Review & synthesis

Ecography 36: 013–026, 2013
doi: 10.1111/j.1600-0587.2012.07773.x
© 2012 The Authors. Ecography © 2012 Nordic Society Oikos
Subject Editor: Ken Kozak. Accepted 23 October 2012

Building megaphylogenies for macroecology: taking up the challenge

Cristina Roquet, Wilfried Thuiller and Sébastien Lavergne

C. Roquet (cristina.roquet@gmail.com), W. Thuiller and S. Lavergne, Laboratoire d'Ecologie Alpine, UMR-CNRS 5553, Univ. Joseph Fourier, Grenoble 1, BP 53, FR-38041 Grenoble Cedex 9, France.

The last decades have seen an upsurge in ecological studies incorporating phylogenetic information with increasing species samples, motivated by the common conjecture that species with common ancestors should share some ecological characteristics due to niche conservatism. This has been carried out using various methods of increasing complexity and reliability: using only taxonomical classification; constructing supertrees that incorporate only topological information from previously published phylogenies; or building supermatrices of molecular data that are used to estimate phylogenies with evolutionary meaningful branch lengths. Although the latter option is more informative than the others, it remains under-used in ecology because ecologists are generally unaware of or unfamiliar with modern molecular phylogenetic methods. However, a solid phylogenetic hypothesis is necessary to conduct reliable ecological analysis integrating evolutive aspects. Our aim here is to clarify the concepts and methodological issues associated with the reconstruction of dated megaphylogenies, and to show that it is nowadays possible to obtain accurate and well sampled megaphylogenies with informative branch-lengths on large species samples. This is possible thanks to improved phylogenetic methods, vast amounts of molecular data available from databases such as Genbank, and consensus knowledge on deep phylogenetic relationships for an increasing number of groups of organisms. Finally, we include a detailed step-by-step workflow pipeline (Supplementary material), from data acquisition to phylogenetic inference, mainly based on the R environment (widely used by ecologists) and the use of free web-servers, that has been applied to the reconstruction of a species-level phylogeny of all breeding birds of Europe.

Over the last decade, a new synthesis between the disciplines of ecology and evolution has been emerging, emphasizing the need to account for potential feedbacks between ecological and evolutionary dynamics of natural systems (Webb et al. 2002, Johnson and Stinchcombe 2007, Lavergne et al. 2010, Mouquet et al. 2012). This is illustrated by the increasing interest shown in integrating phylogenetic data into different areas of ecological research (Table 1), such as studies of community assembly rules (reviewed by Cavender-Bares et al. 2009), large-scale patterns of diversity (Davies et al. 2008), biological invasions (reviewed by Thuiller et al. 2010), or forecasting global change impacts on different facets of diversity (Thuiller et al. 2011). Several methods of varying complexity and reliability have been implemented to integrate evolutionary information into ecological studies, but the development of this new era of 'ecophylogenetics' (Mouquet et al. 2012) has seen a number of methodological impediments (Sanderson and Driskell 2003). This review aims to demonstrate that these limitations can now be overcome and emphasise the need to construct reliable phylogenies based on molecular data, in order to have an accurate phylogenetic hypothesis to work with. The reconstruction of phylogenies from molecular sequences is a vast field that cannot be extensively reviewed here; for this reason we provide a non-exhaustive list of reviews, key works and other useful resources dealing with its main aspects (Supplementary material Appendix 1). Finally, we outline a pipeline (Supplementary material Appendix 2) that can be used as a basic reference to derive large-scale phylogenetic hypotheses, i.e. including several hundreds or thousands of taxa (termed 'megaphylogenies' by Smith et al. 2009). This review is aimed at the increasing number of ecologists with some background on phylogenetic inference, but also at other ecologists who are only users of large phylogenies in order to provide them the basic elements to understand the different step and tools to build large phylogenies.

How phylogenetic data has previously been incorporated

As a first step to incorporate an evolutionary perspective into ecological analyses, some studies have used taxonomical classification as a proxy for phylogenies, as implemented recently to unravel the phylogenetic patterns of introduced

Table 1. Summary of the main areas of ecological research where phylogenies have been intensely used and how the integration of the evolutionary perspective has been done. Example studies for each case are reported.

Ecological research area	How evolution has been integrated	Examples	Branch length assignment	Degree of resolution	Studied taxa and area
Community assembly rules	Taxonomic classification	1) Beche and Statzner (2009) 2) Asner and Martin (2011)	–	Genus	1) Stream invertebrates – USA 2) Canopy plants – Amazonian region
	Supertree	1) Ackerly (2004) 2) Vamosi and Vamosi (2007)	1) Equal branch lengths 2) Obtention of a pseudo-chronogram with published node dates and evenly distribution of branch lengths for remaining nodes	Genus	1) Anarcardiaceae, Ericaceae, Rhamnaceae and Rosaceae – California 2) Predaceous diving beetles (Dytiscidae) – Alberta region
	Phylogenetic inference from molecular data	1) Cavender-Bares et al. (2004) 2) Lovette and Hochachka (2006)	Inference from molecular data	Species	1) Oak species (*Quercus*) – Florida 2) Wood-warblers (Parulidae) – North America
Conservation biology	Taxonomic classification	1) Lockwood et al. (2002) 2) Stuart et al. (2004)	–	Order/Family/Genus	1) Vertebrates – USA 2) Amphibians – global
	Supertree	1) McGoogan (2007) 2) Devictor et al. (2010)	1) – 2) All branch lengths set to 1	Species	1) Primates – Africa 2) Birds – France
	Phylogenetic inference from molecular data	1) Forest et al. (2007) 2) Thomas (2008)	Inference from molecular data	Species	1) Angiosperms – Cape 2) Birds – Britain
Invasion biology	Taxonomic classification	1) Cassey et al. (2004) 2) Diez et al. (2008)	–	1) Family/Subfamily/ Tribe/Genus 2) Genus	1) Parrots (Psittaciformes) – global 2) Plants – Auckland
	Supertree	1) Strauss et al. (2006) 2) Whitney et al. (2009)	1) Obtention of a pseudo-chronogram with published node dates and interpolation of all other interior nodes 2) Equal branch lengths	Species	1) Grasses – California 2) Plants – USA, British Isles
	Phylogenetic inference from molecular data	1) Cadotte et al. (2010) 2) Schaefer et al. (2011)	Inference from molecular data	Species	1) Plants – N California 2) Angiosperms – Azores
Niche evolution	Supertree	1) Diniz-Filho et al. (2010) 2) Dorman et al. (2010)	1) Obtention of a pseudo-chronogram with published node dates and interpolation of all other interior nodes	1) Species 2) Species	1) Carnivora – New World 2) Mammals – Europe
	Phylogenetic inference from molecular data	1) Silvertown et al. (2006) 2) Smith and Donoghue (2010)	Inference from molecular data	Species	1) Plants – Britain 2) *Lonicera* genus – global

species naturalisations (Daehler 2001, Diez et al. 2008). Although appealing because it is relatively easy to implement, such an approach is rather unrealistic since it assumes that intrageneric relatedness is equal for all genera. In other studies, phylogenies have been constructed by assembling, grafting or subsetting published phylogenies (i.e. supertree approach; Bininda-Emonds et al. 2002, Buerki et al. 2011). Supertrees usually provide topological information (i.e. evolutionary relationships), but no information about branch lengths (i.e. quantitative estimates of evolutionary relatedness). This lack of branch-length data is usually dealt with by setting all branch lengths equal to one, which for instance will only allow the estimation of phylogenetic diversity metrics based on the number of nodes separating pairs of taxa (Faith 1992). An alternative approach is to estimate branch 'height', which is described as one less than the number of taxa below a given node. Branch lengths are then calculated as the difference between the heights of successive nodes (Grafen 1989, 1992). An improved third approach is the obtention of a pseudo-chronogram using the BLADJ algorithm (Webb et al. 2008), which fixes some nodes for which there are age estimates available in the literature, and then sets all other branch lengths by placing the rest of the nodes evenly between the dated ones. These approaches (especially the first two) imply that evolutionary rates are homogeneous across the tree, whereas it is now well accepted that rates can vary substantially between different lineages (Hughes and Eastwood 2006).

Recent improvements for the obtention of robust and conservative megaphylogenies

Until now, the construction of large phylogenetic trees based on the simultaneous analysis of character data sets concatenating several regions (i.e. supermatrix approach) has mainly been used for systematic studies (McMahon and Sanderson 2006). The main strength of the supermatrix approach is the direct connection between the character data and the final result, in contrast to the supertree approach, where part of the character information is lost when character datasets are summarised as trees (De Queiroz and Gatesy 2006). However, until recently, analyses of large supermatrices were very limited because of the prohibitive time required for tree heuristic searches and the issue of whether large amounts of missing data in supermatrices would bias phylogenetic inference (Wiens 2003).

The recent development of optimised algorithms for maximum likelihood estimation (RAxML, Stamatakis 2006, Stamatakis et al. 2008; GARLI, Zwickl 2006) has now made possible the analyses of extraordinarily large supermatrices of sequence data. In addition, the constant increase of available molecular data in GenBank (in 2010, nucleotide sequences were available for > 380 000 organisms, Benson et al. 2011), combined with the increasing number of ameliorated algorithms for alignment optimisation and depuration (Castresana 2000, Nuin et al. 2006, Capella-Gutierrez et al. 2009) make possible the improvement of pipelines of analyses able to handle extremely large data sets. Finally, the last decade has seen major advances in consensus knowledge of deep phylogenetic relationships for an increasing number of groups of organisms (e.g. plant families, Davies et al. 2004, Schuettpelz and Pryer 2007, Smith et al. 2011). This means that it is now possible, and indeed recommended, that this information is incorporated as a backbone constraint tree to define monophyletic groups, with the aim of speeding up the analyses and reducing the number of possible artefacts due to data patchiness and/or long-branch attraction (Felsenstein 1978).

Taken altogether, these recent advances make it possible to easily infer robust and up-to-date phylogenetic hypotheses for large species samples, by finding a compromise between speed, simplicity and accuracy (see diagram in Fig. 1). Here, we provide an appraisal of the methodological issues involved in the inference of large phylogenies for ecological studies. A detailed step-by-step workflow primarily intended for ecologists can be found in the Supplementary material Appendix 2.

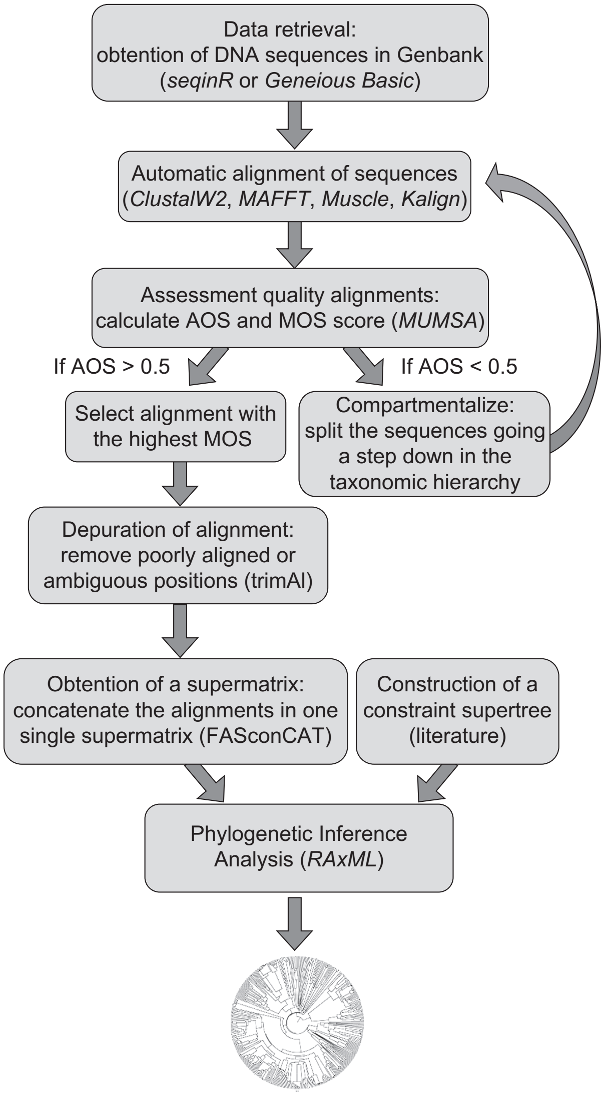
Data retrieval:
obtention of DNA sequences in Genbank
(*seqinR* or *Geneious Basic*)
Automatic alignment of sequences
(*ClustalW2*, *MAFFT*, *Muscle*, *Kalign*)
Assessment quality alignments:
calculate AOS and MOS score (*MUMSA*)
If AOS > 0.5
If AOS < 0.5
Select alignment with the highest MOS
Compartmentalize:
split the sequences going a step down in the taxonomic hierarchy
Depuration of alignment:
remove poorly aligned or ambiguous positions (trimAl)
Obtention of a supermatrix:
concatenate the alignments in one single supermatrix (FASconCAT)
Construction of a constraint supertree (literature)
Phylogenetic Inference
Analysis (*RAxML*)

Figure 1. Diagram showing the iterative steps of the proposed guideline for the inference of phylogenetic trees from molecular data.

Selection and obtention of molecular data

It has been shown that conserved regions (coding loci) are useful in resolving phylogenetic relationships at high taxonomic levels, but usually provide little information at lower ones (Gielly and Taberlet 1994); whereas non-coding regions can resolve, for example, intra-generic relationships (Shaw et al. 2005) but may accumulate too much noise to be aligned consistently in a broad taxonomical group. The combination of several regions with different rates of evolution and a taxonomically clustered alignment for fast-evolving regions (see below) should make it possible to infer species relationships along the terminal branches of the tree. It is usually deemed advisable to survey the literature pertaining to the group of interest as a first step to choosing the regions that may be phylogenetically informative. The 'Phylota browser' (< http://phylota.net/ >, Sanderson et al. 2008) offers a snapshot of the current taxonomic distribution of nucleotide sequences in Genbank, and allows the user to easily download sequences or maximum likelihood trees inferred from these sequences for some particular clades. We have to be aware that, in some cases, unlinked loci (such as regions from different genomes) may result in conflicting phylogenetic signal (Pamilo and Nei 1988); however, this mainly affects groups of closely related species. This phenomenon is usually due to incomplete lineage sorting, i.e. the lack of fixation of gene lineages along a species lineage. This can occur when the ancestor has undergone several speciation events in a short period of time and ancestral polymorphism is not fully resolved when a second speciation event occurs. However, it has been reported that if multiple regions are incorporated, sufficient signal may remain to reconstruct species trees (Maddison and Knowles 2006).

Ecological studies usually focus on an assemblage of organisms (e.g. a community, a species pool for a biogeographic region) and not necessarily on all species of a given clade. Until recently, tools for automatic sequence retrieval were designed mainly for: 1) comparative genomics, e.g. BLAST, which retrieves sequences from Genbank based on similarity (Altschul et al. 1997); 2) systematics, e.g. PowerBLAST, which has the capacity to restrict the search to any level of the NCBI taxonomy index (Zhang and Madden 1997); and 3) comparative biology, for instance the semi-automated pipeline implemented in the PHLAWD package (Smith et al. 2009), which integrates taxonomic hierarchies with iterative alignment procedures to assemble denser data matrices for inferring megaphylogenies. Few recently developed tools suit the particular needs of ecologists. The R environment (R Development Core Team), which is widely used in ecology, comprises the package seqinR that enables the automatic search and retrieval of sequences for a given list of species and a certain region (Charif and Lobry 2007; see the script in Supplementary material Appendix 2). With the use of this package, available sequences can be extracted and deposited in a unique FASTA file (see Supplementary material Appendix 2 for common file formats), together with the accession numbers and a list of missing taxa. It is also possible to limit the number of sequences downloaded for each taxa after some criteria such as sequence length.

Tools for automatic retrieval are highly valuable for large-scale studies, but there are still some difficulties due to poor standardisation of databases (e.g. synonymous wording for gene regions; taxonomic synonyms for controversial taxa) (Bottu 2009). However, in seqinR, the R syntax allows the integration of this uncertainty into sequence retrieval. The Phyutility software (Smith and Dunn 2008) is another alternative to retrieve sequences based on several keywords. A GUI-based solution for sequence retrieval is the free program Geneious Basic (Drummond et al. 2009), in which the search procedure is less automatable but easier to use thanks to its graphical user interface.

When a substantial amount of species included in a study does not occur in GenBank, a genus-level phylogeny may be inferred as a surrogate. In this particular case, an approach that can be considered is to include one species per genus as a representative (Schuettpelz and Pryer 2007), and once the phylogenetic tree has been obtained, to substitute each genus representative by a polytomy of species, which indeed represents unresolved nodes. More sophisticated Monte-Carlo based approaches can also be used to randomly resolve polytomies based on diversification models (Kuhn et al. 2011; see Supplementary material Appendix 2 for an example).

A potential problem that supermatrix analyses (usually with sparse data sets) have to face is rogue taxa, i.e. phylogenetically unstable taxa that can have very divergent placements in a tree set (Wilkinson 1994, Sanderson and Shaffer 2002), leading to lower resolution and support throughout the tree (Smith et al. 2009). One potential source of rogue taxa are taxonomic instability and misspelling errors in Genbank, because it may lead to isolated taxa in the dataset represented by few sequences; thus, once sequences have been obtained, it is important to extract the names of species and check that there are no taxonomic synonyms with an updated reference checklist (Thomson and Shaffer 2010).

Outgroup selection

Phylogenetic inference is sensitive to the outgroup choice (Swofford et al. 1996). The outgroup indicates which nodes in the tree are the oldest, and infers the evolutionary direction of character change in the resultant tree (Maddison et al. 1984). The root of a tree thus represents the common ancestor of all the taxa included in the study. For previously studied groups, the choice of the outgroup can be made on the basis of larger studies in the scientific literature (Sanderson and Shaffer 2002). The outgroup should belong to a clearly distinct lineage with respect to the ingroup sequences, but at the same time it should not be too divergent, if it is to be aligned unambiguously.

Sequence alignments inference

A phylogenetic sequence alignment is a hypothesis about the homology of multiple residues in nucleotide (or protein) sequences. Sequences are usually of different lengths, and gaps (represented as hyphens) are introduced to represent deletions or insertions in the sequences. Since phylogenetic inference relies on the assumption that the characters (data matrix) are homologous, i.e. have the same evolutionary origin, it is crucial to find an alignment that is as accurate as possible, as it will have a significant effect on the quality of the inferred phylogeny (Kress et al. 2009). Sequence alignment can be achieved automatically, but final checking by visual examination is recommended (Morrison 2009). Nowadays there are a large number of automated alignment programs in existence; a comparative analysis of the nine most frequently used (Nuin et al. 2006) indicated that the iterative approach available in MAFFT (Katoh et al. 2005) was the fastest and most accurate, although other algorithms also showed very good results for particular evolutionary scenarios. A recent and promising alignment approach, but more time-consuming, is the phylogeny-aware algorithm implemented in PRANK (Löytynoja and Goldman 2008). This method is able to recognize insertions and deletions as distinct evolutionary events based on previously computed evolutionary guide tree, thus avoiding over-estimation of deletion events.

To obtain an accurate alignment for each region, we suggest that a suitable strategy is to perform alignments using the best performing programs (high accuracy in a reasonable time) according to Nuin et al. (2006), for instance: MAFFT (Katoh et al. 2005), MUSCLE (Edgar 2004), Clustal (Thompson et al. 1994, Larkin et al. 2007) and Kalign (Lassmann and Sonnhammer 2005); all available on free servers (Table 2). As an alternative, the recent software called PRANK (Löytynoja and Goldman 2010) is a very valuable option in the case of relatively small dataset. Once the alignments are obtained, the best alignment for each data partition can be determined using the multiple overlap score provided by MUMSA, which compares and measures the reliability of each alignment based on the principle that pairs of aligned positions that are found in many alignments are more reliable, thus, the alignment with the highest number of these pairs is considered as the most correct one (Lassmann and Sonnhammer 2005).

Inclusion of fast-evolving regions with taxonomic clustering

The reconstruction of a phylogenetic tree for a large taxa sample requires the combining of conserved loci as well as fast-evolving regions to resolve deep and shallow nodes. However, non-conserved regions will probably not be alignable over all taxa where the species sample spans a wide taxonomical spectrum. In this particular case, and to take advantage of information from this type of region, one possible approach is to cluster the alignments taxonomically, and then combine global and taxonomically local alignments for conserved and fast-evolving regions, respectively. Note that each cluster of aligned sequences should overlap with one or more clusters by at least 3 taxa (Thomson and Shaffer 2010).

Algorithms such as those implemented in MUMSA (Lassmann and Sonnhammer 2006) allow checking whether the cumulated sequences for a region are saturated or not (i.e. too divergent to be aligned consistently). MUMSA compares several alignments and computes an average overlap score (AOS) based on the consistency between those alignments. Then, the AOS gives an indication on the degree of divergence between sequences ($AOS < 0.5$ indicates that sequences are very difficult to align and thus it is probable that many positions are saturated by multiple substitutions, Lassmann and Sonnhammer 2006). Where too much divergence is detected, sequences might be split by moving a step down in the taxonomic hierarchy; for instance, if a data partition contains sequences for one given taxonomic order leading to $AOS < 0.5$, then the possibility of splitting the partition into smaller partitions at the family level should be considered. The PHLAWD package also implements an index to detect saturation across a set of sequences, named 'median absolute deviation', which compares uncorrected genetic distances to corrected distances according to a Jukes–Cantor model of molecular substitution (Smith et al. 2009). If alignments appear to be saturated, the alignments are broken up using taxonomic classifications as guides, and separate alignments are carried out for the individual groups delimited in this way.

A promising approach to deal with divergent sequences is the simultaneous inference of sequence alignment and phylogenetic trees implemented in the software SATé (Liu et al. 2009, 2012). This iterative method is based on a divide-and-conquer realignment approach; starting with a existing alignment, it divides the set of sequences into subsets, each subset is re-aligned (with MAFFT, Katoh et al. 2005), and then the alignments are merged together into an alignment on the full set of sequences. However, this algorithm has not been intended for multi-locus data, but still it may be useful to obtain separately more accurate alignments for each region.

Clustering variable regions allows us to maximise the representation of taxa in the final supermatrix, but will also lead to a patchy matrix with a considerable amount of missing data. The impact of missing data on accuracy has been widely studied but no clear consensus has yet emerged. Some empirical and simulation studies are optimistic and argue that the inclusion of taxa even with incomplete sequence data often has positive effects (Driskell et al. 2004; for a revision, see De Queiroz and Gatesy 2006). On the other hand, Sanderson et al. (2010) show that usual coverage for real datasets do not allow to accurately resolve all the nodes of a particular tree, but still it is often possible to distinguish a large fraction of edges in the tree. Another study (Sanderson et al. 2011) shows that, depending on its distribution, missing data can produce a phylogenetic landscape of large sets of different trees with identical optimality scores (called 'terraces'). These terraces can result in heuristic search algorithms requiring an unnecessarily long time to compare trees of one same terrace; further improvement of phylogenetic software should be sought in the future to increase the efficiency of heuristic searches.

Shorter is better: improved alignment with automatic trimming

DNA regions usually do not evolve homogeneously (Whelan 2008). It is common that some parts of an

Table 2. Comparison of different methods discussed in the text for each step to infer dated phylogenies.

Step	Method	Software	Description	References	URL program & webservers
Alignment	Progressive alignment	Clustal W and Clustal X	Most widely used multiple alignment tool. Clustal X is a windows interface program, Clustal W a command-line one. Both are based on the same algorithms.	Larkin et al. (2007)	Download page: <ftp://ftp.ebi.ac.uk/pub/software/clustalw2/> Web server: <www.ebi.ac.uk/Tools/msa/clustalw2/>
	Progressive alignment	Kalign	Uses the Wu–Manber string matching algorithm, which improves accuracy and speed of alignment, specially with distant sequences.	Lassmann and Sonnhammer (2005)	Download and web server page: <http://msa.sbc.su.se/cgi-bin/msa.cgi> Web server: <www.ebi.ac.uk/Tools/msa/kalign/>
	Progressive alignment with iterative refinement	MAFFT	Offers a range of different methods depending on the number and type of sequences (e.g. FFT-NS-2, fast distance calculation recommended for >2000 sequences).	Katoh et al. (2005)	Download page: <http://mafft.cbrc.jp/alignment/software/> Web servers: <http://mafft.cbrc.jp/alignment/server/> <www.ebi.ac.uk/Tools/msa/mafft/>
	Progressive alignment with iterative refinement	MUSCLE	Does progressive alignment using a log-expectation score, and refinement using tree-dependent restricted partitioning.	Edgar (2004)	Download page: <www.drive5.com/muscle/downloads.htm> Web server: <www.ebi.ac.uk/Tools/msa/muscle/>
Alignment quality assessment	Calculation of overlap scores between alignments	MUMSA	Computes an average overlap score (AOS), reflecting the difficulty of aligning the sequences, and a multiple overlap score (MOS) indicating the quality of each alignment. Both scores range between 1 and 0. An AOS score <0.5 indicates sequences are very difficult to align and thus the alignments may be unreliable.	Lassmann and Sonnhammer (2006)	Download and web server page: <http://msa.sbc.su.se/cgi-bin/msa.cgi>
Alignment trimming	Trimming of alignment segments with too many variable positions or gaps	Gblocks	Several parameters can be modified to make the selection of blocks more or less stringent.	Castresana (2000)	Download page: <http://molevol.cmima.csic.es/castresana/Gblocks.html> Web server: <http://molevol.cmima.csic.es/castresana/Gblocks_server.html>
	Trimming of alignment segments with too many variable positions or gaps	trimAl	Adjusts automatically the parameters to optimize the phylogenetic signal-to-noise ratios, making it especially suited for large-scale analyses.	Capella-Gutierrez et al. (2009)	Download page: <http://trimal.cgenomics.org/downloads>
Phylogenetic inference	Maximum likelihood	RAxML	Fast and accurate program to infer phylogenetic trees, can handle huge number of sequences. It implements bifurcating and multifurcating constraint trees. The default model of nucleotide substitution is the general time reversible (GTR).	Stamatakis (2006), Stamatakis et al. (2008)	Download page: <wwwkramer.in.tum.de/exelixis/software.html> 'Black-box' web server: <http://phylobench.vital-it.ch/raxml-bb/> Advanced web server: <www.phylo.org/portal2/>
	Maximum likelihood	GARLI	Fast and accurate program to infer phylogenetic trees, can handle huge number of sequences. It implements bifurcating and multifurcating constraint trees. The assumed model of nucleotide substitution is GTR.	Zwickl (2006)	Download page: <www.nescent.org/wg_garli/Main_Page> Web server: <www.phylo.org/portal2/>
	Bayesian inference	MrBayes	Uses the Markov chain Monte Carlo (MCMC) simulation technique to approximate the posterior probabilities of trees. Backbone tree constraints cannot be provided, however monophyly constraints can be applied defining groups.	Huelsenbeck and Ronquist (2001)	Download page: <http://mrbayes.csit.fsu.edu/download.php> Web server: <www.phylo.org/portal2/>

	Bayesian inference	BEAST	Similar to MrBayes; in addition chronograms (dated phylogenies) can be obtained.	Drummond and Rambaut (2007)	Download page: <http://beast.bio.ed.ac.uk/Main_Page>
Molecular dating with relaxed molecular clock	Penalized-likelihood	r8s	Estimates a chronogram from a given topology with branch lenghts. It is based on the model of rate autocorrelation. Able to handle huge trees.	Sanderson (2003)	Download page: <http://loco.biosci.arizona.edu/r8s/>
		Multidivtime	Estimates a chronogram from sequence data and a given topology. Provides directly confidence intervals for the node ages. It is based on the model of rate autocorrelation. May be unfeasable for huge trees.	Kishino et al. (2001), Thorne and Kishino (2002)	Multidivtime download page: <http://statgen.ncsu.edu/thorne/multidivtime.html> PAML package download page: <http://abacus.gene.ucl.ac.uk/software/paml.html> Step-by-step manual: <www.plant.ch> R implementation (package Lagopus): <www.christophheibl.de/mdt/mdtinr.html>
		BEAST	Estimates a chronogram from sequence data, although a topology can be provided. Various models of rate evolution can be applied. May be unfeasable for huge trees.	Drummond and Rambaut (2007)	Download page: <http://beast.bio.ed.ac.uk/Main_Page>

alignment are well conserved and informative, whereas others are so divergent that the homologous position of gaps cannot be determined. Because the quality of the alignment greatly influences the resulting phylogeny (Morrison and Ellis 1997), it has been recommended that these poorly aligned blocks should be removed (Swofford et al. 1996).

Recent years have seen the development of computerised methods to improve the alignments by trimming them, selecting blocks of conserved regions and removing poorly aligned or ambiguous positions (Gblocks, Castresana 2000; trimAl, Capella-Gutierrez et al. 2009). Simulation studies have shown that trimmed alignments always produce phylogenetic trees that are more accurate (i.e. with better topologies) than, or at least equal to, trees derived from complete alignment (Talavera and Castresana 2007, Capella-Gutierrez et al. 2009). To date, trimAl is probably the best program for analyzing large character data, as it has the possibility to automatically adjust the parameters to improve the phylogenetic signal-to-noise ratio (Capella-Gutierrez et al. 2009). Once the trimmed alignments for all the regions have been obtained, they can be concatenated into a single supermatrix (e.g. using FASconCAT, Kück and Meusemann 2010; or with the R package phylotools, Zhang et al. 2010). The Phyutility software (Smith and Dunn 2008) also allows different sequence manipulation (e.g. concatenating, trimming, fetching).

Phylogenetic inference

The variety of methods available for phylogenetic inference can be intimidating for non-phylogeneticists. There are three groups of methods based on different optimisation criteria: a) distance-matrix methods which convert the differences between sequences into a distance matrix (Saitou and Nei 1987), and are therefore fast but too simplistic (Holder and Lewis 2003); b) maximum parsimony, which is based on the assumption that the most likely tree is the one that minimises the number of mutations to explain the data, thus considering only a minimum evolution scenario (Felsenstein 1978, Edwards 1996); and c) probabilistic methods (Huelsenbeck and Crandall 1997, Huelsenbeck et al. 2001).

The final types of methods, i.e. maximum likelihood (ML) and Bayesian inference (BI), are currently the most used as they have the potential to rigorously explore the landscape of different possible trees; they are also quite accurate for highly divergent sequences and they can account for different models of sequence evolution (Hall 2011). The available models of evolution describe the different probabilities of change from one nucleotide to another along a phylogenetic tree: e.g. the simplest model (JC; Jukes and Cantor 1969) assumes both equal transition rates between all types of nucleotides and equal equilibrium frequencies; whereas the more complex one, the generalised time reversible model (GTR; Tavaré 1986) considers six different transition rates, one for each possible change between different nucleotides, and four nucleotide frequencies. Two additional parameters can be added to these models: a gamma distributed rate heterogeneity and an estimated proportion of invariable sites (+Γ, +I). The GTR model usually is the best fitting-model for real-world data (Stamatakis 2006). Software such as MrModeltest (Nylander 2004) implements hierarchical likelihood ratio tests to select the best-fitting model.

The ML method looks for the tree that, under a specified model of evolution, maximises the likelihood of observing the data (Felsenstein 1981). Until recently, ML could be prohibitive in terms of computational time for large size data; however, programs such as RAxML and GARLI implement optimised search algorithms that allow data with thousands of taxa to be analysed within a reasonable time (Dunn et al. 2008, Hackett et al. 2008, Yarza et al. 2008). Moreover, with both programs it is possible to apply a mixed supermatrix-supertree approach as they allow the use of a non-comprehensive constraint tree.

Bayesian inference was introduced in phylogenetics in the late 1990s (Rannala and Yang 1996). The user provides an alignment and a model of evolution and the program samples the trees with the highest likelihood given these data (Huelsenbeck et al. 2001). It differs from ML in that instead of seeking the tree that maximises the likelihood of observing the data, BI seeks the posterior distribution of trees using Markov chain Monte Carlo (MCMC) methods (Tierney 1994). The MCMC sampling involves 'travelling across parameter space' to produce a set of trees repeatedly visited, with the frequency at which trees are sampled estimating their likelihood (Huelsenbeck et al. 2001). One important issue, if we are to ensure an adequate search of tree space by BI, is to check that the runs have reached convergence, i.e. that the tree topologies obtained are a set of statistically similar trees, sampled in proportion to their true posterior probability distribution (Nylander et al. 2008). Usually a few millions of generations are required to achieve this convergence. Bayesian inference became rapidly popular because it was less demanding computationally than ML, and until the apparition of more efficient ML programs (e.g. RAxML; GARLI), BI programs such as MrBayes (Ronquist and Huelsenbeck 2003) were the only real alternative to parsimony analyses for large datasets (> 100 sequences). However, running BI over thousands of species is at the moment not likely to provide reliable results. For instance, Hackett et al. (2008) reported that reaching convergence was impossible with BI for their dataset of 32 kb for 169 taxa. By now, ML methods are preferable for inferring large phylogenies thanks to recent improvements of the RAxML algorithm for supercomputing environments scaling from hundreds to thousands of cores (Stamatakis et al. 2012).

Constraining heuristic searches by a conservative backbone tree

Some programs of phylogenetic inference allow a non-fully resolved constraint tree to be provided as additional input (e.g. RAxML) or monophyletic groups to be defined (e.g. BEAST, Drummond and Rambaut 2007). Concretely, a constraint tree is a user-defined tree that limits the search space to those trees that are compatible with the constraint tree (Stamatakis 2008). The use of a constraint tree based on well-established relationships among major groups is an interesting means to reduce possible artefacts due to patchy data and speed up the analyses. It is also a way to integrate more evidence in the final tree than the phylogenetic information retained only by genic partitions. There are many cases where a supertree or a phylogenetic tree has been published, establishing the relationships between the main groups, for instance angiosperm and fern families (Davies et al. 2004, Schuettpelz and Pryer 2007, Smith et al. 2011), amphibian genera (Roelants et al. 2007), and bird orders (Hackett et al. 2008). From this type of sources, one can obtain a tree with the taxa of interest (Supplementary material Appendix 2), where only some nodes (i.e. consensus relationships) are resolved, to use it as a backbone to retain resolved well-established, generally deeper nodes. The resulting tree will be part of the input provided to conduct the phylogenetic inference, thus implementing a mixed approach both using supertree and supermatrix approaches.

Clade support assessment

Except for Bayesian inference, phylogenetic inference methods produce only point estimates of the phylogeny. The technique most often used to evaluate the reliability of specific clades in the tree is the bootstrap (BS) analysis (Felsenstein 1985a), which involves resampling characters from the original dataset. More specifically, characters are randomly sampled with replacement to produce a matrix with the same number of taxa and characters as the original one, and a phylogenetic analysis is performed again for a specified number of replications. The BS values show the percentage of times that a clade appears on these replicates. The number of BS replicates needed for accurate estimation of clade support is highly-dataset dependent (typically between 500 and 2000 are conducted), but the highly performant RAxML software has recently incorporated a 'bootstopping' criterion that allows the computation of only the necessary number of replicates to obtain sufficient accuracy (Pattengale et al. 2010). The statistical interpretation of BS is not clear, but it is considered a conservative measure, and thus clades with values above 70% are generally considered as substantially reliable, and above 95% as highly supported (Felsenstein 1985a, Hillis and Bull 1993). Again, conducting BS analyses with a backbone constraint tree will probably help to speed them up and obtain higher supports.

In Bayesian analyses, the frequency of a given clade in a set of trees is the posterior probability of that clade and it estimates the support of this given clade; no bootstrapping is therefore needed to assess the confidence of the estimated topology (Alfaro et al. 2002). The output is a support value for each clade on the final consensus tree, which is termed the posterior probability. Bayesian inference has received some criticisms of putative overcredibility of node supports (Suzuki et al. 2002), mainly associated with inappropriate model choice (Erixon et al. 2003) and failure to allow convergence (Nylander et al. 2008). It is generally considered that only clades with posterior probabilities equal to or higher than 0.95 should be considered as reliable (Huelsenbeck and Rannala 2004).

Finally, rogue taxa is certainly a common issue for the reconstruction of megaphylogenies. Rogue taxa may lead to low statistical support for certain clades, and are usually attributed to ambiguous or insufficient phylogenetic signal in the character data (Sanderson and Shaffer 2002). While there is no optimal solution for dealing with rogue taxa, the most used approach to identify and prune them consists in computing node distances, such as the taxonomic instability index, implemented in Mesquite (Maddison and Maddison 2007). This index measures the variation of pairwise distances between taxon pairs across all bootstrap trees. However, this approach has to rely on an arbitrary threshold to tease apart rogue taxa. A new and promising approach (still not implemented in phylogenetic software) is the one based on algorithms that test for the improvement in consensus trees by pruning one taxon at each time (Pattengale et al. 2010, Aberer et al. 2011).

Accounting and integrating methodological uncertainties

Depending on the data availability in molecular databases, one may infer an incompletely resolved phylogeny, e.g. a genus-level phylogeny where the tips are replaced by a species polytomy (Supplementary material Appendix 2). However, for many analyses it is necessary to have a dichotomic tree, for instance, to identify changes in diversification rates with LASER (Rabosky 2006).

Polytomies can be resolved following different approaches, and the uncertainty associated with polytomies can be taken into account when conducting subsequent analyses with a set of resulting dichotomic trees. The simplest way to work around polytomies is to resolve them randomly with zero-length branches (Felsenstein 1985b), e.g. with the multi2di function in R. However, different methods of resolving polytomies have been described, assigning branch lengths in different ways of increasing complexity: a) to distribute branch lengths evenly (Webb et al. 2008); b) to assign random branch lengths (Thuiller et al. 2011); and c) to apply a specific diversification model (Kuhn et al. 2011; see also the R script stickTips in the Supplementary material Appendix 2). The last option can be applied by permuting the unresolved portions of the tree with a Bayesian MCMC search algorithm based on a diversification model, thus obtaining a pseudo-posterior distribution of completely resolved trees, to which analyses can then be applied (Kuhn et al. 2011).

Dating phylogenies with relaxed molecular clocks

For some eco-evolutionary analyses such as estimating phylogenetic signal in trait data or, more generally, comparative analyses, ultrametric trees are necessary, i.e. a tree with root-to-tip path lengths for all lineages equal to those built under the assumption of a molecular clock. The molecular clock hypothesis was first proposed by Zuckerkandl and Pauling (1965), which postulated that the amount of difference between DNA or protein sequences between two species is proportional to the time elapsed since their evolutionary separation. However, it became rapidly evident that a strict clock model does not fit the studied data in most cases, with pervasive variation in rates of nucleotide substitution (Britten 1986, Li 1997), even in closely related species (Thomas et al. 2006). Several factors can influence the rate of molecular evolution, such as generation time (Bromham et al. 1996, Ohta 2002); metabolic rate (Martin and Palumbi 1993, Gillooly et al. 2005); reproductive mode (Paland and Lynch 2006, Johnson and Howard 2007); or efficiency of DNA repair machinery (Ota and Penny 2003).

In recent decades, molecular dating has become a rapidly developing field, and several methods that incorporate rate heterogeneity have been developed, including nonparametric (Sanderson 1997, Britton et al. 2007) and semiparametric approaches (Sanderson 2002), local clocks (Yoder and Yang 2000), and Bayesian parametric models (Thorne and Kishino 2002, Drummond et al. 2006). Most of them implement models that assume rate autocorrelation among lineages, which means that rates of substitution are likely to be more similar among closely related lineages than in distant ones (Gillespie 1991). The main exceptions are the models described by Drummond et al. (2006) and implemented in the 'BEAST' software (Drummond and Rambaut 2007), which samples rates from a distribution. Non-autocorrelated models suit very fast evolving sequences, such as viruses (Drummond et al. 2006), whereas a comparative analysis of three real-world data sets showed that autocorrelation models fit better (Lepage et al. 2007). We briefly present below the most commonly used methods in the literature (for a wider review see Rutschmann 2006).

Penalized likelihood (PL, Sanderson 2002), implemented in r8s (Sanderson 2003), combines a parametric model with different substitution rates with a nonparametric roughness penalty which costs the model if rates change too quickly from branch to branch. The relative contribution of the two components is determined by a smoothing parameter, which is estimated by a cross-validation (CV) procedure. The CV is computationally intensive for very large trees, because it consists in sequentially removing each terminal branch to estimate the parameters of the model without that branch for a given smoothing parameter, and compare it to the original estimates (Sanderson 2002). A high smoothing value leads to a clock-like model, whereas a low value permits much more rate variation. Once the optimal smoothing value has been determined, the estimation of divergence times is relatively fast. The user provides only a fixed tree topology with branch lengths, which needs to be dichotomous, together with one or several age constraints. Thus, in order to take in account phylogenetic uncertainty, it is better to run the analyses using several trees with a high likelihood or posterior probability (see below).

The bayesian implementation of rate autocorrelation in Multidivtime (Thorne et al. 1998, Kishino et al. 2001, Thorne and Kishino 2002) uses a parametric model to describe the change rate over time with a MCMC procedure to derive the posterior distribution of rates and times. A detailed step-by-step manual (Rutschmann 2005) describes the complete procedure. In contrast to PL, this method is able to account for polytomies (it can therefore be run on a consensus tree), model parameters can be inferred for different regions if specified, and it provides directly estimated ages for each node with a 95% credibility interval; but it is computationally intensive for large data and not feasible for huge datasets.

The bayesian implementation in BEAST (Drummond and Rambaut 2007) directly calculates ultrametric phylogenies based only on sequence data and model parameters, thus it computes at the same time the topology and branch lengths as a function of time. As in Multidivtime, BEAST can analyse multiple data partitions with different substitution models, and provides Bayesian credibility intervals. The main interests of this software are: 1) the correlation of rates between adjacent branches can be tested; 2) it implements several models for molecular sequence variation; and 3) different calibration distributions can be provided (normal, lognormal, exponential or gamma) to model calibration uncertainty instead of simple point estimates or age intervals. However, it is computationally intensive with very high numbers of taxa.

The methods detailed above will probably have serious difficulties (or even won't be able) to cope with huge trees (i.e. with thousands of leaves). By now, the only method that is very fast even with enormous trees is the nonparametric method implemented in PATHd8, which estimates node ages by mean path lengths (i.e. the average of all path lengths from a node to its leaves) while smoothing substitution rates locally when deviations from the global clock are detected by calibrated nodes (Britton et al. 2007).

Sources and uses of calibration information

To obtain absolute time divergence estimates using relaxed-clock approaches, it is necessary to provide one or several age estimates for specific nodes as calibrations. Calibration is critical and highly influential on molecular dating inference (Sauquet et al. 2012). Different types of calibrations have been used to date: fossils, geological events, palaeoclimatic data, and secondary estimations (i.e. estimates from independent molecular dating studies). There is general agreement that the best type of data to calibrate is the fossil record (Magallón 2004), even though it can be subject to sources of errors such as erroneous placement on the phylogeny. A fossil is assigned to a group of extant taxa based on one or more synapomorphies, but it can be assigned to the crown or the stem node of a group. The crown node comprises all extant taxa and their most recent common ancestor, whereas the stem node represents the crown group plus the extinct taxa that diverged after the splitting of the crown group from its closest living relative. Because the fossil record is usually incomplete, the most conservative option is to assign fossils on the stem group node and as a minimum constraint age (Benton and Ayala 2003). When possible, multiple fossils should be used instead of a single one (Benton and Donoghue 2007), and ideally, these fossil constraints should be spread across the tree.

Geological events and palaeoclimatic data (Baldwin and Sanderson 1998) have been used as calibration data assuming that vicariance has led to divergence at a certain node. For this, timings of geological events such as continental splits or the rise of mountain chains are often reported as unique values, but most of these phenomena take place over several millions of years (Garzione et al. 2008). Furthermore, the use of geological or palaeoclimatic data implies that organisms are not available to cross the new barriers through dispersal, whereas in many studies long distance dispersal has been reported (De Queiroz 2005). Oceanic islands have also been used to apply a maximum age constraint on the divergence between endemic species and continental relatives; however, this assumes that the endemic species is a neoendemism and not a relict; moreover, present-day oceanic islands are in some cases only the most recent element of a series of oceanic islands (Christie et al. 1992). Because of these pitfalls, this type of data should be avoided as a source of calibration (Forest 2009, Kodandaramaiah 2011).

Secondary estimations are used usually when the fossil record is nonexistant, but one has to be aware that sources of error generated in the first study remain, and are likely to be multiplied in subsequent analyses. This type of data should be use with care, e.g. using confidence intervals as minimum and maximum values on a given node, otherwise estimates will be of little scientific value (Forest 2009).

Conclusions

Recent years have seen an increasing interest in the testing of ecological hypotheses within a phylogenetic context; e.g. in the study of ecosystem processes (Edwards et al. 2007, Cadotte et al. 2008); in the field of invasion biology (Strauss et al. 2006, Thuiller et al. 2010); conservation biology (Forest et al. 2007, Isaac et al. 2007); and ecophysiology (Moles et al. 2005, Wright et al. 2007). However, although phylogenetic knowledge has been shown to be useful in ecological studies, phylogenies used are often inadequate or too simplistic (e.g. lack of branch-length data). It has thus become a challenge to infer reliable and robust mega-phylogenies, for which necessary data and tools are becoming increasingly available. One issue emerging in parallel will soon be how to visualize such large scale phylogenies (for method development in this direction, see Page 2012). Growing molecular databases such as Genbank, increased consensus of deep phylogenetic relationships and recent improvements in software (e.g. RAxML; GARLI) able to handle huge analyses make it possible to obtain a more solid evolutionary hypothesis with which to work, within a moderate amount of time and using a personal computer.

Acknowledgements – The research leading to these results had received funding from the European Research Council under the European Community's Seven Framework Programme FP7/2007-2013 Grant Agreement no. 281422. The initial discussion around the issue was developed under the ANR DIVERSITALP (ANR-07-BDIV-014) and EVORANGE (ANR-09-PEXT-011) projects. CR was partially supported by a post-doctoral grant from the Fundación Ramón Areces (Spain).

References

Aberer, A. J. et al. 2011. RogueNaRok: an efficient and exact algorithm for rogue taxon identification. – Exelixis-RRDR-2011-10, Heidelberg Inst. for Theoretical Studies, < http://sco.h-its.org/exelixis/publications.html >.

Ackerly, D. D. 2004. Adaptation, niche conservatism, and convergence: comparative studies of leaf evolution in the California chaparral. – Am. Nat. 163: 654–671.

Alfaro, M. E. et al. 2002. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. – Mol. Biol. Ecol. 20: 255–266.

Altschul, S. F. et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. – Nucl. Acids Res. 25: 3389–3402.

Asner, G. P. and Martin, R. E. 2011. Canopy phylogenetic, chemical and spectra1 assembly in a lowland Amazonian forest. – New Phytol. 189: 999–1012.

Baldwin, B. G. and Sanderson, M. J. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). – Proc. Natl Acad. Sci. USA 95: 9402–9406.

Bêche, L. A. and Statzner, B. 2009. Richness gradients of stream invertebrates across the USA: taxonomy- and trait-based approaches. – Biodivers. Conserv. 18: 3909–3930.

Benson, D. A. et al. 2011. Genbank. – Nucl. Acids Res. 39: D32–D37.

Benton, M. J. and Ayala, F. J. 2003. Dating the tree of life. – Science 300: 1698–1700.

Benton, M. J. and Donoghue, P. C. J. 2007. Paleontological evidence to date the tree of life. – Mol. Biol. Ecol. 24: 26–53.

Bininda-Emonds, O. R. P. et al. 2002. The (super)tree of life: procedures, problems, and prospects. – Annu. Rev. Ecol. Syst. 33: 265–289.

Bottu, G. 2009. Sequence databases and database searching – theory. – In: Lemey, P. et al. (eds), The phylogenetic handbook. Cambridge Univ. Press, pp. 33–54.

Britten, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. – Science 231: 1393–1398.

Britton, T. et al. 2007. Estimating divergence times in large phylogenetic trees. – Syst. Biol. 56: 741–752.

Bromham, L. et al. 1996. Determinants of rate variation in mammalian DNA sequence evolution. – J. Mol. Evol. 43: 610–621.

Buerki, S. et al. 2011. Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study. – Syst. Biol. 60: 32–44.

Cadotte, M. W. et al. 2008. Evolutionary history and the effect of biodiversity on plant productivity. – Proc. Natl Acad. Sci. USA 105: 17012–17017.

Cadotte, M. W. et al. 2010. Phylogenetic patterns differ for native and exotic plant communities across a richness gradient in northern California. – Divers. Distrib. 16: 892–901.

Capella-Gutierrez, S. et al. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. – Bioinformatics 25: 1972–1973.

Cassey, P. et al. 2004. Influences on the transport and establishment of exotic bird species: an analysis of the parrots (Psittaciformes) of the world. – Global Change Biol. 10: 417–426.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. – Mol. Biol. Ecol. 17: 540–552.

Cavender-Bares, J. et al. 2004. Phylogenetic overdispersion in Floridian oak communities. – Am. Nat. 163: 823–843.

Cavender-Bares, J. et al. 2009. The merging of community ecology and phylogenetic biology. – Ecol. Lett. 12: 693–715.

Charif, D. and Lobry, J. R. 2007. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. – In: Bastolla, U. et al. (eds), Structural approaches to sequence evolution: molecules, networks, populations. Springer, pp. 207–232.

Christie, D. M. et al. 1992. Drowned islands downstream from the Galapagos hotspot imply extended speciation times. – Nature 355: 246–248.

Daehler, C. C. 2001. Darwin's naturalization hypothesis revisited. – Am. Nat. 158: 324–330.

Davies, T. J. et al. 2004. Darwin's abominable mystery: insights from a supertree of the angiosperms. – Proc. Natl Acad. Sci. USA 101: 1904–1909.

Davies, T. J. et al. 2008. Phylogenetic trees and the future of mammalian biodiversity. – Proc. Natl Acad. Sci. USA 105: 11556–11563.

De Queiroz, A. 2005. The resurrection of oceanic dispersal in historical biogeography. – Trends Ecol. Evol. 20: 68–73.

De Queiroz, A. and Gatesy, J. 2006. The supermatrix approach to systematics. – Trends Ecol. Evol. 22: 34–41.

Devictor, V. et al. 2010. Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. – Ecol. Lett. 13: 1030–1040.

Diez, J. M. et al. 2008. Darwin's naturalization conundrum: dissecting taxonomic patterns of species invasions. – Ecol. Lett. 11: 674–681.

Diniz-Filho, J. A. F. et al. 2010. Hidden patterns of phylogenetic non-stationarity overwhelm comparative analyses of niche conservatism and divergence. – Global Ecol. Biogeogr. 19: 916–926.

Dormann, C. F. et al. 2010. Evolution of climate niches in European mammals? – Biol. Lett. 6: 229–232.

Driskell, A. C. et al. 2004. Prospects for building the tree of life from large sequence databases. – Science 306: 1172–1174.

Drummond, A. J. and Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. – BMC Evol. Biol. 7: 214.

Drummond, A. J. et al. 2006. Relaxed phylogenetics and dating with confidence. – PLoS Biol. 4: 699–710.

Drummond, A. J. et al. 2009. Geneious v5.1. – < www.geneious.com/ >.

Dunn, C. W. et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. – Nature 452: 745–749.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. – Nucl. Acids Res. 32: 1792–1797.

Edwards, A. W. F. 1996. The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. – Syst. Biol. 45: 79–91.

Edwards, E. J. et al. 2007. The relevance of phylogeny to studies of global change. – Trends Ecol. Evol. 22: 243–249.

Erixon, P. et al. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. – Syst. Biol. 52: 665–673.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. – Biol. Conserv. 61: 1–10.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. – Syst. Zool. 27: 401–410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. – J. Mol. Evol. 17: 368–376.

Felsenstein, J. 1985a. Confidence limits on phylogenies: an approach using the bootstrap. – Evolution 39: 783–791.

Felsenstein, J. 1985b. Phylogenies and the comparative method. – Am. Nat. 125: 1–15.

Forest, F. 2009. Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. – Ann. Bot. 104: 789–794.

Forest, F. et al. 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. – Nature 445: 757–760.

Garzione, C. N. et al. 2008. Rise of the Andes. – Science 320: 1304–1307.

Gielly, L. and Taberlet, P. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. – Mol. Biol. Ecol. 11: 769–777.

Gillespie, J. H. 1991. The causes of molecular evolution. – Oxford Univ. Press.

Gillooly, J. F. et al. 2005. The rate of DNA evolution: effects of body size and temperature on the molecular clock. – Proc. Natl Acad. Sci. USA 102: 140–145.

Grafen, A. 1989. The phylogenetic regression. – Phil. Trans. R. Soc. B 326: 119–157.

Grafen, A. 1992. The uniqueness of the phylogenetic regression. – J. Theor. Biol. 156: 405–423.

Hackett, S. J. et al. 2008. A phylogenomic study of birds reveals their evolutionary history. – Science 320: 1763–1768.

Hall, B. 2011. Phylogenetic trees made easy: a how-to manual, 4th ed. – Sinauer.

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 Crandall, K. A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. – Annu. Rev. Ecol. Syst. 28: 437–466.

Huelsenbeck, J. P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogeny. – Bioinformatics 17: 754–755.

Huelsenbeck, J. P. and Rannala, B. 2004. Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. – Syst. Biol. 5: 904–913.

Huelsenbeck, J. P. et al. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. – Science 294: 2310–2314.

Hughes, C. and Eastwood, R. 2006. Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. – Proc. Natl Acad. Sci. USA 103: 10334–10339.

Isaac, N. J. et al. 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. – PLoS One 2: 296.

Johnson, M. T. J. and Stinchcombe, J. R. 2007. An emerging synthesis between community ecology and evolutionary biology. – Trends Ecol. Evol. 22: 250–257.

Johnson, S. G. and Howard, R. S. 2007. Contrasting patterns of synonymous and nonsynonymous sequence evolution in asexual and sexual freshwater snail lineages. – Evolution 61: 2728–2735.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. – In: Munro, H. N. (ed.), Mammalian protein metabolism. Academic Press, pp. 21–123.

Katoh, K. et al. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. – Nucl. Acids Res. 33: 511–518.

Kishino, H. et al. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. – Mol. Biol. Evol. 18: 352–361.

Kodandaramaiah, U. 2011. Tectonic calibrations in molecular dating. – Curr. Zool. 57: 116–124.

Kress, W. J. et al. 2009. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. – Proc. Natl Acad. Sci. USA 106: 18621–18626.

Kück, P. and Meusemann, K. 2010. FASconCAT: convenient handling of data matrices. – Mol. Phylogenet. Evol. 56: 1115–1118.

Kuhn, T. S. et al. 2011. A simple polytomy resolver for dated phylogenies. – Methods Ecol. Evol. 2: 427–436.

Larkin, M. A. et al. 2007. Clustal W and Clustal X version 2.0. – Bioinformatics 23: 2947–2948.

Lassmann, T. and Sonnhammer, E. L. 2005. Kalign – an accurate and fast multiple sequence alignment algorithm. – BMC Bioinform. 6: 298.

Lassmann, T. and Sonnhammer, E. L. 2006. Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. – Nucl. Acids Res. 34: 596–599.

Lavergne, S. et al. 2010. Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. – Annu. Rev. Ecol. Evol. Syst. 41: 321–350.

Lepage, T. et al. 2007. A general comparison of relaxed molecular clock models. – Mol. Biol. Ecol. 24: 2669–2680.

Li, W.-H. 1997. Molecular evolution. – Sinauer.

Liu, K. et al. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. – Science 324: 1561–1564.

Liu, K. et al. 2012. SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. – Syst. Biol. 61: 90–106.

Lockwood, J. L. et al. 2002. A metric for analyzing taxonomic patterns of extinction risk. – Conserv. Biol. 16: 1137–1142.

Lovette, I. J. and Hochachka, W. M. 2006. Simultaneous effects of phylogenetic niche conservatism and competition on avian community structure. – Ecology 87: S14–S28.

Löytynoja, A. and Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. – Science 320: 1632–1635.

Löytynoja, A. and Goldman, N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. – BMC Bioinform. 11: 579.

Maddison, W. P. and Knowles, L. L. 2006. Inferring phylogeny despite incomplete lineage sorting. – Syst. Biol. 55: 21–30.

Maddison, W. P. and Maddison, D. R. 2007. Mesquite: a modular system for evolutionary analysis, version 2.0. – <http://mesquiteproject.org>.

Maddison, W. P. et al. 1984. Outgroup analysis and parsimony. – Syst. Zool. 33: 83–103.

Magallón, S. A. 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. – Int. J. Plant Sci. 165: 7–21.

Martin, A. P. and Palumbi, S. R. 1993. Body size, metabolic rate, generation time, and the molecular clock. – Proc. Natl Acad. Sci. USA 90: 4087–4091.

McGoogan, K. et al. 2007. Phylogenetic diversity and the conservation biogeography of African primates. – J. Biogeogr. 34: 1962–1974.

McMahon, M. M. and Sanderson, M. J. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. – Syst. Biol. 55: 818–836.

Moles, A. T. et al. 2005. Factors that shape seed mass evolution. – Proc. Natl Acad. Sci. USA 102: 10540–10544.

Morrison, D. A. 2009. A framework for phylogenetic sequence alignment. – Syst. Biol. 282: 127–149.

Morrison, D. A. and Ellis, J. T. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. – Mol. Biol. Ecol. 14: 428–441.

Mouquet, N. et al. 2012. Ecophylogenetics: advances and perspectives. – Biol. Rev. doi: 10.1111/j.1469-185X.2012.00224.x

Nuin, P. A. S. et al. 2006. The accuracy of several multiple sequence alignment programs for proteins. – BMC Bioinform. 7: 471.

Nylander, J. A. A. 2004. MrModeltest v2. – Evolutionary Biology Centre, Uppsala.

Nylander, J. A. A. et al. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. – Bioinformatics 24: 581–583.

Ohta, T. 2002. Near-neutrality in evolution of genes and in gene regulation. – Proc. Natl Acad. Sci. USA 99: 16134–16137.

Ota, R. and Penny, D. 2003. Estimating changes in mutational mechanisms of evolution. – J. Mol. Evol. 57: S233–S240.

Page, R. D. M. 2012. Space, time, form: viewing the Tree of Life. – Trends Ecol. Evol. 27: 113–120.

Paland, S. and Lynch, M. 2006. Transitions to asexuality result in excess amino acid substitutions. – Science 311: 990–992.

Pamilo, P. and Nei, M. 1988. Relationship between gene trees and species trees. – Mol. Biol. Evol. 5: 568–583.

Pattengale, N. D. et al. 2010. How many bootstrap replicates are necessary? – J. Comp. Biol. 17: 337–354.

Rabosky, D. 2006. LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. – Evol. Bioinform. 2: 247–250.

Rannala, B. and Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. – J. Mol. Evol. 43: 304–311.

Roelants, K. et al. 2007. Global patterns of diversification in the history of modern amphibians. – Proc. Natl Acad. Sci. USA 104: 887–892.

Ronquist, F. and Huelsenbeck, J. P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. – Bioinformatics 19: 1572–1574.

Rutschmann, F. 2005. Bayesian molecular dating using PAML/MULTIDIVTIME. A step-by-step manual. Version 1.5. – < www.plant.ch >.

Rutschmann, F. 2006. Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. – Divers. Distrib. 12: 35–48.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. – Mol. Biol. Ecol. 4: 406–425.

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. – Mol. Biol. Ecol. 14: 1218–1231.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. – Mol. Biol. Ecol. 19: 101–109.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. – Bioinformatics 19: 301–302.

Sanderson, M. J. and Shaffer, H. B. 2002. Troubleshooting molecular phylogenetic analyses. – Annu. Rev. Ecol. Syst. 33: 49–72.

Sanderson, M. J. and Driskell, A. C. 2003. The challenge of constructing large phylogenetic trees. – Trends Plant Sci. 8: 374–379.

Sanderson, M. J. et al. 2008. The PhyLoTA browser: processing GenBank for molecular phylogenetics research. – Syst. Biol. 57: 335–346.

Sanderson, M. J. et al. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. – BMC Evol. Biol. 10: 155.

Sanderson, M. J. et al. 2011. Terraces in phylogenetic tree space. – Science 333: 448–450.

Sauquet, H. et al. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). – Syst. Biol. 61: 289–313.

Schaefer, H. et al. 2011. Testing Darwin's naturalization hypothesis in the Azores. – Ecol. Lett. 14: 389–396.

Schuettpelz, E. and Pryer, E. 2007. Fern phylogeny inferred from 400 leptosporangiate species and three plastid genes. – Taxon 56: 1037–1050.

Shaw, J. et al. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. – Am. J. Bot. 92: 142–166.

Silvertown, J. et al. 2006. Absence of phylogenetic signal in the niche structure of meadow plant communities. – Proc. R. Soc. B 273: 39–44.

Smith, S. A. and Dunn, C. 2008. Phyutility: a phyloinformatics utility for trees, alignments, and molecular data. – Bioinformatics 24: 715–716.

Smith, S. A. and Donoghue, M. J. 2010. Combining historical biogeography with niche modeling in the Caprifolium clade of Lonicera (Caprifoliaceae, Dipsacales). – Syst. Biol. 59: 322–341.

Smith, S. A. et al. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. – BMC Evol. Biol. 9: 37.

Smith, S. A. et al. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. – Am. J. Bot. 98: 404–414.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. – Bioinformatics 22: 2688–2690.

Stamatakis, A. 2008. The RAxML 7.0.4 manual. – < http://icwww.epfl.ch/~stamatak/ >.

Stamatakis, A. et al. 2008. A rapid bootstrap algorithm for the RAxML web-servers. – Syst. Biol. 75: 758–771.

Stamatakis, A. et al. 2012. RAxML-Light: a tool for computing terabyte phylogenies. – Bioinformatics 28: 2064–2066.

Strauss, Y. E. et al. 2006. Exotic taxa less related to native species are more invasive. – Proc. Natl Acad. Sci. USA 103: 5841–5845.

Stuart, S. N. et al. 2004. Status and trends of amphibian declines and extinctions worldwide. – Science 306: 1783–1786.

Suzuki, Y. et al. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. – Proc. Natl Acad. Sci. USA 99: 16138–16143.

Swofford, D. L. et al. 1996. Phylogenetic inference. – In: Hillis, D. M. et al. (eds), Molecular systematics. Sinauer, pp. 407–514.

Talavera, G. and Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. – Syst. Biol. 56: 564–577.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. – Lect. Math. Life Sci. 17: 57–86.

Thomas, G. H. 2008. Phylogenetic distributions of British birds of conservation concern. – Proc. R. Soc. B 275: 2077–2083.

Thomas, J. A. et al. 2006. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. – Proc. Natl Acad. Sci. USA 103: 7366–7371.

Thompson, J. D. et al. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. – Nucl. Acids Res. 22: 4673–4680.

Thomson, R. C. and Shaffer, H. B. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa and the phylogeny of living tutles. – Syst. Biol. 59: 42–58.

Thorne, J. L. and Kishino, H. 2002. Divergence time and evolutionary rate estimation with multilocus data. – Syst. Biol. 51: 689–702.

Thorne, J. L. et al. 1998. Estimating the rate of evolution of the rate of molecular evolution. – Mol. Biol. Evol. 15: 1647–1657.

Thuiller, W. et al. 2010. Resolving Darwin's naturalization conundrum: a quest for evidence. – Divers. Distrib. 16: 461–475.

Thuiller, W. et al. 2011. Consequences of climate change on the Tree of Life in Europe. – Nature 470: 531–534.

Tierney, L. 1994. Markov chains for exploring posterior distributions. – Ann. Stat. 22: 1701–1762.

Vamosi, J. C. and Vamosi, S. M. 2007. Body size, rarity, and phylogenetic community structure: insights from diving beetle assemblages of Alberta. – Divers. Distrib. 13: 1–10.

Webb, C. O. et al. 2002. Phylogenies and community ecology. – Annu. Rev. Ecol. Evol. Syst. 33: 475–505.

Webb, C. O. et al. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. – Bioinformatics 24: 2098–2100.

Whelan, S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. – Mol. Biol. Ecol. 25: 1683–1694.

Whitney, K. D. et al. 2009. Hybridization-prone plant families do not generate more invasive species. – Biol. Invasions 11: 1205–1215.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. – Syst. Biol. 52: 528–538.

Wilkinson, M. 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. – Syst. Biol. 43: 343–368.

Wright, I. J. et al. 2007. Relationships among ecologically important dimensions of plant trait variation in seven neotropical forests. – Ann. Bot. 99: 1003–1015.

Yarza, P. et al. 2008. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. – Syst. Appl. Microbiol. 31: 241–250.

Yoder, A. and Yang, Z. 2000. Estimation of primate speciation dates using local molecular clocks. – Mol. Biol. Ecol. 17: 1081–1090.

Zhang, J. and Madden, T. L. 1997. PowerBLAST: a new network BLAST application for interactive and automated sequence analysis and annotation. – Genome Res. 7: 649–656.

Zhang, J. et al. 2010. Phylotools: phylogenetic tools for ecologists v0.0.7.4. – <http://CRAN.R-project.org/package=phylotools>.

Zuckerkandl, E. and Pauling, L. 1965. Evolutionary divergence and convergence in proteins. – In: Bryson, V. and Vogel, H. (eds), Evolving genes and proteins. Academic Press, pp. 97–166.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. – PhD thesis, Univ. of Texas.

Supplementary material (Appendix E7773 at <www.oikosoffice.lu.se/appendix>). Appendix 1–2.