Preface

This volume of the *Methods in Molecular Biology* series is entirely devoted to the study of steroid receptor biology. Steroid hormone receptors represent a powerful system for the study of both the most fundamental molecular mechanisms of gene regulation and control and the gross physiological responses of organisms to steroid hormones. Research in this field has brought forth advances in the treatment of cancer, endocrine disorders, and reproductive biology, and allowed elucidation of the fundamental biological mechanisms of gene expression. In *Steroid Receptor Methods: Protocols and Assays*, the reader will find a collection of methods and protocols submitted by many fine steroid receptor researchers from throughout the world. These authors have been instructed to create a highly informative cross-section of the latest research techniques available. The resulting work is timely, useful, and approachable for both the experienced researcher and the novice to the field. Because the steroid receptor family is represented by a wonderfully diverse, yet strongly interrelated set of steroid receptor proteins, *Steroid Receptor Methods* contains protocols for the production and purification of a variety of receptor forms, including the progesterone, glucocorticoid, and androgen receptors. These procedures provide the raw material needed to conduct sophisticated biochemical analysis of receptor properties. Other techniques presented allow the reader to perform biochemical experiments on DNA binding characteristics, hormone binding assays, and protocols using combinatorial chemistry for drug discovery. Because steroid receptor effectiveness is influenced by a variety of cellular proteins, there are included in this volume a series of novel protocols utilizing the latest advances in immunochemistry, yeast two-hybrid screening, fluorescence, and other biochemical and cellular techniques to detect and detail these interactions. These techniques include both in vitro and in vivo approaches to provide the widest possible selection of tools to the modern biological researcher. Finally, in recognition of the growing importance of bioinformatics in biological research, several chapters have been included to guide and assist the modern research biologist in harnessing this increasingly valuable resource. These chapters locate and make accessible to the researcher the diverse computational tools currently available via the Internet. Taken together these chapters provide both novice and experienced researchers alike a set of invaluable tools to advance and extend their research.

Ben Lieberman, PhD

2

Phylogenetic Inference and Parsimony Analysis

Llewellyn D. Densmore III

1. Introduction

Application of phylogenetic inference methods to comparative endocrinology studies has provided researchers with a new set of tools to aid in understanding the evolution and distribution of gene families. Phylogeny, as defined by Hillis et al. ***(1)***, is the "historical relationships among lineages of organisms or their parts (e.g., genes)." Inferring phylogeny is a way of generating a best estimate of the evolutionary history of organisms (or gene families), based on the information (often incomplete, as in a gene sequence) that is available. The use of phylogenetic analyses, specifically those methods that are based on maximum parsimony, has changed the way in which characters and character states are determined and interpreted. Maximum parsimony (often simply called "parsimony") seeks to estimate a parameter based on the minimum number of events required to explain the data. In this type of phylogenetic analysis, the best or optimal tree (generally portrayed as either a cladogram or phylogram, *see* **Note 1**) is that topology which requires the fewest number of character-state changes (*see* below). That tree is arrived at based upon consideration of shared, derived characters. This method assumes that when two taxa (or genes) share a homologous derived character state, they do so because a common ancestor of both had that character state. One goal of phylogenetic analysis that is always implied (and often stated) is to avoid using characters that are homoplastic. Characters that have homoplasy have similarities in character states for reasons other than inheritance from a common ancestor, including convergent and parallel evolution or a reversal of state (e.g., A → G → A).

The most common types of molecular characters that are used in phylogenetic analysis of steroid hormone receptors are the primary sequence positions of DNA or proteins, cDNA sequences derived from RNA, and amino acid

From: *Methods in Molecular Biology, vol. 176: Steroid Receptor Methods: Protocols and Assays*
Edited by: B. A. Lieberman © Humana Press Inc., Totowa, NJ

sequences of proteins inferred from cDNAs. Therefore, in most situations phylogenetic analysis of these sequences is virtually identical to the analysis of sequences in a molecular systematics study attempting to resolve relationships among different taxa. In this chapter, a number of the most commonly applied methods of analyzing such data sets are introduced, emphasizing the phylogenetic approach using parsimony. Although parsimony-based models are emphasized here, other approaches such as maximum likelihood, can also be used for nucleotide-based ***(2)*** or amino acid based ***(3,4)*** phylogeny reconstruction. Maximum likelihood methods are used to evaluate a hypothesis about evolutionary history based on the probability that the proposed model of the evolutionary process and hypothesized history would give rise to the observed data ***(5)***. There are also a number of phenetic approaches (those based on overall character similarity, e.g., unweighted pair group method with averages), some of which are sometimes considered to be more or less phylogenetic methods (e.g., neighbor joining) ***(6)***. All phenetically-based trees (called phenograms) are ultimately generated from similarity measures that are used to estimate genetic distances. Application of these methods certainly may have merit for some studies of steroid hormone receptors, and although the criteria for recovering the sequences and their alignment are literally the same for all of these methods, this discussion is restricted to phylogenetic analyses that are based on maximum parsimony.

Phylogenetic analysis deals with both characters and character states. As noted above, molecular characters are usually the positions of the nucleotides of the DNAs or amino acids for the proteins that are being compared. Virtually all sequence analyses lead to the generation of multistate characters; for nucleotide-based data sets, the character states are normally A, G, C, or T (although a fifth state, which accounts for missing bases, is also often included); for protein data sets, the states would then be the 20 naturally occurring amino acids (again, a state for a gap character could also be included). Multistate characters may be ordered or unordered: They are said to be ordered if a particular state exists between two states (e.g., if mutation to T were required as an intermediate condition during a change from G to A). This requirement is virtually never observed in molecular data, so it is assumed that most nucleotide or amino acid sequence data sets are both multistate and unordered (indicating any state can be reached from any other state).

Homology (inferred common ancestry of genes or gene products) is the characteristic that actually allows one to compare sequences. The two most important types of homology in most molecular data should be distinguished. Orthology assumes that the common ancestry of two sequences can be traced back to a speciation event. Paralogy indicates that the common ancestry of the sequences can be traced back to a gene-duplication event.

A series of sequences that are either orthologs (comparing taxa) or paralogs (comparing lineages of genes), and which all share the same common ancestor, are said to be monophyletic. Monophyletic groups can include gene sequences from different members of a genus or species or related sequences of a gene family (e.g., the estrogen β receptors). In any phylogenetic analysis, it is advisable to employ outgroup comparison. The so-called "ingroup" includes members of a taxon (or genes in a lineage), assumed to be monophyletic. The ingroup sequences can be distinguished from sequences outside of it by having a larger number of shared, derived characters (synapomorphies). Related genes (such as estrogen α receptors when compared to estrogen β receptors) or taxa (such as alligators when compared to crocodiles), might have an evolutionary history similar to the ingroup. They would share fewer synapomorphies with the ingroup members, but would share some number of primitive characters (symplesiomorphies) with the ingroup. Inclusion of these outgroup sequences allows for rooting (*see* **Note 2**) of the phylogenetic tree and verification that all members of the ingroup lineage are more closely related to one another than to some other sequence. At least one outgroup sequence should always be employed in phylogenetic analysis, and in some cases it is important to have two or more (*see* below).

At first glance, the use of primary sequence positions as characters for phylogenetic inference might be considered reasonably straightforward. Examining two purportedly homologous sequences, counting the number of bases or amino acids from one terminus and comparing the two sequences (at say amino acid positions 1–65 for some protein), would allow the absolute number of differences between two sequences to be readily ascertained. However, this simplicity may be misleading. In assessing phylogeny, establishing positional homology is critical and can be complicated. In comparing amino acid sequences, having positional homology indicates not only that both sequences are homologous (e.g., both are estrogen β receptors), but also that every amino acid occurring at a particular position in the protein sequences (e.g., amino acid 43) being compared trace their ancestry to a single position that occurred in the protein sequence of a common ancestor ***(5)***. In all but closely related protein genes and/or the most highly conserved sequences, insertions or deletions probably will have occurred in the nucleotide sequences and thus, often in the amino acid sequences. These must be accounted for by alignment to ensure positional homology. Therefore, proper alignment of sequences, considered by many to be the most critical aspect of molecular phylogeny, will be the first method that is addressed (*see* **Subheading 2.**).

2. Materials

Virtually all researchers have their favorite phylogenetic analysis package(s). For all-around versatility with molecular sequence data, Phylogenetic Analysis Using Parsimony (PAUP*) ***(19)***, a package developed by Swofford, is difficult to beat, especially if one has had a MacIntosh computer. Recently, PC-compatible and UNIX versions have joined the VAX/VMS and Mac OS packages. Reasonable ($85–200 for virtually all operating systems) to acquire through Sinauer Associates (orders@sinauer.com) and menu-driven, it is the most popular phylogenetic analysis package for molecular data. It is the package that my lab uses almost exclusively for phylogenetic analyses. PAUP* has a large number of programs besides those that are parsimony-based and will read a wide range of data input files, including Nexus, PHYLIP, and FASTA.

Perhaps even more versatile, but probably not as easy to use is Felsenstein's Phylogenetic Inference Package (PHYLIP) ***(2)***, a broad package of programs that like PAUP* can perform not only parsimony, but also maximum likelihood and distance analyses. The price of PHYLIP is even more attractive than PAUP*, since it can be acquired at no charge by anonymous ftp from: evolution.genetics.washington.edu (in directory pub/phylip), or by accessing the World Wide Web site: (http://evolution.genetics.washington.edu/phylip.html).

An additional service that Felsenstein has provided at the PHYLIP website is a documented list including 175 programs used for reconstructing relationships. These range from more specialized packages that will primarily perform only alignments (e.g., ClustalW, MacVector, and MALIGN), and deal mainly with genetic distance analyses (e.g., MEGA 2B) or maximum likelihood analyses (e.g., MOLPHY or PAML), to those that allow trees to be interactively manipulated (e.g., MacClade). It also lists those packages that contain a large number of applications (such as PAUP*, PHYLIP, Hennig86, VOSTORG). Included in the documentation for each listing are how to acquire the various programs or packages, a general assessment of the analyses each are able to perform, and any cost that will be incurred.

3. Methods

3.1. Alignment

Possibly the most difficult and poorly understood aspect of phylogenetic analysis is alignment. Local alignment algorithms find all matches in a database search above a certain defined threshold (e.g., 50%). Data bank searches, such as those employed by the National Center for Biotechnology Improvement (NCBI) data bank (http://www.ncbi.nlm.nih.gov/), use several of these algorithms. Two examples are BLAST ***(7)*** and FASTA ***(8)***. The program

"Entrez" available at the NCBI address above allows rapid evaluation of both nucleotide and protein databases. Once genes of interest are identified, Entrez allows location of many similar sequences (however, not necessarily homologous). These can be identified by taxonomic group, terms in titles or abstracts of papers, authors, key words, accession numbers from the database, gene names, and so on. Then the best matches can be extracted and aligned prior to phylogenetic analysis.

Pairwise sequence alignment (which seeks to align two entire homologous regions) is accomplished by the inclusion of gaps, which correspond to insertions or deletions, and balancing these with matches. Most sequence alignment programs are ultimately a derivation of the global alignment program originally developed by Needleman and Wunsch ***(9)***. Aligning sequences can be simple or tedious, depending on the levels of sequence divergence. However, it should be recognized that if one uses enough gaps, ultimately any two sequences can be aligned, therefore gap penalties must be assigned. The gap penalties are typically a combination of both the gap number and the size of the gaps. The former are usually penalized more heavily than gap size because there is no reason to assume that insertion/deletion events will necessarily involve short sequences. In protein-coding sequences, gaps leading to frameshifts are more heavily penalized than those leading to single amino acid substitutions. Gap penalties can be assigned for unequal length sequences, although 5' or 3' gap penalties are typically lower than those found internally.

Changes leading to substitutions also confer alignment cost. This cost can be assigned as one value for all changes or can be based on a matrix of different values, the difference in the cost depends on whether the change leads to a transition or transversion (for nucleotides) or how frequent the change is. For protein sequences, different kinds of changes at the amino acid level (e.g., aliphatic to aromatic amino acid, helix former to helix breaker, and so on) can be assigned different alignment costs. Ultimately, regardless of the sequence alignment that is produced by any computer program, the final alignment should only be accepted after visual inspection, which can lead to alignment changes based on secondary levels of structure at either the nucleotide or amino acid level.

In almost every phylogenetic study, more than two sequences are being examined and there is the requirement for multiple sequence alignment. One approach is to make a series of pairwise alignments, then add all the sequences together. The overall alignment is then the sum of each additional step and compensates by inserting gaps as necessary; one caveat is that this approach is dependent on the order in which the sequences are added. Several ways of overcoming the problem of order dependence have been proposed. One method is to obtain the order of pairwise alignments from clusters in an initial tree generated for a distance matrix across all pairwise alignments ***(10)***. The program called “Clustal” ***(11)*** uses this format, as do several other programs. A similar, but somewhat modified approach is used in the program “TreeAlign” ***(12)***. PILEUP, a program in the Wisconsin Genetics Package sold by the Genetics Computer Group, uses “progressive pairwise alignment” to produce multiple alignments. All are effective, as long as visual inspection verifies the computer-generated alignment.

An alternate strategy is based on the premise that alignment is a constituent part of phylogenetic inference, rather than a treatment that is applied prior to it. The program called “MALIGN” ***(13)*** optimizes multiple alignments by searching for the alignment that minimizes the differences between the sequences. These differences are specified by the defined gap penalties and assigned costs resulting from the substitutions mentioned above. For many studies, the ability of the user to set parameters such as gap weighting and sequence order make this is a very versatile approach. Furthermore, this program outputs aligned sequences that can be used with most all of the major phylogenetic analysis programs.

3.2. Phylogenetic Analysis of Aligned Sequences Using Parsimony

Because most of this discussion is limited to parsimony analysis, it is imperative to identify the important distinctions among the different major types of parsimony and to establish criteria for the use of each, then elaborate on the most widely applied analyses. As stated earlier, parsimony is an optimality approach that seeks to find the minimal tree length. Although there are a number of ways to achieve that goal from the perspective of different algorithms, as Swofford et al. ***(5)*** state, “Algorithms tend to have short life spans,” thus, one needs to be driven by the conceptual framework and not by any specific algorithm.

3.2.1. Common Types of Parsimony and Application for Nucleotide Sequences

1. Fitch parsimony is the simplest type of analysis, which imposes no constraints on character state changes. It allows unordered, multistate changes from any one state to any other state with reversibility ***(14)***.
2. Camin–Sokal parsimony allows multistate, unordered changes, but does not allow reversibility ***(15)***.
3. Transversion parsimony. Because of the higher likelihood of transitions (T → C, C → T, A → G, G → A) over transversions (A or G → C or T [and vice versa]), transitions are ignored and only transversions are used as shared, derived charac-

ters (*see* **Note 3**). These can be recoded as either purines or pyrimidines and Wagner parsimony (*see* **Note 4**) applied.

4. Threshold parsimony, a method developed by Felsenstein ***(2)***, prevents rapidly evolving characters from adding enough length to a tree under consideration to cause it to be rejected. This is accomplished by counting the steps each character must have for a given tree, but not applying these above a specified threshold value. For example, if a character state tree requires seven changes, and the imposed threshold is four, then this character only adds four steps to the tree under consideration. Intuitively, this is an attractive method of extracting phylogenetic information in the presence of several rapidly evolving and potentially homoplastic characters.
5. Generalized parsimony, as the name implies, is the most general type of parsimony analysis, but at the same time is computationally expensive (and therefore often slow). This method assigns a cost for each transformation of every character state to all other states. These are set up in the form of a matrix of weights. In concept, it can include transversions in nucleotide sequences, as well as consider amino acid changes that result from several changes at the nucleotide level ***(5)***.

3.2.2. Common Types of Parsimony Application to Protein Sequences

1. Eck–Dayhoff (Fitch) parsimony, as above, is the simplest type of analysis. Here the genetic code is ignored and there is equal probability for any one amino acid to change to any other ***(16)***.
2. Moore–Goodman–Czelusniak (MGC) parsimony seeks trees requiring the fewest number of nucleotide substitutions at the mRNA level ***(17)***. It generalizes the Fitch parsimony approach to codons, incorporating degeneracy of genetic code and guarantees a minimum number of nucleotide substitutions required by any tree (*see* **Note 5**).
3. PROTOPARS is a program developed by Felsenstein ***(2)***, which includes aspects of both Eck and Dayhoff ***(16)*** and Moore–Goodman–Czelusniak ***(17)*** methods. It does not consider silent mutations, although the genetic code is not ignored (*see* **Note 6**).

For studies of nucleotide-based sequences, generalized parsimony and various modifications of transversion parsimony are probably the most widely applied methods. Threshold parsimony is not used as widely (primarily because of a lack of empirical data on threshold values), although it has the potential to be a valuable tool, especially for closely related sequences or those with mutational hotspots. For studies of protein-based sequences, probably the most widely applied parsimony program is PROTOPARS.

3.3. Finding Optimal Trees

When optimality criteria are outlined as in the previous subheading on types of parsimony, essentially a particular tree is being evaluated under a set of selected criteria (e.g., under transversion parsimony criteria). Finding the optimal tree (or trees) is a different problem, with several approaches that are used to solve it. The most conservative approaches use exact algorithms that typically involve either exhaustive searches or branch and bound searches.

3.3.1. Exact Methods

Exhaustive searches literally evaluate every possible tree topology. In this type of analysis, one starts off with the simplest unrooted association of taxa (three), then adds one taxon per round in all possible combinations (for four taxa, there are three possible trees; for five taxa, 15 possible trees; and so on). This number increases so rapidly that for most studies exhaustive searches are really only practical for eleven or fewer taxa (eleven taxa generate over 35 million possible trees). An advantage of this method is that with all possible trees having been considered, one can look at the frequency distribution of tree lengths (the number of steps required to produce a topology). Near-optimal trees can be identified, so that one can determine whether there are few or many solutions that are close to the most optimal tree ***(5)***.

In most studies, however, even when using a conservative approach to resolve the best tree for the data, it is not necessary to evaluate every single possible topology to find the optimal tree. The so-called "branch-and-bound method" was first applied to phylogenetic analysis by Hendy and Penny ***(18)***. This method adds new groups in all possible combinations, as long as the number of steps involved in the generation of a particular tree is equal to or less than some minimum upper bound of optimality that has been previously chosen. In this way, as new groups are added along a particular branch, if the optimal tree score is exceeded, then the entire branch (from the node that is being evaluated to all terminal groups [located at the ends of branches]) is considered suboptimal (and adding new groups cannot possibly improve the tree score). Thus, no further subsequent consideration along that branch is given (in favor of other branching sequences that do comply with the optimality criterion). In this way, the branch-and-bound still conducts an exhaustive search, but in reality only uses those topologies that can potentially lead to optimal tree resolution. For many data sets of 20 or more gene or amino acid sequences (or taxa), this approach can lead to an exact solution, i.e., a single best tree (or group of trees with identical scores) will be found for that data set.

3.3.2. General Heuristic Methods

Sometimes a data set is so large that the application of exact methods (i.e., exhaustive or branch-and-bound searches) is not practical or feasible in terms of available computing power or time. Then heuristic approaches (*see* **Note 7**) which employ approximate methods can be used. Heuristic tree searches typically use hill-climbing methods ***(5)***. One tree (randomly chosen) starts the process, then that tree is rearranged in a way that the score is improved to the minimum length. Generally for heuristic searches, one chooses some number replicates (e.g., 100, which will probabilistically evaluate many different starting trees), keeping only the shortest tree(s) found. Often, if the data set has enough information content (i.e., is not too noisy), one will find the optimal tree (or some set of equally optimal trees) that might be recovered in much longer branch-and-bound analyses. There are several ways to accomplish heuristic searches. The most commonly applied algorithms are discussed below.

1. Stepwise addition is a common way of producing a starting point for further rearrangement of taxa (or different sequences) to a growing tree. A simple description of stepwise addition follows. Starting with three taxa for the initial tree, the next taxon is added and each of the three trees that are produced is evaluated and the one with the best score is retained. In the next round, another taxon is added to the tree that was retained from the previous round and the best of these five possible trees is retained for the next round, and so on until all of the terminal taxa are added. A problem with this kind of approach is that while the position of taxon A may be optimal at a particular level of addition, if other taxa are subsequently added later on, it could make taxon A's position suboptimal. Furthermore, if two equally optimal trees exist at a particular level, one really should save both and evaluate each under the stepwise criteria. Not all packages will do this. However, stepwise addition algorithms are rapid and if the data are clean (i.e., little homoplasy), then they will quickly come up with the optimal tree with reasonably high frequency.
2. Branch swapping is a process in which stepwise addition can often be improved by choosing sets of predefined rearrangements. The underlying premise is that if one rearranges the tree(s) that are kept at each round (as in the stepwise addition method), then one of these rearrangements may well lead to a better tree that is more likely to be optimal. The three most commonly employed branch-swapping algorithms are nearest neighbor interchange, subtree pruning and regrafting and tree bisection and reconnection. Each uses a slightly different approach to producing the rearrangement. The scope of the present paper precludes the details of each of these rearrangement types to be presented herein, but with analysis packages like PAUP* ***(19)***, they can be easily accessed in a menu-driven fashion.

3.4. Problems of Systematic and Random Error

Evaluating the error component to any analysis is always critical. In phylogenetic inference, the errors in the analysis are primarily due to either systematic error or random error. Swofford et al. ***(5)*** define random error as the deviation between a parameter of a population and an estimate of that parameter due strictly to the sample size used to make that estimate. Thus, random error disappears in an infinite sample. Systematic error is such a deviation caused by incorrect assumptions in the estimate itself, and will not only remain, but can be increased in larger samples.

For parsimony analyses, as long as the number of changes in the sequences being compared is relatively small, then given enough data, the correct phylogeny will be reconstructed. However, when the number of changes increases to the point that there are proportionately more examples of convergent or parallel evolution (increases in homoplasy), parsimony (as well as other approaches) may be less capable of discriminating homoplastic characters. This source of systematic error is probably most serious in phylogenetic trees consisting of both long and short branches ***(20)***. To avoid or at least reduce systematic error, several things can be done. Character weighting (such as differentiating between transversions and transitions as mentioned above) is routinely performed. The elimination of long branches that reflect large divergences can be difficult, but the inclusion of multiple outgroups (which have shared primitive characters) can often diminish these effects. In addition, if there are questions about positional homology, removal of these characters can reduce the problem. Finally, changing the assumptions of the analysis can also diminish systematic error.

From a practical perspective, random error affects all phylogenetic studies, since it can only be eliminated if one collects an infinite amount of data. This unrealistic approach to research can be circumvented in large part by maximizing the extraction of the phylogenetic information by using the most appropriate methods. It is also advisable to use methods that can estimate the sensitivity of the results given the number of samples that are available. Several approaches are useful toward this end: Two of the most commonly applied methods are included here.

3.4.1. Evaluating Hierarchical Structure

The removal of all random covariation in any data set is practically impossible. However, such information constitutes noise and can even lead some phylogenetic methods to choose one tree topology instead of another, although there is no real hierarchical structure in the data to support such a choice. Therefore, it is important to be able to evaluate if there is more hierarchical structure to a data set than would be expected by chance.

Permutation tests are one way of testing for hierarchical structure. From a phylogenetic perspective, they permute the data set by randomizing character states among taxa (or sequences); simultaneously they hold the number of occurrences of any particular character-state constant, which destroys any possible correlation among character-states resulting from phylogenetic signal. If a test statistic from the permuted data set is tested with a null hypothesis generated from a number of permuted data sets, then one can determine whether the null hypothesis of no phylogenetic structure is supported. If the test statistic for the data set being evaluated does not lie in one of the tails (5% level) of the null distribution, then there is a good chance that it arose in the absence of meaningful hierarchical structure ***(5)***.

Another way to test hierarchical structure in a data set is by evaluating the shape of the distribution of all possible trees (or at least a random sample of them). Hillis and Hulsenbeck ***(21)*** showed that as the amount of hierarchical structure in a data set increased, the distribution of tree lengths became more left-skewed, and concomitantly that data sets with little hierarchical structure produced more symmetrical tree-length distributions. The amount of skewness can be quantified using the g_1 statistic. When calculated, if the g_1 statistic is a negative number generally less than –0.5, there is considerable hierarchical structure to the data set.

3.4.2. Individual Branch Support: Bootstrap Analysis and Bremer Support Index

The methods for evaluation of random error discussed above deal primarily with the entire data set and are used to determine whether there is actually a phylogenetic signal or just random noise. As Hillis et al. ***(22)*** point out, "These approaches are designed with hypothesis-generating (rather than hypothesis testing) studies in mind." In other words, there is no previous hypothesis that is being tested, a reliable estimate for the phylogeny of the group is what is being tested. How can the reliability of the reconstructed branches be determined? One of a series of resampling methods, Bootstrap analysis ***(23)***, resamples data points with replacement to form pseudoreplicates of the data set. When one starts with a recovered topology (i.e., an *a priori* hypothesis), the relative number of times that a certain branch is recovered can be ascertained and the support for that branch presented on the tree (generally shown as a percentage). It is advisable to run at least 1000 bootstrap replicates (*see* **ref. *24***, for typical steroid hormone receptor analysis). The bootstrap value should be at least 85% to presume strong support for a branch.

Another approach to the problem of evaluating a branch (or a node) is to use the difference in tree lengths between the shortest trees that contain the monophyletic group that is represented on the branch versus those that do not contain the group. This assessment is called the Bremer Support (or sometimes referred to as the "Decay") Index ***(25)***. For molecular sequence data, this calculation is essentially the number of sequence changes that must occur for a branch to disappear. The greater the number, the higher the level of support for the node and resulting branches. In studies to date, it appears empirically that decay numbers of 10 or higher suggest reasonable support for a node. There is no absolute correlation between the bootstrap value and the Bremer Support

Index probably because of the different ways that these two measures of support are estimated. Thus, many authors choose to use both estimates.

4. Notes

1. Although both are representations of phylogenetic hypotheses, a cladogram is a branching diagram of relationships only, a tree emphasizing the pattern of evolution. Branch length is meaningless in a cladogram. In a phylogram, the branch lengths are proportional to the amount of evolutionary change that has occurred.
2. Most methods of phylogenetic analysis generate an unrooted tree unless directed to do differently. "Unrooted" simply refers to a tree in which the earliest point in time (the location of the common ancestor) is not identified. Outgroup analysis allows a tree to be rooted, based on the taxon (or sequence) that shared a common ancestor with a member of the ingroup most recently. The use of an outgroup taxon is generally advised.
3. Strict transversion parsimony is relatively harsh approach, carrying the presumption that there is little or no valuable information in transitions. Over long periods of divergence, there can be saturation of transitions with respect to transversions, but for recently diverged taxa (or genes), transitions can still retain a great deal of information. Thus, in many cases researchers differentially weight transversions over transitions (while these weights can be calculated in a number of ways, many researchers feel they are best estimated from the ratio of transitions to transversions present in the data set being evaluated).
4. Wagner parsimony is similar to Fitch parsimony, except that the Wagner method allows minimal constraints on character-state changes; the Fitch method allows no such constraints. Possibly the major constraint is that Wagner parsimony assumes interval data, and therefore is highly appropriate for binary and ordered multistate characters (not common in nucleotide or amino acid sequence data sets).
5. In some cases, this method (MGC parsimony) may be considered computational overkill because it pays strong attention to third-position (silent) substitutions that do not cause amino acid changes.
6. Swofford et al. ***(5)*** conclude that the computations required for the general parsimony algorithms in PROTOPARS are simplified with respect to MGC parsimony, because all potential codons that are translated into a particular amino acid are not considered nor are all of the potential synonymous codon assignments to interior nodes.
7. Heuristic methods do not always find the most optimal tree topology. They are limited by the starting tree that is being rearranged and by the order that taxa (or sequences) are added.

Acknowledgments

D. A. Ray commented on earlier versions of this manuscript. The author would also like to thank many colleagues for their patient instruction or insightful discussions on these topics over the years, specifically R. Bradley, W. Brown, H. Dessauer, D. Hillis, R. Honeycutt, A. Kluge, A. Knight, C. Moritz, R. Owen, R. Strauss, D. Swofford and P. S. White.

References

1. Hillis, D. M., Moritz, C., and Mable, B. K. (eds.) (1996) *Molecular Systematics*, Sinauer, Sunderland, MA.
2. Felsenstein, J. (1993) *PHYLIP (Phylogeny Inference Package)*, version 3.57, Department of Genetics, University of Seattle.
3. Kishino, H., Miyata, T., and Hasegawa, M. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31,** 151–160.
4. Adachi, J. and Hasegawa, M. (1992) MOLPHY: programs for molecular phylogenetics I-PROTML: Maximum likelihood inference for protein phylogeny. *Computer Science Monographs, No. 27*, Institute of Statistical Mathematics, Tokyo.
5. Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996) Phlylogenetic inference, in *Molecular Systematics* (Hillis, D. M., Moritz, C., and Mable, B. K., eds.), Sinauer, Sunderland, MA, pp. 407–514.
6. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425.
7. Altschul, S., Gish, W., Miller, W., Myers, E. W., and Lipman, J. (1990) Basic local alignment tool. *J. Mol. Biol*. **215,** 403–410.
8. Pearson, W. R. and Lipman, J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Biol. USA* **85,** 2444–2448.
9. Needleman, S. B. and 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,** 443–453.
10. Feng, D.-F. and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol*. **25,** 351–360.
11. Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci*. **8,** 189–191.
12. Hein, J. (1989) A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when phylogeny is given. *Mol. Biol. Evol.* **6,** 649–448.
13. Wheeler, W. and Gladstein, D. (1994) MALIGN: a multiple sequence alignment program. *J. Hered*. **85,** 417.
14. Fitch, W. M. (1971) Toward defining the course of evolution: minimal change for a specific tree topology. *Syst. Zool.* **20,** 406–416.
15. Camin, J. H. and Sokal, R. R. (1965) A method for deducing branching sequences in phylogeny. *Evolution* **19,** 311–326.
16. Eck, R. V. and Dayhoff, M. O. (eds.) (1966*) Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Springs, MD.
17. Goodman, M. (1981) Decoding the pattern of protein evolution. *Progr. Biophys. Mol. Biol.* **37,** 105–164.

18. Hendy, M. D. and Penny, D. (1982) Branch and bound algorithms to determine minimum evolutionary trees. *Discrete Math*. **96,** 51–58.
19. Swofford, D. L. (1999) *PAUP*: Phylogenetic Analysis Using Parsimony, version 4.0b.2*. Sinauer, Sunderland, MA.
20. Felsenstein, J. (1978) The number of evolutionary trees. *Syst. Zool*. **27,** 27–33.
21. Hillis, D. M. and Hulsenbeck, J. P. (1992) Signal, noise and reliability in molecular phylogenetic analysis. *J. Hered*. **83,** 189–195.
22. Hillis, D. M., Moritz, C., and Mable, B. K. (1996) Applications of molecular systematics, in *Molecular Systematics* (Hillis, D. M., Moritz, C., and Mable, B. K., eds.), Sinauer, Sunderland, MA, pp. 515–544.
23. Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York.
24. Xia, Z., Gale, W. L., Chang, X, Langenau, D., Patino, R., Maule, A. G., and Densmore, L. D. (2000) Phylogenetic sequence analysis, recombinant expression and tissue distribution of a channel catfish estrogen β receptor. *Gen. Comp. Endocrinol*. **118,** 139–149.
25. Bremer, K. (1994) Branch support and tree stability. *Cladistics* **10,** 295–304.