HAGGERTY, S.E. 1994. Superkimberlites: A geodynamic diamond window to the Earth's core. *Earth and Planetary Science Letters*, **122**, 57–69.

KINGSLEY, C. 1890. *Scientific Lectures and Essays*. Macmillan and Co: London and New York.

KNELL, S.J. and LEWIS, C.L.E. 2001. Celebrating the age of the Earth. *In:* Lewis, C.L.E. & Knell, S.J. (eds) *The Age of the Earth: from 4004 BC to AD 2002*. Geological Society, London, Special Publications, **190**, 1–14.

LANKESTER, E.R. 1884. A contribution to the knowledge of *Rhabdopleura*. *Quarterly Journal of Microscopical Science*, **24**, 622–647.

LANKESTER, E. R. 1915. *Some Diversions of a Naturalist*. Methuen & Co Ltd: London (1925 edn.).

LEWIS, C.L.E. 2001. Arthur Holmes' vision of a geological timescale. *In:* Lewis, C.L.E. & Knell, S.J. (eds) *The Age of the Earth: from 4004 BC to AD 2002*. Geological Society, London, Special Publications, **190**, 121–138.

PHIPPS MORGAN, J., RESTON, T.J. and RANERO, C.R. 2004. Contemporary mass extinctions, continental flood basalt, and 'impact signals': are mantle plume-induced lithospheric gas explosions the causal link? *Earth and Planetary Science Letters*, **217**, 263–284.

SPARKS, R.S.J., BAKER, L., BROWN, R.J., FIELD, M., SCHUMACHER, J., STRIPP, G. and WALTERS, A. 2006. Dynamical constraints on kimberlite volcanism. *Journal of Volcanology and Geothermal Research*, **155**, 18–48.

Missing values and Stratigraphy

We have now completed all that is necessary to carry out a computer cladistic analysis. This final article ties up some loose ends that be-devil palaeontologists – stratigraphy and missing values.

The unique property of fossils is that they come with a time dimension that palaeontologists cling to as their unique contribution to the reconstruction of the history of life. Unfortunately, stratigraphy and cladistic analysis have not always sat easily side by side, and the early days of cladistic discussions witnessed fierce arguments over the suitability of including fossils and stratigraphic data. Intuitively we would expect that taxa that occur earlier in the record would show more plesiomorphic states of any particular character and that they would be ancestral to later occurring taxa. But, given the imperfections of the fossil record, neither of these assumptions can be justified.

So what about ancestors? We saw in the introductory article that, for cladograms, ancestors were not recognised since cladograms are atemporal: they are statements about the distributions of characters. Cladistic relationships are expressed solely in terms of sister groups, and the nodes on a cladogram have no connotation of ancestry. Once the cladogram is interpreted as a tree then ancestors come into play – but, for cladists, only in certain circumstances. Figure 1A shows a cladogram translated to a tree by including a time dimension and the stratigraphic occurrence of the individual taxa. Some people will deem it desirable to move a stratigraphically older taxon into an ancestral position relative to another taxon (Figure 1B). There are two caveats i) the putative ancestral taxon should have no autapomorphies, and ii) the stratigraphic ranges of putative ancestor and descendant do not overlap. Some have argued that, in fact, this is a more

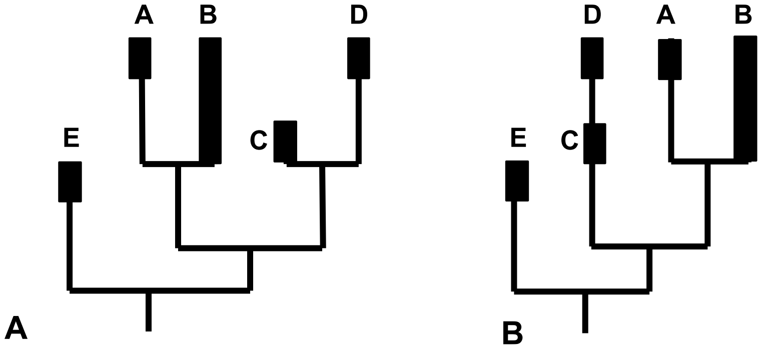

Fig.1. A. Recognising ancestors. Cladogram converted to a tree showing stratigraphic ranges of taxa, black boxes. B. In this circumstance it maybe possible to recognise that C is ancestral to D. See text

robust scientific theory than leaving them as sister groups because such a statement of ancestry and descent can be more easily disproved. Finding an autapomorphy in the presumed ancestor, or finding that the stratigraphic ranges do in fact overlap, will reject the theory.

In the early days, before computers, characters were polarised by using stratigraphy: the earlier occurring state was automatically accepted as the plesiomorphic condition with groupings established on the apomorphic state; a judgement made *a priori*. We have now seen in computer analyses that if a root taxon is chosen this automatically sets the plesiomorphic condition. This could be translated by choosing the oldest taxon as the root of the cladogram (Figure 2).

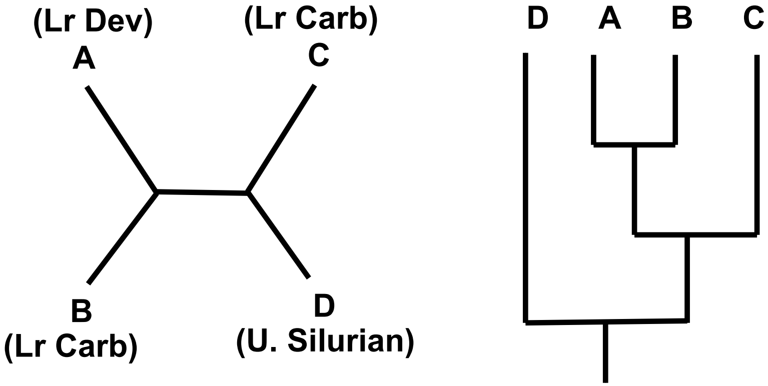

Fig. 2. Rooting the network calculated by PAUP* by using the stratigraphically oldest taxon will determine the shape of the tree.

Another way in which stratigraphy has been used is to consider the stratigraphic distributions of taxa as additional characters to be used alongside morphological characters as part of the tree building process. Steps are added to the length of each of the possible trees for any occurrence of a mismatch between the branching order and the stratigraphic occurrence. Alternative trees are examined and those that minimise the number of steps are chosen as optimal solutions. This procedure is known as stratocladistics. There are several problems with this method: some theoretical (*e.g.* time cannot be considered the same as a morphological character), and some practical (*e.g.* the method is sensitive to how finely we divide the stratigraphic record). If some of you wish to explore the issues further then a debate, moderated by Andy Smith (1998) on the *Nature* website, will lead you to the issues and literature. I'm having nothing more to do with it!

Once stratigraphic data has been added to a phylogenetic tree then this can be put to many uses most of which you probably know (*e.g.* calculating rates of evolution, comparing earth history with phylogeny, calibrating molecular clocks *etc*). There are a plethora of methods that have used different statistics to measure the congruence of the phylogeny with the stratigraphic occurrence of included taxa (see Norell 2001 for a summary of methods). Another way in which it has been used is to arbitrate between two equally parsimonious solutions derived from morphological or molecular data (or combination of the two). Let us say that, as a result of analysis, we ended up with two equally parsimonious trees (Figure 3). We can plot the stratigraphic distribution of the taxa (if the taxa were Recent then we would use the assumed fossil record of these taxa) and then

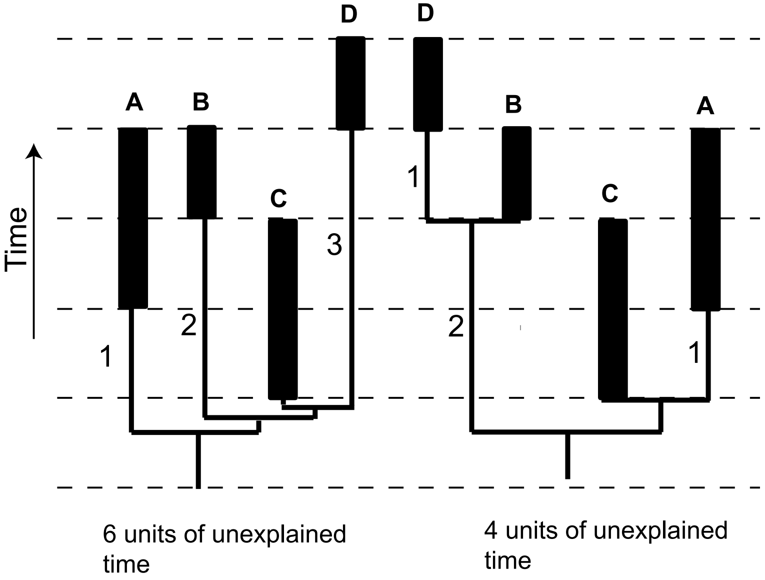

Fig. 3. Using stratigraphy to arbitrate between two equally parsimonious theories of relationship. The tree on the right assumes the least amount of unexplained time and might therefore be chosen on this evidence. The numbers refer to units of time.

calculate the amount of time that we must assume is unfilled by fossils ('unexplained time' in Fig. 3). We would then choose the tree with the least amount of unexplained time as the one to use for further analysis. Note, this is not saying that one tree is more optimal than another. It is simply choosing between two equals when there is no other evidence to hand. For a real example of the method see Smith & Littlewood (1994).

The other major area that palaeontologists have to deal with is missing values. Of course missing data can strike anywhere and for several different reasons. Unfortunately, they tend to be rife in data sets that try to combine modern with fossil taxa. For us there are two principal causes of question marks. First, genuine missing data: the part of the animal has not yet been found, or is unlikely to be found (*e.g.* soft parts). Second, question marks may be introduced due to evolutionary divergence (a phenomenon equally applicable to modern taxa). For example, let us assume that we were trying to establish the relationships between mammalian Orders. It is traditional to include a wealth of characters relating to teeth (presence, shape, height, angle *etc*, of particular tooth cusps). But, none of these characters could be scored for anteaters because they have no teeth (presumed lost). Technically the codes for these characters for anteaters would be 'not applicable' or 'illogical' since the structure to which the variation making up the character refers is not there. For computer cladistic analysis 'not applicable' codes are entered as question marks. That said, it is always a good idea to differentiate between genuine missing data and 'not applicable' in the written data matrix that you publish.

There are three adverse effects that question marks bring to computerised cladistic analyses. First, they can increase the number of equally parsimonious trees. Second, they can destroy resolution among taxa that are known by more complete data (question marks will not alter relationships known by complete data). Third, they can create spurious nodes on trees that have no evidential basis. Note the use of the word 'can'; question marks do not always have these deleterious effects. When PAUP* encounters a question mark it will try and insert real data codes to try and find a most parsimonious solution within the constraints offered by the real data.

Let's take these deleterious effects one by one. First, increasing trees, that I can best explain by an, admittedly old, example. This concerns a study done by Mike Novacek (1992), who was interested in the relationships between the orders of mammals (Figure 4). He analysed the relationships between the twenty recognised orders of living mammals using 88 morphological characters. He obtained eight equally parsimonious cladograms. To this analysis he added seven fossil taxa with varying amounts of missing data (25%–57%). Analysis of this combined matrix resulted in 6,800+ equally parsimonious cladograms (this was the limit of the computer memory in those days) and a strict consensus tree also destroyed resolution of a clade originally recognised to contain primates, tree shrews (Scandentia), flying lemurs (Dermoptera) and bats (Chiroptera). Notice here that it has not changed the topology because the original grouping of primates, tree shrews, flying lemurs and bats is not denied by the polychotomy here.

The situation is actually worse than it looks because of the limitations of the PAUP* algorithm. PAUP* will actually report trees that cannot be supported by any alternative 'real' data that is inserted in place of question marks (for a precise explanation of this see pages 82 – 85 in Kitching *et al*. 1998). These are spurious trees.

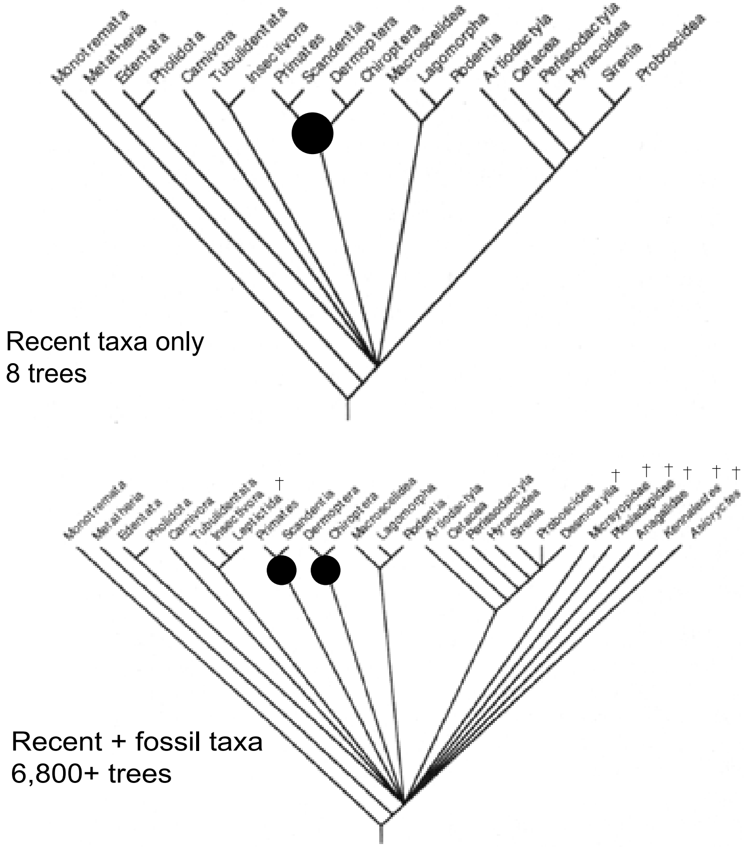

Fig. 4. Introduction of fossils with many missing values can result in significant increase in numbers of equally parsimonious trees and loss of resolution amongst taxa known by complete data. See text. (After Novacek 1992)

So, it would obviously be advantageous to reduce the number of question marks. We could delete characters that show high percentages of '?'. Or we could delete taxa with high percentages of '?'. Or we might recode characters (see later) and eliminate the need for '?'.

Deleting taxa is an obvious ploy: but, in a sense this negates the purpose of the analysis. Closer inspection of the possible disruptive effect that taxa with many '?' can have shows that it is not the ratio of real data to question marks that is important. Rather it is the complement of real data remaining that has the greatest effect. Remember that question marks cannot influence where a taxon is placed. But real data can. If the real data tends to place that taxon in very different parts of the tree it will have a very disruptive effect in collapsing nodes to result in poor resolution. [If you think back to the previous article, the Adams consensus dealt with situations like this]. In situations like this we may decide to delete that taxon from the computer part of the analysis, then place the rogue taxon on the tree afterwards according to the real data that it had.

Mark Wilkinson has thought about this problem a little deeper and has come up with a rule of thumb to advise us on when it may be safe and when it may be unsafe to delete taxa with many question marks. He calls this procedure 'Safe Taxonomic Reduction' and has a program to scan data matrices that will isolate the safe ones (TAXEQ). The routine scans the matrix and looks for taxa that are computationally equivalent, and identifies those that can be safely deleted without affecting the resulting tree(s). Figure 5 illustrates this. Taxon B has exactly the same real codes as the more completely known Taxon A. The inclusion of Taxon B cannot yield any different trees but can only add to the number of trees and mask any signal produced by the real data. It can be safely deleted. Taxon C, however, shows a different real value for character 2 and thus cannot be deleted. This technique has been very successfully applied (Wilkinson & Benton 1996) but, unfortunately, does not always work.

SAFE TAXONOMIC REDUCTION

Taxon A	**1**	**2**	**0**	**?**	**1**	**0**	**1**	**2**	**0**	**1**	**0**	**1**	
Taxon B	**1**	**?**	**0**	**?**	**1**	**0**	**?**	**?**	**0**	**?**	**0**	**1**	**……safe**
Taxon C	**1**	**0**	**0**	**?**	**1**	**?**	**1**	**2**	**?**	**?**	**?**	**1**	**……unsafe**

Fig. 5. Some taxa with many question marks can be safely removed from the matrix in the knowledge they cannot contribute new theories of relationship. See text.

Another tactic that you may be able to use to reduce the number of question marks is to recode some of the data that involves codes that stand for "non-applicable". We met this in the Character Coding article of this series. Suppose we had some taxa with tails and some without. Of those with tails some had red and some blue tails. A common way to code this variation is to use two characters. The first specifies the presence/absence of the tail. The second codes for the colour and assigns a 'non-applicable' code or '?' to taxa lacking tails. It has been shown through simulation analyses that such coding can actually lead to the identification of more parsimonious trees that can only be validated by assigning a colour to those taxa lacking tails (Maddison 1993). A solution to this problem is to use one multistate character where the '0' state is no tail, '1' state = blue tail; and '2' state = red tail (run the character unordered).

The last minefield for question marks that I will mention is the problem of spurious nodes. Sometimes, after analysis sister group pairings may be identified that cannot be justified by the data actually present. In other words, the only reason that the node is there is because the computer has assigned real data in place of question marks. This is not an uncommon occurrence and I have seen several studies where far reaching conclusions are drawn on unjustified sister group pairings. The key is to check the character change output carefully, to make sure that, in all cases, both of the sister groups have real data for at least one of the relevant characters supporting that particular node.

If some of you are interested in following up on discussions of question marks in palaeontological data then I recommend reading five papers that were published sequentially in Volume 23, issue 2 of *Journal of Vertebrate Paleontology* (2003) pages 263–323.

I have come down rather heavily on the question mark. We all use them, sometimes by choice but usually by necessity. In most analyses their effects are relatively benign, particularly if you take care and check what they are doing in any particular analysis. For instance, you could run many analyses sequentially removing and replacing taxa with high percentages of question marks. Those taxa that can be removed without influencing the relationships among the rest are clearly only adding confusion and are best dumped. Ultimately we have to live with them.

This article concludes this short series on cladistic analysis. The intention of these articles is to strip away some of the mystery of cladistic jargon (spell check 'cladistic' and you get 'sadistic') and to allow you to read papers including cladistic analyses without a stiff gin. There are many other aspects to cladistics and techniques grouped under the cladistic rubric (maximum likelihood and Bayesian analysis are two). Usually these are more pertinent to analysis of molecular data. Some of the issues spoken about now centre on the analysis of supertrees. There are many many cladistic analyses of overlapping taxonomic animal and plant groups out there. The issue is how to combine information from these to form one supertree of life.

As a parting shot I will say that, in cladistic analysis, there is a lot of mathematical manipulation, ever more sophisticated. But we must never forget that most of the crucial decisions we have to make are biological/palaeontological. And this is especially true of the delimitation of characters and codes – how we partition the variation that we see. At the very least I hope these articles will allow you to read the results of cladistic analysis with an increased level of critical understanding. Some of you may even want to try it for yourselves. Good Luck!

Peter Forey

REFERENCES

KITCHING, I. J., FOREY, P. L., HUMPHRIES, C. J. and WILLIAMS, D. M. 1998. *Cladistics*. Oxford University Press, Oxford, 228 pp.

MADDISON, W.P. Missing data versus missing characters in phylogenetic analysis. *Systematic Biology*, **42**, 576–81.

NORELL, M. A. 2001. Stratigraphic tests of cladistic hypotheses. *In* Briggs, D. E. and Crowther, P. R. (eds) *Palaeobiology II*. Blackwell Science Ltd, Oxford, pp. 519–522.

NOVACEK, M.J. 1992. Fossils as critical data for phylogeny. *In* Novacek, M. J. and Wheeler, Q. D. (eds). *Extinction and Phylogeny*. Columbia University Press, New York, pp. 46–88.

SMITH, A.B. (ed.) 1998. Is the fossil record adequate? <**http://www.nature.com/nature/debates/fossil/index.html**>

SMITH, A. and LITTLEWOOD, D.T.L. 1994. Paleontological data and molecular phylogenetic analysis. *Paleobiology*, **20**, 259–273.

WILKINSON, M. and BENTON, M. J. 1996. Sphenodontid phylogeny and the problems of multiple trees. *Philosophical Transactions of the Royal Society of London*, **351**, 1–16.