Update

Research Focus

Supertrees join the mainstream of phylogenetics

James A. Cotton[1] and Mark Wilkinson[2]

[1] School of Biological and Chemical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK
[2] Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

Supertree methods are fairly widely used to build comprehensive phylogenies for particular groups, but concerns remain over the adequacy of existing approaches. Steel and Rodrigo recently introduced a statistical model of incongruence between trees, allowing maximum-likelihood supertree inference. This approach to supertree construction will enable hypothesis-testing and model-choice methods that are now routine in sequence phylogenetics to be applied in this setting, and might form an important part of future phylogenetic inference from genomic data.

From consensus to supertrees

Consensus trees, which summarise two or more trees for a set of taxa, are commonplace in phylogenetics. Many biologists will be aware of the most common consensus methods: the strict consensus is used to show the agreement between multiple equally good trees, whereas the majority-rule consensus tree is the standard way of summarising the results from bootstrap replicates or Bayesian phylogenetic analyses. These most familiar methods are even mentioned in undergraduate textbooks [1], but a range of less common, and even exotic, consensus methods exists [2].

Supertree methods differ from consensus approaches: whereas consensus methods summarise agreement between trees only in the special case where they have identical taxa (the same leaves), supertree methods attempt to summarise the agreement between trees in the more general case where they have different (but overlapping) sets of leaves [3]. This extension beyond consensus makes things much more difficult. For example, the strict and majority-rule consensus methods can both be defined in terms of trees sharing particular substructure: 'splits' that partition the leaves into two disjoint sets. The strict consensus tree includes all splits common to every tree in a set, whereas the majority-rule tree includes all splits common to at least half of the trees. In the supertree case these familiar ideas are of little use, because each input tree has different leaves to 'split.' Consequently, there is no immediate supertree analog of even these simple consensus methods, and there are limits on what we can expect any supertree method to achieve [4].

These difficulties have led to most supertree methods being *ad hoc*, rather than designed to have particular desirable properties [5]. An obvious example is matrix representation with parsimony (MRP), which is by far the most widely used supertree method. This employs standard maximum parsimony methods by encoding input tree splits as partial binary characters. The *ad hoc* nature of MRP and most other existing methods means they have several fundamental problems [3,6].

Maximum-likelihood supertrees

The current lack of any completely satisfying supertree methods [4] makes the recent proposal of a maximum-likelihood approach to supertree construction [7] particularly exciting. In a purely theoretical paper, Steel and Rodrigo describe a maximum-likelihood framework for estimating phylogenetic supertrees, in which the likelihood function is the probability of obtaining a particular input tree given an underlying supertree and a simple model of phylogenetic error. This error model proposes that some distance between each input tree and the supertree is distributed exponentially (Figure 1). This simple model makes the method both general and natural. For example, it defines a range of methods that include one generalisation of the majority-rule consensus method to the supertree setting [8]. The methods are also statistically consistent (so that, provided the model is not too wrong, the result will converge on the correct result as the amount of data increases).

Steel and Rodrigo [7] go on to show that MRP is inconsistent under certain conditions, so this paper both adds to the growing literature on problems with current approaches to supertree construction and develops a promising alternative. Moving supertree construction into a likelihood framework is also important because it enables a whole raft of standard methods for hypothesis testing, such as the likelihood ratio test to choose between models, to be used in the supertree setting.

Toward statistical phylogenomics

The approach taken by Steel and Rodrigo has important implications for the growing area of 'phylogenomics': inferring phylogenies from multiple genes. This is becoming more and more important as the increasing amount of genomic sequence data can address a wide range of phylogenetic controversies (e.g. [9,10]). Any phylogenetic inference from multiple sources of data, for example multiple genes in a phylogenomic analysis, must make some assumptions about how estimates vary between different data—that is, about incongruence between the estimates. For example, when alignments from different genes are concatenated, this assumes that the genes evolved along the same tree. This is unrealistic on a genomic scale, because processes such as lineage sorting of alleles, duplication and loss of genes, and lateral gene transfer can all make gene phylogenies differ from the phylogeny for the species from which the genes are sampled [11]. Concatenating genes with different evolutionary histories can lead to incorrect

Corresponding author: Cotton, J.A. (j.a.cotton@qmul.ac.uk).

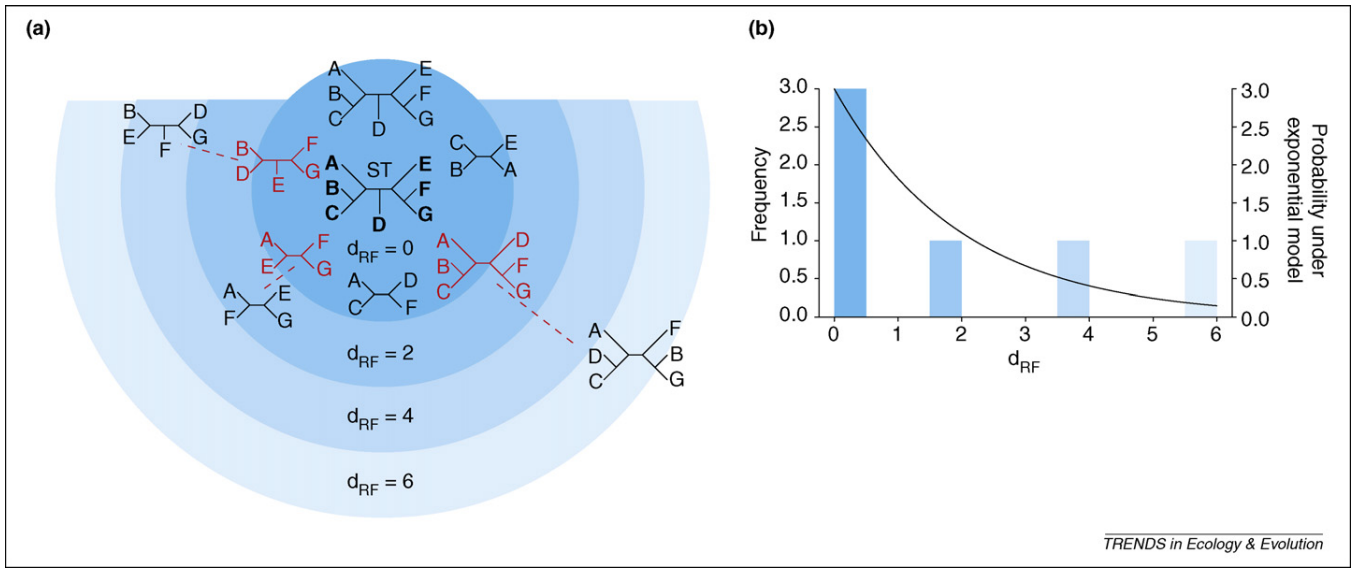

Figure 1. The exponential model. **(a)** The maximum-likelihood supertree (labelled ST, drawn and labelled in bold) for a set of six hypothetical input trees (black lines) on seven taxa. Input trees are shown by Robinson-Foulds distance (d_{RF}) from the supertree, which is the number of splits present in either tree but not in both. Other distances such as the number of subtree-prune and regraft (SPR) operations between the trees could also be used. Red trees in the central circle are identical to the supertree but pruned to match the leaf set of a particular input tree, to which they are connected by dashed lines. Three of the input trees match the pruned supertree exactly, and are shown as black trees in the central circle. **(b)** Steel and Rodrigo's model proposes that these distances are exponentially distributed. The distribution of d_{RF} for these data, and the best-fitting exponential distribution (mean 2.0), shows the probability of each input tree under this exponential model.

estimates of this underlying species phylogeny [12,13], so genome-scale analyses need to address this problem.

In contrast to previous approaches, Steel and Rodrigo's model treats incongruence in a purely formal way, using a distribution of distances between trees, rather than seeking to model the processes that cause incongruence [11,12]. Given the complexities of modelling these processes, we think that this approach seems promising: we know that a mixture of different processes might be acting in any particular case, so modelling any single source of incongruence might be inadequate. This is analogous to models of nucleotide substitution used in phylogenetics that generally make no attempt to model the population genetic processes such as mutation, selection and drift that lead to substitutions becoming fixed [14].

In this context, the supertree model could form part of a hierarchical model, combining the substitution process and incongruence between different loci (e.g. [15]) with the likelihood of a particular supertree being summed across possible tree topologies for each locus. Like most supertree methods, the basic supertree model treats the input trees as fixed, that is, as observations made without error. In fact, the input trees are themselves estimates from data, and this kind of hierarchical model is the most natural way of taking into account the uncertainty in these estimates. Allowing the underlying tree topology to vary between different loci is a substantial extension of current models, which allow model parameters such as nucleotide composition, substitution rates and branch lengths to vary between partitions of the data (Figure 2).

Are supertree methods coming of age?

The history of supertree methods has been one of continual innovation, with a bewildering range of methods proposed,

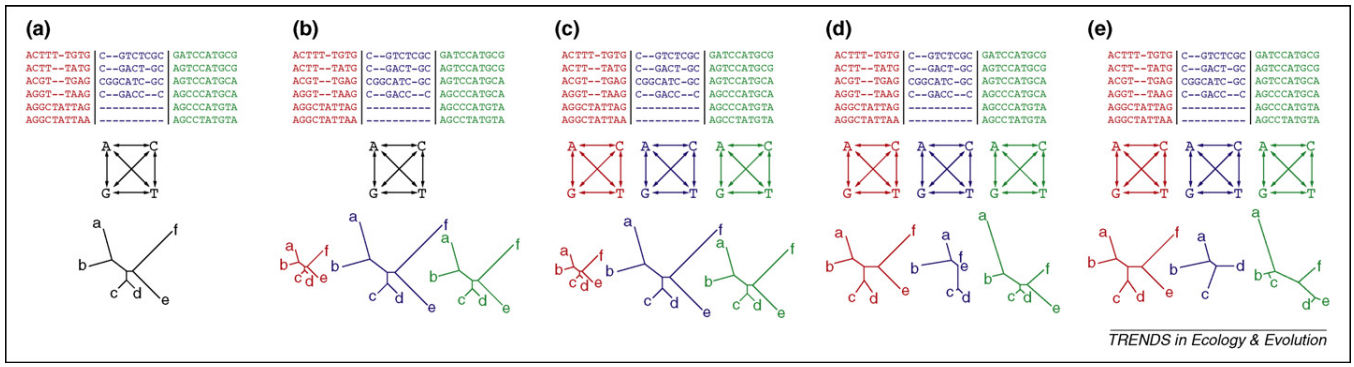

Figure 2. A continuum of phylogenetic models for multilocus data. **(a)** The most basic approach to phylogenetic inference from multilocus data is to concatenate data from a set of genes, and analyse the combined alignment using a single model of sequence evolution to infer a single tree and set of model parameters. **(b)** More-complex analyses might allow the rate of evolution for each gene to differ and **(c)** allow different parameters describing the substitution process for each gene. **(d)** Another step would be to allow rates of substitution to vary on particular branches for each gene, so that each gene has independent branch lengths. **(e)** The most general approach would also allow each gene to have a different topology and set of taxa. In this case, a model of incongruence between gene tree topologies is needed. Colours indicate different data partitions (usually different loci) and the model components (substitution matrices, trees and branch lengths) that apply to those partitions. Model components in black apply across all partitions.

and much less work on understanding the properties or performance of the methods [3]. There is a place for different approaches to reflect the diversity of data and hypotheses being addressed, but it is important that supertree workers justify their choice of method. MRP remains popular largely because of its convenience (as standard phylogenetic software is used to analyse the transformed data), but relying exclusively upon MRP is increasingly difficult to defend.

By proposing a statistical model for the supertree problem, the maximum-likelihood approach is a step in the right direction. Bringing phylogenetic supertrees into the same framework as phylogenetic inference from sequence data should help users focus on the importance of methodological issues, as statistical model selection is now routine when building phylogenies from sequences. In combination with increasingly realistic models of sequence evolution, maximum-likelihood approaches to supertree construction look set to play an important role in the growing field of phylogenomics. Much work remains to be done: it is unclear how well the mathematically convenient exponential distribution models incongruence in real data, and work has only just begun on efficiently estimating optimal trees under this model. Whether or not this particular approach is successful, supertree methods might need to become part of the mainstream toolkit of molecular phylogenetics if systematists are to make proper use of the deluge of genomic sequence data.

Acknowledgements

J.A.C. is supported by an RCUK academic fellowship.

References

1 Page, R.D.M. and Holmes, E.C. (1998) *Molecular Evolution: A Phylogenetic Approach,* Wiley-Blackwell
2 Bryant, D. (2003) A classification of consensus methods for phylogenies. In *Bioconsensus* (Janowitz, M. *et al.*, eds), pp. 163–184, DIMACS AMS
3 Bininda-Emonds, O.R.P. (2004) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life,* Kluwer Academic
4 Steel, M. *et al.* (2000) Simple but fundamental limits for supertrees and consensus methods. *Syst. Biol.* 49, 363–368
5 Wilkinson, M. *et al.* (2007) Properties of supertree methods in the consensus setting. *Syst. Biol.* 56, 330–337
6 Wilkinson, M. *et al.* (2005) The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Syst. Biol.* 54, 419–431
7 Steel, M. and Rodrigo, A. (2008) Maximum-likelihood supertrees. *Syst. Biol.* 57, 243–250
8 Cotton, J.A. and Wilkinson, M. (2007) Majority-rule supertrees. *Syst. Biol.* 56, 445–452
9 Delsuc, F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375
10 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804
11 Slowinski, J.B. and Page, R.D.M. (1999) How should phylogenies be inferred from sequence data? *Syst. Biol.* 48, 814–825
12 Kubatko, L.S. and Degnan, J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24
13 Matsen, F.A. and Steel, M.A. (2007) Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* 56, 767–775
14 Yang, Z. and Nielsen, R. (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.* 25, 568–579
15 Ané, C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24, 412–426

0169-5347/$ – see front matter
doi:10.1016/j.tree.2008.08.006 Available online 18 November 2008