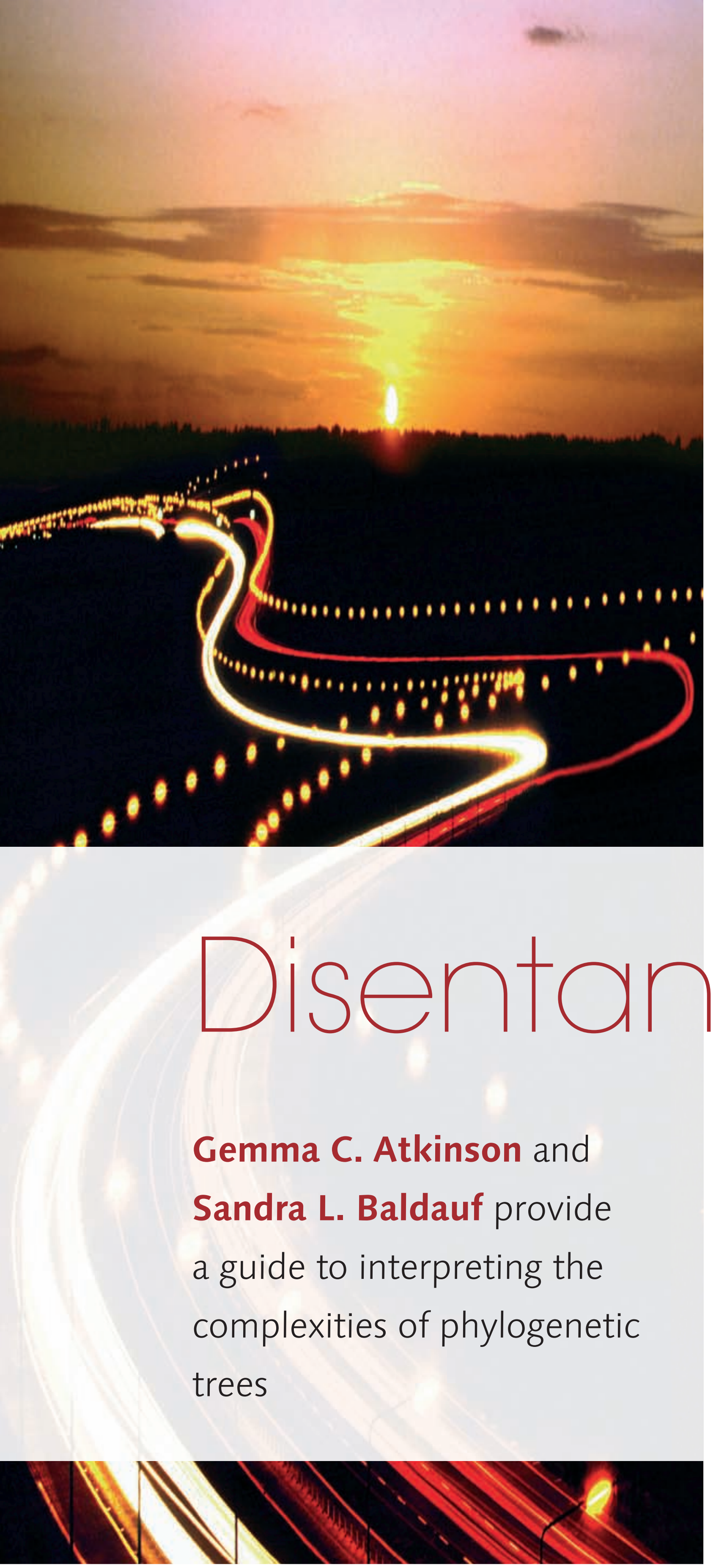
Disentangling the trails of evolution

Gemma C. Atkinson and **Sandra L. Baldauf** provide a guide to interpreting the complexities of phylogenetic trees



Entwined trails of car and road lights, representing the tangled trails of evolution. *Brand X Pictures / Punchstock*

The exponential growth in publicly available molecular sequence data has created a gold mine of encoded information covering billions of years of evolution. One of the most immediate ways to make sense of these data is by the comparison of the sequences in phylogenetic analyses. Getting a simple tree is relatively easy and potentially very informative (if not exactly publishable). A deeper understanding of the methods can help you grow trees that are more reliable and reveal complex evolutionary histories.

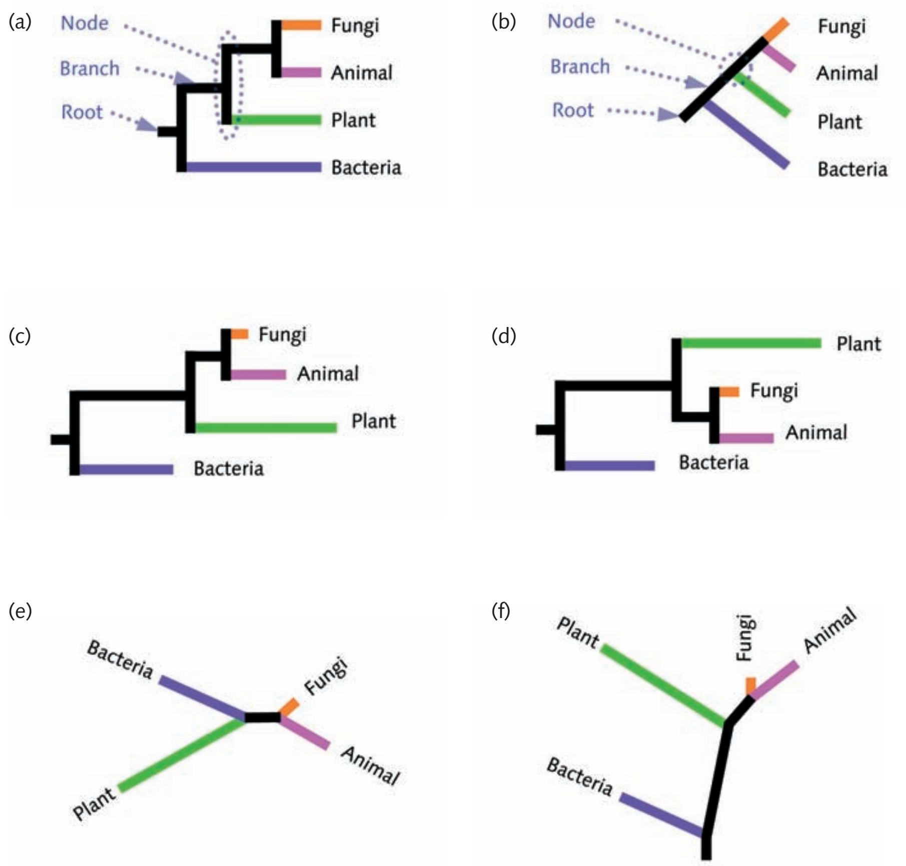
If it is relationships among organisms that you are interested in, it is becoming clear that a tree of organisms based on a single gene or its corresponding protein sequence may not necessarily represent the true history of the organisms in the tree. Rather, it tells the story of the evolutionary history of that particular gene while another gene a few bases away in the genome may have experienced quite a different evolutionary history. This is particularly true in bacterial genomes, where 'illegitimate' trading of genes is rampant (see below). Used with care, and perhaps a healthy dose of scepticism, phylogenetics offers powerful tools for tracing the often entangled evolutionary trails of genes. For bacteria, it is especially useful for revealing phenomena such as lateral gene transfer (LGT), and in all domains of life, gene duplications and, sometimes, true organismal relationships.

This article is intended as a brief tutorial on how to interpret phylogenetic tree diagrams and an introduction to the methods of preparing raw data and building a tree.

Throughout, we wish to emphasize that phylogenetics, like any other branch of science, requires informed judgement in the selection and application of methods to maximize the accuracy of the results.

Reading trees

Leaves, branches and nodes. Phylogenetic trees, or phylogenies, display deduced evolutionary relationships in the form of multiple branching lineages (Fig. 1). The sources of the sequences, or 'operational taxonomic units' (OTUs), are the tips or leaves of the tree. Nodes (junctions where two or more branches join), represent a divergence event (gene duplication or speciation) and also the hypothetical last common ancestor of all the branches arising from them. The further a node is from the tips, the further back in time the divergence happened, with the root of the tree being the most ancient node.



► Fig. 1. Equivalent phylogenetic trees. All these hypothetical trees have exactly the same topology and differ only in presentation style. (a, b) Rooted cladograms, which show branching order but no information on the amount of evolution on each branch; (c, d) rooted phylograms, which show branch lengths proportional to evolutionary distance; and (e) unrooted and (f) rooted radial trees. Note: the node height in trees (a), (c) and (d) is purely a device for even spacing of branches. *G.C. Atkinson*

Orthologues and paralogues. All sequences that share a common ancestor are homologues. Homologues come in several different flavours (Fig. 2). Orthologues are the direct descendents from a single ancestral sequence and are duplicated along with the rest of the genome in each generation. Paralogues arise by gene duplication, thus giving rise to multigene families.

Xenologues. Phylogenetic trees are one of the best means of detecting LGT, which is rampant in bacteria and archaea, and also occurs at a lower, if largely unknown, frequency in eukaryotes. Xenologues are homologues that have undergone LGT and can be identified in phylogenetic trees by their tendency to nest with sequences from the donor's lineage (Fig. 2b).

Growing your own tree

Digging in the databases. A huge, exponentially growing amount of nucleotide and protein sequences is stored in public sequence databases (Table 1), the main ones of which are updated against each other daily. These are usually the first port of call when assembling or augmenting a dataset. As annotations are unreliable, the best way to find homologous sequences is by doing BLAST or FASTA sequence similarity searches. These look for matches to a user-provided query sequence in one or many sequence databases (Table 1).

The alignment: the foundation from which a tree grows. Once you have found your sequences, the next step is to align them (Table 1). A phylogenetic tree is only as good as the alignment it is built on. Multiple sequence alignment programs have continuously improved over the years, but they still tend to perform badly in regions of poor sequence conservation. Here, the human eye is often better at recognizing homologous patterns. Therefore, all alignments should be inspected by eye before building trees (Fig. 3). Only homologous positions should be used to build trees; gaps and ambiguously aligned regions should, as a general rule, be excluded.

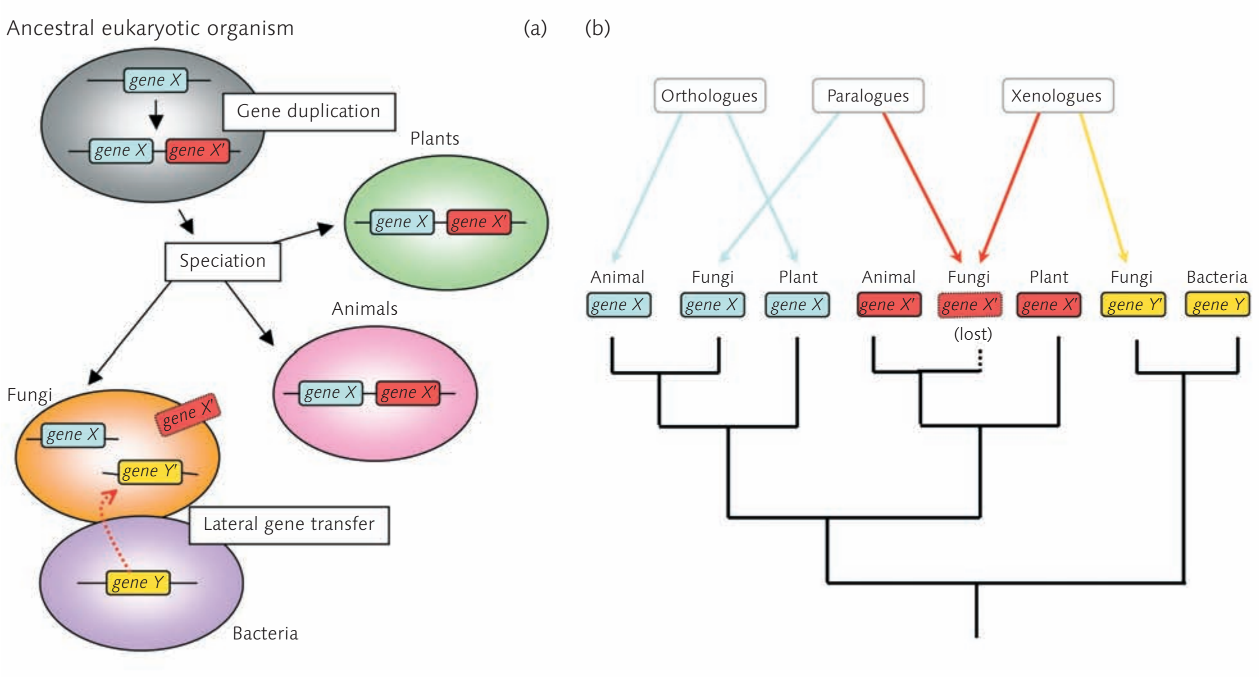
> *With the increasing amount of structural and functional information now available, it is becoming possible to map changes in protein function and structure onto evolutionary trees*

Minimizing distances and hiking through tree landscapes. There are two main classes of methods for building trees: distance and discrete data methods. Distance methods (UPGMA, neighbour joining) summarize all the information between pairs of sequences into a single statistic. This is the distance, essentially the percentage difference, between two sequences. OTUs are clustered into groups in the tree based on these pair-wise distances. Parsimony, maximum likelihood (ML) and Bayesian inference methods are discrete data methods as they treat each column in the alignment as a discrete data point. These methods search 'tree-space', a probabilistic landscape of hills and valleys made up of all possible trees and their 'fitness', in a quest to find the tree or set of trees that best fit the data.

All tree building methods have their strengths and limitations. Therefore, it is a good idea to repeat analyses with multiple methods. Greater confidence in results can be gained if different methods produce congruent trees.

Modelling evolution. All phylogenetic programs assume some evolutionary model to explain amino acid or nucleotide substitution patterns. These are used by the various phylogeny programs to attach weights to different classes of substitutions, rather than assuming all mutations are equally likely. This is important for accurate trees. For example, a mutation that results in an amino acid substituting for another with similar chemical properties tends to have a minimal effect on the function of the protein. Such substitutions are frequent and indicate less evolutionary time than changes at slowly evolving sites; this is taken into account by the gamma rate correction.

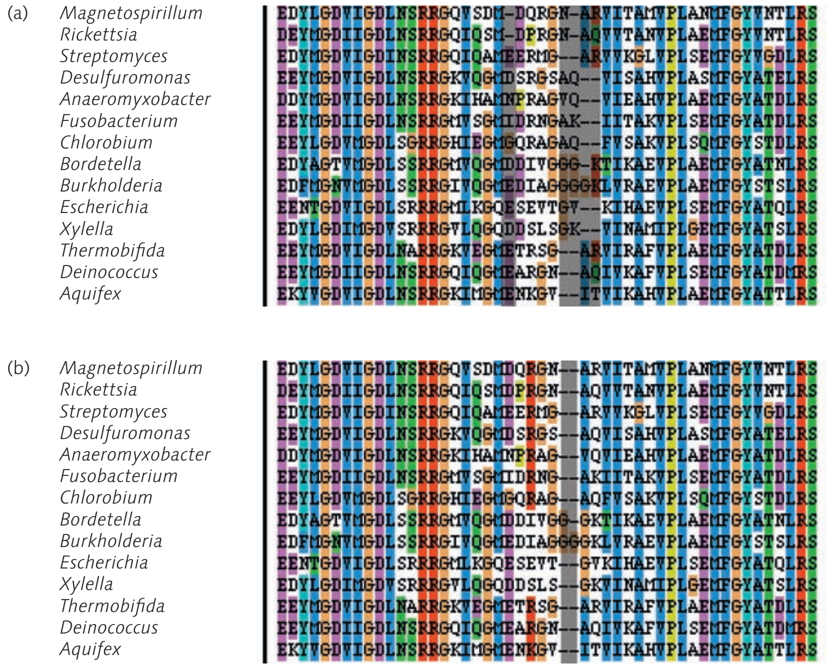
Where is the root? Almost all phylogenetic programs produce unrooted trees from molecular data. However, a root is essential for establishing the order of divergences in a tree. Often, the first diverging sequence in a dataset is unknown. In this case, an outgroup of one or more OTUs can be included



◀ Fig. 2. Orthologues, paralogues and xenologues. Homologous genes can arise by vertical descent or speciation (orthologues), duplication (paralogues), and LGT (xenologues). The latter two can result in non-cannonical phylogenetic trees. (a) Genes *X* and *Y* are ancient homologues, with *X* being the eukaryotic version and *Y* the bacterial one. A hypothetical gene duplication event prior to the origin of plants, animals and fungi leads to the *X'* paralogue, which is inherited by descendent species. In one lineage, *X'* is replaced by *Y'*, a xenologous version of the gene brought about by LGT of gene *Y*. (b) The resulting phylogeny. Since the fungal *X'* gene is lost in this example, this gene would not appear in the phylogeny (dotted line). *G.C. Atkinson*

Table 1. Phylogenetic resources

Online sequence acquisition		
Major databanks	DDBJ	www.ddbj.nig.ac.jp/
	EMBL	www.ebi.ac.uk/embl/
	NCBI	www.ncbi.nih.gov/
	Uni-Prot	www.ebi.uniprot.org/index.shtml
Major genome centres	TIGR	www.tigr.org/
	Sanger	www.sanger.ac.uk/
	JGI	www.jgi.doe.gov/
Genome project portals	GOLD	www.genomesonline.org/
	Eukaryote Genome Portal	www-users.york.ac.uk/~ct505/PhD_Project5/Eukaryote_Homepage.htm
Keyword searches	SRS	http://srs.ebi.ac.uk/
	Entrez	www.ncbi.nlm.nih.gov/Entrez/
Homology searches	BLAST	http://ncbi.nih.gov/BLAST/
	FASTA	www.ebi.ac.uk/fasta33/index.html
Multiple sequence alignment		
	ClustalX	ftp://ftp.ebi.ac.uk/pub/software/
	Muscle	www.drive5.com/muscle/
	T-Coffee	www.ch.embnet.org/software/TCoffee.html
	Bioedit	www.mbio.ncsu.edu/BioEdit/bioedit.html
Phylogenetic analysis		
	PAUP*	http://paup.csit.fsu.edu/
	PHYLIP	http://evolution.genetics.washington.edu/phylip.html
	PHYML	http://atgc.lirmm.fr/phyml/
	MrBayes	http://mrbayes.csit.fsu.edu/
	GARLI	www.bio.utexas.edu/grad/zwickl/web/garli.html
	MEGA	www.megasoftware.net/
	TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
	RAxML	www.ics.forth.gr/~stamatak/index-Dateien/Page443.htm
Comprehensive list		http://evolution.genetics.washington.edu/phylip/software.html



(a)
Magnetospirillum
Rickettsia
Streptomyces
Desulfuromonas
Anaeromyxobacter
Fusobacterium
Chlorobium
Bordetella
Burkholderia
Escherichia
Xylella
Thermobifida
Deinococcus
Aquifex
EDYLGDVIGDLNSRRGQVSDM-DQRGN-ARVITAMVPLANMFGYVNTLRS
DEYMGDIIGDLNSRRGQIQSM-DPRGN-AQVVTANVPLAEMFGYVNTLRS
EDYMGDVIGDINSRRGQIQAMEERMG--ARVVKGLVPLSEMFGYVGDLRS
EEYMGDVIGDLNSRRGKVQGMDSRGSAQ--VISAHVPLASMFGYATELRS
DDYMGDVIGDLNSRRGKIHAMNPRAGVQ--VIEAHVPLAEMFGYATDLRS
EEYMGDIIGDLNSRRGMVSGMIDRNGAK--IITAKVPLSEMFGYATDLRS
EEYLGDVMGDLSGRRGHIEGMGQRAGAQ--FVSAKVPLSQMFGYSTDLRS
EDYAGTVMGDLSSRRGMVQGMDDIVGGG-KTIKAEVPLAEMFGYATNLRS
EDFMGNVMGDLSSRRGIVQGMEDIAGGGGKLVRAEVPLAEMFGYSTSLRS
EENTGDVIGDLSRRRGMLKGQESEVTGV--KIHAEVPLSEMFGYATQLRS
EDYLGDIMGDVSRRRGVLQGQDDSLSGK--VINAMIPLGEMFGYATSLRS
EEYMGDVIGDLNARRGKVEGMETRSG--ARVIRAFVPLAEMFGYATDLRS
EEYMGDIIGDLNSRRGQIQGMEARGN--AQIVKAFVPLSEMFGYATDMRS
EKYVGDVIGDLNSRRGKIMGMENKGV--ITVIKAHVPLAEMFGYATTLRS
(b)
Magnetospirillum
Rickettsia
Streptomyces
Desulfuromonas
Anaeromyxobacter
Fusobacterium
Chlorobium
Bordetella
Burkholderia
Escherichia
Xylella
Thermobifida
Deinococcus
Aquifex
EDYLGDVIGDLNSRRGQVSDMDQRGN--ARVITAMVPLANMFGYVNTLRS
DEYMGDIIGDLNSRRGQIQSMDPRGN--AQVVTANVPLAEMFGYVNTLRS
EDYMGDVIGDINSRRGQIQAMEERMG--ARVVKGLVPLSEMFGYVGDLRS
EEYMGDVIGDLNSRRGKVQGMDSRGS--AQVISAHVPLASMFGYATELRS
DDYMGDVIGDLNSRRGKIHAMNPRAG--VQVIEAHVPLAEMFGYATDLRS
EEYMGDIIGDLNSRRGMVSGMIDRNG--AKIITAKVPLSEMFGYATDLRS
EEYLGDVMGDLSGRRGHIEGMGQRAG--AQFVSAKVPLSQMFGYSTDLRS
EDYAGTVMGDLSSRRGMVQGMDDIVGG-GKTIKAEVPLAEMFGYATNLRS
EDFMGNVMGDLSSRRGIVQGMEDIAGGGGKLVRAEVPLAEMFGYSTSLRS
EENTGDVIGDLSRRRGMLKGQESEVT--GVKIHAEVPLSEMFGYATQLRS
EDYLGDIMGDVSRRRGVLQGQDDSLS--GKVINAMIPLGEMFGYATSLRS
EEYMGDVIGDLNARRGKVEGMETRSG--ARVIRAFVPLAEMFGYATDLRS
EEYMGDIIGDLNSRRGQIQGMEARGN--AQIVKAFVPLSEMFGYATDMRS
EKYVGDVIGDLNSRRGKIMGMENKGV--ITVIKAHVPLAEMFGYATTLRS

▶ Fig. 3. Editing an alignment by eye. (a) A section of a CLUSTAL X (Table 1) alignment of bacterial protein sequences before editing. Five columns contain insertions (shaded columns). (b) The refined alignment. Only two columns contain insertions, a more parsimonious arrangement as fewer insertion or deletion events are assumed. *G.C. Atkinson* in the dataset. This is essentially an external point of reference to identify the oldest node in the tree, which is the outgroup's closest relative. For example, for a tree of mammalian genes, one could use the orthologus gene from a marsupial as the outgroup.

Statistical tests for support – how strong is that branch?

There are various ways to determine how 'strong' a tree is, that is, how much better it is than other possible trees. The most commonly used method is bootstrapping. This involves phylogenetic analyses of multiple random subsamples of your dataset. The bootstrap support for each branch corresponds to the percentage of analyses where that branch appears.

Bayesian inference is becoming increasingly popular in molecular phylogeny. Instead of bootstrapping, this discrete data method uses all the trees it encounters in the tree space to calculate the posterior probability for each branch in a consensus tree of all encountered trees. This is fundamentally very different from bootstrapping and posterior probability and bootstrap support values are not directly comparable. Generally, a bootstrap percentage greater than 70 % can be taken as good support, but only probabilities greater than 95 % are reasonable support for Bayesian inference trees.

Fatal attraction: the curse of long branches. The rate of molecular evolution is not uniform across a tree; some OTUs may have acquired more mutations than others over time and therefore grow longer branches in the tree of their evolutionary history. Tree-building programs have problems with such branches, and tend to group long-branched OTUs together due to spurious sequence similarities. This is the notorious 'long-branch attraction' or LBA. Likelihood methods are on the whole less likely to be duped by variations in evolutionary rate, and the use of more accurate evolutionary models can give any algorithm a helping hand. However, likelihood methods are not beyond failure and there is no perfect model of evolution. Therefore, the easiest remedy for LBA is often to include intermediate sequences to break up long branches. Alternatively, if the long branches are not essential for the questions you are asking of the tree, it is often best just to leave them out.

Conclusions

Phylogenetics allows us to piece together clues left by chance to retrace evolutionary history, but the clues are patchy as most taxa and many genes are extinct. Because of this, phylogenetic methods rely on the experience and judgement of the researcher for the quality of the results they deliver, and knowledge of how to interpret trees is essential for making sense of the patterns within them. The rewards are great, however, as retracing the evolution of a gene with phylogenetic analysis can reveal complex and often surprising stories of molecular history. With the increasing amount of structural and functional information now available, it is becoming possible to map changes in protein function and structure onto evolutionary trees. This will allow us to begin to ask not just how organisms work, but why they do the sometimes seemingly very strange things that they do.

Gemma C. Atkinson and **Sandra L. Baldauf**
Department of Biology, Box 373, University of York,
Heslington, York YO10 5YW, UK (t 01904 328635;
e gca500@york.ac.uk, slb14@york.ac.uk)