ratios but, surprisingly, there is an unexpected change in the predicted outlier (from wallaroo to rabbit). Interestingly, although the Markov model suggests that greater faith be placed in Figure 4(b), the ordinal relationships of Figure 4(a) are consistent with Figure 1 of Arnason et al. (2002).

As a second test of the Markov method for vertebrate species, mtDNA data were aligned for human (a), chimpanzee (b), gorilla (c) and, as an interesting but not too remote outlier, the cat (1). This admitted about 16,500 base pairs in the alignment.

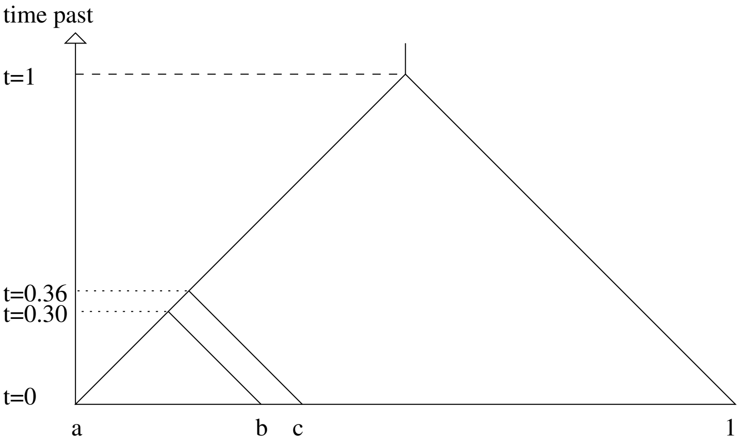

Figure 8: Ordering of the ape species.

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| a | b | **c** | 3 | 0.8308 | 0.4844 | 0.7925 |
| a | b | **1** | 3 | 0.2988 | 0.0477 | 0.9982 |
| a | c | **1** | 3 | 0.3535 | 0.0227 | 0.9860 |
| b | c | **1** | 3 | 0.3614 | 0.0510 | 0.9869 |

The favourable confidence ratios are remarkable, and the resulting tree of Figure 8 matches the generally accepted phylogeny for this quadruple. See Figure 1 of Arnason et al. (2002), for example. Here, the relative divergence time for the human, chimp, gorilla triple is predicted to be 0.30/0.36 from the graph, and is confirmed by the entry 0.83 of the first row in the table above. This is at the upper end of the range 0.62-0.86 suggested by Gagneux and Varki (2001). This example improves our confidence in the policy of using maximal sequence lengths when ordering closely related sets of species.

Given the wealth of paleontological data for hominids, a plausible calibration of the time scale is possible in this case.

# 8 Polytomy

The possibility arises that the separation of three or more taxa from the same branch of a tree is so close in time that the Markov model and the data are not able to resolve the separate binary divergences. This would be the case if both confidence ratios for the same triple are close to one. In such a case it may be helpful to admit a ternary (or higher) node on the tree. In effect this is just an acknowledgement that the Markov model together with the data do not have the power of resolution to form more refined multiple binary divergences.

Searching the results for 48 vertebrates at the website, several triples are found to be candidates for this line of thought. In particular, consider 15 (rubber eel), 16 (African clawed frog), and 45 (Atlantic cod). The tabulated entry is the first row of the next table, and we observe that the relative divergence time is predicted as 0.90

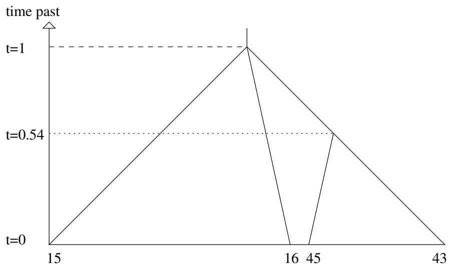

Figure 9: A ternary node

with confidence ratios 1.00 and 0.94; both close to one. This suggests that a ternary node might best represent these taxa in a tree. Several quadruples of taxa were then examined always including this triple with a view to finding plausible 4-trees including the triple node. Such a tree is obtained with the addition of species 43 (sandfish). Here are the results:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| **15** | 16 | 45 | 1 | 0.90 | 1.00 | 0.94 |
| 15 | 16 | **43** | 3 | 1.03 | 0.87 | 0.98 |
| **15** | 43 | 45 | 1 | 0.50 | 0.11 | 0.98 |
| **16** | 43 | 45 | 1 | 0.58 | 0.16 | 0.97 |

The ternary node is consistent with the results for **both** triples $(15, 16, 45)$ and $(15, 16, 43)$. Assuming this polytomy, the predicted relative divergence times for the triples $(15, 16, 45)$ and $(15, 16, 43)$ should agree. They are in fact 0.50 and 0.58 (and their average is indicated in Figure 9).

By adding the species 21 (coelecanth), this example can be developed further to admit a plausible *quaternary* node in a 5-tree. We do not tabulate the results here, but all of the six additional triples are consistent with the topology including a quaternary node for 15, 16, 21, and 43/45. The most significant discrepancy is another forecast for the binary node of Figure 9. The triple $(21, 43, 45)$ predicts a relative divergence time of 0.64 (in contrast to the two preceding predictions of 0.50 and 0.58).

In contrast to this, we check the divergence times for all the triples $(n, 43, 45)$ where $1 \leq n \leq 12$. In this way the divergence of 43 (sandfish) and 45 (cod) can be checked against all the mammals, marsupials, turtles, and birds of this data set. There is a remarkable consistency: these twelve predictions vary between 0.48 and 0.54, and the confidence ratios are uniformly good (as with the last two rows of the preceding table).

Another good example appears subsequently among the ray-finned fish. The 4-tree for the species 35,36,37,44 of Figure 16b contains a well-defined ternary node for the first three species.

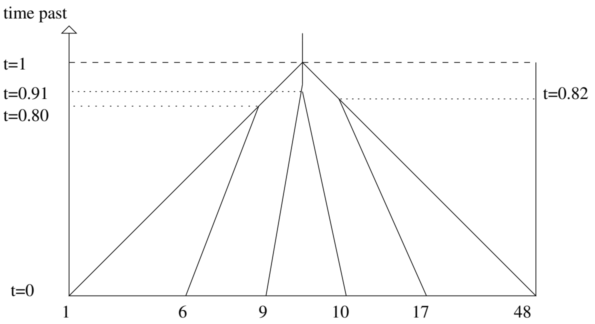

Figure 10: The skeleton tree: mammal, marsupial,turtle,bird,reptile,skate.

# 9 Trees for the vertebrates

Although the full data set for the 48 vertebrate species does not produce a unique well-defined tree, it is possible to obtain reasonably consistent trees for some large sub-groups. In these constructions the implementation of polytomy as described in the preceding section is essential. Close examination of the data reveals that several problems arise with three of the species groups of Appendix A, and these are the birds (10-12), the reptiles/amphibians (13-17), and the lungfish (18-20). Consequently, the first two are omitted from the detailed analysis (but do make an appearance in the initial skeleton tree of Figure 11). In view of the significance of the lungfish in the emergence of terrestrial species (see Brinkman et al., 2004) they are included in Figure 13. Examination of these species requires more refined data, possibly as conducted in Section 4.

We begin with what might be described as a "skeleton" tree (Figure 10) containing one species from each of the groups under consideration. The subsequent trees can be seen as providing more detail on the structure of "branches" of the tree. Recall that the original feature of these trees is the specification of explicit relative divergence times. Careful examination of the numerical results can reveal some inconsistencies in these larger trees, but the topologies and relative divergence times recorded are supported by the greater part of the relevant results. Note first of all that Figure 10 has a ternary node at the root. Indeed, it will be seen there seems to be a limit in time beyond which the model cannot resolve the speciation. This may be a result of using mtDNA. It seems that the root nodes of all of the following trees coincide - but we see this simply as the inability of this model and this data to resolve the tree more precisely. The selected species here are a mammal (1), marsupial (6), turtle (9), bird (ostrich) (10), reptile (salamander) (17), and fish (skate) (48) (see also Figure 3). This figure simply demonstrates that the speciation of the "groups", as we understand them, ocurred in the very distant past (as judged by our model). Our model is more informative in more recent times as speciation of the lineages develops.

Note that in all of the Figures 10-13, the root of the tree is at the same time so that listed relative divergence times are directly comparable. The time associated with the root of the tree of Figure 14 (and Figure 16b) has been adjusted to be consisitent with the other figures. Note also that the recorded times are, in

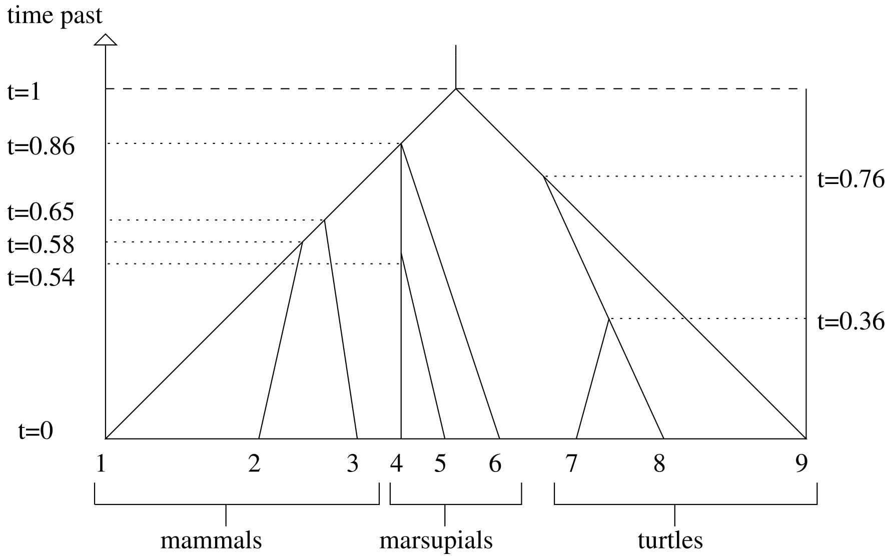

Figure 11: Mammals, marsupials and turtles

general, associated with more than one triple, and the time listed is an average of those recorded in the tabulated data. But these times are recorded only when there is plausible agreement between the several time measurements generated for a particular node. Figure 11 is the tree for the species 1-9, including the mammals, marsupials, and turtles. Notice that it was necessary to include only one multiple (ternary) node in this figure. Figure 5 can be seen as an attempt to resolve the ternary node used here. With the possible exception of this ternary node, the ordinal relationships for the species 1-6 are consistent with those predicted by Arnason et al. (2002). But the anomaly of the tree with re-aligned data (Figure 7) remains.

Figure 12 is a tentative tree structure including the three lungfish (species 18-20), the coelecanth (21, 22), and three other fish. Possibly the most sensitive part of this tree concerns the node marked $t = 0.90$. The numerical results scatter in such a way that this node could reasonably be labelled anything from 0.80 to 1.00 - which can significantly affect the origin of the lungfish and coelecanth viv-a-vis the other taxa. (Indeed, in Figure 13, this time is set at 1.00.) However, the early origin of the lungfish, coelecanth and skate (48) lineage is confirmed.

Figure 13 contains a tree for the species 21-36 including the coelecanth, freshwater fish (23-25), bonytongues (26,27), eels (28,29), carp (30-34) and samples of the ray-finned fish (35,36,45) (see Appendix A for our definitions of these terms). The tree is sketched with a quaternary node at the root. However, the branch supporting 23-29 **may** bifurcate from the root-45 branch at a more recent time, thus reducing the order of the root node (cf. Figure 2 of Arnason et al. (2004)).

Figure 14 is a relatively well-defined tree for the ray-finned fish (36-45), sharks (46, 47) and skate (48) (and contains a quaternary node). Careful comparison of these predictions with those of Arnason et al. (2002 and 2004) is interesting.

Comparison of these results with Figure 3 of Brinkmann et al. (2004) is also

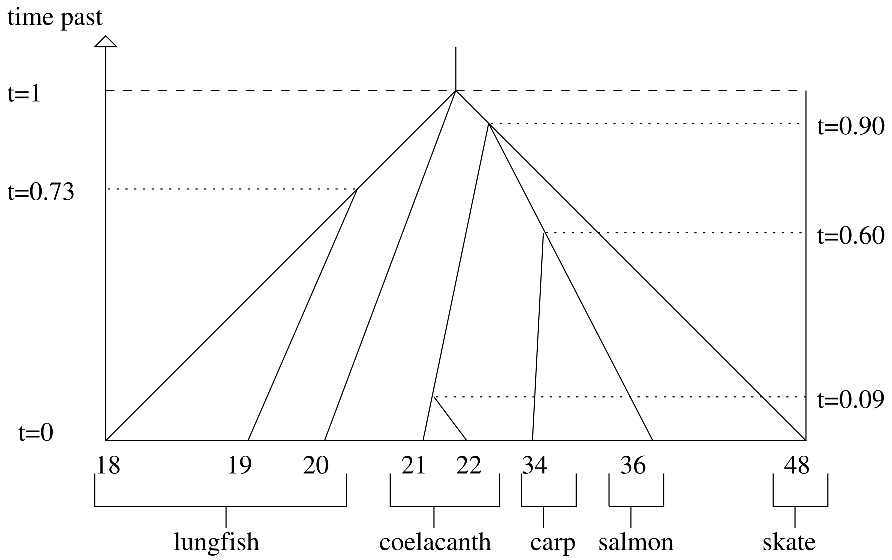

Figure 12: Lungfish, coelecanth and other fish.

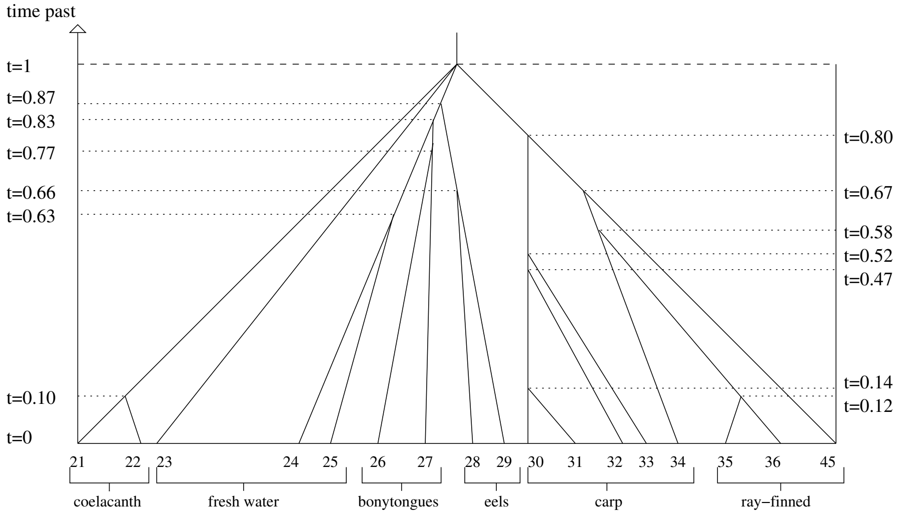

Figure 13: Fish species: first set.

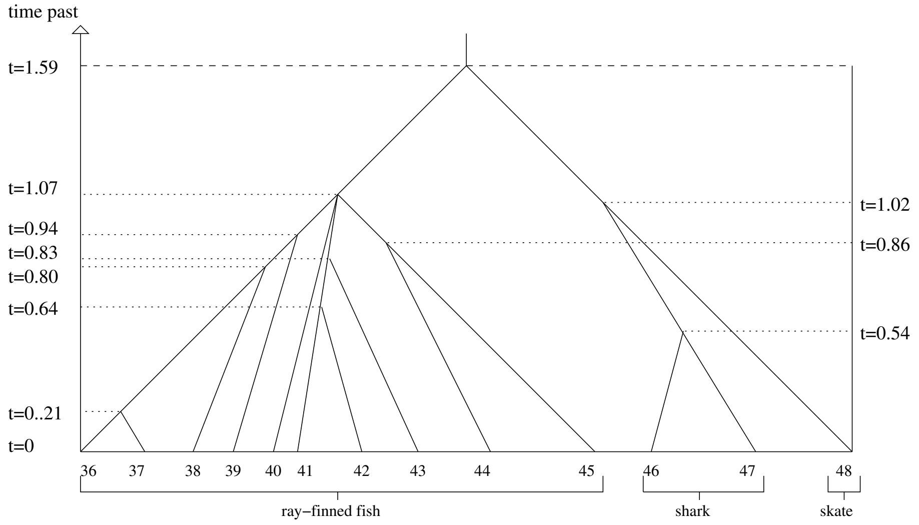

Figure 14: Fish species: second set.

interesting, and not entirely consistent[3]. Note also that the results obtained here with the irreversible Markovian model and those obtaimed with statistical techniques by the Brinkmann group are obtained with the same data. If nothing else, the comparison and the analysis above show the need for improved data.

Another interesting source of comparisons is a set of phylogenetic trees for "ancient fish" drawn up by Inoue et al. (2003) and described by Miya et al. (2003) as "decisively resolved". In that comprehensive study of numerous fish species, two trees are presented for the same set of species, one obtained using "maximum parsimony" (MP) methods (their Figure 2) and the other using "maximum likelihood" (ML) methods (their Figure 3). It is particularly interesting for this discussion because twelve of the species studied by the Inoue group are contained in our data set of 48 species.

Figure 15 provides the tree (with our conventions) deduced in that paper for the twelve species in question. The topologies are the same with both the MP and ML analyses. Relative divergence times are not available by these techniques, so the vertical scale can indicate only the predicted ordering of nodes. Comparable trees obtained by the present Markovian analysis appear in Figures 16a and 16b - presented in two parts for the reader's convenience. We note also that, in Figure 16a, the time $t = 0.91$ is the average of 0.7451, 0.9900, and 0.9970, and cannot be regarded as reliable.

The most notable feature of this comparison is the general agreement on topologies predicted by the three techniques. There is one notable exception, however, in the location of species 35 (rainbow trout) whose predicted ancestry differs substantially in Figures 15 and 16b.

[3] The reader is reminded that definitive lineagies are not expected here; the importance of this work lies in the introduction of a new methodolgy which provides the additional feature of estimated relative divergence times

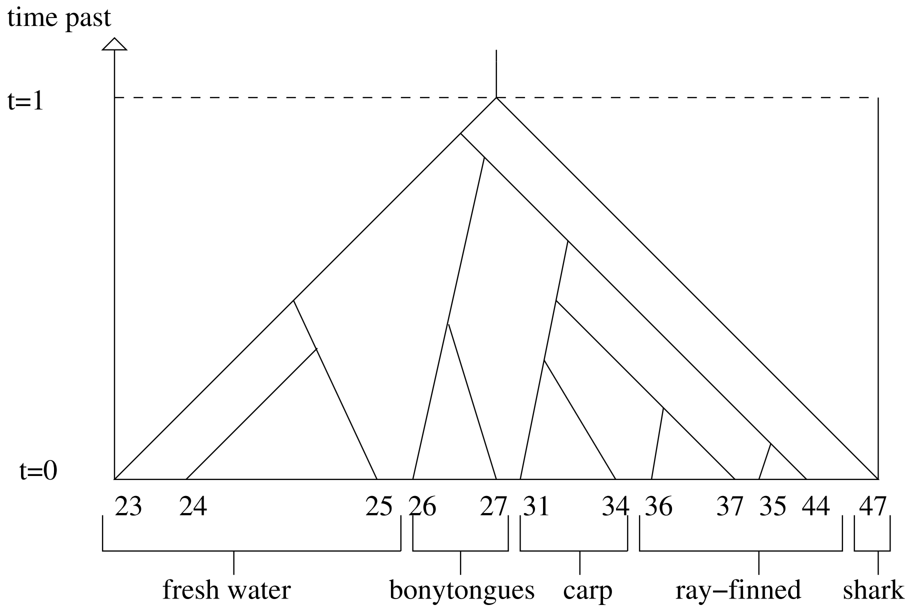

Figure 15: Topological structure of Inoue et al..

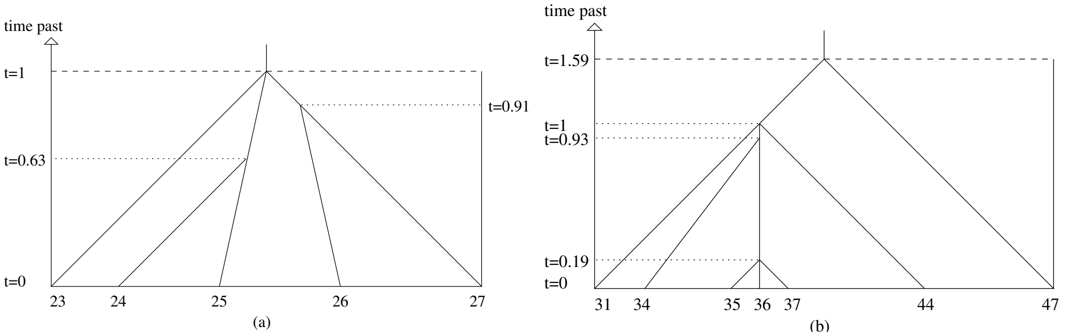

Figure 16: Species studied by Inoue et al. (cf. Fig. 15).

# 10 Conclusions

An irreversible Markov process (used to model the phenomenon of drift of nucleotides in the evolution of species) has been developed and implemented elsewhere by two of the present authors. Here, the usefulness of this model has been illustrated in a number of ways - and avoiding technicalities as far as possible. The implications of the existing theory have been illustrated in the context of some topical problems of phylogeny, and useful insights are obtained. These include preliminary studies of the ordinal relationships of birds (Section 4), three groups of tetrapods (Section 5), three groups of fish (Section 6), groups of mammals (Section 7), and the feasibility of including polytomy in the predicted trees (Section 8). Section 9 contains a comprehensive application of the techniques developed here to the mtDNA data set developed by Brinkman et al (2004) - notwithstanding some apparent weaknesses in that data discussed in Sections 4-6. Some direct comparisons are also made with tree structures proposed recently by Inoue et al. (2003).

In all cases, a remarkable feature is the ability of the model to predict internally consistent relative divergence times.

The results suggest that there are limitations due to the use of *mitochondrial* DNA data. Also, the effects of subjective variations in the alignment processes require further investigation. In particular, there is potential for improved results and further analysis in the following directions:

- Careful study of the effects of sequence length.
- Performing re-alignment for every triple of species. (This would require extension of the existing software.)
- Applying the present technique to protein rather than nucleotide data (also computationally intensive, but requiring no new ideas).
- A mathematical analysis of the sensitivity of predicted divergence times to the data.
- Abandoning homogeneous time dependence in the Markov model and developing more sophisticated time dependence with the objective of resolving high order polytomy phenomena of the distant past - as suggested by the model studied in this paper.

# 11 Acknowledgements

The authors are grateful to Dawn Simon and Derek Sykes for comments and good advice. The work of Ryan Hamilton and Peter Lancaster was partially supported through a grant from the Natural Sciences and Engineering Research Council of Canada.

# Appendix A. 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 |

# 12 Appendix B

DATA FOR FIGURE 2a:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 3 | 0.6951 | 0.1862 | 0.9627 |
| 1 | 2 | 36 | 3 | 0.4696 | 0.1413 | 0.9812 |
| 1 | 2 | 37 | 3 | 0.4695 | 0.1389 | 0.9794 |
| 1 | 2 | 38 | 3 | 0.4613 | 0.1741 | 0.9210 |
| 1 | 4 | 36 | 3 | 0.6863 | 0.2926 | 0.9845 |
| 1 | 4 | 37 | 3 | 0.6828 | 0.2470 | 0.9869 |
| 1 | 4 | 38 | 3 | 0.6663 | 0.2036 | 0.9589 |
| 1 | 36 | 37 | 1 | 0.0988 | 0.0972 | 0.9988 |
| 1 | 36 | 38 | 1 | 0.4150 | 0.1676 | 0.9824 |
| 1 | 37 | 38 | 1 | 0.4056 | 0.1337 | 0.9813 |
| 2 | 4 | 36 | 3 | 0.6670 | 0.2307 | 0.9387 |
| 2 | 4 | 37 | 3 | 0.6618 | 0.1920 | 0.9276 |
| 2 | 4 | 38 | 3 | 0.6416 | 0.1639 | 0.9393 |
| 2 | 36 | 37 | 1 | 0.0943 | 0.0878 | 0.9783 |
| 2 | 36 | 38 | 1 | 0.4009 | 0.1689 | 0.9267 |
| 2 | 37 | 38 | 1 | 0.3908 | 0.0700 | 0.9548 |
| 4 | 36 | 37 | 1 | 0.0969 | 0.0922 | 0.9915 |
| 4 | 36 | 38 | 1 | 0.4050 | 0.2328 | 0.9389 |
| 4 | 37 | 38 | 1 | 0.3934 | 0.1099 | 0.9749 |
| 36 | 37 | 38 | 3 | 0.2514 | 0.2430 | 0.9622 |

DATA FOR THE TETRAPOD/REPTILE CASE:

| | | | outlier | time ratio | cr1 | cr2 |
|---|---|---|---|---|---|---|
| 5 | 6 | 11 | 3 | 0.6957 | 0.2145 | 0.9591 |
| 1 | 5 | 11 | 3 | 0.6999 | 0.3819 | 0.9636 |
| 1 | 6 | 11 | 3 | 0.7537 | 0.3502 | 0.9218 |
| 5 | 6 | 12 | 3 | 0.7126 | 0.2298 | 0.9557 |
| 1 | 5 | 12 | 3 | 0.7059 | 0.3719 | 0.9682 |
| 1 | 6 | 12 | 3 | 0.7530 | 0.3725 | 0.9914 |
| 5 | 6 | 34 | 3 | 0.6534 | 0.2205 | 0.9660 |
| 1 | 5 | 34 | 3 | 0.6793 | 0.3614 | 0.9513 |
| 1 | 6 | 34 | 3 | 0.7087 | 0.3333 | 0.9314 |
| 5 | 6 | 35 | 3 | 0.6649 | 0.2497 | 0.9421 |
| 1 | 5 | 35 | 3 | 0.6811 | 0.3072 | 0.9805 |
| 1 | 6 | 35 | 3 | 0.7116 | 0.2761 | 0.9266 |
| 5 | 6 | 36 | 3 | 0.6728 | 0.2550 | 0.9331 |
| 1 | 5 | 36 | 3 | 0.6946 | 0.3366 | 0.9729 |
| 1 | 6 | 36 | 3 | 0.7234 | 0.3232 | 0.9179 |
| 5 | 6 | 37 | 3 | 0.6809 | 0.2211 | 0.9245 |
| 1 | 5 | 37 | 3 | 0.7003 | 0.3059 | 0.9875 |
| 1 | 6 | 37 | 3 | 0.7270 | 0.2771 | 0.9254 |
| 5 | 6 | 38 | 3 | 0.6586 | 0.2132 | 0.9493 |
| 1 | 5 | 38 | 3 | 0.6759 | 0.3098 | 0.9603 |
| 1 | 6 | 38 | 3 | 0.7200 | 0.2854 | 0.9098 |
| 5 | 6 | 39 | 3 | 0.6460 | 0.1657 | 0.9536 |
| 1 | 6 | 39 | 3 | 0.6955 | 0.2361 | 0.9406 |
| 1 | 5 | 39 | 3 | 0.6689 | 0.2666 | 0.9774 |

# References

Arnason U, Adegoke J A, Bodin K, Born E W, Esa Y B, Gullberg A, Nilsson M, Short R V, Xu X and Janke A (2002) Mammalian mitogenomic relationships and the root of the eutherian tree, Proc. Nat. Acad. Science USA, **99**, no.12, 8151-8156.

Arnason U, Gullberg A, Janke A, Joss J and Elmeroi C (2004) Mitogenomic analyses of deep gnathostome divergence: a fish is a fish, Gene, **333**, 61-70.

Bohl E and Lancaster P (2003) Irreversible Markov processes for phylogenetic models, Numerical Linear Algebra with Applications, **10** : 577-593.

Bohl E and Lancaster P (2006) Implementation of a Markov model for phylogenetic trees, Journal of Theoretical Biology, **239** no. 3: 324-333.

Brinkmann H, Derk A, Zitzler J, Joss J A, and Meyer A (2004) Complete mitochondrial genome sequences of the Australian lungfish, J. Mol. Evol., **59** : 834-848.

Durbin R, Eddy S, Krogh A, and Mitchison G (1998) Biological Sequence Analysis, Cambridge.

Gagneux P, and Varki A (2001) Genetic differences between humans and great apes, Mol. Phylogenet. Evol., **18**, No.1 : 2-13.

Grant T, Faivovich J, and Pol D (2003), The perils of 'point-and-click' systems, Cladistics, **19**, 276-285.

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.

Huelsenbeck J P, Bollback J P, and Levine A M, (2002), Inferring the root of a phylogenetic tree, Syst. Biol. **51**, 32-43.

Inoue JG, Miya M, Tsukamoto K, and Hishida M (2003), Basla actinopterygian relationships: a mitogenetic perspective on the phylogeny of the "ancient fish", Mol. Phylogenet. Evol **26**, 110-120.

Jukes T H, and Cantor C R (1969) Evolution of protein molecules, In Munro H. N. (Ed.), Mammalian Protein Metabolism, **3** : 21-132, Academic press, New York.

Li W-H (1997) Molecular Evolution, Sinauer, Sunderland Mass..

Meyer A, and Zardoya R (2003) Recent advances in the (molecular) phylogeny of vertebrates, Annu. Rev. Ecol. Evol. Syst., **34** : 311-338.

Miya M, Takeshima H, Endo H, Ishiguro N B, Inoue J G, Mukai T, Satoh T P, Yanaguchi M, Kawaguchi A, Mabuchi K, Shirai S M, Nishida M (2003), Mol. Phylogenet. Evol., **26** : 121-138.

Ruvolo M (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets, Mol. Biol. Evol. **14**, No.3 :248-265.

Yang Z (1994) Estimating the pattern of nucleotide substitution, J. Mol. Evol., **39** : 105-111.