Phylogenetic analyses: a brief introduction to methods and their application

David S Horner and Graziano Pesole[†]

Phylogenetic analysis of molecular sequence data plays an increasingly important role in clinical medicine, both in the emerging field of molecular epidemiology and in the rational design of new therapeutic agents. The aims of this review are to introduce some of the methods used to construct phylogenetic trees, to illustrate some of the pitfalls that can introduce artifactual results and to speculate on the long-term importance of this area of computational biology in clinical medicine.

Expert Rev. Mol. Diagn. 4(3), 339–350 (2004)

CONTENTS

[†]*Author for correspondence*
Department of Biomolecular Sciences and Biotechnology, University of Milan, Via Celoria 26, 20133 Milano, Italy
Tel.: +39 25 031 4915
Fax: +39 25 031 4912
graziano.pesole@unimi.it

KEYWORDS:
artefacts, bioinformatics, gene annotation, molecular phylogenetics, sequence alignment, tree reconstruction

Molecular phylogenetics is a relatively new area of research usually associated with organismal, evolutionary and taxonomic studies. However, over the last 10 years, molecular phylogenetic research has moved increasingly towards center stage, as testified by the elevated number of publications in high-profile journals (e.g., 700 articles in *Nature* and *Science* since 1990 retrieved using the search term phylogeny). This statistic is partly justified by the general interest of evolutionary questions to workers in all areas of biological research. However, it also reflects a growing realization of both the importance of a sound evolutionary framework towards biological research and the possible utility of molecular systematic methodologies to highly applied fields including epidemiology, gene therapy, identification of potential drug targets, rational drug design and other problems relevant to clinical medicine. The availability of powerful computers, the provision of user-friendly analytical software and the relative ease of generation of molecular sequence data (along with the advent of whole-genome sequencing and free access to molecular sequence databases, such as Genbank) means that high-quality phylogenetic inferences can, in principle, be made by all but the most computer illiterate researchers. However, even a brief inspection of the literature reveals that inappropriate use of methods and inference of incorrect conclusions from phylogenetic trees remain widespread. Furthermore, conversations with colleagues whose area of specialization lies outside theoretical or applied phylogenetics often reveals a widespread suspicion of the black-box approach to data analysis. The intention of the current review is therefore to:

- Introduce some of the basic terminology and most popular phylogenetic methodologies
- Highlight some of the most significant methodological developments of the last 10 years
- Point out some of the remaining limitations of these methods
- Highlight, using examples, some of the applications of these methodologies in the field of clinical research

The article will progress in essentially this order, affording an introduction to terminology and basic principles for those new to the field (and offering a recap to those with some experience), followed by a slightly more technical explanation of some new developments and limitations to phylogenetic methodologies. Finally, it is hoped that the various strands will be drawn together by illustrating some current applications of the methods outlined. Potential new directions of research which may be explored in the coming years will also be illustrated.

 ISSN 1473-7159

Data selection & alignments

Selection of markers

The first consideration when embarking upon any molecular phylogenetic analysis is the selection of the molecular sequence(s) to be analyzed. The choice can be governed by any number of considerations. If the object of the analysis is to gain insight into some aspect of the evolutionary dynamics of a particular gene, the subject is obviously predetermined. If some indication of an evolutionary relationship between strains or species (e.g., of bacteria or viruses) is sought for an epidemiological study, the choice of phylogenetic markers might be more complex. Faster evolving genes or genomic regions will have a higher probability of exhibiting differences between closely related isolates, thus allowing generation of resolved phylogenetic trees with no multifurcating nodes. However, if isolates are more divergent, there is also a possibility that faster evolving genomic regions will be more difficult to amplify with PCR-based methodologies. Evolutionary information may also be lost from the sequences through multiple unseen substitutions at some sites and through the occurrence of convergent substitutions (homologous alignments in different sequences manifest identical character states as a result of two or more independent substitution events) at others. In such cases, slower evolving genes (or the analysis of inferred protein sequences, which tend to undergo less substitutions than DNA) might be more appropriate.

Selection of phylogenetic markers can also be influenced by the availability of related sequences in public databases. Maximizing sequence sampling generally increases the utility and accuracy of phylogenetic inferences. Availability of suitable sequences often reduces the quantity of new data needed for useful inference. A final consideration in the selection of sequence sampling is that the range of organisms sampled should accurately reflect the total diversity present within the range of organisms under consideration.

An important and recurring concept in phylogenetic analysis is the nature of the evolutionary relationships between a group of sequences. Sequences are said to be homologous if they have descended from a common ancestor. However, sequences can also undergo duplication events and the resulting daughter sequences often undergo subsequent functional divergence. Accordingly, sequences are considered to be orthologous if they are separated from each other on a tree only by speciation events (they are likely to perform the same role in all organisms considered). Sequences that are separated on a tree by gene duplication events (and are thus likely to have assumed different roles) are termed paralagous (FIGURE 1). This distinction is important, particularly when the object of a phylogenetic reconstruction is to establish evolutionary relationships between organisms. If paralagous genes are unknowingly considerd, recovery of incorrect species relationships is likely (FIGURE 1). Thus, gene trees are not always equivalent to species trees. Homology between sequences is an absolute condition; two sequences are either homologous or not. Sequences can share quantifiable levels of sequence identity without being homologous. However, it is important that all sequences included in a phylogenetic reconstruction are homologous (as the aim is to try and trace the evolution of sequences that share a common ancestor). Sequence homology is a necessary but not sufficient condition to carry out an evolutionary analysis. Indeed, two homologous sequences may be so divergent that they do not show any appreciable similarity, thus preventing their reliable alignment. These highly divergent sequences should be excluded in the evolutionary analysis. In practice, the observation of significant identity is generally used to establish homology. High levels of identity and, more importantly, low expectation or e-values (the probability of an equivalent match between two sequences arising by chance rather than shared ancestry) are used as the main indicators of homology [1]. In some cases, shared gene position (synteny) [2] or the presence of a gene in a conserved operon [3] can reinforce the hypotheses of homology.

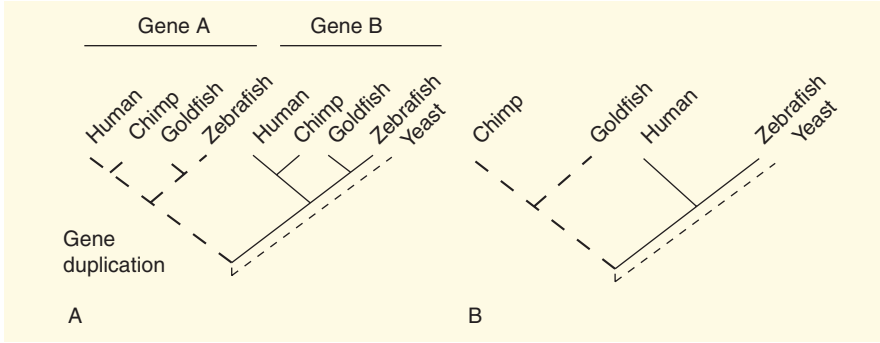

Figure 1. Homology, orthology and paralogy. (A) True phylogeny of a gene family where the tree of animal sequences is rooted using the outgroup yeast. All of the genes shown are homologous (derived from the same common ancestor). A gene duplication prior to the divergence of the lineages leading to fish and mammals leads to the formation of two paralagous gene families (A and B). Within the subfamily A, all comparisons are between orthologous genes (separated only by speciation events). (B) Phylogenetic tree reconstructed from paralagous sequences. The tree correctly illustrates the relationships between the gene sequences considered but does not give a reliable indication of species relationships within animals.

Collection of data

Recovery of reference sequences from public databases often constitutes the first practical step in a phylogeny-based project. Aligned sequences can be extremely useful in the design of primers for PCR amplification of target sequences. Conserved regions of sequence typically allow design of primers that will function for a wider range of organisms. Knowledge of available sampling is also crucial to avoid duplication of work and to direct further sampling. Sequence databases, such as Genbank, can be

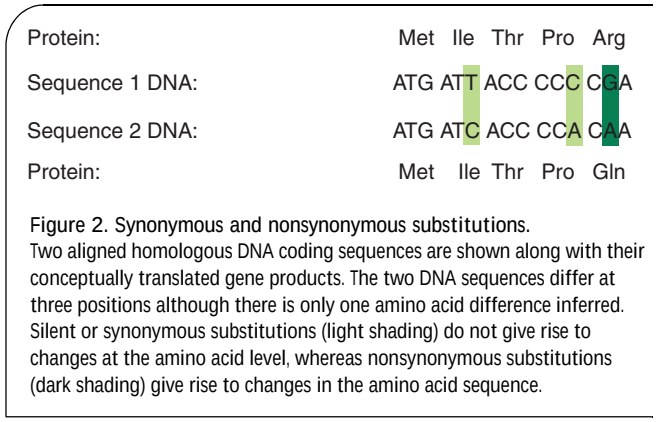

Figure 2. Synonymous and nonsynonymous substitutions.
Two aligned homologous DNA coding sequences are shown along with their conceptually translated gene products. The two DNA sequences differ at three positions although there is only one amino acid difference inferred. Silent or synonymous substitutions (light shading) do not give rise to changes at the amino acid level, whereas nonsynonymous substitutions (dark shading) give rise to changes in the amino acid sequence.

searched using either keywords such as gene and species name or using sequence similarity searches such as Blast, Fasta and their derivatives [1,4]. To compile exhaustive data sets, the use of sequence-similarity searching is strongly recommended as sequences are often incorrectly or badly annotated and thus difficult to find with keyword searches [5].

Sequence alignment

Alignment of molecular sequences is arguably the single most important step in any phylogenetic reconstruction. A multiple sequence alignment is a series of statements of homology [6]. Thus, characters at any given position in an alignment should share common ancestry and descend directly from a single position present in the ancestral sequence. When comparing sequences with low levels of divergence, identification of homologous sites is often a relatively trivial exercise. However, when more divergent sequences are considered, gaps must often be introduced into the alignment to accommodate insertion and deletion events (indels). When DNA sequences from protein-coding genes are aligned, it is best to perform alignment of the inferred amino acid sequences and then perform a codon-by-codon back-alignment of the corresponding DNA sequences (this can be achieved in a variety of sequence-alignment editors or by using on-line tools). Protein sequences are generally more conserved than their DNA coding regions due to the redundant nature of the genetic code. Many substitutions at the third codon position and several at the first codon position do not result in amino acid substitutions (synonymous substitutions). All substitutions at the second position, however, result in alterations in the amino acid sequence (nonsynonymous substitutions) (FIGURE 2). As well as increasing the accuracy of the alignment, preservation of the codon structure of the data is required to estimate the ratio of nonsynonymous to synonymous substitutions in a data set. This ratio is often used as an indicator of the nature of the Darwinian selection acting on sequences and will be discussed later [7]. In general, a low ratio of nonsynonymous to synonymous substitution rates usually indicates purifying selection, acting to prevent changes in amino acid sequences. Increases in the rate of nonsynonymous substitution usually indicates a lack of purifying selection, or indeed selection acting to encourage the generation of novel amino acid sequences (positive selection) [8].

In practice, multiple sequence alignment is usually performed using programs such as Clustal, Malign and others that are widely available on-line (TABLE 1). While these tools are extremely efficient and can save many hours of work when large numbers of sequences are considered, they often rely on heuristic methods and are quite capable of making errors. Manual checking and adjustment of automated alignments in alignments editors, such as SEAL or BioEdit (TABLE 1), is a valuable and worthwhile step.

Finally, secondary structural elements of functional RNAs and structural features of proteins are often more conserved than primary sequence features. If structural information for a biosequence is available, its use can greatly increase the accuracy of multiple sequence alignments [9–11].

Masking

Once a multiple sequence alignment has been created, it is necessary to select which positions will be used for subsequent analyses. As mentioned previously, a sequence alignment makes statements about the homology of amino acids or nucleotides present at each position. If there are gaps in the multiple sequence alignment it can be difficult to be confident that all positions are correctly aligned. Furthermore, the presence of incomplete sequences and variations in length of terminal regions of genes can mean that some alignment positions are poorly sampled, with missing data. Since it is important to include only unambiguously aligned sites [12] and because some phylogenetic tree reconstruction methods do not deal very well with gaps and missing data, it is important to mask the data, or throw away positions that are unlikely to be helpful in subsequent analyses. There are several possible approaches to this process.

- Assessment can be made manually (eyeballing the data). The risk with this approach is the tendency to include positions that agree with the expected conclusion of the researcher and omit those that do not
- Exclude all positions that have gaps in the alignment (and often those immediately flanking them as it may not be clear exactly where the insertion or deletion has occurred)
- Programs such as Gblocks, make objective assessments as to which parts of the alignment are sufficiently conserved and deemed useful for phylogenetic analyses (TABLE 1) [13]

Masking of data is a complicated and controversial part of phylogenetic analysis. There has been a philosophical objection to throwing away data from some members of the phylogenetics community. However, it is clear that in most cases masking increases the accuracy of tree reconstruction [14,15].

Trees & methods used to create them

Trees

At this point it is important to introduce a few terms and concepts relevant to the consideration of phylogenetic trees (FIGURE 3). All trees have branches, nodes and tips (or leaves). Internal nodes are the points where branches split from each other and tips are the points where the evolutionary process has yielded the sequences under analysis. In fact, tips may also be

Table 1. Some useful phylogenetics software.

Program	Function	Website	Ref.
Clustal	Multiple sequence alignment	ftp://ftp-igbmc.u-strasbg.fr	[75]
Malign	Multiple sequence alignment	ftp://ftp.amnh.org/pub/molecular	[76]
T-coffee	Multiple sequence alignment	http://igs-server.cnrs-mrs.fr/~ cnotred/Projects_home_page/t_coffee_home_page.html	[77]
SeAl	Alignment editor (Mac)	http://evolve.zoo.ox.ac.uk/software.html?id = seal	[103]
BioEdit	Alignment editor (Windows)	www.mbio.ncsu.edu/BioEdit/bioedit.html	
Gblocks	Automated alignment masking	http://monstre1.imim.es/~ castresa/Gblocks/Gblocks.html	[13]
PAUP	All DNA analyses, some protein	http://paup.csit.fsu.edu	[14]
PHYLIP	Distance, Likelihood Parsimony	http://evolution.genetics.washington.edu/phylip.html	[26]
TREE_PUZZLE	ML (Quartet puzzleing)	www.tree-puzzle.de	[27]
MrBayes	Bayesian Tree Searching	http://morphbank.ebc.uUSe/mrbayes	[32]
Modeltest	Testing DNA substitution models	http://inbio.byu.edu/Faculty/kac/crandall_lab/modeltest.htm	[41]
CONSEL	Testing tree topologies	www.is.titech.ac.jp/~ shimo/prog/consel	

More complete lists of phylogeny programs are available at [102].

considered as terminal nodes. The lengths of the branches can also indicate the degree of sequence divergence inferred to have occurred between two nodes or between a node and a tip (often expressed in number of substitutions per site). Phylogeneticists often speak of rooted and unrooted trees. The position of the root represents a common ancestor of all sequences analyzed. In reality, almost all frequently used tree reconstruction methods generate unrooted trees as the generation of rooted trees formally requires directional models of sequence substitution. Such models are computationally complex to implement and require *a priori* knowledge of directional substitution biases. In practice, trees are often rooted through the inclusion of outgroup sequences which are known *a priori* to be less closely related to the ingroup sequences than the ingroup sequences are to each other [14]. Thus, the node that connects outgroup sequences to the ingroup can be considered to be the root of the ingroup. The outgroup taxa should not be too distantly related to the ingroup, otherwise distortions of the ingroup topology can occur through long-branch attraction [16,17].

Tree reconstruction

Armed with these basic concepts, the most commonly used methods of phylogenetic tree reconstruction can now be considered. These may be classified in several ways, for example, methods that use an algorithm to generate a tree directly from the data at hand are known as algorithmic approaches; those that evaluate different possible trees and choose the best one according to a predefined set of considerations are known as optimality criterion methods.

Alternatively, tree reconstruction methodologies can be divided between distance and character-based methods. In the former, a model of amino acid or nucleotide substitution is used to construct a matrix of distances between sequences. This matrix forms the basis for the construction of a tree through clustering and grouping of sequences on the basis of their similarities. Character-based methods, such as Likelihood, Bayesian and Parsimony methods, rely on testing the fit and distribution of amino acid or nucleotide (character) changes to the tree in terms of the model of character substitution. Thus, for character-based methods, the preferred tree is the one that best explains the data with respect to the substitution model. These methods must either test all possible trees or, when the number of possible trees is large, use heuristic methods to exclude the trees which are unlikely to present good solutions. Thus, character-based methods generally correspond to optimality criterion methods, while distance methods may be algorithmic or optimality criterion based.

Distance methods & models of sequence evolution

As previously mentioned, distance methods use a model of character substitution to estimate genetic distances between all pairs of sequences (pairwise distances), then estimate a tree to best reconcile these distances. Models of character substitution, which are used as a simple measure of observed substitutions, fail to take into account superimposed multiple changes at individual sites. Models vary in complexity and can differentiate between changes perceived to be common or rare. For example, it is well known that transition substitutions (e.g., purine [A or G] to purine or pyrimidine [C or T] to pyrimidine) are more common than transversions (purine to pyrimidine or *vice versa*). All the possible interconversions between nucleotides are shown in FIGURE 4. The simplest approach, the Jukes Cantor model, assumes that all possible substitutions occur at the same rate and that the underlying frequency of occurrence of all nucleotides is equal [18]. Models of greater complexity incorporate

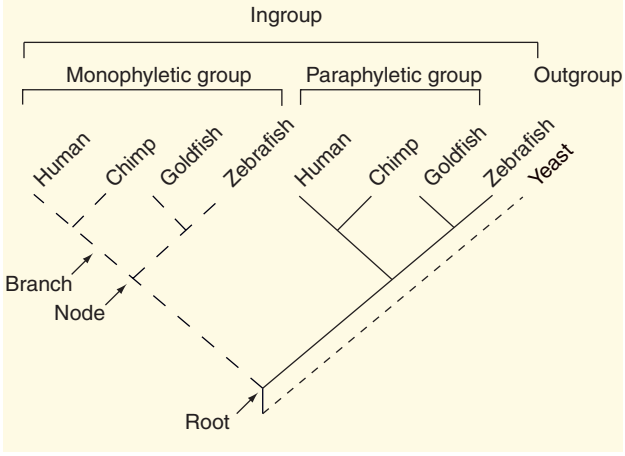

Figure 3. Tree terminology. Considering the same true phylogeny as FIGURE 1, the figure shows some commonly used terminology referring to features of the tree and groups of sequences. The yeast sequence is an outgroup, known *a priori* to be an outlier to the ingroup sequences and chosen to root the tree of the ingroup sequences. The grouping of human, chimpanzee, goldfish and zebrafish sequences is termed monophyletic as it includes all known sequences derived from a single common ancestor. However, the grouping of human, chimpanzee and goldfish sequences is termed paraphyletic as it does not include all known descendents of a common ancestor.

variation in base frequencies and provide different rates for different types of substitutions. The most complex model regularly used is the so-called general time reversible model [19,20], which allows different rates for all six possible nucleotide interconversions. The reversible part of the name implies that, for example, A will change to T at the same rate that T will change to A. Intermediate models have different numbers of variable parameters accommodating either variation in rates for different substitutions or different underlying character frequencies. These parameters can be adjusted to accommodate properties of the data set under analysis. Such models of character evolution are known as time-dependent probabilistic (stochastic) models. They rely on parameter estimates derived from a limited number of sampled sequences. Such samples can be biased and may not provide parameter estimates that accurately reflect the underlying evolutionary process. Accordingly, increased numbers of model parameters result in increased statistical errors.

In the case of amino acid sequences, the number of possible interconversions is much larger and in practice, empirical substitution matrices derived from sequences in databases are usually used to model substitution frequencies. However, it is still possible to incorporate amino acid frequencies from the data set analyzed. Many detailed reviews of nucleotide and amino acid substitution models are available [14,21–24]. For generations of nucleotide distance matrices, almost all available models are implemented in the software PAUP* [25] and DNA-dist [26]. Protein distances are best estimated with TREE-PUZZLE [27] and Prot-Dist [26].

Clustering algorithms including Neighbor Joining and Unweighted Pair Group using Arithmetic Averages (UPGMA) are used to construct trees. Briefly, these algorithms sequentially add sequences to a tree, placing them in such a way as to try to keep the distances between sequences on the tree as similar as possible to the pairwise distances calculated using the distance model. The detailed mechanics of how these algorithms function is beyond the scope of this paper, although Swofford and colleagues have provided good explanations of the methods [14].

Character-based methods

Character-based methods use sequence data more directly in selection of the optimal tree. For example, Parsimony methods select the tree that explains the data observed in terms of the minimal number of possible substitutions. Thus, maximum parsimony generally uses a simple model of sequence substitution (all changes are equally probable). Despite, or perhaps because of, its simplicity, maximum parsimony is becoming less popular. It has been shown that, particularly when more divergent sequences are used, it is prone to recovery of incorrect trees [28]. Maximum likelihood methods seek to identify the single most probable tree (according to the model of substitution employed). However, the Bayesian approaches sample possible trees with a frequency proportional to the likelihood of each tree, finally constructing a consensus of the most frequently observed nodes among the best trees. Bayesian and likelihood (probabilistic) methods have much in common with each other in the sense that both methods can, like distance methods, incorporate complex models of character evolution. However, in probabilistic methods, the parameters associated with the models can be optimized to individual tree topologies. Probabilistic methods thus seek to recover the tree topology, branch lengths and model parameters that would be most likely to generate the observed data. One great advantage of these methods is that they are based upon sound statistics and are amenable to statistical tests of model and topology fit (the relative merit of alternative models and tree topologies can be tested) [29]. Their disadvantage is that they tend to be computationally intensive to use. However, several short cuts have been devised that accelerate their execution with minimal loss of performance. For example, instead of trying to optimize all model parameters and tree topology simultaneously, it is usual to generate a starting tree (e.g., with a distance algorithm) and use this to estimate model parameters. A likelihood search is then performed and parameters are reoptimized on this tree, before further rounds of tree searching and parameter optimization. Some excellent tutorials on how to use likelihood programs, such as PAUP*, this way are available on-line, while many detailed reviews of the methodologies themselves have been published recently [29–31,101].

Number of sequences: exhaustive & heuristic tree search methods

When using character-based methods to analyze many gene sequences (more than approximately 15), the number of possible trees that need to be evaluated to fit the model and data becomes very large (for 53 sequences there are 2.75×10^{80} possible unrooted trees, a number that is bigger than the estimated number of hydrogen atoms in the universe). It is therefore

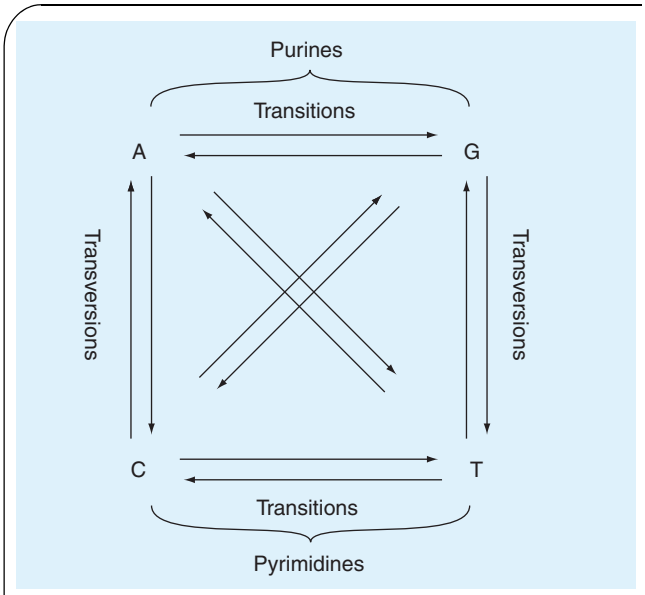

Figure 4. Nucleotide substitution types. Showing the six possible nucleotide interconversions split into transition (purine–purine and pyramidine–pyrimidine) and transversion (purine–pyrimidine interconversions). Empirical evidence indicates that in most systems, transition substitutions occur at least twice as frequently as transversions.

impossible to evaluate all trees, even under simple substitution models. A number of heuristic methods to exclude implausible trees and focus on the most probable parts of tree space have been devised. Many of these are implemented for DNA sequences in the program PAUP*. For protein sequences, the most practical approach is currently to use the Bayesian tree search methods implemented in the program MrBayes [32], or the maximum likelihood method implemented in the program PHYML [33]. These methods are not guaranteed to find the best tree (as mentioned previously, current Bayesian methods usually construct a consensus of the best trees sampled) but have been optimized to an extent that the best tree found often represents a good approximation of the best tree. In general, the most thorough methods use more time and computing power but yield better results. In fact, clustering methods used in distance analysis are also heuristic. The order of addition of distances into a clustering algorithm can have pronounced effects on the final tree topology. Multiple distance additions followed by the use of an optimality criterion (such as selection of the tree with the shorted total sum of squares of branch lengths, as implemented in the program FITCH [26]) is recommended, although not guaranteed to find the best distance tree for a set of sequences.

Testing the robustness of trees & the fit of models

Once a preferred tree topology has been identified, it is important to estimate how well it is supported, how well the model used fits the data and whether competing hypotheses of evolutionary relationships can be excluded.

The most common way to approach the first problem is through the nonparametric bootstrap [34], where trees are estimated from a large number of pseudoreplicates of the original data set. The pseudoreplicate data sets (each of the same length as the original data set) are created by random resampling (with replacement) of alignment positions such that any position might be represented any number of times (or not at all) in any resampled data set. The best tree is estimated from each data set and the proportion of resampled data sets that yield each partition of sequences (node) is calculated. Thus, the bootstrap provides a measure of how consistently the different positions in a data set support different sets of relationships between sequences. While the statistical basis for interpreting bootstrap support values is at best obscure [35], it remains the most widely used measure of tree support. The introduction of Bayesian tree reconstruction methodologies has led to the widespread use of Bayesian posterior probabilities as indicators of branch support. These values represent the frequency with which a given branch is observed in the trees sampled during the search of treespace (thus it is weighted according to the probability associated with each tree in which it was observed). Several recent studies have compared Bayesian posterior probability values with nonparametric bootstrap scores [36–39]. It is now clear that Bayesian posterior probabilities tend to be higher than corresponding bootstrap values, particularly in cases when the model of sequence substitution employed does not adequately describe the evolutionary process.

As mentioned previously, the maximum likelihood phylogenetic framework is statistically based on and thus amenable to statistical tests of tree support and model fit. Specifically, incorporation of additional substitution model parameters is always expected to improve the likelihood score of a tree. However, it is important to check that the use of additional substitution parameters is justified by significant improvements in likelihood scores. Fortunately, it is normally easy to perform such tests with simple Chi-squared tests [40] or with custom software designed for this purpose (TABLE 1) [41]. Furthermore, a series of tests that enable us to decide whether competing hypotheses of relationships (tree topologies) represent significantly worse explanations of the data (given a substitution model) have also been developed. While these tests must be used with care [42], they have now been implemented in user-friendly programs [43] and are accessible to all researchers regardless of whether they possess a strong background in statistics. It is important to implement such controls before drawing strong conclusions from phylogenetic trees. One consequence of using parameter-rich substitution models is an increase in random error of likelihood estimates and consequent widening of confidence intervals regarding parameter values (including tree topologies and branch lengths).

Methodological advances

Computational power

One development of obvious significance to all aspects of computational biology has been the event and availability of more powerful computing facilities. In fact, even sophisticated maximum likelihood phylogenetic analyses can now be performed on relatively standard desktop personal computers. Almost all of the software mentioned in this article is available in Unix/Linux, DOS/Windows or Macintosh OS versions.

Substitution models

The increase in available computing power has also prompted the implementation, and allowed the routine use of, more complex and parameter-rich models of sequence evolution, which more accurately reflect evolutionary processes. Several such advances warrant particular attention here.

The incorporation of underlying base frequencies into models of nucleotide and amino acid substitution has already been discussed. It is also evident that different amino acid or nucleotide sites within a gene or genome accumulate substitutions at different rates. This is demonstrated by the presence of highly conserved and more divergent regions seen in multiple sequence alignments or by the observation that third positions in codons evolve more quickly than first and second. Indeed, failure to take these differences into account has been shown to cause recovery of incorrect phylogenetic trees. In recent years, site-by-site rate heterogeneity has been modeled using the γ–distribution [44], the shape of which is determined by a single parameter, α. The shape parameter can be determined from the sequence data and a guide tree, for example from a distance analysis, and then incorporated into more sophisticated distance or likelihood tree reconstructions in programs such as Phylip, MrBayes, TREE_PUZZLE or PAUP* [15,25,26,32]. Alternatively, the shape parameter can be estimated concurrently with the tree in these programs. Low values of α (below 0.5) indicate a high degree of substitution rate heterogeneity between sites, while higher values suggest that most sites in an alignment are evolving at similar rates. Some sites in an alignment are also likely to be invariable due to strong evolutionary constraints. If there are a significant proportion of such sites it is important to introduce an invariable site category into the model of sequence evolution. The fraction of invariable sites can be estimated with maximum likelihood methods included in the aforementioned programs. As an alternative to the γ–distribution, which is only valid if site rate variation can be modeled accurately by a predetermined distribution, a novel and computationally simple method for the estimation of site-specific relative substitution rates has recently been described [45,46]. The relative substitution rate estimates provided by this approach can be used in the inference of trees by MrBayes.

It has also become clear that some sites are evolving at different rates in different evolutionary lineages, or even in different species. These sites are known as covariotides or covarions [47], or heterotachously evolving sites [48]. Research into this phenomenon is at an early stage and substitution models to correct for it have not yet been widely implemented.

Molecular clock

Both in evolutionary and epidemiological studies it can be desirable to estimate times of evolutionary divergences between sequences, isolates or lineages. To perform such estimates it is necessary to demonstrate that the evolutionary process is clock-like among the sequences considered, that is to say that the rate of evolution is consistent across a phylogenetic tree. For this condition to be fulfilled, it is usually considered that sequences should not be under selection or under constant selection pressure since their divergence. It is also necessary to calibrate molecular clocks with time-points of known divergences, or with sound estimates of rates of substitution (or ideally both). Molecular clock-based estimates for the divergence times of major groups of organisms have recently been proposed [49,50]. Molecular clock analysis remains a controversial area of evolutionary biology with many workers dismissing clock-based conclusions out of hand. While this point of view may be considered rather extreme, it should be stressed that dependence on the molecular clock can only be justified with extensive controls of both divergence times and substitution rates.

Limitations & vulnerabilities of phylogenetic reconstruction

There remain several evolutionary phenomena that can confound even the most sophisticated phylogenetic analysis. Indeed, there are reasons to suspect that they may particularly effect phylogenetic analysis of sequences from medically important organisms. Their manifestation can be difficult to detect and harder still to counteract. However, there are some circumstances when the vigilant researcher might suspect that phylogenetic inference will be prone to problems.

Substitutional saturation

Substitutional saturation is the term used to describe the situation when multiple unseen substitutions have occurred within sequences [28]. Multiple substitutions at a site cannot be detected by conventional pairwise comparisons. Where site-by-site substitution rate variation is extreme and where there is variation of substitution rate between lineages, even the distance correction models described earlier can fail to estimate genetic distances accurately. These considerations mean that distance methods respond badly to susbtitutional saturation and have the tendency to recover incorrect trees. The occurrence of convergent substitutions in different sequences (homoplasy) tends to confound parsimony analyses and while likelihood methods are thought to be less vulnerable to artifacts resulting from saturation, in extreme circumstances they are also prone to errors [51].

One method to detect susbtitutional saturation is to plot, for each pairwise comparison, the number of observed changes (pairwise distance) against the number of changes inferred from a phylogenetic tree (typically a parsimony or likelihood tree). If, for a significant number of comparisons, the model-corrected pairwise distance is much smaller than the number of changes seen on a tree, the sequences must be substantially saturated for substitutions [52].

Compositional biases

Sequences under analysis often have substantially different nucleotide (or amino acid) compositions. This can be caused by underlying mutational biases, a tendency to accumulate A and T residues, for example. If sequences that are relatively

distantly related share strong compositional biases to the exclusion of their true relatives, they will have a tendency to emerge as close relatives in phylogenetic analyses. As with mutational saturation, distance and parsimony methods are most vulnerable to this problem, although likelihood methods are not immune. It is important to perform tests of compositional homogeneity, for example in the program TREE_PUZZLE, to check whether this phenomenon is likely to occur.

Long-branch attraction

Even when homoplasy, mutational saturation and shared compositional biases are not prevelant, extreme variation in substitution rates between sequences (e.g., as a result of changes in selective constraints in some sequences) can result in two or more sequences being extremely divergent from others in an alignment. These sequences, which are likely to manifest long branches in phylogenetic trees, have branch lengths that reflect time since divergence and substitution rates. Such long branches have a tendency to cluster together, often in inappropriately basal positions in phylogenetic analyses [28]. This artifact (as well as the manifestation of saturation and shared composition) is often referred to as long branch attraction. All of these problems can be particularly pronounced in the presence of distant outgroups, especially if they share compositional biases with some ingroup taxa. The recovery of long branches together in basal positions is not in itself conclusive evidence that a long branch problem is present but it should present a warning sign. It is worth spending a moment or two discussing some evolutionary tendencies that seem to be shared among many parasitic (and medically important) microorganisms.

Evolutionary traits of microorganisms

Many obligately parasitic bacteria (e.g., mycobacteria and Spirochaetes) have a reduced genome size with respect to free-living relatives. This may be a consequence of adaptation to life in highly specialized environments or perhaps a consequence of population level pressures to accelerate genome replication times.

It has been noted that a tendency also exists for rapid rates of evolution to occur in parasitic organisms [53–55]. This could be partly caused by pressure to vary the sequences of antigenic proteins to avoid host immune responses. It may also occur as a result of an indirect consequence of reduced genome size (a consequently reduced number of protein–protein interactions occurring within the cell might lead to a reduction of purifying selection at some sites).

In any case, as previously mentioned, accelerated rates of evolution make accurate phylogenetic placement difficult. Some striking examples of this tendency can be seen among eukaryotic microorganisms. Only in the last 2 years have the phylogenetic relationships of the intestinal parasite *Entamoeba histolytica* become clear. For many years it was believed to be a primitive eukaryote that may have diverged prior to the endosymbiotic acquisition of mitochondria. Careful analysis of many gene sequences have now established that it is related to slime molds, such as the model organism *Dictyostelium discoidium* [56]. An even more pronounced example is provided by the microspodidia, obligate intracellular parasites associated with opportunistic infections of immune-suppressed patients. Throughout the 1980s and early 1990s they were believed to be the most pimitive of all eukaryotes [57]. However, it is now clear that despite their long branches in phylogenetic trees, they represent strange and highly modified fungi [58]. There remain many unresolved questions about phylogenetic relationships of parasitic protozoa. For example, the true placement of the flagellate *Giardia intestinalis* and its relatives is still a controversial issue [59].

Lateral gene transfer

In the past few years it has become increasingly clear that among microorganisms (bacteria, microbial eukaryotes and viruses), lateral (or horizontal) gene transfer (LGT) is prevalent [60]. The mechanisms by which genes are transferred from species to species are not well understood and are beyond the scope of this review. However, recombination between coinfecting viruses and the transfer of genomic DNA between bacteria by conjugation are clearly prevalent among these. For example, it has recently been shown that the pathogenic *Escherichia coli* 0157 strains possess up to 1387 genes that are absent from genomes of laboratory strains [61]. Strikingly, many of these genes are involved in pathogenesis and are among the likely targets of phylogenetic analyses performed within epidemiological studies. Likewise, recombination between coinfecting viruses may be responsible for the generation of much viral diversity [62]. Extensive LGT has also been demonstrated in the parasitic protozoa *G. intestinalis*, where it seems that many of the genes implicated may be determinants of pathogenicity [63]. Clearly, a phylogenetic analysis yields only a tree of the evolutionary relationships between the sequences considered. This may not represent the relationships between the organisms from which they originate. If LGT is suspected it can be useful to examine multiple genomic loci independently and compare the trees recovered (with statistical tests). One alarming possibility, particularly in the case of viral sequences, is that recombination could have occurred within the alignment analyzed. In such a case, one part of a sequence under consideration would have a different evolutionary history to the rest, clearly confounding the assumptions upon which phylogenetic analyses are based. Such cases can often be identified through a sliding window phlogenetic approach. Thus, trees are generated for overlapping subsections of the alignment and their congruence (or otherwise) used as indicators of whether (and where) recombination events might have occurred.

Expert opinion

In the future, the developing links between medical science, genomics and evolutionary analyses are expected to provide greater insights into medically relevant problems. However, an

increased dialog between medical researchers and evolutionary biologists will be necessary for these potential benefits to become reality. In the first instance this entails molecular systematists gaining familiarity with the problems and needs of epidemiologists and other medical researchers. A second prerequisite for the development of fruitful collaborations will be medical researchers gaining familiarity with the methods and approaches used by phylogeneticists. It is hoped that this article will help at least this second bottleneck.

Five-year view

As intimated at the beginning of this review, phylogenetic methodologies are finding new uses in a diverse range of biological disciplines. A few established uses will be briefly outlined and the roles that may become increasingly important in the next 5 years will be speculated upon.

Gene annotation

In the age of high-throughput whole-genome sequencing, one of the most obvious applications of phylogenetic reconstruction is in annotating the function of new gene sequences. Until now, predicted genes have typically been assigned the function of the most similar annotated sequence in the public databases. However, given the complications associated with gene duplication, changes in rates of evolution, functional constraints and automated phylogenetic analysis provides the potential to detect orthology between sequences in a more sensitive and accurate manner [64,65]. In particular, the study of evolution of gene families is highly relevant for shedding light on acquisition and specialization of specific functions during evolution and for explaining organism diversity. In this regard, bootstrapping and statistical tests of tree topology allows confidence intervals to be attached to assignments of gene function.

Epidemiology

The field of molecular epidemiology is already centered on the use of phylogenetic methodologies. To give but two examples, phylogenetic reconstruction in conjunction with molecular clock analyses has been used to argue that HIV has been present in human populations since the 1930s rather than being spread through the use of contaminated polio vaccines [66]. Indeed, HIV-1 subtypes are classified on the basis of molecular phylogenetic methodologies.

Haplotype-association studies

The association of haplotypes (closely linked and coinherited clusters of genes) with disease susceptibility loci has long been used as a tool in the identification of disease-linked genes and loci. Incorporation of cladistic informatation (trees explaining the vertical transmission of particular haplotypes) into studies exploring the association of haplotypes with disease susceptibility has been shown to increase the efficiency and sensitivity of such approaches and show great potential to facilitate identification of disease loci [67,68].

Adaptive evolution

As previously mentioned, ratios of synonymous to nonsynonymous substitutions can be used as a measure of selective pressure acting on sequences. Indeed, it is often possible to detect individual amino acids or regions of coding sequences under differential selection in different parts of phylogenetic trees [69–71]. When applied to genes, such as those encoding viral, bacterial or protozoal surface proteins, it is frequently seen that regions which constitute epitopes for host antibodies are under positive selection (selection to generate novel sequences) [72,73]. Conversely, it may well be possible to identify epitopes through the identification of such regions. This process could potentially contribute to the rational design of vaccines in the future.

Structural biology & rational drug design

The study of selective constraints, rates of substitution and even coevolution of groups of amino acids with respect to the 3D structure of proteins offers some exciting possibilities with respect to the rational design of therapeutic compounds and in the prediction of protein–protein interactions [74]. For example, an appreciation of which amino acid substitutions are associated with the development of antibiotic resistance, coupled with knowledge of the tertiary structure of the protein that metabolizes the antibiotic, might collectively allow rational design of modified compounds which will retain efficacy. The availability of large numbers of whole-genome sequences upon which to perform comparative sequence analysis and the advent of structural genomics programs, aimed at providing representatives of all naturally occurring protein folds, has enabled such approaches of drug design to become increasingly plausible.

Evolutionary developmental biology (Evo-Devo) & the evolution of phylogenetics

The study of developmental processes in an evolutionary context is becoming an increasingly popular pursuit. The evolutionary developmental biology approach offers more than the capacity to map the increasing sophistication of evolutionary processes onto phylogenetic trees. For example, correlation of duplications within gene families known to be involved in developmental processes with the advent of novel developmental processes can allow the tentative prediction of gene roles (candidate gene approaches). Like many emerging applications of molecular phylogenetics, the success of such approaches is intimately tied to the advent of high-throughput genome sequencing.

Acknowledgements

This work has been partially funded by Ministero Università e Ricerca, Italy (FIRB Project 'Bioinformatica per la Genomica e Proteomica') and Telethon. DS Horner is funded by Marie Curie category 30 individual fellowship number MCFI-2001-00634.

Key issues

- Phylogenetic reconstruction and bioinformatics analyses that incorporate evolutionary considerations are increasingly important to applied fields such as epidemiology, identification of potential drug targets and rational drug design.
- The increase in the quantity of available molecular sequence data has been mirrored by methodological developments in the field of phylogenetics. Substitution models can now describe substitution rate variation between sites, compositional biases and other eccentricities of individual data sets, while probabilistic methods allow fitting of substitution models to the data and rigorous testing of the robustness of tree topologies.
- Despite these methodological advances, inferences can only be as reliable as the data they are based on. Experimental design and systematic use of controls are vital aspects of molecular phylogenetic studies. The choice of markers, the compilation of satisfactory multiple sequence alignments and the selection of suitable analytical methods are the most critical steps in the generation of good evolutionary inferences.
- Characterization of evolutionary processes acting on specific genes or regions of genes has the potential to reveal critical functional information. For example, residues or domains under negative (purifying) selection are likely to share conserved function while those under positive selection may be associated with the generation of functional diversity or, in the case of pathogens, with evasion of the immune response. Integration of evolutionary and protein structure information promises further such insight.
- The field of molecular epidemiology is already centered on the use of phylogenetic methodologies. For example, HIV-1 subtypes are classified on the basis of molecular phylogenetic methodologies.
- The developing links between medical science, genomics and evolutionary analyses are expected to provide greater insights into medically relevant problems. However, an increased dialog between medical researchers and evolutionary biologists will be necessary for these potential benefits to become reality.

References

Papers of special note have been highlighted as:
- • of interest
- •• of considerable interest

1 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 215(3), 403–410 (1990).

2 Bharathan G, Janssen BJ, Kellogg EA, Sinha N. Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. *Mol. Biol. Evol.* 16(4), 553–563 (1999).

3 Fujibuchi W, Ogata H, Matsuda H, Kanehisa M. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. *Nucleic Acids Res.* 28(20), 4029–4036 (2000).

4 Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* 85(8), 2444–2448 (1988).

5 Karp PD, Paley S, Zhu J. Database verification studies of SWISS-PROT and GenBank. *Bioinformatics* 17(6), 526–534 (2001).

6 Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.* 16(3), 317–330 (2000).

7 Nielsen R. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86(Pt 6), 641–647 (2001).

8 Fay JC, Wu CI. Sequence divergence, functional constraint and selection in protein evolution. *Ann. Rev. Genomics Hum. Genet.* 4, 213–235 (2003).

9 Hickson RE, Simon C, Perrey SW. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Mol. Biol. Evol.* 17(4), 530–539 (2000).

10 Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* 7 (11), 2469–2471 (1998).

11 Espinosa de los Monteros A. Models of the primary and secondary structure for the 12S rRNA of birds: a guideline for sequence alignment. *DNA Seq.* 14(4), 241–256 (2003).

12 Baldauf SL. Phylogeny for the faint of heart: a tutorial. *Trends Genet.* 19(6), 345–351 (2003).

•• Excellent summary of the typical steps taken to perform a good phylogenetic analysis.

13 Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17(4), 540–552 (2000).

• Introduction to objective masking of multiple sequence alignments.

14 Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: *Molecular Systematics.* Hillis DM, Moritz C, Mable BK (Eds), Sinauer: Sunderland, MA, USA, 407–514 (1996).

•• Mathematics underlying phylogenetic reconstruction methods.

15 Morrison DA, Ellis JT. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol. Biol. Evol.* 14(4), 428–441 (1997).

16 Stiller JW, Hall BD. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol. Biol. Evol.* 16(9), 1270–1279 (1999).

• Illustrates the potential problems caused by long-branch attraction.

17 Gribaldo S, Philippe H. Ancient phylogenetic relationships. *Theor. Popul. Biol.* 61(4), 391–408 (2002).

18 Jukes TH, Cantor CR. Evolution of protein molecules. In: *Mammalian Protein Metabolism.* Munro HN (Ed.), Accademic Press, NY, USA, 21–132 (1969).

19 Rodriguez F, Oliver JL, Marin A, Medina JR. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142, 485–501 (1990).

20 Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20(1), 86–93 (1984).

21 Adachi J, Hasegawa M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42(4), 459–468 (1996).

22 Arvestad L, Bruno WJ. Estimation of reversible substitution matrices from multiple pairs of sequences. Models of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 45(6), 696–703 (1997).

23 Dimmic MW, Rest JS, Mindell DP, Goldstein RA. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* 55(1), 65–73 (2002).

24 Yu YK, Wootton JC, Altschul SF. The compositional adjustment of amino acid substitution matrices. *Proc. Natl Acad. Sci. USA* 8, 8 (2003).

25 Swofford DL. PAUP*. In: *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, MA, USA, (2002).

•• The most comprehensive software available for phylogenetic analysis of nucleotide sequences.

26 Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.5c, 1993. Distributed by the author. Department of Genetics, University of WA, USA.

•• Arguably the most comprehensive software package for phylogenetic inference of protein sequences.

27 Strimmer K, von Haeseler A. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13, 964–969 (1996).

28 Huelsenbeck JP. Is the Felsenstein zone a fly trap? *Syst. Biol.* 46(1), 69–74 (1997).

29 Whelan S, Lio P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* 17(5), 262–272 (2001).

• Good review of likelihood and Bayesian phylogenetic inference.

30 Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. *Nature Rev. Genet.* 4(4), 275–284 (2003).

• Good review of likelihood and Bayesian phylogenetic inference.

31 Huelsenbeck JP, Larget B, Miller RE, Ronquist F. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51(5), 673–688 (2002).

32 Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12), 1572–1574 (2003).

•• Allows the implementation of almost all substitution models in Bayesian phylogenetic inference.

33 Guindon S, Gascuel O. A Simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704 (2003).

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

35 Hillis DM, Bull JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192 (1993).

36 Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20, 248–254 (2003).

37 Alfaro M, Zoller S, Lutzoni F. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266 (2003).

38 Erixon P, Svennblad B, Britton T, Oxelman B. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673 (2003).

39 Cummings MP, Handley SA, Myers DS *et al.* Comparing bootstrap and posterior probability values in the four-taxon case. *Syst. Biol.* 52, 477–487 (2003).

40 Huelsenbeck JP, Rannala B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276, 227–232 (1997).

• Simple introduction to the parametric bootstrap.

41 Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14(9), 817–818 (1998).

42 Goldman N, Anderson JP, Rodrigo AG. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49(4), 652–670 (2000).

43 Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247 (2001).

44 Yang Z. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10, 1390–1401 (1993).

45 Mignone F, Horner DS, Pesole G. WebVar: a resource for the rapid estimation of relative site variability from multiple sequence alignments. *Bioinformatics* (2004) (In Press).

46 Horner DS, Pesole G. The estimation of relative site variability among aligned homologous protein sequences. *Bioinformatics* 19(5), 600–606 (2003).

47 Tuffley C, Steel M. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147(1), 63–91 (1998).

48 Lopez P, Casane D, Philippe H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* 19(1), 1–7 (2002).

49 Wang DY, Kumar S, Hedges SB. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B. Biol. Sci.* 266(1415), 163–171 (1999).

50 Hedges SB, Kumar S. Genomic clocks and evolutionary timescales. *Trends Genet.* 19(4), 200–206 (2003).

51 Hasegawa M, Fujiwara M. Relative efficiencies of the maximum likelihood, maximum parsimony and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.* 2(1), 1–5 (1993).

52 Philippe H, Adoutte A. The molecular phylogeny of eukaryota solid facts and uncertainties. In: *Evolutionary Relationships Among Protozoa.* Coombs G, Vickerman K, Sleigh M, Warren A (Eds), Chapman and Hall, UK (1998).

53 Hafner MS, Sudman PD, Villablanca FX *et al.* Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* 265(5175), 1087–1090 (1994).

54 Metzgar D, Wills C. Evolutionary changes in mutation rates and spectra and their influence on the adaptation of pathogens. *Microbes Infect.* 2(12), 1513–1522 (2000).

55 Sogin ML, Silberman JD. Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int. J. Parasitol.* 28(1), 11–20 (1998).

56 Bapteste E, Brinkmann H, Lee JA *et al.* The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium, Entamoeba and Mastigamoeba. Proc. Natl Acad. Sci. USA* 99(3), 1414–1419 (2002).

57 Kamaishi T, Hashimoto T, Nakamura Y *et al.* Protein phylogeny of translation elongation factor EF-1 α suggests microsporidians are extremely ancient eukaryotes. *J. Mol. Evol.* 42(2), 257–263 (1996).

58 Vivares CP Gouy M, Thomarat F, Metenier G. Functional and evolutionary analysis of a eukaryotic parasitic genome. A mitochondrial remnant in the microsporidian Trachipleistophora hominis. *Curr. Opin. Microbiol.* 5(5), 499–505 (2002).

59 Lloyd D, Harris JC. Giardia: highly evolved parasite or early branching eukaryote? *Trends Microbiol.* 10(3), 122–127 (2002).

60 Doolittle WF, Boucher Y, Nesbo CL *et al.* How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358(1429), 39–58 (2003).

61 Perna NT, Plunkett G III, Burland V *et al.* Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409(6819), 529–533 (2001).

62 Hay AJ, Gregory V, Douglas AR, Lin YP. The evolution of human influenza viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 356(1416), 1861–1870 (2001).

63 Andersson JO, Sjogren AM, Davis LA, Embley TM, Roger AJ. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr. Biol.* 13(2), 94–104 (2003).

64 Enault F, Suhre K, Poirot O, Abergel C, Claverie JM. Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes. *Nucleic Acids Res.* 31(13), 3720–3722 (2003).

65 Sicheritz-Ponten T, Andersson SG. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* 29(2), 545–552 (2001).

66 Korber B, Muldoon M, Theiler J *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* 288(5472), 1789–1796 (2000).

67 Seltman H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. *Genet. Epidemiol.* 25, 48–58 (2003).

68 Templeton A, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF. Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics* 156, 1259–1275 (2000).

69 Creevey CJ, McInerney JO. CRANN: detecting adaptive evolution in protein-coding DNA sequences. *Bioinformatics* 19(13), 1726 (2003).

70 Fares MA, Elena SF, Ortiz J, Moya A, Barrio E. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.* 55(5), 509–521 (2002).

71 Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18(8), 1585–1592 (2001).

72 Yang Z. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. *Pac. Symp. Biocomput.* 226–237 (2001).

73 Yang W, Bielawski JP, Yang Z. Widespread adaptive evolution in the human immunodeficiency virus Type 1 genome. *J. Mol. Evol.* 57(2), 212–221 (2003).

74 Valencia A, Pazos F. Prediction of protein–protein interactions from evolutionary information. *Methods Biochem. Anal.* 44, 411–426 (2003).

75 Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4690 (1994).

76 Wheeler WC, Gladstein DS. MALIGN: a multiple sequence alignment program. *J. Hered.* 85, 417–421 (1994).

77 Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302(1), 205–217 (2000).

Websites

101 PAUP FAQ
http://paup.csit.fsu.edu/paupfaq/paupfaq.html
(Viewed April 2004)

102 Phylogeny programs
http://evolution.genetics.washington.edu/phylip/software.html
(Viewed April 2004)

Affiliations

- *Graziano Pesole, PhD*
Department of Biomolecular Sciences and Biotechnology, University of Milan, Via Celoria 26, 20133 Milano, Italy
Tel.: +39 25 031 4915
Fax: +3925 031 4912
graziano.pesole@unimi.it
- *David S Horner, PhD*
Department of Biomolecular Sciences and Biotechnology, University of Milan, Via Celoria 26, 20133 Milano, Italy
Tel.: +39 25 031 4916
Fax: +39 25 031 4912
david.horner@unimi.it