Phylogenetic information of genes, illustrated with mitochondrial data from a genus of gastropod molluscs

SIMON F. K. HILLS*, STEVEN A. TREWICK and MARY MORGAN-RICHARDS

Ecology Group, Institute of Natural Resources, Massey University, Private Bag 11 222, Palmerston North, New Zealand

Received 10 February 2011; revised 11 June 2011; accepted for publication 11 June 2011

A critical assessment of sequencing markers is desirable to ensure that they are appropriate for the speciÞc questions that are to be addressed. This consideration is particularly important where the data set will be used in highly sensitive analyses such as molecular clock studies. However, there is no standard practice for marker assessment. We examined the mitochondrial DNA sequences of a genus of marine molluscs to assess the relative phylogenetic signal of a number of genes using an extension of splits-based spectral analysis. With a data set of almost 8 kb of DNA sequences from the mitochondrial genome of a lineage of marine molluscs, we compared the phylogenetic information content of six protein coding, two ribosomal DNA, and 12 transfer RNA genes. Split-support graphs were used to identify which genes contributed a relatively low signal-to-noise ratio of phylogenetic information. We found that cox2 and atp8 did not perform well for reconstruction at the within-genus level for this lineage. Consideration of nested subsets of taxa improved the resolution of relationships among closely related species by reducing the time frame over which evolutionary processes have occurred, allowing a better Þt for models of DNA substitution. Through this Þne-tuning of available data it is possible to generate phylogenetic reconstructions of increased robustness, for which there is a greater understanding of the underlying signals in the data. We recommend a suitable mitochondrial DNA fragment and new primers for intergeneric studies of molluscs, and outline a general pipeline for phylogenetic analysis. © 2011 The Linnean Society of London, Biological Journal of the Linnean Society , 2011, **104**, 770Ð785.

ADDITIONAL KEYWORDS: Alcithoe Ð data exploration Ð marker assessment Ð molecular evolution Ð phylogenetic splits Ð spectral analysis.

INTRODUCTION

As phylogenetic analysis has become more sophisticated, the molecular evolution of the genes used to infer species evolution has required increased scrutiny. Down-stream analyses, such as molecular-clock studies, are increasingly a common focus in molecular phylogenetics, and such techniques tend to be highly sensitive to incongruent signals in the underlying data (Ho & Phillips, 2009). To ensure the accuracy of inferences made with such techniques it is necessary to have data that will infer robust phylogenetic trees. In order to conÞdently build robust phylogenies one needs to critically assess sequence data to create molecular data sets that are best suited to different levels of divergence.

Comparative approaches allow the phylogenetic utility of markers to be determined (Graybeal, 1994). It is desirable to know whether there is sufficient data for the phylogenetic estimation to reßect the evolutionary history of the entire genome (and therefore the organism), rather than the evolutionary history of

5(toce9hebT&d)-35 197.413 769.851 199.119 769.128 199.119 767

middle ground, i.e. phylogenies of related species, within genera or among sister genera, do not usually address the question of the phylogenetic information content of the markers used. The great majority of phylogenetic studies of animals in this intermediate range have concentrated on mitochondrial cox1, 16S, cytB, and nuclear 18S and 28S sequencing markers. This reliance is largely based on the availability of universal PCR primers, but the reliability of DNA extraction to recover mitochondrial sequences has also contributed. However, universal markers have some limitations. Because of the need to anneal across broad taxonomic ranges, most universal primers target highly conserved DNA fragments. The consequence of this requirement is that the sequencing markers obtained are limited in the taxonomic depths for which they have robust phylogenetic resolution. For example, the universal nuclear markers, such as 18S and 28S, tend to lack resolution for shallow intragenus-level divergence. Conversely, the three mitochondrial genes commonly used for intraspecies analysis (cox1, 12S, and 16S) are more rapidly evolving sequences, but 12S and 16S can be difficult to align for deeper relationships, and cox1 rapidly becomes saturated at third codon positions, and therefore loses resolution (Simon et al., 1994; Roe & Sperling, 2007).

The relative information content of mitochondrial genes has been investigated, and a range of signals has been found in different genes (e.g. Corneli & Ward, 2000; Mueller, 2006; Paton & Baker, 2006). Some studies separate mitochondrial genes into classes based on the level of phylogenetic usefulness (e.g. Zardoya & Meyer, 1996); however, the majority of such studies deal with vertebrate lineages or very broad evolutionary distances (Simon et al., 1994). It is therefore likely that the patterns of gene variability observed are not the same in all data sets. An analysis of the utility and critical selection of the markers to be used to resolve a phylogeny would lend greater conÞdence to the resulting phylogenetic hypothesis, and would provide a foundation from which to assess challenging phylogenetic relationships. Such an analysis is expected to aid marker choice for studies of similar organisms. As the ease and cost-effectiveness of DNA sequencing increases, the reliance on universal primers should diminish. Thus targeting genes suitable for a given type of analysis will be a more feasible strategy, rather than marker selection by convenience. An additional beneÞt of characterizing the phylogenetic utility of markers is to provide information as to the most cost-effective regions to sequence from poor quality DNA samples, such as ancient DNA and extractions from poorly preserved museum specimens.

Assessing the robustness of molecular data sets is not a trivial problem. Robustness can be judged by both congruence among different tree-building methods (where a more robust signal in the data is likely to result in more consistent results from disparate methods) and by the support for inferred clades. High bootstrap values and Bayesian posterior probabilities are often considered to be indicative of ÔtrueÕ tree topology. These measures are only indicative of accuracy if the evolutionary model is accurate; however, this is rarely the case for biological data. As such, misleading signals can occur, high bootstrap values can be obtained for incorrect topologies (Phillips, Delsuc & Penny, 2004), and Bayesian support can be inßated and not representative of the probability of the correct resolution of clades (Simmons, Pickett & Miya, 2004). It is preferable to assess the robustness of a given phylogeny by exploring the signal in the underlying data, independently of the tree. This allows for the assessment of the validity of bootstrap and Bayesian support values, and also for the evaluation of the signals behind clades with low bootstrap and Bayesian support values.

One method of doing this is through the examination of phylogenetic splits, which represent bipartitions of taxa in the DNA data set (Bandelt & Dress, 1992). Any molecular data set will contain one or more sets of compatible splits, and for each compatible set there will be a set of incompatible splits. Any branch in a phylogenetic tree represents a split dividing the represented taxa into two sets. A set of splits is compatible if, when combined, they describe all or part of a fully resolved phylogenetic tree for the taxa involved; if not, they are incompatible (Bryant &

visualization of conßict and support for all signals in a data set, independently of a tree. When referenced to a tree generated from the same data the spectral analysis can be used to diagnose weaknesses in that tree, and reinforce likely true signals. Identifying genes that provide poor phylogenetic information is achieved by a comparison of signal and conßict for splits provided by individual genes. Previous studies have shown the potential of spectral analysis for this purpose (Lento et al., 1995; Wagele & Mayer, 2007). Furthermore, when large sequence data sets exist it is likely that the selection of a subset of the genes that maximize the signal-to-noise ratio will result in phylogenies of greater robustness (Jeffroy et al., 2006).

Here we compare mitochondrial genes from the New Zealand marine gastropod genus *Alcithoe*. These snails are benthic, direct-developing, carnivorous neogastropod molluscs. Molluscs represent a group for which there has been limited molecular phylogenetic analysis considering the extensive species diversity. As such, molluscan phylogenetics has relied heavily on universal markers and, to date, there has been little consideration of the relative information content of these genes. In the *Alcithoe* a prevalence of large intraspeciÞc and low interspeciÞc morphological variation has made the phylogeny of the genus difficult to estimate using morphological characters, and to date there has been no molecular treatment. The taxonomy of both extant and extinct *Alcithoe* is well described, although its stability is subject to the vagaries of morphological characters. Based on shell characteristics, 17 living species are recognized, three of which are subdivided into either two or three subspecies (Bail & Limpus, 2005). There has been a recent increase in the number of extant taxa recognized as a result of the development of new commercial Þsheries and research trips that have yielded new specimens from deeper waters. It is possible that several of these new putative taxa represent local forms of known species. Although it is possible that widespread variable taxa may represent species complexes, the recent history of *Alcithoe* taxonomy is dominated by the synonymy of species, as new samples have bridged apparent morphological gaps.

We sequenced more than 7 kb of mitochondrial DNA from eleven *Alcithoe* species, covering nine genes. Although we aim to infer a robust phylogeny for the *Alcithoe*, our main goal here is to demonstrate a technique for the assessment of the suitability of the mitochondrial genes comprising this data set for intrageneric phylogenetic studies. To do this we will:

1. Explore the signal in each of the genes separately using summary statistics and tree-building methods.
2. Assess the comparative phylogenetic utility of each gene, using splits to examine the relative contribution of signal and noise in a novel approach using spectral analysis to compare the combined spectra of all genes.
3. Make recommendations as to the suitability of the genes comprising this data set for molluscan phylogenetic studies, and recommend a phylogenetic analysis pipeline.
4. Infer a robust phylogeny for the *Alcithoe* using the identiÞed genes.

MATERIAL AND METHODS

Taxon sampling

Table 1. Volute species used to study the phylogenetic information in 9 mitochondrial genes

Genus	Species	Voucher number	Sample Location		GenBank accessions
Alcithoe	aillaudorum	NB 1024	Isle des Pins	New Caledonia	JN379020 JN379030
Alcithoe	arabica	M.279684	Wellington	New Zealand	JN182223
Alcithoe	benthicola	M.183806	Coromandel	New Zealand	JN182217
Alcithoe	Þssurata	M.183785	Coromandel	New Zealand	JN182225
Alcithoe	ßemingi	M.183833	Chatham Rise	New Zealand	JN182218
Alcithoe	fusus	M.279683	Nelson	New Zealand	JN182220
Alcithoe	jaculoides	M.274972	North Island East Coast	New Zealand	JN182221
Alcithoe	larochei	M.274116	North Island East Coast	New Zealand	JN182227
Alcithoe	larochei tigrina	M.183799	Coromandel	New Zealand	JN182224
Alcithoe	lutea	NIWA 30452	Challenger Plateau	New Zealand	JN182219
Alcithoe	pseudolutea	M.183802	Coromandel	New Zealand	JN182222
Alcithoe	wilsonae	M.190062	South Island	New Zealand	JN182228
Cymbiola	pulchra subelongata	M.273459	Queensland	Australia	JN182216
Odontocymbiola	simulatrix	MZSP44320	Cabo Santa Marta	Brasil	JN379019 JN379027
Athleta	studeri	M.273462	Queensland	Australia	JN379024 JN379025
Amoria	hunteri	M.273463	Queensland	Australia	JN182226
Adelomelon	beckii		Mar del Plata	Argentina	JN379023 JN379029
Adelomelon	brasiliana	MACN-In39336	Mar del Plata	Argentina	JN379021 JN379026
Adelomelon	riosi	MZSP32971	Cabo Frio	Brasil	JN379022 JN379028

tr

by up to 1 kb in each direction. The binding sites of primers developed here are shown in Figure 1.

Sequence analysis and phylogenetic reconstruction

Sequences were edited using SEQUENCHER v4.6 (Gene Codes Corporation, Ann Arbor, MI, USA). Alignments were generated in SEQUENCHER and exported in nexus format. SE-AL 2.0a11 (Rambaut, 2002) was used to infer protein sequences from the nucleotide sequences and to reÞne alignments, as appropriate. Ribosomal DNA genes were aligned based on secondary structure. The ribosomal RNA gene 16S was aligned using a published molluscan consensus structure (Lydeard et al., 2000), although we found domain 1 to be too variable to unambiguously align based on this consensus structure, and was aligned based on common secondary structures for volute species returned by Mfold (Zuker, Mathews

& Turner, 1999). As no consensus structure of 12S is available for molluscs, this alignment was based on similarity to structures on the Comparative RNA Web Site (http://www.rna.ccbb.utexas.edu; Cannone et al., 2002) using the secondary structures of Paracentrotus lividus (Lamarck, 1816) and Artemia franciscana Kellogg, 1906. Because of alignment ambiguity with the 5′ and 3′ ends of the chosen model sequences, Mfold was used to infer structures of the volute sequences to use as an alignment guide for these regions. Transfer RNA genes were compared with structures reported for L. cerithiformis (Bandyopadhyay et al., 2006) in order to identify putative stem and loop regions for accurate alignment.

For the purpose of phylogenetic analysis several partitioned subsets of the sequence data were created. In addition to the complete data set a concatenated data set was generated with all intergenic spacers removed, and where an overlap exists the relevant nucleotide positions were included for both genes separately. Each protein coding gene and the two ribosomal RNA (rRNA) genes were each given individual partitions, and the transfer RNA (tRNA) genes were partitioned as a single concatenated set.

Maximum parsimony reconstruction, ModelTest (Posada & Crandall, 1998) and partition homogeneity tests were implemented using PAUP* 4.0 (Swofford, 1998). Consistency indices and partitioned homogeneity tests were also generated in PAUP*. Neighbour-joining trees were constructed using the GENEIOUS tree builder in the GENEIOUS software package (Drummond et al., 2007). Maximum likelihood reconstruction and Bayesian analysis were carried out using PHYML 2.4.4 (Guindon & Gascuel, 2003) and MR BAYES 3.1.2 (Huelsenbeck & Ronquist, 2001), respectively, as implemented in GENEIOUS. Maximum parsimony reconstruction was carried out with default parameters, with the exception that 1000 bootstrap replicates were performed. Alignments that contained gaps were analysed with gaps as missing and with gaps as a Þfth state, in order to assess the effect of gaps on phylogenetic reconstruction within this data set. Maximum likelihood reconstruction was carried out under three sets of modelling parameters for each data set. A simple model (HKY with default parameters in PHYML), an intermediate model (HKY with the averaged parameters for all models, as returned by ModelTest), and the speciÞc model and parameters (or as close as possible using PHYML settings) returned by ModelTest using AkaikeÕs information criterion (AIC) (generally the most complex model).

Bayesian reconstruction was carried out separately for each data partition using both GTR and HKY models with default parameters and with four heated chains of length 1 000 000, sampling every 1000 generations, with a 10% burn-in. Visualization of conßict in the data through analysis of splits and networks was carried out in Splits Tree (Huson & Bryant, 2006). Splits were derived from nucleotide alignments in Splits Tree 4. These splits were transferred into SpectroNet (Huber et al.,.8(9ion)-9491.1(Th-supportn1(.8(9ig00)h5

summary of the details of these sequences is given in Table S2. This DNA fragment represents approximately half the neogastropod mitochondrial genome, and the gene arrangement and order is identical to the 13 neogastropod mollusc mitochondrial genomes published on GenBank to date (e.g. Bandyopadhyay et al., 2006; Simison, Lindberg & Boore, 2006; Cunha, Grande & Zardoya, 2009; McComish et al., 2010). Variability in the length of rRNA and tRNA genes, and intergenic spacer regions, primarily between the two out-group taxa (Cymbiola pulchra and Amoria hunteri) and the Alcithoe in-group, required the inclusion of gaps in the alignment of these sequences. Additionally, nad2 from Cymbiola contained a single amino acid insertion, with respect to the other taxa. However, there was insufficient variability to warrant the exclusion of any coding sequence on the grounds of ambiguity.

Summary statistics

Alignment length, summaries of variability, consistency index, and ModelTest results for the complete data set, each of the nine gene partitions, and the concatenated data set are presented in Table 2. These statistics provide useful general information about the data, and identify genes that might be problematical for phylogenetic reconstruction. Although relatively short (159 bp), atp8 exhibits high variability, but third codon position variability is lower than other genes, whereas second position variability is twice that of any other gene. Overall, low variability is seen in cox1 and

Bayesian analyses) and each of the individual data partitions produced a range of topological solutions. From a total of 88 combinations of tree-building methods, models, and data sets, 25 alternative tree topologies were returned. The complete data set returned the same tree topology under each reconstruction method, each with high support. However, none of the data subsets were as consistent under the alternative tree-building strategies. Each of the individual gene partitions produced several tree topologies. Even the concatenated data set, which only omits 167 bp of intergenic spacer, and where nucleotides in overlapping regions appear twice, produces two different tree topologies. For the individual gene partitions the least conßict in tree topology is seen in cox1 and atp6, which each return only two alternative tree topologies, whereas 16S returns a different tree topology for each of the tree estimation methods and parameter sets used. SH tests of the trees returned for each data subset showed that the different topologies did not give a signiÞcantly better Þt to the DNA sequences. The main areas of conßict in the tree topology are around the placement of Alcithoe wilsonae (Powell, 1933) and the resolution of four closely related taxa: Alcithoe lutea (Watson, 1882), Alcithoe larochei Marwick, 1926, Alcithoe fusus (Quoy and Gaimard, 1833), and Alcithoe Þssurata (Dell, 1963). In addition, two less prevalent inconsistencies were observed: the cox2 gene recovered a sister relationship of Alcithoe jaculoides Powell, 1924 and Alcithoe Arabica (Gmelin, 1791), but with low support. The tRNA set 12S, atp8, and cox2 consistently place Alcithoe tigrina Bail & Limpus, 2005 and Alcithoe pseudolutea Bail & Limpus, 2005 in a clade that is the most recently derived in the phylogeny.

A Neighbor-Net network, derived from the complete data set, illustrates the areas in the Alcithoe phylogeny that are problematic (Fig. 3). This network shows the two regions that are the primary cause of alternative trees. The Þrst is that little resolution is seen regarding the positions of the four most recently diverged Alcithoe species (A. lutea, A. larochei, A. fusus, and A. Þssurata). These taxa differ by between 0.7 and 6.9% in pairwise comparisons, and are responsible for the majority of the alternative topologies observed. The second is that there is a mixed signal in reference to the divergence of A. wilsonae. Two topologies have a similar quantity of signal: one in which A. wilsonae is derived from a lineage leading to the Alcithoe benthicola (Dell, 1963)/Alcithoe ßemingi Dell, 1978 clade; the other where A. wilsonae is independent of this clade. This inconsistency leads to the recovery of four (of a possible 105) topologies that differ in the placement of A. wilsonae

inconsistencies (monophyly of A. arabica and A. jaculoides, and the inconsistent placement of the A. tigrina /A. pseudolutea clade) are revealed to be relatively minor signals that are easily discarded when the whole data set is considered.

Spectral analysis

In order to gain a better understanding of the contribution of signal and noise from each of the gene partitions, we explored the phylogenetic information contained in the nucleotide data by visualizing support of taxa splits using networks and Lento plots for individual genes. These splits represent a summary of the total signal in an alignment, and are not generated assuming any given tree topology. As such they represent a description of the phylogenetic information in the data set that is independent of any reconstruction method or model of DNA evolution. A graph of the summed split support of the gene partitions illustrates the contribution of each of the genes to the total split support and conßict for the total concatenated data (Fig. 4). It is useful to compare the splits depicted in the graph with a reference tree, in this case the tree recovered from the complete data set (see the split key in Fig. 4).

The majority of the splits compatible with the complete data set tree exhibit signiÞcant support, in most cases with contributions from all genes. Many splits representing clades not present in the complete data set tree (incompatible splits) have very little support, large conßict, and tend to rely on signals from only a few genes. These splits are likely to result from homoplasy, and represent noise in the data. The most important splits that are incompatible with the complete data set reference tree describe the alternative topologies identiÞed in the network (Fig. 3).

A single incompatible split refers to a taxon set grouping

A

-0 1

-0.2

A. fusus from A. larochei), and that data from this gene leads to increased splits conßict.

Refinement of analysis

Taking into account the accumulated information about variability, compatibility, signal and noise now generated for this data set, an informed decision can be made as to the most appropriate genes to include to maximize the ratio of signal to noise for robust

divergences rather than generate signiÞcant conßict. Therefore, they were retained in the data set as they contain consequential signal for deeper relationships.

DISCUSSION

Exploration of sequence data

Eleven gene partitions returned a range of DNA sub-

may be informative at the species level but not at the genus level, and the decision as to which level they are informative will be based on the signal-to-noise ratio. Additionally, such genes could be used to resolve some subtrees but not others, based

genes that better obey the conditions of an existing model, rather then attempting to Þt a less well suited model or over parameterize by using multiple models.

ACKNOWLEDGEMENTS

This work was funded as part of the Marsden Fund contract 04 GNS 021, administered by the Royal Society of New Zealand. Additional support was given by the Allan Wilson Centre for Molecular Ecology and Evolution. We appreciatively thank our colleagues James Crampton, Alan Beu, and Bruce Marshall, whose knowledge of the palaeontological history and morphological taxonomy of the Volutidae was invaluable. We gratefully acknowledge those who have provided us with specimens for this study, particularly Bruce Marshall at the Museum of New Zealand, Te Papa Tongarewa. Also Sarah Samadi at the MusŽum national dÕHistoire naturelle, Paris, for samples of A. aillaudorum , and Margaret Richards for collecting our A. arabica specimen. Thanks also to Barbara Holland and Klaus Schliep for helpful discussion regarding the use of splits-based analysis, and to David Penny and two anonymous reviewers for comments on the article.