

## TECHNIQUES FOR MOLECULAR ANALYSIS

# A step by step guide to phylogeny reconstruction

C. Jill Harrison and Jane A. Langdale\*

Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

Received 13 July 2005; revised 20 September 2005; accepted 29 September 2005.

\*For correspondence (fax +44 1865 275147; e-mail jane.langdale@plants.ox.ac.uk).

---

## Summary

**The aim of this paper is to enable those who have never reconstructed a phylogeny to do so from scratch. The paper does not attempt to be a comprehensive theoretical guide, but describes one rigorous way of obtaining phylogenetic trees. Those who follow the methods outlined should be able to understand the basic ideas behind the steps taken, the meaning of the phylogenetic trees obtained and the scope of questions that can be answered with phylogenetic methods. The protocols have been successfully tested by volunteers with no phylogenetic experience.**

**Keywords:** beginners, PAUP\*, protocol, phylogeny.

---

## Aims of phylogeny reconstruction

A phylogeny is the evolutionary history of a group of entities. Given that this can only truly be known in exceptional circumstances, the main aim of phylogeny reconstruction is to describe evolutionary relationships in terms of relative recency of common ancestry. These relationships are represented as a branching diagram, or tree, with branches joined by nodes and leading to terminals at the tips of the tree (Figure 1). The three main types of relationship distinguished are monophyly, paraphyly and polyphyly (Hennig, 1966). Monophyletic and paraphyletic groups have a single evolutionary origin. Monophyletic groups include all the descendants from a single ancestor, as well as that ancestor. If one lineage emerging from a monophyletic group is removed, a paraphyletic group remains. In contrast, polyphyletic groups result from convergent evolution, and the characters that support the group are absent in the most recent common ancestor (Kitching *et al.*, 1998). In gene families these principles approximate to orthology and paralogy (Fitch, 1970). Orthology refers to groups of genes that reveal species phylogeny. Thus, within a monophyletic gene group each species is represented by a single orthologue. In contrast, paralogues reveal the history of a gene family. Thus, within a gene group each species may be represented by a number of paralogues.

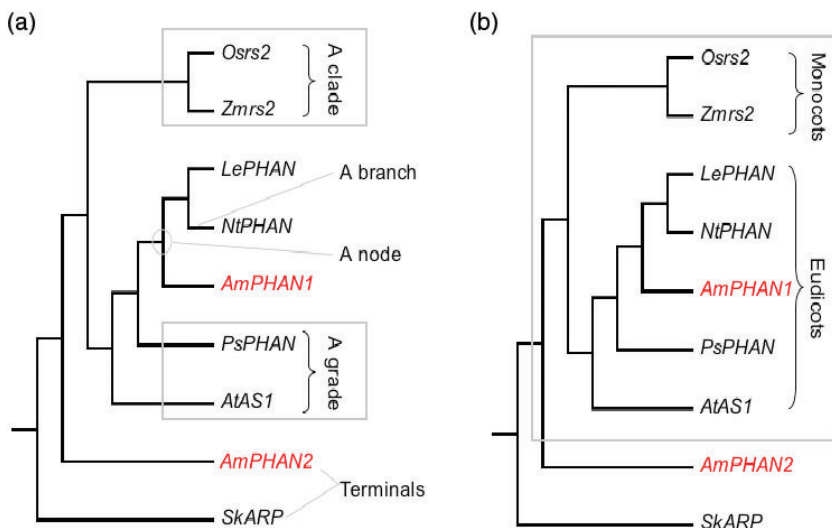
## Overview of phylogenetic analysis

### *Choosing a study group*

Before embarking on phylogeny reconstruction it is important to think about the specific biological question to be answered. It is advisable to sample as densely as possible, to reduce artefactual associations between terminals (Hillis, 1996). For example, if the aim of the reconstruction is to examine when duplications occurred within a gene family from a single species it is appropriate to sample the gene family from that species comprehensively. However, if the aim is to understand how a gene family evolved it is important to sample as widely as possible, not only within but also between species. A good first step is to survey the literature pertaining to the group of interest. This will inform the choice of species and genes to be included in the analysis and will indicate clades in which relationships are likely to be resolved and statistically supported in the resulting phylogeny. It will also highlight clades that should benefit from greater sampling or more attention in alignments.

### *Sampling the group of interest*

There are four main sources of data for reconstructing the phylogeny of a gene family: published sequences of characterized genes (which may not always clearly appear in



**Figure 1.** Some phylogenetic terminology illustrated using a phylogeny of *ARP* genes redrawn from Harrison *et al.* (2005) and rooted on a *Selaginella ARP* gene, *SkARP*.

In the phylogeny, *Osrs2* is the sister terminal to *Zmrs2*, and together these form a monophyletic group. *PsPHAN* and *AtAS1* are paraphyletic with respect to the monophyletic sister group containing *AmPHAN*, *NtPHAN* and *LePHAN*. 'Clade' and 'monophyletic group' can be used interchangeably, as can 'grade' and 'paraphyletic group'.

(a) Terminals, branches, nodes, a clade and a grade are indicated.

(b) To date, *Antirrhinum* is the only species reported to have two *ARP* genes, *AmPHAN1* and *AmPHAN2*, and these are paralogues. The box indicates a monophyletic (orthologous) gene group.

database searches, but should be familiar to the researcher from the literature); gene databases such as NCBI; EST project databases; and unpublished data from colleagues. The advantage of including unpublished sequences from EST databases is that it increases species sampling and also increases the possibility of deducing the point at which gene duplications occurred within the family of interest. In addition to searching a number of different databases, sampling is optimized by using the *TBLASTX* option at the sequence retrieval stage in *BLAST* searches (Altschul *et al.*, 1997). As opposed to other options, this program translates the nucleotide sequence in all six frames and compares the output against all the translated sequences in the database. It therefore maximises the potential for retrieving sequences similar to the gene of interest.

#### Sequence retrieval

The number of sequences retrieved from *BLAST* searches varies depending on the size of the gene family, and what is chosen for inclusion in further analyses will vary accordingly. It is feasible to download all of the sequences for a small gene family, but with a large gene family some selection is required. It is important to be pragmatic at this stage, because analysis of 100 sequences can take days to compute. When sequences are retrieved from *BLAST* searches they are allocated an *e*-score, which is an indication of the degree of similarity between the initial sequence used for searches and the sequence retrieved. The closer the *e*-value is to 0, the higher the degree of similarity between the two sequences. For large gene families there may be a clear cut-off point between the *e*-scores of the group of interest and further gene family members. If there is no clear cut-off, the literature can be used to identify genes in the list of retrieved sequences that, on the basis of their function, are likely to be outside the group of interest. The *e*-scores of

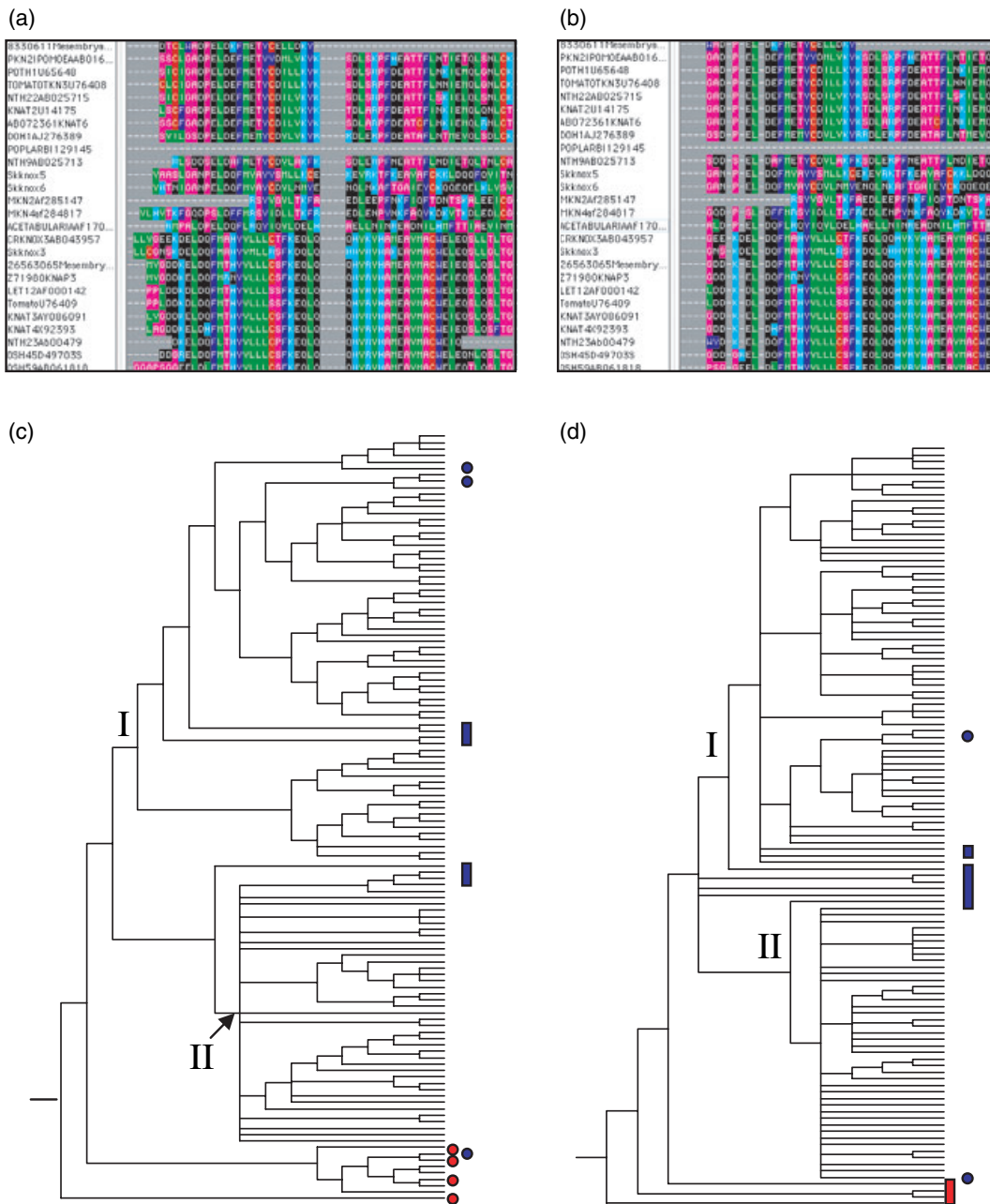
those sequences can then be used as a guide for the cut-off point (also see Hall, 2001).

#### Amino acid or nucleotide data?

Both amino acid and nucleotide data can be analysed to generate a phylogeny and there has been much debate about which is best (e.g. Simmons *et al.*, 2002a,b). The main argument for using amino acid data to infer phylogeny is that there are more possible character states for amino acids as opposed to nucleotides (20 versus 4). For the same reason, alignment of amino acid sequence data is also generally easier. However, the increased number of characters in nucleotide sequences can lead to better resolution of the tree, for example when the gene family of interest has just one highly conserved domain that is barely divergent at the amino acid level. The argument against using nucleotides is that with only four possible character states at one position there is a possibility that terminals may share a character state by chance. However, such 'saturation' at a particular position in the alignment is minimized by thorough sampling (Källersjö *et al.*, 1999). Fortunately, many alignment programs allow transition between sequence types, so both possibilities can be tried with ease.

#### Alignment

Phylogenetic methodology relies on the assumption that the characters used to generate trees are homologous. For gene family phylogenies careful alignment of sequence data fulfils this requirement. Sequence alignment can be achieved automatically or manually. Automatic alignments may fail to correctly identify regions of conservation within a gene, whereas manual alignments allow this but are more labour intensive (see example in Figure 2 and further discussion in Baldauf, 2003). Using published and auto-



**Figure 2.** The effect of misalignment on tree topology.

(a) Part of a KNOX protein sequence alignment generated by CLUSTALW.

(b) Part of an equivalent sequence alignment generated manually.

(c) One of eight most parsimonious trees ( $L = 12949$ ) generated by analysis of the CLUSTALW alignment in (a).

(d) One of 124 most parsimonious trees ( $L = 1892$ ) generated by analysis of manually aligned data in (b).

Trees are rooted on a mouse homeodomain protein sequence. The origin of class I (I), and class II (II) KNOX clades are indicated on the trees. Members of the out-group are indicated in red and genes from species that are taxonomically basal are indicated in blue. Although there were fewer trees generated from the CLUSTALW alignment and the consensus is better resolved, there are major problems with the tree. The main problem is that the in-group is not monophyletic, as indicated by the interspersed placement of the four out-group sequences. Furthermore, the placement of representatives from the most basal species is widespread, whereas they should be together.

matic alignments as a guide to manual alignment offers a good compromise, and is more rigorous than automatic alignment alone. There are many programs for manual alignment, including Se-AL (Rambaut, 1996) and BioEdit

(Tom Hall, Ibis Therapeutics, Carlsbad, CA, USA). These have the benefit of allowing easy transition between nucleotide and translated sequence formats and also output files in useful formats. Both are available free on the web.

*Method of analysis*

Once data are aligned, there are many different types of phylogenetic analysis that can be implemented (recently reviewed by Holder and Lewis, 2003). The type of analysis used will be determined by compromise between the length of computational time and the degree of rigour required. The main techniques are distance, parsimony and likelihood (including Bayesian analysis). All three can be performed using PAUP\* software (Swofford, 2003) that is available from Sinauer Associates Inc. Publishers, Sunderland, MA, USA. There are many alternative programs that perform the same functions and are equally valid to use (see web resource list at the end of this paper).

*Distance.* Distance methods [e.g. neighbour joining (NJ), distance and minimum evolution] calculate pairwise distances between sequences, and group sequences that are most similar. This approach has potential for computational simplicity and therefore speed. However, distance methods do not allow an analysis of which characters contribute to particular groupings. As with other methods, the outcome may depend on the order in which entities are added to the starting tree, but because only one tree is outputted it is not possible to examine conflicting tree topologies. Although distance methods are often useful for making an initial tree, they should be used for final trees with caution. Instead, parsimony and likelihood are preferred because they have the potential to rigorously explore the relationship between the tree and the entities included. Parsimony and likelihood use different criteria to choose the best trees. In these analyses the branches of a starting tree are rearranged to form the tree that minimises the number of character state changes (parsimony) or the tree that best fits the data (likelihood).

*Parsimony.* Parsimony assumes that shared characters in different entities result from common descent. Groups are built on the basis of such shared characters, and the simplest explanation for the evolution of characters is taken to be the correct, or most parsimonious one. With multiple characters, different groupings may be equally plausible, or equally parsimonious, and therefore multiple trees are generated. In such cases, a strict consensus tree should be derived that includes only topologies that are not contradicted in any of the initial trees. If the strict consensus tree is unresolved there is no congruence between initial trees, and thus it is likely that the data used to build the tree are phylogenetically uninformative. A majority rule consensus tree shows nodes that are consistent in half to all of the most parsimonious trees and the percentage of trees in which a given topology exists is shown on the branches. However, since by definition all most parsimonious trees are considered equally good, if any one contradicts the others the node

in question should collapse. Hence majority rule trees are not informative about phylogeny.

*Likelihood methods.* In contrast to parsimony, maximum likelihood analyses compute the probability that a data set fits a tree derived from that data set, given a specified model of sequence evolution. A good first step is to compare the data against a set of models of sequence evolution and choose the one that best describes the observed pattern of sequence variation. Two programs in which this can be performed are Modeltest (Posada and Crandall, 1998) and MrModeltest (Nylander, 2004). Alternatively a user-specified model may be chosen. This model of sequence evolution is then used in the likelihood analysis. The analysis starts with a specified tree derived from the input dataset (for example a NJ tree) and swaps the branches on the starting tree until the tree with the highest likelihood score (i.e. the best probability of fitting the data) is gained. This score is a function both of the tree topology and the branch lengths (number of character state changes). Likelihood analysis therefore allows an explicit examination of the assumptions made about sequence evolution. Likelihood methods are the most computationally demanding techniques for phylogenetic analysis. Currently only nucleotide data sets can be used to perform maximum likelihood (ML) analyses in PAUP\*. Bayesian inference is another likelihood method that is gaining popularity, but this cannot yet be implemented in PAUP\*. Instead, the program MrBayes should be used (Huelsenbeck and Ronquist, 2001; Huelsenbeck *et al.*, 2001). In Bayesian analysis, a further set of assumptions (termed priors) are inputted into the original model and the branch swapping algorithms differ. Likelihood methods produce a number of trees, one of which is usually found to be the most likely tree.

*Rooting trees*

If direct evidence of ancestor–descendant relationships is absent, the direction of change must be inferred by rooting the trees. Unless they are rooted, phylogenetic methods give rise to branching diagrams from which it is impossible to examine the direction in which traits change. In some instances it may not be necessary to root the generated trees. For example, if the hypothesis is to test whether a group of genes are orthologous, and those genes are dispersed amongst other genes on the tree, the hypothesis is essentially refuted. However, in most cases knowledge of the direction of change is fundamental to our understanding of evolutionary processes.

There are two good ways to root molecular trees. Outgroup rooting (Maddison *et al.*, 1984) compares the character states in the group of interest (the in-group) with those in a group that is closely related to, but definitely not in, the in-group (the out-group). These differences are used to infer the direction of character change in the resultant tree.

Out-groups can be selected on the basis of prior knowledge of the group of interest, or may become apparent during alignment (see Figure 3). In gene families, appropriate out-groups share at least one conserved domain with the group of interest. It is often not trivial to find an appropriate out-group because in candidate sequences, domains that can be adequately aligned for phylogenetic analysis may not provide sufficient variance and other domains may be too variant to align. Performing two separate rounds of analysis can solve this problem (Fitter *et al.*, 2002; Harrison *et al.*, 2005). The first round of the analysis should sample as widely as possible within the gene superfamily and should be carried out using only sequence from the most conserved domains. This will allow identification of gene clades that are most closely related to the clade of interest and are potential out-groups for a second round of analysis. The second step should include sequences from the group of interest, plus place-holder representatives from the most closely related clades identified in the first step as potential out-groups. The root should be placed between the most distant of these (selected as the out-group) and the rest (Figures 3, 4). At this stage, it may also be possible to include extra sequence outside the most conserved domain. If the group of interest shares common descent, the sequences chosen as out-groups will all fall outside the in-group (Figure 4).

The second way of rooting is to use duplicated genes, where sequences from one gene clade are used to root another (Simmons *et al.*, 2000). Duplicate gene rooting has the advantage that it can reveal unexpected relationships

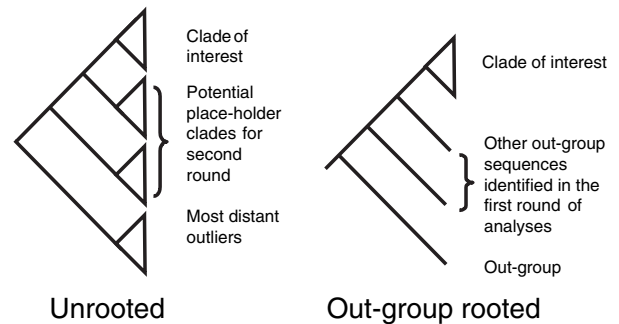


Figure 4. Out-group rooting for a gene family tree.

amongst the species or genes in the main clades where there has previously been ambiguous rooting (Brown and Doolittle, 1995; Gogarten *et al.*, 1989; Iwabe *et al.*, 1989; Mathews and Donoghue, 1999).

Rooting adds an extra node at the base of the tree, and by convention rooted trees are drawn with a stalk at their base. In trees with out-group rooting, the first or specified number of branches on the tree is part of the out-group. Be aware that as unrooted trees also branch at their base, they can be easily mistaken for rooted trees.

Statistical support for trees

As phylogenetic trees represent historical patterns of relationship that are generally incompletely sampled, they are

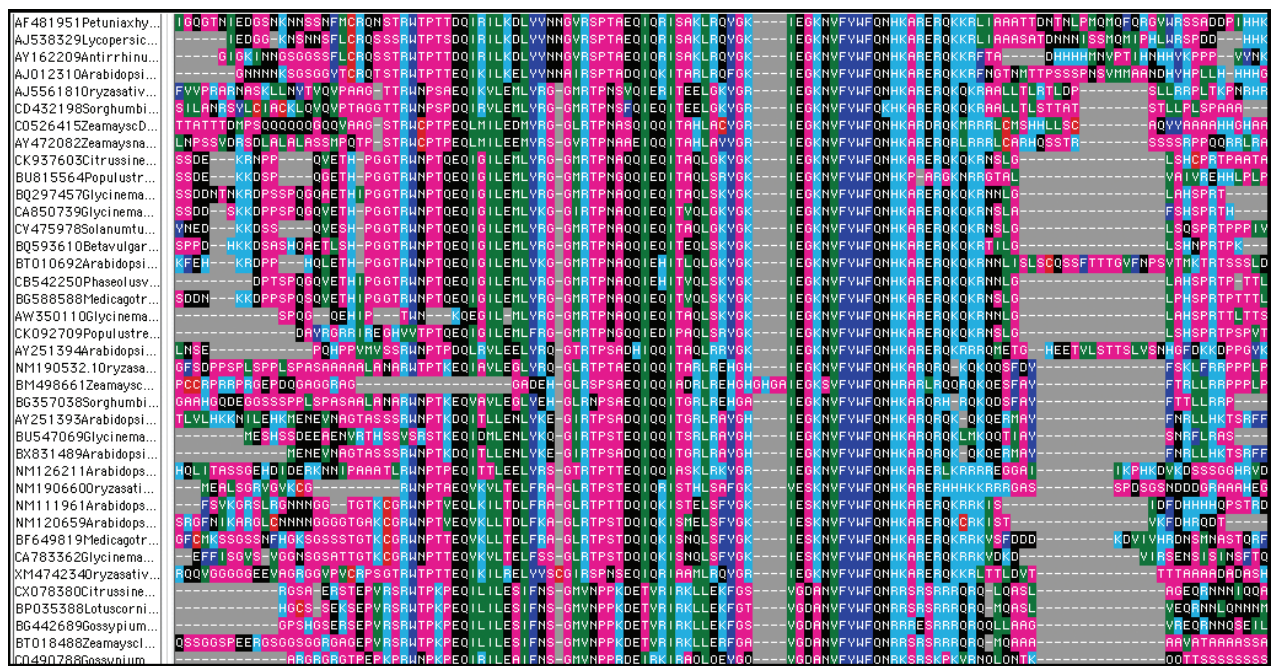


Figure 3. Candidate out-group sequences. Potential out-group sequences are indicated by a drop in sequence identity at the bottom of the alignment. In this instance the last five sequences are likely to comprise the out-group, and a good placement for the root would be between the *Gossypium* (CO490788) sequence and the rest.

hypotheses of relationship not factual depictions of evolutionary relationships. It is therefore hard to assess how accurately they reflect true biological events. However, there are ways in which the robustness of the data used for making the trees can be tested (note that this is not the same as accuracy). The quantifiers most commonly used are bootstrap and jack-knife support values. The bootstrap value shows the percentage of times that a clade appears when individual characters in the data set are randomly removed and replaced with data from another character from the same data set, and the analysis performed again for a specified number of replications (Felsenstein, 1985). The jack-knife value shows the percentage of times that a clade appears when a specified percentage of characters are randomly removed from the data set and the analysis performed again (Farris *et al.*, 1996). Reasonable support is roughly 70% and high support 95% (Felsenstein, 1985; Hillis and Bull, 1993).

In ML analyses, tree support can be evaluated by bootstrapping (which is more time-consuming than with parsimony analyses), and conflict between alternative tree topologies can be examined with less likely trees. In contrast, Bayesian analysis outputs both a tree and the support for that tree together. This is conceptually equivalent to ML with bootstrapping. The support value shown on the nodes of the tree is termed the posterior probability. The posterior is the actual probability of a node being correct. This sounds great, but given that the topology may be prone to sampling and analytical artefacts, and that the probability is dependent on the model used for analysis, it may not reflect evolutionary history!

## How to build phylogenetic trees

### Database searches and sequence retrieval

(i) Search NCBI by opening the BLAST page on the NCBI website (<http://www.ncbi.nlm.nih.gov/BLAST/>). Paste the sequence that you wish to place in the context of a phylogeny into the box that opens and select the TBLASTX option from the 'choose a translation' menu. Use the 'nr' databases. Scroll to the bottom of the page and make sure that the 'sequence retrieval' box is ticked in the format section. Select 'BLAST' and then 'format' to start the search. When the results are returned, scroll through the list to check e-scores and then go through selecting individual sequences to retrieve by checking the boxes next to the gene identifier numbers. Do not select genomic sequences or those with poor e-scores (values closest to 0 reflect greatest similarity to the search sequence). When all the required boxes have been checked, click on 'get selected sequences'. On the next page change 'display' to FASTA, and then 'send' all to file. Save the file to the desktop, open in Word and

then save as a text only (.txt) file. Examine the file and then using a combination of published data and the individual records on NCBI (open sequence file and select 'cds' to get coding sequence), identify the start codon in each sequence. Delete sequence 5' to the ATG from all sequences in the text file, to ensure that all sequences start at the beginning of the first open reading frame.

- (ii) Search EST databases at <http://www.plantgdb.org/PlantGDB-cgi/blast/PlantGDBblast>. Unless you specifically want to, do not search the databases from species that are already genetically well characterized for the gene of interest, as sequences retrieved are likely to be identical to those retrieved from NCBI. Instead access individual databases and check 'EST' and 'cDNA' boxes for each species to be examined. Select the TBLASTX option, paste in the sequence of interest and run the search. Display the retrieved sequences and copy and paste into the text file.
- (iii) Search algal EST databases on <http://www.kazusa.or.jp/en/plant/database.html> and <http://www.chlamy.org/chlamydb.html>. Select TBLASTX or BLASTN, choose the database of interest and run the search. Add the sequences retrieved to the text file.
- (iv) Add any unpublished sequences that you have to the text file.

### Formatting retrieved sequences

- (i) The text file must be formatted so that it is compatible with alignment programs. Open the file in Word and first remove all of the spaces and returns using the 'replace' function under the edit menu.
- (ii) Scroll through the sequences and insert a return between the end of each identifier line and the sequence, and between the end of each sequence and the next identifier line. Delete any unnecessary information in the identifier line, any polyadenylation sequences and any genomic sequences (Figure 5). *Do not remove the species name or the GenBank accession number.*
- (iii) Save as a text file.
- (iv) Next, identify the correct reading frame for each sequence. To do this, open the translation program at <http://us.expasy.org/tools/dna.html> and copy and paste each sequence from the text file in turn into the box. Select 'compact' as the output format and select 'translate sequence'. Examine the output to identify the correct frame (informed by an idea of what conserved motifs are present in the gene family). If necessary, copy and paste the sequence into the reverse complement program at [http://www.bioinformatics.vg/bioinformatics\\_tools/reversecomplement.shtml](http://www.bioinformatics.vg/bioinformatics_tools/reversecomplement.shtml) and

```
(a)
>U32344 gi|1167915:1-1149 Arabidopsis thaliana Shootmeristemless (STM)
mRNA, complete cds
ATGGAGAGTGGTTCCAACAGCACTTCTTGTCCAATGGCTTTTGCCGGGGATAAATAGTGATGGTCCGATGT
GTCCTATGATGATGATGATGATGCCGCCATCATGACATCACATCAACATCATGGTCATGATCATCAACATCA
ACAACAAGAACATGATGGTTATGCATATCAGTCACACCACCAACAAAGTAGTTCCTTTTTCTTCAATCA

(b)
>U32344 ArabidopsisthalianaShootmeristemless
ATGGAGAGTGGTTCCAACAGCACTTCTTGTCCAATGGCTTTTGCCGGGGATAAATAGTGATGGTCCGATGTGTCCTATGATGATGAT
GATGCCGCCATCATGACATCACATCAACATCATGGTCATGATCATCAACATCAACAACAAGAACATGATGGTTATGCATATCAGT
CACACCACCAACAAAGTAGTTCCTTTTTCTTCAATCA
>nextsequence
```

**Figure 5.** Sequence format for alignment.

(a) Downloaded sequence format.

(b) Modified format.

copy and paste the output reverse sequence into the text file in place of the original sequence. Delete nucleotides as appropriate in the text file to adjust the reading frame of each sequence to 5'–3' frame 1. If conserved motifs cannot be identified in any of the six frames, delete the sequence from the file.

- (v) To identify and remove duplicate sequences, open CLUSTALW (Ramu *et al.*, 2003) at <http://www.ebi.ac.uk/clustalw/> (It is equally valid to use CLUSTALX, but we find that versions that are currently freely accessible on the web are unreliable). Copy and paste the sequences from the text file into the alignment window and run the program. When the data are returned, open the alignment file, examine the sequences and identify duplicates. Go back to the initial text file and delete all duplicated sequences, saving the longest in each case (this is tedious to do manually but we have been unable to find a program that does pairwise alignments of the whole set and deletes the shortest sequence every time it finds an identical match).

### Aligning data

Remember that this step is the assessment of character homology and is therefore fundamentally important to the output. Rigorous alignment of a large data set can take months; it is also subjective and therefore it is always worth getting a second opinion. The steps outlined are optimized for the alignment of coding sequences, but can equally be applied to other nucleotide sequences by omitting instructions that refer to amino acid alignment. Alignment protocols differ slightly depending on whether you are using a Macintosh or a PC. Both approaches are outlined below.

### Macintosh

- (i) Download Se-AL from <http://evolve.zoo.ox.ac.uk/software.html?id=seal>. To load your sequences into

the Se-AL program, open the program, choose 'open' under the file menu and select your text file. Files do not open if they are not formatted exactly as shown in Figure 5. Save this Se-AL alignment file as version 1 – it will be the only one where you can revert to DNA sequence.

- (ii) To automatically align sequences in amino acid format, first change the sequence format to translated sequence by changing the alignment type to 'amino acid' under the alignment menu. Select 'export' under the file menu, select 'FASTA' under file format and check 'export alignment as displayed'. Save the FASTA file. Open CLUSTALW again, copy and paste the data from the FASTA file into the alignment window, select 'pir' under output format and run the program. When the data are returned, open the alignment file. Select 'save page as' from the browser file menu and save to the desktop. Reopen the file from within Word and save as a .txt file. Finally, open the text file from within Se-AL. Save the Se-AL file as version 2. You now have the alignment in a form that can be further edited manually.
- (iii) Spend time scanning the alignment for regions of conservation that CLUSTALW may have missed (it is easier to scan the sequence if you select 'use block colours' under the alignment menu). If necessary, manually adjust the alignment (see Baldauf, 2003 for further guidelines). To move entire sequences, double click on the sequence and drag it in either direction. To move blocks of sequence, select the block and drag it.
- (iv) When the alignment is finished, copy and paste the conserved regions into a new file and save it as version 3. Keep the alignment file for reference and publication. If the tree needs to be built using DNA data, go back to the Se-AL version 1 file and manually edit it to look identical to the version 3 file. Change alignment type to 'DNA' under the alignment menu and then save the file as version 4.

- (v) To transfer Se-AL alignment version 3 or 4 files to PAUP\*, select 'export' under the file menu. Select 'NEXUS' under file format and check 'export alignment as displayed'. Save the NEXUS file.

#### PC

- (i) Download BioEdit alignment software from ([http://www.molbiol.bbsrc.ac.uk/reviews/bioedit\\_review.html](http://www.molbiol.bbsrc.ac.uk/reviews/bioedit_review.html)). To load sequences into the program, choose 'open' under the file menu and select your text file. Save this BioEdit file as alignment version 1 – it will be the only one where you can revert to DNA sequence.
- (ii) Select names of sequences with the mouse and change data to amino acid format by selecting 'translate or reverse translate (permanent)' under sequence menu. Save the file as a FASTA file.
- (iii) To automatically align sequences in amino acid format, open CLUSTALW, upload the FASTA file, select 'pir' as the output option and run the program. When the data are returned, open the alignment file, select 'save as' from the browser file menu and save as a text file to the desktop. Finally, open the text file from within BioEdit. Save the BioEdit file as alignment version 2. You now have the alignment in a form that can be further edited manually.
- (iv) Spend time scanning the alignment for regions of conservation that CLUSTALW may have missed (it is easier to scan the sequence if you select 'inverse colours' under the view menu). If necessary, manually adjust the alignment. Move sequences along the alignment individually by clicking on the sequence with the mouse, unclicking and then dragging the sequence across. As it is impossible to move entire sequences bidirectionally, ascertain which sequence has regions of conservation most 3' and align other sequences with respect to that.
- (v) When the alignment is finished, save the file again as alignment version 2. Delete unaligned domains by changing the mode window to 'edit', highlighting the block to be deleted and using the 'delete' key on the keyboard. Save the modified file as BioEdit alignment version 3. Keep the alignment file for reference and publication. If the tree needs to be built using DNA data, go back to the BioEdit version 1 file, select names of sequences with the mouse and change data to amino acid format by selecting 'toggle translation' under the sequence menu. Manually edit the alignment to look identical to the version 3 file. Change back to DNA sequence by selecting names of sequences with the mouse and selecting 'toggle translation' again under the sequence menu. Save the file as BioEdit alignment version 4.
- (vi) To transfer BioEdit version 3 or 4 files to PAUP\*, select 'file', 'export', 'sequence alignment' and then 'nex'.

#### Running a parsimony analysis in PAUP\*

The number of trees resulting from a parsimony search increases hugely with increasing numbers of entities included in the analysis (Felsenstein, 1978). With large data sets it is computationally unfeasible to find all possible trees. Instead, heuristic search strategies are used to examine a subset of the trees. The starting parameters of a heuristic search can affect the outcome of the search, and search strategies can be modified to affect the stringency of the analysis. Changeable parameters include the sequence of addition of terminals to the starting tree, branch swapping algorithms, and the number of replicates of the analysis run. The significance of these and the effects of their alteration are discussed in detail in Felsenstein (2004). The search strategy of Catalán *et al.* (1997) is reasonably stringent and is recommended here, as the order in which entities are added to the starting tree is randomized, emphasis is placed on branch swapping at deep rather than shallow nodes, and the whole search is replicated 1000 times. Again, protocols differ for Macintosh and PC users.

#### Macintosh

- (i) Open the NEXUS file in PAUP\* and click on 'execute' under the file menu. Problems in this step arise from an incorrectly formatted NEXUS file; if new entities (taxa) or characters have been added, the parameters of the NEXUS file (Ntax, Nchar) need to be changed too.
- (ii) Under the analysis menu select 'parsimony' and 'heuristic' search.
- (iii) Set the parameters for the heuristic search. In the box that opens select 'best trees' under keep and then select 'maxtrees'. In the menu that appears, select 'automatically increase by 100' and select OK. Next change the main menu from 'general search options' to 'stepwise addition options'. In the next screen, select 'random' under addition sequence, set replicates to '1000' and select hold '1' tree at each step. Next change the main menu to 'branch swapping options' and select 'TBR'. Select save no more than '2' trees  $\geq$  score '5' for each replicate and 'swap on best trees' only. Select search to start the analysis.
- (iv) When the analysis has finished, note the number of trees, the length of the shortest tree, and save the trees using the 'save trees to file' option under the 'trees' menu. The trees are saved as data, not as a pictorial representation of the tree. If a pictorial representation is required, follow the steps outlined in 7. To reopen the .tre file at a later date, first reopen the original .nex file by selecting 'open' under the file menu and then selecting 'execute'. Under the trees menu, select 'get trees from file', select the .tre file of interest and then select 'get trees'.

- (v) If there is more than one tree, select 'compute the consensus tree' under the trees menu. Select both the strict and majority rule (50%) consensus options.
- (vi) Root the trees by selecting 'root trees' under the trees menu and then 'rooting options'. On the screen that appears select 'root tree at internal node with basal polytomy'. Then select 'define outgroup' and on the next screen, use the mouse to highlight one sequence from the out-group. Select 'to outgroup', then 'OK' in this and the next screen and 'root' in the final screen. Do not use the entire out-group as the root, as this forces the putative in-group to form a monophyletic group.
- (vii) Go to the 'print trees' or 'print consensus trees' option under the trees menu and either print the tree, or select 'preview' and save as a PICT file for manipulation in other programs.

*PC.* To perform optimality analyses in PAUP\* on the PC, commands are required to set the search strategy. Commands are listed in the PAUP\* user manual (commands reference document), available at <http://paup.csit.fsu.edu/down.html#Anchor-58521>. Commands can be typed into the command line at the bottom of the display window, or can alternatively be pasted directly into the NEXUS file. The commands listed below instigate analyses equivalent to those outlined for Macintosh users. To change search strategy parameters refer to the PAUP\* manual.

- (i) Open the NEXUS data file from within PAUP\* in the 'edit' mode.
- (ii) To set analysis parameters for parsimony analysis, scroll to the bottom of the data matrix and after 'End;' hit return and type the following commands:

```
begin paup;
set criterion=parsimony maxtrees=100
increase=auto;
Hsearch start=stepwise addseq=random
nreps=1000 savereps=yes nchuck=2
chuckscore=5 dstatus=none;
savetrees file=[the elected name1] brlens=yes;
set root=outgroup;
outgroup[the name of the out-group taxon that will be
used as the root] /only;
gettrees file=[the elected name1];
contree all/majrule=yes treefile=[the elected
name2] ;
End;
```

Note that square brackets make a command invisible to PAUP\*. The commands listed will result in the production of most parsimonious trees (saved to the elected file) plus a strict consensus and a majority rule tree compiled from the most parsimonious trees (saved to another elected file). The elected names should be inserted as indicated above (minus the square brackets).

Similarly, the out-group identifier should be inserted minus the square brackets.

- (iii) Run the analysis by selecting 'execute' under the file menu. The data will be returned in the two elected files.
- (iv) To view the trees, download the TreeView program from <http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>. Open TreeView, and open the tree file from within it.
- (v) Use the icons at the top of the tree window to choose the plot type (including plots with branch lengths).
- (vi) Root the trees by selecting 'define outgroup' and 'root with outgroup' from the 'tree' menu.
- (vii) To save the tree for manipulation in other graphics programs select 'print trees' under the trees menu and then select 'picture' to save the file.

#### Statistical support for trees

Bootstrap and jack-knife analyses resample the original data set used for phylogenetic inference, and rerun analyses. Thus, the trees obtained may not be as good an estimate of phylogeny as the original trees obtained. Therefore, support values obtained should be manually superimposed on appropriate nodes of parsimony or ML trees.

#### Macintosh

- (i) To obtain bootstrap and jack-knife values open the NEXUS file that was used to make the original tree in PAUP\* and 'execute' it.
- (ii) Select the bootstrap or jack-knife option under the analysis menu. Enter '500 replicates' for each analysis, with '50%' deletion for jack-knife (see Farris *et al.* (1996) and Felsenstein (2004) for a discussion of this).
- (iii) Run the parsimony analysis by repeating steps (ii)–(vii) above, selecting 'print bootstrap consensus' at the last stage. The topology of the trees may differ from the original tree. Manually transfer the bootstrap or jack-knife percentages across from their respective trees to those clades which are also present in the original tree.

#### PC

- (i) To perform a bootstrap analysis with equivalent search parameters to those specified for Macintosh users, open the NEXUS file from within PAUP\* in the 'edit' mode, scroll to the end of the data matrix and type the following commands after 'End;':

```
begin paup;
set maxtrees=100 increase=auto;
bootstrap nreps=500 conlevel=50 treefile=[the
elected name3] keepall=yes cutoffpct=50/
start=stepwise addseq=random nreps=1000
savereps=yes nchuck=2 chuckscore=5
dstatus=none;
```

```
savetrees file=[the elected name3];
set root=outgroup;
outgroup [one of the outgroup taxon names]/only;
End;
```

Note that the elected name should be different from those used in the original analysis. Select 'execute' under the file menu to run the analysis.

- (ii) To perform a jack-knife analysis with equivalent search parameters to those specified for Macintosh users, open the original NEXUS file as above and type the following commands at the end of the file:

```
begin paup;
set maxtrees=100 increase=auto;
jackknife pctdelete=50 nreps=500 conlevel=50
treefile=[the elected name4] keepall=yes
cutoffpct=50 grpfreq=yes/start=stepwise
addseq=random nreps=1000 savereps=yes
nchuck=2 chuckscore=5 dstatus=none;
set root=outgroup;
outgroup [one of the outgroup taxon names]/only;
savetrees file=[the elected name4];
End;
```

Again, the elected name should be different from those used previously. Select 'execute' under the file menu to run the analysis.

- (iii) The display window will show statistical support for nodes of the trees. The topology of the trees may differ from the original tree and therefore bootstrap or jack-knife percentages have to be manually transferred to those clades which are also present in the original tree. Bootstrap and jack-knife values will not transfer to TreeView and thus trees must be printed from the PAUP\* display window using ScreenGrab software ([http://www.sofotex.com/ScreenGrab-1.0-download\\_L2072.html](http://www.sofotex.com/ScreenGrab-1.0-download_L2072.html)) or equivalent. This can involve more than one screen grab per tree!

### Likelihood analysis

For ML trees, the first step is to find a model of sequence evolution that fits the DNA changes in the aligned sequences that are being used. This can be done using the Modeltest program (Posada and Crandall, 1998) or MrModeltest (Nylander, 2004) in conjunction with PAUP\*. We advocate using Modeltest, as it tests more models of evolution than MrModeltest. *Note that only DNA data matrices can be analysed with ML in PAUP\*.*

### Macintosh

- (i) Download Modeltest from <http://darwin.uvigo.es/>.
- (ii) Open the NEXUS file in PAUP\* in the 'edit' mode.
- (iii) After the 'End;' command at the end of the data matrix, type in a return and then:

```
default lscores longfmt=yes;
```

Select 'execute' under the file menu.

- (iv) Open the modeltest folder, the PAUPblock folder within it and the 'modelblock PAUPb10' program. Select 'execute' under the file menu to run the program. The results will automatically be saved as 'modelscores' in the 'PAUPblock' folder.
- (v) Open the modeltest folder, the BIN folder within it and the modeltest program. In the window that appears, select 'file' and then browse to select the modelscores file (the argument box remains empty.) Run the program (click OK). This instructs the program to run through all the different models of sequence evolution featured by the program and to test sequence changes within the data against each of them.
- (vi) When the data are returned, scroll through the file to the Akaike information criterion (AIC) section. Copy the lines of text in this section that start with 'begin paup' and end with 'End;'. Paste this text into the end of the NEXUS file.
- (vii) To run the ML analysis, select 'execute' under the file menu. Select 'likelihood' and 'heuristic search' under the analysis menu. In the window that opens, change 'general search options' to 'starting tree options' and select 'neighbour joining'. Next, change the main menu to 'branch swapping options' and select 'TBR'. Select save no more than '2 trees  $\geq$ score 5' for each replicate and 'swap on best trees' only. Select 'search' to start the analysis.
- (viii) Root the tree and display it as before [steps (vi) & (vii) under parsimony].

### PC

- (i) Download Modeltest from <http://darwin.uvigo.es/>.
- (ii) To perform a likelihood analysis with equivalent search parameters to those specified for Macintosh users, open the NEXUS file in PAUP\* in the 'edit' mode. Remove or bracket out all of the commands that were used for parsimony analyses and then after the 'End;' command at the end of the data matrix type:
 

```
default lscores longfmt=yes;
```

 Save the file as 'likelihood start' and then save a second version as 'likelihood analysis'.
- (iii) Open the 'likelihood start' NEXUS alignment in the 'edit' mode of PAUP\*. After the 'End;' command at the end of the data matrix, paste in the text from the model fit file available at <http://www.plants.ox.ac.uk/langdalelab>.
- (iv) Execute the file to see which model best fits the data. The results will be saved automatically as a 'model scores' file in the folder from which the program was executed.

- (v) Close the likelihood start file.
- (vi) Download the program mtgui from <http://www.genedrift.org/mtgui.php>. Open the mtgui program, click 'select file' and open 'model'. In the window that appears, select 'modeltest'. Scroll through the window that opens to the 'AIC' parameters and copy the 'Lset' commands. Close the modeltest file.
- (vii) Open the likelihood analysis NEXUS file in the edit mode of PAUP\* and type the following commands at the end of the file.

```

Begin paup;
Set criterion=likelihood;
[paste the Lset commands from the previous step here]
Hsearch start=nj nchuck=2 chuckscore=5
dstatus=none;
savetrees format=nexus brlens=yes append=yes
file=[the elected name5];
lscores 1/scorefile=[the elected name5].sf
append=yes;
set root=outgroup;
outgroup [the name of the required taxon]/only;
showtrees all;
End;

```

- (viii) 'Execute' the NEXUS file. The tree produced is a rooted ML tree.

### Description of trees

There are a number of numerical outputs from phylogenetic analyses that describe how the data used to infer phylogeny fit the resultant tree. In particular, the number of trees and tree length, are generally reported. The number of trees is greater with more conflict between possible outcomes. The tree length shows the number of character changes in a given tree. These values can be found in the PAUP\* display window once an analysis is completed and should also be recorded in the saved tree file. Tree length is informative when considering the evolution of particular characters within a group, and therefore is particularly useful where critical evaluation of character evolution is required.

### Interpreting a phylogenetic tree

Once a phylogenetic tree is generated, what inferences can be drawn from it? First look at the strict consensus parsimony tree. A totally unresolved tree (the branches all come out from the same node) may indicate a rapid radiation (e.g. Richardson *et al.*, 2001), insufficient phylogenetic information in the original alignment or incongruence between tree topologies. It should be possible to distinguish between the latter two possibilities by looking at both the alignment and the individual tree topologies. In

the case of insufficient phylogenetic signal from the data, it is possible that the alignment stringency was too high, especially if only the most conserved domains have been accepted. The alignment should be checked to see where it might be possible to include more data. If protein sequence was used to infer phylogeny, the analysis could be rerun using the nucleotide data. In instances of poor resolution, it is also common for the statistical support to be very low. Look at the support values for the clades. If the support values are good to high, you can be reasonably confident that derived tree reflects the data used to generate it.

The next step is to compare the strict consensus parsimony and ML trees to see if they are congruent. Trees that are congruent have the same topology, so that regardless of the order in which terminals appear on the page the branching order and sister group relationships are equivalent. If they are not congruent, it is likely that one or both of the techniques used is suffering from an artefact, such as 'long branch attraction' (Sanderson *et al.*, 2000). This arises when groups are formed on the basis of similarity rather than homology, and may be difficult to detect, but can be minimized by maximising sampling. The outcome of a likelihood analyses may also be subject to artefacts, as it is dependent on the model of molecular evolution specified.

If the parsimony and ML trees are congruent, look at the relationships between the terminals of interest and see how they answer the biological question that was originally posed in terms of orthology and paralogy, and in terms of monophyly, paraphyly and polyphyly (Figure 1). Another useful concept is that of 'sister groups', which refers to clades that are most closely related to each other. For example, in Figure 1 the monocot sequences form a sister group to the eu-dicot sequences. Note that sister group relationships are not dependent on the graphical representation: the sister group relationship is the same whether the monocot or the eu-dicot clade comes at the top of the tree. Remember to bear in mind that the initial taxon sampling and alignment steps can have a significant effect on the quality of the tree generated.

### Alignment and tree presentation

Once alignments and trees are satisfactorily understood they can be prepared for presentation. Alignment and tree files should be placed in the public domain, either as an image file in publications or as an electronic file on a suitable web-based database such as TreeBASE (<http://www.treebase.org/treebase/>). Trees should be displayed and annotated such that they clearly convey the understanding that has been gained from phylogeny reconstruction.

### Further reading

A good review of phylogenetic principles and interpretation of gene family trees is provided by Thornton and DeSalle (2000). An accessible and an in-depth theoretical guide to phylogeny reconstruction are given by Page and Holmes (1998) and Felsenstein (2004) respectively. Finally, Hall (2001) provides a basic practical guide to navigating some phylogenetic programs, with minimal theoretical coverage.

### Useful web resources

Joe Felsenstein's website (<http://evolution.genetics.washington.edu/phylip/software.html>) has a reasonably comprehensive list of links to phylogeny programs available on the web. Some further resources that the testers of this protocol found useful are found on: [http://www.so.e.ucsc.edu/~karplus/compbio\\_pages.html](http://www.so.e.ucsc.edu/~karplus/compbio_pages.html).

### Acknowledgements

We thank Daniel Barker, Sheila McCormick, Elizabeth Moylan, Robert Scotland, Andrew Smith, and two anonymous reviewers for constructive comments on the manuscript. Work in the Langdale group is supported by the BBSRC and the Gatsby Charitable Foundation.

### References

- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
- Baldauf, S.L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends Genet.* **19**, 345–351.
- Brown, J.R. and Doolittle, W.F. (1995) Root of the universal tree of life based on ancient amino-acyl-tRNA synthetase gene duplications. *Proc. Natl Acad. Sci. USA* **92**, 2441–2445.
- Catalán, P., Kellogg, E.A. and Olmstead, R.G. (1997) Phylogeny of Poaceae subfamily Pooideae based on cp *ndhF* gene sequences. *Mol. Phyl. Evol.* **8**, 150–166.
- Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D. and Kluge, A.G. (1996) Parsimony jackknifing outperforms neighbour-joining. *Cladistics*, **12**, 99–124.
- Felsenstein, J. (1978) The number of evolutionary trees. *Syst. Zool.* **27**, 27–33.
- Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
- Felsenstein, J. (2004) *Inferring Phylogenies*. Sunderland Massachusetts: Sinauer Associates.
- Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.
- Fitter, D.W., Martin, D.J., Copley, M., Scotland, R.W. and Langdale, J.A. (2002) GLK gene pairs regulate chloroplast development in diverse plant species. *Plant J.* **31**, 713–727.
- Gogarten, J.P., Kibak, H., Dittrich, P. *et al.* (1989) Evolution of the vacuolar H<sup>+</sup>-ATPase: implications for the origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **86**, 6661–6665.
- Hall, B.G. (2001) *Phylogenetic Trees Made Easy: A How-to Manual for Molecular Biologists*. Sunderland Massachusetts: Sinauer Associates.
- Harrison, C.J., Corley, S.B., Moylan, E.C., Alexander, D.L., Scotland, R.W. and Langdale, J.A. (2005) Independent recruitment of a conserved developmental mechanism during leaf evolution. *Nature*, **434**, 509–514.
- Hennig, W. (1966) *Phylogenetic Systematics*. Urbana, Illinois: University of Illinois Press.
- Hillis, D.M. (1996) Inferring complex phylogenies. *Nature*, **383**, 130–131.
- Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192.
- Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**, 275–284.
- Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
- Huelsenbeck, J.P., Ronquist, F., Nielsen, R. and Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.
- Iwabe, N., Kuma, K.-I., Hasegawa, M., Osawa, S. and Miyata, T. (1989) Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA* **86**, 9355–9359.
- Källersjö, M., Albert, V.M. and Farris, J.S. (1999) Homoplasy increases phylogenetic structure. *Cladistics*, **15**, 91–93.
- Kitching, I.J., Forey, P.L., Humphries, C.J. and Williams, D.M. (1998) *Cladistics (2nd edn): The Theory and Practice of Parsimony Analysis*. Oxford, UK: Oxford University Press.
- Maddison, W.P., Donoghue, M.J. and Maddison, D.R. (1984) Out-group analysis and parsimony. *Syst. Zool.* **33**, 83–103.
- Mathews, S. and Donoghue, M.J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science*, **286**, 947–950.
- Nylander, J.A.A. (2004) *MrModeltest v2. Program Distributed by the Author*. Uppsala University: Evolutionary Biology Centre.
- Page, R.D.M. and Holmes, E.C. (1998) *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science Ltd.
- Posada, D. and Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.
- Rambaut, A. (1996) Se-AL: Sequence Alignment Editor (Se-AL v2.0a11). (Available at: <http://evolve.zoo.ox.ac.uk/>).
- Ramu, C., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.* **13**, 3497–3500.
- Richardson, J.E., Pennington, R.T., Pennington, T.D. and Hollingsworth, P.M. (2001) Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science*, **293**, 2242–2245.
- Sanderson, M.J., Wojciechowski, M.F., Hu, J.-M., Khan, T.S. and Brady, S.G. (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in plants. *Mol. Biol. Evol.* **17**, 782–797.
- Simmons, M.P., Bailey, D.C. and Nixon, K.C. (2000) Phylogeny reconstruction using duplicate genes. *Mol. Biol. Evol.* **17**, 469–473.
- Simmons, M.P., Ochoterena, H. and Freudenstein, J. (2002a) Amino acid vs. nucleotide characters: challenging preconceived notions. *Mol. Phyl. Evol.* **24**, 78–90.
- Simmons, M.P., Ochoterena, H. and Freudenstein, J. (2002b) Conflict between amino acid and nucleotide characters. *Cladistics*, **18**, 200–206.
- Swofford, D.L. (2003) *PAUP\*: Phylogenetic Analysis Using Parsimony (\* and Other Methods), Version 4.0b 10*. Sunderland, Massachusetts: Sinauer Associates.
- Thornton, J.W. and DeSalle, R. (2000) Gene family evolution and homology: genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet.* **1**, 41–73.