CRUSTACEAN ISSUES 18

Decapod Crustacean Phylogenetics

edited by

Joel W. Martin, Keith A. Crandall, and Darryl L. Felder

CRC Press
Taylor & Francis Group

Decapod Crustacean Phylogenetics

Edited by

Joel W. Martin
Natural History Museum of L. A. County
Los Angeles, California, U. S. A.

Keith A. Crandall
Brigham Young University
Provo, Utah, U. S. A.

Darryl L. Felder
University of Louisiana
Lafayette, Louisiana, U. S. A.

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-9258-5 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Decapod crustacean phylogenetics / editors, Joel W. Martin, Keith A. Crandall, Darryl L. Felder.
p. cm. -- (Crustacean issues)
Includes bibliographical references and index.
ISBN 978-1-4200-9258-5 (hardcover : alk. paper)
1. Decapoda (Crustacea) 2. Phylogeny. I. Martin, Joel W. II. Crandall, Keith A. III. Felder, Darryl L.. IV. Title. V. Series.

QL444.M33D44 2009
595.3'8138--dc22 2009001091

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface ix
JOEL W. MARTIN, KEITH A. CRANDALL & DARRYL L. FELDER

I *Overviews of Decapod Phylogeny*

On the Origin of Decapoda 3
FREDERICK R. SCHRAM

Decapod Phylogenetics and Molecular Evolution 15
ALICIA TOON, MAEGAN FINLEY, JEFFREY STAPLES & KEITH A. CRANDALL

Development, Genes, and Decapod Evolution 31
GERHARD SCHOLTZ, ARKHAT ABZHANOV, FREDERIKE ALWES, CATERINA BIFFIS & JULIA PINT

Mitochondrial DNA and Decapod Phylogenies: The Importance of Pseudogenes and Primer Optimization 47
CHRISTOPH D. SCHUBART

Phylogenetic Inference Using Molecular Data 67
FERRAN PALERO & KEITH A. CRANDALL

Decapod Phylogeny: What Can Protein-Coding Genes Tell Us? 89
K.H. CHU, L.M. TSANG, K.Y. MA, T.Y. CHAN & P.K.L. NG

Spermatozoal Morphology and Its Bearing on Decapod Phylogeny 101
CHRISTOPHER TUDGE

The Evolution of Mating Systems in Decapod Crustaceans 121
AKIRA ASAKURA

A Shrimp's Eye View of Evolution: How Useful Are Visual Characters in Decapod Phylogenetics? 183
MEGAN L. PORTER & THOMAS W. CRONIN

Crustacean Parasites as Phylogenetic Indicators in Decapod Evolution 197
CHRISTOPHER B. BOYKO & JASON D. WILLIAMS

The Bearing of Larval Morphology on Brachyuran Phylogeny 221
PAUL F. CLARK

II *Advances in Our Knowledge of Shrimp-Like Decapods*

Evolution and Radiation of Shrimp-Like Decapods: An Overview 245
CHARLES H.J.M. FRANSEN & SAMMY DE GRAVE

A Preliminary Phylogenetic Analysis of the Dendrobranchiata Based on Morphological Characters 261
CAROLINA TAVARES, CRISTIANA SEREJO & JOEL W. MARTIN

Phylogeny of the Infraorder Caridea Based on Mitochondrial and Nuclear Genes (Crustacea: Decapoda) 281
HEATHER D. BRACKEN, SAMMY DE GRAVE & DARRYL L. FELDER

III *Advances in Our Knowledge of the Thalassinidean and Lobster-Like Groups*

Molecular Phylogeny of the Thalassinidea Based on Nuclear and Mitochondrial Genes 309
RAFAEL ROBLES, CHRISTOPHER C. TUDGE, PETER C. DWORSCHAK, GARY C.B. POORE & DARRYL L. FELDER

Molecular Phylogeny of the Family Callianassidae Based on Preliminary Analyses of Two Mitochondrial Genes 327
DARRYL L. FELDER & RAFAEL ROBLES

The Timing of the Diversification of the Freshwater Crayfishes 343
JESSE BREINHOLT, MARCOS PÉREZ-LOSADA & KEITH A. CRANDALL

Phylogeny of Marine Clawed Lobster Families Nephropidae Dana, 1852, and Thaumastochelidae Bate, 1888, Based on Mitochondrial Genes 357
DALE TSHUDY, RAFAEL ROBLES, TIN-YAM CHAN, KA CHAI HO, KA HOU CHU, SHANE T. AHYONG & DARRYL L. FELDER

The Polychelidan Lobsters: Phylogeny and Systematics (Polychelida: Polychelidae) 369
SHANE T. AHYONG

IV *Advances in Our Knowledge of the Anomura*

Anomuran Phylogeny: New Insights from Molecular Data 399
SHANE T. AHYONG, KAREEN E. SCHNABEL & ELIZABETH W. MAAS

V *Advances in Our Knowledge of the Brachyura*

Is the Brachyura Podotremata a Monophyletic Group? 417
GERHARD SCHOLTZ & COLIN L. MCLAY

Assessing the Contribution of Molecular and Larval Morphological Characters in a Combined Phylogenetic Analysis of the Superfamily Majoidea 437
KRISTIN M. HULTGREN, GUILLERMO GUERAO, FERNANDO P.L. MARQUES & FERRAN P. PALERO

Molecular Genetic Re-Examination of Subfamilies and Polyphyly in the Family Pinnotheridae (Crustacea: Decapoda) 457
EMMA PALACIOS-THEIL, JOSÉ A. CUESTA, ERNESTO CAMPOS & DARRYL L. FELDER

Evolutionary Origin of the Gall Crabs (Family Cryptochiridae) Based on 16S rDNA Sequence Data 475
REGINA WETZER, JOEL W. MARTIN & SARAH L. BOYCE

Systematics, Evolution, and Biogeography of Freshwater Crabs 491
NEIL CUMBERLIDGE & PETER K.L. NG

Phylogeny and Biogeography of Asian Freshwater Crabs of the Family Gecarcinucidae (Brachyura: Potamoidea) 509
SEBASTIAN KLAUS, DIRK BRANDIS, PETER K.L. NG, DARREN C.J. YEO & CHRISTOPH D. SCHUBART

A Proposal for a New Classification of Portunoidea and Cancroidea (Brachyura: Heterotremata) Based on Two Independent Molecular Phylogenies 533
CHRISTOPH D. SCHUBART & SILKE REUSCHEL

Molecular Phylogeny of Western Atlantic Representatives of the Genus *Hexapanopeus* (Decapoda: Brachyura: Panopeidae) 551
BRENT P. THOMA, CHRISTOPH D. SCHUBART & DARRYL L. FELDER

Molecular Phylogeny of the Genus *Cronius* Stimpson, 1860, with Reassignment of *C. tumidulus* and Several American Species of *Portunus* to the Genus *Achelous* De Haan, 1833 (Brachyura: Portunidae) 567
FERNANDO L. MANTELATTO, RAFAEL ROBLES, CHRISTOPH D. SCHUBART & DARRYL L. FELDER

Index 581

Color Insert

Phylogenetic Inference Using Molecular Data

FERRAN PALERO[1] & KEITH A. CRANDALL[2]

[1] *Departament de Genètica, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona, Spain*

[2] *Department of Biology, Brigham Young University, Provo, Utah 84602, U.S.A.*

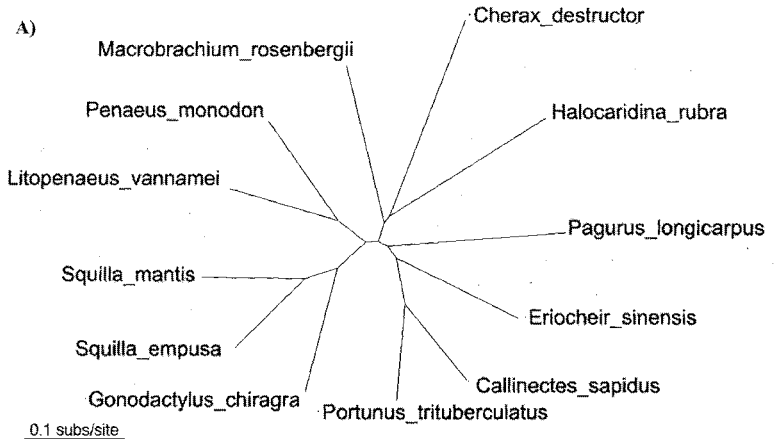
ABSTRACT

We review phylogenetic inference methods with a special emphasis on inference from molecular data. We begin with a general comment on phylogenetic inference using DNA sequences, followed by a clear statement of the relevance of a good alignment of sequences. Then we provide a general description of models of sequence evolution, including evolutionary models that account for rate heterogeneity along the DNA sequences or complex secondary structure (i.e., ribosomal genes). We then present an overall description of the most relevant inference methods, focusing on key concepts of general interest. We point out the most relevant traits of methods such as maximum parsimony (MP), distance methods, maximum likelihood (ML), and Bayesian inference (BI). Finally, we discuss different measures of support for the estimated phylogeny and discuss how this relates to confidence in particular nodes of a phylogeny reconstruction.

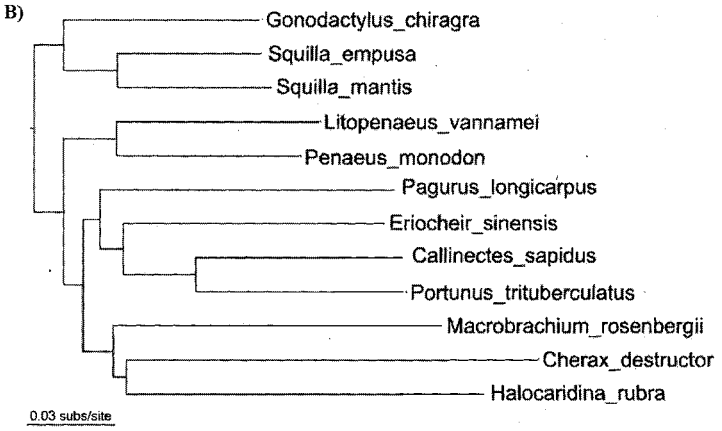
1 INTRODUCTION

The main objective of molecular phylogenetic analysis is to infer the evolutionary history of a group of species and represent it as an hierarchical branching diagram, a cladogram, or phylogenetic tree (Edwards & Cavalli-Sforza 1964). The contemporary taxa in that tree (as opposed to the reconstructed ancestral taxa) are called leaves or terminal tips. Internal nodes represent ancestral divergences into two or more (polytomy) genetically isolated groups (Fig. 1). Clades are characterized by shared possession of uniquely derived evolutionary novelties (synapomorphies). Therefore, phylogenetic analysis can be partially regarded as an attempt to recognize the identity and taxonomic distribution of synapomorphies. These could be any kind of inherited phenotypic or genotypic characteristics; it could be the evolutionary appearance of a nauplius larva or the fixation of a change from guanine to adenine at a particular site in a DNA sequence. Thus, phylogenies become essential tools for comparative biology (Harvey & Pagel 1991).

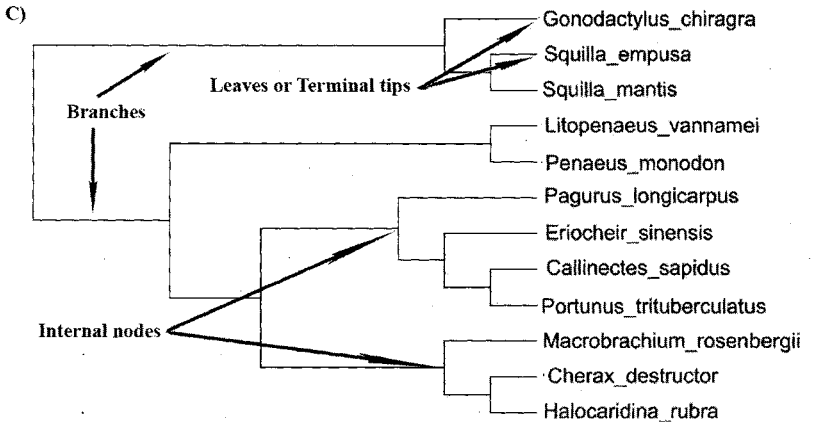
The tree topology is the information on the order of relationships, while the lengths of the branches in the tree can represent the evolutionary distances that separate nodes (phylogram) or not (cladogram). It is important to recognize if branches have been drawn to scale in order to know the relative distance between different species. This is particularly important, since if the sequences do not all evolve at the same rate, it is not possible to have a well-defined time axis on the tree with the standard methods. At this point we should also differentiate between rooted and unrooted trees. Even though biologists tend to think about trees as being rooted and pointing from "lower complexity" to "higher complexity," most phylogenetic methods do not result in a rooted tree (see Modeling Evolution section below). We generally need to define an outgroup by using external evidence not included in the molecular dataset (Weston 1994). Only then can rooted trees inform us about the temporal order of events and about which species have high rates of molecular evolution.



UNROOTED PHYLOGRAM



ROOTED PHYLOGRAM



CLADOGRAM

Figure 1. Phylogenetic trees obtained using a 966bp segment of the cytochrome B gene of several malacostracan crustaceans. (A) Unrooted phylogram, with distance scale bar indicating substitutions per site. (B) Rooted phylogram; the tree was rooted using Stomatopoda species as the outgroup. (C) Cladogram, showing the tree topology only.

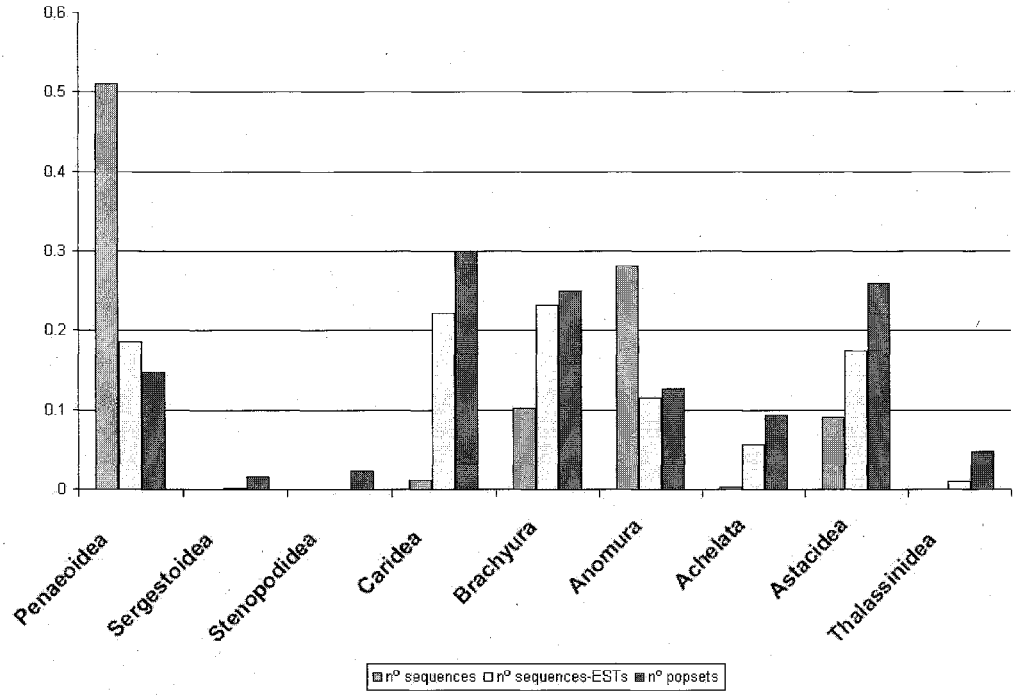


Figure 2. (See Color Figure 1 in the Color Insert at the end of the book.) Decapod sequences in GenBank in April 2008, shown as a proportion of the sequences belonging to the different infraorders relative to the total number of sequences available (355,876), the total number of sequences available after excluding ESTs (337,603), and the relative proportion of population study datasets.

1.1 *Why should we use molecules when we already have morphology-based taxonomies?*

Thanks to the popularization of DNA sequencing techniques, the number of decapod crustacean sequences available in GenBank has increased considerably, even though some infra-orders are still underrepresented (Fig. 2). The amplification of long genomic fragments implies that thousands of new, variable characters are made available for the study of phylogenetic relationships among organisms. This is particularly important for groups with very few characters available for developing morphological matrices (e.g., Rhizocephala) or when homology of morphological characters is particularly difficult to establish (Glenner et al. 2003). Moreover, the widespread use of accurate models of evolution and statistical tests allows us to extract a considerable amount of information from molecular sequence data. With the incorporation of closely related species to our group of interest, DNA sequence data allow polarity to be conferred to our phylogenetic reconstruction and allow us to make inferences on the evolution of molecules and/or the morphological characters themselves. An important advantage of molecular data is its objectivity, since results can be independently reproduced from the sequence data that are deposited in public databases.

However, DNA sequences have the same concerns as morphological traits for phylogeny estimation. Homoplasy can be caused by multiple substitutions occurring on a particular site, and character loss can also happen in gene sequences by insertion–deletion events. Phylogeny reconstruction can aid in the homology determination of molecular characters. Homologous genes may be orthologs, if they separated due to a speciation event, or paralogs, if those gene sequences diverged after gene duplication. In fact, gene duplication has been claimed to play a major role in the evolution of the mitochondrial genome of the Japanese freshwater crab *Geothelphusa dehaani* (Segawa & Aotsuka

2005). Furthermore, DNA sequences obtained from PCR products may correspond to pseudogenes, or non-functional copies. Using a mixture of orthologs and paralogs for phylogenetic reconstruction may point to the wrong topology (making distant taxa cluster together), whereas mixing pseudogenes with functional copies (e.g., nuclear copies of mitochondrial genes or numts) also gives the wrong topology but can make even copies from the same individual seem very distant (Song et al. 2008; Schubart this volume). When dealing with molecular sequences, character homology is incorporated with the sequence alignment, so we must be certain about the homology among nucleotide positions in the alignment.

2 CHARACTER HOMOLOGY AND THE PROBLEM OF SEQUENCE ALIGNMENT

Phylogenetic analysis attempts to reconstruct evolutionary genealogies of species based on similarities and differences. In an alignment of DNA sequences, each aligned site is a separate character with four character states being four nucleotides (A, C, T, G). Carrying out a multiple alignment means to define positional homology, deciding which nucleotide or amino acid positions are homologous for our sequence data. In order to infer the correct topology, nucleotide or amino acid positions must be aligned correctly. However, alignments of distantly related sequences may not be feasible, and different alignment methods often produce variable results depending on the details of the algorithm (Benavides et al. 2007). The most commonly used algorithms employ dynamic programming procedures seeking to maximize the score of the alignment (Needleman & Wunsch 1970). The score is determined by the choice of a matrix of similarities between nucleotides or amino acids and by the assignment of penalties for opening and extending gaps or insertions (Thompson et al. 1994).

Most dynamic programming methods use a greedy approach for progressively aligning pairs of sequences, but hierarchically aligning pairs of sequences is prone to generate biases and dominance by the most similar sequences. Additionally, the alignment tends to be sensitive to the choice of the similarity matrix and of gap penalties. Alternative approaches for aligning sequences include both dynamic programming and motif-finding algorithms. For example, the alignment program MUSCLE (Edgar 2004) first searches regions of similarity refined through iterations and then optimizes the alignment by applying a dynamic programming procedure locally. Since alignment methods are prone to errors, it is customary to manually adjust the alignment or to eliminate positions that are considered to be uncertain (GBLOCKS: Castresana 2000), a procedure that relies somewhat on the judgment of the investigator. Poorly aligned positions may not be homologous or may have been saturated by multiple substitutions and should be eliminated to increase the reliability of the phylogenetic analysis (Swofford et al. 1996; Castresana 2007). However, misalignments can still go undetected, particularly in large-scale analyses and for distantly related sequences.

2.1 *Dealing with gaps*

DNA sequences of homologous genes from distant species usually have unequal lengths and therefore force us to assume particular insertion and deletion events, defining the location of gaps or indels in the alignment. When dealing with protein coding nucleotide sequences, we could translate to the amino acid sequence, which may be easier to align, and then reverse back to the nucleotide sequence. However, the most commonly used genes for phylogenetic inference are non-protein coding genes (i.e., rDNA), and dealing with gaps remains a problem. Most distance-based analyses and, until recently, most likelihood and Bayesian analyses either treated gaps as unknowns or removed the gap containing column(s) from the analyses for pairs of sequences or for all sequences in an alignment (Lutzoni et al. 2000). The specific treatment of gaps in phylogenetic analysis can affect the results (Ogden & Whiting 2003), and several approaches are available for incorporating indel

information into the phylogenetic analysis (Holmes 2005). Indeed, empirical results suggest that incorporating gaps as phylogenetic characters can aid in providing more robust phylogenetic estimates (Egan & Crandall 2008). It has been shown that point estimation of alignment and phylogeny avoids bias that results from conditioning on a single alignment estimate (Lake 1991; Thorne & Kishino 1992).

Within parsimony analysis, gaps may be incorporated as transformations during the cladogram evaluation process (optimization alignment in POY; Varón et al. 2007). It has been shown that in cases where alignment is not totally correct, coding gaps as a fifth state character or as separate presence/absence characters outperforms treating gaps as unknown/missing data nearly 90% of the time (Ogden & Rosenberg 2006). Datasets with higher sequence divergence and polytomies are more affected by gap coding than datasets associated with shallower non-polytomic tree shapes (Ogden & Rosenberg 2007). Redelings & Suchard (2005) describe a statistical method for incorporating indel information into phylogeny estimation under a Bayesian framework. Their method uses a joint reconstruction that simultaneously infers the alignment, tree, and insertion/deletion rates. Estimation proceeds through Markov chain Monte Carlo (MCMC) and naturally accounts for uncertainty in alignments, phylogenies, and other parameters through posterior probabilities. This method is based on a probabilistic model of sequence evolution that contains insertion and deletion events as well as substitution events (Thorne et al. 1991). Gaps are not treated as a fifth character state, since this over-weights the evidence of shared indels by treating an indel of multiple residues as multiple shared indels. Instead, the indel process is separate and independent of the substitution process and allows indels of several residues simultaneously.

3 GENETIC DISTANCES AND SATURATION

Theoretically, if the total number of substitutions between any pair of sequences is known, all the distance methods will produce the correct phylogenetic tree. In practice, this number is almost always unknown. In order to estimate a standardized genetic distance between organisms, we could just count the number of nucleotide differences among sequences and divide that number for the total number of nucleotide positions compared (p distance). However, DNA changes usually do not occur randomly along the sequence because of negative selection acting preferentially over some positions (Frank & Lobry 1999). Besides, if two lineages have been evolving separately for a long time, it is likely that multiple nucleotide substitutions have occurred on a particular position (multiple hits). As mutations accumulate, a point is reached at which there is no further divergence between sequences (mutational saturation). From this point on, it becomes impossible to estimate the evolutionary distance from similarity. This point of mutational saturation may occur at any taxonomic level, depending on the pattern of position-specific variability. Variation of mutation rate patterns among sites, functionally constrained sites, rapidly evolving lineages, and ancient evolutionary events will make the estimates of distances uncertain (Philippe & Forterre 1999). Different molecules evolve at different rates, and some of the fast-evolving genes will be saturated with changes even for closely related taxa. Using fast-evolving genes for phylogenetic inference of distantly related species could provide misleading results. A sensible approach for tackling this problem of saturation would be to use molecular markers that present a slower mutation rate and using an appropriate nucleotide substitution model in order to correct the observed distance for the multiple hits. However, if the gene evolves too slowly, there will be very little variation among the sequences, and there will be too little information to construct a phylogeny. Phylogenetic methods are likely to become unreliable if the sequences are too different from one another, and this should be borne in mind when the choice of gene sequences is made initially. Typically, a combination of genes is needed to accurately reconstruct phylogenetic relationships, with faster-evolving genes resolving close relationships and more slowly evolving genes resolving deeper relationships.

4 MODELING EVOLUTION AND MODEL SELECTION

More complex models, taking into account a variety of biological phenomena, generally provide more accurate estimates of phylogeny regardless of the method (e.g., parsimony, likelihood, distance, Bayesian) (Huelsenbeck 1995). The most common models of DNA evolution include base frequency, base exchangeability, and rate heterogeneity parameters. The parameter values are usually estimated from the dataset in each particular analysis (model selection). Finally, the evolutionary models are defined by matrices containing the relative rates of all possible replacements (transition probability matrix), which allow us to calculate the probabilities of change from any nucleotide to any other nucleotide (Liò & Goldman 1998). Most models assume reversibility of the transition probability matrix so that no inferences about evolutionary direction can be made unless further information extrinsic to the sequences themselves (e.g., fossil record) is supplied.

The base frequency parameters describe the frequencies of the nucleotide bases averaged over all sequence sites and over the tree. These parameters can be considered to represent constraints on base frequencies due to effects such as overall GC content, and they act as weighting factors in a model by making certain bases more likely to arise when substitutions occur. Base exchangeability parameters describe the relative tendencies of bases to be substituted for one another (Fig. 3). These parameters represent a measure of the biochemical similarity of bases, since transitions (i.e., C↔T or A↔G) usually occur more often than transversions (e.g., C↔G) (Brown et al. 1982; but see also Keller et al. 2007). Furthermore, mutation rates vary considerably among sites of DNA and amino acid sequences or among loci, because of constraints of the genetic code, selection for gene function, etc. In fact, we have to consider that if most of the nucleotide positions in our sequences evolve rather slowly or do not change at all (invariant sites), then base changes will tend to accumulate in a few variable sites, and sequence saturation will be reached much more quickly and at a lower divergence than expected under simpler models that do not

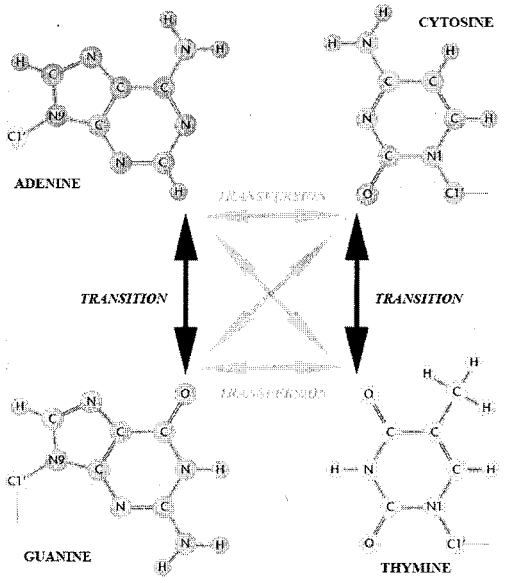


Figure 3. Transition versus Transversion mutations. DNA substitution mutations are of two types. Transitions are interchanges of purines (A–G) or pyrimdines (C–T), which involve bases of similar shape. Transversions are interchanges between purine and pyrmidine bases, which involve exchange of one-ring and two-ring structures.

incorporate rate heterogeneity or a proportion of invariant sites. The most widespread approach to modeling rate heterogeneity among sequence sites is to describe each site's rate as a random draw from a gamma distribution (Yang et al. 1994). The shape of the gamma distribution is controlled by a parameter α. Large values of α suggest that sites evolve at a similar rate, while small values of the parameter α imply higher levels of rate heterogeneity among sites and the presence of many sites with lower rates of evolution. It is also possible to assign specific rates of substitution to different parts of the sequence in order to account for the heterogeneity on the mutation rate (e.g., to the three codon positions of protein coding sequences or to different domains in rRNA).

We can use the likelihood framework to estimate parameter values and their standard errors from the observed data when selecting the optimal model to perform phylogenetic inference (Yang et al. 1994), since comparisons of two competing models are possible using likelihood ratio tests. Competing models are compared (using their maximized likelihoods) with a statistic that measures how much better an explanation of the data the alternative model gives. When the simpler model is a special case of the more complex model, then the required distribution for the statistic is usually a χ^2 distribution with the number of degrees of freedom equal to the difference in the number of parameters between the two models (Goldman 1993). When the models being compared are not nested, as can often be the case for more complex models of sequence evolution, the required distribution can be estimated by Monte Carlo simulation or by parametric bootstrapping (Huelsenbeck & Rannala 1997). Alternatively, one can use different statistical criteria to evaluate alternative models simultaneously (Posada & Buckley 2004).

Complex models describing selection or structure consistently give significantly improved descriptions of the evolution of protein sequences and are especially valuable in giving new insights into the processes of molecular evolution (Porter et al. 2007). Particularly, codon-based models have been developed that describe the evolution of coding sequences in terms of both DNA substitutions and the selective forces acting on the protein product (Nielsen & Yang 1998; Yang et al. 2000). For example, by studying the relationships between rates of synonymous (amino acid conserving) and nonsynonymous (amino acid altering) DNA substitutions, these models have been used successfully to detect where and when positive selection was important (Zanotto et al. 1999). Other models have attempted to associate the heterogeneity of patterns and rates of evolution among sites with the structural organization of RNA. These complex models accommodating RNA secondary structural elements use 16 states to represent all the possible base pairings in stem regions and four states to model loops (Schöniger & von Haeseler 1994).

Finally, while employing multiple alternative models in phylogenetic analysis might be seen as more rigorous, if this approach is to be meaningful there needs to be some quality control on the models employed (Grant & Kluge 2003). Similarly, all methods of phylogenetic inference assume a model of evolution, either implicitly or explicitly. For example, a strict parsimony analysis assumes all character changes are of equal weight. Thus, it becomes incumbent upon the researcher to justify the choice of model, even if it is an implicit model used to describe character evolution. If there are no restrictions on allowable models, virtually any given phylogeny may be found to be supported by some models and refuted by others. The model averaging approach by Lee & Hugall (2006) addresses both issues: a large number of possible models can be employed, but the results of each model are weighted according to its fit, so that the results of implausible models carry little weight on the final estimate. Likewise, statistically testing alternative models of evolution allows one to determine if the addition of more parameters makes a significant improvement in a likelihood score (Posada & Crandall 2001).

5 SEARCHING FOR TREES IN A BROAD TREE SPACE

The reconstruction of a phylogenetic tree using molecular data is an attempt to statistically infer the best estimate of evolutionary relationships given some criterion. While the "true tree" is the goal, what phylogenetic methods actually do is optimize a tree given some model and optimality criterion.

Thus, we are actually searching for not the "true tree" but rather the "optimal tree" and hope that the latter has some relationship to the former. There are two processes involved in this inference: estimation of the topology and estimation of branch lengths for a given tree topology. When a topology is known, statistical estimation of branch lengths is relatively simple, and one can use several statistical methods such as the least squares and the maximum likelihood methods. The problem is the estimation or reconstruction of a topology. The number of possible topologies increases rapidly with the number of sequences (Swofford et al. 1996), and it is generally very difficult to choose the correct topology among them. In phylogenetic inference, a certain optimization principle such as the maximum likelihood (ML) or minimum evolution (ME) principle is often used for evaluating different tree scores and choosing the topology and branch lengths that give an optimal score, so that we need to have tree searching strategies to help us finding the "optimal tree."

Exhaustive search. The exhaustive algorithm evaluates all possible trees. Because it examines all possible topologies, exhaustive searches guarantee the most optimal tree(s), but it is very slow (using 12 taxa, more than 600 million trees are evaluated). The advantage of the exhaustive search is the ability to completely explore the tree space and thereby plot the optimality score distribution. This histogram may indicate the "quality" of your matrix, in the sense that there should be a tail to the left such that few short trees are "isolated" from the greater mass of less optimal trees (but see Kitchin et al. 1998).

Branch and bound. The branch-and-bound algorithm is guaranteed to find all optimal trees, given some criterion (e.g., maximum parsimony). It discards whole classes of trees that it has determined are suboptimal, without the need to examine all of those one by one. The savings is greater the less homoplasy there is in the data. However, in cases where there are many conflicts between information from different characters and much parallelism and convergence, the branch-and-bound strategy does not perform particularly well. Moreover, branch-and-bound methods still have a complexity that is exponential, and it is not recommended to use the branch-and-bound algorithm for datasets with more than 12 taxa.

Heuristic searches. Since most datasets today contain large numbers of sequences, exhaustive and branch-and-bound searches quickly become impractical. We then turn to heuristic searches. Heuristic searches attempt to survey the tree space reasonably well without guaranteeing to find the most optimal tree(s). The key to good heuristic searching is the ability to move around the tree space and spend time exploring reasonable alternative topologies. Thus, a wide variety of branch swapping algorithms has been developed to achieve this goal.

Nearest-neighbor interchange (NNI). This heuristic algorithm adds taxa sequentially, in the order they are given in the matrix, to the branch where they will give least increase in tree length (Robinson 1971; Moore et al. 1973). After each taxon is added, all nearest neighbor trees are swapped to try to find an even shorter tree. Like all heuristic searches, this one is much faster than the algorithms above and can be used for large numbers of taxa, but it is not guaranteed to find all or any of the optimal trees. To decrease the likelihood of ending up on a suboptimal local minimum, a number of reorderings can be specified. For each reordering, the order of input taxa could be randomly permutated and another heuristic search attempted.

Subtree pruning and regrafting (SPR) is similar to NNI, but with a more elaborate branch swapping scheme. In order to find a shorter tree, a subtree is cut off the tree and regrafted onto all other branches in the tree to find the best alternative (Swofford 2003). This is done after each taxon has been added, and for all possible subtrees. While slower than NNI, SPR will often find shorter trees (Felsenstein 2004).

Tree bisection and reconnection (TBR) is similar to SPR, but with an even more complete branch swapping scheme. The tree is divided into two parts, and these are reconnected through every possible pair of branches in order to find a shorter tree. This is done after each taxon is added, and for all possible divisions of the tree (Swofford 2003). TBR will often find shorter trees than SPR and NNI, but it is more time consuming.

The ratchet. Different characters in the data may well recommend different trees to us. To prevent the search from becoming focused on a limited set of trees, it may help to use different starting trees as recommended by various subsets of characters. In the ratchet approach, we pick up some characters and increase their representation by increasing their weight (Nixon 1999; Felsenstein 2004). This moves the search to a tree recommended by this reweighted dataset; then we search from that starting point using the full set of characters.

Given the enormously large size of the tree space even for a small dataset, all we can do is hope that if we have searched for a long time without finding any improvement, then we have probably found the best tree. The problem with long-range moves tends to be that they are rather disruptive, moving the search far from the optimal tree. Most real search programs use a combination of NNIs and slightly longer range moves that have been tested and found to be reasonably efficient at finding optimal trees as quickly as possible. The MCMC method (see below) is a way of searching tree space that allows both uphill and downhill moves, allowing for suboptimal tree topologies to be sampled during the search. Regardless of the optimality criterion used, a key aspect of effective heuristic tree searching is to perform the analysis multiple times with different starting positions to be sure the tree space has been reasonably sampled.

6 INFERENCE METHODS

Ideally, the inference method used will extract the maximum amount of information available in the sequence data, will combine this with prior knowledge of patterns of sequence evolution (included in the evolutionary model), and will deal with model parameters (e.g., the transition/transversion ratio) whose values are not known a priori. The major inference methods for molecular phylogenetics are maximum likelihood, Bayesian inference, distance methods, and maximum parsimony.

6.1 *Maximum likelihood*

Likelihood-based techniques allow a wide variety of phylogenetic inferences from sequence data and a robust statistical assessment of all results. The likelihood of an hypothesis is equal to the probability of observing the data (sequence alignment) if that hypothesis (tree topology) were correct, given the chosen model of sequence evolution (Felsenstein 1981). Thus, a model of nucleotide or amino acid replacement allows the calculation of the likelihood for any possible combinations of tree topology and branch lengths. It permits the inference of phylogenetic trees and also making inferences simultaneously about the patterns and processes of evolution. A great attraction of the likelihood approach in phylogenetics is the existence of a wealth of powerful statistical theory, for example, the ability to perform robust statistical hypothesis tests (see below) and the knowledge that ML phylogenetic estimates are statistically consistent (given enough data and an adequate model, ML will always give the correct tree topology) (Rogers 1997). These strong statistical foundations suggest that likelihood techniques are the most powerful for phylogeny reconstruction and for understanding sequence evolution. Simulation studies show that ML methods generally outperform distance and parsimony methods over a broad range of realistic conditions, and recent developments in distance and parsimony methodology have concentrated on elucidating the relationships of these methods to ML inference and exploiting this understanding to adapt the methods so that they perform more like ML methods (Steel & Penny 2000; Bruno et al. 2000). However, ML suffers from computational intensity, making ML estimation impractical when dealing with several thousands of sequences, but better algorithms are being developed continually that can accommodate an increasingly large number of sequences for ML analyses (Stamatakis et al. 2005).

The ML method is a well-established statistical method of parameter estimation; it gives the smallest variance of a parameter estimate when sample size is large. In the construction of

phylogenetic trees, maximization of the likelihood is done for each topology separately by using a different likelihood function, and the topology with the highest (maximum) likelihood is chosen as an estimate of the true topology. Since different topologies represent different probability spaces of parameters, it is not clear whether the maximum likelihood tree is expected to be the true tree unless an infinite number of nucleotides are examined (Felsenstein 2004). Finally, it should be mentioned that the statistical foundation of phylogeny estimation by ML has not been well established, and some authors have pointed out that topologies are parameters, but these parameters are not included in the likelihood function that is being maximized (Yang 1996a).

6.2 *Bayesian methods*

When inferring phylogenies, we should consider methods that deal directly with ensembles of possible trees, rather than chasing after a single best one, and we should be able to consider the information in the data and any prior information about the probabilities of the events. The fundamental importance of evolutionary models is that they contain parameters, and if specific values can be assigned to these parameters based on observations, such as an alignment of DNA sequences, then biologists can learn something about how molecular evolution has occurred. Although both maximum likelihood and Bayesian analyses are based upon the likelihood function, there are fundamental differences in how the two methods treat parameters. ML makes inferences about the parameters of interest while fixing the values for the other parameters (nuisance parameters). However, Bayesians assign a prior probability distribution to the nuisance parameters and the posterior probability is calculated by integrating over all possible values of those nuisance parameters, weighting each by its prior probability. The advantage of this is that inferences about the parameters of interest do not depend upon any particular value for the nuisance parameters. The disadvantage is that it may be difficult to specify a reasonable prior for the parameters. Nevertheless, when there is a large amount of information in the data and the likelihood function changes rapidly as the parameter values are altered, the choice of prior is not so important and it is possible to use uniform or non-informative priors. All branch lengths could be set as equally likely a priori, and a suitable non-informative choice of prior for base frequencies could be to set all sets of frequencies that add up to one as equally probable.

Markov models are routinely used in several domains of science and do not belong specifically to the Bayesian inference methodology; however, they have revolutionized genetic inferences in many aspects (Beaumont & Rannala 2004). A Markov model is a mathematical model for a process with changes of state over time, in which future events occur by chance and depend only on the current state and not on the history of how that state was reached. In molecular phylogenetics, the states of the process are the possible nucleotides or amino acids present at a given time and position in a sequence, and state changes represent mutations in sequences. Therefore, starting from an evolutionary model and a set of nucleotide frequencies, we can get to an equilibrium at which any state has a probability of occurrence that does not depend on the initial state of the process.

Under the MCMC search in a Bayesian framework, the probability of finding a tree will be proportional to its likelihood multiplied by its prior probability. In that case, the new tree is either accepted or rejected, using a rule known as the Metropolis algorithm. If the likelihood of the proposed tree is larger than the likelihood of the current one, the proposed topology is accepted and it becomes the next tree in the sample. If it is rejected, then the next tree in the sample is a repeat of the original tree. It also allows moves that decrease the likelihood, in order to allow for sampling of suboptimal trees. When the MCMC chain reaches the equilibrium, the probability of observing each tree must be constant. This property is known as detailed balance. It is necessary to strike a balance between moves that alter branch lengths and those that alter topology. If changes are very large, then the likelihood ratio of the states will be far from 1, and the likelihood of accepting the downhill move for sampling suboptimal trees will be very small. Finally, failure to diagnose a lack

of convergence of the MCMC chain will lead to incorrect tree topology estimates (Huelsenbeck et al. 2002).

6.3 *Distance methods*

Distance matrix methods calculate a measure of the distance between each pair of species and then find a tree that predicts the observed set of distances as closely as possible. This leaves out all information from higher-order combinations of character states, reducing the data matrix to a simple table of pairwise distances. Distance methods use the same models of evolution as ML to estimate the evolutionary distance between each pair of sequences from the set under analysis and then try to fit a phylogenetic tree to those distances. The distances will usually be ML estimates for each pair of sequences (considered independently of the other sequences). Disadvantages of distance methods include the inevitable loss of evolutionary information when a sequence alignment is converted to pairwise distances and the inability to deal with models containing parameters for which the values are not known a priori (Steel et al. 1988). We are trying to find the n-species tree that is implied by these distances. The difficulty in doing this is that the individual distances are not exactly the path lengths in the full n-species tree between those two species. Since we are dealing with pairwise distances, we need to be able to find the full tree that does the best job of approximating these individual two-species trees.

In order for distances that are used in these analyses to have the proper expectations, it is essential that they are expected to be proportional to the total branch length between the species. If the distances do not have the linearity property, then wrenching conflicts between fitting the long distances and fitting the short distances arise, and the tree is the worse for them. There are several distance matrix methods available in the literature. Two examples are minimum evolution and neighbor joining.

Minimum Evolution. This method seeks to find the tree with the shortest overall branch lengths. First, the least squares trees are determined for different topologies, and the choice is made among them by choosing the one of shortest total length. Rzhetsky & Nei (1993) showed that if the distances were unbiased estimates of the true distance (many distances are not unbiased), then the expected total length of the true tree was shorter than the expected total length of any other. However, that is not the same as showing that the total length is always shorter for the true tree, as the lengths vary along their expectation. Gascuel et al. (2001) have found cases where the minimum evolution is inconsistent when branch lengths are inferred by weighted least squares or by generalized least squares.

Neighbor Joining. NJ is a clustering method that produces unrooted trees. It works by successively clustering pairs of sequences together. It is related to the UPGMA method of inferring a branching diagram from a distance matrix. Unlike the UPGMA method, NJ can facilitate contemporary tips of uneven length. This makes it a more appropriate tree reconstruction method than UPGMA in those instances when evolution has not proceeded in a strictly clock-like fashion. NJ is guaranteed to recover the true tree if the distance matrix happens to be an exact reflection of a tree. However, in the real world, distances will not be exactly additive, and therefore NJ is just one approximation. Furthermore, the NJ tree may be misleading. If the input distances are not close to being additive, because pairwise distances were not properly calculated or because sequences were not properly aligned, then NJ will give the wrong tree.

NJ is useful to rapidly search for a good tree that can then be improved by other criteria. Ota & Li (2001) use neighbor joining and bootstrapping to find an initial tree and identify which regions are candidates for rearrangement. They then use ML for further refinement. This results in a substantial improvement in speed over pure likelihood methods. Moreover, modifications of NJ have been developed to allow for differential weighting in the algorithm to take into account differences in statistical noise. Gascuel (1997) has modified the NJ to allow for the variances and covariances

of the distances to be proportional to the branch lengths. This is a good approximation provided that the branch lengths are not too long.

6.4 *Maximum parsimony*

The theoretical basis of this method is the philosophical idea that the best hypothesis to explain a process is the one that requires the smallest number of assumptions (Occam's Razor). If there are no backward and no parallel substitutions at each nucleotide site (no homoplasy) and the number of informative nucleotides examined is very large, maximum parsimony (MP) methods are expected to provide the correct (realized) tree. MP assumes that maximizing the congruence among characters will be equal to minimizing incongruence (homoplasy) (Farris 1983). Therefore, computing programs will count the number of mutational changes (steps) we need to explain a particular tree and repeat this counting for thousands of trees. The tree or trees that need a minimum number of changes to explain the relationships between species will be accepted as the most parsimonious tree.

There are two main dynamic programming algorithms for counting the number of changes of state. In both cases, the algorithm does not function by actually placing changes or reconstructing states at the nodes of the tree. The **Fitch algorithm** works for characters with any number of states, provided one can change from any one to any other (Kluge & Farris 1969). Fitch characters are reversible and unordered, meaning that all changes have equal cost. This is the criterion with fewest assumptions, and is therefore generally preferable. The Fitch algorithm can be carried out in a number of operations that are directly proportional to the number of species on the tree, and, therefore, the algorithm is less computationally demanding than other methods. The **Sankoff algorithm** starts by assuming that one has a table of the cost of changes between each character state and each other state. In this case, one computes the total cost of the most parsimonious combinations of events by computing it for each character. Given that a node is assigned a particular character state, we will compute the minimal cost of all the events in the subtree that starts from that node and accept it as the most parsimonious result.

Other algorithms allow us to reconstruct character states at the nodes of the tree. The **Camin-Sokal Parsimony** algorithm (C-S) assumes that we know the ancestral state of the character. In its simplest form, only two states are allowed (presence/absence) and reversals are impossible. One application of C-S parsimony is in the evolution of small deletions of DNA, when we have no reason to believe that they could revert spontaneously. In more complex cases, when deletions overlap and we cannot be entirely sure whether any one of them is present or absent, C-S parsimony would not be appropriate. C-S parsimony infers a rooted tree, since it will favor the placement of the root in one particular part of the tree. In its simplest form, **Dollo parsimony** assumes that there are two states (ancestral/derived). The main difference with C-S parsimony is that in this case the derived state is allowed to evolve only once, but it is allowed to revert to the ancestral state multiple times. The number of these reversions is the quantity being minimized, and it is also an inherently rooted method. In "unweighted" (=equal weighting) MP methods, nucleotide or amino acid substitutions are assumed to occur in all directions with equal or nearly equal probability. In reality, however, certain substitutions (e.g., transitional changes) occur more often than other substitutions (e.g., transversional changes). It is therefore reasonable to give different weights to different types of substitutions when the minimum number of substitutions for a given topology is to be computed. MP methods incorporating a weight matrix for the different types of change are weighted MP methods.

Once the most parsimonious phylogenetic tree has been recovered, we can still wonder about the amount of parallelism or reversal that is found on the tree. A particular character state may have evolved independently in two lineages, and multiple hits may cause a particular nucleotide position to return to an ancestral state. Several indices have been developed to measure the relative amount of homoplasy found in a particular tree. For example, the per-character consistency index

(ci) is defined as m/s, where m is the minimum possible number of character changes (steps) on any tree, and s is the actual number of steps on the current tree. This index hence varies from one (no homoplasy) towards zero (a lot of homoplasy). The ensemble consistency index CI is a similar index, but summed over all characters.

The per-character retention index (ri) is defined as the ratio of (1) the differences between the maximal number of steps for the character on any cladogram and the actual number of steps on the current tree and (2) the differences between the maximal number of steps for the character on any cladogram and the minimum possible number of character changes on any tree (Farris 1989). Therefore, the retention index becomes zero when the site is least informative for MP tree construction, that is, when the difference between the maximal number of steps for the character on any cladogram and the actual number of steps on the current tree is zero.

7 NODE SUPPORT AND TREE COMPARISON

Measures of nodal support provide a useful summary of how well data support the relationships defined by a tree. In the MP approach, the Bremer support (decay index) for a clade can be computed as a measure of the confidence on that particular clade. The Bremer support is the number of extra steps you need to construct a tree (consistent with the characters) where that clade is no longer present. When several genes are included in the analysis, the parsimony-based method of partitioned branch support (PBS) estimates the amount that each dataset contributes to a particular clade support, so that we can estimate the extent to which the data partition supports the most parsimonious tree over trees not including a particular clade (Gatesy et al. 1999). An equivalent "partitioned likelihood support" (PLS) can be obtained for each dataset under a likelihood-based approach (Lee & Hugall 2003). Most measures of nodal support attempt to estimate the degree to which an analysis has converged on a stable result. Of course, high support values do not mean that a node is accurate, only that it is well supported by the data. It is well known that model misspecification and taxon sampling can mislead the analysis (Hedtke et al. 2006).

Currently, the nonparametric bootstrap is one of the most widely used methods for assessing nodal support (Felsenstein 1985). The nonparametric bootstrap is a statistical method by which distributions that are difficult to calculate exactly can be estimated by the repeated creation and analysis of artificial datasets. A number of replicates (typically at least 1000) of the original characters (e.g., sites of a DNA sequence alignment) are randomly produced with replacement, obtaining a new dataset in which some characters are represented more than once, some appear once, and some are deleted. The perturbed datasets are each analyzed in the same manner as for the real data, and the number of times that each grouping of species appears in the resulting profile of cladograms is taken as an index of relative support for that grouping.

Perhaps the best interpretation of the bootstrap is that it quantifies the sensitivity of a node to perturbations in the data (Holmes 2005). However, as commonly implemented, the bootstrap gives a biased estimate of accuracy (Hillis & Bull 1993; Holmes 2005), where accuracy is defined as the probability of obtaining a correct phylogenetic reconstruction (Penny et al. 1992). The statistical theory of bootstrap requires that all positions of an alignment are independently and identically distributed, and this assumption does not apply to nucleotide or amino acid sequences. It is worthwhile to point out the difference between nonparametric and parametric bootstraps. In the nonparametric bootstrap, new datasets are generated by resampling from the original data, whereas in the parametric bootstrap, the data are simulated according to the hypothesis being tested. This well-known bias of the bootstrap has led researchers to seek other methods of estimating nodal support, and perhaps the most popular alternative is Bayesian posterior probability (Larget & Simon 1999; Yang & Rannala 1997). A nodal posterior probability is the probability that a given node is found in the true tree, conditional on the observed data, and the model (including both the prior model and the likelihood model). Early observations of Bayesian inference in phylogenetics

demonstrated a tendency for posterior probabilities to be more extreme than ML nonparametric bootstrap proportions, although the two tended to be correlated (Buckley et al. 2002). Finally, Lewis et al. (2005) demonstrated that if a polytomy exists but is not accommodated in the prior, resolution of the polytomy will be arbitrary and the nodal support indicated by the posterior probability will appear unusually high compared to ML bootstraps. Because we have little knowledge of the goodness of fit between data and model in typical phylogenetic studies (although goodness of fit tests do exist), we have little idea of the seriousness of the problem of model misspecification in current implementations of Bayesian phylogenetic inference. Goodness of fit tests define how well a statistical model fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. The great advantage of the Bayesian posterior probability is that this statistic is drawn from the same distribution that determines the best estimate of tree topology, as opposed to a bootstrap analysis that requires 1000 reruns of the analysis.

7.1 *Statistical tests of tree topologies*

A variety of topology tests has been designed to compare different trees and thereby test alternative hypotheses of phylogenetic relationships. There is a fundamental difference between testing a priori phylogenetic hypotheses versus testing those generated through analyses. The Templeton (1983) test and Kashino-Hasegawa (KH) test (Kishino & Hasegawa 1989) are nonparametric tests designed to compare pairs of topologies selected before a phylogenetic analysis is run, with the Templeton test using a parsimony framework and the KH test using a likelihood framework. However, these approaches may become too liberal when one of the alternative topologies is one estimated from the data (Goldman et al. 2000). In this case, the most widely used parametric test is the Swofford-Olsen-Waddell-Hillis (SOWH) test (Swofford et al. 1996), which uses parametric bootstrapping to simulate replicate datasets that are in turn used to obtain the null distribution. Shimodaira & Hasegawa (1999) have described a non-parametric bootstrap test that directly succeeds the KH test, considering all possible topologies and making the proper allowance for their comparison with the ML topology derived from the same data. Because of the nature of the null hypotheses employed by the nonparametric tests, the Templeton, SH, and KH tests are generally more conservative than the parametric tests (Aris-Brosou 2003; Buckley 2002; Goldman et al. 2000). The more explicit reliance on models of evolution by the parametric tests makes them very powerful tests, yet they are also more susceptible to model misspecification (Buckley 2002; Shimodaira 2002). Bayesian tests of topology are becoming more commonly implemented than the frequentist tests (Aris-Brosou 2003). The Bayesian tests generally rely on Bayes factors to compare marginal likelihoods generated under two hypotheses corresponding to different topologies (Kass & Raftery 1995). The use of Bayes factors in testing topologies will likely receive much greater attention in the future, since it allows for comparison of models that are not hierarchically nested (Nylander et al. 2004).

8 USING MULTIPLE GENES

The best phylogenetic estimates come from using robust inference methods coupled with realistic evolutionary models. However, good estimates of phylogeny ultimately depend on good datasets. The two most obvious ways of increasing the accuracy of a phylogenetic inference are to include more sequences in the data and/or to increase the length of the sequences used. Goldman (1998) showed that adding more sequences to an analysis does not increase the amount of information relating to different parts of the tree uniformly over that tree, whereas the use of longer sequences results in a linear increase in information over the whole of the tree. A potentially powerful approach is to analyze the sequences as a concatenated whole or "meta-sequence." The simplest

analysis would be to assume that all the genes have the same patterns and rates of evolution (Cao et al. 1994). This naïve method should only be used when there is substantial evidence of a consistent evolutionary pattern across all the genes, which can be assessed by statistical tests of different models (as described above). Otherwise, differences amongst gene replacement patterns or rates can lead to biased results. More advanced analyses of concatenated sequences are possible, which allow for heterogeneity of evolutionary patterns among the genes studied (Yang 1996b). This heterogeneity might be as complex as allowing each gene to evolve with different replacement patterns, and with different rates of replacement in all branches of the gene trees (Yang 1997).

The contradictions in the different phylogenetic reconstructions based on analysis of different protein, gene, or noncoding sequences raise questions concerning the variability of evolutionary processes and the reliability of averaging schemes such as sequence concatenation (Teichmann & Mitchison 1999). Lateral transfer, fusion events, and recombination can make the evolutionary relationships among genes unreliable indicators of the phylogenetic relationships among the species. In that case, the Partition Homogeneity Test or incongruence length difference (ILD) test (Farris et al. 1994) could be used for testing if every gene in the analysis is giving a heterogeneous signal under the maximum parsimony framework. However, this heterogeneity can come solely from branch length differences and is not necessarily indicative of topological differences with different data subsets. Finally, in the so-called "total evidence" approach, genes are concatenated end to end, including also information from morphological characters, and the whole dataset is analyzed using parsimony (Ahyong & O'Meally 2004). This has the great advantage of taking into account the different amounts of sequence in different loci and of combining the evidence in a single tree that does not depend on an arbitrary choice of consensus tree method. Still, if different loci have substantially different rates of change, combining them into one dataset obscures evidence that indicates that one locus should be treated differently from another. In order to include this heterogeneity in the phylogenetic analysis, Kolaczkowski & Thornton (2004) recently presented a new mixture model to account for partitioned sequences. Even though there were some concerns about the computational burdens of implementing more complex evolutionary models, these concerns can be accommodated in a likelihood-based analysis. By using MCMC sampling, mixture models and likelihood-based approaches could be used even when evolution is heterogeneous (Pagel & Meade 2004).

9 SUMMARY OF METHODS AND CONCLUSION

> *"The time will come I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of nature."*
>
> Darwin (1857)

Throughout this review, several methods have been introduced that try to infer phylogenetic relationships between species using molecular data. **(1) Maximum parsimony** seeks to find the tree that is compatible with the minimum number of substitutions among sequences. Finding a maximally parsimonious cladogram is usually a computationally intensive task, but for large problems, fast heuristic algorithms can be employed, even though they cannot guarantee to find the optimal cladogram. Parsimony analysis has been criticized for requiring very stringent assumptions of constancy for substitution rates across sites and similar substitution rates among lineages. It has been found that the performance of MP deteriorates when mutational rates differ between nucleotides or across sites (Yang 1996b) or if evolutionary rates are highly variable among evolutionary lineages (Hendy & Penny 1989; DeBry 1992).

As more divergent sequences are analyzed, the overall degree of homoplasy generally increases, and this implies that the true evolutionary tree becomes less likely to be the one with the least number

of changes. Furthermore, when two evolutionary lineages that have undergone a high level of sequence evolution are separated by a short lineage, the long lineages will tend to be spuriously joined in the most parsimonious cladogram produced from the resulting sequence data. Combinations of conditions when this occurs are often called the "Felsenstein zone," and parsimony is particularly affected by this problem because of its inability to deal with homoplasy (Huelsenbeck 1997). Nevertheless, MP methods have some advantages over other tree-building methods. Parsimony analysis is very useful for dealing with morphological characters or some types of molecular data such as insertion sequences and insertion/deletions, and weighted MP methods can be constructed to incorporate information on the evolutionary process.

(2) Distance methods such as neighbor joining seek to reconstruct the tree topology that best represents the matrix of distances between pairs of taxonomic units. As with all greedy methods, the NJ algorithm is not guaranteed to find the globally best solution to a general distance matrix with error (Pearson et al. 1999). In an effort to alleviate this problem, some generalizations of the NJ method have been proposed that explore multiple low-error paths in progressively clustering the sequences (Kumar 1996; Pearson et al. 1999). However, the most serious problem with distance methods is that they require a reliable measure of evolutionary distances between sequences. When evolutionary rates vary from site to site in molecular sequences, distances can be corrected for this variation. When variation of rates is large, these corrections become important. In likelihood methods, the correction can use information from changes in one part of the tree to inform the correction in others, but a distance matrix method is inherently incapable of propagating the information in this way. Thus, distance matrix methods must use information about rate variation substantially less efficiently than likelihood methods (Felsenstein 2004).

(3) Likelihood-based methods permit the application of mathematical models that incorporate our knowledge on typical patterns of sequence evolution, resulting in more powerful inferences. Furthermore, they use a complete statistical methodology that permits hypothesis tests, enabling validation of the results at all stages: from the values of parameters in evolutionary models, through the comparison of competing models describing the biological factors most important in sequence evolution, to the testing of hypotheses of evolutionary relationship. Computer programs for the robust statistical evolutionary analysis of molecular sequence data are widely available (Table 1).

Nevertheless, ML methods do not directly assign probabilities to the parameters, and if one wants to describe the uncertainty in an estimate, one has to repeat the analysis multiple times (bootstrap), increasing the computational cost. In **Bayesian inference**, information can be drawn directly from the simulated joint distribution of parameters at a reasonable computational cost. On the other hand, a review of the current Bayesian phylogenetic literature indicates that much more emphasis needs to be placed on developing more realistic models, checking the effects of the priors, and monitoring the convergence of posterior distributions.

All in all, it should be pointed out that systematic error will confound any tree reconstruction method. Situations such as long-branch-attraction and base-compositional bias are examples of systematic bias. When inferring phylogenies, we try to define the actual succession of divergence events from the present sampled sequences. This means that the actual genes sampled (gain and loss of genes happens, but we rely only on those genes for which homology can be ascertained), species sampled (extinction of intermediate taxa), selection (causing either among-sites or among-loci rate variation), and the population parameters (mutation rates, recombination rates, effective population sizes, etc.) all may influence the strength of the phylogenetic signal. In conclusion, phylogenetic inference should be approached not as a tool for getting a definitive answer for a taxonomical problem, but rather as a tool for asking new questions on the evolution of molecules and morphology in different species and for trying to uncover the causes of such differences in their evolution.

Table 1. A sampling of phylogenetic software to perform evolutionary analyses (see http://evolution. genetics. washington.edu/phylip/software.html for a comprehensive list).

Name	Methods Implemented	Web	Citation
ClustalW	Progressive multiple sequence alignment	http://www.ebi.ac.uk/clustalw/	Thompson et al. 1994
MUSCLE	Progressive alignment and refinement using restricted partitioning	http://www.drive5.com/muscle/	Edgar 2004
POY	Optimization alignment	http://research.amnh.org/scicomp/projects/poy.php	Varón et al. 2007
BAli-Phy	Bayesian inference of alignment and topology	http://www.biomath.ucla.edu/msuchard/bali-phy/index.php	Suchard & Redelings 2006
ModelTest	Model selection	http://darwin.uvigo.es/software/modeltest.html	Posada & Crandall 1998
MrModelTest	Model selection	http://www.abc.se/~nylander/	Nylander 2004
MEGA	Distance, parsimony and maximum likelihood	http:www.megasoftware.net/index.html	Tamura et al. 2007
PAUP	Maximum parsimony, distance matrix, maximum likelihood	http://paup.csit.fsu.edu/	Swofford 2003
PHYLIP	Maximum parsimony, distance matrix, maximum likelihood	http://evolution.genetics.washington.edu/phylip.html	Felsenstein 2005
TNT	Maximum parsimony, ratchet	http://www.zmuc.dk/public/phylogeny/TNT/	Goloboff et al. 2003
Winclada	Maximum parsimony, ratchet	http://www.cladistics.com/aboutWinc.htm	Nixon 2002
PhyML		http://atgc.lirmm.fr/phyml/	Guindon & Gascuel 2003
GarLi	Maximum likelihood using genetic algorithms	http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html	Zwickl 2006
PAML	Maximum likelihood	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang 1997
RAxML-HPC	Maximum likelihood, simple maximum parsimony	http://icwww.epfl.ch/~stamatak/	Stamatakis et al. 2005
MultiDivTime	Dating, molecular clock using Bayes MCMC	http://statgen.ncsu.edu/thorne/multidivtime.html	Thorne & Kishino 2002
BayesPhylogenies	Bayesian inference	http://www.evolution.rdg.ac.uk/SoftwareMain.html	Pagel & Meade 2004
MrBayes	Bayesian inference	http://mrbayes.csit.fsu.edu/index.php	Ronquist & Huelsenbeck 2003

ACKNOWLEDGEMENTS

Thanks are due to P. Abelló, M. Pascual, and E. Macpherson for encouraging the completion of this study. This work was supported by a pre-doctoral fellowship awarded by the Autonomous Government of Catalonia (2006FIC-00082) to FP and by a grant from the US NSF EF-0531762 awarded to KAC. FP is part of the research group 2005SGR-00995 of the Generalitat de Catalunya. Research was funded by project CGL2006-13423 from the Ministerio de Educacion y Ciencia. FP acknowledges EU-Synthesys grant (GB-TAF-1637).

REFERENCES

Ahyong, S.T. & O'Meally, D. 2004. Phylogeny of the Decapoda. Reptantia: resolution using three molecular loci and morphology. *Raffl. Bull. Zool.* 52: 673–693.

Aris-Brosou, S. 2003. Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in presence of conflicting signals under misspecified models. *Syst. Biol.* 52: 781–793.

Beaumont, M. & Rannala, B. 2004. The Bayesian revolution in genetics. *Nat. Rev. Genet.* 5: 251–261.

Benavides, E., Baum, R., McClellan, D. & Sites, J.W. 2007. Molecular phylogenetics of the lizard genus Microlophus (Squamata: Tropiduridae): aligning and retrieving indel signal from nuclear introns. *Syst. Biol.* 56: 776–797.

Brown, W.M., Prager, E.M., Wang, A. & Wilson, A.C. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* 18: 225–239.

Bruno, W.J., Socci, N.D. & Halpern, A.L. 2000. Weighted neighbor-joining: a likelihood-based approach to distance based phylogeny reconstruction. *Mol. Biol. Evol.* 17: 189–197.

Buckley, T.R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51: 509–523.

Buckley, T.R., Arensburger, P., Simon, C. & Chambers, G. K. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. *Syst. Biol.* 51: 4–18.

Cao, Y., Adachi, J., Janke, A., Pääbo, S. & Hasegawa, M. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.* 39: 519–527.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17: 540–552.

Castresana, J. 2007. Topological variation in single-gene phylogenetic trees. *Genome Biol.* 8: 216.

DeBry, R.W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. *Mol. Biol. Evol.* 9: 537–551.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* 32: 1792–1797.

Edwards, A.W.F. & Cavalli-Sforza, L.L. 1964. Reconstruction of evolutionary trees. In: McNeill, J. (ed.), *Phenetic and phylogenetic classification*: 67–76. London: Systematics Association Publication.

Egan, A.N. & Crandall, K.A. 2008. Incorporating gaps as phylogenetic characters across eight DNA regions: ramifications for North American Psoraleeae (Leguminosae). *Mol. Phylogenet. Evol.* 46: 532–546.

Farris J.S. 1983. The logical basis of phylogenetic analysis. In: Platnick, N.J. & Funk, V.A. (eds.), *Advances in Cladistics*: 1–36. New York: Columbia Univ. Press.

Farris, J.S. 1989. The retention index and the rescaled consistency index. *Cladistics* 5: 417–419.

Farris, J.S., Källersj, M., Kluge, A.G. & Bult, C. 1994. Testing significance of incongruence. *Cladistics* 10: 315–319.

Felsenstein, J. 1978. The number of evolutionary trees. *Syst. Zool.* 27: 27–33.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783–791.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates Inc., Massachusetts. 664 pp.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Frank, A.C. & Lobry, J.R. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238: 65–77.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14: 685–695.

Gascuel, O., Bryant, D. & Denis, F. 2001. Strengths and limitations of the minimum-evolution principle. *Syst. Biol.* 50: 621–627.

Gatesy, J., O'Grady, P. & Baker, R.H. 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher-level artiodactyl taxa. *Cladistics* 15: 271–313.

Glenner, H., Lützen, J. & Takahashi, T. 2003. Molecular evidence for a monophyletic clade of asexually reproducing parasitic barnacles: *Polyascus*, new genus (Cirripedia: Rhizocephala). *J. Crust. Biol.* 23: 548–557.

Goldman, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36: 182–198.

Goldman, N. 1998. Phylogenetic information and experimental design in molecular systematics. *Proc. R. Soc. London Ser. B* 265: 1779–1786.

Goldman, N., Anderson, J.P. & Rodrigo, A.G. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49: 652–670.

Goloboff, P., Farris, J.S. & Nixon, K. 2003. TNT: Tree analysis using new technology. Program and documentation, available from the authors, and at http://www.zmuc.dk/public/phylogeny.

Grant, T. & Kluge, A.G. 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. *Cladistics* 19: 379–418.

Guindon, S. & Gascuel, O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52: 696–704.

Harvey, P.H. & Pagel, M.D. 1991. *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Hedtke, S.M., Townsend, T.M. & Hillis, D.M. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* 55: 522–529.

Hendy, M.D. & Penny, D. 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38: 297–309.

Hillis, D.M. & Bull, J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence on phylogenetic analysis. *Syst. Biol.* 42: 182–192.

Holmes, I. 2005. Using evolutionary expectation maximization to estimate indel rates. *Bioinformatics* 21: 2294–2300.

Huelsenbeck, J.P. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44: 17–48.

Huelsenbeck, J.P. 1997. Is the Felsenstein zone a fly trap? *Syst. Biol.* 46: 69–74.

Huelsenbeck, J.P. & Rannala, B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276: 227–232.

Huelsenbeck, J.P., Larget, B., Miller, R.E. & Ronquist, F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol.* 51: 673–688.

Kass, R.E. & Raftery, A.E. 1995. Bayes factors. *J. Amer. Stat. Assoc.* 90: 773–795.

Keller, I., Bensasson, D. & Nichols, R.A. 2007. Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes. *PLoS Genet* 3(2): e22. doi:10.1371/journal.pgen.0030022

Kishino, H. & Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29: 170–179.

Kitchin, I.J., Forey, P.L., Humphries, C.J. & Williams, D.M. 1998. *Cladistics*. Oxford: Oxford University Press.

Kluge, A.G. & Farris, J.S. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18: 1–32.

Kolaczkowski, B. & Thornton, J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431: 980–984.

Kumar, S. 1996. A stepwise algorithm for finding minimum evolution trees. *Mol. Biol. Evol.* 13: 584–593.

Lake, J.A. 1991. The order of sequence alignment can bias the selection of tree topology. *Mol. Biol. Evol.* 8: 378-385.

Larget, B. & Simon, D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16: 750–759.

Lee, M.S.Y. & Hugall, A.F. 2003. Partitioned likelihood support and the evaluation of data set conflict. *Syst. Biol.* 52: 15–22.

Lee, M.S.Y. & Hugall, A.F. 2006. Model type, implicit data weighting, and model averaging in phylogenetics. *Mol. Phylogenet. Evol.* 38: 848–857.

Liò, P. & Goldman, N. 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8: 1233–1244.

Lutzoni, F., Wagner, P., Reeb, V. & Zoller, S. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Syst. Biol.* 49: 628–651.

Moore, G., Goodman, M. & Barnabas, J. 1973. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *J. Theor. Biol.* 38: 423–457.

Needleman, S.B. & Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48: 443–53.

Nielsen, R. & Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929–936.

Nixon, K.C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics* 15: 407–414.

Nixon, K.C. 2002. WinClada ver. 1.00.08. Published by the author, Ithaca, NY.

Nylander, J.A.A. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Nylander, J.A., Ronquist, F., Huelsenbeck, J.P. & Nieves-Aldrey, J.L. 2004. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53: 47–67.

Ogden, T.H. & Whiting, M. 2003. The problem with "the Paleoptera Problem": sense and sensitivity. *Cladistics* 19: 432–442.

Ogden, T.H. & Rosenberg, M. 2006. How should gaps be treated in parsimony? A comparison of approaches using simulation. *Mol. Phylogenet. Evol.* 42: 817–826.

Ogden, T.H. & Rosenberg, M. 2007. Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. *Syst. Biol.* 56: 182–193.

Ota, S. & Li, W.H. 2001. NJML+: An extension of the NJML method to handle protein sequence, data and computer software implementation, *Mol. Biol. Evol.* 18: 1983–1992.

Pagel, M. & Meade, A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence of character-state data. Syst. Biol. 53: 571–581.

Pearson, W.R., Robins, G. & Zhang, T. 1999. Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. *J. Mol. Evol.* 16: 806–816.

Penny, D., Hendy, M.D. & Steel, M.A. 1992. Progress with methods for constructing evolutionary trees. *Trends Ecol. Evol.* 7: 73–79.

Philippe, H. & Forterre, P. 1999. The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* 49: 509–523.

Porter, M.L., Cronin, T., McClellan, D.A. & Crandall, K.A. 2007. Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins. *Mol. Biol. Evol.* 24: 253–268.

Posada, D. & Crandall, K.A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14: 817–818.

Posada, D. & Crandall, K.A. 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50: 580–601.

Posada, D. & Buckley, T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over Likelihood Ratio Tests. *Syst. Biol.* 53: 793–808.

Redelings, B. & Suchard, M. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* 54: 401–418.

Robinson, D.F. 1971. Comparison of labeled trees with Valency Three. *J. Combin. Theor.* 11: 105–119.

Rogers, J.S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 46: 354–357.

Ronquist, F. & Huelsenbeck, J.P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Rzhetsky, A. & Nei, M. 1993. Theoretical foundation of the minimum evolution method of phylogenetic inference. *Mol. Biol. Evol.* 10: 1073–1095.

Schöniger, M. & von Haeseler, A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* 3: 240–247.

Segawa, R.D. & Aotsuka, T. 2005. The mitochondrial genome of the Japanese freshwater crab, Geothelphusa dehaani (Crustacea: Brachyura): evidence for its evolution via gene duplication. *Gene* 355: 28–39.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51: 492–508.

Shimodaira, H. & Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16: 1114–1116.

Song, H., Buhay, J.E. Whiting, M.F. & Crandall, K.A. 2008. DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Nat. Acad. Sci. USA*: 105: 13486–13491.

Stamatakis, A., Ludwig, T. & Meier, H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456–463.

Steel, M.A., Hendy, M.D. & Penny, D. 1988. Loss of information in genetic distances. *Nature* 336: 118.

Steel, M. & Penny, D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17: 839–850.

Suchard, M.A. & Redelings, B.D. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22: 2047–2048.

Swofford, D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C. & Mable, B.K (eds.), *Molecular Systematics*: 407–514. Sunderland: Sinauer Associates.

Teichmann, S.A. & Mitchison, G. 1999. Making family trees from gene families. *Nat. Genet.* 21: 66–67.

Templeton, A.R. 1983. Convergent evolution and nonparametric inferences from restriction data and DNA sequences. In: Weir, B.S. (ed.), *Statistical Analysis of DNA Sequence Data*: 151–179. New York: Marcel Dekker, Inc.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22: 4673–4680.

Thorne, J.L., Kishino, H. & Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33: 114–124.

Thorne, J.L. & Kishino, H. 1992. Freeing phylogenies from artifacts of alignment. *Mol. Biol. Evol.* 9: 1148–1162.

Thorne, J.L. & Kishino, H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* 51: 689–702.

Varón, A., Vinh, L.S., Bomash, I. & Wheeler, W.C. 2007. POY 4.0 Beta 2635. American Museum of Natural History.

Weston, P.H. 1994. Methods for rooting cladistic trees. In: Scotland, R.W., Siebert, D.J. & Williams, D.M. (eds.), *Models in Phylogeny Reconstruction*: 125–155. Oxford: Oxford Univ. Press.

Yang, Z., Goldman, N. & Friday, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11: 316–324.

Yang, Z. 1996a. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11: 367–372.

Yang, Z. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42: 587–596.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13: 555–556.

Yang, Z. & Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: Markov chain Monte Carlo methods. *Mol. Biol. Evol.* 14: 717–724.

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.-M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431–449.

Zanotto, P.M., Kallas, E.Q., de Souza, R.F. & Holmes, E.C. 1999. Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics* 153: 1077–1089.

Zwickl, D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.