INTERNATL MICROBIOL (2000) 3:9–16
© Springer-Verlag Ibérica 2000

REVIEW ARTICLE

David Moreira
Hervé Philippe

UPRES-A 8080, Equipe Phylogénie et Evolution Moléculaires, Université Paris-Sud, France

Molecular phylogeny: pitfalls and progress

Received 13 September 1999
Accepted 15 December 1999

Correspondence to:
David Moreira. División de Microbiología.
Facultad de Medicina. Universidad Miguel
Hernández. Campus de San Juan.
03350 Alicante. Spain.
Tel.: +34-965919451.
Fax: +34-965919457.
E-mail: David.Moreira@umh.es

Summary Molecular phylogeny based on nucleotide or amino acid sequence comparison has become a widespread tool for general taxonomy and evolutionary analyses. It seems the only means to establish a natural classification of microorganisms, since their phenotypic traits are not always consistent with genealogy. After an optimistic period during which comprehensive microbial evolutionary pictures appeared, the discovery of several pitfalls affecting molecular phylogenetic reconstruction challenged the general validity of this approach. In addition to biological factors, such as horizontal gene transfer, some methodological problems may produce misleading phylogenies. They are essentially (i) loss of phylogenetic signal by the accumulation of overlapping mutations, (ii) incongruity between the real evolutionary process and the assumed models of sequence evolution, and (iii) differences of evolutionary rates among species or among positions within a sequence. Here, we discuss these problems and some strategies proposed to overcome their effects.

Key words Phylogeny · Classification · Long branch attraction · rRNA trees · Sequence analysis

Introduction

Over almost five-sixths of its history, the Earth harbored only microbial life. Contrary to the more recent period of metazoan evolution, this vast time span left much less fossil information. The absence of a fossil record hindered the establishment of a reliable framework of the evolutionary relationships among microorganisms. In addition, and especially within prokaryotes, analysis of phenotypic traits proved of little phylogenetic value. Only the definition of low-rank taxa (species, genera, and some families) was meaningful. Attempts to generate higher-order phylogenetic schemes led to rather artificial groupings, such as the "photosynthetic bacteria". This situation was less dramatic for eukaryotic microorganisms. Some important groups could be defined on the basis of particular shared phenotypic traits, even in the rank of phyla and kingdoms, such as the ciliates, apicomplexans, or kinetoplastids. However, the elucidation of their phylogenetic relationships and the delineation of supergroups also remained largely conjecture. Thus, as recently as 1963, Stanier and co-workers sadly declared that "…the ultimate scientific goal of biological classification cannot be achieved in the case of bacteria" [45].

Shortly after, Zuckerkandl and Pauling realized that the primary sequences of nucleic acids and proteins contained a rich source of information about the evolutionary history, which could be retrieved by sequence alignment and comparison [59]. The era of molecular phylogeny was born and, with it, our approach to the analysis of microbial evolution changed. The revolutionary power of these molecular methods became obvious thirteen years later, when Woese and Fox [56] published the analyses of a significant number of organisms based on 16S ribosomal RNA (rRNA) sequences (initially as oligonucleotide catalogues and, later on, as complete sequences). The most significant outcome of this analysis was the recognition of a tripartite division of the living world. A third group encompassing some obscure weird prokaryotes (extreme thermoacidophiles, methanogens, and extreme halophiles) emerged as an independent branch in the tree of life, together with classic bacteria and eukaryotes. The scheme of the three primary kingdoms (Eubacteria, Eukaryotes, and Archaebacteria, later reclassified as the Domains Bacteria, Eucarya, and Archaea) was born. Though still controversial [11, 30, 32], this view currently prevails. Due to their universal distribution and evolutionary conservation, the rRNAs have become the reference markers for studies of molecular phylogeny. Using rRNA comparisons, the phylogenetic relationships among the major prokaryotic and eukaryotic taxa have been obtained [44, 54]. As a consequence, phylogenetic taxonomy has replaced almost completely classical taxonomic approaches, and a huge database of rRNA sequences has been generated.

However, the optimistic view that molecular phylogeny would answer most evolutionary questions was soon challenged. In fact, the subsequent use of alternative phylogenetic markers often resulted in different and conflictive pictures. Confusing results were then obtained, making it difficult to choose the "correct" phylogeny. Hence, at present, the phylogeny of many prokaryotic and eukaryotic taxa remains unsolved. Several factors that account for these incongruencies have a biological origin, such as horizontal gene transfers [31, 49] and gene duplications followed by random independent gene losses [57]. Other factors involved are tree reconstruction artifacts that arise from incorrect assumptions of the evolutionary models applied, and from the limitations of both the data available and the methods currently used for reconstructing molecular phylogenies. In this review, we summarize different problems that may cause artefactual phylogenies. We analyze their origin and suggest possible strategies to elude them or to alleviate their negative effects.

Mutational saturation and the decay of the phylogenetic signal

One of the corollaries of Zuckerkandl and Pauling's work was that sequence divergence occurs in a time-dependent manner so that the accumulation of mutations is proportional to time. This idea permeated the scientific community as the concept of an "evolutionary clock" [20] and is still applied at present. However, when phylogenies covering long-time spans are examined, sequence positions may have accumulated so many mutations that the present bases or amino acids are essentially random, and therefore contain scant or no evolutive information. These are mutationally saturated positions. A similar case may occur for shorter-time spans but fast-evolving positions or fast-evolving sequences. A clear example corresponds to the third codon positions of protein coding sequences, though mutational saturation can even affect complete sequences depending on their particular evolutionary constraints. In fact, the less a molecule is selectively constrained, the faster it will accumulate mutations and the faster the ancient phylogenetic information will change or disappear. When measured in evolutionary time scales, such fast-evolving positions or molecules exhibit a small signal-to-noise ratio and have little phylogenetic value, producing misleading phylogenetic reconstructions. How can mutational saturation be detected?

A simple strategy takes advantage of the ability of the maximum parsimony (MP) or maximum likelihood (ML) methods to correct for multiple substitutions and to estimate their approximate number. The number of substitutions thus inferred for a pair of sequences can be plotted against the number of observed differences for that pair. When this is done for all possible pairs of sequences of a data set, a saturation diagram is obtained [39]. Figure 1 shows a typical saturation diagram calculated for forty 16S rRNA sequences from diverse bacteria. The diagonal of the diagram corresponds to the ideal case of sequences that have undergone no more than one substitution per position, while the points at the right of the diagonal represent sequences that have undergone more than one substitution per position (i.e. multiple substitutions). A significant amount of points is far from the diagonal, indicating that the data set is saturated. In general, data sets corresponding to a broad taxonomic samplings are more saturated than those corresponding to more restricted taxonomic samplings. This is the reason why, usually, the uncertainty is higher for large-scale phylogenies (more saturated) than for lower-order phylogenies (less saturated). Nevertheless, note that in most cases even the MP or ML methods are not able to detect all the multiple substitutions that affect a given data set, since they use oversimplified and often unrealistic models of sequence evolution. Therefore, the number of inferred substitutions is usually an underestimation of the true number of substitutions (although much more accurate than the observed substitutions).

The problem of modeling the evolution of sequences: compositional biases, among-site rate variation, and other sources of phylogenetic inconsistency

Tree reconstruction methods make use of diverse models of sequence evolution, each having different assumptions and degrees of complexity. As commented above, the application of accurate models is crucial to estimate the number of multiple substitutions and, therefore, the problem of mutational saturation. Oversimplified models underestimate this number as well as the level of homoplasy (the fortuitous sharing of character states due

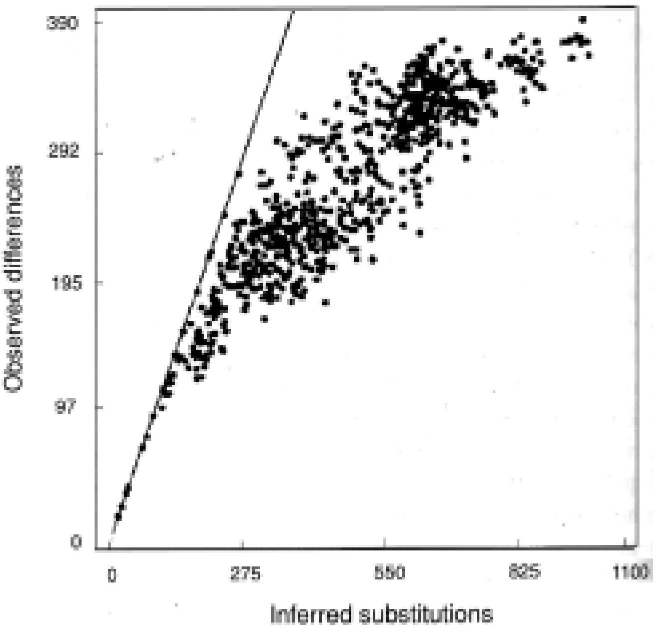

Fig. 1 Mutational saturation diagram for a data set of 16S rRNA sequences from 40 species covering almost all the bacterial taxonomic diversity. The number of inferred substitutions was estimated by maximum parsimony analysis. The diagonal corresponds to the ideal cases for which no more than one substitution per position has occurred

to back replacements or convergence and not to common ancestry) of a data set. Consequently, the similarity of nucleotide composition rather than genuine phylogenetic signal may lead in certain cases to the clustering of unrelated taxa. This was one of the first pitfalls discovered in rRNA phylogenies. For instance, the sequences of some protists that emerge early in rRNA trees turned out to have extreme G + C values, and the reliability of their phylogenetic position was questioned [27]. In contrast to nucleotide sequences, when the amino acid sequences are employed as markers the codon degeneracy supplies a mechanism that naturally absorbs to some extent these G + C drifts. By this reason, some authors consider protein sequences more reliable and suitable for phylogenetic analysis [14, 46]. However, if analyzed with an unfitted model, even the phylogenies constructed using protein sequences may be biased, in this case by amino acid composition. A clear example is the phylogeny of chloroplasts, for which the different relative amino acid compositions found in the distinct lineages constitutes an important biasing factor for tree topology [26].

Researchers have tried to overcome that problem with different approaches. The most obvious is the use of taxonomic samples whose sequences have similar nucleotide or amino acid compositions [5, 23]. However, this is not always possible and, in this case, data can be transformed to reduce the sensitivity of the analysis to compositional bias. For rRNA sequences, bases can be recoded as purine (A, G) vs. pyrimidine (C, T), and only information from transversion events used [55]. For protein sequences, amino acids can also be recoded according to groups of similar biochemical characteristics (e.g. V = I, L = M, K = R, etc.) [35]. Obviously, this approach does not solve all possible biases [24]. Several attempts have been made to develop methods that accommodate biased nucleotide and amino acid frequencies. One example is the use of paralinear distances, which produce values that are more directly comparable to standard genetic distances [21, 46].

In addition to compositional bases, another source of inconsistency arising from oversimplified models of sequence evolution is among-site rate variation. Simplistic models assume that the probability for a position to change is roughly equal for all sites in a sequence. However, some sites are often more prone to undergo changes than others, i.e. the evolutionary rate varies among sequence sites. A clear example is the presence of invariant sites, whose impact on phylogenetic reconstruction was also among the first problems detected. For instance, their inclusion in phylogenetic analyses of chloroplast evolution led to inconsistent results, even when these sites were non-informative for phylogenetic purposes due to their identity in all sequences [25]. The progressive elimination of an increasing percentage of invariant sites produced more consistent results, presumably due to a more uniform distribution of evolutionary rates among sites [25]. Yang proposed a simple model to account for invariant sites in which characters are classified into two categories “not variable” and “variable”, with the same evolutionary rate for all variable sites [58]. However, this model is still oversimplified, since the evolutionary rate can vary enormously among variable sites. A model that copes more successfully with this problem is the gamma rate distribution, which takes into account a wide range of rates. Evolutionary rates and number of positions evolving at different rates can be plotted to generate a curve. A parameter, α, of the gamma distribution accounts for the shape of this curve, and varies between 0 and infinite. If α is large, all sites evolve roughly at the same rate, whereas if it is small, many sites evolve slowly, and only a few of them evolve rapidly. The use of this model allows much more accurate distance calculations than classical models, and is now widely used [58]. A combined approach of invariant sites plus a gamma distribution, which consists of the exclusion of the invariant sites applying a gamma distribution for the variable sites, seems also very useful in some situations [47].

A method that also assumes different rates among sites was specifically developed for the analysis of rRNA sequences [52]. The availability of huge data sets for the different rRNAs has allowed the construction of variability maps for these molecules. The degree of variability of each site is estimated from the aligned sequences. Then, for the inference of phylogenetic trees, a particular weight is given to each site which is inversely proportional to the variability of the site. Slow-evolving sites thus have more weight than fast-evolving ones. This method was successfully used to identify the sister-group of the fast-evolving nucleomorphs, the secondary photosynthetic endosymbionts of the chlorarachniophytes [51].

The search for more accurate models is a very active area of research in phylogenetics. New approaches based on more realistic assumptions are under development. Among these, the covarion model which is most promising. Fitch advanced this model in the early 1970s by stating that the evolutionary rate of each position can vary through time [7]. This possibility is not currently considered by the standard among-site rate variation models, making them sometimes inefficient. Any given site may be variable in some parts of the tree but constant in others. This may explain the differences in the distribution of invariant sites or in the number of variable sites among different lineages. At present, it is probably the most realistic approach to model the non-stationarity of the evolutionary process. In fact, the evolution patterns of several molecules, such as the rRNAs [53] or the elongation factors [28], seem to fit a covarion model. However, the mathematical implementation of this model is very difficult, and tree reconstruction methods based on this model are not yet available [50].

The long branch attraction artifact

Soon after molecular phylogeny techniques become widespread, Felsenstein described a major artifact in the tree reconstruction: the long branch attraction (LBA), which was observed whenever the evolutionary rates were different among the different species [6]. Fast-evolving sequences are more prone to share identical character states by chance (false synapomorphies) than slow-evolving sequences. In the absence of adequate models of sequence evolution, this problem is not corrected, and the fast-evolving sequences will be grouped together irrespective of their true relationships. The parameter combination within which a particular method results are inconsistent because of LBA is referred to as the "Felsenstein-zone" [48]. All current models are more or less oversimplified and, as a consequence, LBA is a very common phenomenon whenever differences of evolutionary rate among sequences exist. An additional problem is that the outgroup is frequently itself a long branch, so that the fast-evolving ingroup sequences are not only attracted among themselves but also by the outgroup. This leads to their misplacement towards the base of the tree. Mutational saturation exacrebates the problem and, if the differences of rate among species are high, the LBA can produce robust misleading phylogenies [40]. Rate differences may correspond to several orders of magnitude [36]. These artifacts are a major source of uncertainty in current molecular phylogeny. Our view of the evolution of certain groups has been unfortunately influenced through unrecognized and incorrect phylogenies resulting from LBA.

An excellent example is the phylum Microsporidia. These protists are unusual and divergent parasites that lack mitochondria and branch at the base of the eukaryotic small subunit (SSU) rRNA tree [23]. Accordingly, for a long time they were considered as living relics of the premitochondrial phase of eukaryote evolution. However, genes of unambiguous mitochondrial origin have been recently discovered in their genomes [8, 18]. Furthermore, several protein markers place them as close relatives of fungi [reviewed in 34]. This relationship is also supported by the presence of chitin in their spore wall and by similarities in their life cycle to that of several fungi [34]. Therefore, these protists are actually a group of fungi that have experienced a strong increase in the evolutionary rate of some genes (such as their SSU rRNA). This is most likely due to their parasitic way of life, characterized by a reduction of metabolic activity and selective constraints, and major changes in the structures of populations [34]. The increased evolutionary rate explains their basal phylogenetic position as a result of an LBA artifact. Another example is found in the bacterial group comprising the thermophilic genus *Thermus* and the radio-resistant genus *Deinococcus*, which branches early in the SSU rRNA bacterial tree [54]. According to Gupta, recent protein data analyses, including the detection of some specific deletions and insertions, suggest that these genera are actually close relatives of the cyanobacteria [12]. In this case, the force leading to the acceleration of the evolutionary rate might have been the adaptation to the harsh environments inhabited by these organisms.

Long branch attraction, tree shape, and the effects of taxonomic sampling

A simple parameter used to describe tree topology is symmetry. A symmetric tree is perfectly bifurcated, whereas an asymmetric tree is ladder-shaped (Fig. 2). The tree symmetry is measured by an index ranging from 0, for a completely symmetric tree, to 1, for a completely asymmetric tree [16]. Since the LBA artifact tends to attract the fast-evolving taxa out of their actual phyletic groups towards the base of the tree, it gives rise for more asymmetric topologies (Fig. 2). This effect is stronger the more distant the outgroup. In fact, when very distant outgroups are used, the fast-evolving taxa usually appear as independent branches at the base of the tree whereas, when close outgroups are used, the fast-evolving taxa often are artefactually grouped together at the base of the tree [40]. Accordingly, very asymmetric trees may suggest the existence of LBA artifacts.

One important consequence of the weakness of the phylogenetic signal due to the combination of the mutational saturation and the unequal evolutionary rates among species is the taxonomic sampling effect [22]. A priori, the relative branching order of the taxonomic groups in a phylogeny should not vary depending on the particular species of these groups included in the analysis. However, very often this is not the case, especially when small data sets are used [38]. For instance, the accepted phylogeny of the α, β, γ, and δ subdivisions of the Proteobacteria displays the topology shown in Fig. 3A, where the β and γ subdivisions are sister-groups. However, by using particular representatives of these subdivisions, phylogenetic trees where the α and β subdivisions are sister-groups can be obtained (Fig. 3B). Underlying these contradictory phylogenies is, once again, the existence of unequal evolutionary rates among species, which produces LBA artifacts. Thus, any group represented by a fast-evolving species is likely to show an equivocal early position in the tree, often misleading the relationships among the other groups.

Can long branches be detected?

An inherent problem in the misplacement of the fast evolving sequences towards the base of the tree by LBA is that these branches are not easily recognized as excessively long (Fig. 2). A possible solution is the a priori analysis of the evolutionary rates of the different sequences in order to detect the fast evolving ones. A mathematically simple test, the relative rate test, was developed for this purpose [42]. It consists of the comparison of the distances of two sequences to a third one, the outgroup. If both distances are equivalent, then it is assumed that the two sequences have similar evolutionary rates. If distances are dissimilar, the most divergent sequence is considered the fastest evolving one. However, it has been recently demonstrated that this test has severe limitations when applied to saturated data sets. Saturation leads to a more severe underestimation of longer pairwise distances relative to shorter distances [9] and, therefore, the number of differences observed between pairs

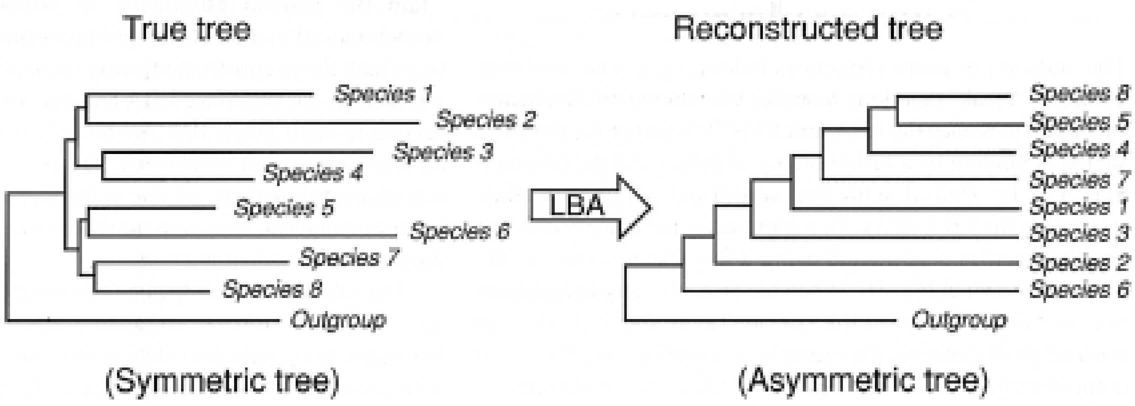

Fig. 2 Tree symmetry and the effects of long branch attraction (LBA). A perfectly bifurcated or symmetric tree (left) may be artefactually retrieved by molecular phylogenetic analysis as a ladder-shaped or asymmetric tree (right) if there are strong differences of evolutionary rate among species. The fast-evolving species are misplaced towards the base of the tree decreasing the tree symmetry

of sequences will not reflect the true number of substitutions. Hence, although their rates may be very different, the result of the relative rate test will be always that these sequences have similar rates. As a result, saturated data sets suggest a false clock-like behavior, especially when using distant outgroups [40].

Alternative methods to cope with the problem of detecting fast-evolving species have been developed. Perhaps the most popular is the use of improved methods for the correction of distances that allow more accurate estimations of the number of substitutions. Several correction methods, based on diverse assumptions about the process of sequence evolution, are currently used. Although their detailed analysis falls out of the scope of this work, it is worth noting that the trees retrieved under the most accurate corrections for a given data set show most clearly the differences of evolutionary rates (i.e. the differences of branch length) among species. For instance, a simple method based on both the removal of the constant positions of the alignment and the construction of the phylogeny using only the variable positions, can be most useful to detect the fast evolving species [25]. However, at present, the most often used approach is the correction for among-site rate variation using a gamma law (see above), since the existence of very dissimilar substitution rates for different positions in a sequence appears ubiquitous [58]. Although the estimation of the correct parameters of the gamma distribution is often problematic, this method has allowed the reconstruction of phylogenies for which the classical methods produced robustly wrong trees [47].

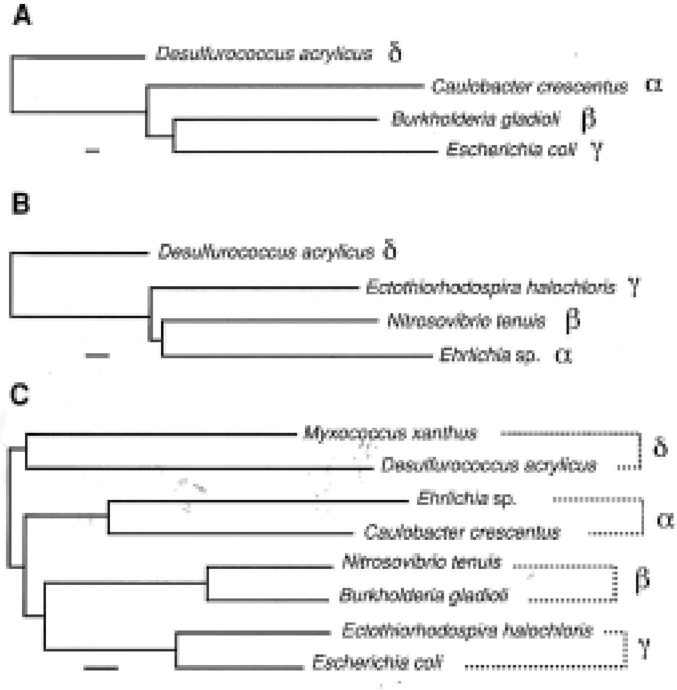

Fig. 3 The effect of taxonomic sampling. (A) A correct 16S rRNA four-species tree where β and γ proteobacterial representatives are sister-groups. (B) An incorrect tree where α and β proteobacterial representatives are sister-groups. (C) The addition of sequences to the incorrect phylogeny allows the retrieval of a correct tree. The bars correspond to 10 nucleotide substitutions

Recently, a new method for the detection of fast evolving sequences, the RASA test, based on statistical measure of phylogenetic signal in character state matrices, has been developed [29]. However, this method is also endowed with important limitations. Slow–evolving sequences are interpreted as fast–evolving ones whenever they are included in data sets with a majority of actually fast–evolving sequences (unpublished observations). A very simple alternative approach that has been proven useful consists of the determination of the number of variable positions for the different taxonomic groups analyzed. Contrary to the current assumptions, the acceleration of the evolutionary rate is due not only to the acceleration of the substitution rate of the variable positions, but also to an increase in the number of variable positions. Therefore, fast–evolving groups are usually characterized by a larger number of variable positions [33]. This amplifies the effect of the LBA artifact, but in turn allows the identification of fast–evolving groups.

How to cope with the long-branch attraction

The addition of more sequences belonging to fast-evolving groups to "break" their long branches was among the first means suggested to reduce the LBA artifact [17]. Some of the positions shared at random by a fast-evolving sequence and the outgroup may not be shared with the additional sequences, thus decreasing the LBA effect. This approach is particularly efficient for phylogenies constructed using a few species, such as the example shown in Fig. 3B. When the proteobacterial phylogeny is constructed using all the species (even those producing mislead phylogenies in the examples containing only four taxa) a correct tree is obtained (Fig. 3C). However, increasing the number of sequences may not completely overcome LBA problems, and may even decrease the consistency of the phylogenies [19]. An example is the protist phylogeny based on elongation factor 1α (EF-1α) sequences, which shows an artefactual polyphyly of ciliates. The addition of several new ciliate sequences did not allow the retrieval of the monophyly of ciliates, but did lead to a decrease in the statistical support in several nodes of the tree [33]. Note that this decrease affected especially the least reliable parts of the tree (e.g. those containing the polyphyletic fast-evolving ciliate branches). This reveals that the decrease of support subsequent to an addition of sequences may be an indicator of artifacts in phylogenies. Statistically well-sustained reliable nodes are less affected by this decrease [33]. Note here that a consistent node is not necessarily a correct node. Several tree reconstruction artifacts, notably the LBA, can produce consistent but incorrect topologies. For instance, the very early emergence of Microsporidia in rRNA trees is consistently supported by high bootstrap values [23]. However, this result is artefactual since Microsporidia are actually close relatives of fungi (see above).

The most effective approach to the study of large numbers of taxa is to focus on slow-evolving sequences [19]. A noticeable example comes from the metazoan phylogeny constructed by using the SSU rRNA, for which only the use of slow-evolving species allows the location of nematodes as sister-group of arthropods and not as early-emerging animals [1]. However, the lack of a priori criteria is a hindrance to foresee the evolutionary rate of any given sequence. The only practical approach to solve this problem is to increase the number of sequences with the hope that this will also include slow-evolving sequences. Careful selection of these among all available sequences can be of great help. At any rate, if slow-evolving sequences are not available, the best approach is the use of as many sequences as possible.

Another major variable that researchers can modify in several cases is the distance to the outgroup. Since the LBA produces the attraction of the fast-evolving taxa towards the base, the choice of the closest outgroup is highly recommended to reduce the impact of this artifact. In addition, the observation of changes of the tree topology related to changes in the outgroup is an evidence of the existence of LBA artifacts. For instance, if we consider again the general phylogeny of eukaryotes, the use of mitochondrial markers often produces much more symmetric trees than those constructed using nuclear sequences [see, for instance, 8, 10, 18]. Since LBA and tree asymmetry are directly related to each other, that asymmetry reveals LBA artifacts affecting eukaryotic phylogeny. The results obtained by using mitochondrial markers are due to the fact that the outgroup, the α-Proteobacteria, seems to be closer to the ingroup than the outgroups for nuclear markers.

The selection of an adequate tree reconstruction method can also help to avoid or alleviate LBA problems. Different reconstruction algorithms do not have the same ability to cope with fast-evolving sequences. In fact, ML methods are the most effective, whereas MP methods seem to be the most sensitive when a small number of taxa is used [13]. However, ML methods are very computationally expensive, which makes their use with large numbers of taxa almost impossible. This is a practical problem, since large data sets usually produce more reliable results.

Our group developed recently a simple method (the Slow-Fast method) that allows the detection of fast-evolving species and provides information about their position in phylogenetic trees [3]. The method is based on the reconstruction of phylogenies by using only the slow-evolving positions of the alignment for each taxonomic group. The number of substitutions is computed for each position within each taxonomic group. Subsequently, different distance matrices (the "Slow-matrices") are built using the information from positions which undergo either no changes, or one, two or even more changes per group. A priori, the phylogenetic information contained in these slow-evolving positions is of the best quality, especially for ancient events. However, this information is frequently overwhelmed by the mutationally saturated fast-evolving positions, which are usually the majority of positions in alignments. By lending more weight to positions containing the most information, the Slow-Fast method facilitates a desaturation of the phylogenetic signal. It has been successfully applied to the problem of the evolution of eukaryotic phyla, revealing that the apparently early-emerging eukaryotic groups are actually fast-evolving, and have been misplaced towards the base of the tree by an LBA artifact. The emerging picture is that, instead of the classical step-by-step model of eukaryotic evolution, all the major eukaryotic phyla appeared within a comparatively short time (the "Big Bang hypothesis") (unpublished data).

Alternative molecular approaches to infer phylogenetic relationships

The problems intrinsic to phylogenetic reconstruction derived from sequence comparison have prompted research on other kinds of molecular markers. The most common approach is to look for rare events, such as insertions or deletions (known as "indels") [2, 11, 15], intron position [41] or retropositional events [43]. In some cases these markers reveal the correct phylogeny whereas the comparisons of the gene sequences containing them produce incorrect inferences. Once again, the Microsporidia are a good example: they appear misplaced at the base of the tree constructed by using the EF-1α sequences, but they share an insert in the EF-1α sequences with fungi and metazoans, which suggests their phylogenetic proximity to these organisms [4]. Nevertheless, even these markers are not absolutely reliable since their information may be biased by different factors. For instance, very short indels should be regarded with caution since they tend to show erratic phylogenetic distributions, as attested by a two amino acid insertion present in the EF-1α of several eukaryotic species that are phylogenetically unrelated (unpublished data). Another source of uncertainty is the possibility of horizontal gene transfer between distantly related organisms. For instance, some Archaea and the Gram-positive bacteria share a specific insertion in the chaperonin Hsp70, which has been claimed as an evidence of their phylogenetic relationship [11]. However, a detailed analysis of the distribution of Hsp70 in archaea and of the phylogeny of the Hsp70 protein family strongly suggests that this is actually a case of horizontal transfer from Gram-positive bacteria to archaea [37].

Conclusions

There have been many advances since Zuckerkandl and Pauling first suggested the use of sequence comparison to infer the evolutionary relationships among organisms. Molecular phylogeny is now one of the most dynamic and rich fields of biology. However, the increasing sophistication of techniques reveals the existence of an important number of structural and computational artifacts. As a consequence, molecular phylogenies should be interpreted with caution, especially when they refer to very long time spans [40] or when other types of information (such as detailed morphological data, or the fossil record) are missing, as occurs for the phylogeny of most microorganisms.

Criteria commonly applied, such as the congruence of the results obtained using different methods, are insufficient to ascertain the accuracy of a given phylogeny, since under very frequently seen conditions (high mutational saturation and evolutionary rate differences among species) the different methods may all be positively misleading [6]. Therefore, it is crucial to choose the most adequate method to analyze a given data set, as well as to apply statistical methods to assess the confidence of phylogenies. Methods providing a large number of alternative trees (such as MP and ML methods) are especially valuable since they allow different statistical analyses and hypothesis testing. In fact, if very different tree topologies are statistically compatible, the selection of a particular tree among the rest can be justified if other sources of information support it. In this regard, the congruence of the results obtained by using different markers (mostly by using markers involved in different cellular functions and, therefore, less prone to coevolution) is especially significant.

Acknowledgments We thank Purificación López for critical reading of the manuscript, and the members of the Equipe Phylogénie et Evolution Moléculaires of the Université Paris-Sud for continuous and fruitful interaction. D.M. has been supported by fellowships from the Spanish Ministerio de Educación y Cultura and the Fondation des Treilles.

References

1. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493
2. Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90:11558–11562
3. Brinkmann H, Philippe H (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artefacts in ancient phylogenies. Mol Biol Evol 16:817–825
4. Embley TM, Hirt RP (1998) Early branching eukaryotes? Curr Op Genet Develop 8:624–629
5. Embley TM, Thomas RH, Williams RAD (1993) Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. System Applied Microbiol 16:25–29
6. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410
7. Fitch WM (1971) The nonidentity of invariable positions in the cytochromes *c* of different species. Biochem Genet 5:231–241
8. Germot A, Philippe H, Le Guyader H (1997) Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. Mol Biochem Parasitol 87:159–168
9. Gojobori T, Ishii K, Nei M (1982) Estimation of average number of nucleotide substitutions when the rate of substitions varies with nucleotide. J Mol Evol 18:414–423
10. Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. Science 283:1476–1481
11. Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. Microbiol Mol Biol Rev 62:1435–1491
12. Gupta RS, Johari V (1998) Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the *Deinococcus-Thermus* group and cyanobacteria. J Mol Evol 46:716–720
13. Hasegawa M, Fujiwara M (1993) Relative efficiencies of maximum likelihood, maximum parsimony, and neighbour joining for estimating protein phylogeny. Mol Phylogenet Evol 2:1–5
14. Hasegawa M, Hashimoto T (1993) Ribosomal RNA trees misleading? Nature 361:23
15. Hashimoto T, Sanchez LB, Shirakura T, Müller M, Hasegawa M (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. Proc Natl Acad Sci USA 95:6860–6865
16. Heard SB (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46:1818–1826
17. Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38:297–309
18. Hirt RP, Healy B, Vossbrinck CR, Canning EU, Embley TM (1997) A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. Curr Biol 7:995–998
19. Kim J (1996) General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing number of taxa. Syst Biol

45:363–374
20. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England
21. Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91:1455–1459
22. Lecointre G, Philippe H, Lê HLV, Le Guyader H (1993) Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol 2:205–224
23. Leipe DD, Gunderson JH, Nerad TA, Sogin ML (1993) Small subunit ribosomal RNA of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. Mol Biochem Parasitol 59:41–48
24. Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D (1999) Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol 16:573
25. Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA 93:1930–1934
26. Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Howe CJ (1998) A covariotide model describes the evolution of oxygenic photosynthesis. Mol Biol Evol 15:1183–1188
27. Loomis WE, Smith DW (1990) Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc Natl Acad Sci USA 87:9093–9097
28. Lopez P, Forterre P, Philippe H (1999) The root of the tree of life in the light of the covarion model. J Mol Evol. 49:496–508
29. Lyons-Weiler J, Hoelzer GA, Tausch RJ (1996) Relative apparent synapomorphy analysis (RASA). I: The statistical measurement of phylogenetic signal. Mol Biol Evol 13:749–757
30. Margulis L, Guerrero R (1991) Kingdoms in turmoil. New Scientist 1761:46–49
31. Martin W (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 21:99–104
32. Mayr E (1998) Two empires or three? Proc Natl Acad Sci USA 95:9720–9723
33. Moreira D, Le Guyader H, Philippe H (1999) Unusually high evolutionary rate of the elongation factor 1a genes from the Ciliophora and its impact on the phylogeny of eukaryotes. Mol Biol Evol 16:234–245
34. Müller M (1998) What are the Microsporidia? Parasitol Today 13:455–456
35. Naylor GJ, Brown WM (1997) Structural biology and phylogenetic estimation. Nature 388:527–528
36. Pawlowsky J, Bolivar I, Fahrni JF, de Vargas C, Gouy M, Zaninetti L (1997) Extreme differences in rates of molecular evolution of Foraminifera revealed by comparison of ribosomal DNA sequences and the fossil record. Mol Biol Evol 14:498–505
37. Philippe H, Budin K, Moreira D (1999) Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family. Mol Microbiol 31:1007–1010
38. Philippe H, Douzery E (1994) The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. J Mammal Evol 2:133–152
39. Philippe H, Sörhannus U, Baroin A, Perasso R, Gasse F, Adoutte A (1994) Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J Evol Biol 7:247–265
40. Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Op Genet Develop 8:616–623
41. Qiu Y-L, Cho Y, Cox JC, Palmer JD (1998) The gain of three mitochondrial introns identifies liverworts as the earliest land plants. Nature 394:671–674
42. Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144–1147
43. Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Khisiro T, Goto M, Munechika I, Okada N (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388:666–670
44. Sogin ML (1991) Early evolution and the origin of eukaryotes. Curr Op Genet Develop 1:457–463
45. Stanier RY, Doudoroff M, Adelberg EA (1963) The microbial world. 2nd edn. Prentice-Hall, Englewood Cliffs
46. Steel MA, Lockhart PJ, Penny D (1993) Confidence in evolutionary trees from biological sequence data. Nature 364:440–442
47. Sullivan J, Swofford DL (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J Mammal Evol 4:77–86
48. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer, Sunderland, Massachussetts, pp 407–509
49. Syvanen M (1994) Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet 28:237–261
50. Tuffley C, Steel M (1998) Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63–91
51. Van de Peer Y, Rensing SA, Maier UG, de Wachter R (1996) Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc Natl Acad Sci USA 93:7732–7736
52. Van de Peer Y, Van der Auwera G, de Wachter R (1996) The evolution of stramenopiles and alveolates as derived by "substitution rate calibration" of small ribosomal subunit RNA. J Mol Evol 42:201–210
53. Waddell PJ, Penny D, Moore T (1997) Hadamar conjugations and modelling sequence evolution with unequal rates across sites. Mol Phylogenet Evol 8:33–50
54. Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
55. Woese CR, Achenbach L, Rouviere P, Mandelco L (1991) Archaeal phylogeny: re-examination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artefacts. Syst Appl Microbiol 14:364–371
56. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090
57. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713
58. Yang Z (1996) Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol 42:294–307
59. Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary

history. J Theor Biol 8:357–366