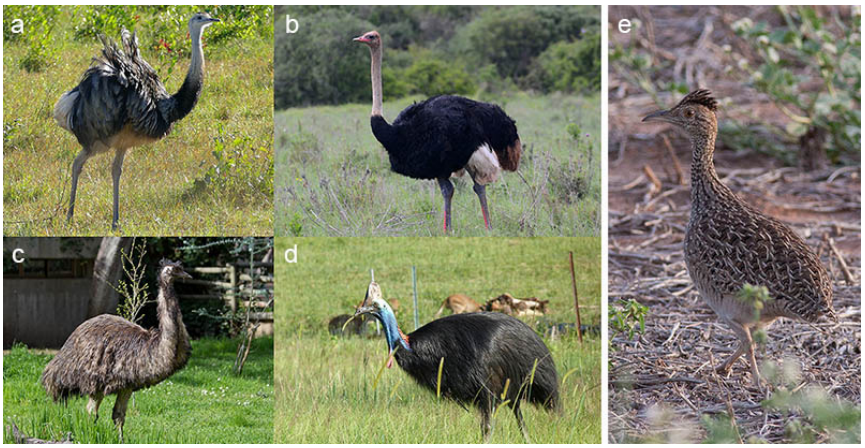
Practical 1.1: The Evolution of Ratite Birds

Sequence alignment and phylogenetic analysis using *MEGA*

Modern birds (class Aves) are a diverse group of vertebrates that comprise about 10,000 living species. Birds are divided into two subclasses: **Neognathae** ('new jaw'), which comprises >99% of all extant avian species, and **Palaeognathae** ('old jaw'), which includes the tinamous and ratites. The split between neognaths and palaeognaths occurred about 100 million years ago, in the middle of the Cretaceous period. This fundamental division in the bird phylogeny is supported by a variety of morphological and genetic data, including analyses of whole-genome sequences.

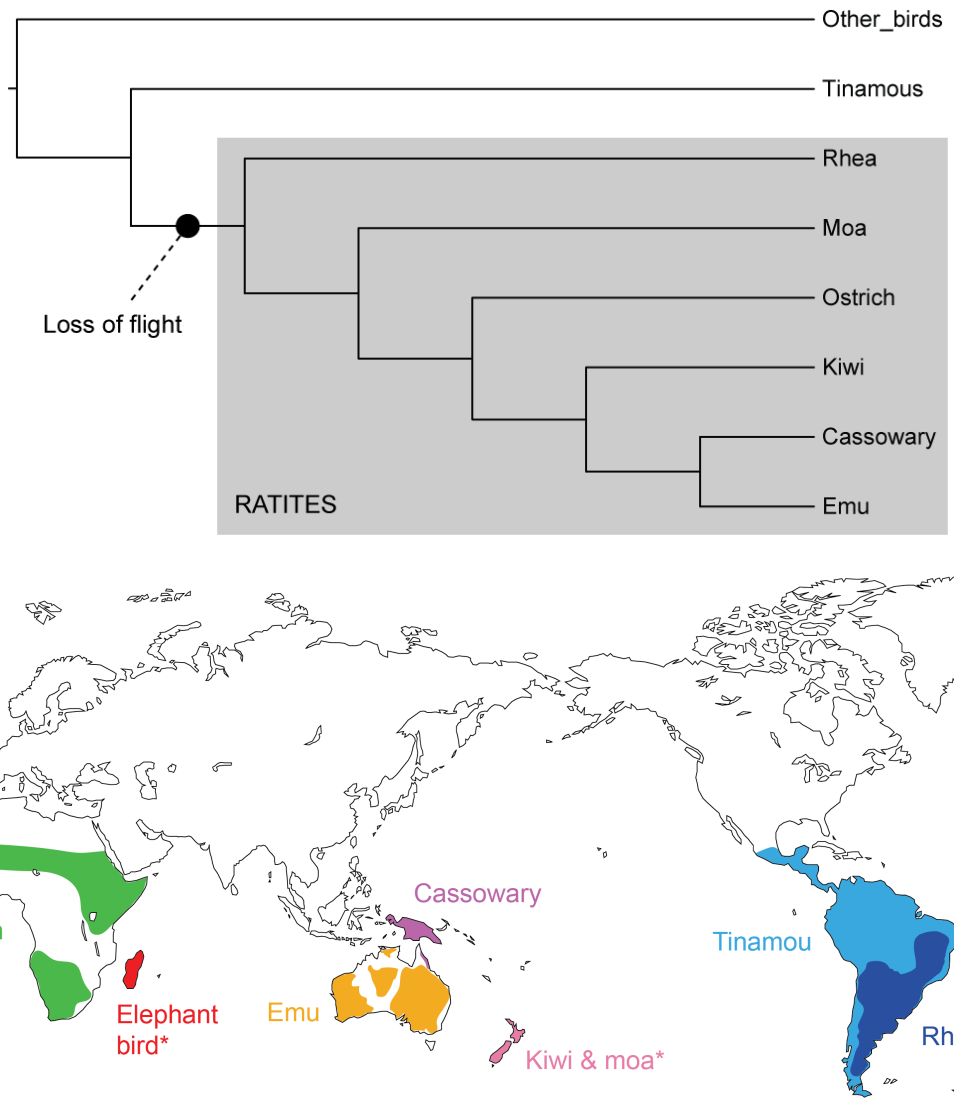
The ratites, all of which have lost the ability to fly, include the ostrich (Africa), rheas (South America), kiwi (New Zealand), emu (Australia), and cassowaries (Australia and New Guinea), as well as the recently extinct moa (New Zealand) and elephant birds (Madagascar). The ratites are the most familiar palaeognaths and are generally large herbivores or omnivores. The palaeognaths also include the South American tinamous, of which there are 47 species in 9 genera. Tinamous are ground-foraging birds that can fly but are not strong flyers.



Representatives of palaeognath birds (clockwise): (a) greater rhea, (b) common ostrich, (b) southern cassowary, (d) emu, and (e) brushland tinamou. Photographs by Bernard Dupont, Kore, Kadellar, Donald Hobern, and Allan Drewitt.

A close relationship between tinamous and ratites is generally accepted, but there has been some uncertainty over their relationships. The flightless ratites have typically been regarded as a monophyletic group, with tinamous being their closest relatives. Accordingly, the most parsimonious interpretation of their evolutionary history is that the ability to fly was lost in the ancestral ratite lineage after it diverged from the lineage leading to tinamous.

Ratites have a distinctive geographic distribution, being found on land masses that were once part of the supercontinent **Gondwana**. They are regarded as quintessential Gondwanan taxa, along with southern beeches (*Nothofagus*) and some groups of freshwater fishes. Researchers have suggested that the diversification of ratites was driven by the break-up of Gondwana. Early molecular studies claimed good agreement with this biogeographic model of **vicariance**, finding a deep divergence between the lineages leading to the ostrich (Africa) and to all other ratites, and with rheas (South America) being a sister taxon to the kiwi (New Zealand) and emu and cassowary (Australia–New Guinea). An opposing view would be that the ratites diversified through **geographic dispersal**.



As seen in the molecular phylogenetic tree above, Cooper *et al*. (2001) estimated that the rhea is the sister taxon to all other ratites. These findings stand in contrast with the vicariance hypothesis for ratite evolution, which suggests that the ostrich should be the sister lineage to all other ratites because Africa was the first to break off from the rest of Gondwana. In addition, the two New Zealand taxa (kiwi and moa) are not closely related to each other, suggesting that their ancestors must have colonised New Zealand independently rather than simply diverging when the continents drifted apart.

A further source of contention has arisen, this time concerning the **relationship of the tinamous** to the ratites. In the past, it was presumed that the flighted tinamous were the sister group to the flightless ratites. In recent years, however, detailed analyses of large amounts of DNA sequence data have produced some surprising results with regard to palaeognath relationships.

In this practical exercise, you will use a phylogenetic approach to investigate these key issues in palaeognath evolution. You will begin by aligning a set of DNA sequences. The data set will then be analysed using two different phylogenetic methods implemented in the free software *MEGA*.

Section A: Sequence alignment

Before you begin, ensure that your computer has a recent version of *MEGA* (version 7, X, or 11) and that you have the DNA data file, **ratites.fasta**. This data file contains sequences of ~2000 nucleotides from 12 birds: Southern Cassowary, Emu, Little Spotted Kiwi, Southern Brown Kiwi, Crested Moa, Eastern Moa, Common Ostrich, Greater Rhea, Andean Tinamou, Grey Tinamou, Highland Tinamou, and Chicken.

In this exercise you will use the free phylogenetic software *MEGA*. This software is able to implement a wide range of sequence analyses, including phylogenetic analysis using neighbour joining and maximum likelihood.

- In *MEGA*, open the alignment file **ratites.fasta** and select "Align". Have a look at the sequences and notice that they generally vary in length. The chicken sequence is the longest (2000 nt).

Q. *Before we conduct any phylogenetic analyses, we need to align these sequences. What is the purpose of sequence alignment?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *What are 'indels'?*

. . . . . .

. . . . . .

- We'll get the computer to align the sequences for us. In the **Alignment** menu, choose "Align by Muscle" and select all of the sequences. Muscle is one of the two alignment algorithms in *MEGA*. The other is ClustalW. These two are probably the most widely used alignment algorithms. We will use Muscle for this practical.

  Change the "Gap Open" penalty to 0. This removes the penalty for inserting gaps in the alignment, meaning that Muscle will be more willing to insert gaps to make the sequences better aligned to each other. Click "OK".

Q. *How long is the sequence alignment?*

. . . . . .

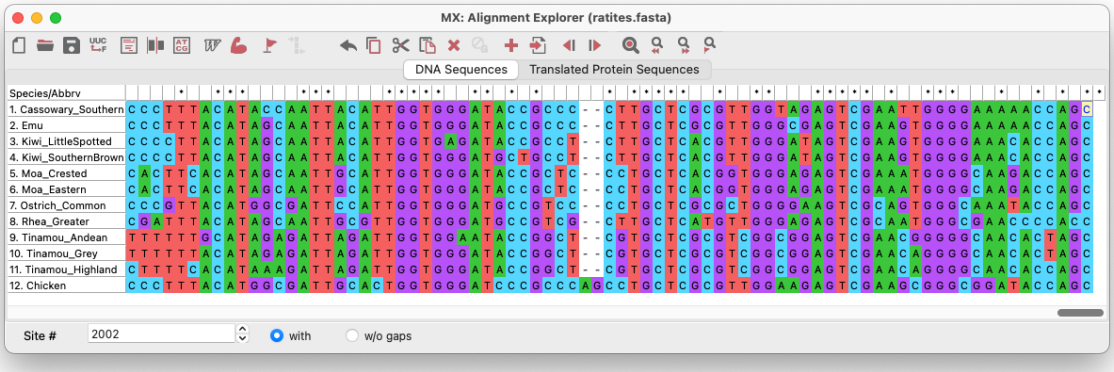
- Now run Muscle again, but with a “Gap Open” penalty of -400.

Q. *How long is the new sequence alignment?*

. . . . . .

It seems that reducing the gap-opening penalty to 0 caused Muscle to insert too many gaps into the data set. The second alignment, produced using a gap-opening penalty of -400, looks more reasonable and we will use this for our analyses.

In this practical we will simply accept the results of the automated alignment, but standard practice is to inspect the alignment to see if the automated method has done a good job. In some cases, visual inspection can reveal sections of the alignment that can be improved. Nowadays, many sequence data sets are far too large for visual inspection to be feasible. For example, phylogenomic data sets often involve hundreds or even thousands of gene alignments. As a consequence, there is an increasing reliance on automatic methods for sequence alignment.



Q. *If one of your sequences had accidentally been shifted to the right by 1 nucleotide (so that it was misaligned by 1 nucleotide compared with the remaining sequences), what would be the consequences for phylogenetic analysis?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Now that we have aligned the DNA sequences, we are ready to analyse the data set.

Section B: Phylogenetic analysis

This section consists of three parts:

- Model selection
- Phylogenetic analysis using neighbour joining
- Phylogenetic analysis using maximum likelihood

Model selection

Before conducting any phylogenetic analyses, we need to identify the best-fitting model of nucleotide substitution for the data set. A key purpose of substitution models is to account for multiple substitutions. Perhaps the most widely used substitution model is the General Time Reversible (GTR) model, which allows different rates for different substitution types. For example, it allows A↔G substitutions to occur at a different rate from C↔G substitutions. The model also allows the four nucleotides to have unequal frequencies.

Q. *Which phylogenetic methods can use an explicit model of nucleotide substitution?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

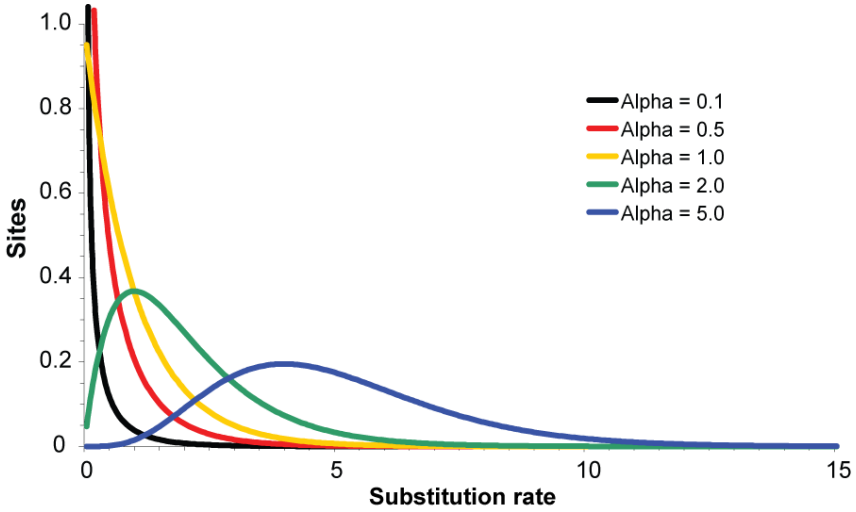
Q. *What is the simplest model of nucleotide substitution?*

. . . . . .

. . . . . .

We can also allow the rate of evolution to vary among the sites in the alignment. By doing this, we are assuming that some sites are under greater selective constrained, whereas others can change more freely without affecting the organism's fitness.

Rate variation among sites is usually modelled using a gamma distribution, which can take a variety of shapes. The shape of the distribution is determined by a single parameter, alpha. When alpha is small (<1), many sites evolve slowly but a small number of sites evolve very quickly. When alpha is large (>1), most sites evolve at about the same rate. For computational reasons, we use discrete rather than continuous gamma distributions. Usually 4 to 6 rate categories are used for the discrete gamma distribution. Increasing the number of rate categories will tend to make the analysis run more slowly, but with little additional benefit.



Conveniently, we can compare different models and select the best-fitting model in *MEGA*.

- In the **Data** menu of the Alignment Explorer window, select "Phylogenetic Analysis". Select "No" when the program asks whether you have "Protein-coding nucleotide sequence data". We are working with noncoding DNA, which does not lead to any protein products.
- Go back to the main *MEGA* window. From the **Models** menu, select "Find Best DNA/Protein Models (ML)" and use the currently active data. This will bring up a box that contains a range of options. Accept the default settings and click on "OK".

MEGA is now computing the likelihood (probability of the data given the model) for 24 different substitution models. Have a look at the results of the analysis. The first column shows a list of the models. The second column shows how many parameters each model has. The third and fourth columns show the scores for two model-selection criteria, the Bayesian information criterion (BIC) and the corrected Akaike information criterion (AICc). For both of these criteria, lower scores indicate better-fitting models. For further details, have a look at the information below the table.

Q. *How are the BIC and AICc calculated?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *What is the best-fitting model(s) according to the BIC and AICc?*

. . . . . .

In the table, the number of free parameters in each model includes the branch lengths, because these need to be estimated in the analysis. For the purpose of model selection, *MEGA* has estimated the phylogenetic tree using a quick neighbour-joining analysis.

Q. *How many branches are in the tree? (hint: an unrooted tree of n tips has 2n-3 branches)*

. . . . . .

In the most parameter-rich model (GTR+G+I), there are 10 free parameters (not including the branch lengths). The simplest model (Jukes-Cantor or JC model) has 0 free parameters.

Q. *What assumptions are made in the Jukes-Cantor model?*

. . . . . .

. . . . . .

Q. *What are the 10 free parameters in the GTR+I+G model?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *For the GTR+G+I model, what is the estimate of the shape parameter of the gamma distribution for this data set? What does this suggest about the degree of rate variation among sites?*

. . . . . .

. . . . . .

. . . . . .

Q. *What is the estimate of the shape parameter of the gamma distribution for this data set when using the GTR+G model? Does this differ from the estimate when using the GTR+G+I model?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

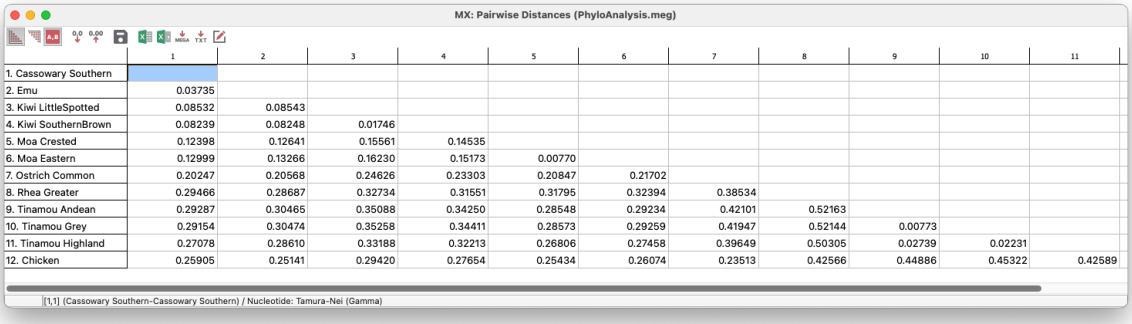
Phylogenetic analysis using neighbour joining

Now we can estimate some phylogenetic trees. We'll do this using two different methods: a distance-based method and maximum likelihood.

Distance-based methods (such as neighbour joining) infer the phylogeny using a matrix of pairwise genetic distances. They are usually very quick because the method does not need to search through tree-space. Instead, distance-based methods employ an algorithm to reconstruct the tree using information from the distance matrix. The most commonly used algorithm is neighbour joining, which is what we will use here.

- From the **Distance** menu, select "Compute Pairwise Distances" and use the currently active data. This will bring up a box that contains a range of options for the neighbour-joining analysis.
- Check that you have the following options selected:
  Model/Method: Tamura-Nei model
  Substitution to Include: d: Transitions + Transversions
  Rates among Sites: Gamma Distributed (G)
  Gamma Parameter: 0.58
  Gaps/Missing Data Treatment: Pairwise deletion

The GTR model is not available as an option here, so we are using the similar Tamura-Nei model instead. Note that we chose 'Gamma Distributed (G)' for 'Rates among sites'. Here we are allowing different nucleotide sites to have different evolutionary rates. We are assuming that these rates follow a gamma distribution. The 'gamma parameter' reflects the degree of rate heterogeneity among sites and was estimated at 0.58 in our model-selection analysis above. Have a look at the resulting matrix of pairwise distances.



	1	2	3	4	5	6	7	8	9	10	11
1. Cassowary Southern											
2. Emu	0.03735										
3. Kiwi LittleSpotted	0.08532	0.08543									
4. Kiwi SouthernBrown	0.08239	0.08248	0.01746								
5. Moa Crested	0.12398	0.12641	0.15561	0.14535							
6. Moa Eastern	0.12999	0.13266	0.16230	0.15173	0.00770						
7. Ostrich Common	0.20247	0.20568	0.24626	0.23303	0.20847	0.21702					
8. Rhea Greater	0.29466	0.28687	0.32734	0.31551	0.31795	0.32394	0.38534				
9. Tinamou Andean	0.29287	0.30465	0.35088	0.34250	0.28548	0.29234	0.42101	0.52163			
10. Tinamou Grey	0.29154	0.30474	0.35258	0.34411	0.28573	0.29259	0.41947	0.52144	0.00773		
11. Tinamou Highland	0.27078	0.28610	0.33188	0.32213	0.26806	0.27458	0.39649	0.50305	0.02739	0.02231	
12. Chicken	0.25905	0.25141	0.29420	0.27654	0.25434	0.26074	0.23513	0.42566	0.44886	0.45322	0.42589

Q. *What do these numbers represent, and in what units are they given?*

. . . . . .

. . . . . .

Q. *Which two taxa are separated by the smallest genetic distance?*

. . . . . .

- From the **Phylogeny** menu, select "Construct/Test Neighbour-Joining Tree" and use the currently active data. This will bring up a box that contains a range of options for the neighbour-joining analysis.
- Check that you have the following options selected:
  Test of Phylogeny: Bootstrap method
  No. of Bootstrap Replications: 100
  Model/Method: Tamura-Nei model
  Substitution to Include: d: Transitions + Transversions
  Rates among Sites: Gamma Distributed (G)
  Gamma Parameter: 0.58
- Click on "OK" to start the neighbour-joining analysis.
- The estimate of the phylogeny will appear in a new window. Check that the tree is rooted between the Chicken and the other taxa. If this is not the case, select the branch leading to Chicken and select "Root the tree on the selected branch" (this is the little tree icon with the red triangle on its left).

Q. *There is a scale bar shown below the tree. What does this measure?*

. . . . . .

. . . . . .

Q. *What is the purpose of including an outgroup taxon (chicken) in the analysis?*

. . . . . .

. . . . . .

. . . . . .

Q. *Is there strong support for the relationships in the tree?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Phylogenetic analysis using maximum likelihood

Now we will try analysing the data set using maximum likelihood. This is a statistical method that was first applied to phylogenetic analysis in the 1970s and formalised in the early 1980s.

In maximum-likelihood analysis, we aim for maximum-likelihood estimates of the parameters and search for the maximum-likelihood tree. Like distance-based methods, maximum likelihood uses an explicit model of nucleotide substitution.

- Go back to the main *MEGA* window. From the **Phylogeny** menu, select "Construct/Test Maximum Likelihood Tree" and use the currently active data. This will bring up a box that contains a range of options for the maximum-likelihood analysis.
- Check that you have the following options selected:
  Test of Phylogeny: Bootstrap method
  No. of Bootstrap Replications: 100
  Substitutions Type: Nucleotide
  Model/Method: General Time Reversible model
  Rates among Sites: Gamma Distributed (G)
  No of Discrete Gamma Categories: 4
  Gaps/Missing Data Treatment: Use all sites
- Click on "OK" to start the maximum-likelihood phylogenetic analysis. See how long it takes to calculate bootstrap support from 100 replicates.

Q. *Did the analysis take much longer than the neighbour-joining analysis? Why do you think that this is the case?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

- The estimate of the phylogeny will appear in a new window. Check that the tree is rooted between the Chicken and the other taxa. If this is not the case, select the branch leading to Chicken and select "Place Root on Branch".

Q. *Are the relationships similar to those estimated using neighbour-joining?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

MEGA has a range of interesting options for visualising the phylogenetic tree. We can now explore some of these with our maximum-likelihood tree.

Try changing the tree format by selecting the first of the tree icons (the sixth button from the left). Select "Circle" to display the tree in circular format. As you can see, this type of tree is more difficult to interpret, because the branching structure of the tree is less clear. Go back to the "Traditional" "Rectangular" tree format.

Now click on the next icon to the right, which is "Toggle scaling of the tree". If you click on this button repeatedly, the displayed tree alternates between a phylogram and a cladogram. The cladogram only shows the relationships among the species; the branch lengths do not reflect any particular biological quantity.

One last thing to examine now is the rooting of the tree. For both the neighbour-joining and maximum-likelihood trees, we placed the root of the tree on the branch leading to the Chicken. This taxon was included in the data set as an outgroup. If we tried to root the tree using a different method, such as midpoint rooting, the results might have been quite different.

- Go to the **View** menu at the top of your screen and select "Root on Midpoint".

Q. *How has this changed the structure of the tree? What is the sister lineage to the Chicken?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *In this example, it seems that midpoint rooting is highly misleading. Why do think that this is the case?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *Of the two trees that you estimated (using neighbour joining and maximum likelihood), which do you think is more reliable, and why?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *Are the relationships consistent with the hypothesis of vicariance (i.e., does the pattern of relationships match the sequence of divergences that we would expect from the break-up of Gondwana)?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *What does the placement of the tinamou suggest about the loss of flight in ratites?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .