12 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer Associates
13 Huelsenbeck, J.P. and Crandall, K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28, 437–466
14 Lewis, P.O. (1998) Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In *Molecular Systematics of Plants II* (Soltis, D.E. *et al.*, eds), pp. 132–163, Kluwer
15 Swofford, D.L. (2000) *PAUP*, Phylogenetic Analysis Using Parsimony* (Sinauer, Sunderland, MA), Version 4.0b3,
16 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556
17 Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724
18 Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736
19 Muse, S.V. and Kosakovsky Pond, S.L. (2000) *HYPHY, Hypothesis Testing Using Phylogenies*, (North Carolina State University, Raleigh) Version 1
20 Tillier, E.R.M. and Collins, R.A. (1995) Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Mol. Biol. Evol.* 12, 7–15
21 Muse, S.V. (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* 139, 1429–1439
22 Schluter, D. *et al.* (1997) Likelihood of ancestor states in adaptive radiation. *Evolution* 51, 1699–1711
23 Mooers, A.Ø. and Schluter, D. (1999) Reconstructing ancestor states with maximum likelihood: support for one- and two-rate models. *Syst. Biol.* 48, 623–633
24 Pagel, M. (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* 48, 612–622
25 Cunningham, C.W. *et al.* (1998) Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.* 13, 361–366
26 Pagel, M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. London B Biol. Sci.* 255, 37–45
27 Qiang, J. *et al.* (1998) Two feathered dinosaurs from northeastern China. *Nature* 393, 753–761
28 Lutzoni, F. and Pagel, M. (1997) Accelerated evolution as a consequence of transitions to mutualism. *Proc. Natl. Acad. Sci. U. S. A.* 94, 11422–11427
29 Larget, B. and Simon, D.L. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16, 750–759
30 Mau, B. and Newton, M.A. (1997) Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 6, 122–131
31 Mau, B. *et al.* (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* 55, 1–12
32 Newton, M. *et al.* (1999) Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. In *Statistics in Molecular Biology* (Seillier-Moseiwitch, F. *et al.*, eds), Institute of Mathematical Statistics
33 Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* 43, 304–311
34 Yang, Z.H. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.* 14, 717–724
35 Metropolis, N. *et al.* (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092
36 Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109
37 Huelsenbeck, J.P. *et al.* (2000) A compound Poisson process for relaxing the molecular clock. *Genetics* 154, 1879–1892
38 Thorne, J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15, 1647–1657
39 Felsenstein, J. (1985) Confidence intervals on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791

Intraspecific gene genealogies: trees grafting into networks

David Posada and Keith A. Crandall

Intraspecific gene evolution cannot always be represented by a bifurcating tree. Rather, population genealogies are often multifurcated, descendant genes coexist with persistent ancestors and recombination events produce reticulate relationships. Whereas traditional phylogenetic methods assume bifurcating trees, several networking approaches have recently been developed to estimate intraspecific genealogies that take into account these population-level phenomena.

During the past decade, the explosion of molecular techniques has led to the accumulation of a considerable amount of comparative genetic information at the population level. At the same time, recent advances in population genetics theory, especially coalescent theory, have generated powerful tools for the analysis of intraspecific data. These two developments have converted intraspecific phylogenies into useful tools for testing a variety of evolutionary and population genetic hypotheses. Several phylogenetic methods, especially NETWORK (see Glossary) approaches, have been developed to take advantage of the unique characteristics of intraspecific data. In this article, we summarize some population genetics principles, explain why networks are appropriate representations of intraspecific genetic variation, describe and compare available methods and software for network estimation, and give examples of their application.

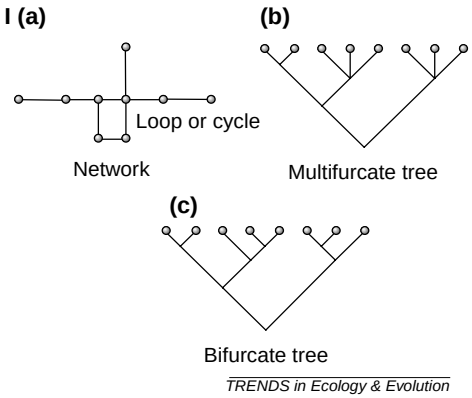
Gene genealogies

Given a sample of GENES, the relationships among them can be traced back in time to a common ancestral gene. The genealogical pathways interconnecting the current sample to the common ancestor constitute a GENE TREE or gene genealogy. A gene tree is the pedigree of a set of genes and exists independently of potential mutations. The only portion of a gene tree that can generally be estimated with genetic data is that portion marked by the (potential) mutational events that define the different ALLELES (Box 1). This lower resolution tree is the allele

http://tree.trends.com 0169–5347/01/$ – see front matter PII: S0169-5347(00)02026-7

Glossary

Additive tree: a tree on which the pairwise distances between haplotypes are equal to the sum of the lengths of the branches on the path between the members of each pair.
Coalescent event: the time inverse of a DNA replication event; that is, the event leading to the common ancestor of two sequences looking back in time.
Gene: a segment of DNA.
Gene tree: the representation of the evolutionary history of a group of genes.
Haplotype or allele: a unique combination of genetic markers present in a sample.
Haplotype space: the collection of points representing the possible different haplotypes. The dimension of this space is the number of characters (L). The number of points (haplotypes) included in the haplotype space is the number of states raised to the number of characters (i.e. for DNA sequences 4^L).
Homoplasy: a similarity that is not a result of common history. It is caused by parallel, convergent or reverse mutations.
Interior haplotypes: those haplotypes that have more than one mutational connection.
Minimum-spanning tree: graph theory construct that connects the n haplotypes, thus a complete network of n-1 branches is built. The tree is 'minimal' when the total length of the branches is the minimum necessary to connect all the haplotypes.
Missing intermediate: an extant haplotype that was not sampled or an extinct ancestral haplotype.
Network: a connected graph with cycles (Fig. Ia).
Patristic or phyletic distances: the distance between haplotypes as inferred from the network. It does not have to be equal to the actual distance (number of differences) between haplotypes.
Phylogeny: hierarchical genetic relationships among species. Arising by speciation.
Robinsonian distance matrix: a matrix with the property that, for any ordered triplet of sequences i, j and k, any distance $d_{ik} \geq \max(d_{ij}, d_{jk})$.
Singletons: haplotypes represented by a single sequence in the sample.
Species tree: the representation of the evolutionary history of a group of species.
Split: the division of the haplotypes into two exclusive sets. For n sequences, there are $2n$-1 possible splits.
Tip haplotypes: those haplotypes that have only a single mutational connection to the other haplotypes within the network.
Tokogeny: nonhierarchical genetic relationships among individuals. Arising by sexual reproduction.
Tree: a connected graph without cycles (cycles are commonly called reticulations or loops by evolutionary biologists) (Fig. Ib,c).



or HAPLOTYPE tree. Gene genealogies are approximated by the estimation of haplotype or allele trees. Given the abundance of methods available for phylogenetic estimation[1], which ones are most appropriate for estimating haplotype trees?

Problems with interspecific methods at the intraspecific level

Evolutionary relationships above and below the species level are different in nature. Relationships between genes sampled from individuals belonging to different species (phylogeny *sensu stricto*) are hierarchical. This is because they are the product of reproductive isolation and population fission over longer timescales, during which mutation combined with population divergence led to the fixation of different alleles and, ultimately, to nonoverlapping gene pools.

By contrast, relationships between genes sampled from individuals within a species (sometimes called TOKOGENY[2]) are not hierarchical, because they are the result of sexual reproduction, of smaller numbers of relatively recent mutations and, frequently, of recombination (Fig. 1). More traditional methods developed to estimate interspecific relationships, such as maximum likelihood, maximum parsimony and minimum evolution, cannot properly take account of the fact that, at the population level, several phenomena violate some of their assumptions. This leads to poor resolution or inadequately portrays genealogical relationships.

Low divergence

By necessity, conspecific individuals diverge later than individuals from different species. Consequently, within-species data sets have fewer characters for phylogenetic analysis, diminishing the statistical power of traditional phylogenetic methods.

Extant ancestral nodes

In natural populations, most haplotypes in the gene pool exist as sets of multiple, identical copies that originated by DNA replication. When one of these copies mutates to a new haplotype, it is extremely unlikely that other copies of the ancestral haplotype also mutate or that all copies of the ancestral haplotype rapidly become extinct. Thus, the ancestral haplotypes are expected to persist in the population and to be sampled together with their descendants. Traditional phylogenetic methods, based on a bifurcating TREE, can detect and artificially represent persistent ancestral haplotypes as occupying a branch of zero length at the basal node of a cluster. However, this approach relies on modifying (e.g. by estimation of branch lengths) an inappropriate model – a bifurcating tree with all haplotypes occupying tips or terminal branches.

Multifurcations

A fact related to the persistence of ancestral haplotypes in the population is that a single ancestral haplotype will often give rise to multiple descendant haplotypes,

David Posada*
Keith Crandall
Dept of Zoology, Brigham Young University, 574 WIDB, Provo, UT 84602, USA.
*e-mail: david.posada@byu.edu

Box 1. Gene trees, haplotype trees and population trees

Gene trees versus haplotype trees

Gene trees are independent of the (neutral) mutation process.They depend on demographic factors such as population size and geographical structure. Because, by definition, neutral mutations do not affect the number of offspring or migration patterns, they do not affect the gene genealogy. Gene trees precisely represent the gene genealogy of a given sample. However, such precise information is usually unknown or even impossible to extract from a sample of extant genes. For example, in Fig. I, there is no way to know that gene A1 is more closely related to A4 than either is to A2. Unless detailed pedigree information is available, which usually is not the case, the only branches of the gene tree that we can estimate are those marked by a mutation and that therefore define haplotypes.

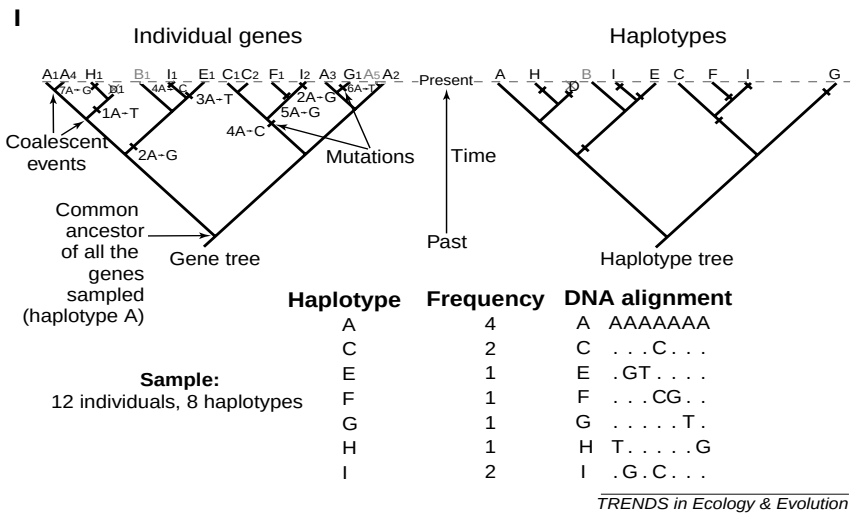
We must be able to see genetic differences (mutations) to determine relationships; therefore, we use haplotype trees most often. With haplotype trees, we cannot see all the coalescent events and can only group genes by their similarities and haplotype classes. Whereas traditional methods often lack the power to solve intraspecific relationships, network approaches offer an appropriate representation of the haplotype relationships, including extinct (Fig. I, haplotype D) or unsampled (Fig. I, haplotype B) haplotype variants.

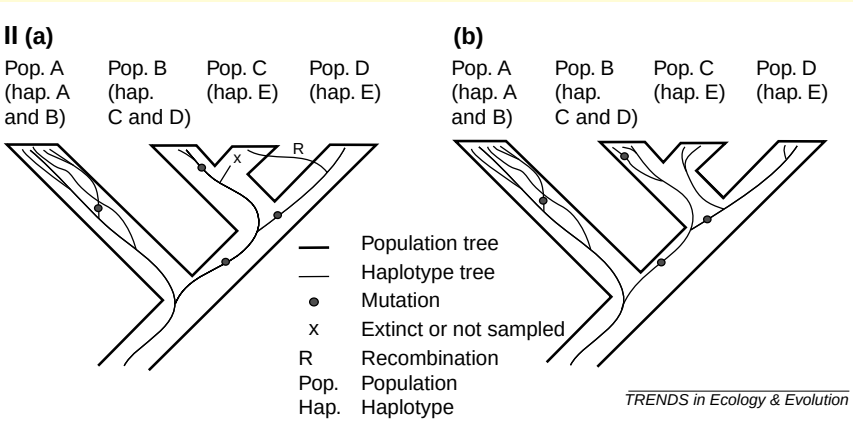
Haplotype trees versus population trees

It is often assumed that the haplotype trees represent exactly the history of the populations sampled. However, haplotypes do not, in general, have the same evolutionary history as population lineages[a]. Disagreement among haplotype trees and population trees can arise from recombination (Fig. IIa) and deep coalescence or lineage sorting (Fig. IIb).

Reference

a Hudson, R.R. (1990) Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7, 1–44





yielding a haplotype tree with true multifurcations. Indeed, population genetics theory predicts that the older the haplotype, the more descendant haplotypes will be associated with it (Box 2).

Reticulation

Evolutionary processes commonly acting at the population level, such as recombination between genes and hybridization between lineages, and HOMOPLASY (Box 1), generate reticulate relationships within the population. Traditional methods, based on bifurcating trees, make no explicit allowance for such reticulations. Instead, for instance, maximum parsimony deals with ambiguities arising from homoplasy by simply selecting a tree that minimizes the number of assumptions of parallel, convergent or reversing mutations without showing where these might have occurred. Recombinants are also typically forced into a nonreticulating tree topology, in which, in some fortunate instances, they might occupy positions intermediate between two clusters. In other cases, the recombinant will be placed in a basal lineage to the clade that includes its most derived parent[3,4].

Large sample sizes

Appropriate sampling can be crucial to phylogenetic studies[5]. Typically, intraspecific studies involve many individuals for comparison, whereas many interspecific phylogenetic studies tend to be based on one representative individual per species. Because of the density of sampling, especially when coupled with low divergence, intraspecific data sets reach considerable sample sizes (>100). These large sample sizes would require excessive computational time for

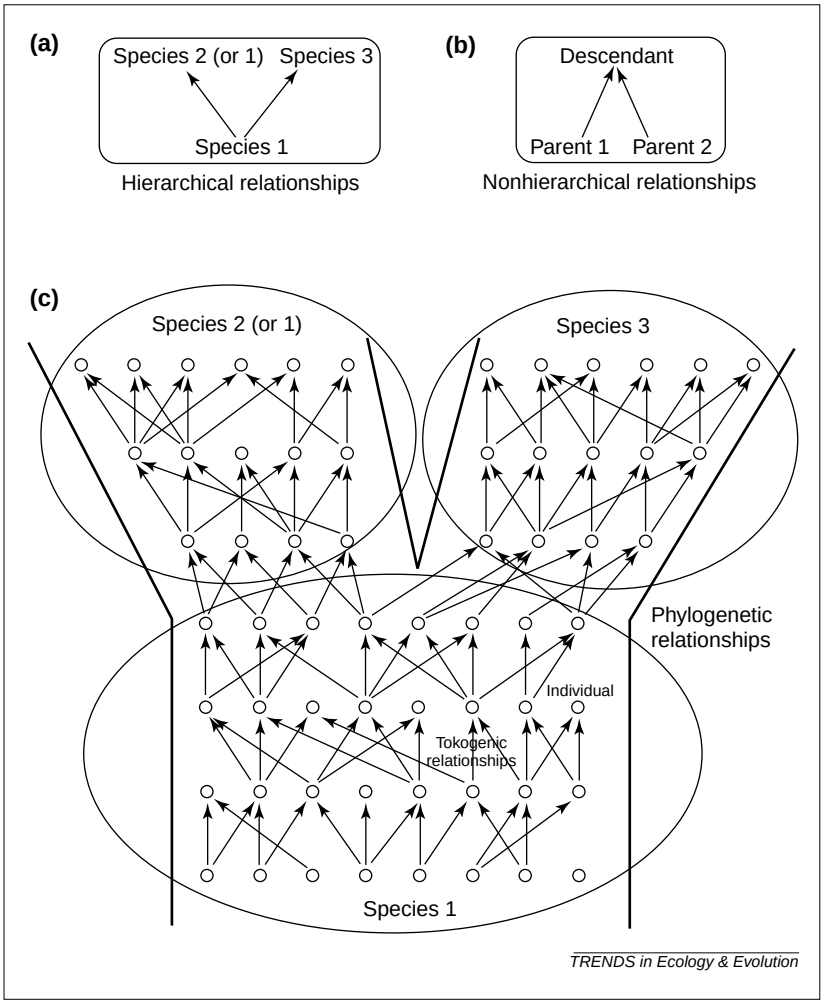


Fig. 1. Tokogeny versus phylogeny. (a) Processes occurring among sexual species (phylogenetic processes) are hierarchical. That is, an ancestral species gives rise to two descendant species. (b) Processes occurring within sexual species (tokogenetic processes) are nonhierarchical. That is, two parentals combine their genes to give rise to the offspring. (c) The split of two species defines a phylogenetic relationship among species (thick lines) but, at the same time, relationships among individuals within the ancestral species (species 1) and within the descendant species (species 2 and 3) are tokogenetic (arrows).

most methods that have been developed for interspecific comparisons.

Solution: network methods

Phylogenetic methods that allow for persistent ancestral nodes, multifurcations and reticulations are needed to take these population phenomena into account. The advantage of networks over strictly bifurcating trees for estimating within-species relationships now becomes obvious. Networks can account effectively for processes acting at the species level and they might be able to incorporate predictions from population genetics theory (Box 2). In addition, networks provide a way of representing more of the phylogenetic information present in a data set (Fig. 2). For example, the presence of loops in a network might indicate recombination. In other cases, loops are the product of homoplasies and precisely indicate the occurrence of reverse or parallel mutations. Most network methods are distance methods, with the common idea of minimizing (with some specific restrictions) the distances (number of mutations) among haplotypes. In other cases, the likelihood function is maximized.

Pyramids

The pyramids technique[6] is an extension of the hierarchical clustering framework. Whereas traditional hierarchical methods such as Unweighted Pair Group with Arithmetic Means (UPGAM) represent a nested set of nonoverlapping clades, pyramids represent a set of clades that can overlap without necessarily being nested. The input data is a (Robinsonian) distance matrix (Box 1). The pyramid is obtained by using agglomerative algorithms. By allowing overlapping clusters, pyramids can be used to represent reticulate events, although these events are only allowed to be placed among terminal nodes that are sister taxa.

Statistical geometry

This was one of the first network approaches for intraspecific phylogenetics to be developed[7]. In this method, haplotypes are considered as geometric configurations in the HAPLOTYPE SPACE. Numerical invariants (some function of the data whose value remains constant) related to the length of the connections (pairwise differences) among haplotypes are assigned to the haplotypes and statistical averages of these invariants are then calculated. Given the value for the invariants, the optimal network connecting the four haplotypes in each quartet is derived. Finally, an average quartet geometry representative of the whole data set is constructed by integrating all the quartet networks. This geometry and the associated statistics can be used, for example, to deduce the degree of tree-likeness of the data, to detect varying positional substitution rates in sequences or to estimate the relative temporal order of haplotype divergence. However, they do not offer an estimate of the sequence genealogy. The statistical geometry incorporates a model of nucleotide substitution through the estimation of haplotype distances and its statistical nature allows a reliable assessment of the derived conclusions.

Split decomposition

Any data set can be partitioned into sets (not necessarily of equal size) of sequences or 'splits'. A network can be built by taking in turn those splits defined by the characters and combining them successively[8]. Each split will define a branch connecting the two partitions delimited by the split. When splits are incompatible (i.e. they define contradictory groupings) a loop is introduced to indicate that there are alternative splits. The split decomposition method is fast, which means that a reasonable number of haplotypes (>50) can be analyzed; that it can be applied to nucleotide or protein data; and that it allows for the inclusion of

Box 2. Predictions from coalescent theory and application to intraspecific phylogenetics

The ancestry of a random sample of *n* genes is often modeled by a stochastic process known as the coalescent. Coalescent theory describes the genealogical process of a sample of selectively neutral genes from a population, looking backwards in time[a]. Several results from coalescent theory related to the frequency and geographical distribution of the haplotypes are relevant to intraspecific phylogenetics.

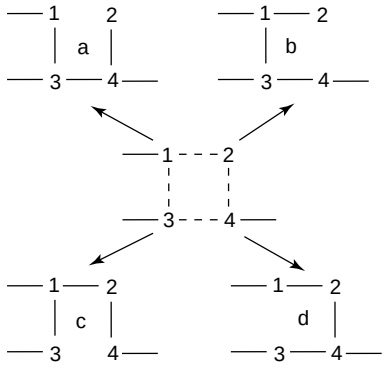
There is a direct relationship between haplotype frequencies and the ages of the haplotypes[b,c]. Specifically, the probability that an allele represented n_i times in a sample of size *n* is the oldest allele in the sample is n_i/n, and the expected rank of the alleles by age is the same as the rank of alleles by frequency. Therefore, high-frequency haplotypes have probably been present in the population for a long time. Consequently, most of the new mutants are derived from common haplotypes, implying that rarer variants represent more recent mutations and are more likely to be related to common haplotypes than to other rare variants[d]. The expectation for the number of alleles in common between the samples[e] implies that the immediate descendents of a new mutation are more likely to remain in the original population than to move to a distant population, unless high levels of gene flow occur. These results can be summarized in five explicit predictions.

- Older alleles (those of higher frequency in the population) have a greater probability of becoming interior haplotypes (those haplotypes that have more than one mutational connection) than younger haplotypes.
- On average, older alleles will be more broadly distributed geographically.
- Haplotypes with greater frequency will tend to have more mutational connections.
- Singletons are more likely to be connected to nonsingletons than to other singletons.
- Singletons are more likely to be connected to haplotypes from the same population than to haplotypes from different populations.

These theoretical predictions make intuitive sense and have been shown to be valid in empirical data sets[f]. However, they assume neutral evolution (and lack of population subdivision) and might not be accurate in the presence of selection. These predictions can also be used to root the network. Some loops or reticulations in the networks can be the result of homoplasies, thus representing the consequence of a lack of power to decide among alternative connections. These loops could be broken at any place, resulting in different networks.

The above predictions can be used to establish which one of the alternative networks is more plausible[f,g]. For example, in Fig. I, if the frequency of haplotype 2 is lower than the frequency of the other haplotypes, resolutions a and b are favored because they result in haplotype 2 as a tip. (Dotted lines represent ambiguous connections.)

I



TRENDS in Ecology & Evolution

References

a Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7, 1–44

b Watterson, G.A. and Guess, H.A. (1977) Is the most frequent allele the oldest? *Theor. Popul. Biol.* 11, 141–160

c Donnelly, P. and Tavaré, S. (1986) The ages of alleles and a coalescent. *Adv. Appl. Prob.* 18, 1–19

d Excoffier, L. and Langaney, A. (1989) Origin and differentiation of human mitochondrial DNA. *Am. J. Hum. Genet.* 44, 73–85

e Watterson, G.A. (1985) The genetic divergence of two populations. *Theor. Popul. Biol.* 27, 298–317

f Crandall, K.A. and Templeton, A.R. (1993) Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* 134, 959–969

g Smouse, P. (1998) To tree or not to tree. *Mol. Ecol.* 7, 399–412

models of nucleotide substitution or amino acid replacement. The method is suitable also to bootstrap evaluation.

Median networks

In the median-network approach[9,10], sequences are first converted to binary data and constant sites are eliminated. Each split is encoded as a binary character with states 0 and 1. Sites that support the same split are grouped in one character, which is weighted by the number of sites grouped. This leads to the representation of haplotypes as 0–1 vectors. Median or consensus vectors are calculated for each triplet of vectors until the median network is finished. For >30 haplotypes, the resulting median networks are impractical to display, owing to the presence of high-dimensional hypercubes. Fortunately, the network can be reduced (i.e. some loops can be solved) using predictions from coalescent theory (Box 2). All the most parsimonious trees are guaranteed to be represented in a median network. Although mainly aimed at mtDNA data, median networks can be estimated from other kinds of data, as long as the data are binary or can be reduced to binary data.

Median-joining networks

The median-joining network method[11,12] begins by combining the MINIMUM-SPANNING TREES (MSTs) within a single network. With a parsimony criterion, median vectors (which represent MISSING INTERMEDIATES) are added to the network. Median-joining networks can handle large data sets and multistate characters. It is an exceptionally fast method that can analyze thousands of haplotypes in a reasonable amount of time and can also be applied to amino acid sequences. However, it requires the absence of recombination, which restricts the application of this method at the population level.

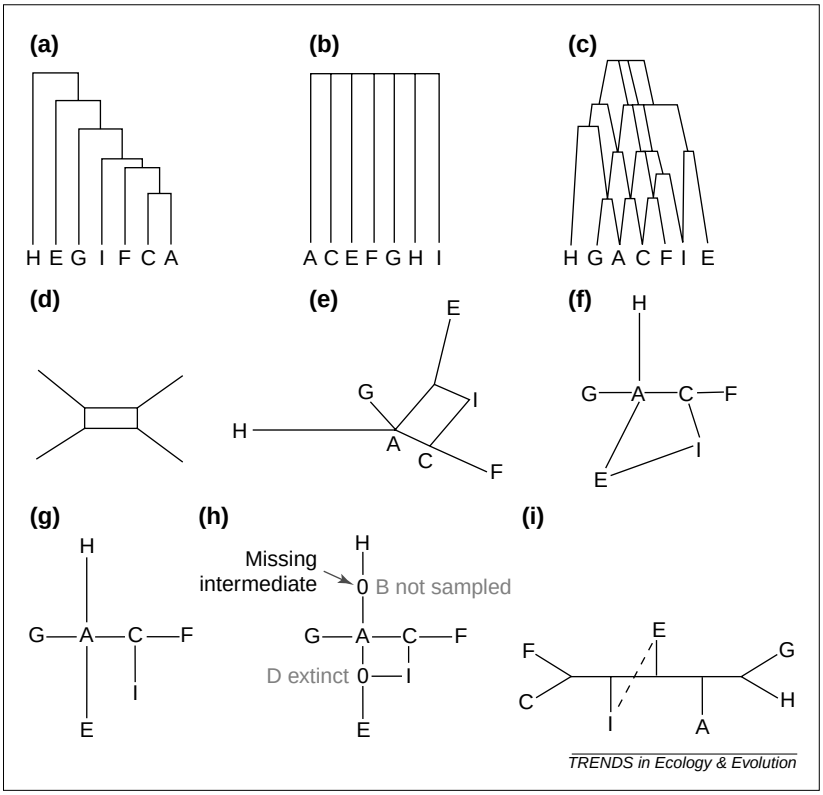


Fig. 2. Phylogenetic estimation. Different phylogenetic reconstruction techniques, including network techniques, were applied to the data set described in Box 1, Fig. I. (a) UPGMA. (b) Maximum parsimony. (c) Pyramid. (d) Statistical geometry (provides an estimate of average quartet topology rather than of the actual genealogy). (e) Split decomposition. (f) Minimum spanning network. (g) Median-joining network. (h) Statistical parsimony. (i) Reticulogram.

Statistical parsimony

The statistical parsimony algorithm[13] begins by estimating the maximum number of differences among haplotypes as a result of single substitutions (i.e. those that are not the result of multiple substitutions at a single site) with a 95% statistical confidence. This number is called the parsimony limit (or parsimony connection limit). After this, haplotypes differing by one change are connected, then those differing by two, by three and so on, until all the haplotypes are included in a single network or the parsimony connection limit is reached. The statistical parsimony method emphasizes what is shared among haplotypes that differ minimally rather than the differences among the haplotypes and provides an empirical assessment of deviations from parsimony. This method allows the identification of putative recombinants by looking at the spatial distribution in the sequence of the homoplasies defined by the network[14].

Molecular-variance parsimony

The overall strategy in the molecular-variance parsimony technique[15] is to use some population statistics as criteria for the choice of the optimal network. Each competing MST is translated into a matrix of PATRISTIC DISTANCES among haplotypes. These matrices are used to compute a set of relevant population statistics: functions of haplotype frequencies, squared patristic distances among haplotypes and geographic partitioning of populations. The optimal MSTs are those from which optimum estimates of population statistics are obtained (e.g. minimizing the molecular variance or the sum of square deviations). This method makes explicit use of sampled haplotype frequencies and geographic subdivisions, and presents the solution in the form of a set of (near) optimal networks.

Netting

This is a distance method that represents all the equally most parsimonious trees for a given data set in a single network[16]. The underlying idea is to join the closest pair of sequences (the pair with the fewest differences). The next sequence that is closest to the first two is joined so that the three pairwise differences are satisfied. Thus, patristic distances necessarily equal the number of differences. If a homoplasy is encountered, a new spatial dimension is added to the graph. Gaps and invariant positions are excluded from the analysis. Because the method tries to satisfy all the distances among haplotypes, the number of dimensions might be high and the display of the network thus becomes complex and difficult. This problem becomes worse as data sets become larger.

Likelihood network

The likelihood-network procedure[17] is based on a directed graphical model for the evolution of sequences along a network. Graphical models are graphs in which nodes are stochastic variables, whereas branches indicate correlations between these

Table 1. Properties of networking algorithms

Methods	Category[a]	Software	Speed	Input data	Model of evolution	Reticulations	Statistical assessment
Pyramids	Distance	Pyramids	Fast	Distances	Yes	Yes	No
Statistical geometry	Distance invariants	Geometry, Statgeom	Fast	Multistate	Yes	Yes	Yes
Split decomposition	Distance parsimony	SplitsTree	Fast	Multistate	Yes	Yes	Yes
Median networks	Distance	No	Slow	Binary	No	Yes	No
Median-joining networks	Distance	Network	Very fast	Multistate	No	No	No
Statistical parsimony	Distance	TCS	Fast	Multistate	No	Yes	Yes
Molecular-variance parsimony	Distance	Arlequin	Fast	Multistate	Yes	Yes	Yes
Netting	Distance	No	Slow	Multistate	No	Yes	No
Likelihood network	Likelihood	PAL	Slow	Multistate	Yes	Yes	Yes
Reticulogram	Least squares	T-rex	Fast	Distances	No	Yes	Yes
Reticulate phylogeny	Least squares	No	Slow	Distances[b]	Yes	Yes	Yes

[a]Details of software programs given in Box 3; [b]Distances estimated from gene frequency data.

Box 3. Software availability

Several software packages implement some of the network methods described in this article.

- PYRAMIDS estimates pyramids from distance matrices. Executables for Windows and Unix are available (http://genome.genetique.uvsq.fr/Pyramids/).There is also a convenient online implementation (http://www.bioweb.pasteur.fr/seqanal/interfaces/pyramids.html). PYRAMIDS was written by J.C. Aude *et al.* (INRA, France).
- STATGEOM calculates the statistical geometry in distance and in sequence space of a set of aligned DNA, RNA, amino acid or binary sequences. Source code with documentation and a Sun SPARC executable are available (http://gwdu17.gwdg.de/~kniesel/statgeom.html). STATGEOM was written by K. Nieselt-Struwe (Dept of Physics, University of Auckland, New Zealand).
- GEOMETRY is a package for nucleotide sequence analysis using the method of statistical geometry in sequence space[a]. The program is available as a DOS executable (http://molevol.bionet.nsc.ru/soft.htm; ftp://ftp.bionet.nsk.su/incoming/molevol/; ftp://ftp.ebi.ac.uk/pub/software/dos/).
- SPLITSTREE implements the split decomposition method[b].This program is currently available as a Mac executable or as a Unix version (ftp://ftp.uni-bielefeld.de/pub/math/splits/).
- NETWORK is a program for estimating median-joining networks.The program is available as a DOS executable (http://www.fluxusengineering.com/sharenet.htm) and was written by A. Röhl (Mathematisches Seminar, Universität Hamburg, Germany).
- ARLEQUIN is a Java program that implements the method of molecular variance parsimony (http://lgb.unige.ch/arlequin/index.php3). ARLEQUIN was written by S. Scheneider *et al.* (Dept of Anthropology, University of Geneva, Switzerland).
- TCS is a Java program for estimating statistical parsimony networks[c] (http://bioag.byu.edu/zoology/crandall_lab/tcs.htm). It allows the estimation of root probabilities.
- PAL is an open-source Java library for molecular evolution and phylogenetics (http://www.pal-project.org/). Its >120 modules allow the fast prototyping of special-purpose analysis programs. PAL has been used to compute likelihoods for networks.The PAL project is led by K. Strimmer (Dept of Zoology, University of Oxford, UK) and A. Drummond (School of Biological Science, University of Auckland, New Zealand).
- T-REX is a C++ program implementing the reticulogram estimation[d]. Windows, Macintosh and DOS executables are available (http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex/index.html).

References

a Kuznetsov, I. and Morozov, P. (1996) GEOMETRY: a software package for nucleotide sequence analysis using statistical geometry in sequence space. *Comput. Appl. Biosci.* 12, 297–301

b Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73

c Clement, M. *et al.* (2000) TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1660

d Makarenkov, V. and Legendre, P. (2000) Improving the additive tree representation of a dissimilarity matrix using reticulations. In *Data Analysis, Classification and Related Methods* (Kiers, H.A.L. *et al.*, eds), pp. 35–46, Springer

variables. A graph might be directed by rooting the network at a specific node and directing all branches away from this node. A given network can be turned into directed acyclic graphs, which allows the computation of its likelihood. Arbitrary phylogenetic networks are evaluated by likelihood and the network with the best likelihood is chosen as the final solution.

Within this framework, the use of ancestral recombination graphs[18] when recombination is present has recently been suggested[19]. The likelihood-network algorithm also allows the simultaneous estimation of parameters (e.g. the recombination rate). The use of a likelihood framework offers the possibility of hypothesis testing, and the statistical comparison of both trees and networks. However, a difference compared with the other methods is that the amount of computing time can be excessively large. Effective search strategies need to be devised before large data sets can be analyzed.

Reticulogram

This procedure[20] is based on the addition of reticulations to a bifurcating tree. The method uses as input a distance matrix and an ADDITIVE TREE inferred from the same distance matrix using one of the classical reconstruction algorithms. An optimality criterion is used to estimate the minimum number of reticulations required to maximize the fit of the network to the data – a least squares loss function computed as the sum of the squared differences between the original and the patristic distances. The minimum of this criterion provides a stopping rule for the addition of reticulation. The computations are repeated recursively for all pairs of nodes (except those already connected) in the network to obtain a globally optimal solution.

Reticulate phylogenies from gene frequencies

This method infers reticulate phylogenies using genetics distances estimated from gene frequency data[21]. The mean squared error (MSE) of a least squares function is used as the optimality criterion to select among possible reticulate phylogenies. Theoretically, all possible phylogenies must be evaluated, and the inferred phylogeny is the one that has the minimum MSE. The use of a least-squares function allows the estimation of the branch lengths in the phylogeny. Two models are defined: a drift model and an extended model with mutation. The drift model is best used in short-term evolution, in which mutation has not played an important role, whereas the mutation model is applicable to long-term evolution.

Strengths and weaknesses of network methods

The comparison of different networking strategies is not straightforward. Comprehensive simulation studies are needed to evaluate the

accuracy and robustness of different methods. In general, distance methods might imply a loss of information by summarizing the difference between haplotypes with one value. Optimality criterion methods (likelihood and least squares) offer a reliable statistical assessment, hypotheses testing and, in some cases, parameter estimation. Most of these methods are conveniently implemented in computer programs, which are in some cases capable of handling many sequences. Several properties of different methods are summarized in Table 1.

Rooting intraspecific phylogenies

In many cases, the problem of biological interest requires a root, or at least some knowledge of the relative ages of haplotypes[22]. Rooting networks is especially difficult because outgroups are often separated from the ingroup by many mutational steps and because individuals within a species are similar to each other. This leads to a lack of power to decide where the rooting should take place. However, predictions from coalescent theory can be applied in intraspecific phylogeny reconstruction to root the network. By definition, the oldest ancestral haplotype is the root of the phylogeny. Given coalescent criteria, this ancestral haplotype is identified as the most frequent. The number of connections and the position of a haplotype in the network can also be used to assign root probabilities. The likelihood-network method provides a way to root a phylogenetic network by choosing a node that produces the most likely network.

Conclusions and future directions

Traditional methods for estimating phylogenies were not designed and might not be adequate for within-species phylogeny. Network approaches can incorporate population processes in the construction or refinement of haplotype relationships. Moreover, networks allow a more detailed display of populational information than strictly bifurcating trees. Although we have focused on the application of networks to sequence data, most methods described here can be applied to proteins or restriction-fragment length polymorphism data. In general, the main interest of intraspecific phylogenies is not in themselves but rather in their applications. They have been used for detecting recombination[23,24], delimiting species[25], inferring modes of speciation[26], partitioning population history and structure[27], and studying genotype and phenotype associations[28].

The development of networks in a likelihood framework, which allows hypothesis testing with a sound statistical basis, is particularly interesting. However, the computational tractability of likelihood networks remains a drawback. Coalescent theory is an active area of research that is likely to yield new predictions that could be used in the development of more refined network approaches, which currently are not based on the coalescent. Most of the methods described here are implemented in computer programs (BOX 3), and other network methods have been reviewed recently[29]. Given the recent interest in intraspecific phylogenies and their applications, we expect to see increased interest in the development and application of intraspecific phylogenetics using network approaches to depict genealogical relationships. Future work will involve comparing methods and testing the robustness of their assumptions.

Acknowledgements
Comments by Chris Simon and François-Joseph Lapointe improved this review. We thank M. Coulthart, A. Whiting, M. Porter and J.W. Sites Jr for helpful suggestions. Our work was supported by NSF-DEB 0073154, the Alfred P. Sloan Foundation and NIH R01-HD34350.

References

1 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer Associates
2 Hennig, W. (1966) *Phylogenetic Systematics*, University of Illinois Press
3 McDade, L. (1990) Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. *Evolution* 44, 1685–1700
4 McDade, L.A. (1992) Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* 46, 1329–1346
5 Lecointre, G. *et al.* (1993) Species sampling has a major impact on phylogenetic inference. *Mol. Phylogenet. Evol.* 3, 205–224
6 Diday, E. and Bertrand, P. (1986) An extension of hierarchical clustering: the pyramidal representation. In *Pattern Recognition in Practice* (Gelsema, E.S. and Kanal, L.N., eds), pp. 411–424, North–Holland
7 Eigen, M. *et al.* (1988) Statistical geometry in sequence space: a method of quantitative sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* 85, 5917
8 Bandelt, H-J. and Dress, A.W.M. (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* 1, 242–252
9 Bandelt, H-J. *et al.* (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141, 743–753
10 Bandelt, H-J. *et al.* (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phylogenet. Evol.* 16, 8–28
11 Bandelt, H-J. *et al.* (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37
12 Foulds, L.R. *et al.* (1979) A graph theoretic approach to the development of minimal phylogenetic trees. *J. Mol. Evol.* 13, 127–149
13 Templeton, A.R. *et al.* (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data III. Cladogram estimation. *Genetics* 132, 619–633
14 Crandall, K.A. and Templeton, A.R. (1999) Statistical methods for detecting recombination. In *The Evolution of HIV* (Crandall, K.A., ed.), pp. 153–176, The Johns Hopkins University Press
15 Excoffier, L. and Smouse, P.E. (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics* 136, 343–359
16 Fitch, W.M. (1997) Networks and viral evolution. *J. Mol. Evol.* 44, S65–S75
17 Strimmer, K. and Moulton, V. (2000) Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.* 17, 875–881
18 Griffiths, R.C. and Marjoram, P. (1997) An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution*. (Donelly, P. and Tavare, S., eds), pp. 257–270, Springer-Verlag
19 Strimmer, K. *et al.* Recombination analysis using directed graphical models. *Mol. Biol. Evol.* (in press)
20 Makarenkov, V. and Legendre, P. (2000) Improving the additive tree representation of a dissimilarity matrix using reticulations. In *Data Analysis, Classification and Related Methods* (Kiers, H.A.L. *et al.*, eds), pp. 35–46, Springer
21 Xu, S. (2000) Phylogenetic analysis under reticulate evolution. *Mol. Biol. Evol.* 17, 897–907
22 Castelloe, J. and Templeton, A.R. (1994) Root probabilities for intraspecific gene trees under neutral coalescent theory. *Mol. Phylogenet. Evol.* 3, 102–113

23 Holmes, E.C. *et al.* (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.* 16, 741–749
24 Templeton, A.R. *et al.* (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* 66, 69–83
25 Shaw, K.L. (1999) A nested analysis of song groups and species boundaries in the Hawaiian cricket genus *Laupala*. *Mol. Phylogenet. Evol.* 11, 332–341
26 Barraclough, T.G. and Vogler, A.P. (2000) Detecting the geographical pattern of speciation from species-level phylogenies. *Am. Nat.* 155, 419–434
27 Gómez-Zurita, J. *et al.* (2000) Nested cladistic analysis, phylogeography and speciation in the *Timarcha goettingensis* complex (Coleoptera, Chrysomelidae). *Mol. Ecol.* 9, 557–560
28 Sing, C.F. *et al.* (1992) Application of cladistics to the analysis of genotype–phenotype relationships. *Eur. J. Epidemiol.* 8, 3–9
29 Lapointe, F-J. How to account for reticulation event in phylogenetic analysis: a review of distance-based methods. *J. Classif.* (in press)

Ecology of sprouting in woody plants: the persistence niche

William J. Bond and Jeremy J. Midgley

Many woody plants can resprout and many ecosystems are dominated by resprouters. They persist *in situ* through disturbance events such as fire, flooding or wind storms. However, the importance of 'persistence' in plant demography has been neglected in favour of 'recruitment'. Thus much research on plant regeneration, conservation and evolution has focused on the importance of safe sites, seed and seedling banks, dispersal and germination with the implied importance of *de novo* replacement rather than persistence. Recent research shows a growing appreciation for the role of sprouting as a form of persistence in a diversity of ecosystems and tradeoffs between the two regeneration modes.

In a seminal paper in 1977, Peter Grubb[1] introduced the concept of the 'regeneration niche' to help explain the coexistence of similar species. Grubb noted that species that share much the same life form, phenology and habitat range might nevertheless have different seedling requirements. For example, 'when a whole large tree is blown over' the large gap thus formed would favour light-demanding seedlings, whereas smaller gaps would favour shade-tolerant recruits. The literature on the ecology of seeds and seedlings has expanded enormously since this time. Seed ecology is widely studied because of its assumed importance in the population growth of plants. Seed traits are included in general systems for classifying plant functional types[2,3]. The same broad recognition has not been given to the mode of persistence of established plants. When a tree is blown over, gaps might not be filled by seedlings but by shoots sprouting from the fallen tree. Sprouts grow much faster than seedlings and can quickly reoccupy their own gaps. Sprouting ability can have major impacts on plant populations: turnover of populations is reduced; the effects of disturbance are minimized; and dependence on seeds for population maintenance might become negligible. Species differ in their sprouting ability, and both strong and weakly sprouting species occur in diverse ecosystems[4,5]. Here, we review studies of the phenomenon emphasizing emerging generalizations and implications for the population biology of woody plants.

William J. Bond*
Jeremy J. Midgley
Dept Botany, University of CapeTown, Private Bag, Rondebosch 7701, South Africa.
*e-mail: bond@botzoo.uct.ac.za

The ecology of sprouting has analogies to clonality[6]. However, although clonal plants generally sprout, only a small fraction of woody sprouters are clonal and capable of vegetative spread. Sprouting response is difficult to quantify partly because there is a continuum of responses to disturbances of varying severity[7]. This could partly account for its absence from general plant strategy schemes (Box 1). After the least severe disturbance, such as damage caused by caterpillar feeding, plants replace lost tissues by sprouting from buds. As the severity of disturbance increases, plants diverge in their ability to recover by sprouting (Fig. 1). Most interest has been in disturbances that can potentially kill plants such as fire, hurricane damage, drought, flooding, herbivory and landslides, as well as anthropogenic disturbance such as forest clearing. In Mediterranean type shrublands, where crown fires are relatively frequent, shrub species are often either killed outright by burning (nonsprouters or 'seeders') or recover vegetatively from roots or stems (sprouters)[4,8]. Classifying sprouting behaviour is harder in forests and woodlands because species often diverge in their sprouting response at different life history stages[5,7,9]. Some species never sprout; in some species, sprouting ability increases with size to reach a maximum in adult stages, although in other species sprouting is common in juveniles but adults are unable to sprout (Fig. 2). Some forest trees retain a bud bank, sprouting continuously with or without disturbance and thus come close to immortality. Examples include *Tilia cordata* (small-leaved lime) in Britain, whose northern populations have survived by sprouting since climates cooled and reproduction ceased 5000 years ago[10]. *Ginkgo biloba* (maidenhair tree) might owe its survival in China to the extraordinary persistence of trees that resprout from specialized basal swellings on the trunk[11].

Sprouting is common, and might be the ancestral state, in woody angiosperms[12]. Most conifers do not sprout although there are exceptions in several unrelated lineages, including species of *Pinus* in the

http://tree.trends.com 0169–5347/01/$ – see front matter PII: S0169-5347(00)02033-4