A decade of progress in plant molecular phylogenetics

Vincent Savolainen and Mark W. Chase

Molecular Systematics Section, Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond TW9 3DS, UK

Over the past decade, botanists have produced several thousand phylogenetic analyses based on molecular data, with particular emphasis on sequencing *rbcL*, the plastid gene encoding the large subunit of Rubisco (ribulose bisphosphate carboxylase). Because phylogenetic trees retrieved from the three plant genomes (plastid, nuclear and mitochondrial) have been highly congruent, the 'Angiosperm Phylogeny Group' has used these DNA-based phylogenetic trees to reclassify all families of flowering plants. However, in addition to taxonomy, these major phylogenetic efforts have also helped to define strategies to reconstruct the 'tree of life', and have revealed the size of the ancestral plant genome, uncovered potential candidates for the ancestral flower, identified molecular living fossils, and linked the rate of neutral substitutions with species diversity. With an increased interest in DNA sequencing programmes in non-model organisms, the next decade will hopefully see these phylogenetic findings integrated into new genetic syntheses, from genomes to taxa.

Phylogenetics – the study of the evolutionary history and relationships of biological taxa – has been revolutionized by DNA sequence data. In the early 1980s, plant physiologists characterized a plastid gene, *rbcL*, encoding the large subunit of ribulose bisphosphate carboxylase (Rubisco), the most abundant enzyme on earth [1]. Because *rbcL* is a key photosynthetic gene, Zurawski and his colleagues were interested in comparing *rbcL* gene sequences from as many taxa as possible, thereby possibly increasing our knowledge of photosynthetic pathways and improving attempts to manipulate photosynthesis, for example, in crops. To achieve this goal, they distributed *rbcL* primers free of charge at a time when all phases of sequencing were costly. As a by-product of this initiative, plant systematists collected *rbcL* DNA sequences for a broad sampling of seed plants (499 species), resulting in one of the first collaborative large-scale phylogenetic analyses, just a decade ago [2]. Since then, several thousand molecular-based phylogenetic analyses have been published for all types of organisms [3]. Rather than reviewing phylogenetic methodologies or the details of ten years in plant phylogenetics, we will concentrate here on some of the major and recent advances, from assembling the general 'tree of life' to the evolution of genes, genomes and the origin of biodiversity. Our discussion emphasizes results from *rbcL* analyses, but we have also included several other relevant publications covering our understanding of plant taxonomy, evolution and methodology.

Towards assembling the 'tree of life': size matters

Over time genome sequences evolve – undergoing mutation and fixation in populations. The extent of the substitution differences in homologous sequences often reflects the evolutionary distinctiveness of organisms with respect to each other; thus, this information can be used to reconstruct molecular phylogenetic trees. Although for prokaryotes a complete-genome approach might be necessary due to the large numbers of horizontal transfers that occurred during early stages of life on Earth [4,5], large-scale multigene-based phylogenetic analyses are practical for many eukaryotes and particularly for plants. In addition, nucleotide changes are roughly clocklike, although the speed at which the clock ticks is usually different between lineages; nevertheless, providing that one can correct for this RATE HETEROGENEITY (see Glossary), nucleotide divergence can also be used as a surrogate for time (Box 1).

Several METHODS TO BUILD PHYLOGENETIC TREES have been developed, but building trees remains a hypercomplex

Box 1. Calibrating molecular phylogenetic trees with fossils

To calibrate molecular phylogenetic trees with fossils (or any biogeographical and tectonic event of known age), several options are available. The simplest way is to look at nucleotide divergence between pairs of extant taxa in a tree, which are the products of molecular change (divergence) that has arisen since these taxa evolved from a common ancestor; this date can be inferred from the fossil record and provides a rate of change that can be used to calculate in turn the ages of all the other nodes of the tree. This procedure, however, assumes a constant molecular clock throughout the tree (i.e. equal rates in each branch from the root), unless it is subdivided into subtrees in which different fossils can be used to provide several estimates for the rates of substitutions in the respective parts. An alternative is to correct first for rate heterogeneity across the tree. For example, it can be appropriate to assume that despite the fact that rates can differ among lineages, they are autocorrelated along lineages from parent to daughter branches, that is, rates are at least partly heritable. Several algorithms can then model the evolution of differential rates along these lineages and can apply some corrections, thereby transforming molecular branch lengths into relative time. Then one fossil calibration point can be used to transform relative time into absolute ages as described above. With more complex algorithms, it is also possible to use simultaneously several fossils for calibrations and to fix minimum, maximum or intervals of ages for some nodes in the tree [55,71].

Corresponding author: Vincent Savolainen (v.savolainen@kew.org).

http://tigs.trends.com 0168-9525/$ - see front matter © 2003 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2003.10.003

Glossary

Angiosperms (flowering plants): plants with flowers and ovules enclosed in an ovary.
Bootstrap: a computational technique in which a percentage of the original data are deleted and randomly resampled to recreate a matrix of the original size, which is used to evaluate support for the groups on the phylogenetic trees.
Convergence: nucleotide changes resulting in identity driven by chance or selection for similar function but not due to common history.
Eudicots: the group of flowering plants with triaperturate pollen.
Functional constraints: the effect of natural selection on DNA to conserve function at the protein level.
Homoplastic changes: any nucleotide changes resulting in identity at a given nucleotide position not due to common history, namely, convergence, parallelism and reversion.
Jackknife: a computational technique in which data points of the original matrix are randomly deleted and the analysis rerun to evaluate clarity of patterns in phylogenetic trees and expressed as percentages of such replicates in which a group of taxa occurs.
Maximum likelihood methods: a computational technique in which phylogenetic trees are built according to models of nucleotide evolution (i.e. incorporating different frequencies of change and nucleotide composition as well as probabilities of change).
Methods to build phylogenetic trees: any of three main categories of computational techniques commonly used to build DNA-based phylogenetic trees: (i) distance methods, in which pairwise genetic distances are used to build trees; (ii) maximum parsimony methods, in which overall nucleotide changes are minimized in the tree-building process (usually with equal probabilities for all changes, but which can also incorporate uneven probabilities much as in maximum likelihood methods); and (iii) maximum likelihood methods (see above). Recently, Bayesian methods have been used in phylogeny inference [71].
Monocots: flowering plants with uniaperturate pollen and parallel leaf venation, comprising palms, grasses, orchids, irises, etc. (Figure 5).
Dicots: a term that referred to the group of plants with two cotyledons (the two specialized leaves that provide nutrients to the growing plantlet) but that phylogenetic studies have shown to be an artificial taxon (Figure 4).
Nonparametric rate smoothing: a computational technique in which rate heterogeneity in DNA sequences is corrected across lineages to make branch lengths proportional to time only.
Rate heterogeneity: the presence of significant difference in the amount of nucleotide changes between lineages or at sites within a DNA region.
Reversion: any nucleotide change that results in restoration of the initial nucleotide (e.g. adenine changing to thymine, and then returning to the original base: that is, thymine back to adenine).
Root: the first split (node) of a phylogenetic tree.
Taxon (pl. taxa): any level in the classification of organisms, for example, species, genus, family and order (Figure 3).
Triaperturate pollen: pollen with three openings, through one of which the pollen tube germinates and transfers the sperm to the ovule.
Ultrametric: referring to a phylogenetic tree in which branch lengths (genetic divergence) are proportional to time only and within which rate heterogeneity among lineages, if any, has been corrected.
Uniaperturate pollen: pollen with a single opening, through which the pollen tube germinates and transfers the sperm to the ovule.
Vascular plants: all plants with tissues specialized for the transport of water, nutrients and minerals.

mathematical problem because the number of solutions (possible trees) that ideally should be evaluated increases exponentially with TAXON number. For example, when using just over 100 taxa, the number of possible trees exceeds the number of particles in the universe. This problem has been brought sharply into focus as a result of large-scale sequencing projects focused on ANGIOSPERM phylogeny and more generally towards assembling the 'tree of life'.

DNA sequences can have rates of substitution at some sites that are so high that the information could be lost due to multiple changes, REVERSIONS and saturation (Figure 1); as a result, sequences from distantly related taxa might be spuriously attracted to each other by one of several forms of 'long branch attraction' [6]. Simulations in four-taxon cases have shown that most tree-building methods would be inconsistent (i.e. converge on a wrong solution) in case of saturation unless methods are used that will correct for unobserved changes [7]. MAXIMUM LIKELIHOOD METHODS have been popular in this respect, but these methods are immense consumers of computer time. If a simple four-taxon case cannot be solved readily, even after sequencing several thousand nucleotides, how can a reliable phylogenetic tree with several thousand taxa be built? An answer came from a study of ribosomal DNA (rDNA) sequences in angiosperms: bigger is better – that is, more taxa are at least as beneficial as longer gene sequences. To evaluate how phylogenetic reconstruction is improved when adding more taxa or nucleotides, Hillis performed a simulation experiment with a large tree [8]. He used a 223-plant taxon, nonclocklike tree based on 18S rDNA as a model tree, and simulated on this tree the evolution of DNA sequences of various lengths. Then, using these artificial sequences in phylogenetic analyses, he asked how many variable nucleotide positions are necessary to recover the model tree. Unexpectedly, he found that as few as 5000 variable base positions (i.e. when all sites changed at least once in the tree) were sufficient to recover in every detail the model tree correctly using maximum parsimony [8]. When Hillis then simulated sequence evolution at rates up to ten times faster, the tree was correctly inferred with even fewer nucleotides [9]. Because the four-taxon studies showed that most phylogenetic methods would fail to recover correct trees if nucleotide change does not follow a constant clock [10], these results at first surprised the phylogenetic community. However, it was quickly realized that larger trees reveal more nucleotide changes overall (there are more branches on which nucleotides can change),

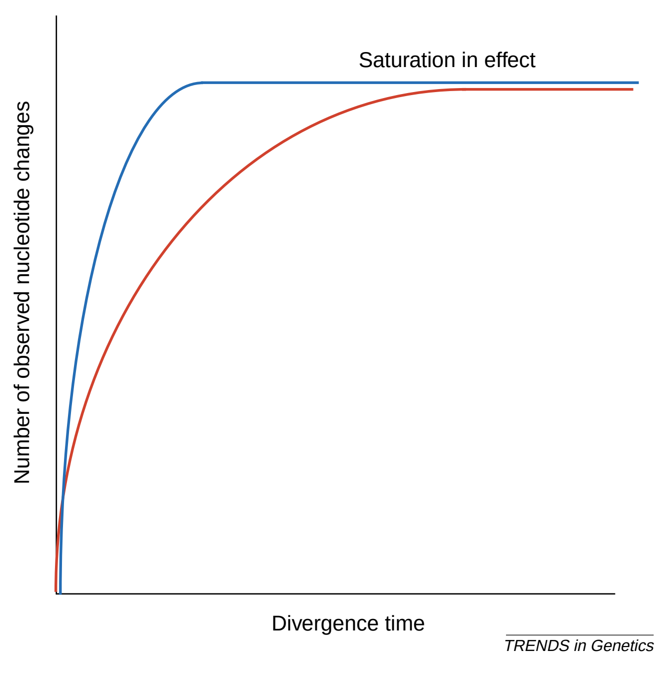

Figure 1. Saturation: when observed nucleotide changes are plotted against time, a plateau is reached when divergence time is great enough for reversions to mask the true number of substitutions; note that DNA sequences with higher substitution rates (blue) reach saturation more quickly than sequences with lower rates (red).

and this makes it easier to recover an accurate phylogenetic signal (Figure 2). In particular, although the number of inferred HOMOPLASTIC CHANGES (i.e. base positions that share nucleotides due to CONVERGENCE and reversion) in larger datasets is higher, and these were at first regarded as 'noise', they can be locally informative: they can reflect relationships in restricted parts of the tree in spite of being globally uninformative (Figure 2). For example, although third-codon positions in protein-coding genes accumulate more changes than first or second positions as a result of the redundancy of the genetic code, they are often more informative than other codon positions in plant datasets (sometimes also including bacteria) [11–14], an observation that contrasts with findings in animals [15] (but see Ref. [16]). These findings have been of immense general importance – outside of angiosperm studies – and they have reorientated strategies used to reconstruct the 'tree of life'.

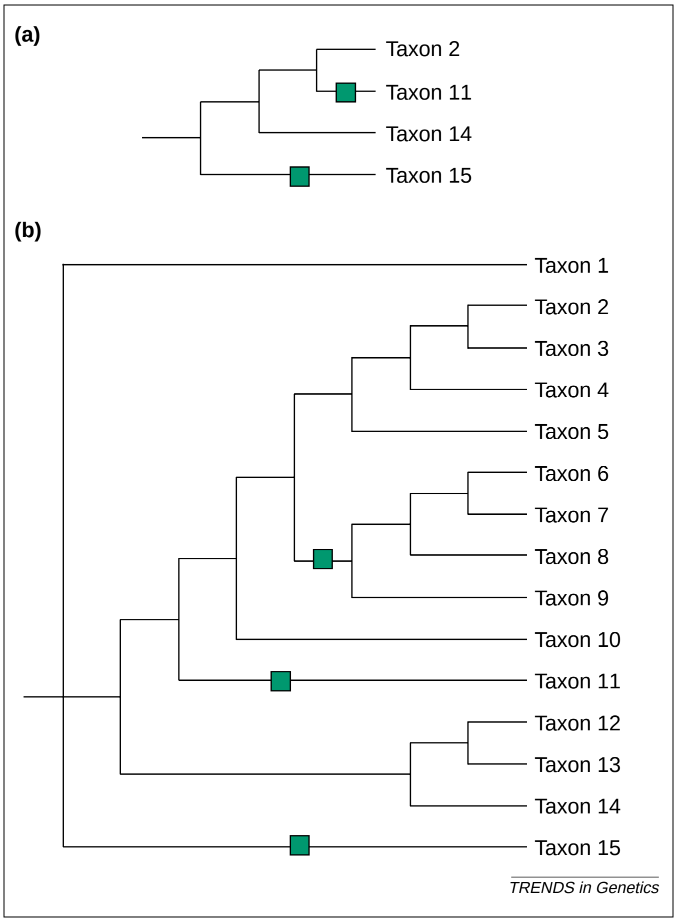

Figure 2. Illustration of how recovery of phylogenetic signal is easier in larger trees. The green square represents a hypothetical substitution (e.g. adenine to cytosine) at one particular site. **(a)** In the smaller tree, this change occurred independently twice, that is, along the branches leading to taxa 11 and 15, and therefore this substitution is a convergence and does not tell anything about evolutionary relationships. **(b)** When additional taxa are added to give the larger tree, this substitution is found on another branch, namely, the ancestor leading to the group of taxa 6 to 9. In this latter case, this change reflects common ancestry despite the fact that overall it is homoplastic. Bigger trees simply have more chance to exhibit such substitutions: that is, substitutions that are 'locally' informative of shared evolutionary history.

Three genomes, one tree

In plant phylogenetics, perhaps one of the most reliable measures of confidence in our trees is the congruence between the information retrieved from the three genomes (plastid, nuclear and mitochondrial). Phylogenetic analyses of angiosperms comprising up to a few thousand taxa (up to 2538 [11]) have been performed with the plastid *rbcL* gene [2,17], *rbcL* combined with plastid *atpB* [13], plastid inverted repeat [18], and various combinations of nuclear rDNA [19–21], nuclear phytochrome genes [22] and mitochondrial *matR* and *atp1* genes [23,24]. Data matrices containing many additional genes have recently been analysed for flowering plants [25]. Although there are sometimes differences of pattern in the published trees, strongly incongruent groupings have rarely been found [26], that is, no contradictory groups depicted in analyses of different genomes received support as measured by the BOOTSTRAP or JACKKNIFE. At the taxonomic level (Figure 3) of families and above, all three genomes appear to be tracking the same evolutionary history. The main factors that could alter detection of historical patterns would be differing structural and FUNCTIONAL CONSTRAINTS (i.e. those caused by strong selection), but combining several genes would be expected to average out such forces operating on individual genes.

There have been reports that DNA sequences from the three genomes evolve at different rates, with those from the nuclear genome being the fastest and those from mitochondrial and plastid DNA the slowest [27]; gene rearrangements are frequent in the mitochondrion, but this does not have an affect on phylogeny reconstruction based on the gene sequences. This situation contrasts with that of animals in which mitochondrial DNA has a higher

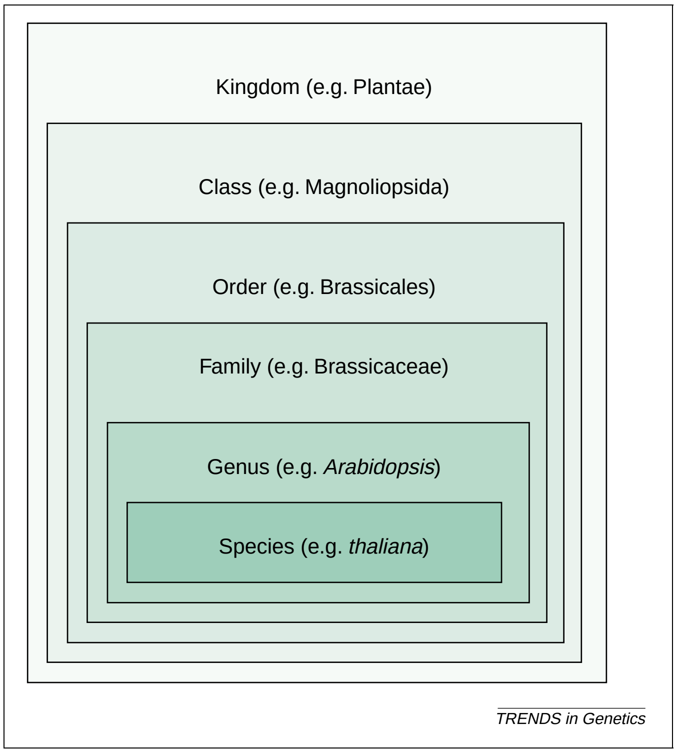
Kingdom (e.g. Plantae)
Class (e.g. Magnoliopsida)
Order (e.g. Brassicales)
Family (e.g. Brassicaceae)
Genus (e.g. *Arabidopsis*)
Species (e.g. *thaliana*)
TRENDS in Genetics

Figure 3. The systematic hierarchical categories of the classification of organisms using the example of *Arabidopsis*.

rate of nucleotide substitution than the nuclear genome and a lower rate of structural evolution than plant mitochondrial and plastid genomes [28]. Furthermore, there has been a great deal of confusion caused by genes being described as rapidly or slowly evolving; for example, 'rapidly evolving' or 'higher rates' could mean higher rates at the same variable sites, more variable sites in some homologous genes or a combination of both [29]. One main issue has been the effect of differential structural and functional constraints, and there have been some concerns about how these might affect phylogenetic inference, especially for the small organellar genomes (with fewer genes) so often used in phylogeny reconstruction and where constraints might be stronger as a result of 'lack of space'. For example, in animals differential functional constraints acting on nonneutral nucleotides of different proteins of the mitochondrial genome have resulted in incorrect evolutionary relationships receiving strong support [28,30]. For anciently diverged plants, concerns have also been raised [31,32], but in angiosperms close examination of plastid genes for their signal content (i.e. nucleotide changes shared due to common history) showed that these genes exhibited evenly distributed phylogenetic information [14] in the different codon positions, amino acids, chemical properties, hydrophobicity and charge, which is the opposite of the animal mitochondrial genome. It is clear that if severe functional constraints were acting on the plastid genome of flowering plants, we would have expected these sites to exhibit changes that not only reflect common history but also convergent changes necessary to preserve function; this was not the case [14]. Therefore, at least for angiosperms, it seems that botanists have made enormous strides in phylogeny inference due to characteristics inherent to the plastid genome (in terms of rates and types of changes at variable sites).

'A rose is still a rose but otherwise everything else in botany has been turned on its head'

Although not as drastic as stated in *The Independent*, 'Botanists reclassify all plants… A rose is still a rose but otherwise everything else in botany has been turned on its head' (pp. 1 and 3, 23 November 1998), botanists have produced the first DNA sequence-based classifications for a major group of organisms. Because angiosperm phylogenetic trees containing several hundreds of taxa were highly congruent although produced by genes in different genomes, botanists decided that it was time to translate the resulting patterns of relationships into a new and comprehensive classification. Rather than a classification reflecting the subjective views of a single author (i.e. based on intuitive ideas of plant evolution), the 'Angiosperm Phylogeny Group' (APG) aimed objectively to interpret published phylogenetic trees and compile them into a hierarchical system at and above the level of family. Their first classification was published in 1998 [33], and an update appeared in early 2003 [34]. The APG classification reflects evolutionary relationships that were newly discovered for ~60% of angiosperm families [33,35]; the main objective of this classification was to maximize information, thus making the system predictive [20].

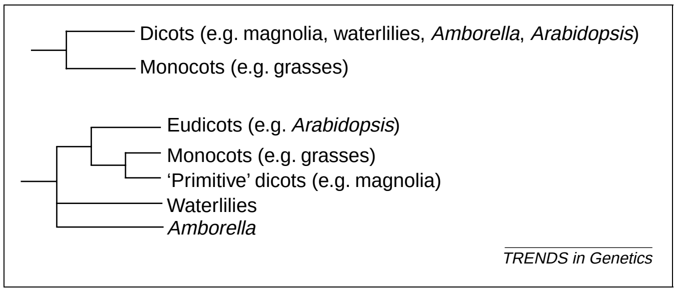
Dicots (e.g. magnolia, waterlilies, *Amborella*, *Arabidopsis*)
Monocots (e.g. grasses)
Eudicots (e.g. *Arabidopsis*)
Monocots (e.g. grasses)
'Primitive' dicots (e.g. magnolia)
Waterlilies
Amborella
TRENDS in Genetics

Figure 4. The major splits within angiosperms as they were viewed before the molecular phylogenetics era (top) and more recently as demonstrated by the use of molecular data (below).

The broad picture of angiosperm relationships has changed, with the first split among angiosperms not being that of MONOCOTS versus DICOTS, as stated in most botany textbooks, but instead one in which the 'primitive dicots' are closer to the monocots, a relationship reflected in their UNIAPERTURATE POLLEN grains versus the TRIAPERTURATE POLLEN of other dicots, the latter now being termed 'EUDICOTS' [33] (Figure 4). Perhaps one of the most spectacular changes of ideas concerns the sacred lotus (*Nelumbo*); because of its morphology and habitat preferences, the lotus was always considered a close relative of the water lilies (Nymphaeaceae), a group of 'primitive' dicots, whereas based on DNA sequence it is a eudicot for which the closest relatives are the northern temperate plane tree (*Platanus*) and the southern-hemisphere *Protea* family (Proteaceae) [20].

Rooting the phylogenetic tree of the angiosperms

Discovering new relationships is of course not only relevant to classification. Finding the ROOT of angiosperms, for example, has been the focus of several studies because it provides a direction and temporal scale for plant evolution (mostly calibrated with the fossil record) (Box 1), thereby permitting the production of explicit hypotheses of how traits such as genome size, colinearity of genes on chromosome arms and development have changed during the past 125 million years. Such ideas can then be used to generate research programmes designed to evaluate such predictions. The large flowers of *Magnolia* were long considered the archetype of the angiosperm flower because of their numerous, spirally arranged floral parts, but it has recently become evident that other flower types are equally as 'primitive' as those of *Magnolia*. These include the flowers of unusual plants such as *Amborella* (but see Ref. [36] for an alternative and controversial view) and *Piper* (the source of black pepper) (Figure 5), which were found to be outside the major clades in phylogenetic trees for angiosperms. It must, however, be stressed that knowing how remnants of basal lineages appear today does not necessarily tell us much about the traits of the ancestral angiosperms [37]. The first flowers could have been different from those of every extant group, and we will not know about them until their fossils have been found. The oldest angiosperm fossils are water lilies [38] and another aquatic plant, belonging to the newly described family Archaefructaceae [39], both ~125 million years old. Molecular systematic studies have refined ideas about

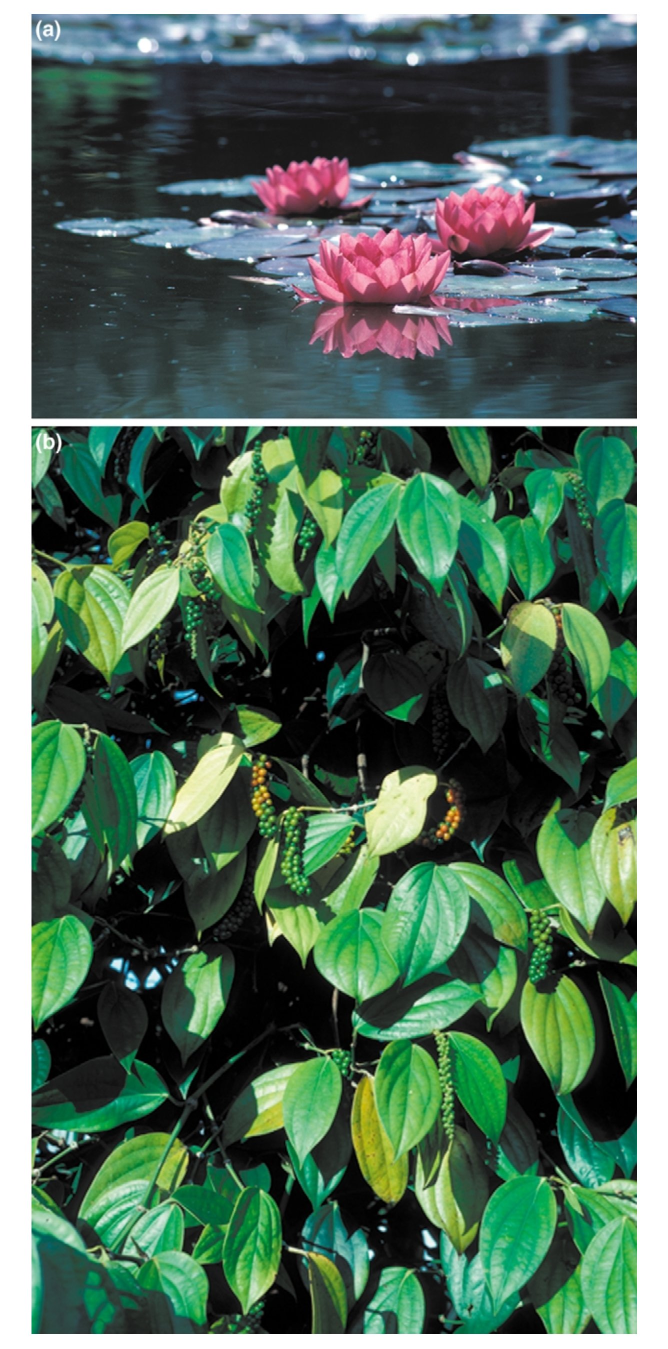

Figure 5. Two potential candidates of the archetype of the angiosperm flower: waterlily (*Nymphaea*, top) and black pepper (*Piper*, below). Photograph, courtesy of P. Gasson, Kew.

which sorts of fossils to look for, but the study of extant lineages alone cannot reveal all that is important for understanding the early angiosperms. For this, studies of fossils are essential.

Genome changes and plant evolution

As described above, it is difficult to infer the floral archetype of the angiosperms solely from knowledge of the phylogenetic relationships of extant species; we can, however, study several other important traits of the early angiosperms in this way, as long as we do not expect them to have been too plastic during the early stages of evolution. For example, by mapping genome sizes gathered from online databases [40] onto the general angiosperm phylogenetic tree, it was possible to infer that the ancestral genome was probably small [41]. How plant genomes increased to the large sizes observed in some modern groups [e.g. >127 pg per unreplicated gametic nucleus (1C value), in *Fritillaria*, a close relative of the lily] remains unexplained, but studies of selfish DNA and other retrotransposable elements could provide key answers [42–47]. At the least, knowing plant relationships can now help pinpoint, which lineages should be of interest, namely, those that have experienced the most drastic expansions or contractions in their genomes, especially because genome change might have provided major contributions to angiosperm radiations. For example, it is known that up to 70% of extant species are descended from taxa in which polyploidization events have occurred [48].

Features of genome evolution have also provided insight into plant phylogeny and vice versa. One example is the striking case of loss of the standard plant telomeric sequences (*Arabidopsis*-type repeats) and their replacement by other categories of repeats. *In situ* hybridization with telomeric probes demonstrated that onion (*Allium*) and aloe (*Aloe*) lack the typical repeats that cap all chromosome arms in the majority of plants [49]. By looking at the DNA-based phylogenetic analysis, it was clear that both species were members of the same order, Asparagales (as redescribed by APG [33], Figure 6) note that in many previous classifications these species were regarded as only distantly related), and therefore most Asparagales genera were examined for absence of the standard telomeric sequences [49]. Apart from a few closely related species of *Ornithogalum* (star of Bethlehem), none of the species between the aloe and onion has the typical plant telomeric repeats. Without phylogenetic information, none of these patterns would have been likely to be investigated in this manner, and clearly ‘tree thinking’ played a key role in this discovery.

Molecular clocks and molecular living fossils

The estimation of divergence times between species is important because it makes it possible to determine the speed of a variety of evolutionary processes, such as chromosome rearrangements, emergence of new forms of viruses and production of new body plans. When Zuckerlandl and Pauling found that the number of amino acid substitutions in haemoglobin was correlated with fossil-based time divergence estimates in vertebrates, the concept of the ‘molecular clock’ was born [50]. However, we now know that this clock ticks at varying speeds between lineages of organisms, and fossil-based versus DNA-based age estimates usually disagree, with molecules generally providing much older ages [51]. Using the broadly sampled angiosperm phylogenetic tree (based on plastid *rbcL* and *atpB* and nuclear 18S rDNA and comprising ~75% of all families [20,21]), NONPARAMETRIC RATE SMOOTHING (NPRS) [52] was applied to correct for rate heterogeneity across lineages, thus making the tree ULTRAMETRIC [53]. This chronogram was calibrated with reliable fossil data (the unique structure of the nuts of oaks and their allies; Box 1), and error estimates for the ages of the nodes of that tree were calculated by reapplying the NPRS protocols to

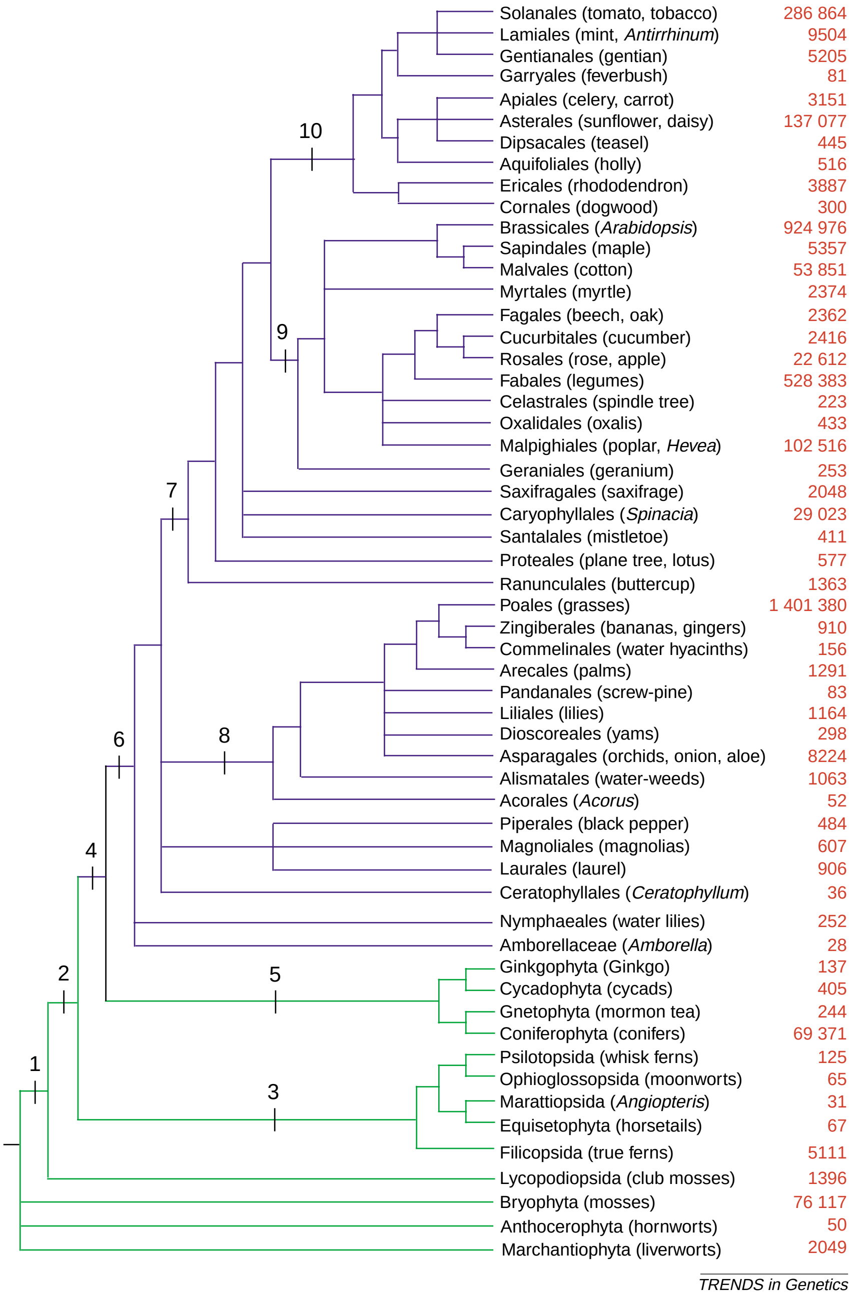
Solanales (tomato, tobacco) 286 864
Lamiales (mint, *Antirrhinum*) 9504
Gentianales (gentian) 5205
Garryales (feverbush) 81
Apiales (celery, carrot) 3151
Asterales (sunflower, daisy) 137 077
Dipsacales (teasel) 445
Aquifoliales (holly) 516
Ericales (rhododendron) 3887
Cornales (dogwood) 300
Brassicales (*Arabidopsis*) 924 976
Sapindales (maple) 5357
Malvales (cotton) 53 851
Myrtales (myrtle) 2374
Fagales (beech, oak) 2362
Cucurbitales (cucumber) 2416
Rosales (rose, apple) 22 612
Fabales (legumes) 528 383
Celastrales (spindle tree) 223
Oxalidales (oxalis) 433
Malpighiales (poplar, *Hevea*) 102 516
Geraniales (geranium) 253
Saxifragales (saxifrage) 2048
Caryophyllales (*Spinacia*) 29 023
Santalales (mistletoe) 411
Proteales (plane tree, lotus) 577
Ranunculales (buttercup) 1363
Poales (grasses) 1 401 380
Zingiberales (bananas, gingers) 910
Commelinales (water hyacinths) 156
Arecales (palms) 1291
Pandanales (screw-pine) 83
Liliales (lilies) 1164
Dioscoreales (yams) 298
Asparagales (orchids, onion, aloe) 8224
Alismatales (water-weeds) 1063
Acorales (*Acorus*) 52
Piperales (black pepper) 484
Magnoliales (magnolias) 607
Laurales (laurel) 906
Ceratophyllales (*Ceratophyllum*) 36
Nymphaeales (water lilies) 252
Amborellaceae (*Amborella*) 28
Ginkgophyta (Ginkgo) 137
Cycadophyta (cycads) 405
Gnetophyta (mormon tea) 244
Coniferophyta (conifers) 69 371
Psilotopsida (whisk ferns) 125
Ophioglossopsida (moonworts) 65
Marattiopsida (*Angiopteris*) 31
Equisetophyta (horsetails) 67
Filicopsida (true ferns) 5111
Lycopodiopsida (club mosses) 1396
Bryophyta (mosses) 76 117
Anthocerophyta (hornworts) 50
Marchantiophyta (liverworts) 2049
10
9
7
8
6
4
2
5
1
3
TRENDS in Genetics

Figure 6. A summary of the terrestrial plant 'tree of life' [20,59] showing vascular plants (all descendants from node 1), which comprise angiosperms (nodes 6–10 depicted in blue) and remaining vascular plants (nodes 1-5 depicted in green). The main groups are leafy plants (node 2), ferns and their allies (node 3), seed plants (node 4), gymnosperms (node 5), flowering plants (node 6), eudicots (node 7), monocots (node 8), rosids (node 9) and asterids (node 10) (time scale not enforced). For flowering plants, most orders are indicated with some of their typical representatives or model organisms. Numbers on the right indicate the number of nucleotides entries held in EBI or GenBank in early November 2002, summing entries from mitochondrial, plastid and nuclear genomes. Several groups have a large number of entries because of the sequencing effort on model organisms. For conifers ~91% of entries belong to *Pinus*; for mosses, *Physcomitrella* (92%); for Malpighiales, *Populus* (92%); for Fabales, *Glycine* (57%) and *Medicago* (33%); for Asterales, *Lactuca* (49%) and *Helianthus* (33%); for Brassicales, *Brassica* (51%) and *Arabidopsis* (48%); and for Poales, *Zea* (26%), *Hordeum* (24%), *Oryza* (23%) and *Triticum* (20%).

bootstrapped DNA matrices. This molecular dating work is the largest published so far in terms of number of taxa (see Refs [54–56] for complementary references). It provides ages for the origin of nearly all angiosperm families, and most of these are in agreement for groups with a good fossil record [57]. However, like most previous studies of molecular clocks, the ages of the deepest nodes were underestimated by the fossil record, whereas the ages of the most recent groups thigh degree of correspondence, for most lineages, between fossil ages and the clock estimates [53] means that the ages of those without a fossil record can now be more reliably estimated than ever before, and this includes the great majority of angiosperm families and orders.

Looking at the VASCULAR PLANTS as a whole, a similar NPRS dating exercise was recently performed [58] using the most comprehensive phylogenetic tree for all lineages of vascular plants based on four genes [59]. Ages were depicted from single genes or combinations, in maximum parsimony and maximum likelihood frameworks, with several fossils of undisputed ages used as calibration points [58]. Many DNA-based age estimates were in agreement with those from fossils, but it was also discovered that certain lineages have drastically decreased their rates of molecular evolution. This was the case, for example, with tree ferns, which were considered to be ‘molecular living fossils’ (see also Ref. [60]), paralleling at the genome level the relative morphological stasis they have exhibited for the past 200 million years [58].

Perspectives

There is no doubt that certain angiosperm lineages have been more successful than others in terms of species production, and several authors have documented these major shifts [61–67], although the factors responsible for increased rates of speciation remain unclear. Now that biodiversity is a major concern for society in general and biology in particular, perhaps only shared with human health, understanding the factors involved in its origin is fundamental. An examination of molecular rates in sister families of angiosperms showed that the more species-rich families have, on average, an increased rate of neutral substitutions in both plastid and nuclear genes [65]. In addition, the more diverse families in terms of morphology also have higher rates of DNA substitution [65], but this was not observed for animals [68]. This higher rate of background mutation (perhaps involving deficient DNA repair and exposure to mutagenic radiation) might affect developmental genes, thereby increasing morphological diversity (although alternative explanations are possible, especially regarding the effects of population size and structure on substitutions). This also holds for odd ecological niches with, for example, parasitism and carnivory in plants being associated with higher substitution rates [69].

Finally, it is clear that a decade of plant phylogenetics has resulted in major steps towards understanding the relationships between genes and species diversity. However, out of around 300 000 species of land plants, only 13 species account for over 81% of all plant nucleotides entries in EMBL and/or GenBank (genome data excluded, Figure 6). Large-scale sequencing projects can help explain the origins of phenotypic diversity [70] and, hopefully, intensive DNA sequencing of non-model organisms during the next decade will lead to new genetic syntheses, the phylogenomic era.

Acknowledgements

We thank Arne Mooers, Emmanuel Douzery, Peter Lockhart, several anonymous reviewers and the Editor for their comments on the manuscript.

References

1 Zurawski, G. *et al.* (1981) The structure of the gene for the large subunit of ribulose 1,5 bisphosphate carboxylase from spinach chloroplast DNA. *Nucleic Acids Res.* 9, 3251–3270
2 Chase, M.W. *et al.* (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. *Ann. MO Bot. Gard.* 80, 528–580
3 Pagel, M. (1999) Inferring historical patterns of biological evolution. *Nature* 401, 877–884
4 Kyrpides, N.C. and Olsen, G.L. (1999) Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry? *Trends Genet.* 15, 298–299
5 Wolf, Y.I. *et al.* (2002) Genome trees and the tree of life. *Trends Genet.* 18, 472–479
6 Hendy, M.D. and Penny, D. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38, 310–321
7 Huelsenbeck, J.P. (1995) Performance of phylogenetic methods in simulation. *Syst. Biol.* 44, 17–48
8 Hillis, D.M. (1996) Inferring complex phylogenies. *Nature* 383, 130–131
9 Hillis, D.M. (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* 47, 3–8
10 Huelsenbeck, J. and Hillis, D. (1993) Success of phylogenetic methods in the four-taxon cases. *Syst. Biol.* 42, 247–264
11 Källersjö, M. *et al.* (1998) Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Syst. Evol.* 213, 259–287
12 Källersjö, M. *et al.* (1999) Homoplasy increases phylogenetic structure. *Cladistics* 15, 91–93
13 Savolainen, V. *et al.* (2000) Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* 49, 306–362
14 Savolainen, V. *et al.* (2002) Phylogeny reconstruction and functional constraints in organellar genomes: plastid *atpB* and *rbcL* sequences versus animal mitochondrion. *Syst. Biol.* 51, 638–647
15 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer Associates Inc.
16 Yoder, A.D. and Yang, Z. (2000) Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* 17, 1081–1090
17 Savolainen, V. *et al.* (2000) Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bull.* 55, 257–309
18 Goremykin, V. *et al.* (1996) Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbcL* data do not support Gnetalean affinities of angiosperms. *Mol. Biol. Evol.* 13, 383–396
19 Soltis, D.E. *et al.* (1997) Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Ann. MO Bot. Gard.* 84, 1–49
20 Soltis, P.S. *et al.* (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402, 402–404
21 Soltis, D.E. *et al.* (2000) Angiosperm phylogeny inferred from a combined data set of 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* 133, 381–461
22 Mathews, S. and Donoghue, M.J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286, 947–950
23 Qiu, Y-L. *et al.* (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402, 404–407
24 Qiu, Y-L. *et al.* (2000) Phylogeny of basal angiosperms: analysis of five genes from three genomes. *Int. J. Plant Sci.* 161, S3–S27
25 Graham, W.W. and Olmstead, R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of basal angiosperms. *Am. J. Bot.* 87, 1712–1730
26 Soltis, D.E. *et al.* (1998) Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. *Syst. Biol.* 47, 32–42
27 Palmer, J.D. (1985) Evolution of chloroplast and mitochondrial DNA in plants and algae. In *Molecular Evolutionary Genetics* (MacIntyre, R.J., ed.), pp. 131–240, Plenum Press
28 Naylor, G.J.P. and Brown, W.M. (1997) Structural biology and phylogenetic estimation. *Nature* 388, 527–528
29 Steel, M. *et al.* (2000) Invariable sites models and their use in phylogeny reconstruction. *Syst. Biol.* 49, 225–232
30 Naylor, G.J.P. and Brown, W.M. (1998) *Amphioxus* mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* 47, 61–76
31 Lockhart, P.J. *et al.* (1998) A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* 15, 1183–1188
32 Lockhart, P.J. *et al.* (1999) Spectral analysis, systematic bias, and the evolution of chloroplast. *Mol. Biol. Evol.* 16, 573–576
33 Angiosperm Phylogeny Group (APG), (1998) An ordinal classification for the families of flowering plants. *Ann. MO Bot. Gard.* 85, 531–553
34 Angiosperm Phylogeny Group (APG), (2003) An update of the

angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* 141, 399–436

35 Chase, M.W. *et al.* (2000) Higher-level classification in the angiosperms: new insights from the perspective of DNA sequence data. *Taxon* 49, 685–704

36 Goremykin, V.V. *et al.* (2003) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* 20, 1499–1505

37 Donoghue, M.J. and Doyle, J.A. (2000) Seed plant phylogeny: demise of the anthophyte hypothesis? *Curr. Biol.* 10, R106–R109

38 Friis, E.M. *et al.* (2001) Fossil evidence of waterlilies (Nymphaeales) in the Early Cretaceous. *Nature* 410, 357–360

39 Sun, G. *et al.* (2002) Archaefructaceae, a new basal angiosperm family. *Science* 296, 899–904

40 Bennett, M.D. (1998) Plant genome values: how much do we know? *Proc. Natl. Acad. Sci. U. S. A.* 95, 2011–2016

41 Leitch, I.J. *et al.* (1998) Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Ann. Bot. (Lond.)* 82, 85–94

42 Palmer, J.D. and Delwiche, C.R. (1998) The origin and evolution of plastids and their genomes. In *Molecular Systematics of Plants II* (Soltis, D.E. *et al.*, eds), pp. 375–409

43 Cho, Y. *et al.* (1998) Explosive invasion of plant mitochondria by a group I intron. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14244–14249

44 Cho, Y.R. and Palmer, J.D. (1999) Multiple acquisitions via horizontal transfer of a group I intron in the mitochondrial cox1 gene during evolution of the Araceae family. *Mol. Biol. Evol.* 16, 1155–1165

45 Palmer, J.D. *et al.* (2000) Dynamic evolution of mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6960–6966

46 Bennetzen, J.L. (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115, 29–36

47 Wendel, J.F. *et al.* (2002) Feast and famine in plant genomes. *Genetica* 115, 37–47

48 Grant, V. (1971) *Plant Speciation*, Columbia Press

49 Adams, S.P. *et al.* (2001) Loss and recovery of *Arabidopsis*-type telomere repeat sequences 5′-(TTTAGGG)$_n$-3′ in the evolution of a major radiation of flowering plants. *Proc. R. Soc. London B Biol. Sci.* 268, 1541–1546

50 Zuckerlandl, E. and Pauling, L. (1962) Molecular disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry* (Kasha, M. and Pullman, B., eds), pp. 189–225, Academic Press

51 Heckman, D.S. *et al.* (2001) Molecular evidence for the early colonization of land by fungi and plants. *Science* 293, 1129–1133

52 Sanderson, M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14, 1218–1231

53 Wikström, N. *et al.* (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. London B Biol. Sci.* 268, 2211–2220

54 Goremykin, V.V. *et al.* (1997) Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. *Plant Syst. Evol.* 206, 337–351

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

56 Britton, T. *et al.* (2002) Phylogenetic dating with confidence intervals using mean path lengths. *Mol. Phylog. Evol.* 24, 58–65

57 Magallón, S. *et al.* (1999) Phylogenetic pattern, diversity, and diversification of eudicots. *Ann. MO Bot. Gard.* 86, 297–372

58 Soltis, P.S. *et al.* (2002) Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4430–4435

59 Pryer, K.M. *et al.* (2001) Horsetails and ferns are a monophyletic group and the closest living relatives to the seed plants. *Nature* 609, 618–622

60 Sanderson, M.J. and Doyle, J.A. (2001) Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. *Am. J. Bot.* 88, 1499–1516

61 Sanderson, M. and Donoghue, M. (1994) Shifts in diversification rate with the origin of angiosperms. *Science* 264, 1590–1593

62 Sanderson, M.J. and Wojciechowski, M.F. (1996) Diversification rates in a temperate legume clade – why are there so many species of *Astragalus* (Fabaceae). *Am. J. Bot.* 83, 1488–1502

63 Barraclough, T.G. *et al.* (1996) Rate of *rbcL* gene sequence evolution and species diversification in flowering plants (angiosperms). *Proc. R. Soc. London B Biol. Sci.* 263, 589–591

64 Savolainen, V. and Goudet, J. (1998) Rate of gene sequence evolution and species diversification in flowering plants: a re-evalutation. *Proc. R. Soc. London B Biol. Sci.* 265, 603–607

65 Barraclough, T.G. and Savolainen, V. (2001) Evolution rates and species diversity in flowering plants. *Evolution* 55, 677–683

66 Magallón, S. and Sanderson, M.J. (2001) Absolute diversification rates in angiosperm clades. *Evolution* 55, 1762–1780

67 Savolainen, V. *et al.* (2002) Is cladogenesis heritable? *Syst. Biol.* 51, 835–843

68 Bromham, L. *et al.* (2002) Testing the relationship between morphological and molecular rates of change along phylogenies. *Evolution* 56, 1921–1930

69 Jobson, R.W. and Albert, V.A. (2002) Molecular rates parallel diversification contrasts between carnivorous plant sister lineages. *Cladistics* 18, 127–136

70 Soltis, D.E. *et al.* (2002) Missing links: the genetic architecture of flower and floral diversification. *Trends Plant Sci.* 7, 22–31

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

Trends in Genetics Editorial Policy

Trends in Genetics is a forum for all aspects of current research in genetics, genomics and development.

Review, Opinion and Research Focus articles are commissioned, peer reviewed and edited to ensure that they are timely, accurate and relevant to the audience of *TiG*. With the exception of the Genome Analysis section, articles in *TiG* are generally invited by the Editor, but proposals are welcomed. We also encourage ideas for Opinion articles and Meeting Reports, as well as Letters on topics of interest to geneticists and developmental biologists.

Prospective authors should send a brief summary (maximum one page) including the article type, expected length and key references to the Editor, who will supply guidelines on manuscript preparation if the proposal is accepted. **The submission of completed articles without consultation is strongly discouraged**. Authors should also note that all opinion and review articles for *TiG* are peer reviewed before acceptance and publication cannot be guaranteed.

Genome Analysis articles include original observations concerning the function, organization and evolution of genomes. All sequences used in the analysis must be available through GenBank. Authors wishing to publish a Genome Analysis paper should send a synopsis of their proposed article directly to the Advisory Editor: Dr E.V. Koonin at National Center for Biotechnology, NIH, Bethesda, MD 20894, USA (e-mail: koonin@ncbi.nlm.nih.gov).

For further information, please contact:
The Editor, *Trends in Genetics*, Elsevier London, 84 Theobald's Road, London, UK WC1X 8RR.
Tel: +44 (0)20 7611 4400; Fax: +44 (0)20 7611 4470; e-mail: tig@current-trends.com