Multilocus Sequence Analysis and Bacterial Species Phylogeny Estimation

Pablo Vinuesa

Abstract
This chapter presents a review of critical factors that have to be considered and evaluated in multilocus sequence analysis (MLSA) in order to make robust estimates of bacterial species phylogenies. The theoretical arguments in favour of the conditional data combination will be presented. I will briefly review criteria for marker selection, and will provide practical advice on the computational aspects, potential pitfalls, and software choices available for each step in a MLSA. For this purpose, a detailed case study using \textit{atpD}, \textit{glnII}, \textit{recA} and \textit{rpoB} sequences of symbiotic root nodule bacteria of the genus \textit{Bradyrhizobium} will be presented. I will discuss and illustrate the use of phylogenetic congruence analysis, and a strategy to evaluate the additivity of the phylogenetic signals in the different partitions. The importance of using multiple isolates per species/lineage, proper model selection and thorough tree searches to get a good estimate of a multispecies phylogeny will be emphasized. Maximum likelihood and Bayesian phylogeny estimation using supermatrices are thoroughly discussed and critically compared with the new, concatenation-independent, Bayesian Estimation of Species Trees (BEST) algorithm, which is based on the multispecies coalescent.

Introduction
Multilocus sequence analysis in the broader context of molecular systematics
Multilocus sequence analysis (MLSA) represents the novel standard in microbial molecular systematics (Gevers et al., 2005). In this context, MLSA is implemented in a relatively straightforward way, consisting essentially in the concatenation of several sequence partitions for the same set of organisms, resulting in a ‘supermatrix’ which is used to infer a phylogeny by means of distance-matrix or optimality criterion-based methods (Devulder et al., 2005; Vinuesa et al., 2005b, 2008; Wertz et al., 2003). This approach is expected to have an increased resolving power due to the large number of characters analysed, and a lower sensitivity to the impact of conflicting signals (i.e. phylogenetic incongruence) that result from eventual horizontal gene transfer events (Escobar-Páramo et al., 2004; Lerat et al., 2003; Rokas et al., 2003).

Although claimed as a new standard in microbial systematics, MLSA is by no means new in the broader context of molecular systematics, as reflected by the rich literature available on how to treat partitioned data. The strategies used to deal with multiple partitions can be grouped in three broad categories: the total evidence, separate analysis and combination approaches (Huelsenbeck et al., 1996). The concatenation approach that dominates MLSAs in the microbial molecular systematics literature is known to systematists working with plants and animals as
the 'total molecular evidence' approach, and has been used to solve difficult phylogenetic questions such as the relationships among the major groups of cetaceans (Hasegawa et al., 1997), that of microsporidia and fungi (Hirt et al., 1999), or the phylogeny of major plant lineages (Soltis et al., 1999), to mention a few classic examples.

The total molecular evidence approach has been criticized because by directly concatenating all available sequence alignments, the evidence of conflicting phylogenetic signals in the different data partitions is lost along with the possibility to uncover the evolutionary processes that gave rise to such contradictory signals (Bull, 1993). The nature of these conflicts is varied, but in the microbial world the strongest conflicting signals often derive from the existence of horizontal gene transfer (HGT) events in the dataset (Boucher et al., 2003; Gogarten and Townsend, 2005; Ochman et al., 2005). If the individuals containing xenologous loci are not identified and removed from the supermatrix prior to phylogeny inference, the resulting hypothesis may be strongly distorted, since standard treeing methods assume a single underlying evolutionary history (Posada and Crandall, 2002; Schierup and Hein, 2000; Vinuesa et al., 2005c). Based on these arguments, the conditional data combination strategy is to be generally preferred in bacterial MLSA (Vinuesa et al., 2005c, 2008).

Gene trees versus species trees
Owing to the stochasticity of gene lineage sorting during the speciation process, no single gene tree is likely to adequately reflect the species phylogeny (Degnan and Rosenberg, 2006; Maddison and Knowles, 2006; Nichols, 2001; Rosenberg, 2002; Rosenberg and Nordborg, 2002). Therefore, single gene trees (i.e. 16S rRNA-based phylogenies) should not be considered equal to species trees (Maddison, 1997; Nichols, 2001; Vinuesa et al., 2005a, 2005b). Theoretically, the best strategy to get a reasonable estimate of the species tree is to consider multiple genealogies inferred from unlinked loci, and to use multiple individuals per species (Rosenberg and Nordborg, 2002; Vinuesa et al., 2008). These important theoretical considerations have been largely neglected in the bulk of the microbial molecular systematics and taxonomic literature (Vinuesa et al., 2005c, 2008). This fact is clearly reflected by the dominant practice of proposing novel bacterial species based on the analysis of a single isolate, the species’ type strain. This specimen is compared with other such closely related type strains by DNA–DNA hybridization and phenotypic traits, providing phylogenetic evidence for the relationships among the ‘species’ under consideration based on a single gene sequence (almost invariably the SSU rRNA gene). This practice forms the so-called polyphasic approach to taxonomy that dominates the field (Stackebrandt et al., 2002; Stackebrandt and Goebel, 1994; Vandamme et al., 1996). In addition to the above mentioned limitations of phylogenies estimated from single strains per species, such phylogenies are of little use in a truly molecular systematics framework, since a minimum of two strains per species would be required for them to form a clade (i.e. a monophyletic group). Ideally, multiple individuals from a single species sampled from different demes (populations) should form monophyletic lineages on the estimated species tree. To have multiple strains per species is also very valuable to detect potential HGT events, and a minimum of 8–12 strains per species or population is required to make population genetic analyses in order to demarcate the species boundaries using evolutionary criteria, and to make inferences about trait evolution, population genetic structures, population subdivision or biogeographic distribution patterns (Didelot and Falush, 2007; Falush et al., 2003; Ivars-Martinez et al., 2008; Martiny et al., 2006; Miragaia et al., 2007; Papke et al., 2003, 2007; Perez-Losada et al., 2006; Ramette and Tiedje, 2007; Vinuesa et al., 2005c, 2008; Whitaker et al., 2003). Owing to space constraints, I will not discuss the use of multilocus data in bacterial population genetics or ecological contexts, referring the interested reader to the references provided above.

What are the basic attributes of good phylogenetic markers for MLSA?

Core versus accessory genes
Many bacterial species have so-called open pangenomes (Tettelin et al., 2005, 2008), with a small fraction of conserved ‘core’ genes and a
large fraction of strain-specific ‘accessory’ genes found only in some individuals of the species. This was first illustrated by the work of Welch and collaborators in a comparative study of three Escherichia coli genomes. They found that only 39% of the single copy genes were shared by the commensal, uropathogenic and enterohaemorrhagic strains compared in that study (Welch et al., 2002). They also showed that most of the accessory genes were clustered as genomic islands interspersed between syntenic orthologous genes and operons, a finding that has been corroborated in more recent studies, showing that the genomic diversity of E. coli represents an open pangenome model containing a reservoir of more than 13,000 genes, many of which may be uncharacterized but important virulence factors (Rasko et al., 2008).

Only the orthologous core loci are useful for MLSA-based inference of multispecies phylogenies and for within-species population genetic analyses (Vinuesa et al., 2005c, 2008). Accessory loci often encode interesting functions for ecological specialization such as virulence, resistance or symbiosis genes which are very informative about ecological attributes of the strains that harbour them. However, they frequently have a contrasting evolutionary history, mode and tempo of evolution when compared with those of the core genes. Therefore, accessory genes are not suitable molecular markers to estimate species phylogenies (Enright et al., 2002; Ogura et al., 2009; Reid et al., 2000; Silva et al., 2005; Vinuesa et al., 2005c; Wirth et al., 2006; Young et al., 2006). This highlights the mosaic nature of many bacterial genomes, which allows rapid genomic adaptation to changing environments (Ogura et al., 2007; Schubert et al., 2009; Sullivan and Ronson, 1998; Touchon et al., 2009; Young et al., 2006).

Ubiquitous core genes, primer design and marker selection for MLSA

A limitation in MLSA and MLST studies is the lack of definition of a ‘universal’ set of loci to target as molecular markers. With ‘universal’ I don’t mean universal primers that would amplify the corresponding loci from whatever organism one may encounter, but conserved orthologous loci across higher taxonomic ranks on which to base the group-specific primer formulations. The lack of such a definition has made it difficult to evaluate the relative merits of different markers and PCR primers used in different studies, often precluding the comparative analysis of evolutionary processes and patterns among closely related species and genera.

Recent estimates of the conserved core genome of all bacteria indicate that it is most likely composed of < 50 genes, of which 34 are ubiquitous (Charlebois and Doolittle, 2004), a figure not much larger than the universal set of single copy orthologous genes shared by the three domains of life, which has been estimated to be 31 genes (Ciccarelli et al., 2006). These figures are very low, and don’t increase much if the ubiquity criterion is relaxed to presence in 90% or 80% of members in each prokaryotic phylum, reaching 60 and 71 genes, respectively (Charlebois and Doolittle, 2004). In a recent phylogenomic analysis of 41 fully sequenced genomes of the Rhizobiales we found 62 of the relaxed set of 71 ubiquitous genes mentioned above (Contreras-Moreira and Vinuesa, in preparation). These genes are potentially useful to develop a broadly applicable MLSA scheme for bacteria in this order. However, 19 of them are < 600 nucleotides long (Table 3.1) and therefore most likely of little value as targets to develop informative molecular markers.

We have recently published the primers4clades web server, which implements an easy to use and highly customizable analysis pipeline to design lineage-specific degenerate PCR primer pairs in an automatic fashion, based on multiple protein sequence alignments using an extended CODEHOP algorithm (Contreras-Moreira et al., 2009). The server does not only provide the user with primer formulations, sorted according to their theoretical quality (measured on the basis of a comprehensive set of thermodynamic criteria), but also estimates the phylogenetic information content of the expected amplicons, as will be explained in detail later. These features greatly help the user in making an informed choice when confronted with alternative primer-pair formulations for a single gene, a situation often encountered with large genes such as ileS or rpoB (Table 3.1). We are currently developing and evaluating primer-pair formulations for MLSA studies in the Rhizobiales and other bacterial taxa (Bacillales, Enterobacteriales, Mycobacteriaceae,

Table 3.1 Mean and median lengths of a set of 62 highly conserved universal orthologous genes found among 41 Rhizobiales genomes, sorted by decreasing length (L.), corresponding to the set of 71 ‘relaxed’ universal orthologous genes shared by Archaea and Bacteria, as defined by Charlebois and Doolittle (2004)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mean L.</th>
<th>Median L.</th>
<th>Gene</th>
<th>Mean L.</th>
<th>Median L.</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpoC</td>
<td>4203.64</td>
<td>4203</td>
<td>trpS</td>
<td>1060.93</td>
<td>1065</td>
</tr>
<tr>
<td>rpoB</td>
<td>4136.36</td>
<td>4140</td>
<td>rpsB</td>
<td>902.79</td>
<td>996</td>
</tr>
<tr>
<td>ilS</td>
<td>3006.43</td>
<td>3006</td>
<td>trxB</td>
<td>969.86</td>
<td>966</td>
</tr>
<tr>
<td>valS</td>
<td>2857.07</td>
<td>2859</td>
<td>ksgA</td>
<td>848.57</td>
<td>861</td>
</tr>
<tr>
<td>topA</td>
<td>2694.21</td>
<td>2736</td>
<td>rplB</td>
<td>835.71</td>
<td>837</td>
</tr>
<tr>
<td>alaS</td>
<td>2679.43</td>
<td>2676</td>
<td>map</td>
<td>827.57</td>
<td>825</td>
</tr>
<tr>
<td>infB</td>
<td>2663.79</td>
<td>2661</td>
<td>uppS</td>
<td>742.93</td>
<td>759</td>
</tr>
<tr>
<td>leuS_v2</td>
<td>2635.07</td>
<td>2631</td>
<td>rplC</td>
<td>705.43</td>
<td>726</td>
</tr>
<tr>
<td>pheT</td>
<td>2424.00</td>
<td>2424</td>
<td>pyrH</td>
<td>725.57</td>
<td>723</td>
</tr>
<tr>
<td>fus</td>
<td>2083.29</td>
<td>2073</td>
<td>rpsC</td>
<td>712.07</td>
<td>708</td>
</tr>
<tr>
<td>thrS</td>
<td>2022.64</td>
<td>2007</td>
<td>rplA</td>
<td>695.14</td>
<td>693</td>
</tr>
<tr>
<td>dnaG</td>
<td>1993.93</td>
<td>2004</td>
<td>rpsD</td>
<td>621.64</td>
<td>618</td>
</tr>
<tr>
<td>ftsH</td>
<td>1926.00</td>
<td>1923</td>
<td>rpsE</td>
<td>574.93</td>
<td>573</td>
</tr>
<tr>
<td>dnaX</td>
<td>1859.36</td>
<td>1875</td>
<td>efp</td>
<td>571.29</td>
<td>567</td>
</tr>
<tr>
<td>argS_v2</td>
<td>1739.79</td>
<td>1794</td>
<td>rplE</td>
<td>559.29</td>
<td>558</td>
</tr>
<tr>
<td>pyrG</td>
<td>1639.50</td>
<td>1632</td>
<td>nusG</td>
<td>532.07</td>
<td>531</td>
</tr>
<tr>
<td>nusA</td>
<td>1613.57</td>
<td>1617</td>
<td>rpsL</td>
<td>476.57</td>
<td>477</td>
</tr>
<tr>
<td>guaA</td>
<td>1606.29</td>
<td>1608</td>
<td>rpsG</td>
<td>471.00</td>
<td>471</td>
</tr>
<tr>
<td>metG</td>
<td>1650.43</td>
<td>1551</td>
<td>rplM</td>
<td>474.21</td>
<td>465</td>
</tr>
<tr>
<td>atpD</td>
<td>1448.57</td>
<td>1431</td>
<td>rplK</td>
<td>429.64</td>
<td>429</td>
</tr>
<tr>
<td>gltX</td>
<td>1443.86</td>
<td>1428</td>
<td>rpsH</td>
<td>399.00</td>
<td>399</td>
</tr>
<tr>
<td>lpdA</td>
<td>1408.29</td>
<td>1407</td>
<td>rpsK</td>
<td>390.43</td>
<td>390</td>
</tr>
<tr>
<td>cysS</td>
<td>1421.57</td>
<td>1401</td>
<td>rplV</td>
<td>388.29</td>
<td>390</td>
</tr>
<tr>
<td>secY</td>
<td>1335.86</td>
<td>1332</td>
<td>rpsL</td>
<td>375.64</td>
<td>372</td>
</tr>
<tr>
<td>proS</td>
<td>1315.93</td>
<td>1320</td>
<td>rpsM</td>
<td>375.86</td>
<td>369</td>
</tr>
<tr>
<td>serS</td>
<td>1349.57</td>
<td>1305</td>
<td>rplN</td>
<td>369.00</td>
<td>369</td>
</tr>
<tr>
<td>glyA</td>
<td>1306.71</td>
<td>1302</td>
<td>rplX</td>
<td>313.50</td>
<td>315</td>
</tr>
<tr>
<td>eno</td>
<td>1273.93</td>
<td>1284</td>
<td>rpsJ</td>
<td>309.00</td>
<td>309</td>
</tr>
<tr>
<td>recA</td>
<td>1087.93</td>
<td>1092</td>
<td>rpsN</td>
<td>306.00</td>
<td>306</td>
</tr>
<tr>
<td>gcp</td>
<td>1106.36</td>
<td>1089</td>
<td>rpsS</td>
<td>279.00</td>
<td>279</td>
</tr>
<tr>
<td>pheS</td>
<td>1090.71</td>
<td>1083</td>
<td>rpmC</td>
<td>205.07</td>
<td>207</td>
</tr>
</tbody>
</table>

A multilocus sequence analysis case study: estimating a species tree for the genus *Bradyrhizobium* using the concatenation approach

In this section I will present a tutorial on how to perform a MLSA. In doing so, I will provide detailed information on what factors to consider in order to design a powerful MLSA for multispecies phylogeny estimation, along with recommendations on software choices available for each step in the analysis for different computing platforms. The data that will be used here are a subset of the *atpD*, *glnII*, *recA* and *rpoB* sequences used in a recently published MLSA-based study of the biogeography and evolutionary genetics of four *Bradyrhizobium* species (Vinuesa et al., 2008). That study analysed the sequence data of 80 soybean nodule isolates obtained from India, Myanmar, Nepal and Vietnam. Thirty three reference strains from a previous study (Vinuesa et al., 2005c) were included in the combined phylogenetic and population genetic approaches used for species demarcation and to estimate the magnitude of evolutionary forces acting within lineages. The aim of the study was to assess the power and practical utility of the multilocus sequence analysis approach (Gevers et al., 2005) for *Bradyrhizobium* molecular systematics and ecological inference. This bacterial genus is considered a ‘taxonomically difficult’ group of organisms due to their highly conserved *rrs* sequences and poor correlation between the groupings formed on the basis of genotypic and phenotypic traits, raising questions about the suitability of the polyphasic taxonomic approach to *Bradyrhizobium* systematics (So et al., 1994; van Rossum et al., 1995).

Multiple sequence alignment, saturation analysis and model selection

A first critical aspect for making good estimates of a gene phylogeny is to obtain a reliable alignment. MLSA is typically based on protein-coding gene fragments. This is of great advantage to ensure a high quality of the multiple sequence alignment because protein alignments are generally much more reliable than nucleotide alignments since the former have 5 times as many character states as the latter. Therefore, if relatively divergent protein-coding sequences are to be aligned, the CDSs should be first translated in order to align them at the protein level. The resulting alignment is then used to generate the underlying codon alignment. There are many programs and algorithms available that generate multiple sequence alignments. Probably the most commonly used programs are those of the Clustal family (Chenna et al., 2003; Larkin et al., 2007). However, newer, faster and more accurate algorithms are implemented in software such as Mafft (Katoh and Toh, 2008) and Muscle (Edgar, 2004), available for all major computing platforms. There are several Windows programs like BioEdit (Hall, 1999) and DAMBE (Xia and Xie, 2001) that allow the translation of CDSs to their encoded protein sequences. Both packages run ClustalW to perform multiple sequence alignments. DAMBE even implements a function that allows the user to align CDSs to the translation products in order to generate the underlying codon alignment. Other options to perform this operation are the RevTrans server or the corresponding Python script that can be freely downloaded (Wernersson and Pedersen, 2003), the PAL2NAL server (Suyama et al., 2006), or simple custom Perl scripts that import BioPerl (Stajich et al., 2002) modules and classes will help in getting this critically important job done on any platform.

If the sequences are relatively divergent, as could be the case when taxa from different genera or families are to be compared, it may be necessary to use the protein alignments instead of the codon alignments as character sources for phylogeny estimation, because DNA alignments get saturated much faster (Page and Holmes, 1998). Software packages like DAMBE (Xia and Xie, 2001) make it easy to get saturation plots for any codon positions in the alignment. If many gaps are found in a multiple sequence alignment, it may be useful to filter out poorly aligned residues by using either GBlocks (Castresana, 2000) or trimAl (Capella-Gutierrez et al., 2009), although this is rarely necessary in typical MLSAs that focus on closely related species or genera.
Once reliable multiple sequence alignments are available for the different gene partitions it is important to select a proper model of sequence evolution (Posada and Buckley, 2004; Posada and Crandall, 2001). One of the most popular programs for automating nucleotide model selection is ModelTest (Posada and Crandall, 1998). Its recent successor jModelTest (Posada, 2008) has the virtue of using the freely available PhyML phylogeny inference software (Guindon and Gascuel, 2003) to calculate the likelihood scores and ML parameter value estimates of each competing model. It is also a cross-platform program since it is written in Java. Furthermore, it implements 5 different model selection strategies, including ‘hierarchical and dynamical likelihood ratio tests’, the ‘Akaike information criterion’, the ‘Bayesian information criterion’, and a ‘decision-theoretic performance-based’ approach. For protein alignments, the equivalent software for selecting the best fitting substitution matrix is ProtTest (Abascal et al., 2005). If you want it all in one program, use ModelGenerator (Keane et al., 2006), a model selection program written in Java, that selects optimal amino acid and nucleotide substitution models from FASTA or Phylip sequences.

If 16S rRNA gene sequences are used, multiple sequence alignments should be performed taking secondary structure criteria into account. There are several good options to do so. ARB (Ludwig et al., 2004) is a dedicated software environment for UNIX/Linux systems that uses curated, pre-aligned rrS sequences to align user-provided sequences (Pruesse et al., 2007). The ARB package comprises additional tools for data import and export, profile and filter calculation, phylogenetic analyses, specific hybridization probe design and evaluation and other components for data analysis. It is therefore the platform of choice for serious work with rrs sequences. It should be noted that although it was initially designed for the analysis of ribosomal RNA data, it can be used for any nucleic and amino acid sequence data as well. Other excellent alternatives for proper rrs sequence alignment are offered by the GreenGenes (DeSantis et al., 2006) and RDP-II (Cole et al., 2007; Maidak et al., 2001) web services. I don’t have space here for discussing the many useful features offered by these web services, but the interested reader can check my online tutorial on using these services, which is accessible from http://www.ccg.unam.mx/~vinuesa/Using_the_GreenGenes_and_RDP_II_servers.html.

Phylogenetic exploration of sequence partitions using distance-matrix reconstruction methods

A convenient way to classify phylogeny inference methods is based on two criteria: (i) the type of data they use to reconstruct the tree(s) (i.e. distance matrices vs. discrete characters) and (ii) the reconstruction strategy (algorithmic vs. tree searches using an optimality criterion) (Felsenstein, 2004a; Swofford et al., 1996). Distance-matrix methods require that the discrete data matrices (i.e. the multiple sequence alignments) are converted to distance matrices. Different nucleotide substitution models or protein substitution matrices can be used to get corrected estimates of the pair-wise genetic distances between all sequences considered (Felsenstein, 2004a; Swofford et al., 1996). A short online tutorial on the use and selection of nucleotide and protein evolutionary models in phylogenetics can be found at the author’s web site (http://www.ccg.unam.mx/~vinuesa/Model_fitting_in_phylogenetics.html).

The big strength of distance-matrix reconstruction methods such as the neighbour-joining (NJ) algorithm (Saitou and Nei, 1987) lies in their speed. Distance matrices for several hundred OTUs can be clustered in a few seconds on a standard desktop or laptop computer. This time scales linearly with the number of bootstrap replicates one may want to analyse in order to get an idea of the impact of sampling error on different clades of the phylogeny (Felsenstein, 1985). These attributes make NJ analyses very suitable for preliminary phylogenetic data exploration. I would recommend that NJ trees are reconstructed for each individual sequence partition, along with a bootstrap analysis, and the resulting trees visually inspected. These analyses can be easily performed on Windows systems using the popular Molecular Evolutionary Genetic Analysis (MEGA4) software package (Tamura et al., 2007). It implements several distance-matrix methods under a restricted set of nucleotide substitution models, including the new Maximum
Composite Likelihood (MCL) method for estimating evolutionary distances between all pairs of sequences simultaneously, with and without incorporating rate variation among sites and substitution pattern heterogeneities among lineages. The MCL method can also be used to estimate transition/transversion bias and nucleotide substitution pattern without knowledge of the phylogenetic tree. It further offers editing of DNA sequence data from autosequencers, mining web databases, performing automatic and manual sequence alignment based on ClustalW, definition of sequence partitions and taxon groups, and performing diverse population genetics analyses and evolutionary hypothesis tests. A powerful tree editor is also integrated. MEGA4 and other native Windows programs can be run in a Linux desktop environment (via the Wine compatibility layer), and on Intel-based Macintosh computers under the Parallels program. Those users that like the command line and automating analyses on UNIX/Linux platforms via simple Shell, Perl or Python scripts will probably favour Joe Felsenstein’s PHYLogeny Inference Package (PHYLIP) (Felsenstein, 2004b), or David Swofford’s PAUP* package (Swofford, 2002), although the latter is not freely available (it is actually the only non-freely available software mentioned in this review). Distance trees reconstructed with any of these packages can be conveniently visualized and edited with tree editors such as Rod Page’s TreeView (Page, 2002) or Andrew Rambaut’s FigTree (Rambaut, 2009), both freely available for all major computing platforms.

Once the NJ trees are available, the user should compare the sequence/taxon composition of well-supported monophyletic lineages (clades) across the different phylogenies in order to identify obvious incongruent clustering of particular strains across them. This is a first and strong indication of potential HGT events. Large bacterial multilocus datasets often contain individuals that are recipients of xenologous alleles at a particular locus (Silva et al., 2005; Vinuesa et al., 2005c, 2008). Otherwise, single long branches that emerge either from a compact and homogeneous clade, or have an isolated position on the tree, with very low bootstrap support values, may be indicative of intragenic sequence mosaics. Such suspicious sequences should be further analysed, searching for evidence of sequence mosaicism (Vinuesa et al., 2005b, 2005c), using algorithms such as the bootscan (Martin et al., 2005) or RDP (Martin and Rybicki, 2000), implemented in the Windows package RDP3 (Martin, 2009).

For the inference of a robust species phylogeny using standard treeing methods and concatenated alignments it is critical that individuals (strains) showing clear evidence of being recipients of xenologous alleles, or harbouring gene mosaics, are removed from the supermatrix. The reason for it is that standard tree reconstruction methods assume a single underlying evolutionary history, reflected in the dichotomously branching pattern of the inferred phylogenetic hypothesis. Organisms (multilocus haplotypes) containing xenologous or chimaeric sequences violate this basic assumption, seriously compromising the accuracy of the phylogenetic estimate (Posada and Crandall, 2002; Schierup and Hein, 2000; Vinuesa et al., 2005b,c). Network graphing methods such as split-decomposition or neighbour-net are useful for visualizing parallel evolutionary pathways, an important alternative to classical treeing methods when recombination or HGT have been at play (Huson and Bryant, 2006; Huson et al., 2007; Kloeper and Huson, 2008; Vinuesa et al., 2005c).

Phylogeny inference under the maximum likelihood criterion using PhyML v3.0

The maximum likelihood (ML) criterion can be defined as the conditional probability of observing the data (D), given a hypothesis or model (H), that is: $L = \Pr(D|H)$. In phylogenetics this equation translates into finding the topology ($\Gamma$), set of branch lengths ($\gamma$) and substitution model parameters ($\sigma$) that maximize the probability of observing the data, i.e. the character states in the multiple sequence alignment. Therefore, we could re-write the expression given above as $L = \Pr(D|\Gamma, \gamma, \sigma)$. Phylogenetic inference under the ML criterion is therefore an optimization problem, although a difficult one. The difficulty derives from the nature of the phylogenetic model itself. Branch lengths and parameters of the substitution model are continuous variables, whereas the topology is a discrete parameter. Furthermore,
as more taxa/sequences are added, the number of unique, unrooted and strictly bifurcating topologies increases factorially, as indicated by the following expression: \((2n-5)!/(n-3)2^{n-3}\). Therefore, for four taxa we have three distinct bifurcating topologies, for eight taxa this number increases to 10,395, and for only 22 taxa it is somewhere around \(3 \times 10^{23}\), like ‘a mol of trees’! Searching for the ML tree therefore relies on sophisticated, heuristic methods that combine both discrete and continuous optimization procedures to explore a likelihood function of high dimensionality, which for most biological datasets defines a rugged landscape with multiple peaks. In practical terms this means that for large datasets tree searches often get stuck in local maxima. Since the search algorithms are heuristic, there is no guarantee of finding the global (best) ML phylogeny (Felsenstein, 2004a; Swofford et al., 1996). Therefore it is important to make a thorough search of the tree space, which can be most efficiently performed by launching multiple searches starting from different seed trees, generated for example by a sequential random addition algorithm (Swofford, 2002; Swofford et al., 1996). These seed trees are used to explore the astronomically large space of tree topologies that exists for large datasets by doing topological rearrangements of the seed trees by either nearest-neighbour interchange (NNI), subtree pruning and regrafting (SPR) or tree bisection and reconnection (TBR) moves or algorithms (Swofford et al., 1996). Each of these moves explores alternatives of the tree space, but the neighbourhood defined by the TBR moves is much larger than that defined by SPR moves, which in turn defines a much larger neighbourhood than NNI rearrangements of the original topology on which they operate. Hence the computing times required by these algorithms are as follows: TBR > SPR > NNI. Alternative topologies found by any of these moves that improve the score of the optimality criterion (e.g. parsimony or likelihood) are retained and used as the seed for the next round of branch swappings, until no further increase in the likelihood function is attained.

The most efficient algorithms are those that provide the best trade-off between their ability to maximize the likelihood function (escaping as many local maxima as possible) and the processor time used to achieve this. PhyML is one of the most efficient modern ML tree search algorithms currently available (Guindon and Gascuel, 2003). The last version of the software is PhyML v3.0 (Guindon et al., 2008), which implements three branch swapping methods (NNI, NNI+SPR and SPR only), while v2 only implemented NNI moves, and was therefore much more prone to get stuck in local maxima than the new version run with NNI+SPR or pure SPR moves (Guindon et al., 1999; Hordijk and Gascuel, 2005). By default, both PhyML v2 and v3 initiate the search from a BioNJ distance tree (Gascuel, 1997). To perform thorough searches of tree space, it is possible to use random trees in addition to the default BioNJ tree to seed the searches from multiple distinct points in tree space, a strategy that almost invariably finds better trees than those found in searches starting from the BioNJ topology. We have recently demonstrated this in a MLSA analysis of *Bradyrhizobium* strains that involved 62 multilocus haplotypes, for which there are \(~1.945514 \times 10^{181}\) distinct unrooted bifurcating topologies (Vinuesa et al., 2008).

Another advantage of PhyML v3 is that it implements fast approximate likelihood ratio tests (aLRTs) of branch support (Anisimova and Gascuel, 2006; Guindon et al., 2009), which is an efficient alternative to the traditional and time consuming bootstrap analysis (Vinuesa et al., 2008). In the latter publication we proposed an aLRT-based measure of the phylogenetic information content of different sequence partitions, which I will introduce in the next section. PhyML also implements classical bootstrap analyses. Finally, it works with both protein and DNA sequences, implementing many protein distance matrices and the full range of 203 models of nucleotide substitution of the GTR+I+G family. Therefore, PhyML v3 is an excellent, fast and accurate, all purpose tree search program under the maximum likelihood criterion, which can be run on all major computing platforms. The casual user can run PhyML searches remotely over the Internet using different web servers (Dereeper et al., 2008; Guindon et al., 2005). Another very useful, multipurpose and free multipurpose package that runs PhyML using a graphical user interface is SeaView4 (Gouy et al., 2009), which
I will introduce in a section below on alignment concatenation.

Looking again at the expression of the likelihood function \[ L = \Pr (D | I; \gamma, \sigma) \] makes it clear that in addition to the critical issue of doing a thorough search of tree space (more and more important as the number of taxa to analyse increases), it is also important to select the best fit substitution model for the data at hand. As mentioned in a previous section, this can be conveniently done with programs such as jModelTest (Posada, 2008) or ProtTest (Abascal et al., 2005). These programs do not only select for the best approximating model, but also provide ML estimates of the corresponding parameters (i.e. transition/transversion rate ratio, alpha value of the gamma distribution to model among-site rate heterogeneity, etc). When using PhyML, we only need to provide the selected model (i.e. its parameterization) to the program, and it will optimize the parameter values along with the topology and branch lengths in multiple iterations of topology, branch length and substitution model parameter optimizations.

Computing phylogenetic congruence and signal content of sequence partitions under the maximum likelihood criterion

As mentioned in the introduction, gene trees are expected to differ from each other and from the species tree because of the stochastic nature of the gene lineage sorting process during speciation. Lineage sorting occurs much earlier than speciation, and this difference in timing of the cladogenesis process varies from gene to gene, resulting not only in different gene tree topologies but also notable differences in branch lengths (Maddison, 1997). It is therefore important to gather the evidence from multiple loci to get a better estimate of the species phylogeny in which the gene trees are embedded (Edwards, 2009). There is the potential, however, that the subset of loci selected for a MLSA has strongly conflicting signals (Papke et al., 2007). The evolutionary process that generated them is often difficult to identify, as it could be the result of a mixture of forces or processes such as cross-species HGT or hybridization, mutational saturation, hidden paralogy or retention of ancestral polymorphisms due to incomplete lineage sorting (Degnan and Rosenberg, 2006; Maddison, 1997; Maddison and Knowles, 2006). This may result in a poorly resolved species tree if it is estimated from a concatenated supermatrix (Jeffroy et al., 2006; Kubatko and Degnan, 2007; Vinuesa et al., 2008), as generally done in microbial MLSA studies (Devulder et al., 2005; Lerat et al., 2003; Rokas et al., 2003; Vinuesa et al., 2008). If the signals from the different sequence partitions are strongly conflicting, then the estimated gene trees will not only be incongruent with each other, which is actually expected based on coalescence theory (Degnan and Rosenberg, 2006; Rosenberg, 2002; Rosenberg and Nordborg, 2002), but also with different subsets of concatenated partitions. In the worst case, the conflicting signals in the different partitions don’t contribute to a net increase in overall phylogenetic resolving power after concatenation and the inferred phylogeny won’t be a good estimate of the underlying species tree (Jeffroy et al., 2006; Kubatko and Degnan, 2007; Vinuesa et al., 2008). Given these potential problems, it is convenient to statistically evaluate the degree of conflict or ‘cooperativeness’ of the signals in the different sequence partitions to be used for the construction of a supermatrix. This information can then be used to make an informed choice of the partitions to use for the construction of the supermatrix (Jeffroy et al., 2006).

Once good estimates of the individual gene trees have been obtained, there are several methods to evaluate the significance of the phylogenetic congruence between them (Felsenstein, 2004a). PAUP* (Swofford, 2002) implements parsimony-specific methods such as the ‘incongruence length difference’ or ILD test (Farris et al., 1994), which has been strongly criticized (Darlu and Lecointre, 2002; Dolphin et al., 2000; Dowton and Austin, 2002), and the so-called ‘paired sites tests’, such as the t-test, Wilcoxon signed-rank test, and the RELL or Kishino–Hasegawa (KH) test (Felsenstein, 2004a). The basic idea behind these tests is that two trees can be compared for either their parsimony or likelihood scores. The expected log-likelihood of a tree is the average log-likelihood we would get per site as the number of sites grows without limit. Assuming independence of character evolution, if two trees have equal expected log-likelihood or
parsimony scores, then the differences in scores at each site will be drawn independently from some distribution whose expectation is zero. Statistical tests can be performed to test whether the mean of these differences is zero, therefore testing the hypothesis that both topologies under comparison are phylogenetically congruent, i.e. equally good estimates. If this null hypothesis is rejected, the topology with the best global parsimony or likelihood score is significantly better than the competing topology. These tests, in their one-sided version, are only adequate for pair-wise comparison of topologies between the most parsimonious or maximum likelihood trees and an alternative tree (e.g. a constrained topology or one found by a different reconstruction method). They are not suitable, however, to perform multiple tree comparisons, not even if some sort of correction for multiple tests such as the simple Bonferroni correction is applied (Goldman et al., 2000; Shimodaira and Hasegawa, 1999). For multiple tree comparisons the user should rather use the Shimodaira–Hasegawa test (SHT) (Shimodaira and Hasegawa, 1999), which implements a resampling method that approximately corrects for testing multiple trees. However, this test loses power if there are many highly unlikely trees in the set of trees to compare. Shimodaira and Hasegawa (1999) warn against including too many trees in the analysis, as it will dilute the power of discriminating among the plausible trees. A more powerful test is the ‘approximately unbiased test’ (Shimodaira, 2002), implemented in CONSEL under the maximum likelihood criterion, along with the classical KH and SH tests (Shimodaira and Hasegawa, 2001). The SH, KH and a expected likelihood weights (ELWT) tests (Strimmer and Rambaut, 2002) are also implemented in TREE-PUZZLE (Schmidt et al., 2002), a powerful quartet-based maximum likelihood phylogenetic inference and hypothesis testing program, freely available for all platforms.
Concatenation of individual sequence alignments and supermatrix filtering

After the generation of reliable multiple sequence alignments and the performance of saturation, model selection, phylogenetic congruence and signal additivity analyses, the user is ready to concatenate selected alignments into a ‘supermatrix’ or ‘superalignment’, the dominant strategy used by microbial systematists to infer species trees.

This step can be easily done with Windows programs such as BioEdit (Hall, 1999) or DAMBE (Xia and Xie, 2001). This operation can be performed on multiple platforms with the excellent and versatile SeaView4 package (Gouy et al., 2008), which, among other goodies, integrates multiple sequence alignment generation and manipulation, with phylogeny inference under distance, parsimony and maximum likelihood criteria. It also implements simple but convenient functions for tree visualization and edition. Alternatively, you could use web services such as FaBox (Villesen, 2007), or write your own simple Perl scripts, eventually making use of the ‘cat’ method of the Bio::Align::Utilities module from the BioPerl suite (Stajich et al., 2002). An important note to make at this point is that you should, of course, have the sequences equally sorted in all partitions before concatenating them, otherwise you will be generating ‘in silico’ sequence chimaeras that will seriously compromise the accuracy of the species tree estimation.

Once concatenated, and before starting the proper estimation of the species phylogeny, the sequences should be collapsed to haplotypes, that is, repeated multilocus sequences should be dereplicated. Again, Windows programs such as DAMBE (Xia and Xie, 2001) will do this automatically. A convenient cross-platform multiple sequence alignment editor and analysis workbench is Jalview2 (Waterhouse et al., 2009), which can also collapse repeated sequences to haplotypes. The FaBox web server can also do this task (Villesen, 2007). Writing ad hoc Perl code for this job will allow you to perform this operation on as many files as needed in an automatic fashion.

Some phylogeny programs such as MrBayes (Ronquist and Huelsenbeck, 2003), RAxML (Stamatakis, 2006) or PAUP* (Swofford, 2002) allow for partitioned models. In order to make use of such models, which provide a much better fit (i.e. are more realistic) than selecting a single best-approximating model for the entire supermatrix (Nylander et al., 2004; Posada and Crandall, 2001; Vinuesa et al., 2005b,c), the user needs to keep track of the coordinates of the individual partitions within the supermatrix.
Table 3.2 Relative performance of individual molecular markers and some of their combinations assessed by computing descriptive statistics of Shimodaira-Hasegawa-like P values of bipartition support values for the corresponding maximum likelihood phylogenies inferred with PhyML

<table>
<thead>
<tr>
<th>Partition</th>
<th>mean</th>
<th>median</th>
<th>std_dev</th>
<th>variance</th>
<th>%_NRB</th>
<th>%_PRB</th>
<th>%_MRB</th>
<th>%_WRB</th>
<th>%_HRB</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpD483</td>
<td>0.74</td>
<td>0.82</td>
<td>0.24</td>
<td>0.0600</td>
<td>32.35</td>
<td>32.35</td>
<td>23.53</td>
<td>5.88</td>
<td>5.88</td>
</tr>
<tr>
<td>rpoB771</td>
<td>0.73</td>
<td>0.83</td>
<td>0.27</td>
<td>0.0711</td>
<td>26.47</td>
<td>35.29</td>
<td>20.59</td>
<td>11.76</td>
<td>5.88</td>
</tr>
<tr>
<td>glnII591</td>
<td>0.83</td>
<td>0.86</td>
<td>0.19</td>
<td>0.0355</td>
<td>11.76</td>
<td>38.24</td>
<td>17.65</td>
<td>20.59</td>
<td>11.76</td>
</tr>
<tr>
<td>recA510</td>
<td>0.73</td>
<td>0.88</td>
<td>0.30</td>
<td>0.0930</td>
<td>26.47</td>
<td>14.71</td>
<td>41.18</td>
<td>14.71</td>
<td>2.94</td>
</tr>
<tr>
<td>glnIIrecA</td>
<td>0.85</td>
<td>0.94</td>
<td>0.23</td>
<td>0.0538</td>
<td>14.71</td>
<td>8.82</td>
<td>29.41</td>
<td>11.76</td>
<td>35.29</td>
</tr>
<tr>
<td>glnIIrecArpoB</td>
<td>0.84</td>
<td>0.95</td>
<td>0.23</td>
<td>0.0507</td>
<td>17.65</td>
<td>14.71</td>
<td>17.65</td>
<td>11.76</td>
<td>38.24</td>
</tr>
<tr>
<td>atpDglnIIrecA</td>
<td>0.85</td>
<td>0.96</td>
<td>0.26</td>
<td>0.0674</td>
<td>14.71</td>
<td>5.88</td>
<td>26.47</td>
<td>23.53</td>
<td>29.41</td>
</tr>
<tr>
<td>atpDglnIIrpoB</td>
<td>0.85</td>
<td>0.96</td>
<td>0.23</td>
<td>0.0550</td>
<td>14.71</td>
<td>11.76</td>
<td>17.65</td>
<td>35.29</td>
<td>20.59</td>
</tr>
<tr>
<td>atpDrecArpoB</td>
<td>0.86</td>
<td>0.96</td>
<td>0.20</td>
<td>0.0410</td>
<td>14.71</td>
<td>14.71</td>
<td>17.65</td>
<td>26.47</td>
<td>26.47</td>
</tr>
<tr>
<td>atpDglnIIrecArpoB</td>
<td>0.87</td>
<td>0.97</td>
<td>0.20</td>
<td>0.0406</td>
<td>20.59</td>
<td>0.00</td>
<td>20.59</td>
<td>26.47</td>
<td>32.35</td>
</tr>
</tbody>
</table>

*aThe mean, median, standard deviation and variance of SH-like P values were computed on ML trees inferred with PhyML v3 under the ‘best’ branch swapping method and best fitting substitution models selected with jModelTest.

*bPercentage of non resolved bipartitions, with P < 0.70.

*cPercentage of poorly resolved bipartitions, with 0.70 ≤ P < 0.85.

*dPercentage of moderately resolved bipartitions, with 0.85 ≤ P < 0.95.

*ePercentage of well resolved bipartitions, with 0.95 ≤ P < 0.99.

*fPercentage of highly resolved bipartitions, with P ≥ 0.99.
Phylogenetic inference using partitioned models under ML and Bayesian optimality criteria

A current limitation of PhyML v3.0 is that it does not implement partitioned models (Nylander et al., 2004; Vinuesa et al., 2005b,c), that is, it will not allow defining a best-fitting model for each gene partition. TreeFinder (Jobb et al., 2004) is an easy-to-use analysis environment for molecular phylogenetics under the maximum likelihood criterion written in ANSI C and Java, that implements fast heuristics for tree searching and evaluation of phylogenetic hypotheses using large datasets. In addition, it provides a user-friendly graphical interface and a phylogenetic programming language. Among its many features, it implements partitioned models and a broad range of phylogenetic congruence tests. Another powerful, versatile and very fast maximum likelihood tree search algorithm that allows for partitioned models is implemented in RAxML (randomized accelerated maximum likelihood for high performance computing). It is a sequential and parallel program for inference of large phylogenies under the maximum likelihood criterion (Stamatakis, 2006). Low-level technical optimizations such as efficient memory use, and the use of the GTR+CAT approximation as replacement for GTR+Gamma yield a program that is extremely fast and accurate, allowing the inference of ML trees using multiple sequence partitions and hundreds or even thousands of protein or nucleotide sequences. It also runs on two publicly accessible web-servers (Stamatakis et al., 2008). All these programs are freely available.

An alternative to maximum likelihood within the realm of statistical phylogenetics is the Bayesian framework (Holder and Lewis, 2003; Huelsenbeck et al., 2001, 2002). MrBayes is currently the most popular software implementation for Bayesian estimation of phylogenies (Ronquist and Huelsenbeck, 2003). It is freely available for all major platforms from http://mrbayes.csit.fsu.edu/. Excellent tutorial materials are available from that site. MrBayes implements a Metropolis–Hastings coupled Markov-chain Monte Carlo (MCMC) tree sampler that permits the combination of information from different partitions evolving under different stochastic evolutionary models. This allows the user to analyse heterogeneous data sets consisting of different data types – e.g. morphological, nucleotide, and protein – and to explore a wide variety of structured models, mixing partition-unique and shared parameters (Nylander et al., 2004; Vinuesa et al., 2005b, 2005c). In addition to standard tree searches, MrBayes also implements constrained searches and the analysis of selection at the molecular level using codon-based substitution models. The program employs MPI to parallelize Metropolis coupling on Macintosh or UNIX clusters, making it suitable for the analysis of large datasets.

Figure 3.2 shows a Bayesian species tree inferred from our multilocus sequence data set using MrBayes and the command block shown below:

```
BEGIN mrbayes;
log start filename=37U_atpD_glnII_recA_rpoB_bygene_RUN1.log1 append;
outgroup Rho_pal_1;
charset atpD = 1-483; charset glnII = 484-1074;
charset recA = 1075-1584; charset rpoB = 1585-2355;
partition by_gene = 4: atpD, glnII, recA, rpoB;
set autoclose=yes nowarnings=yes; set partition = by_gene;
Prset ratepr = variable;
Lset applyto =(all) nst=6
rates=invgamma Ngammacat=4;
unlink shape=(all) pinvar=(all)
statefreq=(all) revmat=(all);
mcmc ngen=3000000 printfreq=10000
samplefreq=100 nchains=4 nruns=2
temp=0.18;
```

This block specifies four sequence partitions, each evolving at its own rate, under the best-fitting GTR+I+G model selected for each partition by MrModelTest 2.3 (Nylander, 2004). Notice that by unlinking the state (nucleotide) frequencies, substitution rates (revmat), shape parameter of the gamma distribution (shape), and proportion of invariant sites (pinvar), the values for each of these model parameters will be optimized independently for each of the four partitions. The MC3
Figure 3.2 Bayesian species phylogeny estimated under gene-partitioned, best fitting models, based on a supermatrix of 2355 characters for a selection of 37 unique atpD+glnII+recA+rpoB sequences (multilocus haplotypes) of representative strains of 6 Bradyrhizobium species previously studied by Vinuesa et al. (2008). The numbers on the nodes denote the Bayesian posterior probabilities (PP)/Shimodaira–Hasegawa-like (SH) bipartition support values. Asterisks indicate any of these values when < 0.90. The Bayesian PP values were computed from 10,000 post-burnin trees pooled from two independent, convergent, MC3 simulations run for $3 \times 10^6$ generations. The MrBayes block used to run the analysis is presented in the main text. The SH-like support values were obtained with PhyML v3, run with the ‘best’ branch swapping mode (see text for the details). The bar scale represents the expected number of substitutions per site, under the mixed model.

The chain will be run for $3 \times 10^6$ generations, sampling parameters from the chain every 100th in order to avoid excessive autocorrelation between successive samples. The MCMC command settings actually indicate that two replicate runs will be executed, each with three heated chains (heat parameter set to 0.18) and a cold chain. These same settings were used for two additional runs, so that we ended up with data from six independent chains (three runs, each with two replicates, executed on three processors).

One of the critical issues in Bayesian MCMC-based analyses is to obtain convincing evidence that the Markov chain has been run long enough so that the collection of trees and parameters sampled from it are a good representation of the

posterior probability distribution (Huelsenbeck et al., 2002). This evidence can be obtained indirectly from different sources, but none is truly conclusive. The key consideration here is to run several replicate MC^3 chains, each starting from independent random trees (the default setting in MrBayes). The first check that should be performed is to examine the so-called trace or generation plots for all parameters, in which the log probability of the data given the parameter values are plotted against the number of generations (MCMC cycles), as shown in Figure 3.3. In the beginning of the run the log likelihood of the cold chain typically increases rapidly. This phase of the run is referred to as the burnin, and the samples taken from the MCMC chain during these early steps of the search phase are typically discarded. Ideally, however, the chain should be run long enough, so that the burnin phase has no noticeable effect on the results. Once the likelihood of the cold chain stops to increase and starts to randomly fluctuate within a more or less stable range, the run may have reached stationarity. At this stage, the cold and heated chains should ideally be exchanging states frequently. As a rough rule of thumb, an efficient Metropolis–Hastings MCMC sampler will have acceptance rates somewhere in the range of 10–70%. If they are much lower, a first possibility to improve this misbehaviour of the chains is to set a lower temperature value, which is set to temp=0.2 by default. Figure 3.3 reveals that two of the chains (labelled as run1_rep1 and run2_rep1) converged at a relatively late point of the runs (generations > 2,250,000), while chains run1_rep2 and run3_rep2 did not find the region of higher posterior probabilities.

Figure 3.3 Log likelihood traces (generation plots) of six MC^3 chains, each run with three heated chains and a cold chain, as indicated in the MrBayes block shown in the text. Runs labelled as run2_rep2 and run3_rep1 were used for the Bayesian species phylogeny estimation presented in Figure 3.2. These plots were generated with R (R Development Core Team, 2009) Note the good mixing of the chains, as revealed by the fast random fluctuation of the lnL values within a more or less stable range, indicating that the cold and heated chains are exchanging states frequently. Runs labelled as run1_rep2 and run3_rep2 got stuck in a suboptimal region of parameter space. The y-axis was severely shortened in order to present a better resolution of the chains in the region of higher posterior probabilities.
hit by the other four chains. In these analyses
the heating parameter was set to temp=0.18,
but should probably have been lowered further
to ~0.15. This would allow the heated chains
to explore peaks in parameter space that are ‘further
away’, that is, separated from currently sampled
local maxima by ‘deeper valleys’. Tracer (Rambaut
and Drummond, 2007) is a very convenient tool
for examining multiple trace files (the *.p param-
eter files generated by MrBayes) at once, providing
convenient summary statistics for each run,
along with marginal density, joint marginal and
trace plots for each parameter. It calculates the
effective sample size (ESS), that is, the number of
effectively independent draws from the posterior
distribution that the Markov chain is equivalent
to. This parameter should be well over 100,
the higher the better (ESS for the global lnL of
run3_rep1 is 993.6577). Tracer is written in Java
and is therefore cross-platform.

What we can learn from this analysis is that
when using relatively large datasets and complex
partitioned models, such as the one used in this
analysis (GTR+I+G model, unlinked for each
partition), long MCMC runs have to be per-
formed and replicated as many times as possible.
MrBayes’ default values of $1 \times 10^6$ generations
($\text{mcmc ngen}=1000000$) and two replicates ($\text{nruns}
= 2$) are clearly too low for such analyses and
should at least be triplicated, as done herein.

Based on the trace plots shown in Figure 3.3
we should limit our downstream analysis to the
data sampled from chains run3_rep1 and run2_rep2,
discarding at least the first 7000 samples of
the latter run as burnin. MrBayes provides the
important and convenient $\text{sump}$ and $\text{sumt}$ func-
tions to summarize the model parameters and
trees sampled from selected chains, respectively.
The $\text{sump}$ function can be executed for single runs
using the following line: $\text{‘sump filename=\text{file_}
basenam}_\text{run2_rep2 burnin}=7001 \text{ nruns}=1\text{’}$.
Notice that by default MrBayes will execute the
$\text{sump}$ command on all replicate runs issued from
a single MrBayes command block. However, only
the two replicates of run 2 found the region of
highest posterior probability. For this particular
run and replicates we could have issued the com-
mand $\text{‘sump filename=\text{file basenam}_run2}
burnin=22501 \text{ nruns}=2\text{’}$ in order to summarize
the parameters from both replicates, although
discarding the first 22501 suboptimal samples
of run2_rep2 (see Figure 3.3). The $\text{sumt}$ output
includes a rough generation plot and a table with
summary statistics for the model parameters for
each partition, including their 95% credibility
intervals. The $\text{sumt}$ command has a similar syntax,
allowing the user to get a summary of the sampled
trees ($\text{‘sumt filename = file basenam}_\text{run3_}
\text{repl burnin}=3001 \text{ nruns}=1\text{’}$). The latter com-
mand would write a consensus tree to disk with
the *.con extension based on 27000 sampled
trees ($3 \times 10^6$ generations sampled every 100th)
from run3_rep1. This is the file that the user
should open with a tree editor such as TreeView
or FigTree to visualize the estimated Bayesian
phylogeny with branch support values indicated
as posterior probabilities.

Further convergence diagnostics can be
generated by MrBayes from selected pairs of
independent runs by issuing the $\text{comparetree}$
command using the following syntax: $\text{‘compare-
tree file}\_\text{filename1=\text{file basenam}_run3_}
\text{repl1.t file}\_\text{basename}_\text{run2_}
\text{repl2.t burnin}=7501\text{’}$.
This command will display a rough bivariate
plot for a random subsample of clade probability
values found in filename1 and those for the same
clades found in filename2. It will also display
rough generation plots for different tree-to-tree
distance measures. Figure 3.4 shows the bivari-
ate plot of clade probabilities for run3_rep1 vs.
run2_rep2, according to the code line shown
above. This plot and the Pearson correlation
analysis shown (Figure 3.4) were generated with
R (R Development Core Team, 2009) using the
*.comp.dist file written to disk by MrBayes as the
source data. This analysis shows a very strong and
significant correlation of the support values for
the same clades in both runs, suggesting that the
corresponding trees were sampled from the same
posterior distribution.

The AWTY (Are We There Yet?) online
convergence diagnostics tool can further aid in
deciding if the chains have been run long enough
and have converged (Nylander et al., 2008). It
provides a graphical representation of additional
parameters such as convergence rates of poste-
rior split probabilities and branch lengths. AWTY
therefore complements the diagnostics described
and available in Tracer and MrBayes.
As mentioned before, convergence diagnostics can never prove that the globally best phylogeny has been found. The possibility exists that there are still better tree islands to be found, but it should be clear at this point that long runs, well selected substitution models, clean data matrices and careful selection of priors and proposals are key conditions for making robust Bayesian inferences of phylogeny. However, we can be quite confident in our estimate (Figure 3.2), because in addition to the evidence gathered from diverse convergence diagnostics, we did also make the inference under the ML criterion using PhyML v3. Both trees are nearly identical, the Robinson–Foulds (R–F) symmetric difference between them being 2, as computed with the Treedist program from the Phylip package (Felsenstein, 2004b). That is, they differ in only a single branch from each other at the base (root) of the tree. This difference is irrelevant, since MrBayes places the outgroup as a basal polytomy, whereas PhyML does not. Furthermore, the SH-like branch support values estimated with PhyML were mapped on the Bayesian tree shown in Figure 3.2. The first number on the bipartitions of the phylogeny corresponds to Bayesian posterior probabilities, whereas the second number corresponds to the SH-like support values. Bayesian posterior probabilities (PP) of clade support have been criticized as providing ‘overcredibility’ (Alfaro et al., 2003; Buckley, 2002; Douady et al., 2003; Erixon et al., 2003), whereas the SH-like tests are known to be very conservative (Shimodaira and Hasegawa, 1999), as mentioned in the section on phylogenetic congruence tests. Therefore we could take the PP and SH-like support values as upper and lower credibility bounds of branch support, respectively. Since both values are highly significant for most bipartitions, we can safely conclude that the species phylogeny estimate presented in Figure 3.2 is robust and well resolved.

**Figure 3.4** Bivariate plot of clade probabilities for run3_rep1 vs. run2_rep2 (see Figure 3.3) computed by the ‘comparerep’ command of MrBayes, according to the command shown in the text. This plot and the Pearson correlation analysis were generated with R (R Development Core Team, 2009). The f(T|X) labels on the axes denote the clade posterior probabilities of the two runs compared.
Prospect: estimating species trees from multilocus data without concatenation – an example using BEST

Recent empirical studies have demonstrated that concatenation of sequences from multiple genes prior to phylogenetic analysis often results in inference of a single, well-supported phylogeny (Devulder et al., 2005; Lerat et al., 2003; Rokas et al., 2003; Vinuesa et al., 2008). Theoretical work, however, has shown that the coalescent can produce substantial variation in single-gene histories, reflected in a high diversity of topologies and branch lengths across gene trees (Degnan and Rosenberg, 2006; Degnan and Salter, 2005; Rosenberg, 2002). Using simulation, these ideas were recently tested by Laura Kubatko and James Degnan to examine the performance of the concatenation approach under conditions in which the coalescent produces a high level of discord among individual gene trees and showed that it leads to statistically inconsistent estimation in this setting (Kubatko and Degnan, 2007). Furthermore, they showed in that study that the use of the bootstrap to measure support for the inferred phylogeny can result in moderate to strong support for an incorrect tree under such conditions. These results highlight the importance of incorporating variation in gene histories into multilocus phylogenetics (Degnan and Rosenberg, 2009; Edwards, 2009). This variation in genealogies (topologies and branch lengths) cannot be taken into account using the concatenation strategy, not even when partitioned models are used with program such as MrBayes, since they impose a common topology and set of branch lengths on the estimate made from each partition. The concatenation approaches estimate a phylogeny that reflects some average of gene trees but do not explicitly estimate species trees. The latter are the trees that really matter to systematists (Edwards, 2009; Edwards et al., 2007).

This shortcoming of standard treeing methods to infer species trees based on supermatrices has been recently addressed by several groups, who have developed and released first versions of software that will provide estimates of the species trees from multilocus sequence data, without concatenation.

Liang Liu and Dennis Pearl developed BEST (Bayesian Estimation of Species Trees) (Edwards et al., 2007; Liu, 2008; Liu and Pearl, 2007; Liu et al., 2008). It is based on a modified MrBayes code base that implements a Bayesian hierarchical model to jointly estimate gene trees and the species tree from multilocus sequences. The technique of simulated annealing is adopted along with Metropolis coupling as performed in MrBayes to improve the convergence rate of the Markov Chain Monte Carlo algorithm. BEST takes advantage of the information contained in multiple gene trees to perform a hierarchical Bayesian analysis to estimate the topology of the species tree, divergence times in coalescent units and ancestral population sizes. The BEST model is based on two conditional probability distributions: the probability distribution \( f(D|G) \), which is the previously discussed likelihood function used to estimate gene trees, and the probability distribution \( f(G|S) \), which is the likelihood of gene trees (G) conditional on the species tree (S), as derived from the multispecies coalescent (Degnan and Rosenberg, 2009). The algorithm samples from the joint posterior distribution over a set of gene trees and the species tree. Therefore the model involves a larger number of parameters to be estimated from the data than in a standard Bayesian or maximum likelihood gene tree search. This additional computational burden can make it very challenging for the BEST algorithm to converge if the datasets are large and ‘misbehaved’. In particular, the BEST model (as all other currently available species tree inference algorithms) assumes that there is no HGT, hybridization, gene flow or intragenic recombination in the data, but free recombination between loci (Liu, 2008). It is likely that many microbial multilocus datasets violate some of these assumptions. It is therefore critical that the data sets are carefully chosen to minimize the impact of systematic bias and to improve convergence rate among replicate runs, as exemplified below.

Figure 3.5A shows the species tree estimate obtained with BEST using the type strains of each taxon in a replicate run using only the glnII, recA and rpoB partitions, default prior values, the HKY85+G substitution model for each partition, \(10 \times 10^6\) MC\(^3\) generations with two chains \((temp=0.17)\), sampling every 100th
and discarding 25% of the samples as burnin. Exclusion of the *atpD* locus was critical to achieve convergence. This and previous studies revealed that this sequence partition is conflictive in several respects, such as being particularly prone to HGT and intragenic recombination (Vinuesa et al., 2005c, 2008), yielding the gene tree with the strongest topological discordance among those analysed (see Figure 3.1).

The species tree shown in Figure 3.5A is highly resolved, all partitions having a posterior probability ≥ 0.98. When compared with the ML tree shown in Figure 3.5B, estimated with PhyML v3.0 for the corresponding supermatrix under the HKY+G model using a BioNJ and 5 random starting trees and the ‘best’ search mode (NNI+SPR), a single difference is apparent, namely, that *B. japonicum* I and Ia are sister species in the former but not in the latter tree. This difference may seem trivial, but highlights the difference between estimating some sort of an average gene tree from a concatenated alignment and estimating a species tree that takes into account topological and branch length heterogeneity in the gene trees underlying the species tree. Pairwise Robinson–Foulds symmetric distances between the BEST species tree, and the PhyML trees for the *glnII*+, *recA*+ and *rpoB* supermatrix and the *glnII*, *recA* and *rpoB* partitions were 2, 4, 0 and 6, respectively, illustrating the diversity of topologies that can be found even within small datasets such as this case study.

It will be interesting to evaluate the impact of adding more loci and strains (alleles) per species on the estimation of the species tree using the traditional concatenation approach and the more rigorous species tree inference using *ad hoc* algorithms such as BEST (Liu, 2008). A maximum likelihood alternative called STEM (Species Tree Estimation using Maximum likelihood) has been
recently released (Kubatko et al., 2009). STEM provides a capability for searching the space of species trees for a collection of $k$ species trees with high likelihood, where $k$ is set by the user. These and other new software packages and methods represent promising and exciting avenues for future MLSA-based research. Extensive empirical and simulation work is now required to identify potential pitfalls of these methods and aspects of the current software implementations that need to be improved to make them broadly useful for practicing microbial systematists and evolutionary biologists.

Acknowledgements
I would like to thank the editors for their invitation to write this chapter, as well as for their suggestions and criticism on a first draft of the manuscript. I gratefully acknowledge funding from the Autonomous University of Mexico through grant DGAPA/PAPIIT-UNAM IN201806–2, and CONACyT-Mexico through grant P1–60071.

References


